Made available by Hasselt University Library in https://documentserver.uhasselt.be

Inference of the generalized-growth model via maximum likelihood estimation: A reflection on the impact of overdispersion

Peer-reviewed author version

GANYANI, Tapiwa; FAES, Christel & HENS, Niel (2020) Inference of the generalized-growth model via maximum likelihood estimation: A reflection on the impact of overdispersion. In: JOURNAL OF THEORETICAL BIOLOGY, 484 (Art N° 110029).

DOI: 10.1016/j.jtbi.2019.110029
Handle: http://hdl.handle.net/1942/30494

Inference of the generalized-growth model via maximum likelihood estimation: a reflection on the impact of overdispersion

Tapiwa Ganyani[a,*], Christel Faes[a], Niel Hens[a,b]

[a] *Interuniversity Institute for Biostatistics and statistical Bioinformatics, UHasselt (Hasselt University), Diepenbeek, Belgium*
[b] *Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium*

Abstract

Recently, the generalized-growth model was introduced as a flexible approach to characterize growth dynamics of disease outbreaks during the early ascending phase. In this work, by using classical maximum likelihood estimation to obtain parameter estimates, we evaluate the impact of varying levels of overdispersion on the inference of the growth scaling parameter through comparing Poisson and Negative binomial models. In particular, under exponential and sub-exponential growth scenarios, we evaluate, via simulations, the error rate of making an incorrect characterization of early outbreak growth patterns. Simulation results show that the ability to correctly identify early outbreak growth patterns can be affected by overdispersion even when accounted for using the Negative binomial model. We exemplify our findings using data on five different outbreaks. Overall, our results show that estimates should be interpreted with caution when data are overdispersed.

Keywords: generalized growth model, sub-exponential growth, early epidemic growth phase, maximum likelihood estimation, overdispersion.

1. Introduction

Over the last few decades, mechanistic models of infectious diseases have been helpful to gain insights about key disease transmission parameters as well as to estimate the

[*] Corresponding author
Email address: tapiwa.ganyani@uhasselt.be (Tapiwa Ganyani)

evolution of epidemics (1). These models typically study disease outbreak data based on the assumption that the early epidemic phase follows exponential growth dynamics in the absence of susceptible depletion or intervention measures (2). Nevertheless, slower-than-exponential (sub-exponential) growth patterns have been observed for infectious diseases such as, measles, HIV/AIDS, foot and mouth disease, Ebola and influenza (see for example, (3; 4)). This phenomenon can occur due to, among other reasons, clustering of infectives or reactive behavioural changes which tend to reduce the number of contacts per unit time, transmission route (contact versus airborne) and spatial heterogeneity (see for example, (3; 5; 6; 7)).

Accordingly, more detailed mechanistic models (e.g., models with time varying contact rates, metapopulation models and network models) have been developed to describe sub-exponential growth (see review in (7)). Nevertheless, potential causes of this phenomenon are often difficult to identify and quantify and hence to incorporate in a model (2). In (3), a simple phenomenological model known as the generalized-growth model (GGM) was introduced to model incidence counts. It allows relaxing the exponential growth assumption using a growth scaling parameter that is able to capture constant growth ($p = 0$), sub-exponential growth ($0 < p <1$) and exponential growth ($p = 1$).

An accurate quantification of departure from exponential growth dynamics is essential for public health decision-making as it can inform the likely “signature feature” of the threat level (3). For example, in (8) researchers use incidence data of the 2014-16 West Africa Ebola outbreak to demonstrate that the growth scaling parameter can be a useful indicator of final epidemic size, which may have significant implications to guide outbreak control. Also, accurate knowledge about the extent of departure from exponential growth dynamics can be important when developing detailed epidemic models which explicitly capture transmission dynamics; depending on the magnitude of the growth scaling parameter, several potential pathways likely to contribute to slower than exponential growth can be incorporated and validated. Moreover, when the "true" underlying epidemic growth pattern is slower than exponential, models that assume exponential growth dynamics may overestimate the evolution of the epidemic; as such, incorporating departures from exponential growth may be important in order to obtain more accurate estimates (3; 9; 10; 11; 12; 13).

Estimation of GGM parameters has proceeded mainly via the least-squares method (ordinary nonlinear regression) (3; 9; 10; 11; 12; 13). It is well known that this method may not perform well when data are heteroscedastic; it may lead to different conclusions about the same data set compared with a method of estimation which allows heteroscedasticity (see for example, (14; 15; 16)). In this work we use classical maximum likelihood (ML) estimation to estimate GGM parameters. ML methods are the standard inference approach in the statistical literature owing to their useful statistical properties, as for instance, consistency (true parameter value that generated the data recovered for sufficiently large samples) and efficiency (lowest-possible variance of parameter estimates achieved asymptotically) (see for example, (16; 17)). Also, ML methods can be easily applied to simple statistical models which allow for heteroscedasticity. Moreover, as count data often show overdispersion (see for example, (18; 19; 20)), through the choice of an appropriate statistical model, ML methods can offer a direct way to determine levels of overdispersion present in the data. The first aim of this study, therefore, is to demonstrate the application of classical ML methods to estimate GGM parameters. Due to the count nature of incidence data, we formulate the GGM as a generalized nonlinear model (GNM) (18) and fit it using a Poisson model (assuming no overdispersion) and a Negative binomial (NB) model (accommodating overdispersion).

The Poisson distribution assumes mean and variance are equal and can yield misleading conclusions when variability exceeds the mean; in the latter case, the NB distribution performs better (see for example, (18; 19; 20)). As such, it is imperative to study the estimation of GGM parameters (particularly the essential parameter, p) when data are overdispersed. The second aim, therefore, is to compare inference of the growth scaling parameter using these two models both when the mean and variance are equal as well as in the presence of varying levels of overdispersion. In particular, we make the comparison under the scenarios of exponential as well as sub-exponential growth. We study *Type I error* which we define as concluding sub-exponential growth in an exponential growth scenario as well as *Type II error* which we define as concluding exponential growth in a sub-exponential growth scenario. We illustrate our analyses using simulated data and exemplify findings using data on five different outbreaks.

2. Methods

2.1. Generalized growth model

The GGM is given by the differential equation:

$$\frac{dC(t)}{dt} = C'(t) = r * C(t)^p, \tag{1}$$

where $r * C(t)^p$ is the incidence curve at time t; $C(t)$ is cumulative number of cases at time t, r is the growth rate, and $p \in [0, 1]$ is a growth scaling parameter. If $p = 0$, the model describes constant incidence over time and the cumulative number of cases grows linearly; if $p = 1$, the model describes an exponential growth pattern and; $0 < p < 1$ describes a sub-exponential growth pattern (3).

In real life, epidemic data are always observed in discrete time intervals rather than continuous time (e.g. daily, weekly, monthly) and they reflect aggregated information between consecutive reporting periods. Hence, a discrete approximation of (1) can be formulated as,

$$\frac{C(t+h) - C(t)}{h} = r * C(t)^p, \tag{2}$$

where h denotes length of the discrete time step between periods. Taking the limit as $h \to 0$ in (2) we obtain (1). The number of individuals who become infected during the time interval $[t,\ t+h)$ is given by $C(t+h) - C(t)$.

2.2. Formulating the GGM as a generalized nonlinear model

Generalized linear/nonlinear models are flexible generalizations of ordinary least-squares regression that allow for outcome variables to have error distribution models other than a normal distribution. A GNM consists of three elements, namely, (i) a probability distribution from the exponential family; (ii) a nonlinear predictor and; (iii) a link function (18).

Let y_t denote the observed number of individuals who become infected in the interval $[t,\ t+h)$. Due to the count nature of y_t it is natural to assume that heterogeneity in y_t can be modeled using a Poisson model, i.e.,

$$y_t|C(t) \sim \text{Poisson}\big(\mu_t\big), \tag{3}$$

where $\mu_t = r * C(t)^p * h$ is the nonlinear predictor with identity link function. Conditional on $C(t)$ as well as on model parameters, incidence counts y_t are independent. An important property of the Poisson distribution is that μ_t is the mean and variance of y_t (equidispersion). With (3), heterogeneity is incorporated through observed variation (with no contribution from unobserved heterogeneity), i.e., variation as a result of observing incidence at different time points. Unique time points t have unique (fixed) mean values μ_t, however, equal or approximately equal incidence counts occur at more or less the same time point. Hence, observed y_t are random fluctuations which closely resemble the qualitative behaviour of the mean trend μ_t.

In real life the equidispersion assumption does not usually hold, oftentimes the variability of the data is greater than that predicted by the Poisson model (overdispersion) (18). In the statistical literature overdispersion is typically motivated in terms of unobserved heterogeneity. Instead of assuming that, at a given time t, μ_t is fixed, it is replaced by a distribution of μ_t's, i.e., $\tilde{\mu}_t = \xi_t * \mu_t$, where ξ_t is a random error term uncorrelated with $C(t)$, so that (3) is modified to,

$$y_t|C(t), \xi_t \sim \text{Poisson}\big(\xi_t * \mu_t\big). \tag{4}$$

The ξ_t term can be viewed as representing combined effects of unobserved variables that have been omitted from the model (unmodeled processes) or as another source of pure randomness (see, e.g, (15)). It relaxes the equidispersion assumption such that equal or approximately equal incidence counts do not have to occur at more or less the same time point. Note that when ξ_t is chosen to be a white-noise process the conditional independence assumption is preserved.

The noise intensity of ξ_t has an impact on the qualitative behaviour of y_t. When the noise intensity tends to zero, fluctuations will closely resemble the mean trend as in (3). On the other hand, high noise intensity increases, at any given time, the frequency of counts which are smaller or greater than μ_t resulting in fluctuations which may not closely resemble the mean trend μ_t (overdispersion). Under the assumption that ξ_t is a gamma white-noise process, averaging over ξ_t model (4) reduces to,

$$y_t|C(t) \sim \text{NB}\left(\frac{\theta^{-1}}{\mu_t + \theta^{-1}}, \theta^{-1}\right), \tag{5}$$

where θ (>0) is the dispersion parameter representing the level of overdispersion. Under this parameterization of the NB distribution the conditional mean and variance are given by,

$$E\Big(y_t|C(t)\Big) = \mu_t \text{ and,} \tag{6}$$

$$Var\Big(y_t|C(t)\Big) = \mu_t + \theta\mu_t^2, \tag{7}$$

respectively. As $\theta \to 0$ the model is equivalent to the Poisson model (4). As $\theta \to \infty$ the model results in a pure random walk as the noise completely swamps the deterministic solution (21; 22). As such, the NB model can be used to model count data with varying degrees of overdispersion (23). Note that $\theta < 0$ corresponds to a scenario of under-dispersion (mean greater than variance); since this scenario is not common in practice we do not consider it.

2.3. Model inference

2.3.1. Point estimation

Estimation of model parameters is performed within the framework of classical maximum likelihood theory. Let Θ denote the parameter vector, i.e., $\Theta = \{r, p\}$ for the Poisson model (3), or, $\Theta = \{r, p, \theta\}$ for the NB model (5). The likelihood function for the Poisson model is given by,

$$\begin{aligned} L\big(\Theta|y_t\big) &= \prod_{t=1}^{n} p(y_t|\Theta) \\ &= \prod_{t=1}^{n} \frac{\mathrm{e}^{-\mu_t}\mu_t^{y_t}}{y_t!}. \end{aligned} \tag{8}$$

For the NB model the likelihood function is given by,

$$\begin{aligned} L\big(\Theta|y_t\big) &= \prod_{t=1}^{n} p(y_t|\Theta) \\ &= \prod_{t=1}^{n} \frac{\Gamma\big(y_t+\theta^{-1}\big)}{y_t!\Gamma(\theta^{-1})}\left(\frac{\theta^{-1}}{\theta^{-1}+\mu_t}\right)^{\theta^{-1}}\left(\frac{\mu_t}{\theta^{-1}+\mu_t}\right)^{y_t}. \end{aligned} \tag{9}$$

In either case, the log-likelihood function can be maximized using standard optimization methods to obtain the maximum likelihood estimate $\hat{\Theta}$. We maximize the log-likelihood function using the Nelder-Mead algorithm which is implemented in the `mle2` function of the `R` package `bblme` (24). To prevent negative values for $\hat{\Theta}$ we estimate all parameters on the log scale and exponentiate the resulting ML estimates.

Note that in our analyses we do not assume that the length of the data reporting interval is related to h; we fix $h = 0.001$ so that (2) provides a close approximation to (1). To evaluate the likelihood, we difference $C(t)$ at the time points which correspond to the data reporting interval.

2.3.2. Interval estimation

In the maximum likelihood framework standard errors for constructing confidence intervals are typically obtained via inverting the Hessian matrix (matrix of second order derivatives of the log-likelihood). Standard errors obtained this way rely on the assumption that $\hat{\Theta}$ is approximately normally distributed (see, e.g., (17)). This assumption may not hold for short time series, moreover, it may yield confidence intervals which contain negative values for non-negative parameters. One possible way to circumvent this issue is to construct confidence intervals using the parametric-bootstrap approach (25). A $100(1-\alpha)\%$ confidence interval for Θ which has an exact coverage in the neighbourhood of $\hat{\Theta}$ is obtained as follows:

1. fit (3) or (5) to the incidence time series y_t to obtain $\hat{\Theta}$;
2. fix $\Theta = \hat{\Theta}$ in (3) or (5) and generate new incidence time series y_t^b from the distribution, where $b = 1, \ldots, B$;
3. for $b = 1, \ldots, B$ fit (3) or (5) to the incidence time series y_t^b to obtain $\hat{\Theta}^b$.
4. The $100(1-\alpha)\%$ confidence interval for Θ is given by $\left(\hat{\Theta}^b_{\frac{\alpha}{2}},\ \hat{\Theta}^b_{1-\frac{\alpha}{2}}\right)$.

2.3.3. Model comparison

The Akaike's information criterion (AIC) is a popular likelihood based model comparison tool. It is calculated as,

$$\text{AIC} = -2\log L\big(\Theta|y_t\big) + 2k, \qquad (10)$$

where k is the dimension of Θ. It acts as a penalized log-likelihood criterion, providing a trade off between a good fit (high value of log-likelihood) and complexity (models with larger k are penalized more than those with smaller k). Among a set of candidate models the 'best' model is the one with the smallest AIC (26). A general rule of thumb is that models that differ in AIC by more than two units are generally considered to 'differ' in terms of fit.

3. Results

3.1. Simulation studies

Our simulation studies serve two purposes. First, through inspecting sample-to-sample behaviour (bias, sample-to-sample variability and coverage) of parameter estimates, we evaluate performance of the estimation procedure, i.e., the extent to which model parameters are well estimated. We calculate bias as the difference between the average of the sample parameter estimates and the true value of the parameter. Sample-to-sample variability is calculated as the standard deviation of the sample parameter estimates - this gives an idea of the extent to which sample-to-sample estimates are concentrated around the average of the estimates. Coverage is calculated as the proportion (%) of the time that the 95% confidence interval contains the true value. In principle, if all sources of variabililty are correctly accounted for then coverage should equal the confidence level of the interval.

Secondly, we assess the *Type I* and *Type II errors* through inspecting the proportion (%) that an incorrect conclusion is reached. A priori, we take the position that the early ascending phase follows exponential growth dynamics, i.e., we test - $H_0 : \ p = 1$ against $H_1 : \ p < 1$. Given an appropriate confidence interval, H_0 is rejected when the confidence interval excludes the null hypothesized value $p = 1$ (see for example, (27)). Note that since H_1 is one-sided we calculate a one-sided confidence interval for p, i.e., (-∞, $\hat{p}_{1-\alpha}$). Table 1 summarises relations between truth/falseness of H_0 and conclusions reached upon fitting the GGM. In the exponential growth scenario, a *Type I error* occurs when the upper confidence limit is less than one. In the sub-exponential growth scenario, a *Type II error* occurs when the upper confidence limit includes one.

Table 1: Definition of *Type I* and *Type II* errors

	$H_0 : p = 1$	
	True	False
fail to reject H_0	correct inference	*Type II error*
reject H_0	*Type I error*	correct inference

We simulate data with varying levels of dispersion: $\theta = \{0.001, 0.1, 0.2, 1\}$. For a given mean, the smallest value, $\theta = 0.001$, leads to a NB distribution which is practically indistinguishable from a Poisson distribution (23). The remaining θ values, chosen on an ad-hoc basis, represent more levels of variability than that predicted by the Poisson distribution.

For each scenario defined by θ, a simulation proceeds as follows:

1. set $C(0) = 1$;
2. in (1) set $\left(r^{-1} \text{ day},\ p\right)$ equal to some "true" values and the ascending phase $t = 0, \dots, 20$ days;
3. solve (1) to obtain the deterministic solution of the cumulative incidence $C(t)$ and, calculate the incidence $C'(t)$;
4. set $\mu_t = C'(t)$ in (5) and generate 250 time series.
5. For each generated time series obtain $\hat{\Theta}$ and construct a 95% bootstrap confidence interval (we set $B = 250$).

For each time series, we estimate $\hat{\Theta}$ for increasing lengths of the ascending phase (15:20 days) so as to assess the sample to sample behaviour of estimates for different lengths of the time series. At each ascending phase length, we fit both the Poisson and NB models so as to assess how ignoring or accounting for overdispersion may impact conclusions.

We perform three sets of simulations. The first is the exponential growth setting where we set an ad hoc parameter set: $\{r = 0.3,\ p = 1\}$. The last two are sub-exponential growth settings, i.e. one where p is closer to one: $\{r = 0.43,\ p = 0.9\}$ and, another where p is further below one: $\{r = 0.9,\ p = 0.7\}$. Since a value of $p < 1$ yields slower-than-exponential growth (smaller case counts), we increase the value of r from 0.3 so as to obtain, across the three settings, comparable numbers of cumulative cases at time $t = 20$, i.e., r values are increased such that in the deterministic solution $C(20)$ is between 400-500.

3.1.1. Exponential growth setting

Figure 1 shows data simulated under the exponential growth assumption. As expected, when θ is very small (=0.001), simulated data closely resemble the deterministic curve. On the other hand, as the value of θ gets larger, simulated data deviate from the deterministic curve, in some cases appearing to rise sub-exponentially.

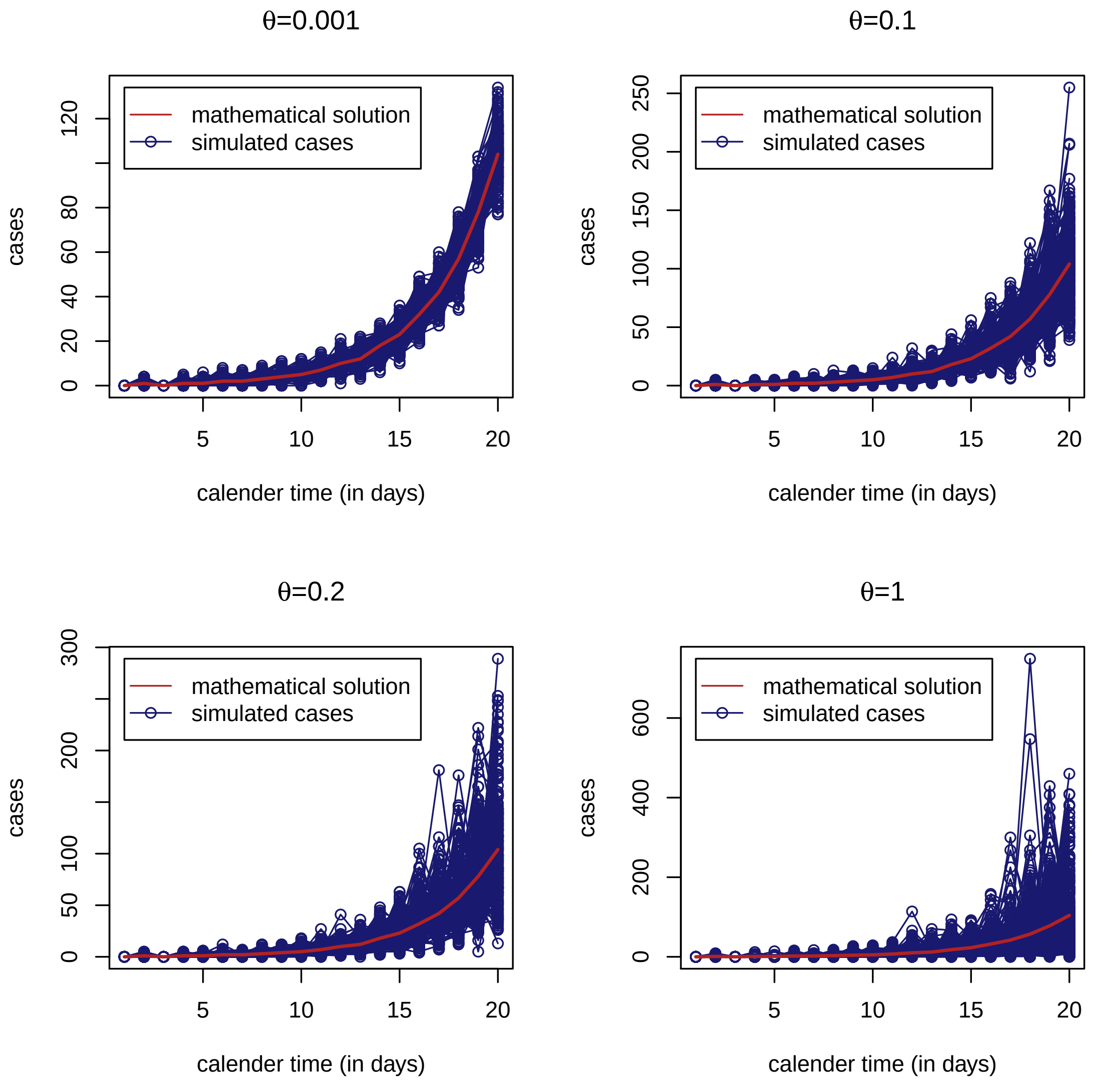

Figure 1: Exponential growth setting: the blue lines represent 250 realizations (simulated datasets) of the GGM. The red lines represent the deterministic solution of the GGM. Each graph corresponds to a different value of the dispersion parameter, $\theta = \{0.001, 0.1, 0.2, 1\}$. For each case, values of the growth rate and the growth scaling parameter are set at $r = 0.3$ and $p = 1$, respectively.

Table 2 presents a summary of the sample to sample behaviour of parameter estimates for the exponential growth setting: for the NB and Poisson models, under different levels of dispersion and, for increasing ascending phase lengths.

Table 2: Exponential growth setting: simulation summary by length of ascending phase for each of the four settings defined by θ: bias, standard deviation, coverage (%) as well as % of simulations in which p is significantly less than one (sub-exponential growth). True values of the growth rate and the growth scaling parameter are: $r = 0.3$ and $p = 1$.

model	dispersion	length of	bias			st. deviation			coverage (%)			*Type I error*
	parameter	ascending phase	r	p	θ	r	p	θ	r	p	θ	$p < 1$ (%)
NB	$\theta = 0.001$	15	0.014	-0.012	0.012	0.054	0.071	0.036	94.400	95.200	27.600	4.800
		16	0.006	-0.003	0.008	0.044	0.058	0.029	96.000	95.200	28.400	4.800
		17	0.007	-0.006	0.006	0.036	0.044	0.021	96.000	95.200	30.400	4.800
		18	0.006	-0.005	0.005	0.032	0.038	0.015	94.400	94.800	30.800	5.200
		19	0.003	-0.002	0.003	0.026	0.030	0.010	95.600	95.600	40.400	4.400
		20	0.001	0.000	0.002	0.021	0.023	0.007	94.400	96.400	50.000	3.600
	$\theta = 0.1$	15	0.010	-0.002	-0.020	0.067	0.094	0.113	92.800	93.600	62.000	6.400
		16	0.007	-0.001	-0.023	0.059	0.081	0.100	90.400	92.800	67.600	7.200
		17	0.004	0.001	-0.020	0.053	0.069	0.098	90.400	94.000	67.600	6.000
		18	0.002	0.003	-0.017	0.045	0.056	0.084	91.200	94.400	69.600	5.600
		19	0.003	0.000	-0.017	0.041	0.049	0.075	91.200	95.600	71.200	4.400
		20	0.002	0.001	-0.017	0.036	0.042	0.065	92.800	95.600	73.600	4.400
	$\theta = 0.2$	15	0.014	-0.004	-0.034	0.078	0.110	0.176	91.600	93.200	72.400	6.800
		16	0.012	-0.004	-0.033	0.068	0.093	0.150	92.400	94.000	72.800	6.000
		17	0.011	-0.005	-0.028	0.059	0.076	0.133	93.600	93.200	76.800	6.800
		18	0.009	-0.004	-0.032	0.054	0.068	0.120	94.400	92.400	77.600	7.600
		19	0.008	-0.004	-0.032	0.050	0.059	0.116	93.200	91.200	73.200	8.800
		20	0.008	-0.005	-0.029	0.045	0.051	0.109	92.800	92.400	75.200	7.600
	$\theta = 1$	15	0.034	-0.014	-0.134	0.142	0.197	0.655	94.000	93.200	74.800	6.800
		16	0.025	-0.007	-0.121	0.127	0.163	0.575	93.200	93.600	76.400	6.400
		17	0.022	-0.006	-0.121	0.116	0.142	0.522	91.200	94.000	76.400	6.000
		18	0.020	-0.007	-0.101	0.106	0.129	0.492	89.200	92.800	78.000	7.200
		19	0.017	-0.007	-0.094	0.098	0.114	0.487	92.800	91.600	78.400	8.400
		20	0.013	-0.005	-0.093	0.088	0.097	0.453	94.400	92.000	76.800	8.000
Poisson	$\theta = 0.001$	15	0.013	-0.012	-	0.053	0.070	-	95.200	94.400	-	5.600
		16	0.005	-0.001	-	0.044	0.058	-	94.800	95.200	-	4.800
		17	0.007	-0.005	-	0.035	0.043	-	95.600	94.000	-	6.000
		18	0.006	-0.005	-	0.032	0.038	-	95.200	93.600	-	6.400
		19	0.003	-0.002	-	0.026	0.029	-	95.200	94.800	-	5.200
		20	0.001	0.000	-	0.021	0.023	-	95.200	96.800	-	3.200
	$\theta = 0.1$	15	0.011	-0.004	-	0.068	0.097	-	86.400	88.400	-	11.600
		16	0.008	-0.001	-	0.061	0.084	-	82.800	88.800	-	11.200
		17	0.007	-0.002	-	0.056	0.074	-	81.200	86.000	-	14.000
		18	0.003	0.002	-	0.048	0.061	-	81.600	87.200	-	12.800
		19	0.006	-0.002	-	0.047	0.056	-	74.800	82.400	-	17.600
		20	0.005	-0.002	-	0.042	0.047	-	72.400	82.400	-	17.600
	$\theta = 0.2$	15	0.019	-0.009	-	0.086	0.121	-	75.600	84.400	-	15.600
		16	0.016	-0.006	-	0.076	0.104	-	77.200	82.400	-	17.600
		17	0.015	-0.009	-	0.066	0.086	-	77.200	80.800	-	19.200
		18	0.013	-0.006	-	0.063	0.081	-	69.200	75.600	-	24.400
		19	0.013	-0.008	-	0.060	0.071	-	65.600	75.600	-	24.400
		20	0.014	-0.010	-	0.056	0.063	-	61.200	71.200	-	28.800
	$\theta = 1$	15	0.058	-0.032	-	0.166	0.233	-	60.000	66.000	-	34.000
		16	0.044	-0.020	-	0.151	0.198	-	51.600	66.800	-	33.200
		17	0.044	-0.018	-	0.147	0.181	-	44.800	64.000	-	36.000
		18	0.054	-0.035	-	0.143	0.162	-	40.800	55.600	-	44.400
		19	0.055	-0.043	-	0.134	0.142	-	38.400	56.400	-	43.600
		20	0.054	-0.040	-	0.132	0.134	-	29.200	50.800	-	49.200

Negative-binomial model: In general, the estimation procedure yields satisfactory estimates for parameters r and p. For all considered levels of dispersion, estimates are well within the range of their true values - small bias (close to zero), low sample to sample variability as well as high coverage ($\approx$ above 90%). Bias and precision improve as longer time series are used in estimating parameters. It is worth noting that sample to sample variability is higher for larger amounts of dispersion - this is expected as more dispersion leads to increased variability around the deterministic trend.

The estimated *Type I error* is between approximately 4%-8% which is close to expected (5%). This is an indication that, in the scenario of exponential growth, when data are equidispersed or overdispersed, fitting the NB model leads to a small probability of reaching an incorrect conclusion that an outbreak is unfolding sub-exponentially.

On the other hand, the dispersion parameter tends to be underestimated and highly variable even more when estimated with shorter time series. Bias and precision improve with longer time series. Coverage tends to be low, however, it improves with longer time series - this is expected since longer time series decrease bias, moreover, a decrease in bias leads to more accurate confidence intervals since bootstrap samples will be more heterogeneous. From this observation it therefore seems plausible to conclude that the reduction in the coverage of r and p, from about 95%, when there is no overdispersion, to about 91% in the presence of overdispersion is due to underestimation of dispersion. Note that the overestimation of the dispersion parameter in the case of equidispersion ($\theta = 0.001$) is due to outliers leading to a high expected value relative to the true value (see supplement).

Poisson model: Similar to the NB model, the estimation procedure yields satisfactory estimates for parameters r and p. However, in the presence of overdispersion, i.e., $\theta = \{0.1, 0.2, 1\}$, due to model misspecification (equidispersion assumption), the quality of estimates is affected - bias and variability are larger compared with NB model results. Note that the larger sample-to-sample variability in the Poisson model does not show in a confidence interval since confidence intervals reflect sampling error as implied by the underlying assumptions of the statistical model; as such, the larger sample-to-sample variability does not imply wider confidence intervals.

When there is no overdispersion, as expected, the Poisson model yields approximately similar results as the NB model. Also, as is the case with the NB model, for all amounts of dispersion considered, bias and precision improve with longer time series. It is worth noting that coverage is lower in the presence of overdispersion compared with the NB model. This is due to the fact that the Poisson model underestimates variability present in a data set leading to narrower confidence intervals which more often exclude the true value.

With respect to *Type I error*, in the presence of overdispersion, also due to narrow confidence intervals, the error rate is higher than expected (5%). This implies that there is a higher estimated probability of incorrectly concluding sub-exponential growth when in fact growth is exponential - when $\hat{p} < 1$, overly narrow confidence intervals more often exclude one.

In terms of goodness-of-fit, in the presence of overdispersion, the NB model outperforms the Poisson model (see supplement). In the absence of overdispersion, both models agree as expected, i.e, the difference in AIC is about 2 units due to the extra parameter.

3.1.2. Sub-exponential growth setting

In the scenario of sub-exponential growth two simulations are conducted, i.e., one where the true value of the growth scaling parameter is close to one, and another where it is further below one. Different lessons are drawn from the two.

Growth scaling parameter p close to one

Figure 2 shows data simulated under the sub-exponential growth assumption ($p = 0.9$). As before, when θ is very small (=0.001), simulated data closely resemble the deterministic curve, moreover, simulated data deviate from the deterministic curve as the value of θ gets larger.

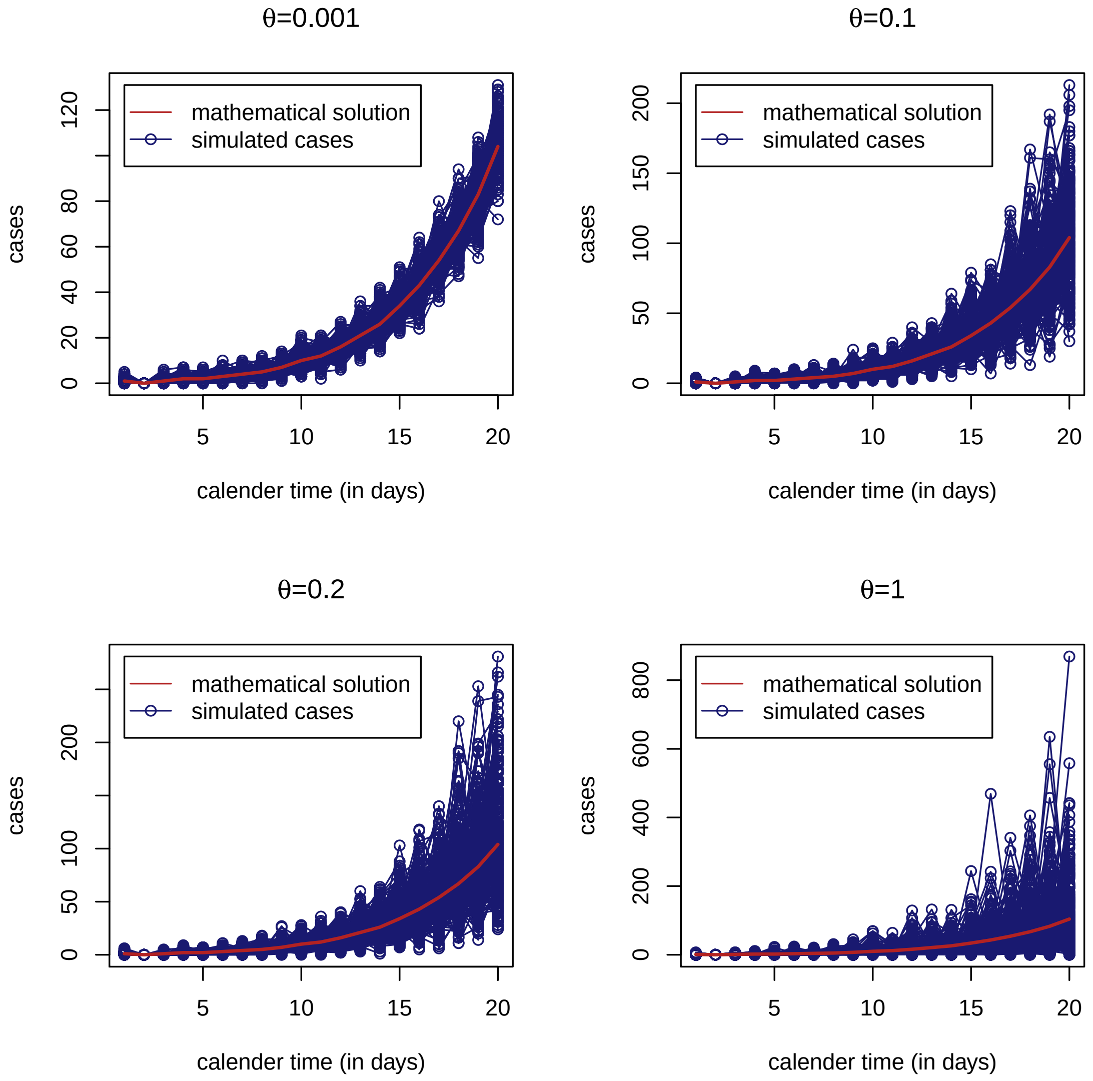

Figure 2: Sub-exponential growth with p close to one: the blue lines represent 250 realizations (simulated datasets) of the GGM. The red lines represent the deterministic solution of the GGM. Each graph corresponds to a different value of the dispersion parameter, $\theta = \{0.001, 0.1, 0.2, 1\}$. For each case, values of the growth rate and the growth scaling parameter are set at $r = 0.43$ and $p = 0.9$, respectively.

Table 3 presents a summary of the sample-to-sample behaviour of parameter estimates: for the NB and Poisson models, under different levels of dispersion and, for increasing ascending phase lengths. Note that here, the true value of the growth scaling parameter is $p = 0.9$, hence, given that the confidence interval includes the true value, $p < 1$ is the correct conclusion.

Table 3: Sub-exponential growth with p close to one: simulation summary by length of ascending phase for each of the four settings defined by θ: bias, standard deviation, coverage (%) as well as % of simulations in which p is not significantly less than one. True values of the growth rate and the deceleration parameter are: $r = 0.43$ and $p = 0.9$.

model	dispersion	length of	bias			st. deviation			coverage (%)			*Type II error*
	parameter	ascending phase	r	p	θ	r	p	θ	r	p	θ	$p = 1$ (%)
NB	$\theta = 0.001$	15	0.007	-0.001	0.007	0.063	0.052	0.023	94.800	96.800	31.200	42.000
		16	0.004	0.001	0.005	0.055	0.043	0.017	95.600	95.600	32.000	30.000
		17	0.002	0.002	0.003	0.048	0.036	0.010	97.200	96.400	34.400	16.400
		18	0.002	0.001	0.002	0.044	0.031	0.008	93.600	94.400	29.600	7.600
		19	0.006	-0.003	0.002	0.038	0.026	0.006	92.800	94.000	38.400	1.600
		20	0.003	-0.001	0.001	0.034	0.022	0.005	94.800	96.000	41.600	0.000
	$\theta = 0.1$	15	0.014	-0.003	-0.022	0.086	0.073	0.082	92.400	94.000	73.600	52.400
		16	0.012	-0.003	-0.024	0.079	0.066	0.074	90.800	95.200	75.200	44.800
		17	0.011	-0.003	-0.022	0.074	0.059	0.067	90.800	94.800	74.000	38.000
		18	0.008	-0.001	-0.018	0.065	0.050	0.062	92.800	93.600	78.400	33.600
		19	0.008	-0.002	-0.016	0.061	0.045	0.057	91.200	91.600	80.400	19.600
		20	0.007	-0.002	-0.014	0.057	0.041	0.053	90.800	93.600	82.000	17.600
	$\theta = 0.2$	15	0.015	-0.001	-0.042	0.106	0.088	0.143	89.200	93.200	72.000	62.400
		16	0.011	0.002	-0.042	0.097	0.078	0.126	89.600	93.600	70.800	56.400
		17	0.011	0.000	-0.043	0.088	0.069	0.115	89.200	94.000	72.000	48.000
		18	0.007	0.003	-0.036	0.081	0.061	0.104	91.200	94.000	74.800	44.400
		19	0.005	0.003	-0.035	0.073	0.053	0.097	91.200	95.200	75.600	35.600
		20	0.006	0.002	-0.031	0.068	0.048	0.092	92.400	93.600	76.400	33.200
	$\theta = 1$	15	0.035	-0.005	-0.056	0.190	0.149	0.519	92.800	94.400	80.800	82.400
		16	0.028	-0.003	-0.060	0.175	0.134	0.488	90.400	96.000	82.800	78.400
		17	0.021	-0.001	-0.069	0.151	0.118	0.461	90.400	94.800	82.000	78.800
		18	0.016	0.003	-0.060	0.145	0.108	0.438	91.600	94.800	83.200	74.800
		19	0.014	0.003	-0.056	0.138	0.100	0.405	92.400	93.600	86.400	71.600
		20	0.009	0.005	-0.040	0.125	0.090	0.389	89.600	94.400	86.000	73.600
Poisson	$\theta = 0.001$	15	0.006	-0.001	-	0.062	0.052	-	95.600	97.200	-	38.800
		16	0.004	0.001	-	0.055	0.043	-	94.800	94.000	-	26.000
		17	0.001	0.002	-	0.047	0.035	-	95.200	96.400	-	12.400
		18	0.003	0.000	-	0.043	0.031	-	95.200	94.000	-	6.800
		19	0.007	-0.003	-	0.037	0.026	-	95.600	92.800	-	0.400
		20	0.003	-0.001	-	0.034	0.022	-	94.000	95.600	-	0.000
	$\theta = 0.1$	15	0.015	-0.003	-	0.093	0.080	-	79.200	87.200	-	42.800
		16	0.015	-0.004	-	0.088	0.072	-	78.400	84.000	-	31.600
		17	0.012	-0.002	-	0.083	0.066	-	76.000	84.000	-	22.800
		18	0.007	0.001	-	0.075	0.056	-	73.200	83.600	-	18.800
		19	0.012	-0.003	-	0.075	0.054	-	73.600	81.600	-	12.400
		20	0.011	-0.003	-	0.070	0.048	-	68.000	79.600	-	8.800
	$\theta = 0.2$	15	0.015	0.001	-	0.112	0.097	-	74.400	84.400	-	43.200
		16	0.012	0.003	-	0.105	0.088	-	71.200	82.800	-	36.800
		17	0.015	-0.001	-	0.098	0.080	-	69.200	82.000	-	27.200
		18	0.009	0.004	-	0.096	0.075	-	60.800	78.000	-	27.200
		19	0.008	0.003	-	0.090	0.065	-	57.600	79.200	-	21.600
		20	0.011	0.000	-	0.083	0.060	-	58.000	71.600	-	12.400
	$\theta = 1$	15	0.065	-0.018	-	0.230	0.184	-	54.000	67.200	-	40.800
		16	0.060	-0.015	-	0.224	0.173	-	48.800	66.000	-	36.400
		17	0.048	-0.016	-	0.190	0.144	-	44.800	64.000	-	33.200
		18	0.055	-0.017	-	0.211	0.142	-	37.200	60.800	-	33.200
		19	0.067	-0.026	-	0.214	0.133	-	39.600	60.400	-	27.600
		20	0.065	-0.027	-	0.202	0.124	-	32.000	56.400	-	26.000

Negative-binomial model: In terms of performance of the estimation procedure, bias, sample-to-sample variability and coverage point in the same direction as already discussed in the previous section. The major difference is with respect to error rates - the *Type II error* is considerably high. However, for each level of dispersion, the *Type II error* decreases with longer time series - the decrease is considerably more when there is less overdispersion. This can be explained as follows. Given that the true value of p is close to one, longer time series are required in order to gain sufficient precision to correctly conclude that $p < 1$ - when a time series is short, precision is low resulting in wide confidence intervals which contain one. Moreover, as overdispersion leads to greater variability "substantially" longer time series will be required for high levels of overdispersion and vice versa.

Poisson model: Looking at performance of the estimation procedure, similar results are observed as before in terms of bias, sample-to-sample variability and coverage. As is the case with the NB model, the *Type II error* is considerably high and it decreases with longer time series. Note that the error rates are lower for the Poisson model compared with the NB model due to narrower confidence intervals. The implication is not that the Poisson model gives better results since its coverage is lower.

Growth scaling parameter p further below one

Figure 3 shows data simulated under the sub-exponential growth assumption ($p = 0.7$). As before, when θ is very small (=0.001), simulated data closely resemble the deterministic curve, moreover, simulated data deviate from the deterministic curve as the value of θ gets larger.

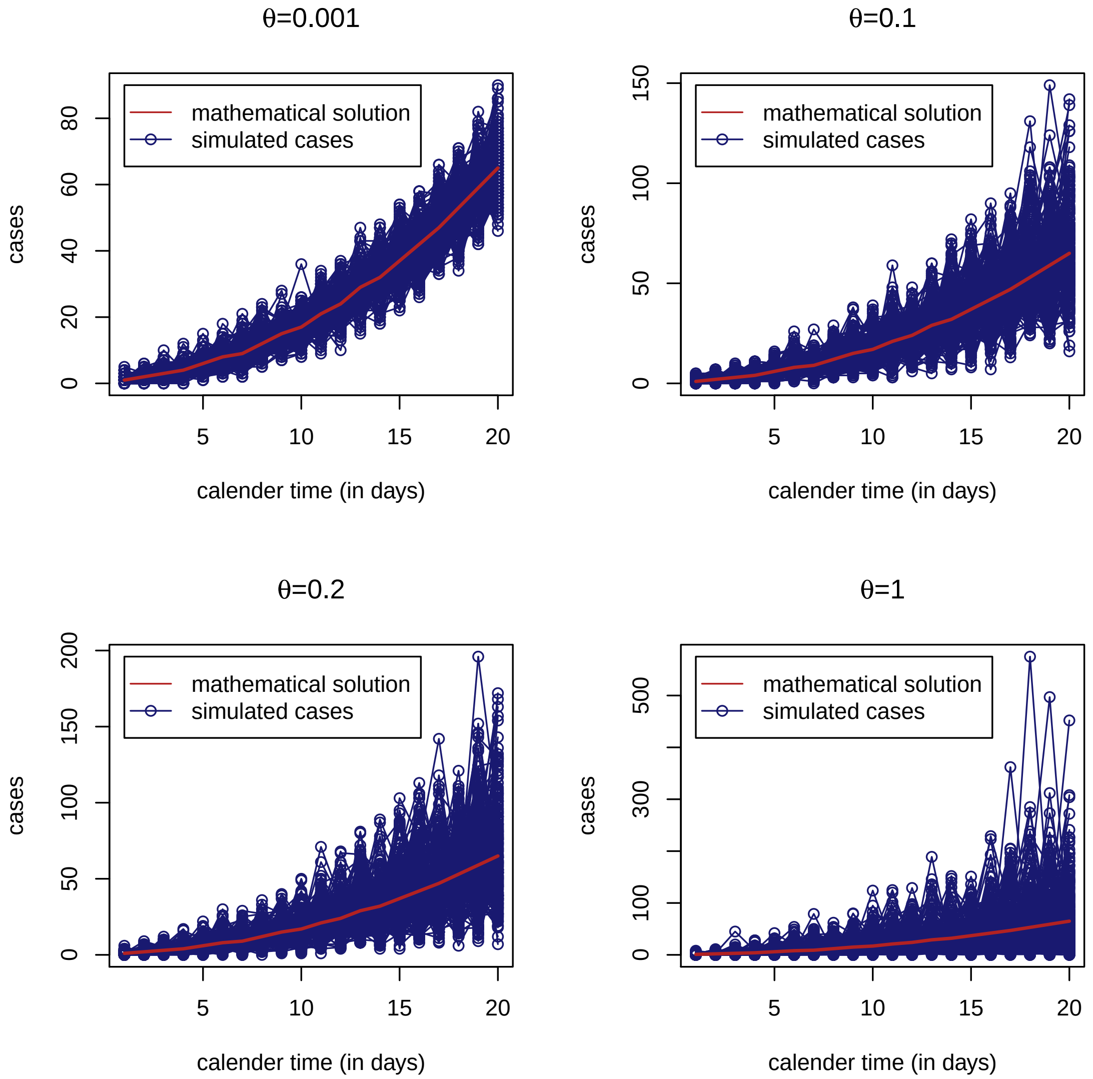

Figure 3: Sub-exponential growth with p further below one: the blue lines represent 250 realizations (simulated datasets) of the GGM. The red lines represent the deterministic solution of the GGM. Each graph corresponds to a different value of the dispersion parameter, $\theta = \{0.001, 0.1, 0.2, 1\}$. For each case, values of the growth rate and the deceleration parameter are set at $r = 0.9$ and $p = 0.7$, respectively.

Table 4 presents a summary of the sample-to-sample behavior of parameter estimates: for the NB and Poisson models, under different levels of dispersion and, for increasing ascending phase lengths. The true value of the growth scaling parameter is $p = 0.7$, hence, given that the confidence interval includes the true value, $p < 1$ is the correct conclusion.

Table 4: Sub-exponential growth with p further below one: simulation summary by length of ascending phase for each of the four settings defined by θ: bias, standard deviation, coverage (%) as well as % of simulations in which p is significantly less than one (sub-exponential growth). True values of the growth rate and the deceleration parameter are: $r = 0.9$ and $p = 0.7$.

model	dispersion	length of	bias			st. deviation			coverage (%)			*Type II error*
	parameter	ascending phase	r	p	θ	r	p	θ	r	p	θ	$p = 1$ (%)
NB	$\theta = 0.001$	15	0.002	0.003	0.004	0.129	0.041	0.013	96.400	97.600	33.600	0.000
		16	0.007	0.001	0.004	0.121	0.037	0.011	94.400	96.400	35.200	0.000
		17	0.005	0.001	0.003	0.111	0.033	0.010	96.000	97.200	34.800	0.000
		18	0.004	0.001	0.003	0.105	0.030	0.009	95.200	96.800	35.600	0.000
		19	0.000	0.002	0.002	0.098	0.027	0.008	94.400	97.200	35.600	0.000
		20	0.000	0.002	0.002	0.093	0.025	0.007	94.000	97.200	35.200	0.000
	$\theta = 0.1$	15	0.010	0.004	-0.019	0.198	0.064	0.068	89.600	96.000	77.600	0.400
		16	0.005	0.005	-0.016	0.180	0.058	0.066	91.600	97.200	80.000	0.000
		17	0.005	0.004	-0.016	0.169	0.052	0.060	91.200	97.600	82.000	0.000
		18	0.003	0.005	-0.015	0.163	0.049	0.056	92.400	94.800	83.600	0.000
		19	0.005	0.004	-0.015	0.156	0.046	0.053	92.400	96.800	82.800	0.000
		20	0.008	0.002	-0.015	0.149	0.043	0.050	92.000	96.000	81.600	0.000
	$\theta = 0.2$	15	0.031	0.000	-0.027	0.226	0.070	0.112	95.600	94.400	78.800	0.400
		16	0.030	-0.001	-0.026	0.224	0.067	0.106	92.800	94.800	78.000	0.000
		17	0.032	-0.002	-0.026	0.220	0.064	0.098	92.400	94.000	80.000	0.000
		18	0.028	-0.001	-0.026	0.207	0.059	0.092	91.600	93.600	80.400	0.000
		19	0.027	-0.001	-0.022	0.201	0.056	0.089	90.000	94.000	81.200	0.000
		20	0.028	-0.002	-0.021	0.193	0.053	0.085	93.200	95.200	81.200	0.000
	$\theta = 1$	15	0.069	0.005	-0.153	0.477	0.136	0.366	88.800	94.000	78.800	24.400
		16	0.072	0.001	-0.133	0.465	0.125	0.355	90.400	94.800	81.600	15.600
		17	0.061	0.005	-0.131	0.463	0.119	0.344	88.400	93.600	82.400	13.200
		18	0.060	0.004	-0.124	0.481	0.113	0.338	89.200	94.400	82.400	8.800
		19	0.067	0.002	-0.105	0.498	0.107	0.337	88.000	96.000	83.200	6.400
		20	0.049	0.003	-0.102	0.393	0.099	0.320	90.400	97.200	84.400	5.200
Poisson	$\theta = 0.001$	15	0.003	0.002	-	0.129	0.041	-	96.400	98.400	-	0.000
		16	0.009	0.000	-	0.121	0.037	-	95.600	97.200	-	0.000
		17	0.007	0.000	-	0.110	0.033	-	95.600	98.400	-	0.000
		18	0.005	0.000	-	0.104	0.030	-	94.400	96.800	-	0.000
		19	0.001	0.001	-	0.097	0.027	-	94.400	96.400	-	0.000
		20	0.002	0.001	-	0.093	0.025	-	94.800	97.600	-	0.000
	$\theta = 0.1$	15	0.017	0.003	-	0.213	0.070	-	81.200	90.000	-	0.400
		16	0.013	0.004	-	0.198	0.064	-	75.600	89.600	-	0.000
		17	0.013	0.003	-	0.183	0.057	-	78.400	89.600	-	0.000
		18	0.008	0.005	-	0.181	0.054	-	76.800	89.200	-	0.000
		19	0.013	0.002	-	0.174	0.050	-	74.800	85.600	-	0.000
		20	0.017	0.001	-	0.168	0.049	-	72.800	83.200	-	0.000
	$\theta = 0.2$	15	0.064	-0.009	-	0.259	0.077	-	76.800	81.600	-	0.000
		16	0.060	-0.007	-	0.268	0.080	-	70.000	80.400	-	0.000
		17	0.063	-0.008	-	0.266	0.076	-	65.200	77.600	-	0.000
		18	0.053	-0.006	-	0.245	0.069	-	67.600	75.600	-	0.000
		19	0.051	-0.005	-	0.247	0.069	-	66.400	76.800	-	0.000
		20	0.051	-0.006	-	0.229	0.064	-	60.800	73.600	-	0.000
	$\theta = 1$	15	0.184	-0.019	-	0.609	0.157	-	47.600	66.000	-	4.000
		16	0.209	-0.023	-	0.663	0.157	-	43.200	66.800	-	4.400
		17	0.186	-0.014	-	0.656	0.160	-	35.200	62.400	-	4.400
		18	0.195	-0.018	-	0.699	0.147	-	38.000	60.000	-	3.600
		19	0.228	-0.026	-	0.728	0.146	-	38.800	58.000	-	3.200
		20	0.223	-0.027	-	0.643	0.140	-	31.600	55.600	-	2.000

Negative-binomial model: As before (Tables 2 and 3), the estimation procedure performs well in terms of bias, sample-to-sample variability and coverage. In terms of *Type II error*, since here the true value of p is further below one, with moderate to no overdispersion, i.e., $\theta = \{0.001, 0.1, 0.2\}$, the data provide sufficient precision such that almost all confidence intervals exclude one (0% error rate). However, for high levels of overdispersion, i.e., $\theta = 1$, as observed in Table 3, *Type II error* is high and it decreases with longer time series.

Poisson model: In terms of the estimation procedure, performance is similar as observed in Tables 2 and 3. With respect to *Type II error*, as the true value of p is further below one, due to narrow confidence intervals, all confidence intervals except for $\theta = 1$ exclude one (0% error rate). In the case of $\theta = 1$ we again see that *Type II error* decreases with longer time series.

3.2. Real life data examples

Figure 4 shows five real outbreaks used for benchmarking findings from the simulation study: 2015 Zika outbreak in Antioquia, Colombia (obtained from (11)); 2014 Ebola outbreaks in Tonkolili, Sierra Leone and Margibi, Liberia (obtained from (28)); 1918 Influenza outbreak in San Francisco, USA (obtained from (29)) and; 2015 Zika outbreak in Girardot and Sanandres, Colombia (obtained from (30)).

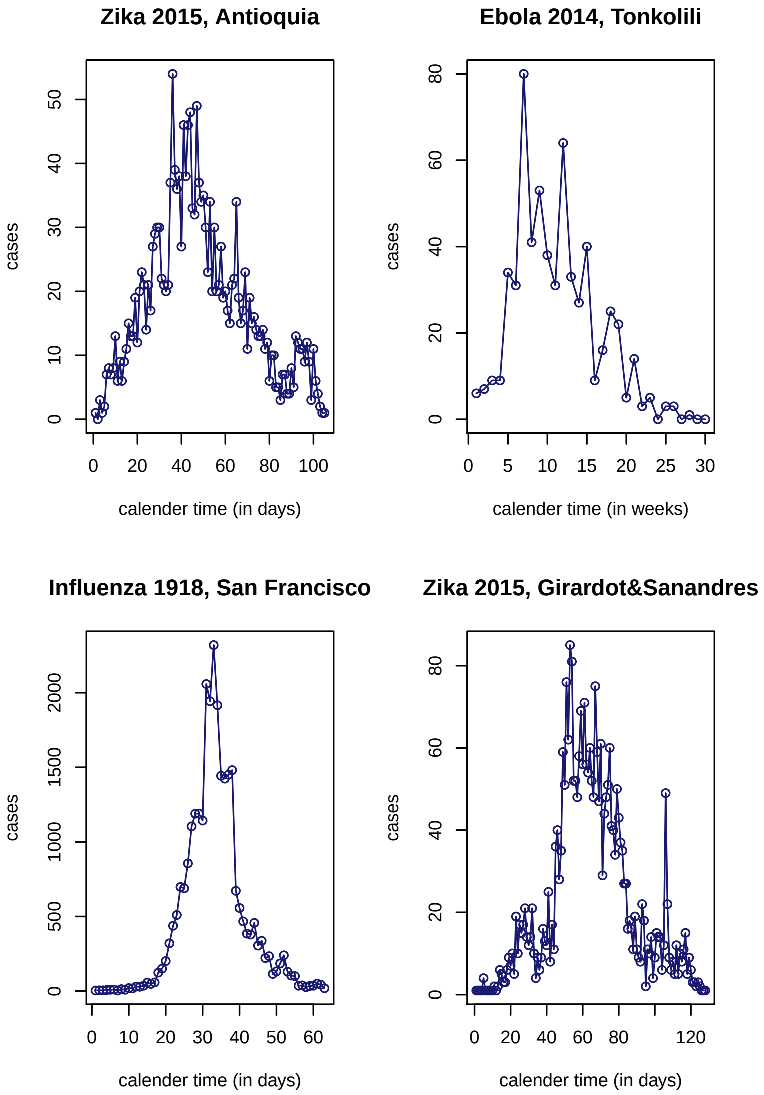

Figure 4: Studied real life epidemics. From top left to bottom right: Zika 2015, Antioquia; Ebola 2014, Tonkolili; Ebola 2014, Margibi; Influenza 1918, San Francisco and; Zika 2015, Girardot and Sanandres.

Table 5 presents parameter estimates of the five real outbreaks: for the NB and Poisson models, for increasing ascending phase lengths chosen on an ad hoc basis. Below we describe the results.

Zika, Antioquia: The NB and Poisson models strongly agree in terms of goodness-of-fit - the latter is slightly better with an AIC difference of about two. Moreover, parameter estimates are also in agreement with indications of no overdispersion and both models point to sub-exponential growth. This outbreak can be compared to the simulation study where we set p far below one ($p = 0.7$) with no overdispersion ($\theta = 0.001$). Confidence intervals become narrower with longer time series and *Type II error* is 0%. From this analysis it can be concluded that this epidemic unfolded sub-exponentially.

Ebola, Tonkolili: The NB and Poisson models are in agreement in terms of both goodness-of-fit and parameter estimates. Both models support conclusion of exponential growth. This outbreak can be compared to the exponential growth simulation study ($p = 1$) with no overdispersion $\theta = 0.001$ where p estimates are close to one with longer time series leading to a gain in precision and *Type I error* about 5%. From this analysis it can be concluded that this epidemic unfolded exponentially.

Table 5: Real life epidemics results: parameter estimates, confidence intervals and AIC values obtained when fitting the GGM to selected ascending phase lengths.

data	temporal resolution	model	length of ascending phase	estimate (95% confidence interval)			AIC
				$\hat{r}$	$\hat{p}$	θ	
Zika 2015, Antioquia	days	NB	15	1.271 (0.721, 2.224)	0.487 (0.322, 0.679)	2.235x10^{-7} (2.429x10^{-12} , 1.158x10^{-1})	67.978
			16	1.197 (0.741, 1.958)	0.512 (0.352, 0.652)	1.005x10^{-13} (5.378x10^{-12}, 9.074x10^{-2})	73.106
			17	1.207 (0.750, 1.992)	0.509 (0.355, 0.655)	2.740x10^{-8} (2.790x10^{-12}, 7.299x10^{-2})	77.533
		Poisson	15	1.271 (0.785, 2.436)	0.487 (0.296, 0.641)	-	65.978
			16	1.198 (0.749, 2.174)	0.512 (0.354, 0.654)	-	71.105
			17	1.207 (0.737, 2.070)	0.509 (0.353, 0.657)	-	75.533
Ebola 2014, Tonkolili	weeks	NB	5	0.058 (0.016, 0.155)	1.127 (0.804, 1.574)	7.373x10^{-7} (2.015x10^{-9}, 1.046x10^{-1})	28.214
			6	0.133 (0.041, 0.429)	0.840 (0.467, 1.201)	0.038 (9.177x10^{-10}, 1.337x10^{-1})	35.535
			7	0.091 (0.040, 0.183)	0.965 (0.764, 1.195)	0.032 (1.611x10^{-8}, 1.019x10^{-1})	44.110
		Poisson	5	0.058 (0.019, 0.172)	1.128 (0.762, 1.514)	-	26.214
			6	0.138 (0.055, 0.318)	0.827 (0.583, 1.102)	-	34.478
			7	0.083 (0.054, 0.137)	0.990 (0.849, 1.112)	-	43.329
Ebola 2014, Margibi	weeks	NB	9	0.085 (0.060, 0.126)	1.040 (0.902, 1.184)	0.111 (2.324x10^{-8}, 2.772x10^{-1})	50.551
			10	0.099 (0.065, 0.149)	0.963 (0.814, 1.117)	0.206 (3.890x10^{-8}, 4.372x10^{-1})	62.920
			11	0.111 (0.076, 0.164)	0.912 (0.784, 1.037)	0.206 (4.377x10^{-8}, 4.506x10^{-1})	73.702
		Poisson	9	0.091 (0.075, 0.113)	1.014 (0.934, 1.089)	-	56.810
			10	0.134 (0.107, 0.168)	0.855 (0.787, 0.927)	-	97.223
			11	0.153 (0.128, 0.195)	0.807 (0.739, 0.861)	-	109.526
Influenza 1918, San Francisco	days	NB	19	0.523 (0.397, 0.680)	0.846 (0.780, 0.915)	0.048 (2.114x10^{-6}, 1.152x10^{-1})	128.395
			20	0.468 (0.371, 0.590)	0.876 (0.817, 0.934)	0.049 (0.002, 0.094)	140.551
			21	0.420 (0.338, 0.520)	0.906 (0.858, 0.957)	0.063 (0.011, 0.116)	155.586
		Poisson	19	0.403 (0.328, 0.495)	0.910 (0.862, 0.960)	-	138.538
			20	0.354 (0.301, 0.421)	0.944 (0.904, 0.982)	-	150.940
			21	0.294 (0.262, 0.332)	0.991 (0.962, 1.018)	-	173.960
Zika 2015, Girardot and Sanandres	days	NB	51	0.448 (0.342, 0.602)	0.677 (0.605, 0.744)	0.197 (0.084, 0.296)	300.744
			52	0.434 (0.333, 0.556)	0.686 (0.624, 0.751)	0.197 (0.085, 0.294)	310.479
			53	0.415 (0.329, 0.535)	0.700 (0.640, 0.762)	0.207 (0.101, 0.305)	322.354
		Poisson	51	0.384 (0.323, 0.450)	0.714 (0.678, 0.755)	-	384.406
			52	0.362 (0.311, 0.421)	0.730 (0.698, 0.764)	-	397.382
			53	0.326 (0.285, 0.369)	0.758 (0.729, 0.789)	-	424.424

Ebola, Margibi: For this data set the NB and Poisson models do not agree in terms of goodness-of-fit, the former fits better. Dispersion estimates indicate some evidence of overdispersion enough to cause differences between the two models; the estimates are however highly variable as they are estimated from short time series (Section 3.1). Parameter estimates differ, moreover, at weeks 10 and 11, the former concludes exponential growth whereas the latter concludes sub-exponential growth. These results can be compared to the exponential growth simulation study ($p = 1$) with overdispersion $\theta = 0.2$ (Table 2) where, in the NB model, p estimates are close to one with longer time series leading to a gain in precision and *Type I error* about 5%. On the other hand, the Poisson model supports the conclusion of sub-exponential growth due to too narrow confidence intervals but not because it is better than the former since it may have a high *Type I error* (>15%) based on the simulation study.

This outbreak can also be compared to the sub-exponential growth setting where we set the true value of the growth scaling parameter close to one ($p = 0.9$) but the time series is short such that the NB model yields wide confidence intervals which include one and a considerably high *Type II error* (Table 3). On the other hand, the Poisson model yields too narrow confidence interval leading to the conclusion of sub-exponential growth.

Influenza, San Francisco: For this outbreak the NB model fits better compared to the Poisson model. There is some evidence of overdispersion enough to cause differences between the two models, however, there is large uncertainty associated with the dispersion estimates which could be a result of estimating this parameter with few data (Section 3.1). The NB model yields smaller p estimates and supports the conclusion of sub-exponential growth whereas, at day 21, the latter concludes exponential growth. These results can be compared to the sub-exponential growth setting where we set the true value of p close to one and the data give sufficient precision for the NB model to yield confidence intervals which exclude one and hence, a lower *Type II error* (Table 3). On the other hand, at day 21, the Poisson model yields a larger p estimate whose narrow confidence interval includes one.

Zika, Girardot and Sanandres: For this outbreak the NB model provides a better fit than the Poisson model and there is evidence of overdispersion. Both models support the conclusion of sub-exponential growth for all lengths of the ascending phase, however, the Poisson model yields narrower confidence intervals. These results can be compared to the sub-exponential growth setting with ($p = 0.7$ and $\theta = 0.2$) where *Type II error* is 0% for both models but with the Poisson model resulting in lower coverage due to narrower confidence intervals.

4. Discussion and Conclusions

The growth scaling exponent p is an essential parameter of the GGM. On account of this parameter, the GGM can help in identifying epidemic growth patterns of unfolding infectious disease outbreaks, refining existing epidemic models as well as estimating the evolution of epidemics and final epidemic size (see for example, (3; 8; 9; 10; 11; 12; 13)). As such, good parameter estimates of p are required. This paper makes two contributions to inference using the GGM. First, against the background that estimation of GGM parameters has mainly proceeded via the least-squares method, it demonstrates the application of classical ML estimation which has some advantages, e.g., desirable theoretical properties as well as offering a direct way to account for overdispersion. Secondly, the paper evaluates, when using the ML method, the impact of varying levels of overdispersion on the inference of the growth scaling parameter through comparing a Poisson model which assumes equidispersion and a NB model which allows variability to be larger than the mean. In particular, we evaluate *Type I* and *Type II error* rates in exponential and sub-exponential growth scenarios respectively.

We are certainly not the first to compare the estimation of GGM parameters using different statistical models nor to apply a likelihood-based method when estimating GGM parameters in anticipation of overdispersion. For example, in (31) (published while this paper was under review) researchers explore, in the presence of overdispersion, the quality of estimates by comparing a ML method (using a Poisson model) with the method of least-squares (implicitly ordinary nonlinear regression assuming homoscedasticity). Also, in (8) researchers use a ML method to estimate growth scaling parameters of Ebola epidemics at the level of administrative areas during the 2014-16 Ebola epidemic in West Africa.

In general, classical maximum likelihood method performs well at recovering model parameters - bias and precision improve with longer time series. Parameter estimates based on the NB and Poisson models are approximately equal under the assumption of equidispersion but slightly differ when data exhibit overdispersion due to different underlying assumptions of the models. In the case of the NB model, the dispersion parameter tends to be underestimated (even more with shorter time series). The underestimation is unsurprising as it is well known that ML methods generally tend to underestimate variance parameters since they do not adjust for the fact that the mean is estimated from the same data (see for example, (23; 32)). The underestimation implies that more homogeneous datasets are simulated when constructing bootstrap confidence intervals leading to narrower intervals and hence lower coverage. A bias-corrected MLE method for estimating the dispersion parameter was proposed in (33), here, we do not consider it - we expect that correcting the bias (underestimation) will provide more uncertainty to parameter estimates of the NB model hence main conclusions should remain the same. In the case of overdispersed data, the Poisson model underestimates variability leading to too narrow confidence intervals. When data exhibit overdispersion the NB model outperforms the Poisson model in terms of goodness-of-fit.

In the scenario of exponential growth, with or without overdispersion, the NB model performs approximately at the expected 5% *Type I error* rate. The Poisson model also performs approximately at 5% *Type I error* rate when there is no overdispersion; in the presence of overdispersion, it results in error rates which considerably exceed 5% due to too narrower confidence intervals which exclude the true value one. With respect to sub-exponential growth, error rates can be considerably high even if the NB model is used; they depend on how close to one the true value of the growth scaling parameter is, length of the time series as well as the level of overdispersion. For values of p closer to one, longer time series are required to obtain low *Type II error* rates (even more when the data exhibit a great amount of overdispersion). On the other hand, error rates of the Poisson model are lower due to too narrower confidence intervals, however, this should not be taken to indicate that it is better than the NB model. The high *Type II error* rates indicate that results should be interpreted with caution.

Analyses of real outbreak data yield results which are comparable to some of our simulation results. The Zika (Antioquia) and Ebola (Tonkolili) outbreaks exemplify, more or less in an equidispersion setting, scenarios of sub-exponential and exponential growth, respectively. In this case, ignoring or accounting for overdispersion does not impact conclusions about the description of the growth pattern and similar conclusions are reached when the GGM is fitted to longer time series. On the other hand, the Ebola (Margibi) and Influenza (San Fransisco) outbreaks exemplify, in an overdispersion setting, exponential and sub-exponential growth, respectively. For each of these data, depending on whether or not overdispersion is accounted for, different conclusions about the description of the growth pattern are reached. The Zika (Girardot and Sanandres) outbreak exemplifies, in an overdispersion setting, a scenario of sub-exponential growth where both models point to the correct conclusion but with the Poisson model yielding narrower confidence intervals which may potentially exclude the "true" value.

Though we draw these similarities with findings from simulation studies, it is worth pointing out that our work is based on some simplifying assumptions. For example, the Poisson-Gamma mixture (4) is chosen for convenience as it results in a closed-form distribution (5). Though the NB distribution often fits biological data well (34), in our analysis of real outbreak data, this convenience does not guarantee that it offers the "best" fit - it may well be that the gamma mixing distribution is a wrong probability model for the unexplained variation in the expected counts (14). The fit might well be significantly improved by using a larger family of mixture distributions, for instance, the Poisson-Generalized Inverse Gaussian distribution (20). We suggest extensions in this direction for further research.

Also, the GGM is a simple model which does not capture features of the biological processes behind a disease in question. Growth scaling parameters can also be incorporated in mechanistic models (see for example, (4) and references therein). Overdispersion can be introduced in these models by randomizing transmission parameters (see for example, (35; 36)), therefore, a similar study could be performed using a mechanistic model as further research. Nevertheless, mechanistic models introduce additional complexities that can distract from the objectives of this paper. For example, since the likelihood can have no closed-form distribution upon randomizing transmission parameters (see (36) and references therein), inference then proceeds via more computationally intensive ML methods which require tuning requirements (37; 38) therefore complicating large-scale simulations.

Overall, our results demonstrate that classical ML estimation performs well for estimating GGM parameters. The NB model outperforms the Poisson model; we therefore recommend that the Poisson model be avoided even if overdispersion is unsuspected - in the absence of overdispersion, the NB model yields estimates and confidence intervals which are practically indistinguishable to those obtained using the Poisson model. The ability to correctly identify early growth patterns of outbreaks can be affected by overdispersion. While the NB model accommodates overdispersion, *Type I* and *Type II errors* can be high due to the degree of overdispersion present in the data, small sample sizes as well as how close to one the true growth scaling parameter is. Hence, results should be interpreted with caution especially when dealing with short time series or when data are highly overdispersed as growth patterns may be ambiguous.

Acknowledgements

The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government department EWI. The authors are greatful to Gerardo Chowell for constructive comments and suggestions on the manuscript.

References

[1] F. Brauer, Mathematical epidemiology: Past, present, and future, Infectious Disease Modelling 2 (2) (2017) 113–127. doi:10.1016/j.idm.2017.02.001.

[2] G. Chowell, C. Viboud, L. Simonsen, S. M. Moghadas, Characterizing the reproduction number of epidemics with early subexponential growth dynamics, Journal of The Royal Society Interface 13 (123) (2016) 20160659. doi:10.1098/rsif.2016.0659.

[3] C. Viboud, L. Simonsen, G. Chowell, A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks, Epidemics 15 (2016) 27–37. doi:10.1016/j.epidem.2016.01.002.

[4] T. Ganyani, C. Faes, G. Chowell, N. Hens, Assessing inference of the basic reproduction number in an SIR model incorporating a growth-scaling parameter, Statistics in Medicine 37 (29) (2018) 4490–4506. doi:10.1002/sim.7935.

[5] V. Capasso, G. Serio, A generalization of the kermack-McKendrick deterministic epidemic model, Mathematical Biosciences 42 (1-2) (1978) 43–61. doi:10.1016/0025-5564(78)90006-8.

[6] N. D. Barlow, Non-linear transmission and simple models for bovine tuberculosis, Journal of Animal Ecology 69 (4) (2000) 703–713. doi:10.1046/j.1365-2656.2000.00428.x.

[7] G. Chowell, L. Sattenspiel, S. Bansal, C. Viboud, Mathematical models to characterize early epidemic growth: A review, Physics of Life Reviews 18 (2016) 66–97. doi:10.1016/j.plrev.2016.07.005.

[8] T. Ganyani, K. Roosa, C. Faes, N. Hens, G. Chowell, Assessing the relationship between epidemic growth scaling and epidemic size: The 2014–16 ebola epidemic in west africa, Epidemiology and Infection 147 (oct 2018). doi:10.1017/s0950268818002819.

[9] G. Chowell, C. Viboud, J. M. Hyman, L. Simonsen, The western africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates, PLoS Currents (2014). doi:10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261.

[10] G. Chowell, C. Viboud, Is it growing exponentially fast? – impact of assuming exponential growth for characterizing and forecasting epidemics with initial near-exponential growth dynamics, Infectious Disease Modelling 1 (1) (2016) 71–78. doi:10.1016/j.idm.2016.07.004.

[11] G. Chowell, D. Hincapie-Palacio, J. Ospina, B. Pell, A. Tariq, S. Dahal, S. Moghadas, A. Smirnova, L. Simonsen, C. Viboud, Using phenomenological models to characterize transmissibility and forecast patterns and final burden of zika epidemics, PLoS Currents (2016). doi:10.1371/currents.outbreaks.f14b2217c902f453d9320a43a35b9583.

[12] G. Chowell, Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts, Infectious Disease Modelling 2 (3) (2017) 379–398. doi:10.1016/j.idm.2017.08.001.

[13] L. Dinh, G. Chowell, R. Rothenberg, Growth scaling for the early dynamics of HIV/AIDS epidemics in brazil and the influence of socio-demographic factors, Journal of Theoretical Biology 442 (2018) 79–86. doi:10.1016/j.jtbi.2017.12.030.

[14] W. Gardner, E. P. Mulvey, E. C. Shaw, Regression analyses of counts and rates: Poisson,

overdispersed poisson, and negative binomial models., Psychological Bulletin 118 (3) (1995) 392–404. doi:10.1037/0033-2909.118.3.392.

[15] S. J. Long, J. S. Long, J. Freese, Regression models for categorical dependent variables using Stata, Stata press, 2006.

[16] I. J. Myung, Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology 47 (1) (2003) 90–100. doi:10.1016/s0022-2496(02)00028-7.

[17] L. Held, D. S. Bové, Applied Statistical Inference, Springer Berlin Heidelberg, 2014. doi:10.1007/978-3-642-37887-4.

[18] P. McCullagh, J. Nelder, Generalized Linear Models, London: Chapman and Hall, 1989.

[19] G. J. S. Ross, D. A. Preece, The negative binomial distribution, The Statistician 34 (3) (1985) 323. doi:10.2307/2987659.

[20] P. Hougaard, M.-L. T. Lee, G. A. Whitmore, Analysis of overdispersed count data by mixtures of poisson variables and poisson processes, Biometrics 53 (4) (1997) 1225. doi:10.2307/2533492.

[21] T. Coulson, P. Rohani, M. Pascual, Skeletons, noise and population growth: the end of an old debate?, Trends in Ecology & Evolution 19 (7) (2004) 359–364. doi:10.1016/j.tree.2004.05.008.

[22] M. J. Keeling, P. Rohani, Modeling Infectious Diseases in Humans and Animals, Princeton University Press, 2008.

[23] J. O. Lloyd-Smith, Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases, PLoS ONE 2 (2) (2007) e180. doi:10.1371/journal.pone.0000180.

[24] B. Bolker, Maximum likelihood estimation and analysis with the bblme package, https://cran.r-project.org/web/packages/bbmle/vignettes/mle2.pdf, accessed: 7 March 2017 (2017).

[25] B. Efron, Bootstrap methods: Another look at the jackknife, The Annals of Statistics 7 (1) (1979) 1–26. doi:10.1214/aos/1176344552.

[26] G. Claeskens, N. L. Hjort, Model selection and model averaging, New York: Cambridge University Press., 2008.

[27] J. H. Steiger, Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis., Psychological Methods 9 (2) (2004) 164–182. doi:10.1037/1082-989x.9.2.164.

[28] WHO, World health organization ebola data and statistics 2016, http://apps.who.int/gho/data/node.ebola-sitrep.ebola-country?lang=en, accessed: 1 June 2016 (2016).

[29] G. Chowell, H. Nishiura, L. M. Bettencourt, Comparative estimation of the reproduction number for pandemic influenza from daily case notification data, Journal of The Royal Society Interface 4 (12) (2007) 155–166. doi:10.1098/rsif.2006.0161.

[30] Z. N. Kamvar, J. Cai, J. R. Pulliam, J. Schumacher, T. Jombart, Epidemic curves made easy using the r package incidence, F1000Research 8 (2019) 139. doi:10.12688/f1000research.18002.1.

[31] K. Roosa, , R. Luo, G. C. and, Comparative assessment of parameter estimation methods in the presence of overdispersion: a simulation study, Mathematical Biosciences and Engineering 16 (5) (2019) 4299–4313. doi:10.3934/mbe.2019214.

[32] W. W. Piegorsch, Maximum likelihood estimation for the negative binomial dispersion parameter, Biometrics 46 (3) (1990) 863. doi:10.2307/2532104.

[33] K. Saha, S. Paul, Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter, Biometrics 61 (1) (2005) 179–185. doi:10.1111/j.0006-341x.2005.030833.x.

[34] L. J. Willson, J. L. Folks, J. H. Young, Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter *k*, Biometrics 40 (1) (1984) 109. doi:10.2307/2530749.

[35] C. Bretó, D. He, E. L. Ionides, A. A. King, Time series analysis via mechanistic models, The Annals of Applied Statistics 3 (1) (2009) 319–348. doi:10.1214/08-aoas201.

[36] C. Bretó, Modeling and inference for infectious disease dynamics: A likelihood-based approach, Statistical Science 33 (1) (2018) 57–69. doi:10.1214/17-sts636.

[37] M. Fasiolo, N. Pya, S. N. Wood, A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology, Statistical Science 31 (1) (2016) 96–118. doi:10.1214/15-sts534.

[38] A. A. King, D. Nguyen, E. L. Ionides, Statistical inference for partially observed markov processes via the `R` package pomp, Journal of Statistical Software 69 (12) (2016). doi:10.18637/jss.v069.i12.