Estimation of the Reproduction Number, Generation Time Distribution and Serial Interval Distribution

Boris Luttikhuizen and Joris de Jong

June 2021

1 Introduction

The COVID-19 epidemic has had and still has a disastrous impact on people everywhere. Since the beginning of 2021, the world has been revolving around the SARS-CoV-2 virus. Countless organizations have put in an immense effort into mapping the characteristics of the virus, which is necessary for making decisions on how to control the disease. This paper is a comparative study of some of the latest and most promising methods with regard to the effective reproduction number, generation time distribution and serial interval distribution of COVID-19.

The *effective reproduction number* of COVID-19 helps governments determine the rate at which the virus is spreading through their countries. This so-called R value is determined by the average number of secondary infections caused by a primary case. In part, it can also determine the effect of the measures taken by the government. However, there is a variety of methods available. Most of the estimates of R are based on the number of daily infections and either the generation time distribution or the serial interval distribution.

The *generation time distribution* is the probability distribution of the generation interval, which is the time between the infection of a primary and secondary case. One could assume that the generation interval is equal to the serial interval, which is the time between symptoms onset in a primary and secondary case. However, it has been shown that this results in a biased estimation of the effective reproduction number [2]. Nevertheless, one would like to infer the generation time distribution from the serial interval distribution, because it is easier to obtain in practice. For this, we need to consider the so-called *incubation time*, which is the time between infection and symptom onset. We will discuss two methods that estimate the generation time in this report. The first method estimates the generation time from a data set containing serial intervals and the second method estimates the generation time from a smaller data set containing the generation intervals.

This report discusses different methods for estimating: (1) the effective reproduction number and (2) the generation time distribution and/or the serial interval. The aim of this report is to assess which methods are the most reliable and therefore the most suitable for basing decisions on.

2 Methodology

As mentioned in the introduction, this paper discusses the estimation of the effective reproduction and the generation time. As either the generation time distribution or the serial interval distribution is required to estimate the effective reproduction number, this section starts with the methodology for the generation time distribution and serial interval distribution and then continues with the methodology for estimating the effective reproduction number.

2.1 Generation Time Distribution and Serial Interval Distribution

Most of the methods for the computation of the effective reproduction number require information about either the serial interval (time between symptom onset in a primary and secondary case) or the generation interval (time between infection of a primary and secondary case). Two different methods for the computation of these distributions will be evaluated in this section.

2.1.1 Ganyani et al.

In this subsection, we discuss the methods in the paper by Ganyani et al [6]. As mentioned there is a difference between the generation interval and the serial interval that depends on the incubation period. Ganyani et al. refer to a paper by Britton and Tomba [2], which shows that the assumption that there is no difference between the serial and generation interval has the potential to lead to a bias in the calculation of the effective reproduction number. Ganyani et al. use a Markov Chain Monte Carlo (MCMC) approach in their estimation of the generation time and serial interval distributions.

The data used by Ganyani et al. contains symptom onset data and cluster information for COVID-19 cases in Singapore (91 cases) and China (135 cases) at the start of 2020. Some of the cases are supplied with their infector, but for the missing links they use the supplementary information. The possible infectors within a cluster are determined based on contact information and symptom onset date.

Under the assumption that the incubation period is time-independent, the serial time Z_i of a person i infected by a person $v(i)$ can be written as

$$\begin{aligned} Z_i &= (t_i + \delta_i) - (t_{v(i)} + \delta_{v(i)}) \\ &= (t_i - t_{v(i)}) + (\delta_i - \delta_{v(i)}) \\ &= X_i + Y_i, \end{aligned}$$

where t_i and δ_i are the time of infection and incubation period of person i respectively. The assumption that the $X_i \overset{IID}{\sim} f(x;\Theta_1)$ and $\delta_i \overset{IID}{\sim} k(\delta,\Theta_2)$ is made, resulting in a distribution $Y_i \sim g(y_i;\Theta_2)$. An observation of the serial interval, z_i, can then be written as $z_i = x_i + y_i$, combining this with the earlier mentioned distributions implies that z_i follows a distribution with parameters $\Theta = \{\Theta_1, \Theta_2\}$. For the density of this distribution, Ganyani et al. refer to Mood et al. [9]

$$h(z;\Theta) = \int_{-\infty}^{\infty} f(z-y;\Theta_1)g(y;\Theta_2)dy.$$

This integral does not necessarily have a closed-form solution, which is why Ganyani et al. estimate it using Monte Carlo methods. The integral can be approximated as follows

$$\begin{aligned} h(z;\Theta) &= \int_{-\infty}^{\infty} f(z-y;\Theta_1)g(y;\Theta_2)dy \\ &= E_Y[f(z-y;\Theta_1)] \\ &\approx \frac{1}{J}\sum_{j=1}^{J} f(z-y_j;\Theta_1), \end{aligned}$$

where J is the number of Monte Carlo samples and y_j is sampled from $g(y, \Theta_2)$. The likelihood produced by this density is as follows

$$L(\Theta) = \prod_{i=2}^{n} \frac{1}{J} \sum_{j=1}^{J} f(z_i - y_j|\Theta).$$

A matrix of possible infectors is determined for each of the missing links and they are assigned with equal probability in each of the iterations to estimate the generation time and serial interval. The algorithm proceeds along the following two steps:

1. Sample the missing links with equal probabilty.
2. Propose a new θ.
3. Accept or reject the proposed θ
4. Repeat until desired number of iterations.

The assumption made by Ganyani et al. is that both the generation interval and the incubation period follow a gamma distribution, i.e. $f(x; \Theta_1) = \Gamma(a_1, \beta_1)$ and $k(x; \Theta_2) = \Gamma(a_2, \beta_2)$. The parameters θ_2 are fixed at $(a_2 = 3.45; \beta_2 = 0.66)$ according to results in [14]. These parameters correspond to a mean incubation period of 5.2 days and a standard deviation of 2.8 days (they analyze the robustness of their methods by considering two alternative settings). The prior distributions of Θ_1 are chosen to be uniform and the selection of any possible missing link within a cluster is assigned equal probability. The θ_1 parameters are sampled using a Metropolis-Hastings sampling scheme.

2.1.2 Nishiuraa et al.

Another method for deriving the serial interval distribution can be found in the paper by Hiroshi Nishiuraa, Natalie M. Lintona and Andrei R. Akhmetzhanov [10]. In their study, they estimated the serial interval distribution by studying the time of symptom onset between 28 infectee-infector pairs gathered from publications on COVID-19 in 2020. The reliability of the publications was ranked as either "certain" or "probable" and estimates are provided for both the "certain" and complete data sets. Similarly to Ganyani et al., they use a Bayesian approach to obtain estimates of the serial interval distribution. The likelihood used by Nishiuraa et al. is [12]

$$L(\Theta) = \prod_{i} \int_{E_{L,i}}^{E_{R,i}} \int_{S_{L,i}}^{S_{R,i}} g(e) f(s-e) ds de,$$

where, in contrast to the last section, i does not represent a single case but an infectee-infector pair. For these pairs, $(E_{R,i}, E_{L,i})$ and $(S_{R,i}, S_{L,i})$ are the intervals for symptom onset in infector and infectee respectively. The paper uses intervals instead of exact dates, as it is often the case that the exact moment of symptom onset is unknown. The $g(.)$ density models the exposure, which is said to an uninformative uniform distribution and the $f(.)$ density is assumed to one of three distributions: Log-Normal, Gamma or Weibull.

They also account for the fact that all of their data was gathered during the early phase of the pandemic using the so-called right truncated pdf

$$f'^{(s-e,e)} = \frac{f(s-e)}{\int_0^{T-e} \frac{r \exp(-ru)}{1-r \exp(-ru)} F(T-e-u) du},$$

where T represents the time point of the latest observation and r is the exponential growth rate (estimated at 0.14 by Jung et al. [8]). They provide both estimates for the truncated and non-truncated case.

They sample from the posterior distributions of the parameters using the *rstan* package (MCMC). This allows them to estimate the most likely parameters for each of the three distributions. To compare the distributions, they use the widely applicable information criterion (WAIC). WAIC is a generalization of AIC for models that do not have a positive definite Fischer Information matrix for all possible parameters, which can be the case for the fits provided by the *rstan* package.

2.2 Effective Reproduction Number

The spread of a virus is partly determined by the reproduction number, which indicates the number of people that got infected from a single infection. However, as a society is dynamic and prevention measures are taken to contain the virus, the reproduction number can be estimated over several points in time, thus we are interested in the time-dependent reproduction number, or *effective reproduction number* R. There are several methods for estimating R, all of which have been implemented in a package in the programming language R. In our research, we use three different R packages to estimate the effective reproduction number R: R0, EpiEstim, and EpiNow2. For simplicity, *R0* will be referred to as *R0* and similar for *EpiEstim* and *EpiNow2*.

R0 is written as part of a paper by T. Obadia et al.[11] in which they argued that several methods for estimating transmission parameters during a virus outbreak have been proposed, but that no software exists that implements these methods, allowing for easy comparison. They filled that gap and wrote R0 containing methods based on exponential growth, maximum likelihood estimation, sequential Bayesian methods, and a method provided by Wallinga and Teunis [13].

EpiEstim was written as part of an article by A. Cori et al.[3] in which the reproduction number is estimated by studying the ratio of the number of new infections at time t to the total infectiousness of all infected cases at time t. To support their method, the R package EpiEstim was written.

EpiNow2 is the most current package for estimating the time-varying reproduction number as well as other time-varying epidemiological parameters. It makes use of roughly the same methods as EpiEstim, however, it also accounts for a number of assumptions that EpiEstim does not.

The three R packages will be discussed, starting with R0, then continuing with EpiEstim, and concluding with EpiNow2.

2.2.1 R0

R0 contains five methods for estimating the reproduction number. However, four out of those five methods are for estimating the time-independent reproduction number. Only the Wallinga and Teunis method provides an algorithm for estimating the time-dependent reproduction number. Therefore, only this method will be discussed here.

The Wallinga and Teunis method was introduced in a paper by Wallinga and Teunis [13]. At each time step, the method determines which infection networks could have produced the observations and averages over them. The probability p_{ij} that person i who was infected at time t_i was infected by person j who was infected at time t_j can be calculated as follows

$$p_{ij} = \frac{N_i w(t_i - t_j)}{\sum_{k \neq i} N_i w(t_i - t_k)},$$

where N_i is the number of new infections at time t_i and w is the generation time distribution. See Section 2.1 for more details on the generation time distribution. The effective reproduction number of case j is then calculated by summing over i, as these are the probabilities of secondary cases being infected by case j, i.e. the effective reproduction number for j is given by

$$R_j = \sum_i p_{ij}.$$

This allows us to calculate the effective reproduction number at time t, by averaging over all cases that had onset at time t, i.e.

$$R_t = \frac{1}{N_t} \sum_{j:t_j=t} R_j.$$

The advantage of this method is that only the generation time distribution and the number of daily infections is needed. However, a major disadvantage is that R_t requires data beyond time t.

2.2.2 EpiEstim

The requirement of data beyond time t to estimate R_t is not ideal for estimating current effective reproductive numbers. This drawback is mentioned in the paper by A. Cori et al.[3] and was motivation for an alternative method. A. Cori et al. estimate R_t in a manner that only depends on the observations up to time t and an estimated generation time distribution. R_t is estimated by the ratio of the number of new infections at time t to the total infectiousness of all infected cases at time t, i.e.

$$R_t = \frac{N_t}{\sum_{s=1}^{t} N_{t-s} w_s}.$$

The advantage over the Wallinga and Teunis method is that only data up to time t is required to estimate R_t. However, a disadvantage of this method is that it makes a number of assumptions that may result in a biased estimation. These assumptions will be discussed in the next section, regarding EpiNow2.

2.2.3 EpiNow2

In the description of the EpiNow2 documentation, it states that the current best practices based on an article by Gostic et al.[7] are used. In the article by Gostic et al., a few methods are discussed and compared. Furthermore, it discusses the drawbacks and limitations of each method and how to overcome them. Finally, it shows how certain assumptions can lead to a bias in the estimated effective reproduction number. EpiNow2 uses the same method as A. Cori et al., but it implements the solutions given by Gostic et. al. to overcome the drawbacks and limitations. Furthermore, it provides good visualization tools for the resulting outputs. We will now discuss the assumptions that A. Cori et al. make and how EpiNow2 attempts to overcome them. We will not go into elaborate detail on these implementations, so for more details, the reader is referred to the article by S. Abbott et al. [1].

The first assumption that can lead to a bias in the estimation of the effective reproduction number is the delay in the data. Due to delays caused by for example symptom onset, virus detection, or reporting numbers, there will most likely be a delay in the data. Therefore, the data that is observed is most likely not fully representative of the actual data. This is accounted for by sampling a delay in time from a fitted distribution and shifting the observed time by the sampled delay. In case there is data available about the delay in symptom onset and case notification, the distribution from which the delays are sampled is fitted to an exponential and gamma distribution. The best-fitted distribution will then be used to sample the delay from.

The second assumption that must be taken into account, is so-called right-truncation. Right-truncation addresses the absence of recent infections, causing the latest values of the number of new cases to be underestimated. This needs to be accounted for when computing R, because there is most likely an underestimation of the latest number of new cases. Covid-19 has a mean incubation time of 5.2 with a deviation of 2.8 according to [14], hence it can be assumed that there will very likely be an underestimation of the number of new cases in at least the previous 5.2 days and possibly more. This is accounted for by a binomial upscaling of the number of reported cases close to the present time.

The third assumption is the incompleteness of data. There will always be cases that are not observed, which will affect the reproductive number. Furthermore, the observations can also not be assumed to be taken with a constant thoroughness. For example, it is likely that the observation of new infections was less thorough at the beginning of the epidemic. This change over time of things like testing availability or interest in testing causes the reproductive number to be biased. Unfortunately, this is an assumption that is not accounted for in EpiNow2. The reason is that there is no direct data available to estimate the thoroughness of detecting new cases. Another assumption would have to be made to overcome this, which could simply be wrong. Therefore, EpiNow2 makes the assumption that the effectiveness of detecting new cases is constant. Note that the initial bias will reduce over time as the testing efforts are increased at the beginning of the pandemic, but will roughly stagnate at a certain point in time.

EpiNow2 has implemented techniques that address the above assumptions and they build those techniques

around the methods used in EpiEstim. Therefore, we expect EpiNow2 to work best compared to R0 and EpiEstim.

3 Data

In this section, the data that will be used in our simulations is discussed. First, the data for the generation time distribution and the serial interval distribution are discussed. Second, the data for the reproduction number is discussed.

3.1 Generation Time Distribution and Serial Interval Distribution

The data used for the estimation of the generation time distribution and the serial interval distribution differs between the two methods. We will first discuss the data used by Ganyani et al.

3.1.1 Data Ganyani et al.

As mentioned in the methodology section, Ganyani et al use symptom onset data and cluster information from COVID-19 cases in Singapore and Tianjin China. Both of the data sets were recorded in the first two months of 2020. Some of the illustrative cases in the Singapore data set, which contains a total of 91 cases, are shown in Table 1. The Singapore data set used by Ganyani et al. was obtained from the

Cluster	Case	Symptom Onset	Known Contacts	Gender	Age	Notes
-	1	21/01/20	3	M	66	Wuhan national
-	2	21/01/20	13	F	53	Wuhan national
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	8	24/01/20		F	56	Married to 9 / Wuhan national
1	9	24/01/20		M	56	Married to 8 / Wuhan national

Table 1: Cluster data from Singapore.

Ministry of Health of Singapore. The data contains missing information for over half of the cases. For the cases that were not provided with detailed contact information, it was assumed that they could have been infected by any other person in the cluster. The cases that were known to be Wuhan nationals, or known to be in close contact with Wuhan nationals, were assumed to have been infected in Wuhan or by a person from Wuhan. For some cases, like 8 and 9 in Table 1, more detailed information like their partner is available.

The first 5 out of a total of 105 cases in the Tianjin data set are given in Table 2.

CaseID	ClusterID	Source	Source 2	source	symptom onset	relationship	gender	age
TJ34	1	0	3,4,9,22,25,7,11,13,14,17,18,24,28 34,36,6,29,32,5,10,13,21,42,26,38	mall	22/01/2020	sales assistant	0	52
TJ36	1	1	1	TJ34	24/01/2020	husband of TJ34	1	53
TJ37	1	0	0	import (unsure)	25/01/2020	sales assistant	0	43
TJ43	1	0	0	import	21/01/2020	sales assistant	0	35
TJ48	1	0	1,3,4,9,22,25,10,13,21,42,26,38	mall	24/01/2020	customer	0	26

Table 2: Cluster data from Tianjin China

The data from Tianjin, China, contains information that helps decide who the possible infectors could have been for a new infection. Similar to the Singapore data set, this information is used to determine the possible missing links. This is done by creating a vector of possible infectors for each of the cases. When the infector is known, the possible infector vector only contains the true infector. The results of this process, for the first 5 cases, is shown in Table 3. Observe how for case 5 they did not allow for negative serial intervals, which is not in line with reality. In a sensitivity analysis Ganyani et al. study the impact of this decision by leaving out the restriction.

Case	Possible Infectors
1	4
2	1
3	0
4	0
5	1, 3, 4, 42

Table 3: Possible infectors for Tianjin cluster.

3.1.2 Data Nishiuraa et al.

The first 5 cases of the data used by Nishiuraa et al. is given in Table 4. The cases are also accompanied by

CaseID	SI Classification	Diagnosis Country	Infector Onset	Infectee Onset	EL	ER	SL	SR
1	Certain	Vietnam	1/17/2020	1/20/2020	48	47	50	51
2	Probable	Vietnam	1/17/2020	1/19/2020	48	47	49	50
3	Certain	South Korea	1/22/2020	1/26/2020	53	52	56	57
4	Certain	Germany	1/24/2020	1/26/2020	55	54	56	57
5	Certain	Germany	1/24/2020	1/26/2020	55	54	56	57

Table 4: Data used by the Nishiuraa et al. algorithm.

their source, the reliability of these sources is determined by whether or not they are, for example, official government publications or peer-reviewed articles. The interval bounds for symptom onset are already included in the data because they have been extracted from the articles beforehand. An interesting observation is the fact that the data set used by Nishiuraa et al. is not recorded in a single country.

3.2 The Effective Reproduction Number

The data that we will use for estimating the effective reproduction number is the daily number of confirmed infections and the generation time distribution. As the generation time distribution has been discussed in the previous subsection, we will now describe the daily number of infections. We use the data that is provided by John Hopkins University on their Github repository[1][5]. This data consists of csv files containing the cumulative number of infections, deaths, and recovered for a large number of countries. Here, recovered refers the someone who was infected with COVID-19, but has since recovered. Furthermore, we use data consisting of the population size of each country from World Bank Open Data[2]. The population data set need only be downloaded once and this is done manually, but the COVID-19 infections data is updated every day, so we provide code that downloads the data sets from the Github repository before preprocessing. The preprocessing of the data mainly consists of removing inconsistent or empty values and then restructuring it into an easy-to-use object.

After preprocessing, the data will be a list of 169 country objects, where each country object contains its latitude and longitude (for completeness), a time series with the cumulative number of infections, deaths, recovered, and the population size. In this report we will use the data of the Netherlands as an example. However, it must be noted that the methods and simulations discussed in this report can be extended to any of the other 168 countries.

As mentioned in the previous section, the input for R0, EpiEstim, and EpiNow2 consists of, among other things, the number of daily confirmed COVID-19 cases. From our data, it is computed by subtracting the cumulative number of confirmed cases from itself with a lag of 1 day. Figure 1 shows the daily number of confirmed cases, as well as a 7-day moving average in the period from the 27th of February until the 16th of June in the Netherlands.

[1] `https://github.com/CSSEGISandData/COVID-19`

[2] `https://data.worldbank.org/indicator/SP.POP.TOTL?end=2019&start=2019`

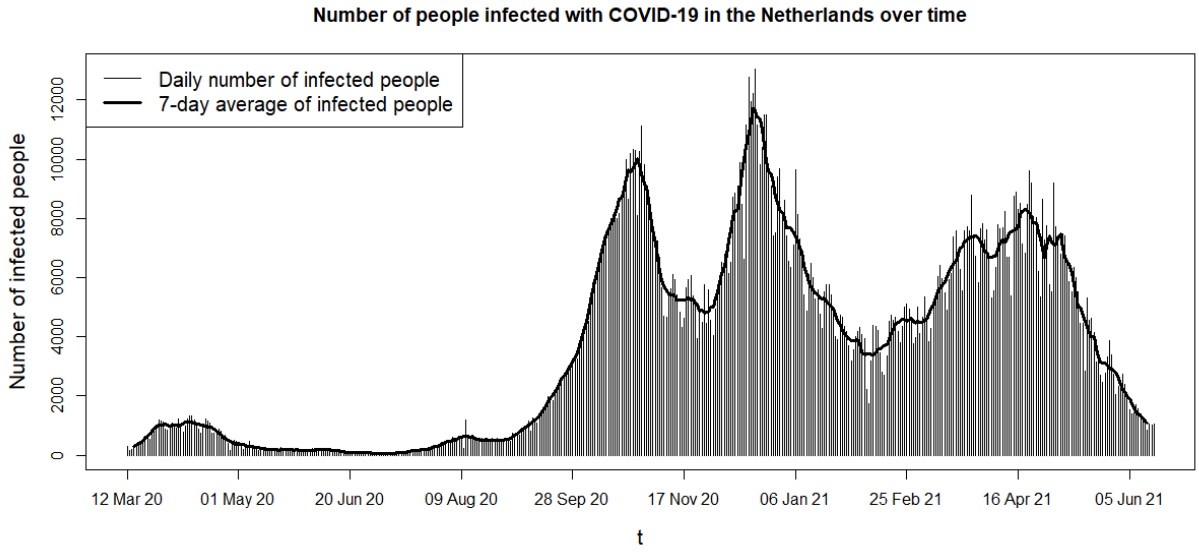


Figure 1: The number of daily COVID-19 infections from the 27th of February until the 3th of June. The black bars show the observed infections and the thick black line shows the 7-day moving average.

Since we use the COVID-19 data for the Netherlands, we can compare our results with the effective reproduction number reported by the RIVM, the Dutch National Institute for Public Health and the Environment. Therefore, the figures showing our results will also show the RIVM-reported effective reproduction number. In Figures 5, 6, and 7, the RIVM-reported effective reproduction number is depicted in red with its corresponding 95% confidence interval in light red. It must be noted that the effective reproduction number has a relatively large uncertainty in the period from April 2020 until July 2020. This is due to the fact that the number of daily infections in that period was very low. In addition, it was still around the beginning of the epidemic, i.e., testing for infections was not as thorough as it is currently. This amounted to lots of uncertainty of the effective reproduction number during that period. Therefore, we expect our estimation of the effective reproduction number to deviate from the RIVM-reported effective reproduction number during that period.

4 Results

The results in this section are divided up into two parts. The first part discusses the results we found regarding the generation time and serial interval distribution. The second part discusses our results for the effective reproduction number. All the supplementary code can be found at this Github repository [4].

4.1 Generation Time Distribution and Serial Interval Distribution

The results for the estimation of the generation time distribution and the serial interval distribution will be discussed in the current subsection. We will first discuss the results of the simulations based on the method by Ganyani et al. and afterward we discuss the results of the simulations based on the method by Nishiuraa et al.

4.1.1 Ganyani et al.

The methods used by Ganyani et al. can be studied using the code provided on their GitHub. To make the code more understandable to ourselves and the reader, we decided to go through it step by step and added annotations where clarification was needed. We also removed and modified parts of the code, as parts were not needed for our current research or produced unclear results. The modified code can be found in the aforementioned repository [4]. To compare our results with those found in the paper, we decided to use the same MCMC settings, i.e., 3000000 runs, a burn-in period of 100000, and a thinning of 200. This means that, for the total trajectory of size 3000000, the first 100000 values are discarded and after that only every 200-th value is stored. Ganyani et al. were pointed to the fact that their list of possible infectors for the Singapore data can lead to impossible cycles. This, in combination with the computational cost of the MCMC simulations (about 10 hours), motivated us to focus our research on the Tianjin data set. The trajectories of the Markov chains are stored during the iterations and can be used to generate a trajectory plot of the simulation. In Figure 2 we plotted the trajectory of the mean and variance of the generation time for the Tianjin data set.

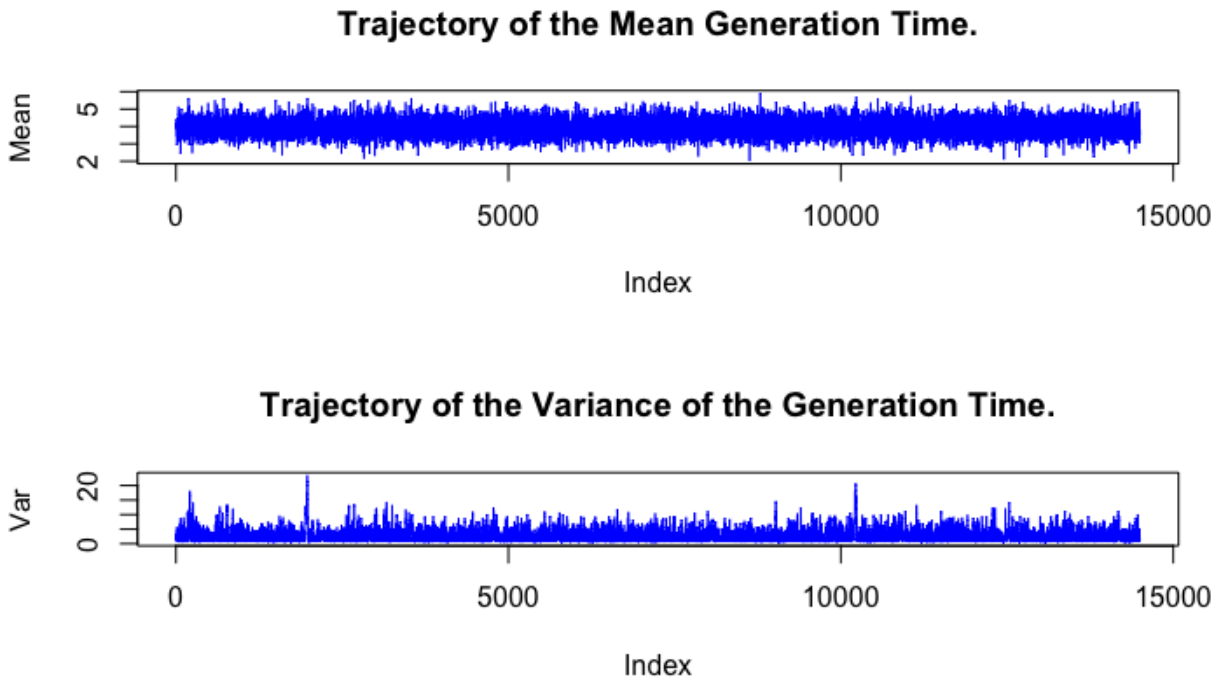
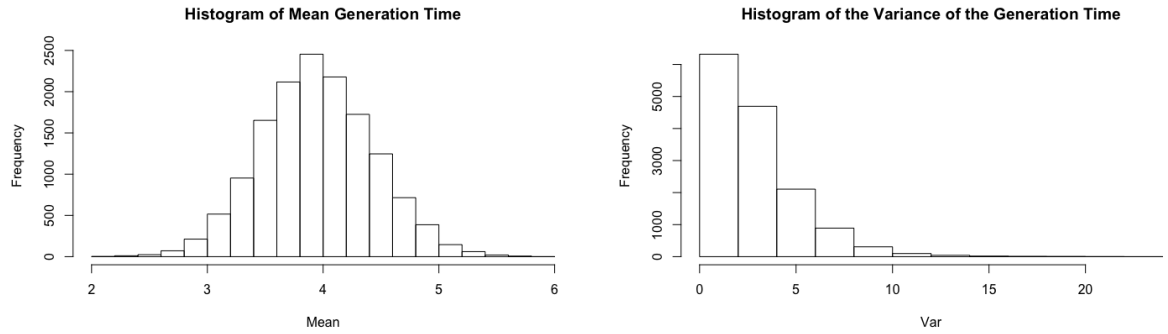


Figure 2: Markov trajectories for the mean and variance of the GI.

The trajectories in Figure 2 show a trend centered (asymmetric in the case of the always positive variance) around certain mean values, to illustrate this better we included histograms of the sampled values, these are given in Figure 3.



(a) Mean Generation Time.

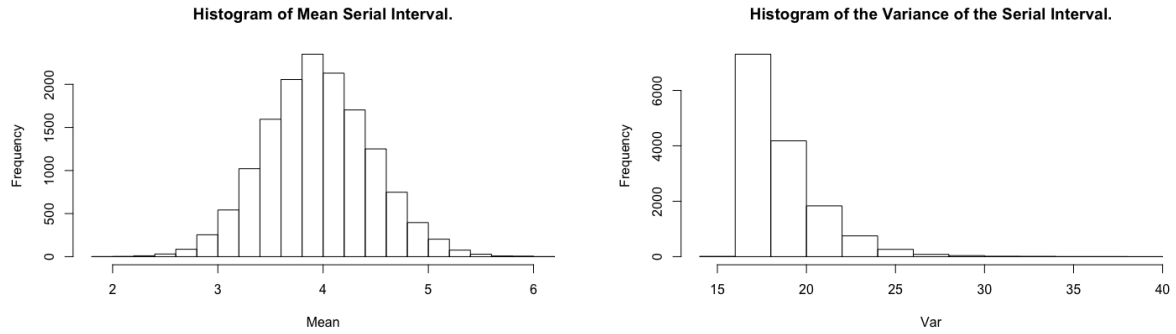
(b) Variance of the Generation Time.

Figure 3: Histograms for the Generation Time.

The code by Ganyani et al. also infers the serial interval, Z_i, from the generation time, X_i, in each iteration using the fact that

$$mean(Z_i) = mean(X_i + Y_i) = mean(X_i) + mean(Y_i) = mean(X_i) + mean(\delta_i - \delta_{v(i)}),$$
$$Var(Z_i) = Var(X_i) + Var(Y_i) + 2Cov(X_i, Y_i) = Var(X_i) + Var(\delta_i) + Var(\delta_{v(i)}),$$

where the variance of the incubation time is 2.8^2 as mentioned in a previous section and the $mean(\delta_i - \delta_{v(i)})$ is determined using 600 samples from the incubation time distribution $\Gamma(\alpha_2, \beta_2)$. Using this information, we can generate histograms for the mean and variance of the serial interval, which can be found in Figure 4.



(a) Mean Serial Interval.

(b) Variance of the Serial Interval.

Figure 4: Histograms for the Serial Interval.

We observe that the distribution of the mean generation time and mean serial interval is very similar, which is expected from the equation in section 2.1.1. However, the variance for the serial interval is significantly larger for the serial interval, which is caused by the variance in the incubation time. An important observation to make is that the mean serial interval tends to be positive, which suggests that there is not a lot of presymptomatic infection. The fact that the mean generation time is smaller than the mean incubation time, however, suggests that this should not be the case. Ganyani et al. included a sensitivity analysis,

which confirmed that allowing for negative serial intervals indeed has a large effect. The mean and quantiles of the posterior sample for the Generation Time and Serial Interval, generated by the MCMC simulation, can be found in Table 5.

	Mean	2.5 % quantile	50 % quantile	97.5 % quantile
Mean GT	3.949002	3.019692	3.939081	4.917231
Standard Deviation GT	1.601301	0.7408664	1.5166738	2.9409230
Mean SI	3.951415	2.988443	3.938707	4.971621
Standard Deviation SI	4.305885	4.028509	4.240318	4.932447

Table 5: Tianjin China

Observe that the difference in magnitude of the variance of the GT and SI is confirmed by the values in Table 5. We also conclude that the mean generation time and mean serial time based on the Tianjin data set are approximately 3.95, which is in line with the results in [6].

4.1.2 Nishiuraa et al.

The code that was provided by Nishiuraa et al. was written in both R and Python and the scripts were given in Jupyter notebooks files (.ipynb). The provided code made use of bash scripts, which can be difficult to work. Therefore, we rewrote the code such that it can be run in only R. The core of the code is the open-source software STAN, which was accessed through the R package *rstan*. Furthermore, the code was cleaned and annotated to allow for clarity and easier adaptability. The number of iterations was set to the default value of 2000. One advantage of *rstan* is that it allows for the use of multiple cores. Lastly, WAIC was used for comparison, which was explained in Section 2.1.2. The results of our simulations are shown in Table 6 and they represent the most likely generation time distributions for the log-normal, gamma, and Weibull distribution with and without truncation. We observe that both with and without truncation the

Distribution	WAIC	Truncation	Mean	SD	2.5 % quantile	50 % quantile	97.5 % quantile
Log-Normal	224.2	Yes	4.62	2.86	3.71	4.62	6.05
Gamma	225.4	Yes	4.91	2.83	3.91	4.86	6.21
Weibull	226.8	Yes	4.89	1.94	3.94	4.84	6.08
Log-Normal	127.9	No	4.57	2.72	3.70	4.53	5.75
Gamma	129.1	No	4.75	2.67	3.83	4.72	5.93
Weibull	130.6	No	4.80	1.95	3.88	4.76	5.95

Table 6: Best fitting GT for each distribution option.

Log-Normal has the lowest WAIC and therefore performs "best". This is an important result, because we saw in an earlier section that Ganyani et al. assumed that the generation time was Gamma distributed. We also observe that the truncation results in larger mean generation times, which can be influential in government decisions. The results that we found are in line with those produced by Nishiuraa et al. [10].

4.2 Effective reproduction Number

In this subsection, the results are shown for estimating the effective reproduction number using R0, EpiEpstim, and EpiNow2. Furthermore, our estimates will be compared to the effective reproduction numbers reported by the RIVM. It should be noted that each package has its own visualization of the results, where EpiEstim and EpiNow2 are the most intuitive and clear. However, for the sake of comparison, the results are plotted as in the same visualization format, which we created.

4.2.1 R0

The effective reproduction number was estimated using R0 and the results are depicted in Figure 5. It shows the estimation of the effective reproduction number (black line) along with its 95% confidence interval

(gray area). Furthermore, it shows the RIVM-reported effective reproduction number (red line) with its 95% confidence interval (light red area) to allow for comparison.

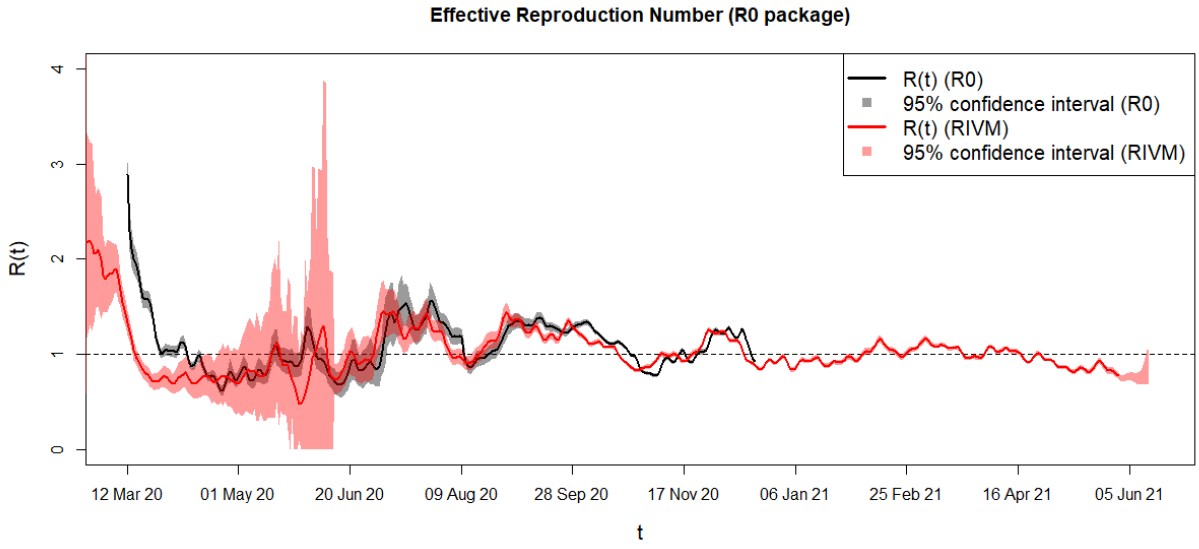


Figure 5: An estimation of the effective reproduction number ($R(t)$) from the beginning of March 2020 until mid-June 2021 using R0. The black line depicts the estimation and they grey area shows its 95% confidence interval. The red line is given by the estimated $R(t)$ according to the RIVM and the light red area shows its 95% confidence interval.

It can be observed that $R(t)$ is not estimated for the whole period. This is due to the disadvantage of the Wallinga and Teunis method, as described in Section 2.2.1, which is that data beyond time t is required to compute $R(t)$. Therefore, the Wallinga and Teunis method is not suitable for making decisions based on current events. R0 is more suitable for an reproduction number analysis after an epidemic has occurred. It offers functions that are more appropriate for computing the initial reproduction number R_0, which is assumed to be intrinsic to the disease and constant over time. It is different from the effective reproduction number. Although R0 contains a method for estimating the effective reproduction number, R0 is more useful for estimating R_0 rather than $R(t)$.

4.2.2 EpiEstim

The effective reproduction number was estimated using EpiEstim and the results are depicted in Figure 6.

In contrast to Figure 5, the estimation of $R(t)$ with EpiEstim was over the whole time range rather than only a part of the time range, as can be seen in Figure 6. First and foremost, it can be observed that the EpiEstim estimation (black line) follows roughly the same trend as the RIVM estimation (red line). In fact, the black line seems to be a few days ahead of the red line. The reason for this is the delay in the data, which is one of the assumptions discussed in Section 2.2.3. The EpiEstim assumes that the observed number of infections is representative for the actual number of infections, but, as we argued in Section 2.2.3, it is more realistic that there is a delay in the data taken into account.

Another observation of Figure 6 is that the estimated $R(t)$ shows a relatively small confidence interval (grey area) compared to the confidence interval in Figure 5. It is almost not visible in Figure 6, whereas in Figure 5 it is much more present. Normally, a small confidence interval is desired as it narrows down the estimation. However, in the case of this pandemic, we believe that EpiEstim shows a confidence interval that is not fully representative of the actual uncertainty of the effective reproduction number.

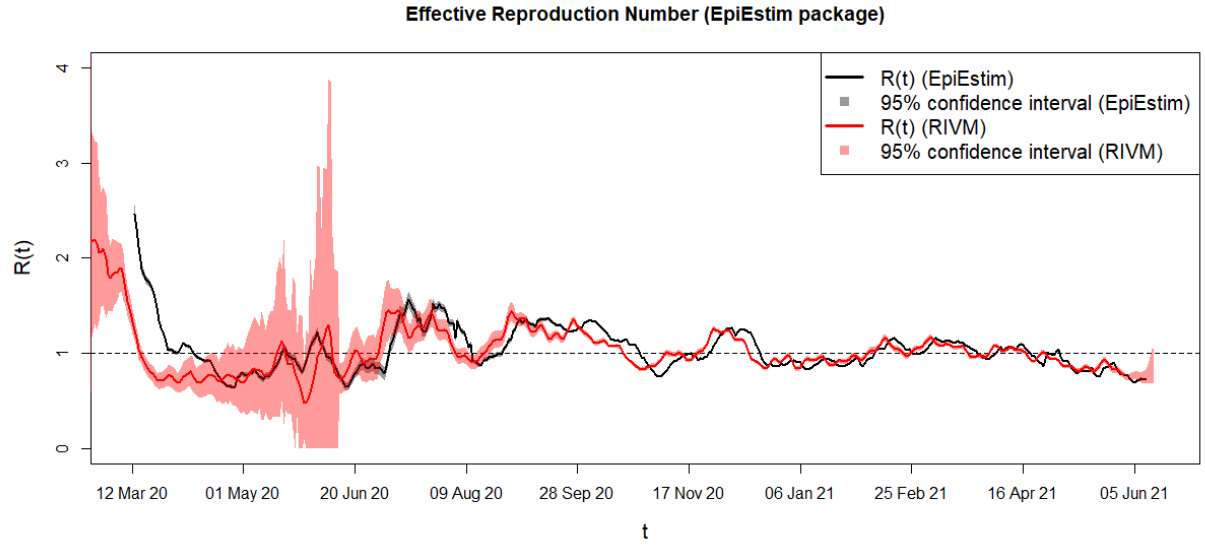


Figure 6: An estimation of the effective reproduction number ($R(t)$) from the beginning of March 2020 until mid-June 2021 using EpiEstim. The black line depicts the estimation and they grey area shows its 95% confidence interval. The red line is given by the estimated $R(t)$ according to the RIVM and the light red area shows its 95% confidence interval.

4.2.3 EpiNow2

The effective reproduction number was estimated using EpiNow2 and the results are depicted in Figure 7 and it shows some remarkable features.

The estimated $R(t)$ (black line) largely follows the trend of the RIVM-reported $R(t)$ (red line). It deviates strongly from March 2020 up to August 2020. However, as mentioned in Section 3.2, there is little to be inferred about the comparison in this time range. After August 2020, the black line seems to follow the red line. The black line is smoother than the red line, which is mainly attributed to the assumptions discussed in Section 2.2.3. Before determining the effective reproduction number, EpiNow2 estimates the true number of infections along with their true times of infection. This estimation can be seen in the visualization tools provided by EpiNow2. The visualization for the estimated number of infections is omitted in this report, but can be seen in the *output* folder of supplementary code [4]. It clearly shows a smoother curve of the number of daily infections.

It must be remarked that EpiNow2 uses the same methods as EpiEstim, except that those methods are complemented with a variety of extensions in EpiNow2. This means that the functions provided by EpiNow2 can be adjusted such that the same results are achieved as EpiEstim. All extensions make it possible to adjust to any use case and the documentation of EpiNow2 describes many use cases. On the other hand, the running time of EpiNow2 with the default settings ($\approx 15min$) was significantly longer than the running time of R0 ($\approx 30s$) and EpiEstim ($\approx 1s$). Nevertheless, this can be reduced significantly by adjusting certain parameters, where, again, the EpiNow2 documentation provides numerous examples on how to adjust the parameters to achieve this. In addition, to shorten the running time of the simulations we note that the simulation for estimating $R(t)$ need only be run once over all the time steps in the data. If the goal is to update the current effective reproduction number daily, then the simulation can be executed with only the last few months. This improves the running time of the simulations significantly.

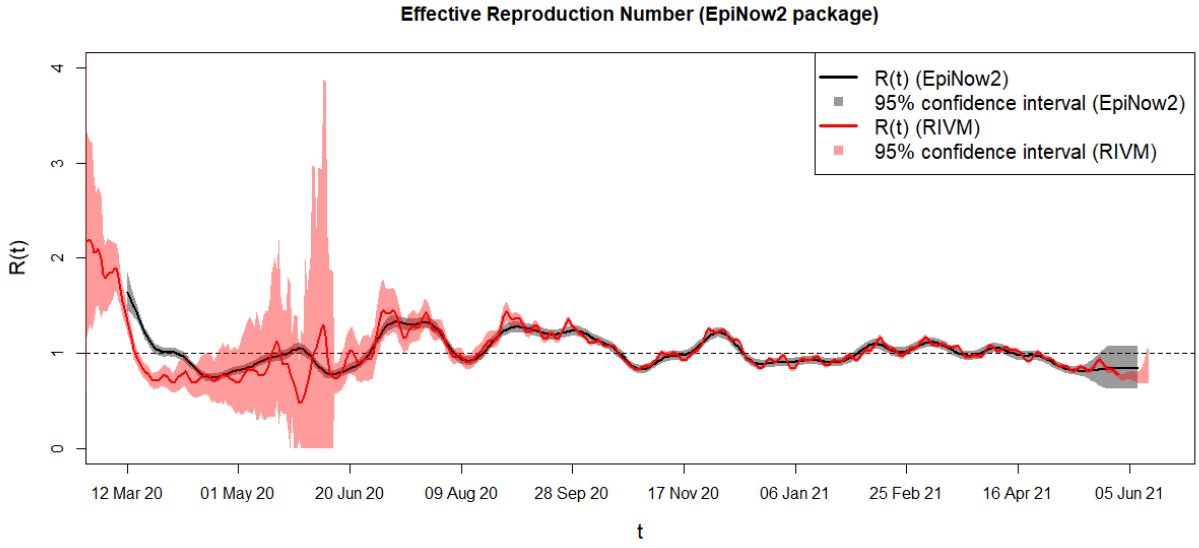


Figure 7: An estimation of the effective reproduction number ($R(t)$) from the beginning of March 2020 until mid-June 2021 using EpiNow2. The black line depicts the estimation and they grey area shows its 95% confidence interval. The red line is given by the estimated $R(t)$ according to the RIVM and the light red area shows its 95% confidence interval.

5 Conclusions

This section comprises the conclusions of the generation time distribution and serial interval distribution, and the effective reproduction number. In both cases, the consequences of the results are summarized and are concluded with our recommendations for reliable models and software.

5.1 Generation Time Distribution and Serial Interval Distribution

During our research, we found that the study of the generation time distribution and the serial interval distribution is often based on relatively small incomplete data sets. A problem with gathering data for research is that contact tracing and determining the exact moment of infection and symptom onset is often nearly impossible. Ganyani et al. refer to the work by Britton et al. [2] who showed that, under the assumption that the incubation period distribution remains constant over time, the generation time and serial interval should have approximately equal means, but different variances. Our simulations produced results that were in line with these claims. Throughout the calculations of the Serial Interval, we fixed the incubation period at a mean value of 5.2 days and a standard deviation of 2.8, this was motivated by the sensitivity analysis by Ganyani et al. [6], which showed that small changes in the incubation time don't have a major effect on the serial interval distribution. An important thing that was pointed out by Ganyani et al. is that allowing for negative serial intervals, i.e. presymptomatic transmission, significantly reduces the mean serial interval [6]. The research by Nishiuraa et al. introduced us to the idea of right-truncation, which identifies the inherent problem of studying a pandemic during the initial growth phase. The results show that this right-truncation consistently leads to higher estimates of the mean generation time. The method by Ganyani et al. found a mean generation time of around 3, whilst the one found by Nishiuraa et al. was closer to 4.7. Using the methods by Nishiuraa et al., we found that on the provided dataset the Log-Normal distribution performed best both with and without truncation. This is an interesting result as it was assumed by Ganyani et al. that the generation time distribution follows a Gamma distribution. The models by Ganyani et al. also assumes a homogeneous mixture of people within clusters, which is not necessarily realistic.

5.1.1 Recommendation

We recommend being careful when choosing generation time and serial interval distributions during the study of the spread of a pandemic. Due to the difficult collection of data, models often require a lot of assumptions. Another big problem with the data sets used in this paper is that the data was gathered in an early stage of the pandemic in specific parts of the world (early 2020). In recent times it has become clear that there are multiple mutations of the COVID-19 virus all over the world, which behave in different ways. To produce more reliable results it is advised that local governments study the spread of the mutations of the virus local to them in a more controlled environment. Studying more recent data would also resolve the problem of right-truncation and allow for easier modelling. Having a better idea of the generation time distribution is an important tool in better understanding the spread of COVID-19.

5.2 Effective reproduction Number

First, as we have seen in Section 4.2, R0 is not suitable for estimating current effective reproduction numbers. It is, however, appropriate for estimating the initial reproduction number and analyzing an epidemic after it has already occurred. In addition, the package was easy to use and the running time was quite fast. The number of simulations in the functions could be adjusted to tilt the balance between computational cost and accuracy.

Second, we found that EpiEstim was the easiest to use, it provided results with the shortest running time, and the results were similar to the results of R0. Moreover, the current effective reproduction number could also be estimated. On the other hand, the disadvantages were also clearly visible when compared to the RIVM-reported effective reproduction number.

Lastly, EpiNow2 provides the same functionalities as EpiEstim and numerous more. With all those extensions, the running time is also significantly increased and the package becomes slightly harder to use. Nevertheless, we found that EpiNow2 is very well documented, which makes up for the large number of parameters. Additionally, the documentation also provides tools to speed up the simulations. The EpiNow2 estimation of the effective reproduction number was found to be the most in line with the RIVM-reported effective reproduction numbers.

5.2.1 Recommendation

For making decisions based on the effective reproduction number, we recommend using EpiNow2. This software package is the most reliable, due to its many adjustable parameters that can be set to fit any data set of daily number of infections.

For a quick intuition of the effective reproduction number, we recommend EpiEstim, because of its easy implementation and quick results. However, it must be noted that the estimated effective reproduction number will most likely be biased.

References

[1] S. Abbott, J. Hellewell, R. Thompson, K. Sherratt, H. Gibbs, N. Bosse, J. Munday, S. Meakin, Emma L. Doughty, J. Y. Chun, Y. Chan, F. Finger, Paul Campbell, A. Endo, C. Pearson, A. Gimma, T. Russell, S. Flasche, A. Kucharski, R. Eggo, and S. Funk. Estimating the time-varying reproduction number of sars-cov-2 using national and subnational case counts. 2020.

[2] T. Britton and G. Scalia Tomba. Estimation in emerging epidemics: biases and remedies. *Journal of the Royal Society Interface*, 16, 2019.

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

[4] J. de Jong and B. Luttikhuizen. Estimation of the reproduction number and generation timedistribution. `https://github.com/jorispdejong/StatisticalConsulting`, 2021.

[5] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet. Infectious Diseases*, 20:533 – 534, 2020.

[6] Tapiwa Ganyani, C. Kremer, D. Chen, Andrea Torneri, C. Faes, J. Wallinga, and N. Hens. Estimating the generation interval for coronavirus disease (covid-19) based on symptom onset data, march 2020. *Eurosurveillance*, 25, 2020.

[7] K. Gostic, L. McGough, E. Baskerville, S. Abbott, K. Joshi, C. Tedijanto, R. Kahn, R. Niehus, J. Hay, P. M. De Salazar, J. Hellewell, S. Meakin, J. Munday, N. Bosse, K. Sherratt, R. M. Thompson, L. White, J. Huisman, J. Scire, S. Bonhoeffer, T. Stadler, J. Wallinga, S. Funk, M. Lipsitch, and S. Cobey. Practical considerations for measuring the effective reproductive number, rt. *medRxiv*, 2020.

[8] S. Jung, A. Akhmetzhanov, Katsuma Hayashi, N. Linton, Y. Yang, Baoyin Yuan, T. Kobayashi, Ryo Kinoshita, and H. Nishiura. Real-time estimation of the risk of death from novel coronavirus (covid-19) infection: Inference using exported cases. *Journal of Clinical Medicine*, 9, 2020.

[9] A. Mood. Introduction to the theory of statistics / by alexander m. mood, franklin a. graybill, duane c. boes. 1963.

[10] H. Nishiura, N. Linton, and A. Akhmetzhanov. Serial interval of novel coronavirus (covid-19) infections. *International Journal of Infectious Diseases*, 93:284 – 286, 2020.

[11] T. Obadia, R. Haneef, and P. Böelle. The r0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, 12:147 – 147, 2012.

[12] N. Reich, J. Lessler, D. Cummings, and R. Brookmeyer. Estimating incubation period distributions with coarse data. *Statistics in medicine*, 28 22:2769–84, 2009.

[13] J. Wallinga and P. Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160:509 – 516, 2004.

[14] J. Zhang, M. Litvinova, W. Wang, Yan Wang, Xiaowei Deng, Xinghui Chen, M. Li, Wen Zheng, L. Yi, Xinhua Chen, Q. Wu, Y. Liang, X. Wang, J. Yang, K. Sun, I. Longini, M. Halloran, P. Wu, B. Cowling, S. Merler, C. Viboud, A. Vespignani, M. Ajelli, and H. Yu. Evolving epidemiology of novel coronavirus diseases 2019 and possible interruption of local transmission outside hubei province in china: a descriptive and modeling study. *medRxiv : the preprint server for health sciences*, 2020.