Estimation of the Distribution of the Individual Reproduction Number: The Case of the COVID-19 Pandemic

A. Braumann[a], J. Krampe[b], J.-P. Kreiss[a] and E. Paparoditis[c]

[a] TU Braunschweig, Germany; [b] University of Mannheim, Germany; [c] University of Cyprus, Cyprus

Abstract

We investigate the problem of estimating the distribution of the individual reproduction number governing the COVID-19 pandemic. Under the assumption of a Negative Binomial distribution, we focus on constructing estimators of the parameters of this distribution using reported infection data and taking into account issues like under-reporting and the time behavior of the infection and of the reporting processes. To this end, we extract information from regionally dissaggregated data reported by German health authorities, in order to estimate not only the mean but also the variance of the distribution of the individual reproduction number. In contrast to the mean, the variance also depends on the unknown under-reporting rate of the pandemic. A bootstrap method to construct confidence intervals for the parameters of interest is presented and the hypothesis of a Negative Binomial distribution is empirically examined. The estimates obtained allow not only for a better understanding of the time-varying behavior of the expected value of the individual reproduction number but also of its dispersion and for a discussion of the implications of different policy interventions. Our methodological investigations are accompanied by an empirical study of the development of the COVID-19 pandemic in Germany, which shows a strong overdispersion of the individual reproduction number.

1. Introduction

The individual reproduction number R is commonly used in epidemiology to quantify the transmission of a disease. R describes the number of secondary infections caused by a single SARS-CoV-2-positive individual. Of special interest is the expectation

$\mathrm{E}(R)$ of the reproduction number, often denoted as R_0 and called *basic reproduction number*. Notice that the expectation $\mathrm{E}(R) = R_0$ is only one parameter of the distribution of the random variable R. Even though we treat R as a random variable, the reproduction number also plays an important role in deterministic modeling in epidemiology which typically is based on ordinary differential equations including the Kermack-McKendrick epidemic model SIR (Susceptible-Infectious-Removed) and the SEIR (Susceptible-Exposed-Infectious-Removed) model. For an introduction to the reproduction number R and especially to the Basic Reproduction Number R_0 we refer to Chowell and Brauer (2009).

Estimators for R_0 or $R_{0,t}$, where the index t takes into account possible changes over time, on the basis of observed non in-depth case numbers can be found in many papers in the literature. A fundamental alternative would be to estimate $\mathrm{E}(R) = R_0$ from in-depth tracking of infection-chains. Dehning et al. (2020b) discuss a model-free estimation of the reproduction number $R_{0,t}$ and compare it with the standard techniques applied by the Robert Koch Institute (RKI), which is the German government's central scientific institution in the field of biomedicine. Quite important for the various RKI-estimators is the so-called generation duration or generation time.

We will follow the approach of developing methodology on the basis of non in-depth reported case numbers but we will not only focus on the expectation but also on the entire distribution of the reproduction number R. Cori et al. (2013) and also Lloyd-Smith et al. (2005a) together with the associated supplementary material (Lloyd-Smith et al. (2005b)) suggested a Negative Binomial distribution to model the stochastic behavior of R. For COVID-19-pandemic data, the Negative Binomial distribution is also used in Althouse el al. (2020) and Endo et al. (2020). Of special interest is the ability of the Negative Binomial distribution to include possible dispersion, which is rather likely to be present in the COVID-19-pandemic. Dispersion means that the standard deviation or variance of a distribution may vary independently of the mean. The latter for example is not possible for the also often used Poisson distribution, for which variance and mean coincide. We refer to Azmon et al. (2014) for Poisson modeling when describing methodology to estimate the reproduction number R.

We present estimators of important parameters of the distribution of the individual reproduction number R for the COVID-19-pandemic on the basis of non in-depth infection data provided on a daily basis in Germany by the RKI. We argued that it is important to estimate not only the expectation but also the dispersion or equivalently

the variance of the individual reproduction number R. We present assumptions under which we are able to give reasonable estimates of the variance of R which as a consequence allows us to estimate dispersion. To this end it appears to be necessary to rely on COVID-19-data on district (Landkreis) level in Germany. Furthermore, we present a bootstrap method to construct confidence intervals for the parameters of interest and we examine the hypothesis of a Negative Binomial distribution using the aforementioned infection data. As a result it will be seen that the parameters of the distribution of R indeed have changed over time and that dispersion (even over-dispersion) is rather relevant. We will make the implication of this clear.

2. Preliminaries

Cumulative data about newly reported cases, totally infected cases, fatalities as well as a 7-days incidence rate per 100,000 inhabitants can be found on the website `http://corona.rki.de`. The reported cases are based on positive laboratory testing of SARS-CoV-2 and are denoted as COVID-19-cases irrespective of COVID-19-symptoms. The data is available separately per local states (Bundesländer), districts (Landkreise), age-groups and gender, to name only a few. When considering daily COVID-19-cases it is important to carefully distinguish for individuals the time t at which an SARS-CoV-2-infection took place, the time at which the case was first reported to the (local) health authorities (if it was laboratory confirmed at all) and the time at which the case finally was reported to RKI.
For the investigations in this paper we typically consider times t at which an infection with SARS-CoV-2 takes place. It is most likely that the difference of the time of infection of a case, which in the end will be reported, and the time of reporting follows a distribution over a couple of days. On average, we assume a time shift of $\tau = 7$ days, which seems to be reasonable.

For this paper the relevant times are the infection time and the time of the first reporting. We denote by RKI_s the number of COVID-19-cases first reported at time s to the health authorities. Then, roughly speaking the number of infections at time t, which in the end will be reported, and $\mathrm{RKI}_{t+\tau}$ are strongly related to each other. Figure 1 displays the daily data RKI_t for the period April 2020 to September 2021 together with a moving average smoothing over 7 days. Averaging over 7 days seems to be appropriate since a very strong periodic behavior of RKI_t over the week is observed.

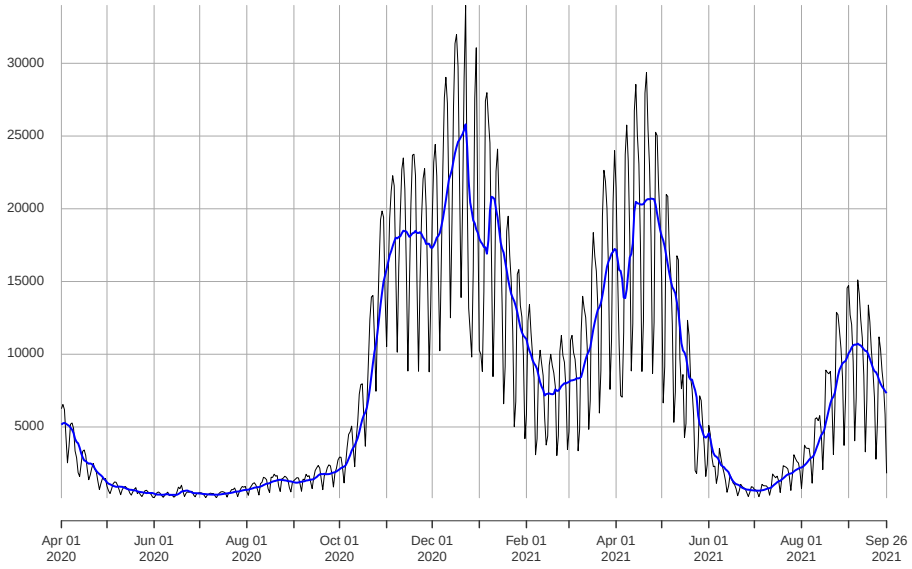

Figure 1. Daily numbers of reported COVID-19-cases, denoted by RKI_t, in black. A moving average over seven days of RKI_t is given in blue.

3. Methodology

The Negative Binomial distribution with parameters $p \in (0,1)$ and $r \in \mathbb{N}$ is well-known in statistics for modeling the number of failures in a sequence of independent Bernoulli(p)-trials until the r-th success occurs. It has been extended to parameters $p \in (0,1)$ and $r \in (0,\infty)$ (we use the abbreviation $\mathcal{NB}(p,r)$) via a consideration of a classical Poisson-distribution with random parameter λ distributed according to a specific Gamma-distribution.

The Negative Binomial distribution allows for a more flexible modeling of rare events. For $R \sim \mathcal{NB}(p,r)$ we have

$$R_0 = \mathrm{E}(R) = \frac{r\,(1-p)}{p} \quad \text{and} \quad \mathrm{Var}(R) = \frac{r\,(1-p)}{p^2}. \tag{1}$$

The Negative Binomial distribution possesses a coefficient of variation $\mathrm{CV}(R) = \mathrm{Var}(R)/\mathrm{E}(R) = 1/p$, and allows for modeling the distribution of the reproduction number R with dispersion. The dispersion parameter κ is defined through

$$\mathrm{Var}(R) = R_0 \cdot \Big(1 + \frac{R_0}{\kappa}\Big), \tag{2}$$

which leads for the Negative Binomial distribution to $\kappa = r$ (cf. Lloyd-Smith et al. (2005a)). The dispersion parameter κ and the coefficient of variation both are widely used to quantify the size of the variance given the expectation of a random variable.

We will model the number R_t of secondary infections caused by an individual COVID-19-case with infection time (day) t via a $\mathcal{NB}(p_t, r_t)$-distribution. If we further denote the number of newly infected cases at time t by N_t, we then are faced with a total of

$$\sum_{i=1}^{N_t} S_{i,t}, \tag{3}$$

secondary infections in the future. Here $S_{1,t}, S_{2,t}, \ldots$ denote i.i.d. random variables distributed according to $\mathcal{NB}(p_t, r_t)$.

For several reasons not all of these future cases will be reported to the German health authorities and subsequently will not show up in the RKI-statistics of newly laboratory-confirmed COVID-19-cases. One major, but not the only reason for this under-reporting is that a substantial number of SARS-CoV-2-infections are asymptomatic. The studies Buitrago-Garcia al. (2020) and Oran and Topol (2020) report that percentages of asymptomatic cases vary between 20% to 45%. These values coincide with results from a study in the community of Kupferzell (Germany), where a percentage of 24.5% of asymptomatic cases has been observed, cf. Santos-Hövener et al. (2020).

Although it seems difficult to assess the exact rate of under-reporting, this rate is of course a relevant quantity when investigating the development of the pandemic. Some studies state rates of not reported cases up to 80%. See for example Streeck et al. (2020) and Santos-Hövener et al. (2020). Rahmandad et al. (2020), in a study across 86 nations, found out that under-reporting varies substantially over countries, where for Germany, the estimated ratio of actual to reported cases is about 6 to 7.

We denote the proportion of COVID-19-cases reported to the German health authorities by $p_{0,t}$, and allow this rate to vary (slowly) with time. A value of $p_{0,t} \approx 0.2$ seems realistic in the light of the aforementioned studies.

From a number of N_t newly infected individuals at time t we therefore will see within the statistics of RKI only a binomial thinned selection of N_t, which we denote by N_t'. According to our assumption of a reporting rate of $p_{0,t}$ and because of a not small

number of cases it is reasonable to assume that, approximately,

$$\frac{N_t'}{N_t} \approx p_{0,t}. \tag{4}$$

It is worth mentioning that the number of reported cases N_t' out of infections happened at time t do not show up within the RKI-statistics neither at time t nor at a single time point in the future. Rather, occurrence in the RKI-statistics will spread over a span of days.

This further means that from the total number $\sum_{i=1}^{N_t} S_{i,t}$ of secondary infections caused by N_t primary infections we only observe $\sum_{i=1}^{N_t} \widetilde{S}_{i,t}$ laboratory-confirmed cases with the statistics of RKI, where $\widetilde{S}_{i,t}$ possesses a Binomial-distribution with parameters $S_{i,t}$ and $p_{0,t}$. Equivalently, the number of reported SARS-CoV-2 secondary infections out of a cohort of N_t primary infected individuals can be written as

$$\sum_{i=1}^{N_t} \sum_{j=1}^{S_{i,t}} Z_{i,j}, \tag{5}$$

where $(Z_{i,j}, i, j \in \mathbb{N})$ is a family of i.i.d. Bernoulli$(p_{0,t})$-distributed random variables. Success, i.e. $Z_{i,j} = 1$, means that a secondary infected individual gets a positive COVID-19-test at some day in the future.

Before further elaborating on this point let us take a look at the distribution of the total numbers of secondary infections $\sum_{i=1}^{N_t} S_{i,t}$ and reported secondary infections $\sum_{i=1}^{N_t} \widetilde{S}_{i,t}$. Since it is known that the Negative Binomial distribution is additive, we immediately obtain, assuming independence of the single cases, that $\sum_{i=1}^{N_t} S_{i,t} \sim \mathcal{NB}(p_t, N_t \cdot r_t)$. Moreover, $\widetilde{S}_{i,t}$ is a Binomial-thinning of $S_{i,t}$. The property that Binomial-thinning changes the parameter but not the family of Poisson-distributions carries over to the family of Negative Binomial distributions. In fact, the following holds true: If $X \sim \mathcal{NB}(p, r)$ with parameters $p \in (0, 1)$ and $r > 0$ and if $Z_1, Z_2, \ldots$ are i.i.d. Bernoulli(p_0) variables, then

$$Y := \sum_{j=1}^{X} Z_j \sim \mathcal{NB}(q, r) \quad \text{with} \quad q = \frac{p}{p + p_0 - p_0 \cdot p}.$$

This implies that we end up with

$$\widetilde{S}_{i,t} \sim \mathcal{NB}(q_t, r_t), \quad i = 1, 2, \ldots, \tag{6}$$

with

$$q_t = \frac{p_t}{p_{0,t} + p_t - p_{0,t} \cdot p_t}. \tag{7}$$

The parameter q_t depends on the percentage $p_{0,t}$ of SARS-CoV-2-infections reported to the health authorities. Furthermore,

$$\sum_{i=1}^{N_t} \widetilde{S}_{i,t} \sim \mathcal{NB}(q_t, N_t \cdot r_t). \tag{8}$$

So far we focused on time points t at which SARS-CoV-2-infections take place. As already mentioned, these time points t should not be confused with the time points at which SARS-CoV-2-infections are first reported to the health authorities (recall that we denoted the number of COVID-19-cases first reported at time point t to the health authorities by $\mathrm{RKI_t}$). We argued that there is a (random) time shift between these two time points with a likely mean of $\tau = 7$.

To keep it simple and still take time shifts as well as random fluctuations of reporting delays over a span of days into account, we make the following two assumptions

$$\sum_{s=0}^{6} N'_{t-s} \approx \sum_{s=0}^{6} \mathrm{RKI_{t+\tau-s}} \tag{9}$$

and

$$\sum_{s=0}^{6} N'_{t-s} \approx \sum_{s=0}^{6} \sum_{i=1}^{N_{t-4-s}} \widetilde{S}_{i,t-4-s}. \tag{10}$$

The first assumption means that newly infected cases, which are of a type that will be reported in the future, summed up over a week approximately will occur in the RKI-statistics also within a week but shifted by τ days to the future, while the second assumption is a relaxation of $N'_t \approx \sum_{i=1}^{N_{t-4}} \widetilde{S}_{i,t-4}$. The latter assumption would mean that secondary infections occur with a fixed time delay of 4 days to the primary infection. Instead of such a strict assumption, (10) means that the two quantities are roughly the same if they are summed up over a week. Here the number 4 can be viewed as generation time of the virus.

4. Estimation of Parameters

Based on the considerations of the previous section it follows that the smoothed estimate of the mean of the reproduction number published on a daily basis by RKI, and denoted by $\widehat{R}^7_{0,t}$ fulfills

$$\widehat{R}^7_{0,t} := \frac{\sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t-s}}}{\sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t-4-s}}} \approx \frac{\sum_{s=0}^{6} N'_{i,t-\tau-s}}{\sum_{s=0}^{6} N'_{t-4-\tau-s}} \approx \frac{\sum_{s=0}^{6} \sum_{i=1}^{N_{t-4-\tau-s}} \widetilde{S}_{i,t-4-\tau-s}}{\sum_{s=0}^{6} N'_{t-4-\tau-s}}, \tag{11}$$

cf. (9) and (10). It is important to note that $\widehat{R}^7_{0,t}$ reflects the reproduction behavior approximately 14 days ago. A plot of $\widehat{R}^7_{0,t}$ together with a moving average over 7-days is given in Figure 2.

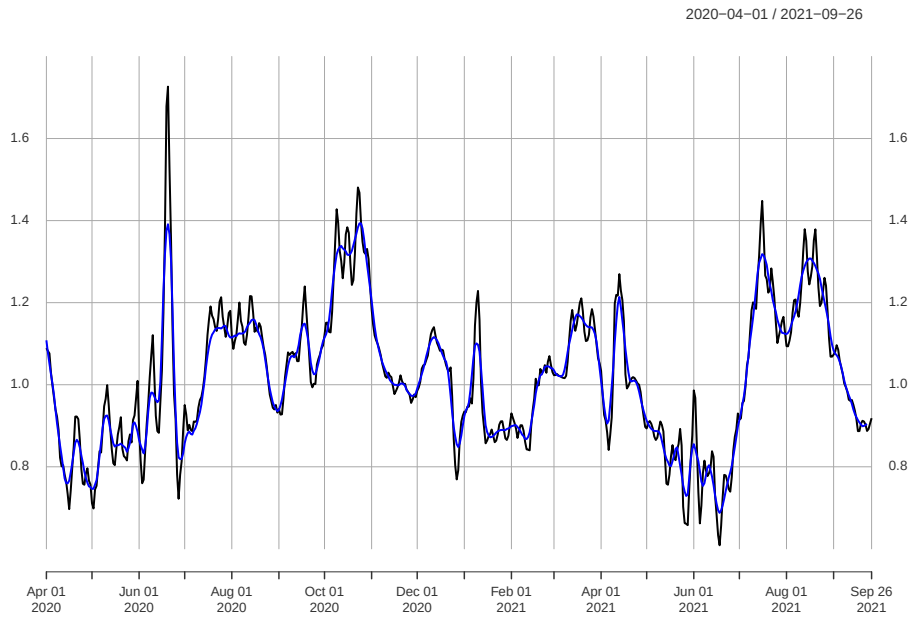

Figure 2. Estimated mean $\widehat{R}^7_{0,t}$ of the reprodution number (black line) with a moving average of order 7 (blue line).

From (8) we have that the distribution of the numerator $\sum_{s=0}^{6} \sum_{i=1}^{N_{t-4-\tau-s}} \widetilde{S}_{i,t-4-\tau-s}$, given the numbers $N'_{t-4-\tau-s}, s = 0, \ldots, 6$, approximately is (we have $q_{t-4-\tau-s} \approx q_{t-\tau-7}$ and $r_{t-4-\tau-s} \approx r_{t-\tau-7}, s = 0, \ldots, 6$)

$$\mathcal{NB}(q_{t-\tau-7}, r_{t-\tau-7} \cdot \sum_{s=0}^{6} N_{t-\tau-4-s}), \tag{12}$$

with (conditional on $\sum_{s=0}^{6} N_{t-\tau-4-s}$) expectation

$$\sum_{s=0}^{6} N_{t-\tau-4-s} \cdot \frac{r_{t-\tau-7} \cdot (1 - q_{t-\tau-7})}{q_{t-\tau-7}}. \tag{13}$$

Using the further approximation from (4) this leads to the following value that is estimated by $\widehat{R}^{7}_{0,t}$

$$\frac{\sum_{s=0}^{6} N_{t-4-\tau-s}}{\sum_{s=0}^{6} N'_{t-4-\tau-s}} \cdot \frac{r_{t-\tau-7}\,(1 - q_{t-\tau-7})}{q_{t-\tau-7}} = \frac{1}{p_{0,t-\tau-7}} \cdot \frac{r_{t-\tau-7}\,(1 - q_{t-\tau-7})}{q_{t-\tau-7}}. \tag{14}$$

Fortunately, we obtain by simple algebra and using (7), that the expectation $R_{0,t-\tau-7}$ of the reproduction number $R_{t-\tau-7}$ equals

$$E(R_{t-\tau-7}) = \frac{1}{p_{0,t-\tau-7}} \cdot \frac{r_{t-\tau-7}\,(1 - q_{t-\tau-7})}{q_{t-\tau-7}} = \frac{r_{t-\tau-7}\,(1 - p_{t-\tau-7})}{p_{t-\tau-7}}. \tag{15}$$

In order to be able to estimate both parameters r_t and p_t of a Negative Binomial fit to the distribution of the reproduction number R_t we further need an estimator of $\mathrm{Var}(R_t)$. For this we need somehow replicates of realizations of R_t, which we will obtain from reported COVID-19-cases on district level from Germany. In total, Germany is divided into about 401 districts with population numbers ranging from $34,193$ to $3,669,491$. For each district RKI provides daily COVID-19-cases along the same guidelines as for Germany as a whole. As before we denote the number of newly infected (not necessarily reported!) COVID-19-cases by $N_{t,\ell}$, where t counts the day (time) and $\ell = 1, \ldots, L = 401$ denotes the number of the district. The total number of secondary infections caused by $N_{t,\ell}$ primary infected individuals then follows a $\mathcal{NB}(p_t, N_{t,\ell} \cdot r_t)$-distribution. Following the same arguments as in Section 3 we obtain that the total number of reported SARS-CoV-2- secondary infections out of a number of $N_{t,\ell}$ primary infections in district ℓ, which we denote by $N'_{t,\ell}$, is distributed according to $\mathcal{NB}(q_t, N_{t,\ell} \cdot r_t)$, cf. (12) and (11).

In order to relate the number of daily first reported cases $\mathrm{RKI}_{\mathrm{t},\ell}$ within district ℓ with the total number of secondary infections $N'_{t,\ell}$, for which the reporting is spread over some of days, we assume in accordance with (9) and (10) for each district $\ell = 1, 2, \ldots, L$

$$\sum_{s=0}^{6} N'_{t-s,\ell} \approx \sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t}+\tau-\mathrm{s},\ell}, \tag{16}$$

and

$$\sum_{s=0}^{6} N'_{t-s,\ell} \approx \sum_{s=0}^{6} \sum_{i=1}^{N_{t-4-s,\ell}} \widetilde{S}_{i,t-4-s}. \tag{17}$$

Our main focus, when turning to reported COVID-19-cases on district level, is to obtain an estimator of the variance $\mathrm{Var}(R_{t-\tau-7}) = r_{t-\tau-7} \cdot \big(1 - p_{t-\tau-7}\big)/p^2_{t-\tau-7}$. Because of the approximation of the distribution of the numerator of $\widehat{R}^7_{0,t,\ell}$, cf. (11), by a $\mathcal{NB}(q_{t-\tau-7}, r_{t-\tau-7} \cdot \sum_{s=0}^{6} N_{t-\tau-4-s,\ell})$-distribution together with the approximation

$$\frac{\sum_{s=}^{6} N'_{t-\tau-4-s}}{\sum_{s=0}^{6} N_{t-\tau-4-s}} \approx p_{0,t-\tau-7}, \tag{18}$$

cf. (4) and also (14), we obtain for given $\sum_{s=0}^{6} N_{t-\tau-4-s} = \sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t-4-s},\ell}/\mathrm{p}_{0,\mathrm{t}-\tau-7}$, i.e., the denominator is considered fix, that

$$\mathrm{Var}(\widehat{R}^7_{0,t,\ell}) \approx \frac{1}{\sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t-4-s},\ell}\, \mathrm{p}_{0,\mathrm{t}-\tau-7}} r_{t-\tau-7}(1 - q_{t-\tau-7})/q^2_{t-\tau-7}. \tag{19}$$

This means, the variance is heterogeneous among the districts with a factor $1/\sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t-4-s},\ell}$. Taking this into account leads to the estimator (20), which is an estimator for $\mathrm{Var}(\widetilde{S}_{i,t-\tau-7})/p_{0,t-\tau-7}$, i.e., the variance of the number of *reported* SARS-CoV-2-secondary infection cases from a single infected individual scaled by $1/p_{0,t-\tau-7}$.

$$\widehat{\mathrm{Var}(\widetilde{S}_{t-\tau-7})} := \frac{1}{L} \sum_{\ell=1}^{L} \sum_{s=0}^{6} \mathrm{RKI}_{\mathrm{t-4-s},\ell} \left(\widehat{\mathrm{R}}^7_{0,\mathrm{t},\ell} - \widehat{\mathrm{R}}^7_{0,\mathrm{t}} \right)^2. \tag{20}$$

Since the distribution of $\widetilde{S}_{i,t}$ is $\mathcal{NB}(q_t, r_t)$ we obtain

$$\begin{aligned} \frac{1}{p_{0,t}} \cdot \mathrm{Var}(\widetilde{S}_{i,t}) &= \frac{r_t \cdot (1 - q_t)}{p_{0,t} q_t^2} \\ &= \frac{r_t \cdot (1 - p_t)}{p_t^2} \cdot (p_{0,t} + p_t - p_{0,t} p_t) \\ &= \mathrm{Var}(R_t) \cdot (p_{0,t} + p_t - p_{0,t} p_t), \end{aligned} \tag{21}$$

that is, for $p_t > 0$,

$$\mathrm{Var}(R_t) = \frac{1}{(p_{0,t} + p_t - p_{0,t} p_t)} \cdot \frac{1}{p_{0,t}} \mathrm{Var}(\widetilde{S}_{i,t}). \tag{22}$$

As it is seen, and in contrast to the expectation (cf. (15)), the variance of the reproduction number based on COVID-19-cases reported to the health authorities depends on

the unknown reporting rate $p_{0,t}$. Since the reporting rate $p_{0,t}$ cannot be estimated from reported data we only can calculate variance estimators for a variety of assumed reporting rates $p_{0,t}$ (see Figure 3). Based on estimators $\widehat{\mathrm{E}R_{t-14}} := \widehat{R}^7_{0,t}$, cf. (11), with $\tau = 7$, $\widehat{\mathrm{Var}(\widetilde{S}_{t-14})}$ as given in (20) and because of (22) together with explicit expressions of expectation and variance of the Negative Binomial distribution $\mathcal{NB}(p_{t-14}, r_{t-14})$ assumed for R_{t-14}, we finally are led to the following estimators $\widehat{p}_{t-14}$ and $\widehat{r}_{t-14}$ of the parameters of this distribution for any assumed and fixed value of the reporting rate $p_{0,t-14}$ and the choice of $\tau = 7$:

$$\widehat{p}_{t-14} = \frac{\widehat{R}^7_{0,t} \cdot p_{0,t-14}}{\widehat{\mathrm{Var}(\widetilde{S}_{t-14})} - \widehat{R}^7_{0,t} \cdot (1 - p_{0,t-14})} \quad \text{and} \quad \widehat{r}_{t-14} = \frac{\widehat{R}^7_{0,t} \cdot \widehat{p}_{t-14}}{1 - \widehat{p}_{t-14}}\,. \tag{23}$$

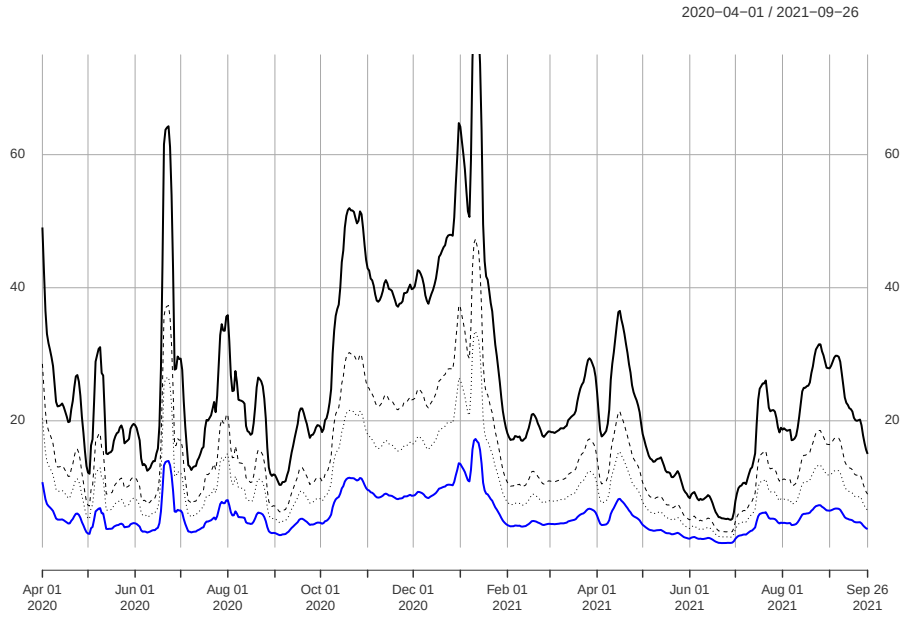

Figure 3. The estimates of the variances $\mathrm{Var}(\widetilde{S}_{t-\tau-7})/p_{0,t-\tau-7}$ (in blue) and $\mathrm{Var}(R_{t-\tau-7})$ for reporting rates $p_0 = 0.2$ (black solid line), $p_0 = 0.35$ (black dashed line) and $p_0 = 0.5$ (black dotted line).

5. Bootstrap Confidence Intervals

It is important to construct confidence intervals for the unknown mean $R_{0,t}$. According to our previous discussion, the numerator of the estimator $\widehat{R}^7_{0,t}$ approximately satisfies

for $\tau = 7$

$$\sum_{s=0}^{6} RKI_{t-s} \sim \mathcal{NB}\big(q_{t-14}, r_{t-14} \cdot \sum_{s=0}^{6} RKI_{t-s-4}/p_{0,t-14}\big); \tag{24}$$

see also the discussion before and after equation (18) for the same property for observations obtained at the district level. Recall that this distribution depends on the unknown under-reporting rate $p_{0,t-14}$. Based on expression (24) the following parametric bootstrap procedure is proposed for constructing a confidence interval for the mean $R_{0,t-14}$ of the individual reproduction number.

Step 1: For $p_{0,t-14}$ given and estimates $\widehat{q}_{t-14}$ and $\widehat{r}_{t-14}$, we generate for $t = 15, 16, \ldots, n$, pseudo random variables $(\sum_{s=0}^{6} RKI_{t-s})^*$ distributed as

$$\big(\sum_{s=0}^{6} RKI_{t-s}\big)^* \sim \mathcal{NB}\big(\widehat{q}_{t-14}, \widehat{r}_{t-14} \cdot \big(\sum_{s=0}^{6} RKI_{t-s-4}\big)^*/p_{0,t-14}\big), \tag{25}$$

using the starting values

$$\big(\sum_{s=0}^{6} RKI_{t-s-4}\big)^* = \sum_{s=0}^{6} RKI_{t-s-4},$$

for $t = 15, 16, 17$ and 18.

Step 2: Calculate for $t = 15, 16, \ldots, n$ the pseudo estimator

$$\widehat{R}_t^* = \frac{\big(\sum_{s=0}^{6} RKI_{t-s}\big)^*}{\big(\sum_{s=0}^{6} RKI_{t-s-4}\big)^*}.$$

Step 3: Repeat Step 1 and Step 2 a large number of times, say B times, and denote by

$$\widehat{R}_{t,1}^*,\ \widehat{R}_{t,2}^*, \ldots,\ \widehat{R}_{t,B}^*$$

the pseudo-random variables obtained for $t \in \{15, 16, \ldots, n\}$.

Step 4: For a desired $1 - \alpha$ confidence level, let $Q_1 = [B * \alpha/2]$ and $Q_2 = [B * (1 - \alpha/2)]$. A $(1 - \alpha)100\%$ confidence interval for $R_{0,t}$ is then given by

$$\big[\widehat{R}_{t,(Q_1)}^*,\ \widehat{R}_{t,(Q_2)}^*\big],$$

where $\widehat{R}_{t,(1)}^*, \widehat{R}_{t,(2)}^*, \ldots, \widehat{R}_{t,(B)}^*$ denotes the ordered values of the random sample $R_{t,i}^*$, $i = 1, 2, \ldots, B$, generated in Step 3.

Figure 4 shows the 7-days moving average estimate of $R_{0,t}$ constructed using $\widehat{R}^7_{0,t}$ (see also Figure 2), together with the corresponding 95% pointwise confidence intervals constructed using the bootstrap algorithm proposed in this section.

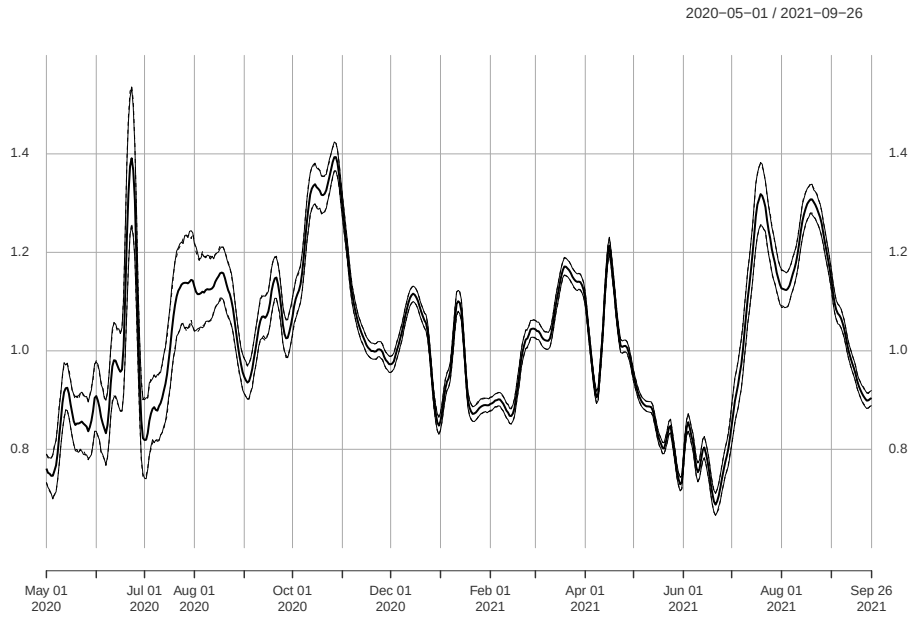

Figure 4. 7-days moving average estimates of $R_{0,t}$ together with 95% pointwise confidence intervals for two different reporting rates $p_0 = 0.2$ (solid) and $p_0 = 0.5$ (dotted) which can hardly be distinguished.

6. Validation of The Negative Binomial Hypothesis

We first investigate the suitability of the assumed Negative Binomial distribution for describing the random behavior of the individual reproduction number. Toward this end and as for estimating the variance, we focus on reported COVID-19-cases on district level, i.e., on the observations $RKI_{t,\ell}$. This allows for getting replicates of the random variable of interest and, therefore, for testing the hypothesis that the individual reproduction number follows a Negative Binomial distribution. Recall that in our discussion in Section 3, it was assumed that the sum of reported cases over the time points $t, t-1, \ldots, t-6$ in district ℓ, that is, $\sum_{s=0}^{6} RKI_{t-s,\ell}$, satisfies ($\tau = 7$),

$$\sum_{s=0}^{6} RKI_{t-s,\ell} \sim \mathcal{NB}(q_{t-14}, r_{t-14} \cdot \sum_{s=0}^{6} N_{t-11-s,\ell}).$$

Assuming that $p_{0,t}$ remains essentially constant in the range of τ days, we get using (4) and (9) that

$$\sum_{s=0}^{6} N_{t-11-s,\ell} \approx p_{0,t-14} \sum_{s=0}^{6} N'_{t-11-s,\ell} \approx p_{0,t-14} \sum_{s=0}^{6} RKI_{t-4-s,\ell}.$$

That is, in terms of the observed RKI data, the assumption we have to test translates to

$$\sum_{s=0}^{6} RKI_{t-s,\ell} \sim \mathcal{NB}\big(q_{t-14},\ \frac{1}{p_{0,t-14}} \cdot r_{t-14} \cdot \sum_{s=0}^{6} RKI_{t-4-s,\ell}\big). \tag{26}$$

In order to select from the existing data appropriate samples for testing the above assumption, we proceed as follows. We first select all districts for which the average of reported cases at time points $t-4, t-5, \ldots, t-11$ is approximately the same. Practically, this means that we consider districts for which

$$\frac{1}{7} \sum_{s=0}^{6} RKI_{t-4-s,\ell} \in [15, 25]. \tag{27}$$

We have experienced that chosen a number of average daily infections at district level outside the above interval, leads to the selection of a relatively small number of districts, that is to a small sample size. Let $L_{t,S}$ be the total number of districts at time point t satisfying condition (27) and let $\{1, 2, \ldots, L_{t,S}\}$ be the corresponding set of districts. From the total number of time points t available, we further only consider those time points, for which $L_{t,S} \geq 75$. This ensures that a sufficiently large number of districts is available for testing the hypothesis of interest. After applying this selection procedure to the $RKI_{t,\ell}$ data, we end up with a total of $T = 45$ data points t for which the corresponding condition (27) and $L_{t,S} \geq 75$ is satisfied.

The problem of testing the goodness-of-fit of a Negative Binomial distribution $\mathcal{NB}(p, r)$ when both parameters p and r are unknown, has been considered by some authors in the literature; see Meintanis (2005) and Best et al. (2009). Meintanis (2005) proposed a test based on the comparison of the empirical probability generating function with that of the Negative Binomial distribution with estimated parameters. Best et al. (2009) considered tests based on the comparison of third and fourth order moments. In the following we focus on the test proposed by Meintanis (2005) and we also report results for the test proposed by Best et al. (2009).

We use the test statistic proposed by Meintanis (2005) with suggested parameter

$a = 5$. To obtain critical values for this test, a parametric bootstrap procedure is used. More specifically, i.i.d. random samples of length $L_{t,S}$ are generated from a $\mathcal{NB}(\widehat{q}_t, \widehat{r}_t)$ distribution, where

$$\widehat{q}_t = \frac{\widehat{r}_t}{\widehat{r}_t + \overline{Y}_{L_{t,S}}} \quad \text{and} \quad \widehat{r}_t = \frac{L_{t,S}^{-1} \sum_{j=1}^{L_{t,S}} Y_{t,j}^2}{S_{L_{t,S}}^2 - \overline{Y}_{L_{t,S}}}.$$

The distribution of the test statistic under the null is then estimated using the distribution of the same test statistic calculated using the bootstrap pseudo random sample.

Applying the above test to the RKI data sets selected according to the described procedure, the null hypothesis of a Negative Binomial distribution has been rejected at the 5% level in only 17 out of the 182 different data sets considered. Qualitatively the same result is obtained, if one uses the test proposed by Best et al. (2009). Applying this test leads to a rejection of the the null hypothesis at the 5% level, in only 3 out of the 182 data sets selected. To summarize, our testing procedures find, therefore, no evidence against the assumption that the random behavior of the individual reproduction number R is governed by a Negative Binomial distribution.

7. Empirical Results

We present the estimated parameters of the Negative Binomial distribution obtained for Germany based on the RKI data set using the method developed and the time period April 1, 2020 to September 26, 2021. As mentioned in Section 4, the parameter estimates depend on the unknown reporting rate $p_{0,t}$. We present results for three possible reporting rates, i.e., $p_{0,t} \in \{0.2, 0.35, 0.5\}$. The estimated parameters p_t and r_t are given in Figures 5 and 6 respectively. Note that the dispersion parameter for the Negative Binomial distribution coincides with the parameter r_t (cf. Lloyd-Smith et al. (2005a)). Therefore, Figure 6 is also an illustration of the behavior of the dispersion parameter κ_t over time.

The parameter estimates can be translated into probabilities that an individual case causes a given number of secondary infections over its entire infectious period. We present in Figure 7 the probability that an individual causes no infections, in Figure 8 the probability that an individual causes one to five infections, and in Figure 9 the probability that an individual causes 20 or more infections.

Note that over the entire period non-pharmaceutical measures were in place such

as mandatory wearing of face masks in public areas, detected cases and contacts were quarantined, etc. This also can be seen in the estimates of the parameter R_0. Over this time period, the average of R_0 is 1.017, and consequently, far less than the reproduction rate without any measures, which is estimated as 3.32 by Alimohamadi et al. (2020) in a meta-study. Over the entire time period a strong overdispersion can be observed irrespective of the reporting rate. Smaller reporting rates lead in general to smaller parameter values for r_t. Furthermore, over both summer periods larger parameter values for r_t can be observed than in the fall periods. The overdispersion can be well displayed using probabilities. During the summer period 2020 the probability that an individual cases causes no infection is given by $60\% - 80\%$ and it rises in the fall period to $80\% - 90\%$. A similar behavior can be observed for 2021. Additionally, the probability that an individual case causes 20 or more infections almost doubles from summer to fall and peaks in October 2020 with values about $1.5\% - 2\%$. In contrast, the probability that an individual case causes one to five infections almost halves from summer to fall 2020 with values in summer of about $10\% - 20\%$. For the summer and early fall 2021 the situation looks a bit different. This may be due to the vaccination (which was absent in 2020) and as a consequence to a clear reduction in social distancing measures and to an increasing holiday travel behavior.

Endo et al. (2020) estimated the overdispersion parameter r_t of the Negative Binomial distribution as 0.1 with a 95% confidence interval from 0.04 to 0.2. Note however, that their considered time period is January and February of 2020. In that time period non-pharmaceutical measures such as obligatory face masks in public areas were not yet in place or less strict than in the time period considered here. Since in our considered time period non-pharmaceutical measures were less strict in Germany during the summer 2020, it seems most reasonable to compare the results obtained in the summer period, June 1, 2020 to October 31, 2020, with the results obtain by Endo et al. (2020). In the summer period, r_t takes values in the range of 0.025 to 0.22 and for a reporting rate of 0.35, we obtain values in the range of 0.044 to 0.157 with a mean value of 0.095. Hence, the obtained values coincide well with the values obtained in Endo et al. (2020).

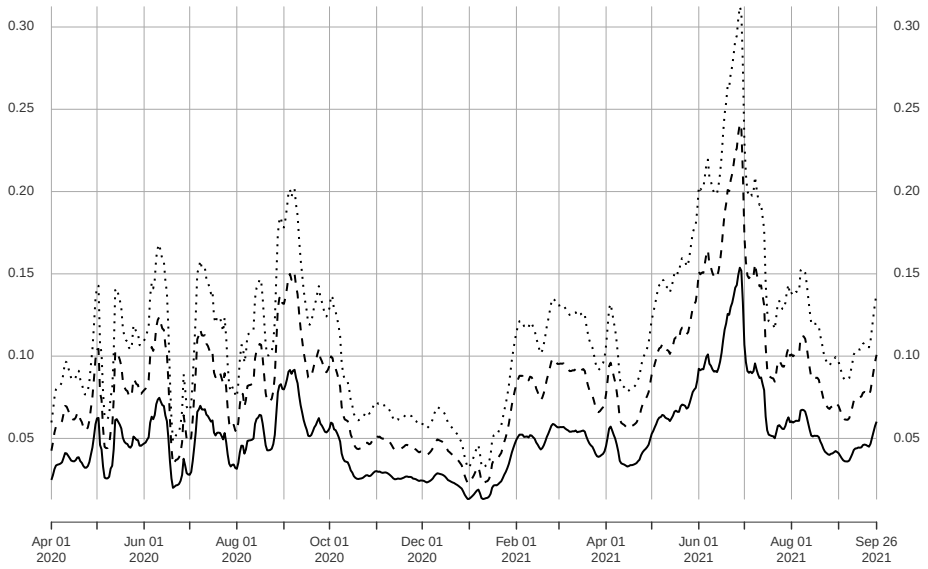

Figure 5. The estimated parameter p_t for $p_0 = 0.2$ (black solid line), $p_0 = 0.35$ (black dashed line), $p_0 = 0.5$ (black dotted line).

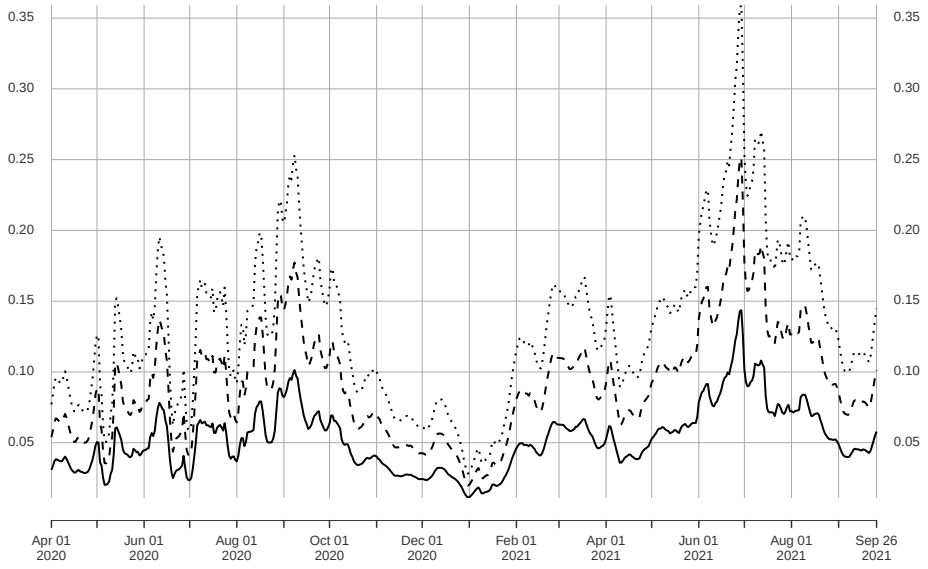

Figure 6. The estimated parameter r_t (which coincides with the dispersion parameter κ_t) for $p_0 = 0.2$ (black solid line), $p_0 = 0.35$ (black dashed line), $p_0 = 0.5$ (black dotted line).

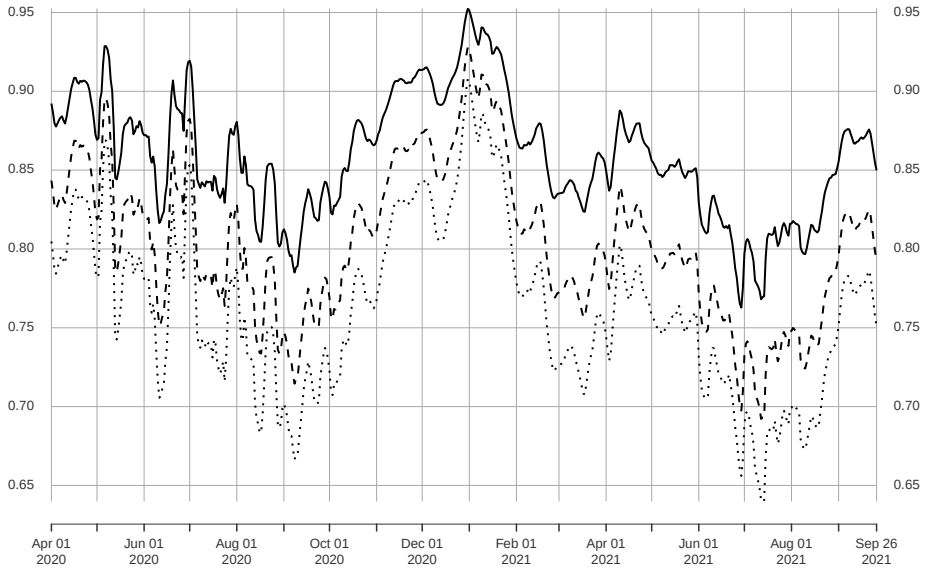

Figure 7. The probability that an individual causes no secondary infection for $p_0 = 0.2$ (black solid line), $p_0 = 0.35$ (black dashed line), $p_0 = 0.5$ (black dotted line)..

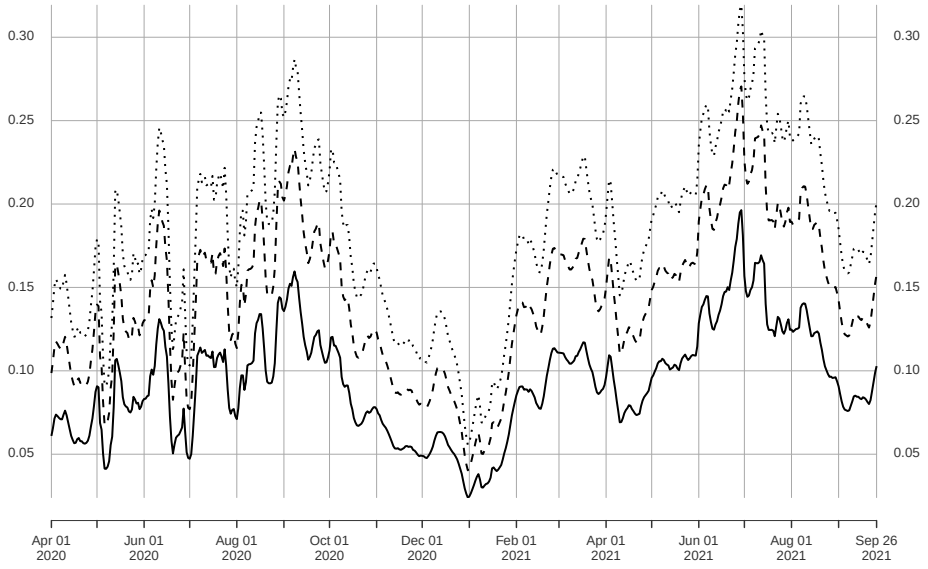

Figure 8. The probability that an individual causes between 1 and 5 secondary infections for $p_0 = 0.2$ (black solid line), $p_0 = 0.35$ (black dashed line), $p_0 = 0.5$ (black dotted line)..

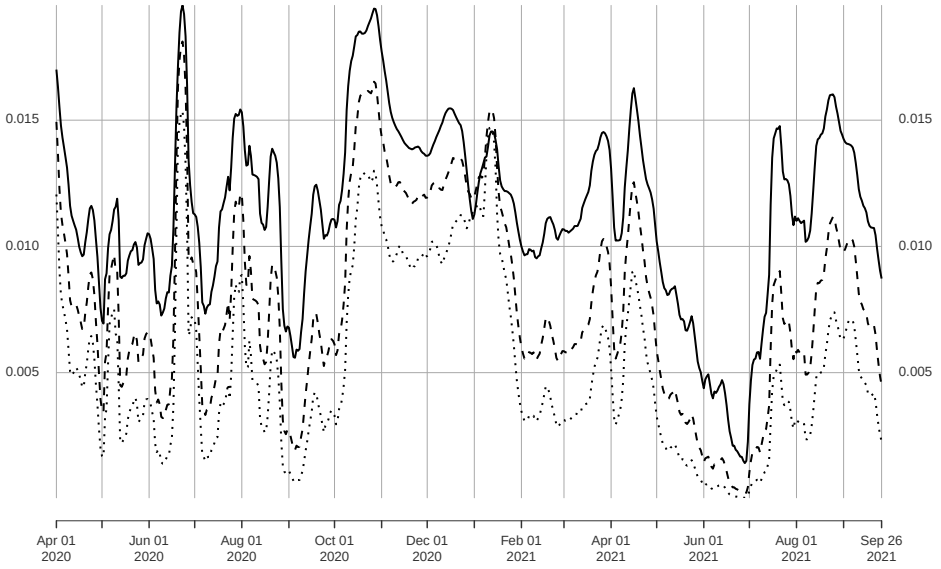

Figure 9. The probability that an individual causes 20 or more secondary infections for $p_0 = 0.2$ (black solid line), $p_0 = 0.35$ (black dashed line), $p_0 = 0.5$ (black dotted line)..

8. Appendix

Lemma 1. *Let $X \sim \mathcal{NB}(p, r)$ with parameters $p \in (0, 1)$ and $r > 0$. If $Z_1, Z_2, \ldots$ are i.i.d. Bernoulli(p_0) variables, then*

$$Y := \sum_{j=1}^{X} Z_j \sim \mathcal{NB}(q, r) \quad \textit{with} \quad q = \frac{p}{p + p_0 - p_0 \cdot p}.$$

Proof: Notice first that for $X = n$ given, the distribution of $Y|X = n$ is $Binomial(n, p_0)$. Furthermore, if $X \sim \mathcal{NB}(p, r)$ then $X \sim Poisson(\lambda)$ with $\lambda \sim Gamma\big(r, p/(1-p)\big)$; see (**??**). From these we get for $k \in \mathbb{N} \cup \{0\}$,

$$\begin{aligned} P(Y = k) &= \sum_{n=k}^{\infty} P(Y = k|X = n) \cdot P(X = n) \\ &= \sum_{n=k}^{\infty} \begin{pmatrix} n \\ k \end{pmatrix} p_o^k (1 - p_0)^{n-k} \int_0^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} \cdot f_{r, \frac{p}{1-p}}(\lambda) d\lambda \\ &= \int_0^{\infty} \frac{(p_0 \cdot \lambda)^k e^{-\lambda}}{k!} \sum_{s=0}^{\infty} \frac{(1 - p_0)^s \lambda^s}{s!} \cdot f_{r, \frac{p}{1-p}}(\lambda) d\lambda \end{aligned}$$

$$= \int_0^\infty \frac{(p_0 \cdot \lambda)^k e^{-p_o \cdot \lambda}}{k!} \cdot f_{r,\frac{p}{1-p}}(\lambda) d\lambda$$
$$= \int_0^\infty \frac{\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!} \cdot \frac{1}{p_0} f_{r,\frac{p}{1-p}}\Big(\frac{\widetilde{\lambda}}{p_0}\Big) d\widetilde{\lambda},$$

where the last equality follows using the substitution $\widetilde{\lambda} = p_o \cdot \lambda$. Since

$$\frac{1}{p_0} f_{r,\frac{p}{1-p}}\Big(\frac{\widetilde{\lambda}}{p_0}\Big) = \frac{1}{\Gamma(r)}\Big(\frac{p}{(1-p)p_0}\Big)^r \widetilde{\lambda}^{r-1} e^{-\widetilde{\lambda}\frac{p}{(1-p)p_0}}$$
$$= f_{r,\frac{q}{1-q}}\big(\widetilde{\lambda}\big),$$

where $q = p/(p + p_0 - p_0 \cdot p)$, we get

$$P(Y = k) = \int_0^\infty \frac{\widetilde{\lambda}^k e^{-\widetilde{\lambda}}}{k!} \cdot f_{r,\frac{q}{1-q}}(\widetilde{\lambda}) d\widetilde{\lambda},$$

which is the probability function of the $\mathcal{NB}(q, r)$ distribution.

□

References

Alimohamadi, Y., Taghdir, M., and Sepandi, M. (2020). The estimate of the basic reproduction number for novel coronavirus disease (COVID-19): a systematic review and meta-analysis. Journal of Preventive Medicine and Public Health.

Althouse, B. M., Wenger, E. A., Miller, J. C., Scarpino, S. V., Allard, A., Hébert-Dufresne, L., and Hu, H. (2020). Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2. arXiv preprint arXiv:2005.13689.

Azmon, A., Faes, C. and Hens, N. (2014). On the Estimation of the Reproduction Number Based on Misreported Epidemic Data. *Statistics in Medicine* **33**, 1176–1192.

Best, D.J., Rayner, C.W. and Thas, O. (2009). Anscombe's Test of Fit for the Negative Binomial Distribution. *Journal of Statistical Theory and Practice* **3**, 555–565.

Buitrago-Garcia, D. C., Egli-Gany, D., Counotte, M. J., Hossmann, S., Imeri, H., Salanti, G. and Low, N. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis *PLoS medicine **17(9)**, e1003346.*

Chowell, G. and Brauer, F. (2009). The Basic Reproduction Number of Infectious Diseases: Computation and Estimation Using Compartmental Epidemic Models.

In: G. Chowell, J.M. Hyman, L.M.A. Bettencourt and C Castillo-Chavez (eds.) *Mathematical and Statistical Estimation Approaches in Epidemiology*, Springer, Dordrecht, 1–30.

Cori, A., Ferguson, N.M., Fraser, C. and Cauchemez, S. (2013). A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology* **178**, 1505–1512.

Dehning, J., Zierenberg, J., Spitzner, F.P., Wibral, M., Neto, J.P., Wilczek, M. and Priesemann, V. (2020a). Inferring Change Points in the Spread of COVID-19 Reveals the Effectiveness of Interventions. *Science* **369**, 1–9.

Dehning, J., Spitzner, F.P., Linden, M.C., Mohr, S.B., Neto, J.P., Zierenberg, J., Wibral, M., Wilczek, M. and Priesemann, V. (2020b). Model-Based and Model-Free Characterization of Epidemic Outbreaks. *MedRxiv* Preprint doi: `https://doi.org/10.1101/2020.09.16.20187484`

Endo, A., Abbott, S., Kucharski, A. J., and Funk, S. (2020). Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Research, 5(67), 67.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E. and Getz, W.M. (2005). Superspreading and the Effect of Individual Variation on Disease Emergence. *Nature* **438**, 355–359.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E. and Getz, W.M. (2005). Superspreading and the Effect of Individual Variation on Disease Emergence: Supplementary Information.

Verlautbarung des Robert-Koch-Instituts (2020). Erläuterung der Schätzung der zeitlich variierenden Reproduktionszahl R. *Robert-Koch-Institut*.

Meintanis, S.G. (2005). Transform Methods for Testing the Negative Binomial Hypothesis. *Statistica* **LXV**, 293–300.

Oran, D. P. and Topol, E. J. (2020). Prevalence of Asymptomatic SARS-Cov-2 Infection. *Annals of Internal Medicine*.

Rahmandad, H., Lim, T.Y. and Sterman, J. (2020). Estimating COVID-19 Under-Reporting Across 86 Nations: Implications for Projections and Control. *MedRxiv* Preprint doi: `https://doi.org/10.1101/2020.06.24.20139451`

Santos-Hövener, C., Neuhauser, H.K, Schaffrath Rosario, A., Busch, M., Schlaud, M., Hoffmann, R., Gößwald, A., Koschollek, C., Hoebel. J., Allen. J., Haack-Erdmann, A., Brockmann. S., Ziese, T., Nitsche, A., Michel, J., Haller, S.,

Wilking, H., Hamouda, O., Corman, V.M., Drosten, C., Schaade, L., Wieler, L., CoMoLo Study Group, Lampert, T. (2020). Serology- and PCR-Based Cumulative Incidence of SARS-CoV-2 Infection in Adults in a Successfully Contained Early Hotspot (CoMoLo study), Germany, May to June 2020. *Euro Surveill.* **25(47)**:pii=2001752. `https://doi.org/10.2807/1560-7917.ES.2020.25.47.2001752`

Streeck, H., Schulte, B., Kümmerer, B.M. et al (2020). Infection Fatality Rate of SARS-CoV-2 in a Super-Spreading Event in Germany. *Nature Communications* **11**, article number: 5829 (2020). https://doi.org/10.1038/s41467-020-19509-y