IMAGE PROCESSING ON LINE

Published in Image Processing On Line on YYYY–MM–DD.
Submitted on YYYY–MM–DD, accepted on YYYY–MM–DD.
ISSN 2105–1232 © YYYY IPOL & the authors CC–BY–NC–SA
This article is available online with supplementary materials,
software, datasets and online demo at
https://www.ipol.im/pub/pre/301/

An empirical algorithm to forecast the evolution of the number of COVID-19 symptomatic patients after social distancing interventions

Luis Alvarez[1], Miguel Colom[2], and Jean-Michel Morel[2]

[1] Departamento de Informática y Sistemas, Universidad de Las Palmas de Gran Canaria, Spain
(lalvarez@ulpgc.es)
[2] Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-94235, Cachan, France
({miguel.colom-barco, jean-michel.morel}@ens-paris-saclay.fr)

PREPRINT November 16, 2020

Abstract

We present an empirical algorithm to forecast the evolution of the number of COVID-19 symptomatic patients after social distancing interventions. The algorithm is based on a low dimensional model for the variation of the exponential growth rate that decreases after the implementation of the social distancing measures. From the observable data given by the number of tested positive, our model estimates the number of infected hindcast introducing in the model formulation the incubation time. We also use the model to follow the number of infected patients who later die using the registered number of deaths and the distribution time from infection to death. We present some experiments to show the ability of the model to properly forecast the epidemic spread at the beginning of the epidemic outbreak when very little data and information about the coronavirus were available. In the case of France, we obtain a correct estimate of the peak of the new cases of tested positives 9 days in advance and only 7 days after the implementation of a strict lockdown. Moreover, using an extended model we study the timeline of the first wave obtaining a precise knowledge of the chronology of the main epidemiological events during the full course of the first wave in South Korea, Italy, France, Spain, Germany, United Kingdom, The New York state and USA. In particular we estimate the number of days the coronavirus was in free circulation before the social distancing measures take effect. Moreover, using the results obtained in the different regions, we obtain that the initial exponential growth rate of the epidemic, when the coronavirus is in free circulation, is around the value 0.250737.

Source Code

To simplify the presentation of the code we use a single file where all basic procedures are included, from the management of data to the parameter optimization. The reviewed source code and documentation for this algorithm are available from the web page of this article[1]. Compilation and usage instruction are included in the README.txt file of the archive.

Keywords: COVID-19, forecast, social distancing

1 Introduction

In this work we propose an empirical parametric model to forecast the evolution of the number of COVID-19 symptomatic patients, $N(t)$, after social distancing interventions. This study presents a numerical analysis of the effect of the confinement phase of the pandemic. It attempts to predict the evolution of the number of cases and deaths, based on past observations and assuming that the social distancing policy is steady or evolves slowly. Hence the main assumptions we take are:

1. The evolution of the cumulative number of contaminated patients, $y(t)$, grows at an exponential rate (that we name a), during a period of time t_0. We thus have $y'(t) = ay(t)$. Then, after social distancing measures are imposed the exponential rate $r(t)$ (such that $y'(t) = r(t)y(t)$) decreases until it attains the value 0 at time t_1. In this study, we considered first the following type of evolution for the exponential rate:

$$r_1(t) = \begin{cases} a & if \quad 0 \leq t \leq t_0 \\ b & if \quad t \in (t_0, t_1] \\ 0 & if \quad t > t_1 \end{cases} \tag{1}$$

   but we realized that the next parametric model, with the same number of unknowns, was more flexible and accurate:

$$r(t) = \begin{cases} a & if \quad 0 \leq t \leq t_0 \\ a\left(\frac{t_1 - t}{t_1 - t_0}\right)^{\gamma} & if \quad t \in (t_0, t_1] \\ 0 & if \quad t > t_1. \end{cases} \tag{2}$$

   In the first model the parameters for $r_1(t)$ are a, b, t_0 and t_1, and the parameters for $r(t)$ are a, γ, t_0 and t_1. The values for a, b and γ are always positive. The larger the value of γ the stronger the effect of the social distancing measures on the growth of $N(t)$.

2. The evolution of the number $N(t)$ of the symptomatic patients at time t depends on the evolution law of contaminated patients, and on the law of the incubation period.

3. At the beginning of the epidemic outbreak, the data of tested positive patients provided by most countries can be assumed to concern mostly symptomatic patients. This is a reasonable assumption in the countries where tests were performed only on patients which show some symptoms. It is important to point out that the available databases about the coronavirus expansion make no distinction between infected subjects which show symptoms or not. If we assume that the number of symptomatic patients is proportional to the number of registered infected subjects, the model still works. This is a reasonable assumption as long as a country keeps the same infection test policy. If a country changes its testing policy and starts testing

[1] https://www.ipol.im/pub/pre/301/

more subjects, then many non-symptomatic subjects are going to be included in the dataset, which can strongly deteriorate the accuracy of any observational model. This is why our model (2) for the decay of the exponential rate is merely empirical, and aims at the simplest formulation possible.

4. The social distancing measures are taken at the beginning of the epidemic and there are many more exposed subjects than infected and recovered ones, so that we can assume that the variation of the symptomatic patients only depends on the existing contaminated patients and the influence of the social distancing measures (see below the relation with the SIR model).

Regarding the distribution of the incubation period, Lauer et al. in [6], using the data of 181 patients approximate the distribution of the incubation period as a log-normal distribution. The cumulative distribution function of this log-normal is given by

$$F(t) = \begin{cases} \int_0^t \frac{e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}} dx & if \quad t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

with $\mu = 1.621$ and $\sigma = 0.418$.

The rest of the paper is organized in the following way: in section 2 we study the solution of equation (4). In section 3, we analyze the relation of this model with the usual SIR model. In section 4, we present a short discussion about the lack of reliability of the available data of the COVID-19 spread. In section 5, we present the algorithm proposed to fit the model to the data. In section 6, we present an extension of the empirical model to the forecast of the number of deaths. Section 7 presents some experiments which focus on two aspects of the the epidemic spread: the ability of the model to predict the epidemic evolution in advance and the study of the full course of the first wave. Finally section 8 concludes.

2 The empirical evolution model

The continuous version of the evolution of contaminated subjects, $y(t)$, following an exponential grow, $r(t)$, is given by the very basic differential equation:

$$y'(t) = r(t)y(t).$$

This equation can be solved explicitly, and in the case of $r(t)$ given by (2) the solution is

$$y(t) = Ce^{\int_0^t r(s)ds} = \begin{cases} Ce^{at} & if \quad t \in [0, t_0] \\ Ce^{at_0} e^{\frac{a}{\gamma+1}\left((t_1-t_0)-\left(\frac{t_1-t}{t_1-t_0}\right)^{\gamma}(t_1-t)\right)} & if \quad t \in [t_0, t_1] \\ Ce^{at_0} e^{\frac{a}{\gamma+1}(t_1-t_0)} & if \quad t > t_1, \end{cases} \tag{4}$$

we notice that C and t_0 are closely related because if we change t_0 by $t_0 - T$, C by Ce^{aT} and t_1 by $t_1 - T$, the solution $y(t)$ does not change. Therefore t_0 is an "abstract" time and we can not say that the coronavirus has been actually in free circulation for t_0 days. What we can say is that the coronavirus has been in free circulation until $y(t)$ reaches the value Ce^{at_0} but, from the above formula, we can not decide the actual starting time of the epidemic outbreak.

The asymptotic state of the number of contaminated subjects is

$$Lim_{t\to\infty}\ y(t)=Ce^{at_0}e^{\frac{a}{\gamma+1}(t_1-t_0)}, \tag{5}$$

and it is attained at t_1. Therefore the impact of the social distancing measures is determined by the value :

$$M_{a,\gamma,t_1,t_0}=\frac{a}{\gamma+1}(t_1-t_0). \tag{6}$$

The smaller this value, the more effective the social distancing interventions. We notice that the peak in the new daily contaminated patients is obtained when $y'(t)$ changes sign which corresponds to $y''(t)=0$. Using a straightforward computation we obtain that the peak is attained at

$$t_{peak}=t_1-\left(\frac{\gamma}{a}(t_1-t_0)^{\gamma}\right)^{\frac{1}{\gamma+1}}$$

The evolution of symptomatic subjects, taking into account the cumulative distribution of the incubation time, $F(t)$, is given by

$$N(t)=\int_0^t y'(s)F(t-s)ds. \tag{7}$$

We observe that since $F(t)$ converges to 1 when t goes to ∞, then

$$Lim_{t\to\infty}N(t)=Ce^{at_0}e^{\frac{a}{\gamma+1}(t_1-t_0)}-C.$$

Notice that $y(0)=C$, $N(0)=0$ and there is a delay between the evolutions of $y(t)-C$ and $N(t)$, corresponding to the time required by contaminated subjects to become symptomatic. Notice that the official number of cases is related to $N(t)$ which is the variable that can be observable. $N(t)$ does not follow, in general an exponential growth. In fact:

$$a\approx\log\frac{y(t+1)}{y(t)}\neq\log\frac{N(t+1)}{N(t)}=\log\frac{\int_{t_0}^{t_1}y'(s)F(t+1-s)ds}{\int_{t_0}^{t_1}y'(s)F(t-s)ds}.$$

In practice, the observable data is the number of registered tested positive patients. An extra time is required from the moment the patient shows symptoms until the test is done and it is finally recorded as tested positive. This time strongly depends on the in-country logistics. In this work we assume that this time is about 2 days, so in expression 7 we replace $F(t-s)$ by $F(t-s-2)$. This modification does not change the profile of $N(t)$, it simply represents an extra delay between the evolution of contaminated subjects and the evolution of the registered tested positive subjects.

2.1 An extended model to track different trend modifications

The exponential growth given by equation (2) is very simple and it is useful to compute an estimation of $y(t)$ after an strict lockdown is implemented, this estimation covers from the epidemic outbreak until a certain time after the daily peak. However if we want to go further and to approximate the evolution for a longer time we need to extend the model to have more flexibility in order to fit the epidemic spread. In fact, the above basic model can be easily extended in the following way: let $\{t_0^k,t_1^k\}_{k=1,..,K}$ be 2 increasing sequences of real numbers satisfying $t_0^{k+1}<t_1^k$. $\{t_0^k\}$ represent times

where a change is expected in the evolution trend of the epidemic. Then the exponential growth (2) can be extended in the following way:

$$r(t) = \begin{cases} a_1 & if \quad t \in [0, t_0^1] \\ a_k \left(\frac{t_1^k - t}{t_1^k - t_0^k} \right)^{\gamma_k} & if \quad t \in (t_0^k, t_0^{k+1}] \quad k = 1, 2, .., K-1 \\ a_K \left(\frac{t_1^K - t}{t_1^K - t_0^K} \right)^{\gamma_K} & if \quad t \in (t_0^K, t_1^K] \\ 0 & if \quad t > t_1^K. \end{cases} \tag{8}$$

Notice that $r(t)$ can be discontinuous at t_0^k because a relaxation of social distancing measures will definitely produce an abrupt modification in the exponential growth of the epidemic. We point out that the function $r(t)$ is always decreasing except at the possible points of discontinuity t_0^k . In particular, the model is not well adapted to scenarios where the growth rate can grow continuously, such as a second epidemic wave.

3 Relation with the SIR model

The basic SIR model separates the population in three compartments: $S(t)$ (the number of susceptible), $I(t)$ (the number of infectious), and $R(t)$ (the number of recovered). It should be mentioned in this model that the number of dead is negligible. We can also consider that R(t) is the sum of recovered and deceased. Each member of the population typically progresses from susceptible to infectious to recovered. The basic SIR model to estimate $S(t)$, $I(t)$ and $R(t)$ is the following system of ordinary differential equations:

$$\begin{array}{l} \frac{dS}{dt} = -\beta \frac{I}{I+S+R} S \\ \frac{dI}{dt} = \left(\beta \frac{S}{I+S+R} - \gamma \right) I = \left(R_0 \frac{S}{I+S+R} - 1 \right) \gamma I \\ \frac{dR}{dt} = \gamma I, \end{array}$$

where β and γ are parameters which depend on the particular disease. $R_0 = \frac{\beta}{\gamma}$, named the reproductive number, is one of the key parameters in transmission models and it represents the number of secondary infections that arise from a typical primary case in a completely susceptible population. Notice that $S(t)$, the number of susceptible subjects, is a decreasing function. When the ratio between $S(t)$ and the total population satisfies

$$\frac{S(t)}{I(t) + S(t) + R(t)} = \frac{1}{R_0},$$

we obtain $\frac{dI}{dt}(t) = 0$. Hence the peak of infected subjects is attained, and from that time, the number of infected subjects starts decreasing. Notice that the larger R_0, the larger the time required to attain the infection peak. We observe that in our model, in the evolution of contaminated patients, $y(t)$, we include the infected and recovered subjects, so $y(t) = I(t) + R(t)$ and then using the SIR model we obtain that

$$\frac{dy}{dt}(t) = \beta \frac{S(t)}{I(t) + S(t) + R(t)} (y(t) - R(t)). \tag{9}$$

The SIR model with constant γ and β and the conclusions about the peak of infected subjects make sense only if the virus propagates freely across time, but everything changes if we impose social distancing measures to the population. A natural way to include human interventions in the SIR model is to replace β by a time dependent function $\beta(t)$. This strategy has been used by different authors in different contexts using extended versions of the SIR models. For instance in [3], the authors propose the following exponential type function:

$$\beta(t) = \beta_1 + (\beta_0 - \beta_1)e^{-q(t-t_0)_+}$$

another exponential type function has been introduced in [7]:

$$\tau(t) = \tau_1 e^{-\mu(t-N)_+}$$

In [1] the following rational function is proposed:

$$\beta(t) = \beta_0(1 - \rho(t - t_0)/t)$$

In [4] the author proposes the function

$$\tau(t) = \tau_0[1 - \mu(t - N)_+]_+. \tag{10}$$

We observe that this is a particular case of the function (2) defining $r(t)$ where $a = \tau_0$, $\mu = 1/(t_1 - t_0)$, $N = t_0$ and $\gamma = 1$. The only difference of this function with $r(t)$ is that in $r(t)$ we add the power γ to modulate the way the exponential growth rate decreases.

In this work we assume that social distancing interventions govern the evolution of contaminated subjects rather than the SIR dynamic and we replace equation (9) by

$$\frac{dy}{dt}(t) = r(t)y(t).$$

Therefore we include in the term $r(t)$ the impact of the human interventions, the influence of the ratio between $S(t)$ and the total population and the influence of $R(t)$. The latter makes sense if we are at the beginning of the pandemic (so $R(t) \approx 0$) or if we assume that $R(t)$ is proportional to $y(t)$. By focusing just on the number of contaminated subjects we reduce the complexity of the problem and we avoid to deal with the balance between infected, exposed and recovered patients which is very difficult to estimate properly due to the lack of accuracy in the data we can manage about the number of infected subjects. We point out that in our model $y(t)$ is the number of infected patients which show symptoms, which is the data most countries provide when using PCR tests.

4 A discussion about the reliability of the existing data about the coronavirus expansion in terms of the evaluation of the impact of social distancing interventions

Tested positive subjects: First, we stress again that what we can observe is the evolution of tested positive subjects, which is quite different from the evolution of contaminated subjects. This value strongly depends on the testing policy which can change across the time. If the testing policy does not change too much during the period of time used to estimate the model, our forecast will still be valid to some extent. This value has the advantage that it is the first one to react to the installation of social distancing measures.

Symptomatic tested positive subjects: With the existing variety of testing policies, this value seems to be more reliable than just tested positive subjects. On the one hand official data make no distinction between symptomatic and asymptomatic tested positive subjects. On the other hand, when the health system is overwhelmed, many symptomatic subjects not requiring hospitalization are simply sent home without testing.

Number of deaths: Theoretically, this is a reliable data, but when the health system is overwhelmed a significant number of patients die without being counted as affected by the coronavirus, so the accuracy of this data depends on the capacity of the health system to properly count the deaths. This is far from being the case when the health system is completely overwhelmed.

Number of hospitalizations or number of patients in intensive care: Again, theoretically, these data are more reliable than the number of tested positive, but again, in the case of a health system completely overwhelmed, the quality of these data is strongly deteriorated. Another issue with these data is the way they are provided. In some cases, the official data refer to the current situation where the patients which leave the hospital or reanimation are removed from the statistics.

Another important issue in the data quality is the time required for a new case to be included in official statistics. For example, if new PCR positive tests and new antibody positive tests are added at the same day, the quality of the data deteriorates seriously. Indeed, both detection correspond to infections at very different past times! Even using only PCR tests, the time from the presentation of symptoms to inclusion in official statistics must be taken into account. In Spain, this time is distributed with a median of 6 days; in 25% of cases it is even more than 10 days. This delay deteriorates the usability of the data, and hinders a short-term prediction of the evolution of the epidemic.

5 The algorithm

As discussed in the previous section, the data we use are far for being reliable. In our approach we use a very simple model with few parameters in the hope that the simplicity of the model can compensate in some way for the lack of accuracy of the data and provide a big picture of the evolution of pandemic expansion after social distancing interventions. We observe that if we have not enough data after the implementation of social distancing measures, the parameters γ and t_1 cannot be computed properly from the data. To reduce the uncertainty in the calculation of these parameters, we can set "a priori" the expected value of the effectiveness of the containment effectiveness given by M_{a,γ,t_1,t_0}, defined in (6), based on the values obtained for other countries which have implemented previously the same kind of social distancing measures. That is, we can constrain the effectiveness to satisfy

$$M_{a,\gamma,t_1,t_0} = \frac{a}{\gamma+1}(t_1 - t_0) = M_0 \tag{11}$$

where $M_0 > 0$.

5.1 Model discretization

We estimate $N(t)$ by discretizing equation (7) in the following basic way

$$N(t) = \int_0^t y'(s)F(t-s)ds \approx \sum_{k=0}^{k=t_1-1} (y(k+1)-y(k))F(t-(k+0.5)), \tag{12}$$

where $y(t)$ is given by (4).

5.2 Parameter adjustment

Given a dataset, $D(t)$, of the number of symptomatic patients across the time for a region, we fix the parameters by minimizing the quadratic mean error

$$Error_{[t_{min},t_{max}]}(C,a,\gamma,t_0,t_1,\tilde{t}) = \frac{1}{\sum_{t=t_{\min}}^{t_{\max}} w(t)} \sum_{t=t_{\min}}^{t_{\max}} w(t)\left(D(t)-N(t+\tilde{t})\right)^2, \tag{13}$$

where $\tilde{t}$ is the translation of $N(t)$ to fit $D(t)$. The interval $[t_{\min}, t_{\max}]$ is the range of values we use to fit the parameters of the model. We assign the following weight $w(t)$ to each data value in the model estimation:

$$w(t) = (t - t_{\min} + 1)^{\alpha} \tag{14}$$

where $\alpha \geq 0$. When $\alpha = 0$ all points in the dataset have the same weight ($w(t) \equiv 1$). The higher the value of α the more weight it will be giving to the latest values of the dataset. To adjust the model parameters, we use a Newton-Raphson type method combined with an extensive search exploration of potential parameter interval values.

Computation of C: we observe that C is a scale factor and if the other parameters of the model are given, C can be estimated by equating to zero the derivative of the error (13) with respect to C, which yields the following expression for C:

$$C = \frac{\sum_{t=t_{\min}}^{t_{\max}} w(t)D(t)N_1(t+\tilde{t})}{\sum_{t=t_{\min}}^{t_{\max}} w(t)N_1(t+\tilde{t})N_1(t+\tilde{t})} \tag{15}$$

where $N_1(t+\tilde{t}) = N(t+\tilde{t})$ is computed using $C = 1$ and the other given parameters. We point out that the values of C and t_0 are very related, in terms of the evolution of contaminated subjects, $y(t)$. Indeed the values $C' = Ce^a$ and $t_0' = t_0 - 1$ provide the same results as C and t_0.

5.3 Computing the minimum of $Error_{[t_{min},t_{max}]}(C,a,\gamma,t_0,t_1,\tilde{t})$

First, we observe that, in general, due to the strong variation of the available data, the quadratic error, $Error_{[t_{min},t_{max}]}(C,a,\gamma,t_0,t_1,\tilde{t})$ can have several local minima. To avoid getting trapped in a spurious local minimum, we use a basic optimization strategy, where we combine massive evaluations of $Error_{[t_{min},t_{max}]}(C,a,\gamma,t_0,t_1,\tilde{t})$ in large discrete intervals with a basic Newton-Raphson type method to improve locally $a,\gamma,\tilde{t}$. To simplify the complexity, we use (15) to express C as a function of the rest of parameters, that is, $C \equiv C(a,\gamma,t_0,t_1,\tilde{t})$, so the quadratic error becomes $Error_{[t_{min},t_{max}]}(a,\gamma,t_0,t_1,\tilde{t}) \equiv Error_{[t_{min},t_{max}]}(C(a,\gamma,t_0,t_1,\tilde{t}),a,\gamma,t_0,t_1,\tilde{t})$. The times t_0 and t_1 are computed in integer precision and the rest of parameters in floating precision. The computation of t_0 and t_1 in integer precision has little influence on the final result because small variations of t_0 are mostly compensated modifying C and small variations of t_1 are mostly compensated modifying γ.

MAIN STEPS OF THE OPTIMIZATION ALGORITHM

- Step 1: Computation of $t_{\min}$ and t_{max}. We define $t_{\min}$ as the the time when the data starts to grow with an exponential growth with a minimum value of 10, that is:

  $$t_{\min} = \min\{t : D(t-2) > 10 \text{ and } D(t) > 1.1D(t-1) \text{ and } D(t-1) > 1.1D(t-2)\}$$

  and we fix $t_{\max} = min\{t_{min}+N_1, t_c\}$, where N_1 is a parameter of the algorithm to fix the number of days used to compute the model. t_c is the max available time in the data set observation.

- Step 2: Initial estimation of $\tilde{t}$. We fix initially the following reference values for the rest of parameters: $a = 0.13$, $\gamma = 2$, $t_0 = 17$ and $t_1 = 52$. Then $\tilde{t}$ is computed initially in integer precision as

  $$\tilde{t}_0 = \underset{k \in N \cap [k_{\min}, k_{\max}]}{\arg\min} \; Error_{[t_{min}, t_{max}]}(a, \gamma, t_0, t_1, k)$$

  where $[k_{\min}, k_{\max}]$ has been fixed experimentally as $[k_{\min}, k_{\max}]$ =$[-26 - t_{\min}, 10 - t_{\min}]$.

- Step 3: Computing an initial minimum evaluating the energy in parameter intervals. For each parameter $p \in \{a, \gamma, t_0, t_1, \tilde{t}\}$ we define a discrete interval $I_p = \{p_1, .., p_{N_p}\}$ (in the case of $p = \tilde{t}$, I_p is a neighborhood of $\tilde{t}_0$ computed above) and we define the set $\mathcal{I} = I_a \times I_\gamma \times I_{t_0} \times I_{t_1} \times I_{\tilde{t}}$. We compute a first minimum, P_0, as

  $$P_0 = \underset{(a,\gamma,t_0,t_1,\tilde{t}) \in \mathcal{I}}{\arg\min} \; Error_{[t_{min}, t_{max}]}(a, \gamma, t_0, t_1, \tilde{t})$$

  This "brute force" technique has the advantage that it can be easily implemented using parallelization and to a certain extent, it avoids getting trapped in spurious local minima. Once P_0 is computed, it is improved using a basic Newton-Raphson method to optimize $a, \gamma, \tilde{t}$.

- Step 4. Improving iteratively the minimum location: For $k = 0, 1, 2, ..$ we use a small discrete neighborhood, N_{P_k}, of P_k, and we define P_{k+1} as

  $$P_{k+1} = \underset{(a,\gamma,t_0,t_1,\tilde{t}) \in N_{P_k}}{\arg\min} \; Error_{[t_{min}, t_{max}]}(a, \gamma, t_0, t_1, \tilde{t})$$

  after this initial estimation, P_{k+1} is improved using the Newton-Raphson method. Iterations stop when

  $$\frac{Error(P_k) - Error(P_{k+1})}{Error(P_k)} < TOL$$

  where TOL is a convergence parameter (we fix $TOL = 10^{-6}$ in the algorithm implementation). This iterative procedure allows to improve the minimum estimation. In particular, it allows the minimum to go beyond the initial parameter interval $\mathcal{I}$.

As quoted before, at the beginning of the epidemic spread, when not much data is available it can be useful to fix the expected value of effectiveness of the containment effectiveness given by M_{a,γ,t_1,t_0} defined in (6), in that case the value of M_{a,γ,t_1,t_0} becomes a parameter of the algorithm and this value constraints the parameter optimization steps of the algorithm.

5.4 Adaptation of the algorithm to the extended model

In the case where the exponential growth is given by the extended model (8), we compute the unknowns of the model, given by C, $\tilde{t}$ and $t_0^k, t_1^k, a_k, \gamma_k$ for k=1,..,K, in the following way:

1. We compute C, $\tilde{t}$ and $t_0^1, t_1^1, a_1, \gamma_1$ using the algorithm explained above.

2. for each k=2,..,K, we compute $t_0^k, t_1^k, a_k, \gamma_k$ iteratively in the following way:

   - $t_{max} = min\{t_{min} + \sum_{i=1}^{k} N_i, t_c\}$. (where N_k is a parameter of the algorithm to fix how many days we consider to compute the model k)
   - $t_0^k = \tilde{t} + t_{max} - N_k$
   - We compute t_1^k, a_k, γ_k by minimizing the quadratic error in the interval $[t_{min}, t_{max}]$ with respect to t_1^k, a_k and γ_k.

5.5 Description of the online DEMO parameters

The parameters we use in the online DEMO interface are the following:

1. *Type of data*: it can be tested positive or deaths.

2. *Number of days to compute the basic model*: this parameter is denoted by N_1 in the text. It represents the number of days used to compute the model after the number of cases starts to grow exponentially (that is t_{min}).

3. *Constraining lockdown effectiveness*: if this option is activated by the user, then she/he can constrain the lockdown effectiveness given by equation (6).

4. *Use extended model to fit trend modification*: if this option is activated by the user, then she/he can choose the number of days used to compute two extra extended models. These parameters are denoted by N_2 and N_3 in the text. If the value of one of these parameters is zero, then no extended model is computed. Using two extended models and the basic model (the first one) we can manage situations where the pandemic outbreak initially starts to growth exponentially and then, due to lockdown measures the growth rate starts to decrease (this is managed by the basic model), then the growth rate changes its trend, because, for instance, the test capacity of the country improves (this can be managed by the first extended model) and finally the evolution stabilizes around a baseline (this can be managed by the second extended model). Many countries have followed these three phases when a strict lockdown has been implemented. We point out that the model we propose is not expected to simulate properly the impact of mild social distancing measures or of a second wave.

5. *Weight in least squares fitting*: the parameter α in equation (14).

6. *The country or uploaded data used* .

6 Forecasting the number of deaths

We can easily extend the model to the case of the evolution of the number of deaths. In this case $y(t)$ represents the number of contaminated subjects who die and $N(t)$ the registered number of deaths. The only thing we have to change is the cumulative distribution $F(t)$. In that case we have to use the infection-to-death time distribution. In [2], the authors model this distribution as the sum of two independent random times, both being Gamma distributed with mean 5.1 days and coefficient of variation 0.86 and 18.8 days and a coefficient of variation 0.45 respectively. The infection-to-death distribution is therefore given by

$$i_{frm} \cdot (Gamma(5.1, 0.86) + Gamma(18.8, 0.45))$$

where i_{frm} is the population averaged over the age structure of a given country. As i_{frm} is a constant factor, we can assume that $i_{frm} = 1$ because in our model, this factor will be compensated by the constant factor C. Therefore, changing, in expression (7), $F(t)$ by the cumulative distribution of the infection-to-death time distribution we can follow the evolution of the number of deaths.

In the same way, assuming that we know the time distribution of other registered values as for instance, the infection-to-hospitalization time distribution, we can forecast, using the same model, the evolution of the corresponding registered value. We point out that the time distribution of the COVID-19 registered values is a topic under investigation and the results can change in the next future. For instance, the study presented in [9] suggests that there are two sub-populations in delays between hospitalization and death: individuals that die quickly upon hospital admission (15% of fatal cases, mean time to death of 0.67 days) and individuals who die after longer time periods (85% of fatal cases, mean time to death of 13.2 days). The combination of Gamma distributions presented above does not reflect this behavior. In the official Spain report [8] using the information of 9765 patients, it is estimated that, in the case of men, the time from the onset of symptoms to death has a median of 11 days with quartiles $Q1 = 7$ and $Q3 = 16$. In the case of women, these values are median$= 10$, $Q1 = 6$ and $Q3 = 14$. Based on the values for men an women, we approximate, experimentally, the distribution of the time from the onset of symptoms to death as a $lognormal(\mu = 2.351375257, \sigma = 0.6011434688)$ distribution. The median of this distribution is $= 10.5$, $Q1 = 7$ and $Q3 = 15.75$. So we can approximate the distribution of the time from infection to death as the following mixtures of lognormal distributions

$$\text{lognormal}(\mu = 1.621, \sigma = 0.418) \; + \; \text{lognormal}(\mu = 2.351375257, \sigma = 0.6011434688), \tag{16}$$

where the first one corresponds to the time infection to the onset of symptoms (see (3)). In Fig. 1, we compare the profile of the distributions using the mixture of Gammas proposed in [2] and the one obtained using the mixture of lognormals (16). We point out that they are quite different. Using the mixture of Gammas, a patient takes considerably more time to die from the infection. In Fig. 2 we compare the forecasts obtained by the proposed model using the infection to death time distribution proposed in [2] and the one obtained using (16). We observe that the forecast of deaths are quite similar but the forecasts of fatally affected subjects are very different. We believe that the one obtained by the lognormals is more plausible because in the other one the number of fatally affected subjects goes to zero too quickly with respect to the evolution of deaths. In the IPOL online demo we use the one obtained by the mixture of lognormals. However, we believe that this approximation is not very accurate either, and as quoted before, we think that the knowledge and accuracy of the time distribution of the basic epidemic factors will be improved in the near future.

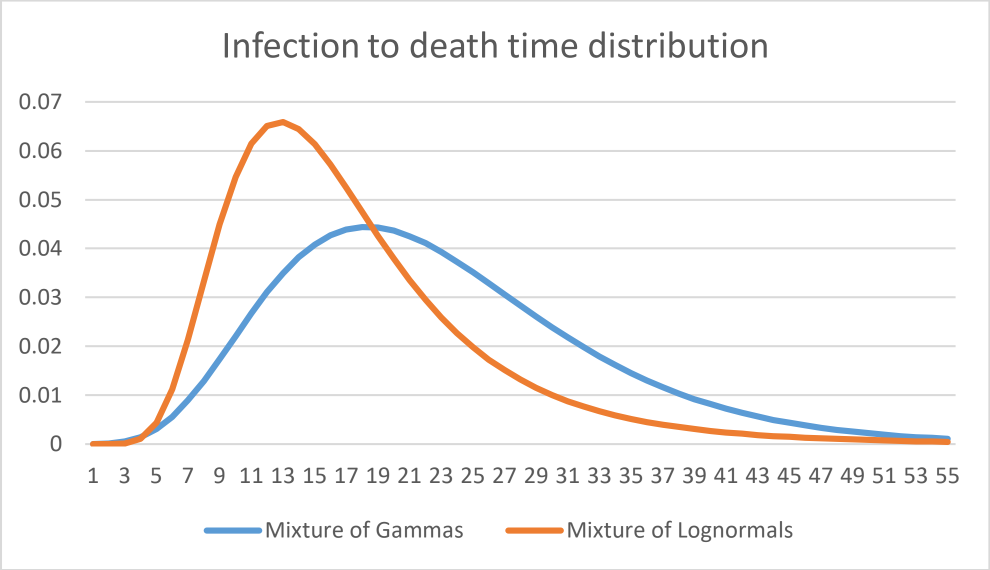


Figure 1: Comparison of the infection to death time distribution using the mixture of Gammas proposed in [2] and the one obtained using the mixture of lognormals (16).

7 Experiments

In the case of the basic model, we are going to focus the attention of our experiments on studying the ability of the model to predict the evolution of the epidemic in the early stages of the epidemic spread in the case a strict lockdown is implemented.

We will use the extended model to study the evolution of the epidemic in the first wave, establishing a timeline of epidemiological events that we will try to follow from the adjustment of the parameters of the extended model to the complete evolution of the epidemic during the first wave.

Therefore, due to its simplicity, we use the basic model to predict the epidemic spread in the early stages when little information is available, and we use the extended model to study the timeline of main epidemiological events in the first wave. Due to its greater complexity, the extended model can better explain the full course of the epidemic during the first wave, which includes an exponential growth phase followed by a decay phase and finally a stabilization around a baseline.

We used the dataset of the evolution of tested positive patients for the different countries from the web page https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide. All the experiments presented can be reproduced using the online DEMO.

7.1 Experiments with the basic model. The ability to predict the epidemic evolution in advance.

To study the ability of the model to predict the evolution of the epidemic in the early stages of the epidemic spread we focus on the study of the model estimation, using the available data up to a given date, of the following epidemiological events:

- The date the peak of new daily cases is reached.

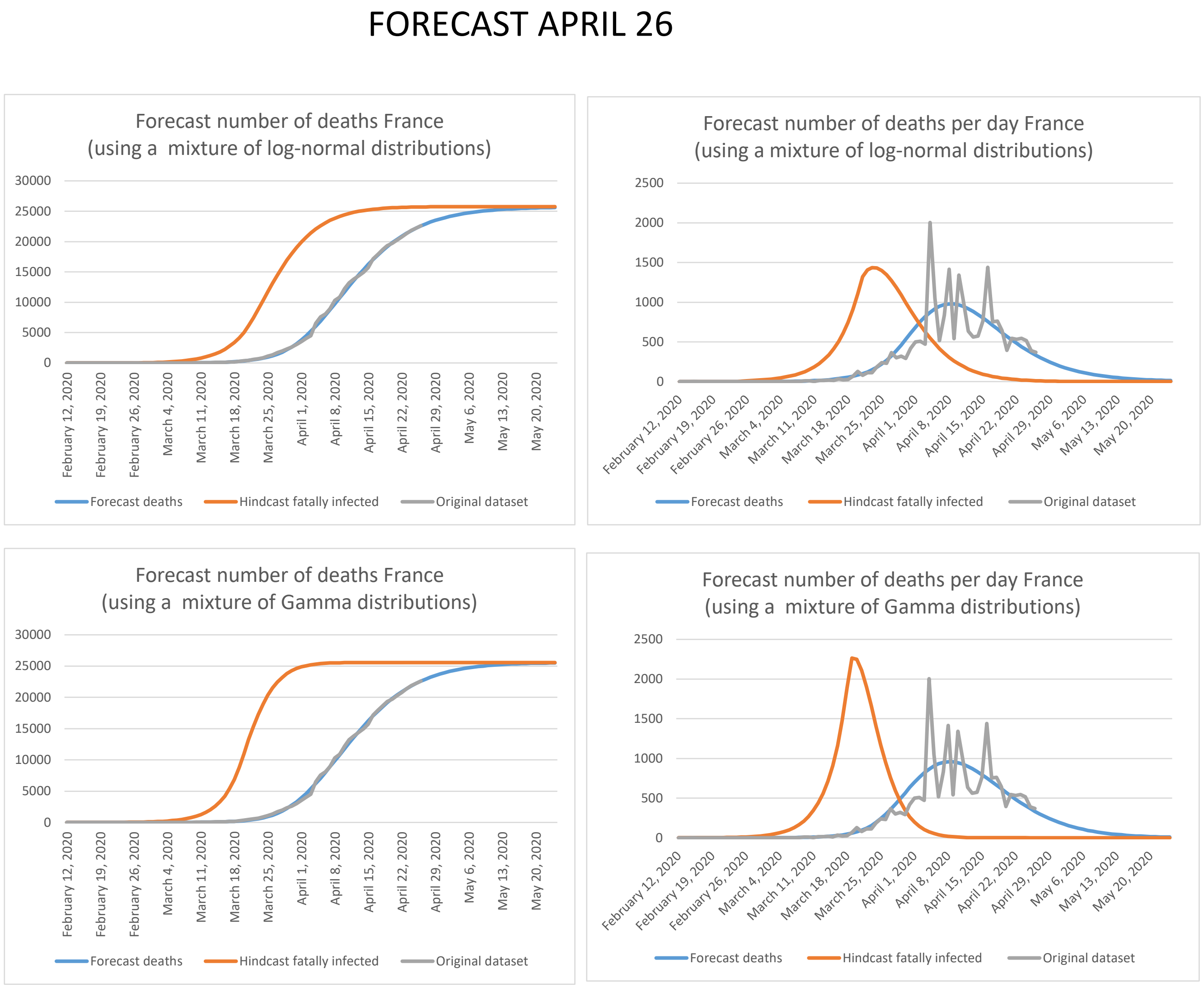


Figure 2: Comparison of the forecasts obtained by the proposed model using the infection to death time distribution proposed in [2] and the one obtained using (16).

- The value of the number of daily cases at the peak.
- The 7-day average of the accumulated number of cases three weeks after reaching the peak.

We are going to use the case of France in this study. The case of France is quite challenging because both the evolution of the number of registered infected and the evolution of the number of deaths show strong fluctuations near the peak of new daily cases. The parameter N_1 indicates the number of days used to compute the model parameters. Therefore, modifying N_1 we obtain the model estimation using the data up to a given date which depends on the value of N_1. In Fig. 3 and 5 we illustrate the basic model for the number of tested positive and deaths and some particular values of N_1. We will compare the results obtained by leaving free the effectiveness of the social distancing measures given by M_{a,γ,t_0,t_1} and setting "a priori" the value of M_{a,γ,t_0,t_1}. In all the experiments presented in this section we use as regularization weight parameter $\alpha = 0$.

In Fig.4 and 6 we show the estimate of the epidemiological events explained above using the available data up to a given date. We point out that a strict lockdown was implemented in France by March 17, the peak of the daily new cases was reached at April 2 (for tested positives) and in April 9 (for deaths). In the case of tested positives, we observe that, without constraining the value of M_{a,γ,t_0,t_1}, we obtain a reasonable estimate of the epidemiological events since April 3 (that is 1 day after reaching the peak). However, if we fix $M_{a,\gamma,t_0,t_1} = 1.5$, we obtain a reasonable estimate since March 24, just 7 days after the lockdown implementation and 9 days before reaching the peak, which means that the model was able to predict the events quite correctly in advance. In fact, the error in the estimation of March 24 was quite small (just 2.92% in the estimation of the 7-day average 3 weeks after the peak).

In the case of deaths, the results are not so good. On April 4, a large number of new deaths were recorded in France, which produced a great disturbance in the results of the model. In fact, the estimate of April 3 is much better than that of April 4, and until April 7 (two days before the peak) it does not begin to provide a reasonable estimate of the events analyzed. In this case setting the value of M_{a,γ,t_0,t_1} does not contribute much and in fact a correct approximation is obtained before using the model without setting the value of M_{a,γ,t_0,t_1}.

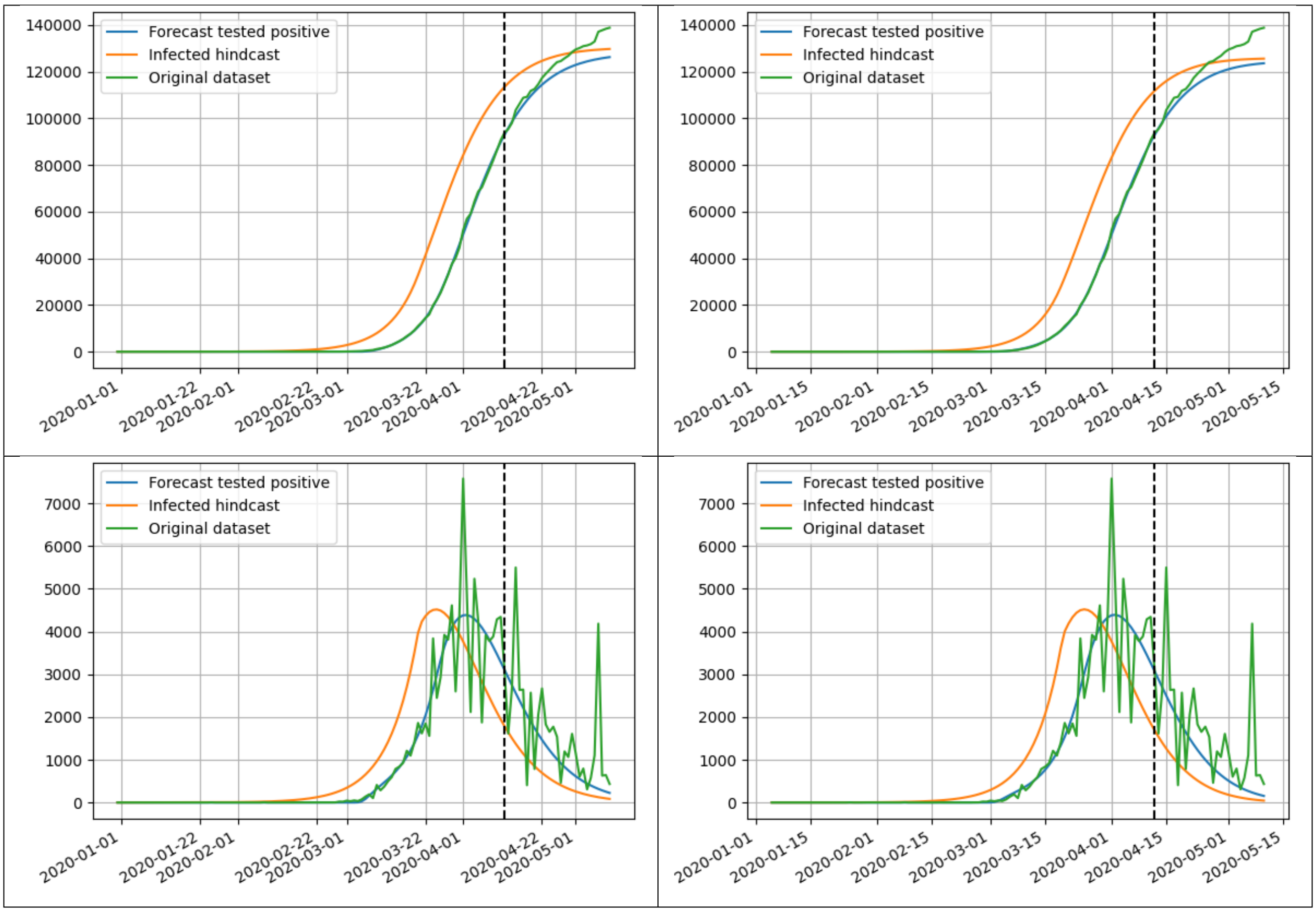


Figure 3: Basic model for tested positive in France using $N_1 = 45$. On the left we present the accumulated and daily cases when the effectiveness is free and on the right when the effectiveness is fixed to $M_{a,\gamma,t_1,t_0} = 1.5$. The algorithm uses the data up to the date represented by the vertical black line.

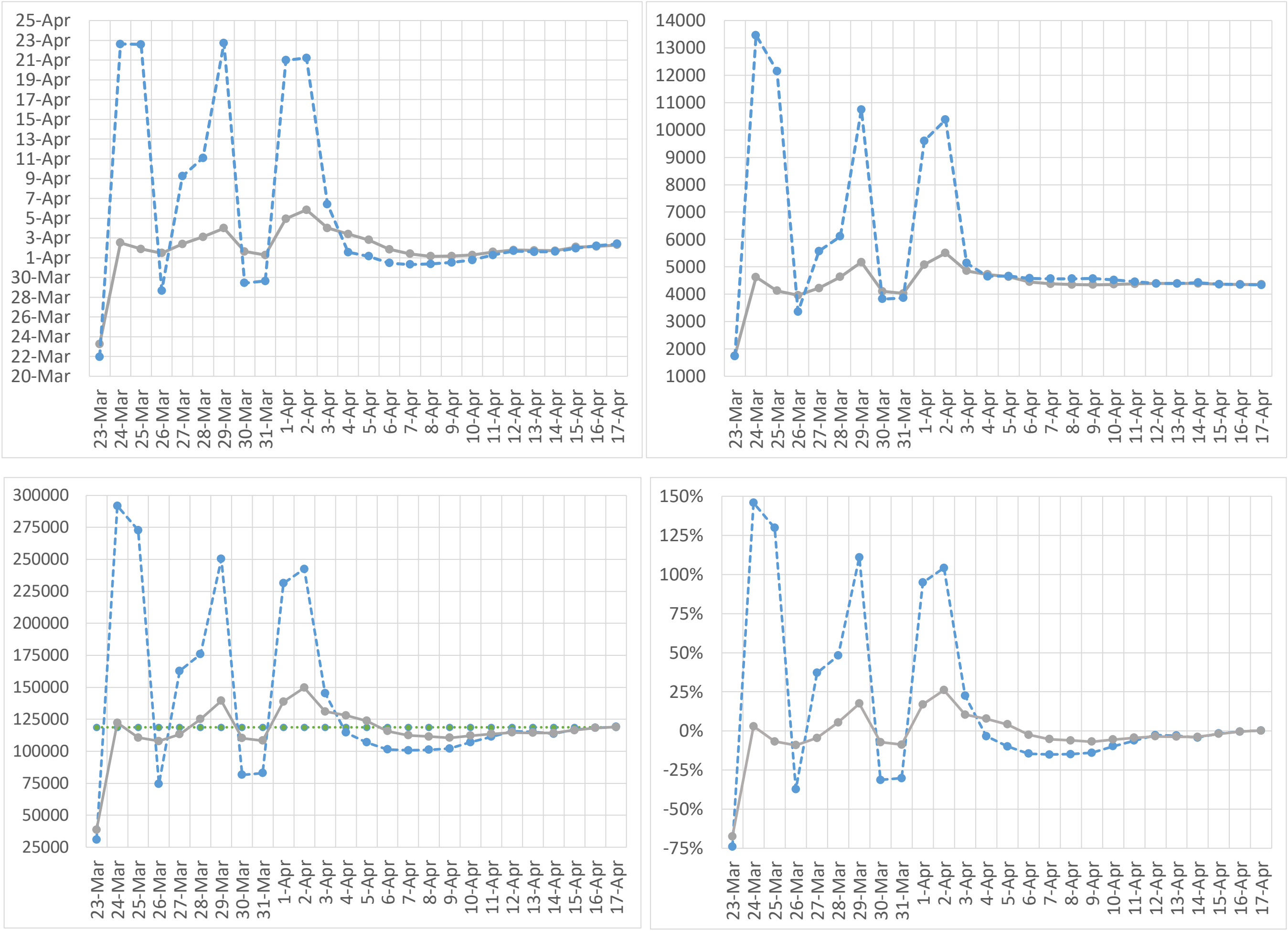


Figure 4: Some epidemiological indicators computed by our model using the registered number of new infected up to the date in the horizontal axis. The solid lines represent the estimate constraining the effectiveness $M_{a,\gamma,t_0,t_1} = 1.5$, the dashed line, the estimate without constraining M_{a,γ,t_0,t_1}, and the dotted line represents the average of the cumulative registered number of infected between April 20 and April 26. From left to right and from top to down, we present: (i) the estimated date of the daily peak of the new tested positive, (ii) the estimated value of the daily number of tested positive in the peak, (iii) the estimate of the average of the cumulative tested positive between April 20 and April 26, and (iv) the error, in percentage, between the estimate and actual value of the average of the cumulative tested positive between April 20 and April 26.

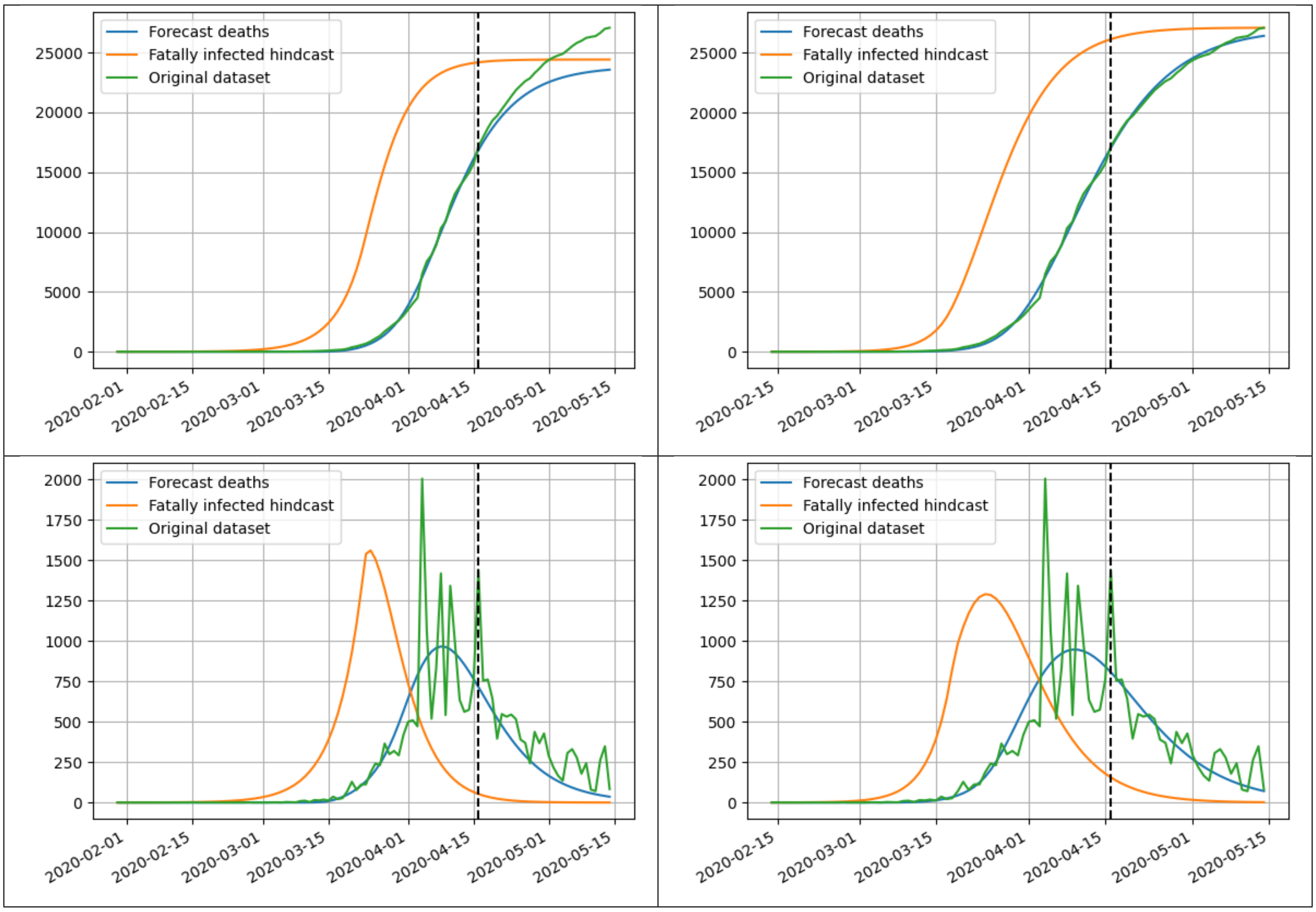


Figure 5: Basic model for deaths in France using $N_1 = 34$. On the left we present the accumulated and daily cases when the effectiveness is free and on the right when the effectiveness is fixed to $M_{a,\gamma,t_1,t_0} = 2$. The algorithm uses the data up to the date represented by the vertical black line.

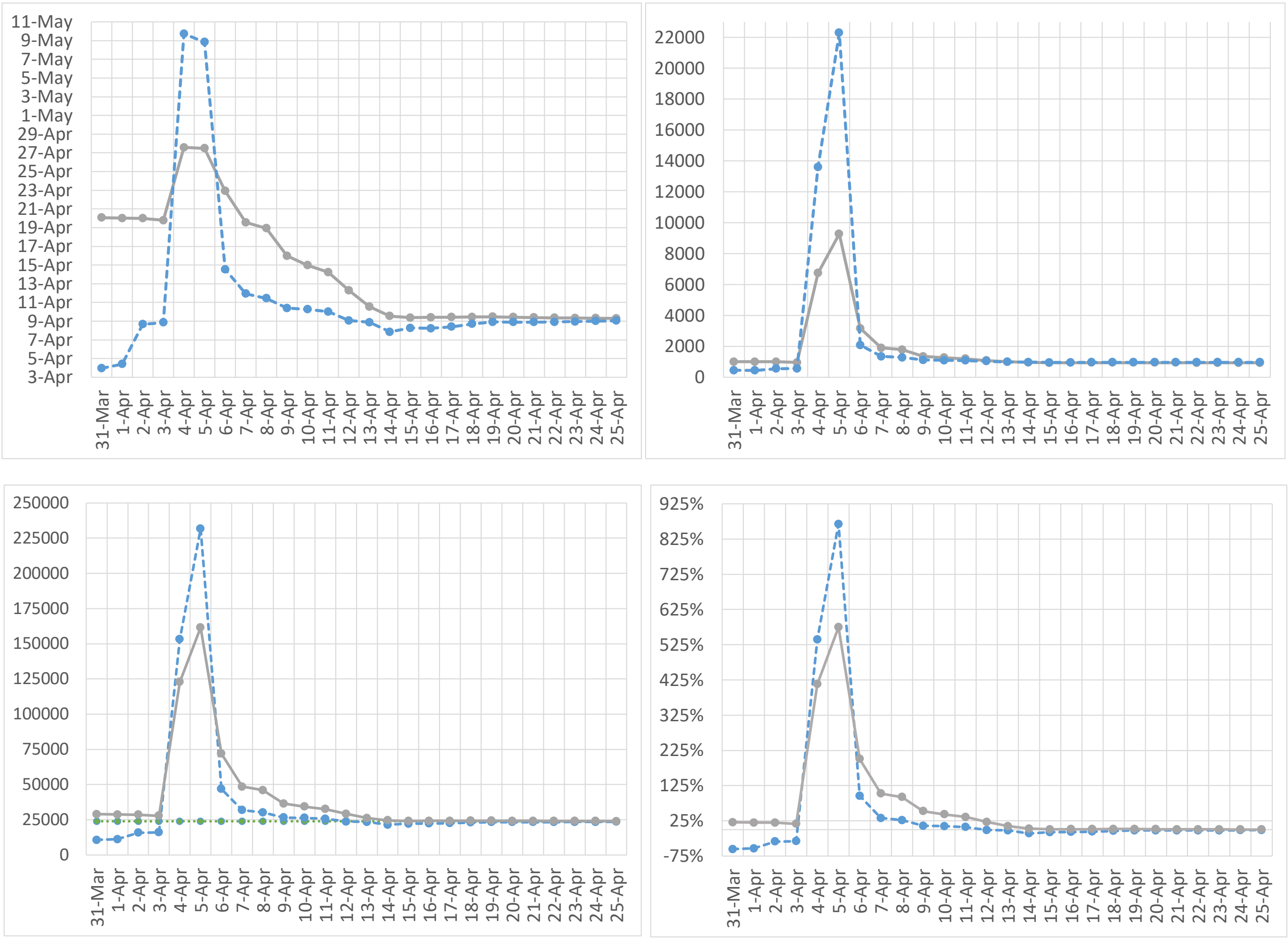


Figure 6: Some epidemiological indicators computed by our model using the registered number of deaths up to the date in the horizontal axis. The solid lines represent the estimate constraining the effectiveness $M_{a,\gamma,t_0,t_1} = 1.5$, the dashed line, the estimate without constraining M_{a,γ,t_0,t_1}, and the dotted line represents the average of the cumulative registered number of deaths between April 27 and May 3. From left to right and from top to down, we present: (i) the estimated date of the daily peak of the new deaths, (ii) the estimated value of the daily number of deaths in the peak, (iii) the estimate of the average of the cumulative registered number of deaths between April 27 and May 3, and (iv) the error, in percentage, between the estimate and actual value of the average of the cumulative registered number of deaths between April 27 and May 3.

7.2 Experiments with the extended model. Understanding the full course of the epidemic during the first wave.

We are going to use the extended model to try to better understand the evolution of the first epidemic wave in South Korea, Italy, Spain, France, United Kingdom, USA and New York state. Using the proposed model, among other things, we are going to estimate the number of days that the virus has been circulating freely before the effect of the social distancing measures take effect, as well as the exponential growth rate of infected in this phase of free circulation. We are going to study the effect of a strict lockdown implemented in the early stages of the epidemic spread, the time it takes to reach the peak of daily cases and the time it takes to divide by two the number of cases reached at the peak. We also compare the results obtained for tested positive and deaths, which provide interesting information on the testing capacity of these countries at the beginning of the epidemic spread.

For each country we manually chose the parameters of the model (see subsection 5.5) to get the best fit between the daily registered data and the model prediction given by $N'(t)$ (see (7)). In Fig. 8, 9, 10 and 11 we illustrate the results obtained and the model parameters for each country. For each country, the value of M_{a,γ,t_1,t_0} (see (11)) only appears if it is manually fixed in the algorithm.

To study the epidemic spread, we will use the timeline (see Fig. 7) given by the following epidemiological events:

- **Outbreak**: Start date of the first epidemic wave. In the case of infected, we consider that the epidemic wave begins when the accumulated number of observable cases reaches 1 subject per 100,000 inhabitants (1 subject per 1,000,000 in the case of deaths). In fact we have two estimates of the outbreak date. The one obtained using the real data-set communicated by the countries (that we name data outbreak) and the one obtained from the model approximation of the observable data given by $N(t)$ (that we name model outbreak). In general, there is little difference between the two estimates in the countries studied. In the case of New York state there are 4.4 days of difference in the case of infected due to the fact that very few cases were detected at the beginning of the epidemic.

- **SDM**: Social Distancing Measures. We will assume that the countries implement social distancing measures to control the epidemic. We pay particular attention to cases where a strict lockdown is implemented at the beginning of the epidemic spread because that allows us to study the impact of a strict lockdown as the first measure to control the epidemic. Except in the case of a strict lockdown, this event does not have a specific date associated with it because it can correspond to a variety of measures taken at different times.

- **FTM**: first trend modification in reaction to **SDM**. We assume, in agreement with our model, that at the beginning of the epidemic the coronavirus was in free circulation and that the number of infected grew exponentially with a constant growth rate, a_1, until the accumulated number of infected reached the value $Ce^{a_1 t_0^1}$ and from that moment, in reaction to the social distancing measures the growth rate began to decrease. We calculate the date when this reaction begins to be observable by the model as the time t such that $N(t) = Ce^{at_0^1} - C$. This is a useful information because it tells us the observable reaction time to the social distancing measures. We consider that the coronavirus was in free circulation from the date of the start of the outbreak (computed using the model) until the first trend modification.

- **PEAK**: date when the maximum of $N'(t)$ is reached in the first wave, which represents the peak of the daily number of observable tested positive or deaths.

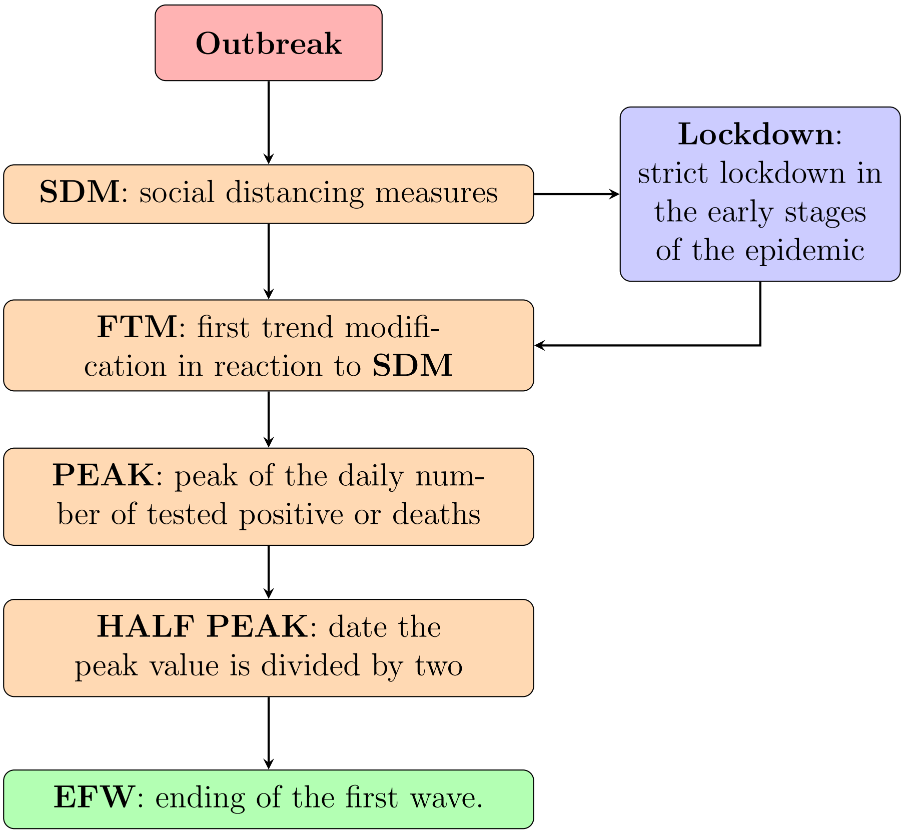

Outbreak
SDM: social distancing measures
Lockdown: strict lockdown in the early stages of the epidemic
FTM: first trend modification in reaction to SDM
PEAK: peak of the daily number of tested positive or deaths
HALF PEAK: date the peak value is divided by two
EFW: ending of the first wave.

Figure 7: Timeline of events we consider in the first wave of the COVID-19 epidemic spread.

- **HALF PEAK**: date after reaching the peak when the number of daily cases at the peak is divided by two.

- **EFW**: ending of the first wave. In the case of infected, we observe that over time, the number of new daily cases stabilizes around a baseline that changes between different countries. We manually set the value of the parameter N_2 so that this stabilization time corresponds to $t_{min} + N_1 + N_2$. We consider this time as the date of the end of the first wave. In the case of the deceased, this stabilization is not easily observed and we set the date of the end of the first wave as the time the number of new daily cases reaches the value of 1 deceased per 1,000,000 inhabitants.

In tables 2 and 3 we present a summary of the dates of the timeline events we obtain for the different countries for tested positive and deaths. Below each date, in brackets, we write the number of days passed since the previous event. Furthermore, for each event we include the value of the daily number of cases predicted by the model for that day (normalized to the country's population) and below that value, in brackets, we write the ratio between this value and the same value in the previous event. This gives us an idea of the growth rate of the number of daily cases between one event and another. In table 4 we compare between the results obtained for the tested positive and deaths. Below we will present a discussion of the most relevant results of these large tables, both globally and country by country.

7.2.1 Analysis country by country

South Korea: Since the initial outbreak, the virus was in free circulation only for 7.38 days, then the model starts to react to the **SDM** and reached the daily peak 1.61 days later and the number of cases at the peak (1.24) was divided by two 6.37 days later. A very good performance in controlling the epidemic spread. South Korea monitored the spread of the epidemic using technological resources, such as tracking credit card use, checking CCTV footage and an efficient centralized contact-tracing using cellular phones. The number of deaths has been very small and has not been studied because it is not statistically significant.

Italy: On 8 March 2020, a strict lockdown was imposed in the Lombardy region and later, on March 11, a strict lockdown was imposed in the whole country. Since the initial outbreak, the virus was in free circulation for 17.31 days (7.16 days after the second lockdown). The peak of daily cases was reached 7.75 days later and the number of cases at the peak (9.526) was divided by two 25.09 days later. The fact that the lockdown in Lombardy was implemented 3 days before the lockdown in the whole country modifies the reaction time with respect to the second lockdown.

Spain: On 14 March 2020, a strict lockdown was imposed in the whole country. Since the initial outbreak, the virus was in free circulation for 16.03 days (8.7 days after the lockdown). The peak of daily cases was reached 5.29 days later and the number of cases at the peak (17.01) was divided by two 13.08 days later. The first really significant measure to control the epidemic that Spain and France implemented was the lockdown. Therefore, these two countries are a good example to study how a strict lockdown affects the free circulation of the virus.

France: On 17 March 2020, a strict lockdown was imposed in the whole country. Since the initial outbreak, the virus was in free circulation for 17 days (9.31 days after the lockdown). The peak of daily cases was reached 6.76 days later and the number of cases at the peak (6.62) was divided by two 16.04 days later.

Germany: Since the initial outbreak, the virus was in free circulation for only 8.43 days. The peak of daily cases was reached 14.59 days later and the number of cases at the peak (6.532) was divided by two 14.70 days later. A good performance in controlling the epidemic spread.

United Kingdom: Since the initial outbreak, the virus was in free circulation for 13.16 days. The peak of daily cases was reached 15.35 days later and the number of cases at the peak (7.23) was divided by two 41 days later.

New York State: Since the initial outbreak, the virus was in free circulation for 15.94 days. The peak of daily cases was reached 14.63 days later and the number of cases at the peak (50.75) was divided by two 19.96 days later.

USA: Since the initial outbreak, the virus was in free circulation for 13.59 days. The peak of daily cases was reached 11.78 days later. In USA each state followed different strategies at different times, globally we cannot say that a first wave has been completed in USA and then, in this case, we do not study the evolution in USA after the first peak.

7.2.2 Global Analysis

At the beginning of the epidemic spread, among the different countries studied, only South Korea and Germany had a testing capacity that allowed to correctly follow the evolution of the epidemic. Indeed, these two countries obtain a similar and high initial exponential growth rate (around 0.25) that can be obtained because of a good testing capacity which is able to track the epidemic spread. Furthermore, in the case of Germany there are other indicators that suggest this fact, such as the

	S. Korea	Italy	France	Germany	UK
Method proposed in this paper	0.2525	0.1359	0.1282	0.2478	0.1684
Method proposed in [7]	0.24	0.18	0.16	0.27	0.17

Table 1: Comparison of the estimation of the initial exponential growth rate computed using our method and the method proposed in [7].

high delay in the number of days between the outbreak of infected and the outbreak of deaths and between the dates at the peaks of daily cases of infected and deaths. The mortality ratio obtained by Germany is also a good sign of the testing capacity. An important key to the success of these two countries has been the anticipation reflected in the small number of days the virus circulated freely before social distancing measures began to take effect.

In the rest of the countries, the testing ability was not sufficient to follow the evolution of the epidemic during the first wave. In short, the numbers of infected was strongly underestimated. In these countries the initial exponential growth rate is lower than 0.2, there is little delay between the outbreaks of infected and deaths and the peaks and in all these countries the mortality ratio is artificially high.

In [7] the authors computed an initial exponential growth rate, (a_1 following our notation), for different countries. In table 1 we show some comparison results. The results are reasonably similar considering that the techniques are quite different. In [7] the calculation was carried out directly on the registered data of infected and we do it on the infected hindcast (notice that we always estimate the infected hindcast) and in [7], the authors use a time interval set manually and we use a time interval calculated automatically when minimizing the quadratic error.

In Italy, Spain and France, who implemented a lockdown in the early stages of the epidemic spread, it took between 8 and 10 days to begin to notice the effect of this measure (a little earlier in Italy due to the lockdown in Lombardy). This is quite a reasonable delay considering the incubation time and the administrative time required to register the cases. Afterwards, it took between 5 and 7 days to reach the peak of daily cases (a little longer in Italy due to the Lombardy effect again). In Germany, UK and New York it took between 14 and 16 days to reach the peak of daily cases since the effect of **SDM** began to be noticed. This suggests that the lockdown accelerated this phase considerably. The later time to divide by two the number of daily cases reached at the peak is highly variable among all countries and it cannot be clearly concluded that the initial lockdown improves results.

The social distance effectiveness measure given by M_{a,γ,t_1,t_0} is very good in South Korea ($M_{a,\gamma,t_1,t_0} = 1$). In countries who implemented a strict lockdown it is about 1.5 (it is a little higher in Italy because of the early Lombardy lockdown) and in the other regions it is higher than 2. Even in Germany the value is quite high (2.9554). The reason is that this measure of effectiveness considers the evolution of the epidemic spread after the **SDM** starts to be noticed. So in the case of Germany the **SDM** starts to be noticed very soon but the effect of the **SDM** was not so clear cut as in the case of a strict lockdown.

In the study of the evolution of the number of deaths, it is highlighted that the initial exponential growth rate of those infected who later die is very similar among all countries except Germany. The reason could be that the number of deaths is independent of the ability to perform tests, but the calculation of the initial exponential growth rate requires, for the deaths, that the virus has been circulating freely for a sufficient number of days, which happens in all countries except Germany and this affects the evolution of the number of deaths. We observe that the initial exponential growth rate for those infected who later died estimated in Italy, Spain, France, UK, New York State and USA are very similar to each other and in turn very similar to the growth rate of those infected who tested positive in South Korea and Germany. The mean of all these values obtained in these countries is 0.2507375 with a standard deviation of only 0.0024433264. This seems to indicate that the initial exponential growth rate when the virus circulates freely is about 0.2507375. One of the key parameters in the initial epidemic dynamics is the doubling time t_D which according with our estimation corresponds to

$$e^{0.2507375 \cdot t_D} = 2 \qquad \rightarrow \qquad t_D = 2.764433643.$$

This estimate is consistent with the result shown in [5] where the authors estimate that $t_D \in [1.86, 2.96]$. Another key parameter of the initial epidemic dynamics is the reproduction number $R0$. There are a variety of techniques to compute $R0$ from the initial exponential growth rate, a_1, and some other epidemiological information like the serial time or the mean recovery time $1/\gamma$. If we assume that the mean recovery time is $1/\gamma \in [7, 14]$ then we have that

$$R0 = a_1 \frac{1}{\gamma} = 0.2507375 \; \frac{1}{\gamma} \in [1.7551625, 3.510325].$$

An evident key to success or failure to control the epidemic in the first wave was the anticipation in decision-making that is reflected in the number of days that the virus had been circulating freely before the social distancing measures began to take effect. According to our calculations, each day that the virus circulated freely, the number of infected was multiplied by a factor of $1.285 = e^{0.2507375}$. This indicates the enormous importance of anticipation when taking social distancing measures.

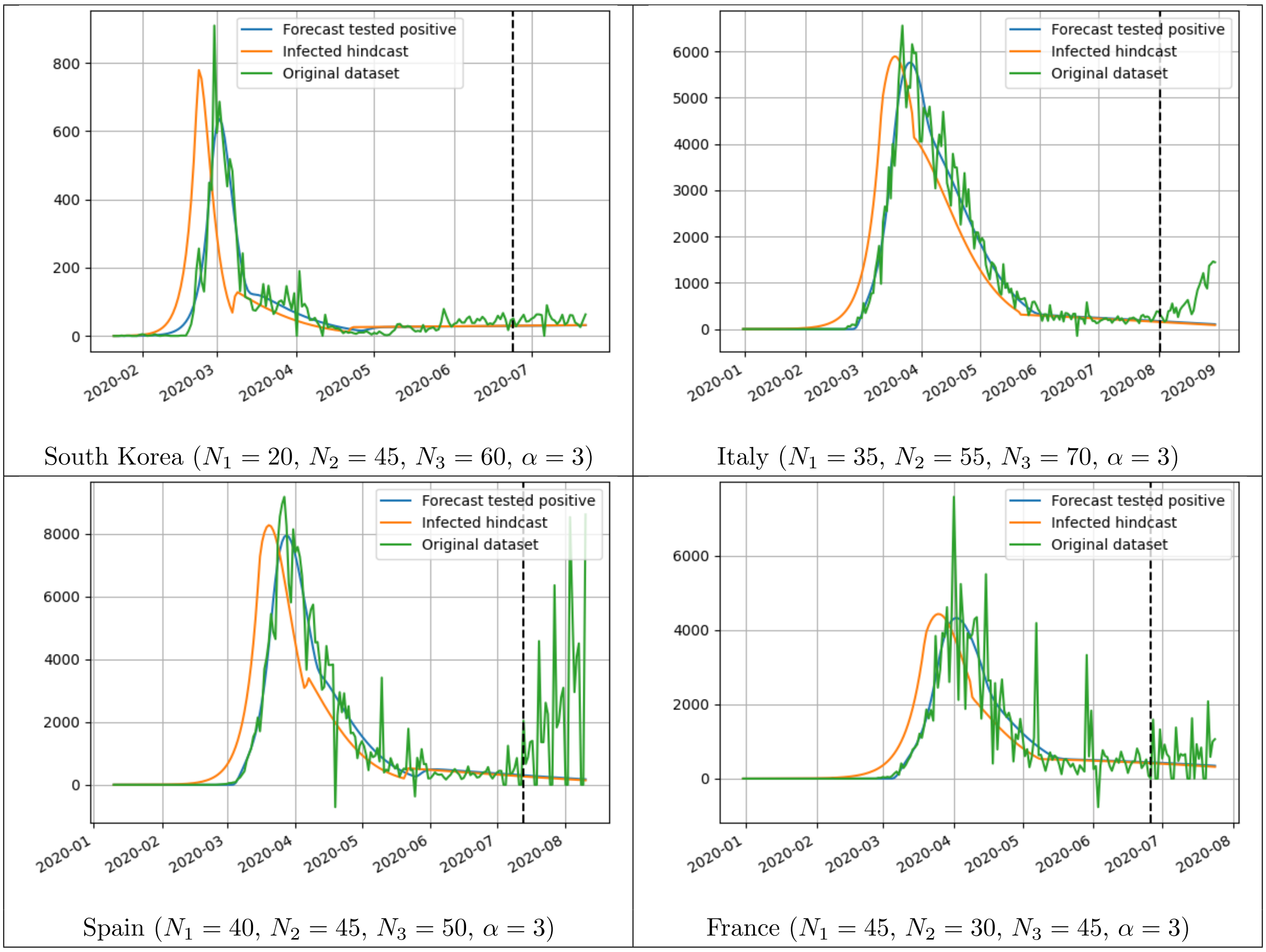
8 Conclusion

The proposed algorithm for the basic model is able to forecast quite well the evolution of the epidemic spread in its early stage when little information is available and strict social distancing measures are implemented. In the case of infected, if we fix manually the value of M_{a,γ,t_0,t_1} using the one obtained for other countries where similar social distancing measures have been imposed, we can anticipate the daily peak and the accumulated number of tested positive 3 weeks after the daily peak.

We have used the extended model to study in detail the timeline of epidemiological events during the full course of the first wave of the epidemic in South Korea, Italy, Spain, France, Germany, United Kingdom, New York state and USA. We obtain that one of the key parameters in the success of early control of the epidemic is the number of days that the coronavirus circulated freely before the social distancing measures began to take effect. In that sense, only South-Korea and Germany successfully anticipated the first wave. The testing capacity of the rest of the countries was not sufficient to correctly follow the growth of the epidemic. On the other hand, from the analysis of exponential growth in the early stage of the epidemic, we have obtained that the exponential growth rate in this phase of the epidemic is around 0.2507375. This determines the famous $R0$, the reproduction number of the coronavirus.

A critical reader will have noticed that contrarily to the SIR models, our model based on the $r(t)$ law is empirical. This is justified by two facts that we have stressed:

a) Given the huge observation noise it is better to work with a very low dimensional model, so that we estimate a very few empirical parameters, rather than the many that cannot actually be estimated;



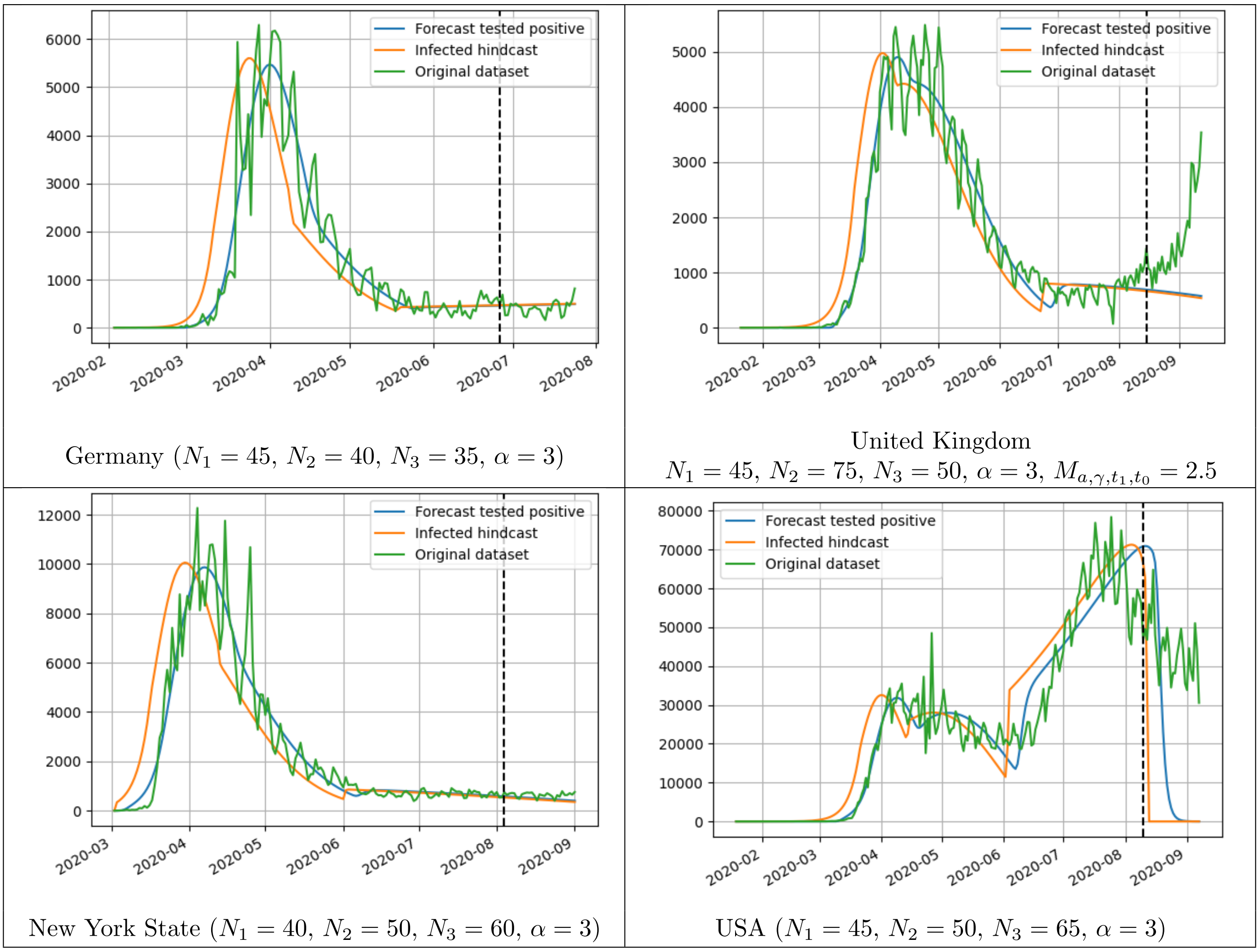
South Korea ($N_1 = 20$, $N_2 = 45$, $N_3 = 60$, $\alpha = 3$)

Italy ($N_1 = 35$, $N_2 = 55$, $N_3 = 70$, $\alpha = 3$)

Spain ($N_1 = 40$, $N_2 = 45$, $N_3 = 50$, $\alpha = 3$)

France ($N_1 = 45$, $N_2 = 30$, $N_3 = 45$, $\alpha = 3$)

Figure 8: Extended model for tested positive applied to South Korea, Italy, Spain and France. Below the plot of each country we show the parameters of the model manually chosen to get the best fit between the daily registered data and the model prediction.



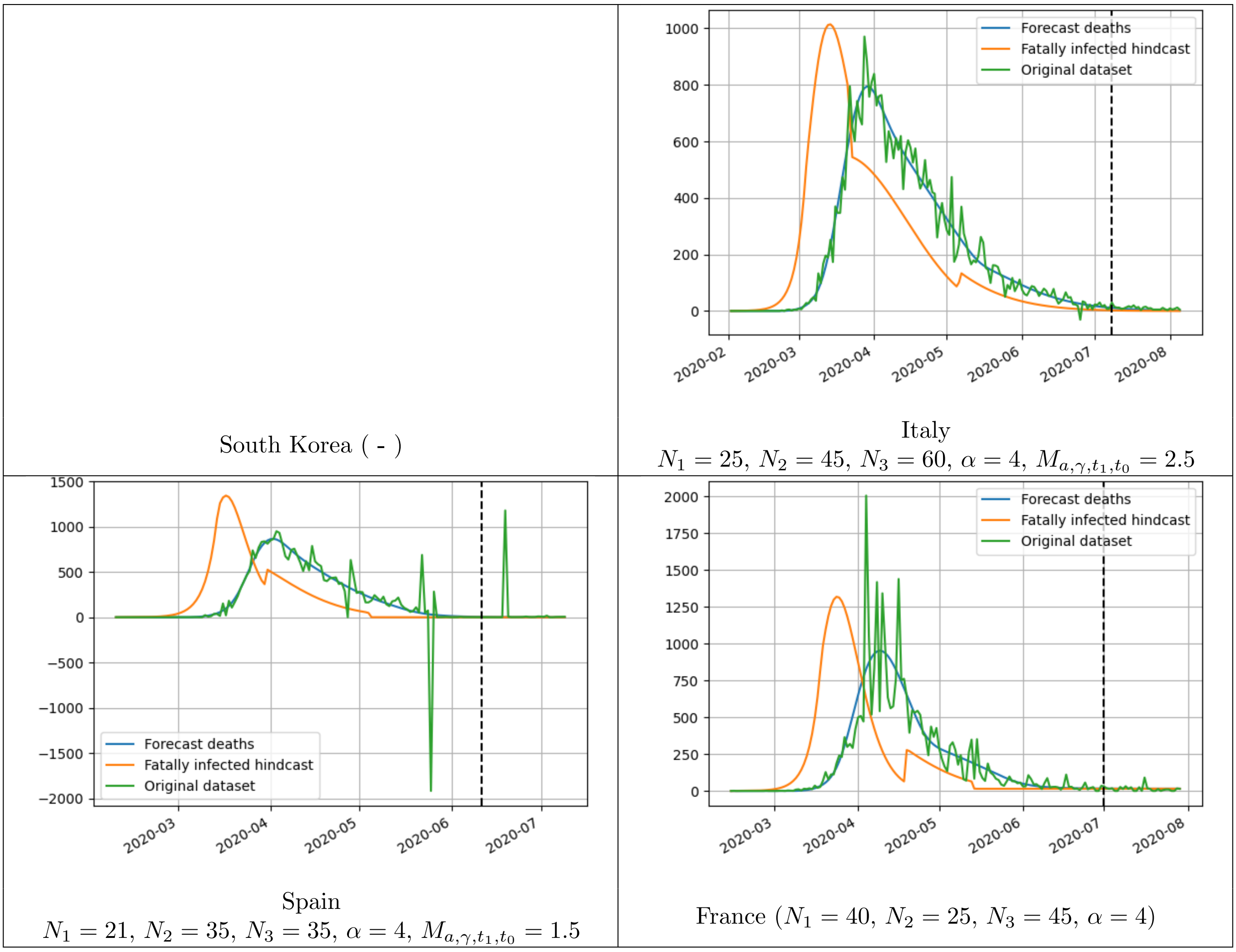
Germany ($N_1 = 45,\ N_2 = 40,\ N_3 = 35,\ \alpha = 3$)

United Kingdom
$N_1 = 45,\ N_2 = 75,\ N_3 = 50,\ \alpha = 3,\ M_{a,\gamma,t_1,t_0} = 2.5$

New York State ($N_1 = 40,\ N_2 = 50,\ N_3 = 60,\ \alpha = 3$)

USA ($N_1 = 45,\ N_2 = 50,\ N_3 = 65,\ \alpha = 3$)

Figure 9: Extended model for tested positive applied to Germany, United Kingdom, New York State and USA. Below the plot of each country we show the parameters of the model manually chosen to get the best fit between the daily registered data and the model prediction. The value of M_{a,γ,t_1,t_0} (see (11)) only appears if it is manually fixed in the algorithm.



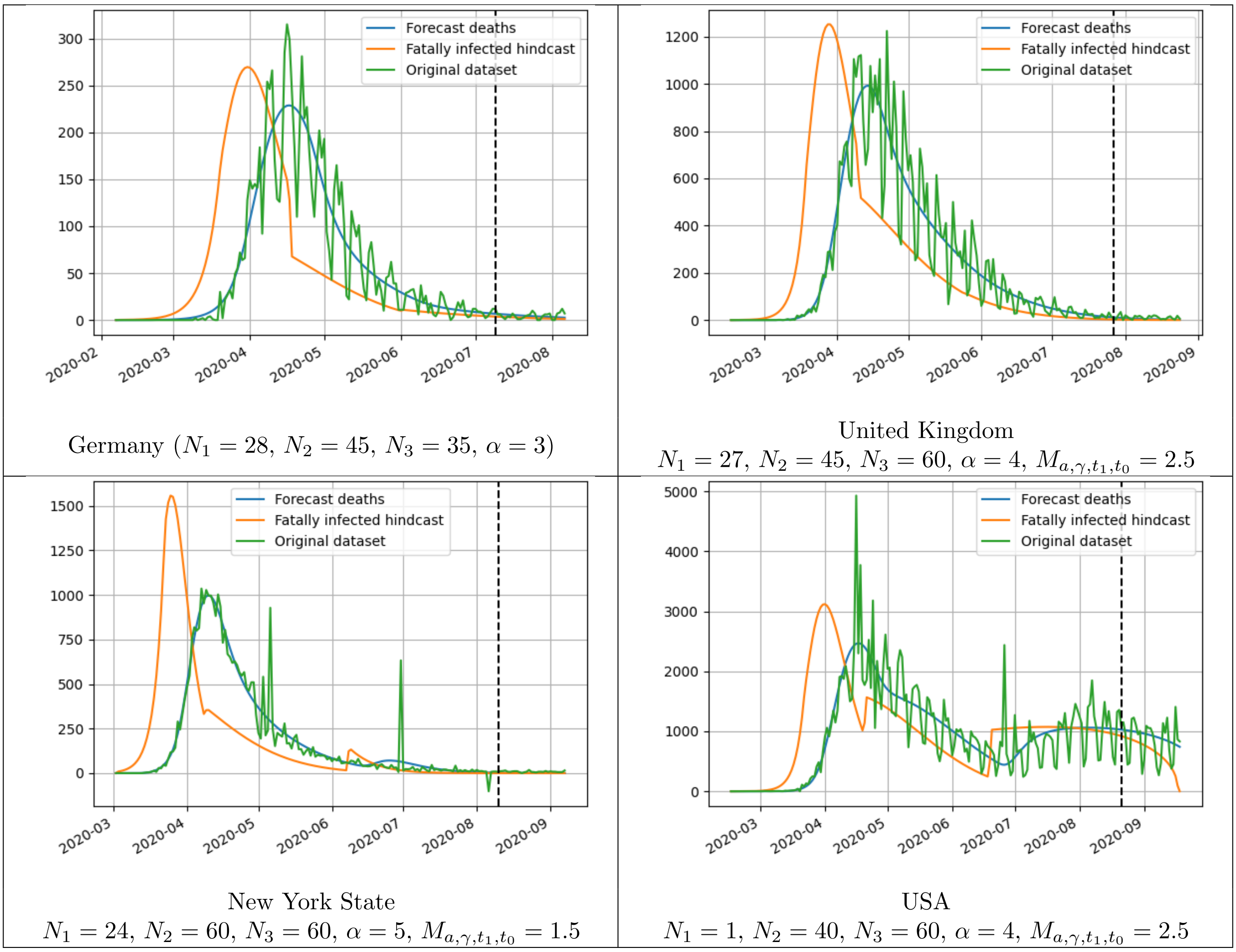
South Korea (-)

Italy
$N_1 = 25,\ N_2 = 45,\ N_3 = 60,\ \alpha = 4,\ M_{a,\gamma,t_1,t_0} = 2.5$

Spain
$N_1 = 21,\ N_2 = 35,\ N_3 = 35,\ \alpha = 4,\ M_{a,\gamma,t_1,t_0} = 1.5$

France ($N_1 = 40,\ N_2 = 25,\ N_3 = 45,\ \alpha = 4$)

Figure 10: Extended model for deaths applied to Italy, Spain and France. Below the plot of each country we show the parameters of the model manually chosen to get the best fit between the daily registered data and the model prediction. The value of M_{a,γ,t_1,t_0} (see (11)) only appears if it is manually fixed in the algorithm. We do not include the case of South Korea because the number of deaths were too small to be significant from an statistical point of view.



Germany ($N_1 = 28,\ N_2 = 45,\ N_3 = 35,\ \alpha = 3$)

United Kingdom
$N_1 = 27,\ N_2 = 45,\ N_3 = 60,\ \alpha = 4,\ M_{a,\gamma,t_1,t_0} = 2.5$

New York State
$N_1 = 24,\ N_2 = 60,\ N_3 = 60,\ \alpha = 5,\ M_{a,\gamma,t_1,t_0} = 1.5$

USA
$N_1 = 1,\ N_2 = 40,\ N_3 = 60,\ \alpha = 4,\ M_{a,\gamma,t_1,t_0} = 2.5$

Figure 11: Extended model for deaths applied to Germany, United Kingdom, New York State and USA. Below the plot of each country we show the parameters of the model manually chosen to get the best fit between the daily registered data and the model prediction. The value of M_{a,γ,t_1,t_0} (see (11)) only appears if it is manually fixed in the algorithm.

TESTED POSITIVE	S. Korea	Italy	Spain	France	Germany	UK	New York	USA
Model Outbreak	Feb-21	Feb-29	Mar-6	Mar-9	Mar-9	Mar-12	Mar-9	Mar-14
Data Outbreak	Feb-22 (+0.666)	Feb-27 (-2.04)	Mar-5 (-1.02)	Mar-7 (-1.92)	Mar-7 (-1.11)	Mar-11 (-1.23)	Mar-12 (+3.3)	Mar-15 (+1.27)
Data Outbreak: daily value per 100,000/h	0.267	0.144	0.295	0.174	0.204	0.251	1.818	0.513
Lockdown	-	Mar-11 (+12.18)	Mar-14 (+8.35)	Mar-17 (+9.61)	-	-	-	-
Lockdown: daily value per 100.000/h		2.89 (x 20.05)	3.51 (x 11.91)	1.7 (x 9.75)				
initial exponential growth rate	0.2525	0.1359	0.1645	0.1282	0.2478	0.1684	0.1949	0.1869
FTM: first trend modification	Feb-29 (+6.72)	Mar-18 (+7.16)	Mar-22 (+8.70)	Mar-26 (+9.31)	Mar-17 (+9.54)	Mar-25 (+14.39)	Mar-25 (+12.65)	Mar-27 (+12.31)
FTM: daily value per 100.000/h	1.15 (x 4.31)	7.181 (x 2.49)	13.58 (x 3.86)	5.27 (x 3.10)	2.062 (x 10.11)	3.31 (x 13.18)	21.34 (x 11.74)	5.03 (x 9.81)
PEAK: date	Mar-1 (+1.61)	Mar-25 (+7.75)	Mar-27 (+5.29)	Apr-2 (+6.76)	Apr-1 (+14.59)	Apr-9 (+15.35)	Apr-8 (+14.63)	Apr-8 (+11.78)
PEAK: daily value per 100.000/h	1.24 (x 1.08)	9.526 (x 1.33)	17.01 (x 1.25)	6.62 (x 1.26)	6.532 (x 3.17)	7.23 (x 2.18)	50.75 (x 2.38)	9.626 (x 1.91)
HALF PEAK: date	Mar-8 (+6.37	Apr-20 (+25.09)	Apr-10 (+13.08)	Apr-18 (+16.04)	Apr-15 (+14.70)	May-5 (+41.02)	Apr-28 (+19.96)	-
EFW: date end first wave	Apr-25 (+47.65)	May-24 (+34.00)	May-24 (+43.93)	May-12 (+23.89)	May-22 (+36.23)	Jun-26 (+36.00)	Jun-7 (+39.25)	-
EFW: daily value per 1,000,000/h	0.031 (x 0.05)	0.853 (x 0.18)	0.614 (x 0.07)	1.04 (x 0.31)	0.531 (x 0.16)	0.571 (x 0.16)	3.198 (x 0.13)	-
Measure Efectiveness	1.0000	1.62846	1.4927	1.4972	2.9554	2.5	2.6237	2.2872

Table 2: Timeline of the epidemic spread and epidemiological indicators for tested positive in different countries.

DEATHS	S. Korea	Italy	Spain	France	Germany	UK	New York	USA
Model Outbreak	-	Mar-3	Mar-10	Mar-16	Mar-18	Mar-17	Mar-18	Mar-23
Data Outbreak	-	Mar-3 (-0.07)	Mar-11 (+1.17)	Mar-13 (-3.09)	Mar-22 (+4.13)	Mar-17 (-0.76)	Mar-18 (+0.09)	Mar-21 (-1.31)
Data Outbreak: daily value per 1.000.000/h		0.29	0.513	0.11	0.312	0.24	0.446	0.188
Lockdown	-	Mar-11 (+7.70)	Mar-14 (+2.02)	Mar-17 (+3.76)	-	-	-	-
Lockdown: daily value per 1.000.000/h		2.06 (x 7.19)	0.92 (x 1.80)	0.38 (x 3.50)				
initial exponential growth rate	-	0.2498	0.2517	0.2486	0.1673	0.2497	0.2560	0.2498
FTM: first trend modification	-	Mar-16 (+5.94)	Mar-27 (+13.76)	Mar-31 (+14.52)	Apr-02 (+11.35)	Mar-31 (+14.61)	Apr-5 (+18.87)	Apr-4 (+13.10)
FTM: daily value per 1,000,000/h	-	6.173 (x 2.99)	15.44 (x 16.72)	9.45 (x 24.54)	1.56 (x 4.98)	6.83 (x 28.45)	42.70 (x 95.74)	3.49 (x 18.54)
PEAK: date	-	Mar-29 (+12.36)	Apr-1 (+5.06)	Apr-9 (+8.58)	Apr-16 (+13.83)	Apr-13 (+12.03)	Apr-10 (+5.03)	Apr-17 (+13.08)
PEAK: daily value per 1,000,000/h	-	13.142 (x 2.13)	18.52 (x 1.20)	14.58 (x 1.54)	2.73 (x 1.76)	14.62 (x 2.14)	51.325 (x 1.20)	7.459 (x 2.14)
HALF PEAK: date	-	Apr-25 (+27.11)	Apr-21 (+19.77)	Apr-23 (+13.97)	May-4 (+17.30)	May-4 (+20.79)	Apr-25 (+14.91)	-
EFW: ending first wave	-	Jun-9 (+45.25)	May-21 (29.96)	May-28 (+35.80)	May-9 (+5.22)	June 24 (+50.73)	Jul-18 (+83.69)	-
Measure Efectiveness	-	2.5	1.5	1.9297	2.1677	2.5	1.5	2.5

Table 3: Timeline of the epidemic spread and epidemiological indicators for deaths in different countries.

	S. Korea	Italy	Spain	France	Germany	UK	New York	USA
Lag between data outbreaks	-	4.48	6.33	5.85	14.68	5.93	5.59	6.45
Lag between peaks	-	3.39	4.83	7.03	15.73	3.83	2.22	8.53
Fatality rate ratio peak values	-	13.80%	10.89%	22.03%	4.18%	20.17%	10.11%	7.75%

Table 4: Comparison between the spread of tested positive subjects and the spread of of deaths in different countries. We compare the lag between the start of the outbreaks, the lag between the peaks and the fatality rate computed as the ratio between the values in the peaks.

b) The virtue of such an empirical model is that it may cope not only with noise, but also with a variation of the very definition of observed variables. This variation definitely happens. Indeed, the various administrations are progressively changing the way they make their statistics about the observed cases. They also adapt their testing policy, and ultimately they also adjust their containment policy. Thus, an adapted parametric approach to the prediction might be adequate to overcome all these limitations.

References

[1] Loli Piccolomiini E. and Zama F, *Monitoring italian covid-19 spread by an adaptive seird model.* preprint medRxiv, DOI: 10.1101/2020.04.03.20049734, 2020.

[2] Seth Flaxman et al., *Estimating the number of infections and the impact of nonpharmaceutical interventions on covid-19 in 11 european countries.* Imperial College COVID-19 Response Team, https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-Europe-estimates-and-NPI-impact-30-03-2020.pdf, 2020.

[3] Chowell G., Hengartner NW., Castillo-Chavez C., Fenimore PW., and Hyman JM., *The basic reproductive number of ebola and the effects of public health measures: the cases of congo and uganda*, Journal of Theoretical Biology, 229 (2004), pp. 119–126.

[4] Josselin Garnier, *Quantification d'incertitudes bayesienne pour les modeles de propagation d'epidemie de type covid19.* Presentation, GdT Maths4covid19, Laboratoire Jacques-Louis Lions, 2020.

[5] Marek Kochańczyk, Frederic Grabowski, and Tomasz Lipniacki, *Super-spreading events initiated the exponential growth phase of covid-19 with 0 higher than initially estimated*, R. Soc. open, sci.7200786 (2020).

[6] Stephen A. Lauer, PhD * MS, BA * Kyra H. Grantz, MHS Qifang Bi, Forrest K. Jones, MPH, MHS Qulu Zheng, PhD Hannah R. Meredith, Andrew S. Azman, PhD, Nicholas G. Reich PhD, and PhD Justin Lessler, *The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application*, Annals of Internal Medicine, (2020).

[7] Z. Liu, P. Magal, and G. Webb, *Predicting the number of reported and unreported cases for the covid-19 epidemics in china, south korea, italy, france, germany and united kingdom*, Journal of Theoretical Biology, 509 (2021), p. 110501.

[8] Spain Red Nacional de Vigilancia Epidemiológica, *Informe sobre la situación de covid-19 en españa. informe n. 25. 23 de abril de 2020.* Official report, 2020.

[9] Henrik Salje, Cécile Tran Kiem, Noémie Lefrancq, Noémie Courtejoie, and Paolo Bosetti et al., *Estimating the burden of sars-cov-2 in france.* HAL Id: pasteur-02548181, 2020.