M2 EEE Empirical Project

Understanding an Epidemic: Theory and Application in the Covid-19 Case

---

Andrew Boomer,

Camille Calandre,

Nikita Marini,

Jacob Pichelmann,

Luca Poll

Supervised by:

Prof. Nour Meddahi

Abstract

In this project we use a combination of two different types of epidemiological models to study the efficacy of a popular public health metric of the novel coronavirus, SARS-COV2 (Covid-19) under different methodological and disease conditions. To this end, we simulate several outbreaks, taking different assumptions for the underlying infectiousness of the disease, and estimating the rate of transmission in these scenarios. We then forecast the progression of future cases given our estimators. We find that using a non-parametric estimator of the rate of transmission provides robustness when the shape of the underlying distribution of infectiousness is unknown. We also note that the assumption of stable conditions makes disease outbreak forecasting in a generalized sense more difficult.

Table of Contents

1 Introduction

The global pandemic brought on by the novel coronavirus, known as SARS-COV2 or Covid-19 has had far reaching public health, social, and economic consequences. It has affected every country on earth to varying degrees, and has resulted in an unprecedented reduction in social interactions and economic activity. In terms of loss of life and infected people, it has been the largest global public health crisis in a century (Khalili et al., 2020). Because of its far ranging social impacts, understanding epidemiological Covid-19 models and their economic ramifications has been of particular interest to economists (see e.g. Toda (2020), Atkenson (2020), and Baker et al. (2020) for early attempts of linking the predictions of an epidemiological model to their economic consequences).

In this project, we survey the infectious disease modeling literature to understand how outbreaks are simulated, estimated, and mitigated. In particular, we focus our attention on the Time Since Infection model developed by Fraser (2007) and expanded by Cori et al. (2013), which appears to be the backbone of one of the most important indexes of the evolution of a pandemic; the reproduction number. In Section 2 and Section 3 of this paper, we will analyze a series of statistical properties relating to this value and its estimates by simulating infections under different scenarios and applying some of the estimating techniques found in the surveyed literature. Section 4 explores the possibility to expand the results previously derived to the (real-world) case of an evolving virus that gives rise to new variants with different levels of contagiousness. Lastly, in Section 5 we apply some of these modeling techniques to estimate rates of transmission and make short-term forecast using French Covid-19 infections data.

2 Data and Modeling Context

One of the main indicators used to monitor the evolution of the COVID-19 pandemic in many countries is the reproduction rate, often denoted as $R(t)$ (Istituto Superiore di Sanità, 2021). This number, which represents the number of people an infected individual at time t can be expected to infect assuming constant conditions (Fraser, 2007), is crucial to identify whether the pandemic is increasing or decreasing[1]. In order to arrive at an analytical formulation of the reproduction rate, however, it is necessary to specify a model of infection dynamics.

[1] corresponding to an $R(t) > 1$ or $R(t) < 1$, respectively.

2.1 The Time Since Infection model

In line with the methodology used in our home countries, which employ an estimation technique for $R(t)$ based on Cori et al. (2013), we begin by specifying a Time Since Infection (TSI) model, first introduced by Fraser (2007). In particular, we assume that the transmission of the virus is a Poisson process, where the probability that an individual infects another one over a very small time interval $[t, t + \Delta]$ is given by $\beta(t, \tau)\Delta$. By aggregating infections at the population level, we obtain the the mean incidence $I(t)$, which is given by:

$$I(t) = \int_0^t \beta(t, \tau) I(t - \tau) d\tau \tag{1}$$

The interpretation of equation 1 goes as follows: the average number of individuals infected at time t is given by the (integrated) sum of the individuals infected between 0 and t, weighted by a function $\beta(t, \tau)$ that depends on the calendar time t and the time since the onset of the infection on a given individual τ. The function $\beta(t, \tau)$ represents the transmissibility of the virus and it reflects how much the pathogen is expected to shed at a given point in time t, given that it has had a certain amount of days τ to reproduce inside the body of the host (time since infection). In the case of COVID-19, it is reasonable to assume a single-peaked function, reflecting a period of high infectivity due to a large reproduction of the virus in the host body, followed by a recovery period in which the infected individual does not have "enough virus" inside of him/her to infect others.[2] The comparison of single and double peaked infectivity profiles, $\beta(t, \tau)$ is shown in Figure 1.

A crucial assumption that will drive the rest of our identification strategy is that the infectivity as a function of the time since infection τ does not depend on calendar time. In other words, the average infectivity of an individual who has been infected for, e.g. 5 days, will be the same irrespective of whether we observe it during the summer or during the winter, during a lockdown or during a period of no generalized social distancing measures. This rather strong assumption allows us to separate the function $\beta(t, \tau)$ into two separate functions, $\phi_1(t)$ and $\phi_2(\tau)$. Fraser (2007) shows how these functions can be normalized such that $\phi_1(t)$ is equal to the instantaneous reproduction number $R(t)$ and that $\phi_2(\tau)$ reflects the distribution of how the infections are spread over the time since infection τ, $\omega(\tau)$. The former depends only on

[2] This shape, however, may not apply to other diseases, such as HIV for instance, which often presents two peaks: one in the early stages of infection and a second one in the period preceding the death of the infected individual.

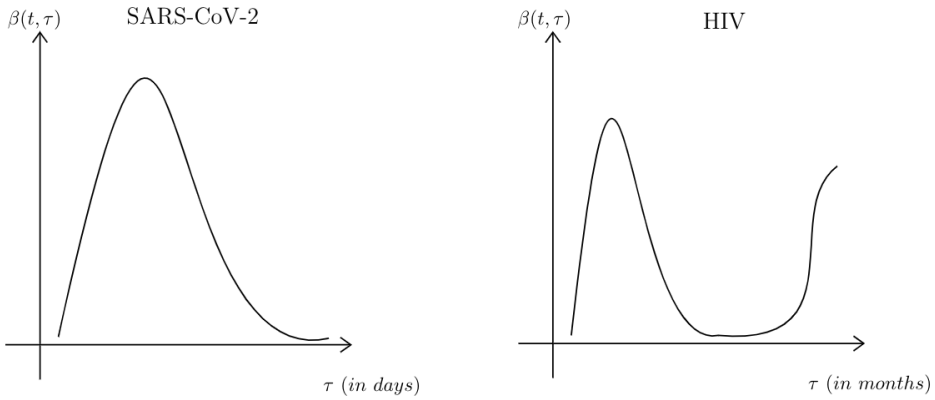


Figure 1: Rough sketches of examples of two different $\beta(t,\tau)$

calendar time while the latter depends on time since infection. These identities allow us to rewrite Equation 1 as

$$I(t) = \int_0^t \underbrace{R(t)\omega(\tau)}_{\beta(t,\tau)} I(t-\tau)d\tau \tag{2}$$

As the first element is not dependent on τ, it can be pulled out of the integral, so that the expression becomes:

$$I(t) = R(t) \int_0^t \omega(\tau) I(t-\tau)d\tau \tag{3}$$

Under this formulation, we can indeed interpret $R(t)$ as the reproduction number at time t and $\omega(\tau)$ as a distribution of how likely these secondary infections are, given the time since infection. Truncating $\omega(\tau)$ at τ_m such that the transmissibility after day τ_m is equal to zero leaves us with the final model:

$$I(t) = R(t) \int_0^{\tau_m} \omega(\tau) I(t-\tau)d\tau \tag{4}$$

where we assume that $\omega(\tau) = 0$ for $\tau > \tau_m$. This implies in other words that an infected individual will not be able to pass the infection after the time since infection τ_m.

2.1.1 Discretizing the TSI Model

Lastly, it is important to transform Equation 3 so that it can be applied to a real-world scenario, where incidence is reported at daily intervals; i.e., with t representing one day, rather than an

infinitely small interval of time. Using Δ as our "very small" interval, we can approximate the integral as a discrete sum as follows:

$$\dot{I}(u) = R(t)\Delta \sum_{n=1}^{N} \omega(n\Delta) I(t - n\Delta) \tag{5}$$

with $t = N\Delta$[3]. For instance, if we assume that infections happen once every 24^{th} of a day (every hour), we obtain:

$$\tilde{I}(t) = \sum_{n=1}^{24} \dot{I}(t - \Delta + n\Delta) \tag{6}$$

where the tilde denotes that the incidences are now obtained as the sum of a discretized process.

2.1.2 Deriving $\widehat{R}(t)$

Rearranging terms, we obtain an expression of $R(t)$ than can be estimated from the data:

$$\widehat{R}(t) = \frac{\tilde{I}(t)}{\sum_{n=1}^{\tau_m} \tilde{I}(t - \tau)\, \omega(\tau)} \tag{7}$$

Equation 7 highlights how the estimation of $\widehat{R}(t)$ is dependent on the knowledge of $\omega(\tau)$ which, as mentioned above, is a distribution specific kernel denoting the probability that a secondary infection will happen after τ days since the infection of an index case. Since this distribution is not known to the researcher, however, the estimator in Equation 7 is not feasible. In order to estimate $\widehat{R}(t)$, the distribution of $\omega(\tau)$ firstly has to be estimated which results in the feasible estimator for $\widehat{R}(t)$ that can be expressed as follows:

$$\widehat{R}(t) = \frac{\tilde{I}(t)}{\sum_{n=1}^{\tau_m} \tilde{I}(t - \tau)\, \widehat{\omega}(\tau)} \tag{8}$$

2.2 Serial interval estimation

Many of the European countries we investigated take $\omega(\tau)$ to be a Gamma distribution, with mean of approximately 6 days and a variance of 2 (as in the Italian case, Cereda et al., 2020). In other words, the probability of the event of a secondary infection will depend on the time since infection τ of the index case, where the average τ will be 6 days and its variance 2 days. These

[3]Note that for $n = 0$ the product $\omega(0)I(t) = 0$ which is why we start the summation at $n = 1$.

estimates are derived from studies that try to describe the distribution of what is known as the serial interval of a disease, and on whose methodology we will focus in the current section.
In order to estimate the function $\omega(\tau)$ the ideal strategy would be to collect data about every infection and to note down after how many days each index case infects a secondary case, holding contact rates constant. This interval of time is known as generation time of a disease. In reality, however, this information is hard to retrieve, as the exact time of the infection is mostly unknown. Instead, researchers gather data on the serial interval, which is defined as the amount of time that elapses between the onset of the symptoms in the index case and the onset of symptoms in the secondary case. Figure 2 gives a graphical representation of the difference between the two concepts. Notice, however, that if incubation times (the time that elapses between the infection and the manifestation of symptoms) are independently and identically distributed -a rather strong assumption, then the two values will be the same.

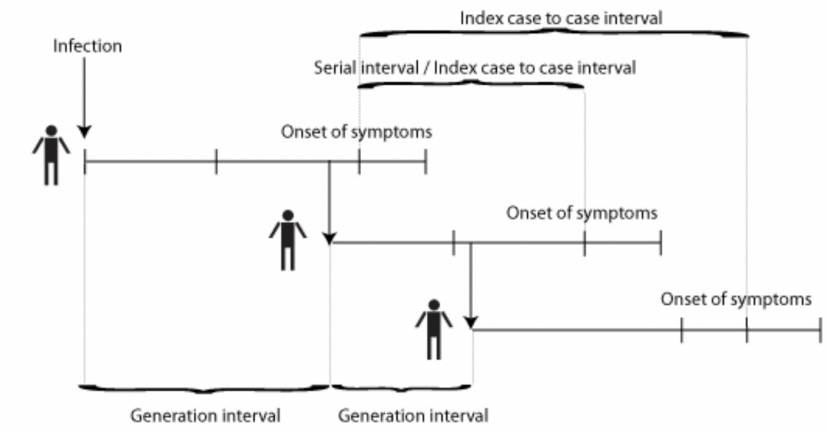


Figure 2: Serial interval and generation time. The vertical arrow indicates an infection event: the amount of time between two consecutive vertical arrows (i.e., from index to secondary case) is defined as the generation time (or generation interval), while the time elapsed between the onset of the symptoms in the index case and in the secondary case is called the serial interval. Source: Vink (2010).

Once observations about each infector-infectee pair are gathered, most studies adopt a parametric approach and estimate a set of parameters via Maximum Likelihood, where the fitted distribution is most often either a Gamma, a Weibull, a Log-Normal (Nishiura, Linton, and Akhmetzhanov, 2020) or in some cases a Normal distribution (which allows to account for negative values in the cases in which the symptoms manifest themselves in the secondary case before they appear in the index case (Ali et al., 2020)). Figure 3 reports the results of a literature review conducted by Rai, Shukla, and Dwivedi (2021), who surveyed the literature regarding the

estimation of serial intervals for the COVID-19 case between January and August 2020, mostly on Chinese data. The resulting average estimated mean of the serial interval is around 5.2 days, with the 95% confidence interval being $[4.37, 6.02]$ days.

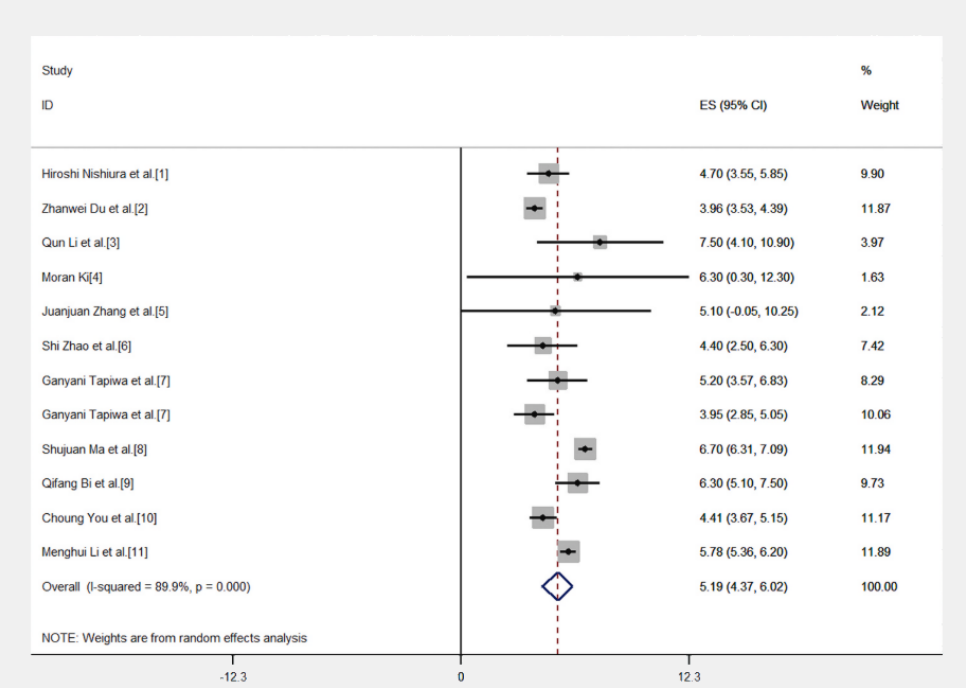


Figure 3: Serial Interval estimation review in Rai, Shukla, and Dwivedi (2021). The first column represents the individual serial interval study while the second column reports the estimated mean with the 95% confidence interval. The third column reports the individual weight that was assigned to the respective study by the authors.

2.3 Simulating a serial interval study

In this section we are concerned with the study of the properties of the estimates of a typical serial interval study. In particular, we will look at how well the mean and variance of $\omega(\tau)$ are estimated in the context of only few observations -in many real-world studies, the average number of infector-infectee pairs observed ranges between 20 to 50 [(Griffin et al., 2020), (Rai, Shukla, and Dwivedi, 2021)]. Additionally, we are interested in the aggregation error; i.e., the bias that is introduced in the estimation by the fact that the realizations of an event -in our case, a secondary infection- are observed over finite intervals of time (most often at a daily level), while the true process happens continuously, over infinitesimally small intervals.

In order to define the reference point of our estimates, we need to simulate infection data, by creating several infector-infectee pairs. To do so, we must specify the process that generates the

infections at an individual level. As the aggregated process of infections is derived under the assumption that they follow a Poisson distribution (*cf.* Equation 1), we make the assumption that this is also true at the micro level. Therefore, the average number of infections generated over the interval of length Δ^* $[t, t+\Delta^*]$ by an individual who has been infected for τ days is given by $\mathcal{P}(\lambda_i)$, where λ_i can be interpreted as the infectivity of individual i and depends on the calendar time and the interval length:

$$\lambda_i(t, \Delta^*) = R(t) \int_{t-\Delta^*}^{t} \omega(\tau) d\tau, \qquad i = 1, ..., N \tag{9}$$

$R(t)$ is assumed to be the same across all individuals and $\omega(\tau)$ is a density function described by a vector of parameters θ, which is our object of interest.

Again, in a real-world scenario only daily infections are reported, and we therefore need to derive a discretized version of Equation 9. This is obtained via a transformation similar to the one employed in Equation 5, which allows to write the individual infection process as $\mathcal{P}(\tilde{\lambda}_i)$, where $\tilde{\lambda}_i$ is given by:

$$\tilde{\lambda}_i(t, \Delta) = R(t)\, \Delta \sum_{k=1}^{1/\Delta} \omega\left(\tau + \Delta - k\Delta\right), \qquad i = 1, ..., N \tag{10}$$

Having obtained the expected number of infections generated throughout one day by an individual, we are ready to estimate the serial interval distribution $\omega(\tau)$ by means of Maximum Likelihood. In particular, we are interested in the vector of parameters $\hat{\theta}$ of a Gamma distribution (namely, its shape and its rate) that maximize the likelihood of observing the infector-infectee pairs in the sample at hand. The choice of a Gamma distribution is in line with the approach followed by the literature, albeit other specifications can also be attempted (among the most used ones despite the Gamma, are the Weibull and the Normal distributions, where the latter is employed in the rare cases where negative values are present in the sample).

2.3.1 Properties of the serial interval estimates

In the context of the studies described above, different sources of bias need to be accounted for, each of which will impact the final estimation of $R(t)$ differently. A first potential bias arises from a fact that has been duly stressed so far, namely that while infections occur in continuous time (i.e., in infinitesimally small intervals), they are only reported on a daily basis. This will

give rise to what we shall call the "aggregation error". If all infections happened in a single moment of the day, once per day, this error would be nonexistent, as the generating process would be equivalent to the one observed. However, this is not case, and we therefore expect that the farther the true process is from its observed, aggregated outcome, the greater the bias of our estimate.

Secondly, we have already mentioned that the sample size in a typical serial interval study tends to be relatively low, with some estimated serial intervals being based on fewer than 10 pairs of infection (e.g. Li et al. (2020) and Huang et al. (2020)). This will of course impinge on the precision of the estimated mean of the distribution of $\omega(\tau)$, which will very likely have large confidence intervals. (Li et al. (2020) for instance report a mean serial interval of 7.5 days, with a 95% CI of [5.3, 19]). Small-sample bias will therefore also have to be evaluated.

Lastly, we are concerned about possible misspecification. Most studies employ a parametric approach to the estimation of $\omega(\tau)$, most often assuming that the underlying distribution is a Gamma. Different specifications of this distribution will of course lead to different estimates. In order to cope with this issue, we also propose to estimate $\omega(\tau)$ non parametrically[4].

Nonetheless, it is important to remember that all the biases listed above are only relevant in the larger framework of the estimation of $R(t)$. For instance, if the true underlying process is described by a Weibull distribution with a vector of parameters θ_1, but we estimate it via a Gamma distribution defined by the vector of parameters θ_2, this "misspecification error" is only relevant if the resulting estimate of $R(t)$ is likewise biased. Therefore, in order to assess the relevance of these potential sources of bias described above, we propose to evaluate them in light of the error that they give rise to in the final estimation of $R(t)$.

In practice, we will simulate an epidemic according to the process described by Equation 5 and compute the mean squared difference between $\widehat{R}(t)$ and its true value $R(t)$ for different estimates of $\omega(\tau)$. In order to assess the properties of the estimator under different scenarios, we consider 5 specifications of the true $R(t)$: a constant $R(t)$ of 1.8, an increasing one that starts at 1.5 and gets up until 3, a decreasing one following the inverse path, an $R(t)$ described by a polynomial of degree 5, depicting two waves of the pandemic, and an inverted U-shaped $R(t)$ with a maximum value of 4 and a minimum of 0.

[4]In particular, we will use a Gaussian kernel with a bandwidth chosen according to normal scale rule: $h = 1.059\hat{\sigma}^{1/5}$, where $\hat{\sigma}$ is the standard deviation of the observation in our simulated sample.

2.3.2 Aggregation Error

The results of the estimation of $\omega(\tau)$ and of its impact on $\widehat{R}(t)$ are summarized in Figure Figure 4, where the main focus lays primarily on the evaluation of the aggregation error. For each of the 5 specifications of the true $R(t)$, 1,000 serial interval studies were simulated for each value of Δ in Equation 10 ranging between 1 and 1/100. The average estimates of the mean and variance of the serial interval (obtained via Maximum Likelihood Estimation by fitting a Gamma distribution on our data[5]) are then plotted against the level of aggregation $1/\Delta$ (upper panels). The bottom panels, on the other hand, report the shape of the true $R(t)$ used in that particular specification and the Mean Squared Error of the $\widehat{R}(t)$ that is obtained using as the denominator in Equation 8 the $\widehat{\omega}(\tau)$ estimated in the preceding step (i.e., described by the ML estimates $\hat{\mu}$ and $\hat{\sigma}^2$ for each value of $1/\Delta$):

$$MSE_{R(t)} = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{R}(t) - R(t)\right)^2 \tag{11}$$

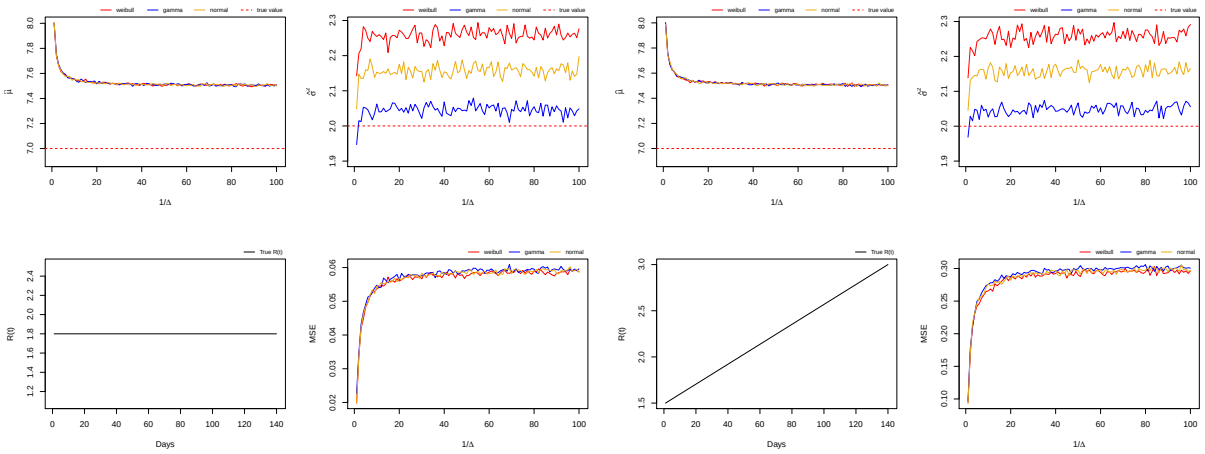
with T the number of days in our simulated pandemic.

The results point out several interesting features: firstly, the top left panel of each specification shows that at low values of Δ (high values of $1/\Delta$) the estimated mean of the serial interval is highly biased, with the bias decreasing as the level of aggregation increases. We explain this result as follows: for low values of Δ the simulated infections in the serial interval study are the results of few draws from a Poisson distribution. Therefore, the number of infections that are generated at each point in time (e.g., at each day for $1/\Delta = 1$) is subject to a great degree of randomness. Conversely, when the infections are simulated over smaller intervals (i.e., when Δ increases), the aggregation that is done at the daily level becomes an increasingly better estimate of the expected number of infections on that given day. Nonetheless, a positive bias in the estimated mean persists when $\Delta \to \infty$, which is in the order of approximately 7% of the true value of μ[6]. This result is robust across different functional forms of the true $\omega(\tau)$ used in the generation of the infections, represented by a each different color in the graph.

The estimated variance $\hat{\sigma}^2$, on the other hand, shows an opposite behavior: while the aggregation error does not seem to cause a significant bias in its estimation, the misspecification one does; in particular, the ML estimator obtained by fitting a Gamma distribution through our daily

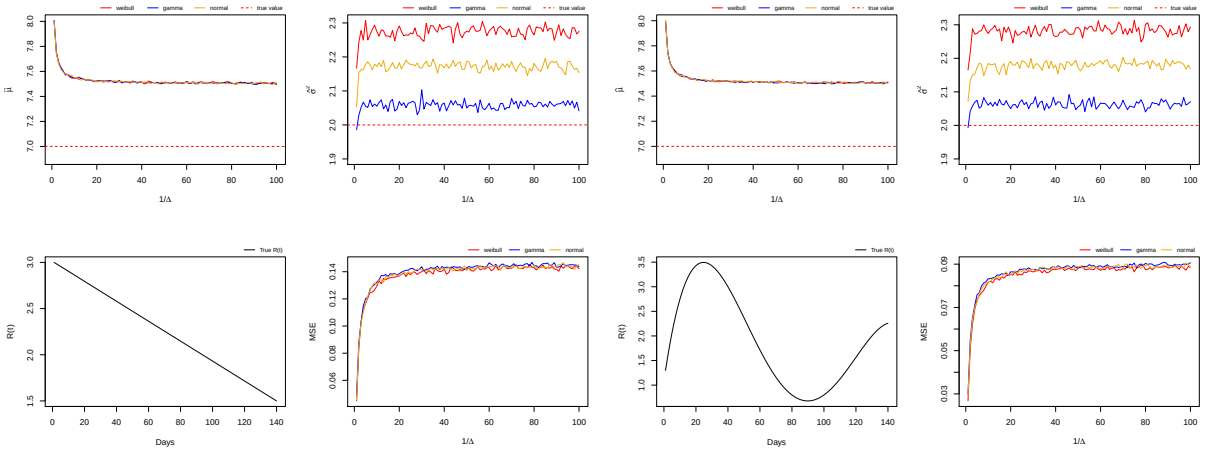
[5] notice that a Gamma distribution $\Gamma(\alpha, \beta)$ is defined by a shape and a rate parameter α and β, respectively, which we transform to mean and variance according to the following: $\mu = \frac{\alpha}{\beta}$ and $\sigma^2 = \frac{\alpha}{\beta^2}$

[6] as a robustness check, we specified a serial interval of 10 days, which resulted in a bias of 5.5% of the true value.



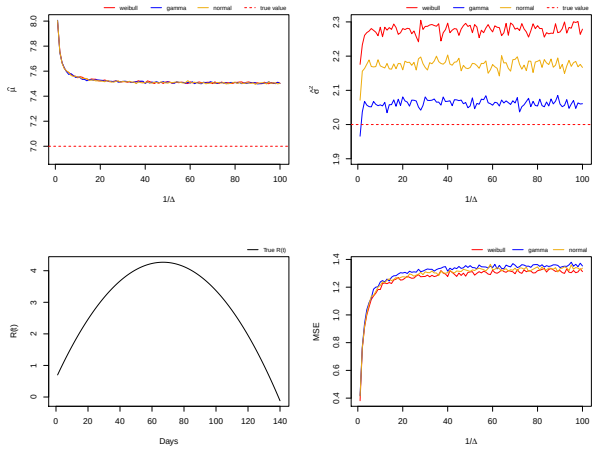
(a) Constant $R(t)$

(b) Increasing $R(t)$



(c) Decreasing $R(t)$

(d) Multiple-waves $R(t)$



(e) Inverted-U $R(t)$

Figure 4: Results of the serial interval study simulation: each group of four panel represent a different study, based on a given specification of $R(t)$. The top-left panel in each group displays the estimated mean of the serial interval for a given interval (expressed as a fraction of a day) used to simulate the infections; the top-right panel reports the estimated variance of the serial interval; the bottom-left panel shows the form chosen for the $R(t)$ that generates the pandemic over 140 days and on which $\widehat{R}(t)$ is computed; the Mean Squared Error of the estimate of $R(t)$ is displayed in the bottom-right panel of each graph.

observations performs best when (unsurprisingly) the true process is also generated by a Gamma distribution -the blue line in the graphs, while when the serial interval is generated using a Normal or a Weibull distribution -the golden and red lines, respectively, the bias increases (with the Gamma distribution being a better estimate for the Normal case than the Weibull one).

Lastly, and most importantly, we focus on what these results imply in terms of the estimation of $R(t)$. This is shown in the bottom right panel of each graph, alongside a visualization of the shape of the $R(t)$ throughout the whole pandemic. Right away, we observe that the shape of the MSE function is the same across the 5 specifications of the true $R(t)$, with its lowest value corresponding to low values of Δ. In order to make sense of this, we must look back at Equation 5, which describes how the infections are generated in discrete time, and to Equation 6, which offers an example about how the daily aggregation is constructed, so to match the way infections are reported in real world. Indeed, when $\Delta = 1$, Equation 6 simplifies to $\dot{I}(t)$, meaning that the observed new infections correspond to the ones generated by the data. Conversely, as Δ becomes smaller, this correspondence is lost and $\widehat{R}(t)$ will consequently suffer some bias, as we indeed observe in the bottom right panels of each graph. Focusing on the magnitude of the bias, we see that it ranges between a minimum of 0.06 -for an unrealistic constant $R(t)$, and 1.4 for the inverted-U scenario. The most realistic case of the "multiple waves" $R(t)$, depicted in Figure 4d, however, reports a bias of less than 0.1, even when our simulated infections happen at a frequency of approximately 15 minutes (i.e., $1/\Delta = 96$) and irrespective of the underlying distribution of $\omega(\tau)$.

We therefore conclude that the strategy employed in serial interval estimation is naturally limited in the identification of the true $R(t)$, This is due to the reporting of cases at a daily basis, which leads to the presented aggregation error. Given the different settings of $R(t)$ under which the pandemic evolves, however, this bias might be negligibly small.

2.3.3 Misspecification and Small Sample Bias

Departing from the aggregation error to the remaining two potential sources of a bias, the misspecification error and the small sample, we once again conducted a Monte Carlo study. Hereby, a pandemic was simulated according to Equation 4. This was repeated 1000 times respectively for the five different settings of $R(t)$, three different true underlying distributions for the serial interval (Gamma, Normal, Weibull) and five different serial interval study sample sizes

(10, 20, 50, 100, 500). We firstly simulated the outbreak, on which we then conducted two serial interval studies. One serial interval was estimated by fitting a gamma distribution as previously stated, while in the second serial interval study we employed a nonparametric approach. Lastly, we used the obtained serial interval estimates in order to estimate $\widehat{R}(t)$ through the feasible estimator specified in Equation 8. Similar to Subsubsection 2.3.2, we were not only interested in how well the shape of the true serial interval was estimated, but also in the resulting effect onto the estimates of $\widehat{R}(t)$. In order to evaluate the results, the respective mean of $\hat{\mu}_\omega$, $\hat{\sigma}^2_\omega$ and $MSE_{R(t)}$ along with their standard deviations were recorded and are reported in Table 1.

In the results, we can easily observe aggregation error: the three different serial interval distributions were specified with $\mu_\omega = 7$ and both approaches, the Gamma as well as the Nonparametric (NP) approach consistently estimate $\hat{\mu}_\omega$ at ≈ 7.5 across the three different true serial interval distributions, the different specifications for $R(t)$ as well as the varying sample sizes. While the estimated $\hat{\mu}_\omega$ is very similar for the two approaches, the Gamma approach tends to estimate $\hat{\sigma}^2_\omega$ on average better across the specifications. The aggregation error is consistent across the different sample sizes, where we see besides a decreasing standard deviation overall no improvement in the identification of the serial interval distribution parameters. Considering the average $MSE_{R(t)}$, we observe across the different specifications that in most cases the Mean Squared Error drops initially comparably much when we go from 10 to 20 observations, but then decreases at a slowing rate. This could be interpreted as evidence that the small sample bias reduces quickly as we increase the number of infector - infectee pairs in the serial interval study. Given the costly and challenging nature of such study in a realistic environment, a sample of 20 to 50 observations might hence already prove as a sufficient trade off for obtaining consistent estimates of $R(t)$. If we proceed to compare the average $MSE_{R(t)}$ not across the respective sample sizes but across the different specifications and the true serial interval distributions, it is evident that the nonparametric approach has a on average considerably smaller Mean Squared Error than the gamma approach. This confirms not directly the presumption of the misspecification error since the gamma approach seems to perform relatively consistent across the specifications, but nevertheless emphasizes the on average better performance of the nonparametric approach compared to the parametric approach. Lastly, as in the previous scenarios, we see that the average Mean Squared Error differs considerably across the different specifications of $R(t)$. In the following sections we will exploit these differences further and extend the scope of our analysis.

		True dist:	Gamma						Normal						Weibull					
		Evaluation:	$\hat{\mu}_\omega$		$\hat{\sigma}^2_\omega$		$MSE_{R(t)}$		$\hat{\mu}_\omega$		$\hat{\sigma}^2_\omega$		$MSE_{R(t)}$		$\hat{\mu}_\omega$		$\hat{\sigma}^2_\omega$		$MSE_{R(t)}$	
	R type	Sample	Gamma	NP	Gamma	NP	Gamma	NP	Gamma	NP	Gamma	NP	Gamma	NP	Gamma	NP	Gamma	NP	Gamma	NP
Mean	Constant	10	7.530528	7.530530	1.98	2.73	0.063	0.008	7.527872	7.527867	2.08	2.74	0.062	0.008	7.533368	7.533382	2.15	2.70	0.062	0.008
(SD)	Constant	-	0.361192	0.361206	0.72	1.00	0.031	0.010	0.338012	0.338020	0.79	0.97	0.029	0.009	0.335597	0.335601	0.86	0.94	0.030	0.009
Mean	Constant	20	7.502261	7.502266	2.01	2.56	0.058	0.006	7.518890	7.518895	2.14	2.60	0.059	0.007	7.509000	7.508994	2.26	2.60	0.058	0.006
(SD)	Constant	-	0.242539	0.242533	0.50	0.65	0.020	0.006	0.242486	0.242493	0.58	0.65	0.021	0.006	0.243930	0.243938	0.63	0.62	0.021	0.006
Mean	Constant	50	7.525607	7.525621	2.06	2.43	0.059	0.006	7.529128	7.529121	2.18	2.44	0.059	0.006	7.513229	7.513225	2.29	2.45	0.057	0.006
(SD)	Constant	-	0.155940	0.155946	0.32	0.39	0.013	0.004	0.155007	0.154997	0.36	0.38	0.013	0.004	0.150059	0.150066	0.38	0.35	0.013	0.004
Mean	Constant	100	7.518910	7.518907	2.06	2.33	0.058	0.005	7.519391	7.519383	2.19	2.34	0.057	0.006	7.517806	7.517806	2.31	2.36	0.057	0.006
(SD)	Constant	-	0.110185	0.110182	0.23	0.27	0.009	0.003	0.106294	0.106302	0.25	0.26	0.009	0.002	0.112679	0.112685	0.29	0.26	0.009	0.003
Mean	Constant	500	7.521478	7.521461	2.08	2.15	0.058	0.005	7.520649	7.520647	2.20	2.15	0.057	0.005	7.522504	7.522497	2.30	2.15	0.057	0.006
(SD)	Constant	-	0.047470	0.047486	0.10	0.11	0.004	0.001	0.048056	0.048059	0.11	0.10	0.004	0.001	0.049741	0.049739	0.13	0.11	0.004	0.001
Mean	Increasing	10	7.537182	7.537190	1.97	2.70	0.323	0.038	7.530039	7.530039	2.01	2.64	0.322	0.040	7.514963	7.514977	2.14	2.69	0.313	0.039
(SD)	Increasing	-	0.345965	0.345971	0.73	1.04	0.158	0.046	0.351629	0.351629	0.76	0.90	0.166	0.049	0.357246	0.357214	0.86	0.94	0.164	0.047
Mean	Increasing	20	7.535426	7.535431	2.03	2.58	0.309	0.031	7.530139	7.530140	2.12	2.56	0.305	0.031	7.516108	7.516101	2.23	2.58	0.297	0.030
(SD)	Increasing	-	0.241316	0.241321	0.51	0.68	0.107	0.029	0.241046	0.241038	0.56	0.62	0.109	0.029	0.238677	0.238676	0.61	0.62	0.109	0.029
Mean	Increasing	50	7.523183	7.523183	2.08	2.46	0.296	0.027	7.521716	7.521718	2.16	2.42	0.294	0.027	7.522059	7.522063	2.29	2.45	0.290	0.027
(SD)	Increasing	-	0.154315	0.154313	0.32	0.39	0.067	0.017	0.156079	0.156081	0.34	0.35	0.069	0.018	0.150450	0.150449	0.42	0.38	0.069	0.018
Mean	Increasing	100	7.522774	7.522777	2.06	2.33	0.294	0.026	7.528007	7.527998	2.21	2.36	0.293	0.026	7.520409	7.520407	2.30	2.35	0.287	0.026
(SD)	Increasing	-	0.108679	0.108683	0.22	0.27	0.046	0.012	0.109668	0.109674	0.26	0.26	0.049	0.013	0.106146	0.106143	0.28	0.26	0.048	0.012
Mean	Increasing	500	7.518666	7.518667	2.08	2.15	0.290	0.026	7.519748	7.519749	2.20	2.16	0.287	0.026	7.520410	7.520395	2.30	2.15	0.285	0.026
(SD)	Increasing	-	0.050378	0.050366	0.10	0.11	0.021	0.006	0.046680	0.046682	0.12	0.11	0.020	0.005	0.048176	0.048174	0.13	0.11	0.022	0.006
Mean	Decreasing	10	7.516377	7.516375	2.02	2.65	0.147	0.016	7.525596	7.525599	2.12	2.65	0.149	0.017	7.516641	7.516644	2.18	2.60	0.147	0.017
(SD)	Decreasing	-	0.282519	0.282512	0.56	0.76	0.061	0.017	0.283809	0.283811	0.69	0.79	0.063	0.018	0.284191	0.284189	0.73	0.76	0.064	0.018
Mean	Decreasing	20	7.516985	7.516986	2.03	2.50	0.144	0.014	7.520245	7.520244	2.16	2.52	0.143	0.014	7.528175	7.528186	2.25	2.51	0.144	0.015
(SD)	Decreasing	-	0.193632	0.193640	0.40	0.51	0.040	0.010	0.196857	0.196853	0.46	0.50	0.042	0.011	0.201582	0.201580	0.53	0.50	0.045	0.012
Mean	Decreasing	50	7.522167	7.522168	2.09	2.40	0.142	0.013	7.517061	7.517060	2.17	2.37	0.140	0.013	7.522478	7.522480	2.29	2.39	0.140	0.013
(SD)	Decreasing	-	0.128649	0.128649	0.25	0.31	0.026	0.007	0.121981	0.121978	0.30	0.30	0.025	0.007	0.128694	0.128683	0.34	0.31	0.028	0.007
Mean	Decreasing	100	7.525352	7.525346	2.07	2.28	0.142	0.013	7.518204	7.518212	2.19	2.29	0.140	0.013	7.520611	7.520618	2.29	2.28	0.139	0.013
(SD)	Decreasing	-	0.089412	0.089409	0.18	0.22	0.018	0.005	0.089155	0.089151	0.22	0.22	0.019	0.005	0.089938	0.089940	0.23	0.21	0.019	0.005
Mean	Decreasing	500	7.520271	7.520269	2.08	2.12	0.141	0.013	7.520390	7.520395	2.20	2.12	0.139	0.013	7.522594	7.522601	2.31	2.12	0.139	0.013
(SD)	Decreasing	-	0.039314	0.039318	0.08	0.09	0.008	0.002	0.039354	0.039361	0.09	0.09	0.008	0.002	0.040082	0.040093	0.11	0.09	0.009	0.002
Mean	Polynomial	10	7.536923	7.536919	2.06	2.63	0.093	0.010	7.507964	7.507963	2.13	2.60	0.090	0.009	7.527502	7.527505	2.24	2.62	0.091	0.010
(SD)	Polynomial	-	0.255325	0.255315	0.54	0.72	0.035	0.009	0.264414	0.264411	0.59	0.66	0.036	0.009	0.254861	0.254865	0.66	0.67	0.037	0.010
Mean	Polynomial	20	7.524714	7.524708	2.04	2.47	0.090	0.008	7.522947	7.522950	2.15	2.47	0.089	0.008	7.518435	7.518433	2.27	2.49	0.088	0.008
(SD)	Polynomial	-	0.180439	0.180429	0.37	0.47	0.024	0.006	0.176911	0.176913	0.42	0.44	0.024	0.006	0.178367	0.178376	0.46	0.44	0.025	0.006
Mean	Polynomial	50	7.518350	7.518350	2.07	2.35	0.088	0.008	7.521516	7.521515	2.19	2.35	0.087	0.008	7.523152	7.523160	2.31	2.37	0.087	0.008
(SD)	Polynomial	-	0.111997	0.111995	0.24	0.29	0.015	0.004	0.108983	0.108979	0.26	0.26	0.015	0.004	0.111560	0.111570	0.30	0.27	0.016	0.004
Mean	Polynomial	100	7.521479	7.521485	2.08	2.27	0.088	0.008	7.523630	7.523630	2.20	2.26	0.087	0.008	7.523848	7.523837	2.30	2.26	0.086	0.008
(SD)	Polynomial	-	0.079666	0.079666	0.17	0.20	0.010	0.003	0.078820	0.078826	0.18	0.18	0.011	0.003	0.080932	0.080931	0.20	0.18	0.011	0.003
Mean	Polynomial	500	7.522251	7.522250	2.09	2.12	0.087	0.008	7.519763	7.519771	2.19	2.11	0.086	0.008	7.521230	7.521216	2.31	2.11	0.086	0.008
(SD)	Polynomial	-	0.035953	0.035954	0.07	0.08	0.005	0.001	0.035172	0.035171	0.08	0.07	0.005	0.001	0.036668	0.036681	0.10	0.08	0.005	0.001
Mean	U-inverted	10	7.507583	7.507588	1.97	2.54	1.366	0.117	7.528817	7.528802	2.13	2.61	1.397	0.133	7.532644	7.532648	2.21	2.59	1.395	0.140
(SD)	U-inverted	-	0.258917	0.258910	0.54	0.71	0.536	0.123	0.271078	0.271068	0.60	0.68	0.589	0.141	0.266924	0.266927	0.69	0.70	0.612	0.150
Mean	U-inverted	20	7.516705	7.516707	2.04	2.48	1.337	0.105	7.523812	7.523814	2.17	2.50	1.335	0.112	7.531701	7.531703	2.27	2.49	1.338	0.119
(SD)	U-inverted	-	0.178175	0.178180	0.40	0.50	0.360	0.081	0.184435	0.184432	0.44	0.46	0.392	0.090	0.187108	0.187096	0.48	0.45	0.423	0.099
Mean	U-inverted	50	7.517843	7.517841	2.07	2.36	1.317	0.101	7.516782	7.516788	2.21	2.39	1.291	0.101	7.514866	7.514860	2.28	2.36	1.274	0.103
(SD)	U-inverted	-	0.115051	0.115047	0.24	0.29	0.227	0.050	0.114954	0.114966	0.28	0.28	0.242	0.055	0.114709	0.114704	0.30	0.28	0.250	0.057
Mean	U-inverted	100	7.517600	7.517610	2.07	2.26	1.309	0.102	7.517269	7.517269	2.20	2.28	1.287	0.102	7.517938	7.517937	2.31	2.28	1.265	0.103
(SD)	U-inverted	-	0.082225	0.082220	0.17	0.20	0.165	0.038	0.082054	0.082052	0.20	0.20	0.171	0.039	0.077589	0.077602	0.22	0.19	0.170	0.039
Mean	U-inverted	500	7.520172	7.520185	2.08	2.11	1.308	0.107	7.522110	7.522117	2.20	2.12	1.291	0.109	7.517782	7.517791	2.31	2.12	1.259	0.107
(SD)	U-inverted	-	0.036730	0.036722	0.08	0.08	0.073	0.017	0.036103	0.036098	0.08	0.08	0.076	0.018	0.035719	0.035707	0.10	0.08	0.078	0.018

Table 1: Monte Carlo Simulation of Serial Interval Study: The table reports the mean and the standard deviation of the evaluation parameters $\hat{\mu}_\omega$, $\hat{\sigma}^2_\omega$ and $MSE_{R(t)}$ for differing specifications of the pandemic and the serial interval study. Namely the true underlying serial interval distribution, different assumptions on the shape of $R(t)$ and the sample size of the serial interval study. The table furthermore compares estimates obtained via the presented parametric "Gamma" approach to a nonparametric (NP) approach.

3 Modeling an Outbreak

Having understood the properties of the estimator of the reproduction number, $\widehat{R}(t)$, we can now move on to the simulation of a whole outbreak, in order to assess the behavior of the estimator at each point throughout the epidemic. In order to accurately simulate the progression of an outbreak, we need to be able to account for the changing susceptibility of the population. The TSI model described in Section 2 takes the assumption of an infinite population, which, as noted in Fraser (2007), may be valid in the beginning of an outbreak. To account for the changing susceptibility of the population in later stages of the pandemic, however, we can turn to a different group of epidemiological models known as compartmental models, the most well known of which is the SIR model.

3.1 Compartment Models: SIR

The basic discrete time SIR model is a compartmental model, categorizing the total population into three subgroups: (1) S = Susceptible, (2) I = Infected, and (3) R = Recovered. In this framework, Chudik, Pesaran, and Rebucci (2020) note that this is simply an accounting exercise, and when we allow the proportion of the population in each subgroup to vary over time, we get the identity $N = S_t + I_t + R_t$. Progressing further, these groups are related to each other with medically informed parameters specific to each disease. In the basic framework, these two parameters are β and γ, where β represents the rate of transmission, and γ the recovery time (Chudik, Pesaran, and Rebucci, 2020). The SIR model with these parameters, is therefore a set of discrete equations outlining the change in the number of people in each subgroup.

$$S_{t+1} - S_t = -\beta \frac{S_t}{N} I_t \tag{12}$$

$$I_{t+1} - I_t = (\beta \frac{S_t}{N} - \gamma) I_t \tag{13}$$

$$R_{t+1} - R_t = \gamma I_t \tag{14}$$

In this basic framework, it can be seen that given a set of initial conditions (N, β, γ) this SIR model is deterministic. Further parameters can be added to account for certain mitigation efforts (e.g., to model social distancing as in Chudik, Pesaran, and Rebucci (2020)). Additionally, the β parameter could be allowed to vary over time or location to account for regional or time-

varying differences in the rate of transmission. The basic SIR model can easily be extended by partitioning the population into more narrow subgroups. Common examples of these extra groups are $E =$ Exposed, and $D =$ Dead. New parameters would have to be added to account for transitions to/from these groups into/out of the other subgroups in the compartmental model.

3.2 Time Since Infection vs. Compartment models

As noted before, the assumption of an infinite susceptible population is restrictive and cannot be used to accurately simulate the progression of new infections. For this reason, to model the outbreak, we alter the TSI equation to include an adjustment for the changing susceptibility of the population as new individuals become infected. This adjustment implicitly morphs the TSI into a quasi-compartmental model as we then track both the number of infected and the number of susceptible over time, resulting in a kind of SI-TSI model.

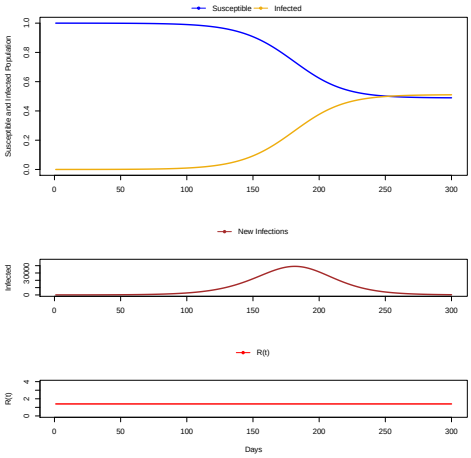
In order to simulate an outbreak, we therefore assume infections to evolve with the dynamic described by the TSI model that was presented in Section 2, where we additionally account for compartments, giving rise to:

$$\dot{I}_{SI}(u) = R(t)\Delta \sum_{n=1}^{N} \omega(n\Delta) I(t - n\Delta) \times \frac{S(n\Delta)}{N} \tag{15}$$

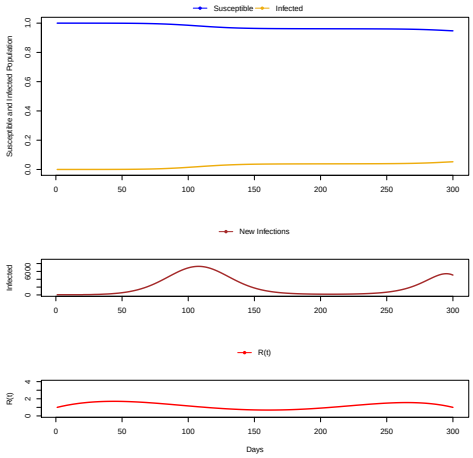
$$\tilde{I}_{SI}(t) = \sum_{i=1}^{24} \dot{I}_{SI}(t + \Delta - i\Delta) \tag{16}$$

The state of the pandemic described by $R(t)$ is estimated using the same estimator as in Equation 8. The compartmental TSI model described in Equation 16 is an SI model, and is used to generate a simulated series of cases. We use this simulated set of cases to test our estimator of $R(t)$ given knowledge of the DGP. We take the constant $R(t)$ and the multiple-waves $R(t)$ scenarios described earlier to understand the performance of our estimator on the data simulated with this SI model.

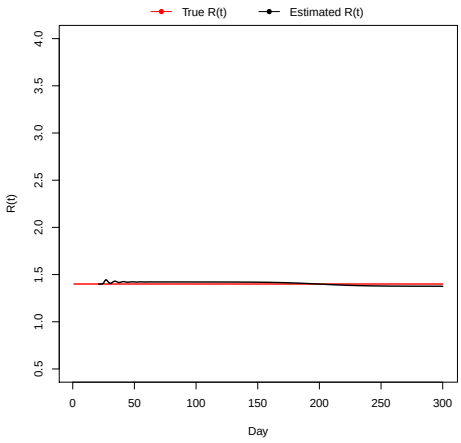
Figure 5 shows the results of a simulated outbreak where the serial interval is described by a Gamma distribution whose parameters are unknown to the econometrician and, like in the TSI case, are retrieved via a serial interval study. We can see that the $R(t)$ estimator performs slightly better in the constant $R(t)$ case in Figure 5c compared to the multiple-waves $R(t)$ case depicted in Figure 5d, albeit both estimators seems to perform rather well -these results are in line with the ones displayed in Figure 4. In particular, the estimator performs the worst



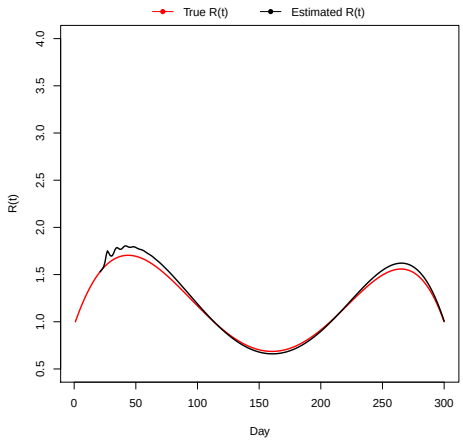
(a) Constant $R(t)$: outbreak simulation



(b) Multiple-waves $R(t)$: outbreak simulation



(c) Estimate of constant $R(t)$



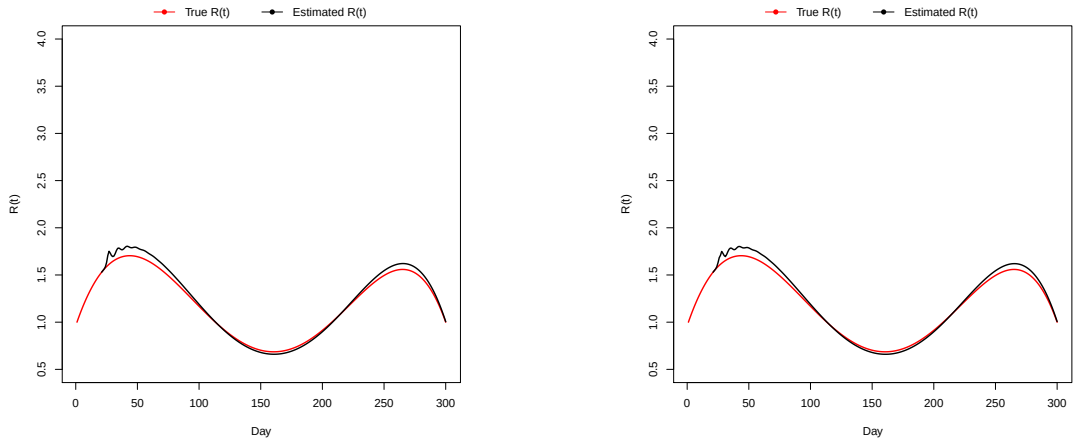
(d) Estimate of multiple-waves $R(t)$

Figure 5: Simulation of an SI model: the top panels display the evolution of an outbreak which follows two different processes for the true reproduction number as well as a different size in population (this is simply the case in order to obtain two waves and avoid infecting the whole population at an earlier stage in panel (b)). Underneath, the resulting $\widehat{R}(t)$ is displayed together with its true value.

at the minimum and maximum of the $R(t)$ function in Figure 5d. One explanation for this comes from the stable conditions assumptions that the TSI model makes in order to be able to separate the infectiousness $\beta(t, \tau)$. $R(t)$ is a proxy for future rates of change in cases, so when $R(t)$ is increasing, the future rate of change in cases is increasing, which appears to result in the overestimate of the instantaneous $R(t)$ that we see in Figure 5d. We see the inverse scenario as $R(t)$ is decreasing, with an underestimate on the way down.

Finally, we go one step further and relax the assumption that we have knowledge of the underlying serial interval distribution type. In Figure 6 we simulate an SI-TSI model where the underlying distribution is a weibull. We then compare the $R(t)$ estimates if we assume a gamma

distribution where the parameters are estimated from the serial interval study, and a non-parametric $R(t)$ estimator. We can see that when we don't know the shape of the underlying distribution, the non-parametric estimator performs just as well as the parametric one, and may therefore be preferred if there is no prior knowledge of the true, underlying process.



(a) Parametric estimate

(b) Nonparametric estimate

Figure 6: Estimated $R(t)$ under different Serial Interval estimation techniques.

4 Understanding an Evolving Virus

Late December 2020, authorities started to inform about evolutions of the SARS-CoV-2 genome sequences. A diversification of the virus due to evolution and adaptation processes has been observed globally. The rapid implementation of open-source sharing of virale genome sequences have eased real-time detection and tracking of variants. All viruses observe continuous mutations, and most mutations in the viral genome that emerge then quickly recede as they share the same characteristics of the main lineage. However, if a mutation provides selective advantage to the variant, it is considered variants of concern (VOC). A variant of Concern is defined by at least one of the following characteristics: a higher transmissibility, infection fatality or reinfection rate than the original lineage, or if current drugs and vaccine are inefficient against it.

In Europe, the British variant, known as variant B.1.1.7 rapidly expanded.[7] The lineage was detected in November 2020 in the South East region of England. The B.1.1.7 carries a mutation

[7]Other potentially dangerous variants are reported by the WHO, such as the Brazilian variant (named P.1), and the South Africa variant (B.1.351) at the time of the writing.

in the S protein, producing negative results for the S-gene target when sequencing PCR tests. Hence, the absence of S gene detection, called S-gene target failure (SGTF) can serve as a proxy for identifying B.1.1.7 cases[8]. The British variant was classified as variant of concern as it displays higher transmissibility rate than the original lineage. Though the number of studies are limited, studies found scientific evidence for an increase in R_t by a factor between 1.35 and 1.75 [(Volz et al., 2021), (Davies et al., 2021), (Sarah et al., 2020)].

Shinde, Bhorat, and Lalloo (2021) on the other hand provide evidence of immune escape for the South-African variant: individuals previously infected with preexisting variants have a degree of susceptibility to reinfection. The study shows that having antibodies from the original lineage do not protect against infection by the B.1.351 variant, with a risk of reinfection of 5%.

Studying variants in modelling epidemiology is particularly relevant, as even with a number of cases slowing down and an effective reproduction number less than 1, if the effective reproduction number of the new variant is greater than 1, the number of infected will eventually increase toward a new disease wave (Ramos et al., 2021). In other words, in contrast with common thoughts, $R_t < 1$ is not enough for having the spread of the disease under control, if a more contagious new variant is active. It is important to understand how mutations may affect the spread of covid-19 to better formulate public health responses, and in particular whether variants require changes in existing measures for disease monitoring and containment.

4.1 Modeling VOC: SII TSI Model

To explicitly model and visualize the spread of VOC within the greater pandemic, we extend the TSI SI model specified in Section 3 to include two infection compartments. This yields a Susceptible-Infected1-Infected2 (SII) compartmental model. Figure 7a displays total cases, daily cases, and $R(t)$ estimation for the SII simulated cases. The variant is set to start later in the simulated outbreak, so that we see it quickly take over the original lineage after day 150. Of particular interest is the right panel of Figure 7a, where we can see that failing to track the two disease lineages separately (the black R(t) estimate line), would lead policy makers to poorly forecast the future disease progression.

Figure 7b extends this to another scenario, which we call the "panic" scenario[9]. The trajectory

[8]Other mutations can cause SGTF, but lineage B.1.1.7 has been found to represent more than 96% of cases with SGTF

[9]The "panic" scenario corresponds to the modelling of $R(t)$ as a polynomial of degree 5 as presented in Subsection 2.2

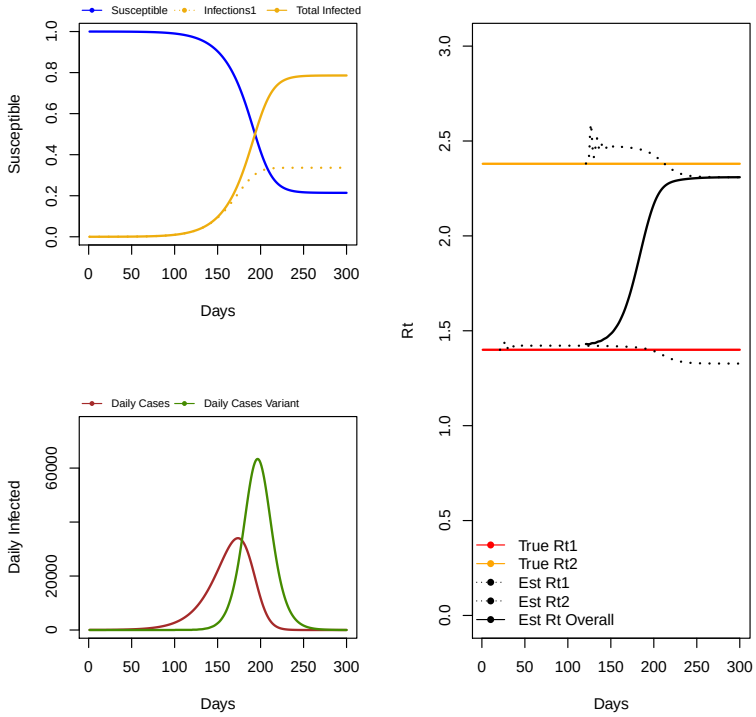
of $R(t)$ for the original lineage is such that the outbreak is very severe at first, which leads to caution from people in response, followed by a slight relaxing of that caution later on. The variant is again modeled to have a constant $R(t)$ implying again that the standard cautionary measures people took which were effective against the original lineage did not impact the rate of transmission of the VOC.

In Figure 7a and Figure 7b it is initially tempting to look at the upper left panel of both Figures, and use the dotted Infections1 line as a counterfactual on the trajectory of the original lineage if the VOC had never existed. However, since within this modelling framework the original lineage and the VOC both eat into each others susceptible populations, the introduction of the VOC actually caused the outbreak of the original lineage to die out faster than it otherwise would have. Therefore, the counterfactual scenario without the VOC would be in between the ending percentage of infections from the original lineage and the percentage of both lineages. While not perfect, this does provide upper and lower bounds which is useful for a more detailed counterfactual analysis.

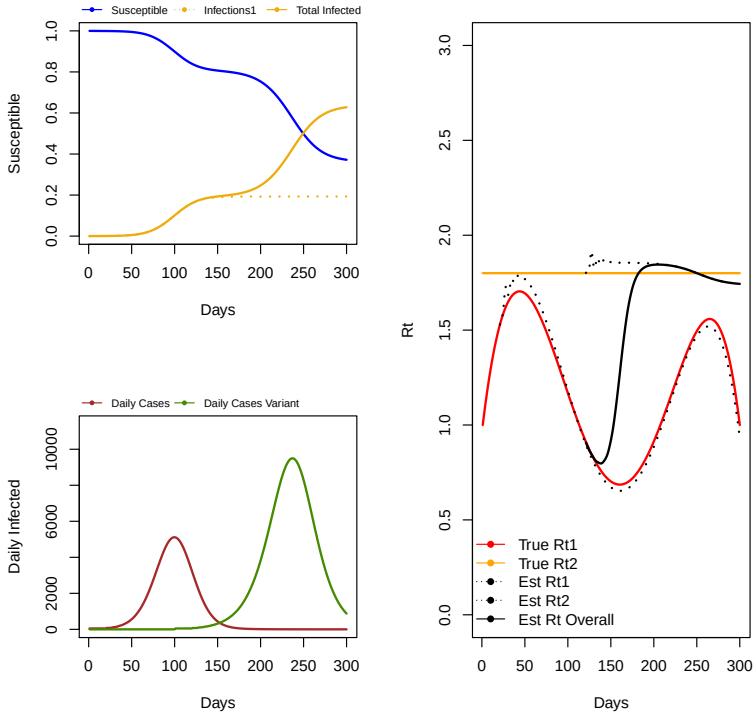
5 Application to Covid-19 Case Data

Having set the theoretical framework, we now apply the outlined models and estimation techniques to real-world French data. This empirical exercise uses two different datasets: the first one contains total daily cases and daily new cases, and the second one daily new cases of each variant. Both datasets are available on the website of *Sante Publique France*.[10] France started recording daily cases on February 2^{nd} 2020. Data on virus variants, however, has only been recorded since February 12^{th} 2021. Recording variant cases is a challenging exercise. In fact variants are detected only through genome sequencing. Fortunately, France provides the necessary data to study variants as sequencing is applied to all positive PCR test. In France, the British variant, known as variant B.1.1.7 is now the most frequent lineage. However, there are two other variants of concern circulating: the Brazilian variant (P.1), and the South Africa variant (B.1.351). France reports the daily percentage of sequenced test that are British variants, South African and Brazilian variant together, original lineage and all undetermined variants. Using these two dataset, we estimate $R(t)$ and forecast the future spread of the virus, first in a context of single lineage, and then taking into account the British variant circulating.

[10] https://www.data.gouv.fr/en/organizations/sante-publique-france/



(a)



(b)

Figure 7: Simulated outbreak in the presence of a variant: the two panels on the left of each graph display the evolution of the epidemic under two different specifications of the $R(t)$'s of each variant. The true shape of these reproduction numbers is depicted in golden and red in the right panel, together with the estimated overall $R(t)$ -solid red line, and the individual $R(t)$'s -dashed black lines, which is only feasible if infections are reported separately for each variant.

5.1 From theory to empirics

In order to forecast future infections in the context of the compartmental TSI model we need a forecast of $R(t)$. The TSI model builds on the assumption of static conditions during the period of concern (as discussed in detail in Subsection 2.1). This assumption permits the separation of infectiousness into a function dependent on time since infection, and a function dependent on calendar time. Moreover, this assumption has implications for the type of forecasting model we use to forecast $R(t)$ within each calendar time regime. That is, we cannot specify a model that would allow for changes in $R(t)$ that resulted from changing conditions. For this reason, within each of our estimated date ranges, we assume that the behavior of $R(t)$ follows a white noise process around a mean value.

The presence of noise in the real world estimation of $R(t)$ comes from a few sources we assume to be independent of calendar time. These are (1) intra-weekly variation in the level of testing (the weekend effect), (2) variation in the number of false positives, (3) variation in the overall level of testing, and (4) variation in the proportion of asymptomatic cases. These errors will have a direct and immediate effect on the observed daily cases, which will in turn introduce noise into our estimation of $R(t)$. So, for our forecasted cases in the short term horizon, we assume $R(t)$ in date regime T follows

$$R_T(t) = \mu_T + \epsilon_T(t) \quad \text{where} \quad \epsilon_T(t) \sim \mathcal{N}(0, \sigma_T^2)$$

Following from this model, the predicted value for $R_T(t)$ within this calendar time regime is the sample mean of the estimated $R(t)$ values. To verify that this is a plausible model, we need to understand the basic time series dynamics of the $R(t)$ estimates, including the presence of autocorrelation and whether the series is stationary.

Figure 8 shows the daily $R(t)$ estimates in blue, as well as the 7-day rolling mean of $R(t)$ for the entire observed time period in red. While we cannot make the assumption of stable conditions for the entire period, visualizing the whole series is a useful first step. The 7-day rolling mean introduces extra persistence (by design) in the data, which complicates testing for autocorrelation and stationarity. For this reason, while the 7-day rolling average makes sense to use when forecasting in order to remove the weekend effect, we focus on the daily estimates to

verify the validity of Subsection 5.1.

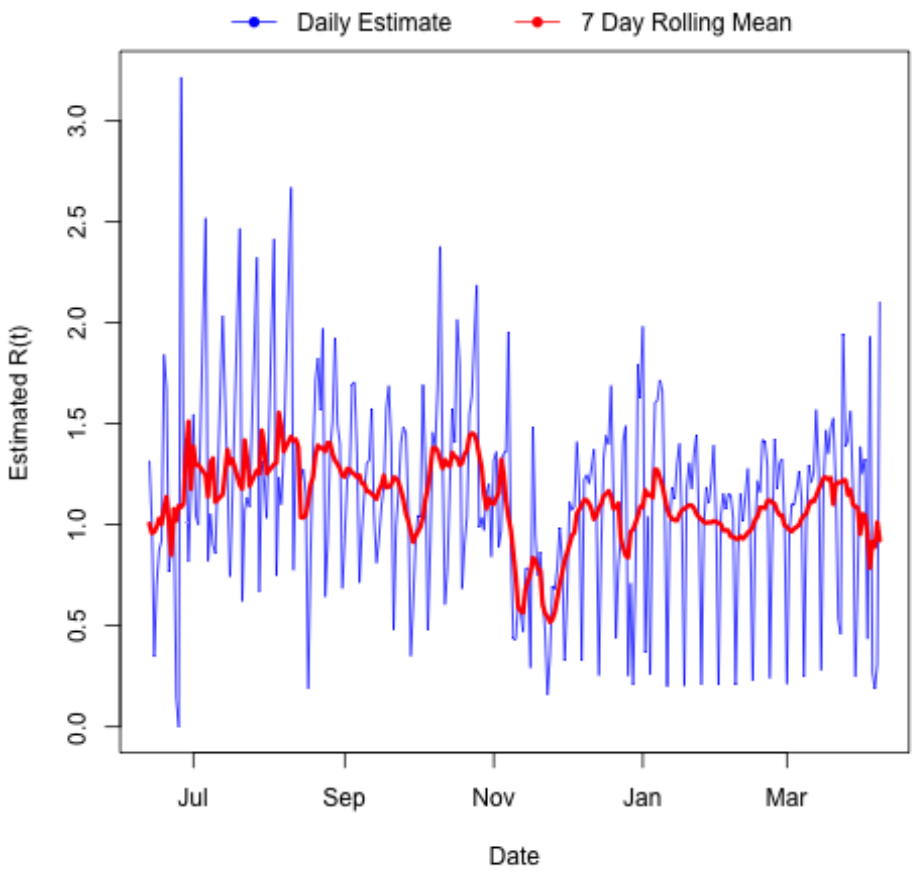
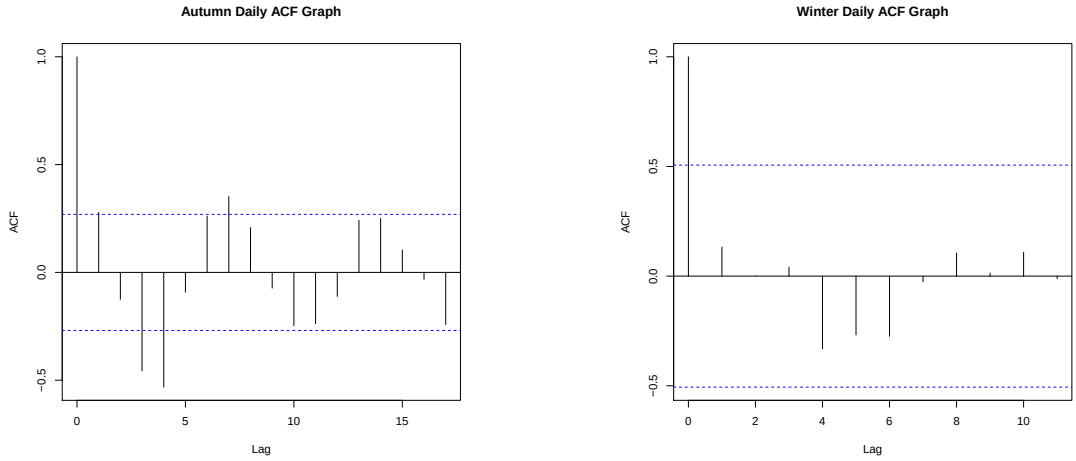


Figure 8: $R(t)$ Estimate Full Time Period

In each case of both the autumn and winter, we find that the best ARIMA model for the daily estimate time series includes only an intercept term, while the 7-day rolling mean series includes a first difference, implying that the 7-day series' both contain a unit root. We additionally run KPSS tests on both the Autumn and Winter $R(t)$ time series and fail to reject the null hypothesis of stationarity. Figure 9a displays the ACF plot of the daily estimated $R(t)$ series for Autumn while Figure 9 represents the Winter period. Although the autumn time period shows some autocorrelation, the best model is still only an intercept term in both cases.

5.2 Empirical setting

We identify periods in the data where no major policy changes took place, i.e. periods of static conditions. We first estimate R_t on a low-circulating period, with few non-pharmaceutical interventions: July 15^{th} 2020 to October 1^{st} 2020. During this period, the virus was considered



(a) (b)

Figure 9: ACF Graphs Daily Data

under control, and no lockdown or curfew were implemented. Second, we apply our estimator to a period of a high level of social distancing: October 15^{th} 2020 to January $1st$ 2021. On October 14^{th}, France first introduced a curfew in major urban areas, quickly follow by a general curfew in 38 departments on October 22^{nd}, and by a strict lockdown combined with a curfew on October 30^{th}. While the lockdown ended on December 15^{th}, social distancing was maintained with a 6pm curfew. These periods then constitute a model consistent environment to employ the TSI based R_t estimator shown in Equation 7.

Generally, as described in Subsection 2.2, a crucial parameter when estimating R_t is the serial interval, usually estimated using infected-infectee data. However, as the study data used to investigate the serial interval is not publicly available, we take the distribution estimated by *Sante Publique France* as given. At the beginning of the epidemic, countries had to obtain information on the serial interval as quickly as possible in order to conduct analysis, and many countries did not conduct their own studies. Due to this time constraint, France decided in March 2020 to rely on Nishiura, Linton, and Akhmetzhanov (2020) for the serial interval estimation. While more timely studies of serial interval where conducted on french data, we will follow the official estimation which still relies on the initial study and take $\widehat{\omega}(\tau)$ to be a Gamma distribution with mean 4.8 and standard deviation 2.8.

5.3 Estimating $R(t)$

For the sake of conciseness $\widehat{R}(t)$ over all periods of interest is shown in Figure 8. We observe that over the summer period, $R(t)$ progressively decreases but remains greater than 1. Some irregularities occur around August 15^{th}. Our estimator is very sensitive to the number of daily cases, which corresponds to the number of people that tested positive on a given day. As August 15^{th} is a public holiday, and also corresponds to one of the most common weeks for vacation, people got tested less around that date, hence spuriously impacting the number of daily cases. The second period of interest (November-December 2020) corresponds to a high social-distancing period. As expected, this non-pharmaceutical intervention leads to a sharp decrease in $\widehat{R}(t)$: our estimates fall below 1. Notably the estimated values of $R(t)$ match the official estimates provided by *Sante Publique France*.[11]

5.4 Forecasting Results

Using the sample mean of the estimated values for $R(t)$ in the respective periods of interest we can forecast the spread of the virus by employing the simulation framework outlined in Section 3. The forecasting period is 20 days following the last date for which $R(t)$ has been estimated, i.e the respective ending dates of periods we assume to have static conditions. The initial number of cases is taken from the actual data. $\widehat{\omega}(\tau)$ is again assumed to be a Gamma distribution with mean 4.8 and standard deviation 2.8.

The results for the first forecasting period, i.e. the 20 days following October 1^{st} 2020 are shown in Figure 10. Despite smoothing the trend the forecasted number of cases generally matches the observed cases. Linking this back to reality we can see that the Autumn period without intervention indeed led to the now so-called second wave of Covid. In this context a standard TSI model proved to be a sufficient tool to predict the future rise of case numbers. It hence remains a puzzle why policy makers failed to react in a timely manner. However, answering this question is outside of the scope of this project and likely the answer is not to be found in any model framework where humans are assumed to behave rationally.

However, applying this forecasting framework to the second period of interest yields a much different picture and direly highlights its limitations, as seen in Figure 11. As we estimate $R(t)$ to be < 1 during the chosen lockdown period our model predicts a falling number of cases for

[11] https://www.santepubliquefrance.fr/dossiers/coronavirus-covid-19/coronavirus-chiffres-cles-et-evolution-de-la
block-266151

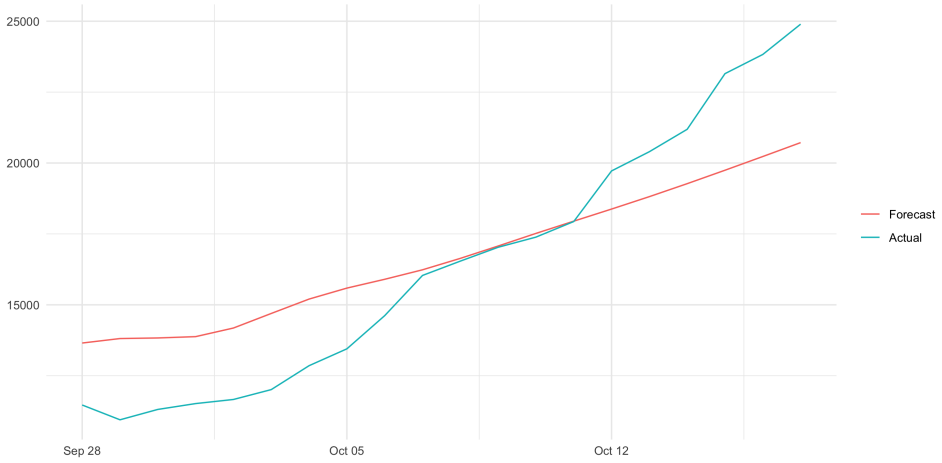


Figure 10: 20 Day Forecast Autumn 2020

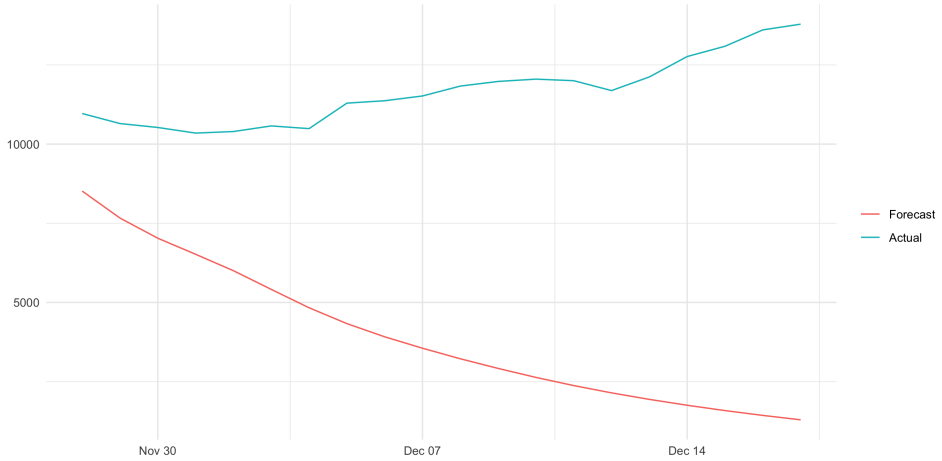


Figure 11: 20 Day Forecast Winter 2020

the subsequent 20 days. In reality, however, there were multiple factors at play that prevented the pandemic from coming to a halt. First, we argue that after January 1^{st} 2021 we observe in the data the lax adherence to social distancing rules over the Christmas holidays. Second, at this point in time the British virus variant was already circulating in France without being accounted for in the $R(t)$ estimation. We hence expect our values for $\widehat{R}(t)$ to be downward biased for this period.

To illustrate this point further we apply the same forecasting procedure on data from spring 2021. In spring 2021, the British variant was actively circulating in France, and it's increased infectivity made it the main lineage circulating, taking over the original lineage. We again estimate $R(t)$ and conduct the forecasting procedure on the total number of daily cases, without distinguishing between variants. The resulting 20 days forecast is displayed in Figure 12. As

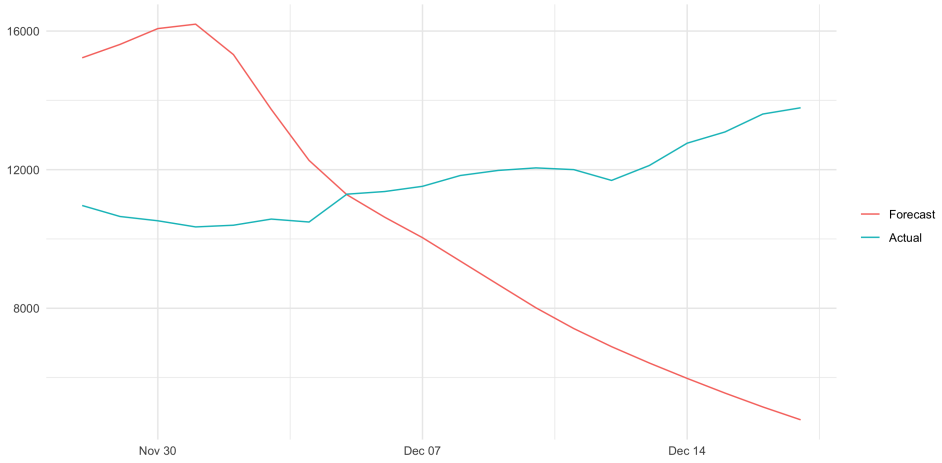


Figure 12: 20 Day Forecast Spring 2020

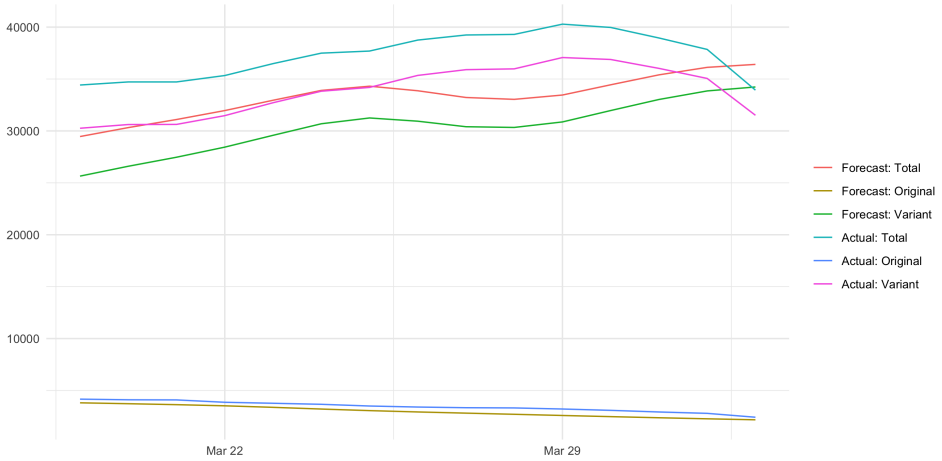


Figure 13: 20 Day Forecast Variant

expected, $\widehat{R}(t)$ is downwards biased and we in turn observe that the model predicts falling number of cases.

Contrarily, if we make use of case data by variant to employ our forecast procedure we are much more successful obtaining forecasts that are closer to observed cases. Indeed, as Figure 13 clearly shows, the model now manages to account for the fact that the number of original lineage cases is declining while the dominant new variant is driving cases up. These forecasting results highlight the importance of taking variants into account when estimating $\widehat{R}(t)$. As the different variants have different $R(t)s$, failing to take them into account in the model biases the results.

6 Conclusion

This paper presented and discussed some of the underlying difficulties when modelling and estimating the dynamics of a pandemic using a widespread TSI model. Departing from a detailed theoretical presentation of the model and its adoption to data observed in the real world, potential sources of biases as well as the properties of a parametric and a nonparametric estimator of the instantaneous reproduction number $R(t)$ were discussed in different settings. It was shown that an accurate specification of the serial interval estimate is crucial for obtaining a good estimate of $R(t)$. Furthermore, the case of simultaneous pandemics through variants was considered and the consequence of not observing a variant with a higher infectivity in the estimation of $R(t)$ was stressed. Since a pandemic is constantly evolving, the need to adopting the estimation techniques and updating the disease related parameters frequently is crucial for obtaining precise information about the current state of a pandemic. Especially the profound and timely study of the underlying serial intervals across different regions is crucial as well as the early detection of variants.

References

Ali, S. T., L. Wang, E. H. Lau, X. K. Xu, Z. Du, Y. Wu, G. M. Leung, and B. J. Cowling (2020). "Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions". In: *Science* 369.6507, pp. 1106–1109. ISSN: 10959203. DOI: 10.1126/science.abc9004.

Atkenson, A. (2020). "What Will Be The Economic Impact of Covid-19 in the US?" In: *NBER Working Paper Series* 53.9, pp. 1689–1699. ISSN: 1098-6596. arXiv: arXiv:1011.1669v3.

Baker, S., N. Bloom, S. Davis, and S. Terry (Apr. 2020). "COVID-Induced Economic Uncertainty". In: *National Bureau of Economic Research*. ISSN: 0898-2937. DOI: 10.3386/w26983. URL: www.worlduncertaintyindex.com,.

Cereda, D. et al. (2020). "The early phase of the COVID-19 outbreak in Lombardy, Italy". In: *arXiv*. arXiv: 2003.09320.

Chudik, A., M. H. Pesaran, and A. Rebucci (Apr. 2020). "Voluntary and Mandatory Social Distancing: Evidence on COVID-19 Exposure Rates from Chinese Provinces and Selected Countries". In: *National Bureau of Economic Research*. DOI: 10.3386/w27039. URL: http://www.nber.org/papers/w27039.pdf.

Cori, A., N. M. Ferguson, C. Fraser, and S. Cauchemez (2013). "A new framework and software to estimate time-varying reproduction numbers during epidemics". In: *American Journal of Epidemiology* 178.9, pp. 1505–1512. ISSN: 00029262. DOI: 10.1093/aje/kwt133.

Davies, N. G. et al. (2021). "Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England". In: *Science* 3055.March, eabg3055. ISSN: 0036-8075. DOI: 10.1126/science.abg3055.

Fraser, C. (2007). "Estimating individual and household reproduction numbers in an emerging epidemic". In: *PLoS ONE* 2.8. ISSN: 19326203. DOI: 10.1371/journal.pone.0000758.

Griffin, J., M. Casey, Á. Collins, K. Hunt, D. McEvoy, A. Byrne, C. McAloon, A. Barber, E. A. Lane, and S. More (Nov. 2020). "Rapid review of available evidence on the serial interval and generation time of COVID-19." In: *BMJ open* 10.11, e040263. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2020-040263.

Huang, L., X. Zhang, X. Zhang, Z. Wei, L. Zhang, J. Xu, P. Liang, Y. Xu, C. Zhang, and A. Xu (June 2020). "Rapid asymptomatic transmission of COVID-19 during the incubation period demonstrating strong infectivity in a cluster of youngsters aged 16-23 years outside Wuhan and characteristics of young patients with COVID-19: A prospective contact-tracing study". In: *Journal of Infection* 80.6, e1–e13. ISSN: 15322742. DOI: 10.1016/j.jinf.2020.03.006.

Istituto Superiore di Sanità (2021). *Monitoraggio Fase 2 Report settimanale Report 48 Sintesi nazionale*. Tech. rep. Istituto Superiore di Sanità. URL: https://www.iss.it/primo-piano/-/asset_publisher/o4oGR9qmvUz9/content/id/5477037.

Khalili, A., E. Petersen, M. Koopmans, U. Go, D. H. Hamer, N. Petrosillo, F. Castelli, M. Storgaard, and S. Al Khalili (2020). "Personal View Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics". In: *The Lancet Infectious Diseases* 20, e238–e244. DOI: 10.1016/S1473-3099(20)30484-9. URL: www.thelancet.com/infection.

Li, Q. et al. (Mar. 2020). "Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia". In: *New England Journal of Medicine* 382.13, pp. 1199–1207. ISSN: 0028-4793. DOI: 10.1056/NEJMoa2001316. URL: http://www.nejm.org/doi/10.1056/NEJMoa2001316.

Nishiura, H., N. M. Linton, and A. R. Akhmetzhanov (2020). "Serial interval of novel coronavirus (COVID-19) infections". In: *International Journal of Infectious Diseases* 93, pp. 284–286. ISSN: 18783511. DOI: 10.1016/j.ijid.2020.02.060. URL: https://doi.org/10.1016/j.ijid.2020.02.060.

Rai, B., A. Shukla, and L. K. Dwivedi (2021). "Estimates of serial interval for COVID-19: A systematic review and meta-analysis". In: *Clinical Epidemiology and Global Health* 9.August 2020, pp. 157–161. ISSN: 22133984. DOI: 10.1016/j.cegh.2020.08.007. URL: https://doi.org/10.1016/j.cegh.2020.08.007.

Ramos, A. M., M. Vela, M. R. Ferrández, A. B. Kubik, and B. Ivorra (2021). "Modeling the impact of SARS-CoV-2 variants and vaccines on the spread of COVID-19". In: *ResearchGate preprint* January. DOI: 10.13140/RG.2.2.32580.24967/2. URL: https://www.researchgate.net/publication/348559868_Modeling_the_impact_of_SARS-CoV-2_variants_and_vaccines_on_the_spread_of_COVID-19.

Sarah, W., Z. J. Madewell, Y. Yang, I. M. L. Jr, M. E. Halloran, and N. E. Dean (2020). "Increased infections, but not viral burden, with a new SARS-CoV-2 variant". In: *medRxiv* 165, pp. 1–13.

Shinde, V., Q. Bhorat, and U. Lalloo (2021). "Preliminary Efficacy of the NVX-CoV2373 Covid-19 Vaccine Against the B.1.351 Variant". In:

Toda, A. A. (Mar. 2020). "Susceptible-Infected-Recovered (SIR) Dynamics of COVID-19 and Economic Impact". In: *arXiv*. arXiv: 2003.11221. URL: http://arxiv.org/abs/2003.11221.

Vink, M. A. (2010). "The generation interval of infectious diseases". PhD thesis. Utrecht University, pp. 1–44.

Volz, E., S. Mishra, M. Chand, J. C. Barrett, R. Johnson, S. Hopkins, A. Gandy, A. Rambaut, and N. M. Ferguson (2021). "Transmission of SARS-CoV-2 Lineage B.1.1.7 in England: Insights from linking epidemiological and genetic data". In: *medRxiv*, p. 2020.12.30.20249034.