On an interval prediction of COVID-19 development based on a SEIR epidemic model

Denis Efimov, Rosane Ushirobira

▶ **To cite this version:**

Denis Efimov, Rosane Ushirobira. On an interval prediction of COVID-19 development based on a SEIR epidemic model. [Research Report] Inria. 2020. hal-02517866v6

HAL Id: hal-02517866
https://inria.hal.science/hal-02517866v6

Submitted on 3 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On an interval prediction of COVID-19 development based on a SEIR epidemic model

Denis Efimov[1] and Rosane Ushirobira[1]

***Abstract*—In this report, a revised version of the well-known mathematical outbreak SEIR model is used to analyze the epidemic's course of COVID-19 in eight different countries. The proposed model enhancements reflect the societal feedback on pandemic and confinement features. The parameters of the SEIR model are identified by using publicly available data for France, Italy, Spain, Germany, Brazil, Russia, New York State (US), and China. The identified model is then applied for the prediction of the SARS-CoV-2 virus propagation under different conditions of confinement. For this purpose, an interval predictor is designed allowing variations and uncertainties in the model parameters to be taken into account. The code and the utilized data are available in Github.**

I. Introduction

The SEIR model is one of the most elementary compartmental models of epidemics [1]. It is very popular and widely used in different contexts [2]. This model describes the evolution of the relative proportions of four classes of individuals in a population of constant size, see a general scheme given in Fig. 1. Namely, the susceptible individuals S, capable of contracting the disease and becoming infectious; the asymptomatic E and symptomatic I infectious, capable of transmitting the disease to susceptible; and the recovered R, permanently immune after healing or dying (if the number of deaths is of special interest, then an additional compartment D can be introduced). Such a simple model represents well a generic behavior of epidemics (plainly as a series of transitions between these populations), and a related advantage consists in a small number of parameters to be identified (three transition rates in Fig. 1: σ, γ and b). This is an important outcome in the case of a virus attack with a limited amount of data available. That is the case of the current worldwide situation[1] under the presence of the SARS-CoV-2 virus.

There are many types and variations of SEIR models [1] (*e.g.*, in the simplest case the classes E and I are modeled at once, leading to a SIR model). A specificity of COVID-19 pandemics is the global confinement imposed by the majority countries, which influences the virus dynamics [3]. In the literature, numerous suggestions are answering the question of how to reflect the confinement characteristics in the mathematical models [4], [5], [6].

A. Considered model

This note is based on a modified SEIR discrete-time model proposed in [7], where it has been used to model the trend of the epidemic of COVID-19 in China (other similar SIR/SEIR-type models used recently for modeling SARS-CoV-2 virus are [8], [9], [10], [11]). The model we adopt in this work is as follows (we do not consider the influence of the natural birth and mortality, since for the short period of analysis considered here the population may be assumed quasi-constant):

$$
\begin{aligned}
S_{t+1} &= S_t - b\frac{\left(p_{t-\tau_p}I_t + r_{t-\tau_r}E_t\right)}{N}S_t, && \text{(1a)}\\
E_{t+1} &= (1-\sigma-\sigma')E_t + b\frac{\left(p_{t-\tau_p}I_t + r_{t-\tau_r}E_t\right)}{N}S_t, && \text{(1b)}\\
I_{t+1} &= (1-\gamma-\mu)I_t + \sigma E_t, && \text{(1c)}\\
R_{t+1} &= R_t + \gamma I_t + \sigma' E_t, && \text{(1d)}\\
D_{t+1} &= D_t + \mu I_t, && \text{(1e)}
\end{aligned}
$$

where $t \in \mathbb{N}$ (the set of non-negative integers) is the time counted in days ($t = 0$ corresponds to the beginning of measurements or prediction), $N = S+E+I+R+D$ denotes the total population, the parameter $0 < \gamma < +\infty$ represents the recovery rate, $0 < \mu < +\infty$ is the mortality rate, the parameter $0 < b < +\infty$ corresponds to the rate of the virus transmission from infectious/exposed to susceptible individuals during a contact, $0 < \sigma, \sigma' < +\infty$ are the incubation rates at which the exposed develop symptoms or directly become recovered without a viral indication, $0 \le p_t < +\infty$ corresponds to the number of contacts for the infectious I (it is supposed that infected people with symptoms are in quarantine, then the number of contacts is decreased), $p_t \le r_t < +\infty$ is the number of contacts per person per day for the exposed population E (in the presence of confinement and depending on its severity, this number is time-varying), and $\tau_p, \tau_r > 0$ are the delays in the reactions of the compartments on variations of quarantine conditions (we assume that if $t < \tau_p$ or $t < \tau_r$, then $p_{t-\tau_p} = p_0$ or $r_{t-\tau_r} = r_0$, respectively). Compared to the model in [7], the inflow/outflow variables from/to other regions for each state are not considered in our analysis.

Remark 1. In the versions $1-4$ of this report (they can be found in HAL), the compartment D was not explicitly considered in the model (it was a part of R) and we used $\sigma' = 0$ and $\tau_p = 0$.

B. Societal feedback and confinement influence in the model

To consider the reaction of society on confinement and virus propagation, we introduced the delays τ_p and τ_r in the seclusion inputs p_t and r_t, respectively.

The idea behind τ_r is that after the activation of the quarantine, several days pass before changes in the disease

[1]Denis Efimov and Rosane Ushirobira are with Inria, Univ. Lille, CNRS UMR 9189 - CRIStAL, F-59000 Lille, France, Denis.Efimov, Rosane.Ushirobira@inria.fr

[1]This note was initially written on March 23 and updated June 3, 2020.

propagation become detectable (such an effect can be easily observed in the data for all analyzed countries). Roughly speaking, the increase in the number of infected individuals E and I is predefined by the number of contacts in the previous days, when the confinement was not yet imposed, for example.

We assume that during the phase of active lockdown, $r_{t-\tau_r} = p_{t-\tau_p}$ always holds, *i.e.*, the number of contacts for asymptomatic E and symptomatic I infected populations is the same (it corresponds to the case of the society that follows well the government requirements).

The delay τ_p is used to model the clustering effect of the confinement: under restrictions on displacement activities, people are constrained to stay in their district and visit a limited number of attractions (shops, pharmacies, hospitals, *etc*.). So the population can be considered to be divided into smaller groups, where after some time the chances to meet an infected person start to decay (*e.g*., in such a group, there is no infected person, or the individual was isolated, or the whole group can be infected, but anyway the virus propagation is basically stopped).

Remark 2. Another way of inclusion of societal feedback on virus development is the substitution:

$$b \longrightarrow \frac{b}{1+\eta I_t},$$

where $0 < \eta < +\infty$ is a tuning parameter. In this case, we model the effect of natural augmentation of confinement strictness, when society becomes aware of the problem following the increased number of infected or dead people (I implicitly represent them, or it can also be explicitly replaced with D). To this end, we decrease the virus transmission rate b with the growth of the number of infected/dead individuals. This variant has been tested, but we prefer to use the delays τ_p and τ_r since the parameter identification is simpler in this case.

C. Model parameters

Therefore, the SEIR model (1) has seven parameters to be identified or assigned: σ, σ', τ_p, τ_r, γ, μ and b.

1) Generic observations: The parameters σ, γ, μ and b represent, respectively, the rate of changes between the states E to I, I to R, I to D and S to $E+I$ (as in Fig. 1). The parameters σ and σ' have a physical meaning: $\sigma = \frac{1}{T_S}$ and $\sigma' = \frac{\kappa}{T_S}$, where T_s is the average duration of the virus incubation period after contamination, which can be well-identified in patients, and $\kappa \in [0,1)$ is the ratio of recovering period for the patients with the mild form of COVID-19, which can be also found in sufferers. Similarly, the delays τ_r and τ_p, which are also of order T_s, have a natural origin. The numbers of contacts in p_t and r_t (with or without (relaxed) confinement) can be evaluated heuristically based on the population density and social practices (for prediction, different their profiles can be selected for testing).

2) Known or accepted quantities: The incubation period T_s, that is widely reported in the literature for COVID-19 studies, is considered to be between 2 and 14 days [7], or in a more specialized research between 2 and 12 days [12], so we assume

$$\frac{1}{12} \leq \sigma \leq \frac{1}{2}.$$

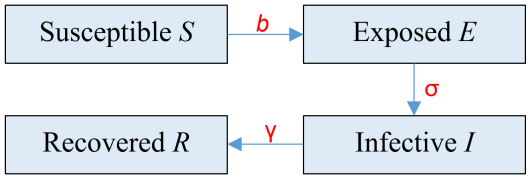

Fig. 1. A schematic representation of SEIR model

It also implies that the delays can be selected in the corresponding limits:

$$2 \leq \tau_r \leq 12,\ \tau_r + 2 \leq \tau_p,$$

where the condition $\tau_p > \tau_r$ entails that the clustering starts to be important after the effect of confinement becomes significant (adding an incubation period).

The numbers of contacts have to be selected separately for each country. For example, we may take the values of [7] and make some reduction related with a smaller population density in the considered countries:

$$\begin{aligned} p_Q &= 3 \text{ (number of contacts in quarantine)},\\ p_N &= 15 \text{ (number of contacts in normal mode)},\\ p_R &= 10 \text{ (number of contacts in relaxed quarantine)},\\ p_C &= 0.1 \text{ (number of contacts under clustering)}. \end{aligned}$$

Then the input $p_t \in \{p_Q, p_C\}$ and $r_t \in \{p_Q, p_N, p_R, p_C\}$ for all $t \in \mathbb{N}$.

The identification of the model parameters must be performed using statistics published by authorities[2]. As a worthy remark, many researches devoted to the estimation and identification of SIR/SEIR models were developed by now, and several in the last few years, like [13], [14], [15], [16], [17], to mention a few.

D. Uncertainty and prediction

Since the measured data and parameters contain numerous uncertainties and perturbations, it is then difficult to carry out a reasonable prediction based on the simulation of such a model with fixed values of parameters (also taking into account the model simplicity and generality). However, the interval predictor and observer framework [18], [19], [20], [21], [22] allows a set of trajectories to be obtained corresponding to the interval values of parameters and inputs, increasing the model validity without augmenting its complexity. This approach has already been applied to different SEIR models (see, *e.g*., [23], [24], [25]). In this report, we apply the interval predictor method for the considered SEIR model (1) to improve its forecasting quality.

Remark 3. It is important to highlight that the interval predictor framework used here is not the only method oriented on improving the reliability of prediction applying SEIR models. Usually, as in [7], [8], [9], [10], [11], [26], stochastic and multi-agent simulation approaches are used. In those cases,

[2]As in Imperial College Report, for example.

by assuming that the parameters and initial conditions are distributed with some *given* probability, multiple numeric experiments are performed to reconstruct the behavior of possible trajectories of the system. As a first remark, such a methodology needs more computational effort for its realization. Secondly, additional information on the form of the probability distribution for all parameters and variables is necessary, demanding either extra hypotheses or more measured data for estimation. As it is currently demonstrated by the SARS-CoV-2 virus attack, it is difficult to obtain such data quickly during the epidemic development. Contrarily to these approaches, the interval predictor method does not use these additional assumptions on probability distributions, and it has been also proposed to estimate a *guaranteed* interval inclusion of trajectories with minimal computational effort, by the cost of a more complicated mathematical analysis and design [22].

E. Structure of the report

The outline of this note is as follows. The model validation and identification are first presented considering France as the study case, next the same results are reported for Italy, Spain, Germany, Brazil, Russia, New York State (US), and China. To this end, we describe the measured data applied for parameter identification in Section II, together with some hypotheses used in the sequel. In the versions $1-4$ of this report (check them in HAL), the parameters obtained in [7] for China were tested for France and validated on the data, which demonstrated the need for identification and model adaptation for each country. The method for parameter identification is presented in Section III, with validation and some experiments on the influence of confinement parameters on COVID-19 development. An interval predictor is designed in Section IV, which allows us to evaluate the situation in France under the variation of parameters and initial states. The accuracy of interval prediction is also evaluated by using a part of data for identification and another part for verification. The results of the application of the proposed identification routine and the interval predictor for other countries are given in Section V. Final discussions and remarks are provided in Section VI.

II. Available data on COVID-19 in France

The current population in France is $N = 67064000$[3].

A. Used data-sets

The data available from public sources[4], and which is used in this note, for the time between March 12th and 2nd June is provided in Github[5], where $\mathscr{I}$, $\mathscr{D}$, and $\mathscr{R}$ represent the number of total detected infected, deceased and recovered individuals, respectively.

Not all cases can be detected and documented by public health services, so there is a ratio between populations I and $\mathscr{I}$, R and $\mathscr{R}$, D and $\mathscr{D}$, which is denoted in this work by α. The interval of admissible values for α is estimated from different sources as follows[6]:

$$5 \leq \alpha \leq 25.$$

Formally, such a ratio α has to be time-varying and different for I, D and R. Due to strict and similar requirements of health services in almost all considered countries, in this report we take the following hypotheses:

$$I_t = \alpha_1(\mathscr{I}_t - \mathscr{D}_t - \mathscr{R}_t),\ R_t = \alpha_2\mathscr{R}_t,\ D_t = \alpha_3\mathscr{D}_t, \qquad (2)$$

i.e., the number of active infected cases and the related recovered individuals can be masked due to the complexity of examination and the actual confirmation of the virus presence, while availability or not of post-mortem tests can influence the number of registered passing. An additional reason is that in many cases the virus symptoms result in a mild reaction of patients (approximately 80% of cases, see the sources above), hence maybe with no official virus confirmation in such a situation. In this work, we assume modest values for these parameters:

$$\alpha_2 = \alpha_1,\ \alpha_3 = 1,$$

then, roughly speaking, such a choice corresponds to the registration of deaths exactly (see also [10]) with the same error for recovered and infected individuals (the exclusion was made only for US). A technique to identify α_1 from the measurements of $\mathscr{I}$, $\mathscr{R}$ and $\mathscr{D}$ is described by CMMID (see the last footnote) giving for France

$$\alpha_1 = 16.7.$$

Remark 4. A way to determine α_3 is given in https://github.com/sebastianhohmann.

So, by fixing α_1, α_2, and α_3, the three variables of the model (1), I, D, and R, are available from the beginning of the epidemics via (2).

B. Fixed values of parameters

For parameters, we select an average value for the incubation rate:

$$\sigma = \frac{1}{7}$$

to simplify further identification (the variation in this value can be taken into account later in the interval predictor), then

$$\sigma' = \kappa\sigma,\ \kappa = 0.1,$$

and we assume that there is a very slow transfer from exposed E to recovered R directly without symptom exposition. The values of the delays are chosen as

$$\tau_r = 5,\ \tau_p = \tau_r + 20,$$

and the algorithm for their identification is discussed below. The procedure for the identification of γ, μ, and b is also given in the next section.

[3] www.en.wikipedia.org/wiki/Demographics_of_France.

[4] See www.santepubliquefrance.fr/, www.geodes.santepubliquefrance.fr and www.data.gouv.fr.

[5] Sources: Eficiens, Johns Hopkins University.

[6] See, for example, Coronalinks-3-19-20, or a dedicated analysis in Imperial College of London, [27], CMMID or University of Melbourne.

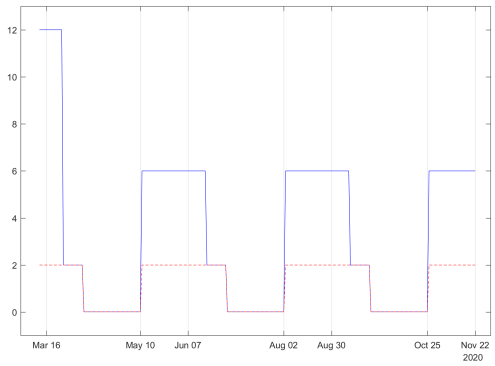

Fig. 2. Variation of the number of contacts $p_{t-\tau_p}$ and $r_{t-\tau_r}$

C. Scenario of confinement

In [8], the theory of a cyclic application of quarantine regimes of different severity is evaluated for COVID-19. By iterating the periods of complete isolation for everybody (*suppression*), which decelerates the virus advancement, with a time of mild regulation (*mitigation*), which allows the economy balance to be maintained on an arguable level, and when only fragile parts of the population are isolated, it is possible to attenuate the material consequences of epidemics while decreasing the load on health services. Following this idea, for simulation, we consider a cyclic scenario of confinement (*e.g.*, with 8 weeks of strict quarantine and 4 weeks of a relaxed one), which is further periodically repeated. For the chosen model, this scenario impact only the input variables $p_{t-\tau_p}$ and $r_{t-\tau_r}$, an example of their behavior is shown in Fig. 2 (by red dash and blue solid lines, respectively).

Remark 5. In other words, r_t and p_t can be considered as a sort of control for the virus propagation, by imposing different periods and strictness levels for the confinement for compartments I and E. A more detailed analysis may also take into account the age or the geographic distribution.

III. Parameter identification

For the parameter identification, we assume that the incubation rates σ and σ' are fixed as above, and that the symptomatic infectious I_t, the dead D_t, and the recovered R_t persons are measured for the first $J > 0$ days of the virus attack as in (2) for $t = 0, 1, \dots, J$.

In this section, we begin by discussing approaches to the identification of the delays τ_p and τ_r. Then, the method for the identification of the mortality rate μ, the recovery rate γ and the infection rate b is presented. Finally, the model (1) with the obtained values for the parameters is validated in simulations.

A. Delay identification

We propose two approaches for the estimation of τ_p and τ_r.

1) Method 1: From the dynamics of (1b), the increment of E_t (*i.e.*, $E_{t+1} - E_t$) is directly proportional to $p_{t-\tau_p}$ and $r_{t-\tau_r}$. The number of contacts $r_{t-\tau_r}$ instantaneously changes its value after the imposition of the quarantine (it jumps from p_N to p_Q). Since $\tau_p > \tau_r$ and $p_{t-\tau_p} = r_{t-\tau_r}$ in confinement, the signals $p_{t-\tau_p}$ and $r_{t-\tau_r}$ jump next from p_Q to p_C, and the same occurs after the suppression of the confinement (from p_C to p_Q or p_R), see Fig. 2. It implies that the increment of E_t shows discontinuities in these time instants. The variable E_t is not available for measurements, but the same (filtered) behavior is also observed in the increment of the variable I_t. Since both variables, I_t and E_t in (1) have an exponential rate of changes, then the signal

$$dI_t = \ln(I_t) - \ln(I_{t-1})$$

for $t = 2, \dots, J$ should have a step-like form (the logarithm of the increment of an exponentially growing or decaying signal is a constant) with the change of value in the time instant $t_c \geq 2$, when a modification of the confinement rules starts to influence the variable I_t. Therefore, the delay can be estimated as (with a mild ambiguity in this work we use the same symbol to denote a parameter and its estimate)

$$\tau_r = t_c - t',$$

where $t' \geq 0$ is the instant of application of the new confinement rule. Hence, to estimate the value t_c the following algorithm is proposed:

$$t_c = \underset{t=3,\dots,J}{\arg\min} \sqrt{\sum_{\ell=2}^{J} |dI_\ell - dI_\ell^t|^2},$$

where

$$dI_\ell^t = \begin{cases} \frac{1}{t-2} \sum_{s=2}^{t-1} dI_s & ,\text{if } \ell < t \\ \frac{1}{J-t+1} \sum_{s=t}^{J} dI_s & ,\text{if } \ell \geq t \end{cases}$$

is a step-like varying signal, which jumps at the instant t. The main drawback of this approach is the noise in the measurements (as for any approach that indirectly use a derivative estimation).

Remark 6. Note that if the values of γ and μ are known (see below how we can estimate them), then using (1c) the variable $E_t = \frac{1}{\sigma}(I_{t+1} - (1-\gamma-\mu)I_t)$ can be reconstructed from the measurements, and the same approach can be applied to the increment $dE_t = \ln E_t - \ln E_{t-1}$, which explicitly depends on $p_{t-\tau_p}$ and $r_{t-\tau_r}$. Unfortunately, we have a very noisy data for COVID-19, so the calculated variables E_t contain a lot of perturbations, and the above (derivative-based) approach does not provide a reliable estimation using dE_t.

2) Method 2: This method also uses the estimated values of E_t (see below for the detailed description), but it does not use (approximated) derivatives. The idea of this approach is based on the observation that $\ln E_t$ (the variable E_t is exponentially growing) can be approximated by a straight line for any constant values of $p_{t-\tau_p}$ and $r_{t-\tau_r}$:

$$\ln E_t = at + b,$$

for some $a, b \in \mathbb{R}$. Such an approximation filters the noise contrarily the derivative-based method presented in the previous subsection. Then the initial phase of the epidemics can be decomposed on three intervals of time:

$$\mathscr{T}_1 = [0, T_1 + \tau_r),\ \mathscr{T}_2 = [T_1 + \tau_r, T_1 + \tau_p),\ \mathscr{T}_3 = [T_1 + \tau_p, T_2],$$

where T_1 is the day of confinement activation, T_2 is the day of commutation to the relaxed quarantine, and on each interval

$$\ln(E_t) = a_i t + b_i,$$

for $t \in \mathscr{T}_i$ and some coefficients $a_i, b_i \in \mathbb{R}$ with $i = 1,2,3$, is a reliable approximation. The coefficients a_i, b_i can be calculated using the Least Square Method (LSM), or any other approach of solving this system of linear equations with known reconstructed values of E_t. Next, we can calculate the instants of these lines intersection:

$$\tau_r = \frac{a_2 b_1 - a_1 b_2}{a_2 - a_1} - T_1,\ \tau_p = \frac{a_3 b_2 - a_2 b_3}{a_3 - a_2} - T_1.$$

Note that the intervals $\mathscr{T}_i$, $i = 1,2,3$ are unknown (their definitions depend on the values of τ_p and τ_r), then we can introduce two tuning parameters $Z \in (0, \tau_r)$ and $J_Z \in (0, J)$ such that

$$\hat{\mathscr{T}}_1 = [0, Z),\ \hat{\mathscr{T}}_2 = [J_Z - Z, J_Z),\ \hat{\mathscr{T}}_3 = [J - Z, J]$$

are the estimates for $\mathscr{T}_1$, $\mathscr{T}_2$ and $\mathscr{T}_3$, respectively, which are utilized for calculation of a_i, b_i. Actually, these auxiliary parameters can be rather easily selected having the plot of $\ln(E_t)$ in sight.

This method provides rather good guesses for τ_p and τ_r, as we demonstrate at the end of this section. In general, due to the noise, these estimates are very sensitive to the data.

B. Rates identification

From the equation (1e), we can identify the value of the mortality rate μ:

$$\mu = \frac{D_{t+1} - D_t}{I_t},$$

whose LSM estimation is

$$\mu_k = \frac{\sum_{t=0}^{J-k-1} I_t (D_{t+1} - D_t)}{\sum_{t=0}^{J-k-1} I_t^2}$$

for $k = 0, 1, \ldots, K$, where $0 < K < J - 1$ is the number of the last days used for identification (in this work we selected $K = J - 10$). Another possible approach is the moving window estimation:

$$\mu_k = \frac{\sum_{t=k}^{k+K_w} I_t (D_{t+1} - D_t)}{\sum_{t=k}^{k+K_w} I_t^2}$$

for $k = 0, 1, \ldots, K$ with $K = J - K_w$, where $K_w > 1$ is the window length. Then the average value is used for further analysis and design:

$$\mu = \frac{1}{K+1} \sum_{k=0}^{K} \mu_k = 5.3345 \times 10^{-4}.$$

Since $\sigma' = \kappa\sigma$, multiplying the equation (1c) by κ and subtracting it from (1d), we can identify the value of the parameter γ:

$$\gamma = \frac{R_{t+1} - R_t - \kappa I_{t+1} + \kappa(1-\mu) I_t}{(1+\kappa) I_t},$$

whose LSM estimation is

$$\gamma_k = \frac{\sum_{t=0}^{J-k-1} I_t (R_{t+1} - R_t - \kappa I_{t+1} + \kappa(1-\mu) I_t)}{(1+\kappa) \sum_{t=0}^{J-k-1} I_t^2}$$

for $k = 0, 1, \ldots, K$, or the moving window estimation:

$$\gamma_k = \frac{\sum_{t=k}^{k+K_w} I_t (R_{t+1} - R_t - \kappa I_{t+1} + \kappa(1-\mu) I_t)}{(1+\kappa) \sum_{t=k}^{k+K_w} I_t^2}$$

for $k = 0, 1, \ldots, K$ with $K = J - K_w$. As for μ the average value is used for further analysis and design:

$$\gamma = \frac{1}{K+1} \sum_{k=0}^{K} \gamma_k = 0.0184.$$

Next, the sum of equations (1c) and (1c) allows us to calculate the related number of asymptomatic infectious (σ and σ' are chosen, μ is estimated):

$$E_t = \frac{1}{\sigma + \sigma'} \left(I_{t+1} - (1-\mu) I_t + R_{t+1} - R_t \right),$$

while the number of susceptible individuals can be evaluated using the total population:

$$S_t = N - I_t - R_t - E_t - D_t.$$

At this point, having derived quantities E_t we can estimate the delays τ_r and τ_p using one of the methods presented above. Finally, from the equation (1b) we can derive the infection rate (for the selected values p_Q, p_N, σ and σ'):

$$b = N \frac{E_{t+1} - (1 - \sigma - \sigma') E_t}{\left(p_{t-\tau_p} I_t + r_{t-\tau_r} E_t \right) S_t},$$

whose LSM estimation is

$$b_k = N \frac{\sum_{t=0}^{J-k-1} \left(p_{t-\tau_p} I_t + r_{t-\tau_r} E_t \right) \left(E_{t+1} - (1 - \sigma - \sigma') E_t \right) S_t}{\sum_{t=0}^{J-k-1} \left(p_{t-\tau_p} I_t + r_{t-\tau_r} E_t \right)^2 S_t^2}$$

for $k = 0, 1, \ldots, K$, or the moving window estimation version:

$$b_k = N \frac{\sum_{t=k}^{k+K_w-1} \left(p_{t-\tau_p} I_t + r_{t-\tau_r} E_t \right) \left(E_{t+1} - (1 - \sigma - \sigma') E_t \right) S_t}{\sum_{t=k}^{k+K_w-1} \left(p_{t-\tau_p} I_t + r_{t-\tau_r} E_t \right)^2 S_t^2}$$

for $k = 0, 1, \ldots, K$ with $K = J - K_w$, then the identified value is again the average of these estimates:

$$b = \frac{1}{K+1} \sum_{k=0}^{K} b_k = 0.0918.$$

Remark 7. Due to measurement noise, the derived values of E_t, γ_k and b_k can be negative (that is physically impossible), then a previous positive estimate can be taken into account, *i.e.*, $E_t = E_{t-1}$, or only positive quantities for average calculation can be used: $b = \frac{1}{K+1} \sum_{k=0}^{K} \rho_k b_k$ with $\rho_k = 0.5(\mathrm{sign}(b_k) + 1)$ (it is 0 for negative b_k and 1 otherwise).

C. Results of identification

The obtained values γ_k, b_k and μ_k (solid lines) together with the selected average estimates γ, b and μ (dot lines), and the signal $\ln(E_t)$ (solid line) with approximations $a_i t + b_i$ (dash lines) are shown in Fig. (3). As we can conclude from these results, the identification of the value of γ is rather reliable and converging; the mortality rate μ follows the gravity of the outbreak (it was maximal during the most difficult period of virus propagation at the beginning of April); the value of b is more complicated to estimate since it depends on all

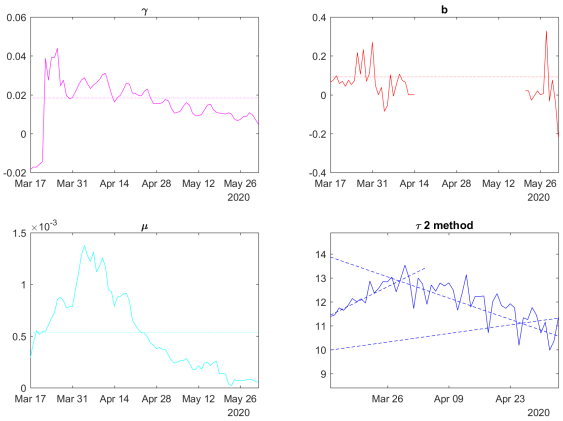

Fig. 3. The identified parameters for France

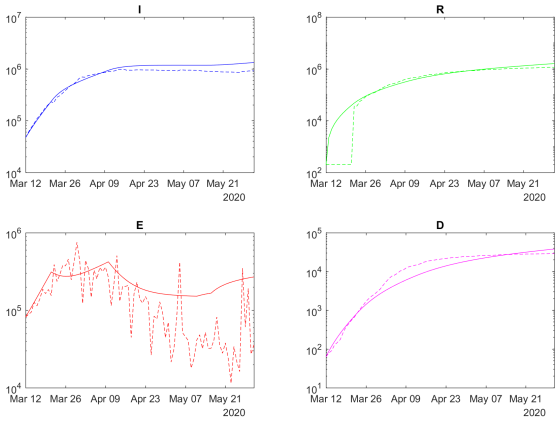

Fig. 5. The results of verification with identified parameters

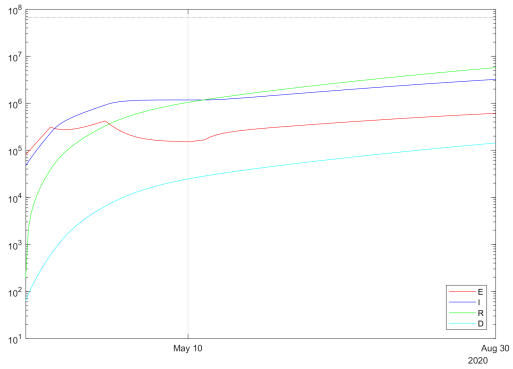

Fig. 4. The results of simulation with identified parameters

quantities (we are stopping the identification if $p_{t-\tau_p}$ and $r_{t-\tau_r}$ are sufficiently small to avoid very noisy results, see the empty space in the middle); finally, the delay τ_r is noticeable from the plot, but the delay τ_p is still hard to detect at the current stage (we can use a nominal value in such a case).

D. Simulation and validation

The simulation results of the model (1) with the identified values of parameters are given in Fig. 4 (for better visibility, all populations are plotted in the logarithmic scale in this report), a zoomed comparison of the measured and reconstructed data is shown in Fig. 5 (as we can see the measured data for I, R and D has a smooth shape, while the reconstructed variable E, which was also used for identification, is rather noisy). In this case, the model can approximate reasonably well the virus propagation, since the identified parameters are consistent with the available statistics for France.

Unfortunately, the obtained curves also demonstrate the lack of efficiency of the confinement, the number of asymptomatic infectious can be reduced quickly, but symptomatic patients may persist a long time giving rise to a second wave. Of course, this conclusion can be related to a probable weak validity of the model for the decreasing phase of the outbreak.

Let us enlarge the validity of the prediction based on (1) by considering intervals of admissible values for the parameters and initial conditions.

IV. Interval prediction

In the previous section, we achieved a prediction for the model (1) with identified values of the parameters $b, \gamma, \mu, \tau_p, \tau_r$ and selected values for $\alpha_1, \alpha_2, \alpha_3, \sigma, \sigma'$. The initial values for the state of the model S_0, I_0, E_0, D_0, and R_0 were chosen from measured/reconstructed sets. However, due to the generic structure of the model, uncertainties in the values of the auxiliary parameters, and noises in the measured information, it is obvious that the reliability of the obtained prognosis is limited. A way to overcome such a weakness consists in taking into account intervals of admissible values for all variables and parameters used for simulation, hence enlarging the validity of the model. In such a case, we calculate/evaluate the sets of the resulted trajectories. In this work, we use for this purpose the interval prediction framework [22].

A. Explanation of idea

The idea of the interval prediction can be illustrated on a simple scalar system:

$$x_{t+1} = a_t x_t + d_t,$$

where $x_t \in \mathbb{R}_+$ is a non-negative system state, whose initial conditions belong to a given interval:

$$x_0 \in [\underline{x}_0, \overline{x}_0];$$

$a_t \in \mathbb{R}_+$ and $d_t \in \mathbb{R}$ are uncertain inputs, which also take values in known intervals:

$$a_t \in [\underline{a}_t, \overline{a}_t],\ d_t \in [\underline{d}_t, \overline{d}_t]$$

for all $t \in \mathbb{N}$. So, we assume that $\underline{x}_0 \leq \overline{x}_0$, $0 \leq \underline{a}_t \leq \overline{a}_t$ and $\underline{d}_t \leq \overline{d}_t$ are known for all $t \in \mathbb{N}$. The imposed non-negativity constraints on x_t and a_t correspond to the case of the model (1). We would like to calculate the lower $\underline{x}_t$ and upper $\overline{x}_t$

predictions of the state x_t of this system under the introduced hypotheses on all uncertain variables, requiring the relations:

$$\underline{x}_t \le x_t \le \overline{x}_t \quad \forall t \in \mathbb{N}.$$

The theory of interval observers and predictors [22], [28] answers this question, and a possible solution (that utilizes the non-negativity of x_t and a_t) is as follows:

$$\begin{aligned}\underline{x}_{t+1} &= \underline{a}_t \underline{x}_t + \underline{d}_t,\\ \overline{x}_{t+1} &= \overline{a}_t \overline{x}_t + \overline{d}_t,\end{aligned}$$

which is rather straightforward. To substantiate the desired interval inclusion for x_t by $\underline{x}_t, \overline{x}_t$, we can consider the lower $\underline{e}_t = x_t - \underline{x}_t$ and the upper $\overline{e}_t = \overline{x}_t - x_t$ prediction errors, whose dynamics take the form:

$$\begin{aligned}\underline{e}_{t+1} &= (a_t x_t - \underline{a}_t \underline{x}_t) + (d_t - \underline{d}_t),\\ \overline{e}_{t+1} &= (\overline{a}_t \overline{x}_t - a_t x_t) + (\overline{d}_t - d_t),\end{aligned}$$

then it is easy to verify that the terms $d_t - \underline{d}_t, \overline{d}_t - d_t$ are non-negative by the definition of $\underline{d}_t, \overline{d}_t$, and the terms $a_t x_t - \underline{a}_t \underline{x}_t, \overline{a}_t \overline{x}_t - a_t x_t$ have the same property for $t = 0$ by the definition of $\underline{a}_t, \overline{a}_t$ and $\underline{x}_0, \overline{x}_0$, therefore, $\underline{e}_1 \ge 0$, $\overline{e}_1 \ge 0$ (that implies $x_1 \in [\underline{x}_1, \overline{x}_1]$) and the analysis can be iteratively repeated for all $t \in \mathbb{N}$. Let us apply this method to the model (1) (clearly, each equation there has the form as above).

B. Equations of interval predictor and its properties

To this end, we assume that all parameters belong to the known intervals (for simplicity we do not deviate the values of τ_p, τ_r and κ):

$$\begin{aligned}\sigma \in [\underline{\sigma}, \overline{\sigma}],\ \gamma \in [\underline{\gamma}, \overline{\gamma}],\ b \in [\underline{b}, \overline{b}],\\ p_t \in [\underline{p}_t, \overline{p}_t],\ r_t \in [\underline{r}_t, \overline{r}_t] \quad \forall t \in \mathbb{N},\end{aligned} \tag{3}$$

together with the initial conditions in (1):

$$\begin{aligned}S_0 \in [\underline{S}_0, \overline{S}_0],\ I_0 \in [\underline{I}_0, \overline{I}_0],\ E_0 \in [\underline{E}_0, \overline{E}_0],\\ D_0 \in [\underline{D}_0, \overline{D}_0],\ R_0 \in [\underline{R}_0, \overline{R}_0],\end{aligned} \tag{4}$$

where non-negative values $\underline{\sigma}, \overline{\sigma}$, $\underline{\gamma}, \overline{\gamma}$, $\underline{b}, \overline{b}$, $\underline{p}_t, \overline{p}_t$, $\underline{r}_t, \overline{r}_t$, $\underline{S}_0, \overline{S}_0$, $\underline{I}_0, \overline{I}_0$, $\underline{E}_0, \overline{E}_0$, $\underline{D}_0, \overline{D}_0$ and $\underline{R}_0, \overline{R}_0$ are obtained from the ones used in the previous section by applying $\pm\delta\%$ deviation from those nominal quantities. Then, applying the approach explained just above, we derive the equations of the interval predictor:

$$\begin{aligned}\underline{S}_{t+1} &= \left(1 - \overline{b}\frac{\left(\overline{p}_{t-\tau_p}\overline{I}_t + \overline{r}_{t-\tau_r}\overline{E}_t\right)}{N}\right)\underline{S}_t,\\ \underline{E}_{t+1} &= \left(1 - (1+\kappa)\overline{\sigma} + \underline{b}\frac{r_{t-\tau_r}}{N}\underline{S}_t\right)\underline{E}_t + \underline{p}_{t-\tau_p}\underline{b}\frac{I_t S_t}{N},\\ \underline{I}_{t+1} &= (1 - \overline{\gamma} - \overline{\mu})\underline{I}_t + \underline{\sigma}\underline{E}_t,\\ \underline{R}_{t+1} &= \underline{R}_t + \underline{\gamma}\underline{I}_t + \kappa\underline{\sigma}\underline{E}_t,\\ \underline{D}_{t+1} &= \underline{D}_t + \underline{\mu}\underline{I}_t,\\ \overline{S}_{t+1} &= \min\left\{N, \left(1 - \underline{b}\frac{\left(\underline{p}_{t-\tau_p}\underline{I}_t + \underline{r}_{t-\tau_r}\underline{E}_t\right)}{N}\right)\overline{S}_t\right\},\\ \overline{E}_{t+1} &= \min\left\{N, \left(1 - (1+\kappa)\underline{\sigma} + \overline{b}\frac{\overline{r}_{t-\tau_r}}{N}\overline{S}_t\right)\overline{E}_t + \overline{p}_{t-\tau_p}\overline{b}\frac{\overline{I}_t\overline{S}_t}{N}\right\},\\ \overline{I}_{t+1} &= \min\left\{N, (1 - \underline{\gamma} - \underline{\mu})\overline{I}_t + \overline{\sigma}\overline{E}_t\right\},\\ \overline{R}_{t+1} &= \min\left\{N, \overline{R}_t + \overline{\gamma}\overline{I}_t + \kappa\overline{\sigma}\overline{E}_t\right\},\\ \overline{D}_{t+1} &= \min\left\{N, \overline{D}_t + \overline{\mu}\overline{I}_t\right\},\end{aligned}$$

where $\underline{S}_t, \overline{S}_t$, $\underline{I}_t, \overline{I}_t$, $\underline{E}_t, \overline{E}_t$, $D_t, \overline{D}_t$ and $\underline{R}_t, \overline{R}_t$ are the lower and upper interval predictions for S_t, I_t, E_t, D_t and R_t, respectively.

Theorem 8. *For the model* (1) *satisfying the relations* (3) *and* (4) *with*

$$2\overline{b}\sup_{t\in\mathbb{N}}\overline{r}_t \le 1,\ \overline{\sigma} \le \frac{1}{1+\kappa},\ \overline{\gamma} + \overline{\mu} \le 1, \tag{6}$$

the interval predictor (5) *guarantees the interval inclusions for the state of* (1) *for all* $t \in \mathbb{N}$*:*

$$\begin{aligned}S_t \in [\underline{S}_t, \overline{S}_t],\ I_t \in [\underline{I}_t, \overline{I}_t],\ E_t \in [\underline{E}_t, \overline{E}_t],\\ D_t \in [\underline{D}_t, \overline{D}_t],\ R_t \in [\underline{R}_t, \overline{R}_t]\end{aligned}$$

with boundedness of all predictions for all $t \in \mathbb{N}$*:*

$$\underline{S}_t, \overline{S}_t, \underline{I}_t, \overline{I}_t, \underline{E}_t, \overline{E}_t, \underline{D}_t, \overline{D}_t, \underline{R}_t, \overline{R}_t \in [0, N].$$

Proof: By direct calculation, we can check that

$$\begin{aligned}&\underline{b}\frac{\left(\underline{p}_{t-\tau_p}\underline{I}_t + \underline{r}_{t-\tau_r}\underline{E}_t\right)}{N} \le b\frac{\left(p_{t-\tau_p}I_t + r_{t-\tau_r}E_t\right)}{N} \le \overline{b}\frac{\left(\overline{p}_{t-\tau_p}\overline{I}_t + \overline{r}_{t-\tau_r}\overline{E}_t\right)}{N},\\ &\underline{b}\frac{\underline{r}_{t-\tau}}{N}\underline{S}_t - (1+\kappa)\overline{\sigma} \le b\frac{r_{t-\tau}}{N}S_t - (1+\kappa)\sigma \le \overline{b}\frac{\overline{r}_{t-\tau}}{N}\overline{S}_t - (1+\kappa)\underline{\sigma},\\ &\underline{p}_{t-\tau_p}\underline{b}\frac{\underline{I}_t\underline{S}_t}{N} \le p_{t-\tau_p}b\frac{I_t S_t}{N} \le \overline{p}_{t-\tau_p}\overline{b}\frac{\overline{I}_t\overline{S}_t}{N},\\ &\underline{\sigma}\underline{E}_t \le \sigma E_t \le \overline{\sigma}\overline{E}_t,\\ &\underline{\gamma}\underline{I}_t \le \gamma I_t \le \overline{\gamma}\overline{I}_t,\\ &\underline{\mu}\underline{I}_t \le \mu I_t \le \overline{\mu}\overline{I}_t\end{aligned}$$

due to (3) and (4) for $t = 0$. Since (recall that $r_t \ge p_t$, $\overline{I}_t + \overline{E}_t \le 2N$, thus $\underline{S}_t \ge 0$)

$$\begin{aligned}&\overline{b}\frac{\left(\overline{p}_{t-\tau_p}\overline{I}_t + \overline{r}_{t-\tau_r}\overline{E}_t\right)}{N} \le \overline{b}\overline{r}_{t-\tau_r}\frac{\overline{I}_t + \overline{E}_t}{N} \le 2\overline{b}\overline{r}_{t-\tau_r} \le 2\overline{b}\sup_{t\in\mathbb{N}}\overline{r}_t,\\ &1 - (1+\kappa)\overline{\sigma} + \underline{b}\frac{\underline{r}_{t-\tau_r}}{N}\underline{S}_t \ge 1 - (1+\kappa)\overline{\sigma},\end{aligned}$$

we obtain that

$$1 \geq \overline{b}\frac{\left(\overline{p}_{t-\tau_p}\overline{I}_t + \overline{r}_{t-\tau_r}\overline{E}_t\right)}{N},\ 1+\underline{b}\frac{\underline{r}_{t-\tau_r}}{N}\underline{S}_t \geq (1+\kappa)\overline{\sigma}$$

due to (6), then as we demonstrated above

$$S_1 \in [\underline{S}_1, \overline{S}_1],\ I_1 \in [\underline{I}_1, \overline{I}_1],\ E_1 \in [\underline{E}_1, \overline{E}_1],$$
$$D_1 \in [\underline{D}_1, \overline{D}_1],\ R_1 \in [\underline{R}_1, \overline{R}_1],$$

and such a verification can be repeated for all $t \in \mathbb{N}$. In the same way we can show that if the relations

$$\underline{S}_t \leq \overline{S}_t,\ \underline{I}_t \leq \overline{I}_t,\ \underline{E}_t \leq \overline{E}_t,\ \underline{D}_t \leq \overline{D}_t,\ \underline{R}_t \leq \overline{R}_t$$

are satisfied for some $t \in \mathbb{N}$, then they also hold for $t+1$ in (5), and to substantiate boundedness of the state of the interval predictor, it is enough to guarantee that $\overline{S}_t$, $\overline{I}_t$, $\overline{E}_t$, $\overline{D}_t$ and $\overline{R}_t$ doe not exceed N (as it is done by construction in (5)), while non-negativity of $\underline{S}_t$, $\underline{I}_t$, $\underline{E}_t$, $\underline{D}_t$ and $\underline{R}_t$ is ensured by (6). ■

Remark 9. The boundedness of the state of (5) established in Theorem (8) does not imply the stability of the internal dynamics of the interval predictor (it is also a reason to impose the explicit saturation in (5)), which is a frequent and challenging problem for these predictors [21], [28].

Remark 10. The dynamics of lower and upper interval predictions are interrelated through the update equations of $\underline{S}_t, \overline{S}_t$. Thus, the dimension of the predictor (5) is twice higher than in the system (1). The values of the variables $\underline{S}_t, \overline{S}_t$ can be evaluated using the population equation $S_t + E_t + I_t + R_t + D_t = N$:

$$\underline{S}_t = N - \overline{I}_t - \overline{E}_t - \overline{R}_t - \overline{D}_t,$$
$$\overline{S}_t = N - \underline{I}_t - \underline{E}_t - \underline{R}_t - \underline{D}_t,$$

which however does not isolate the dynamics of lower and upper interval predictions. In addition, preliminary simulations show that such a modification leads to more conservative results, so we keep (5) for all further utilization.

Remark 11. Note that the calculated by (5) interval prediction estimates are also valid in the case of time-varying parameters in (1) satisfying the interval inclusions (3), which enlarges reliability of the proposed interval outbreak prediction.

C. The results of numeric experiments

The simulation results of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 6 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the results of simulation obtained in the previous section, the circles depict measured and reconstructed data points used for identification and validation). The width of the predicted interval of admissible values for the state of (1) is growing, which is related with a high level of uncertainty reflected by δ and chosen for these simulations.

As we can conclude from these curves, under sufficiently big deviations of the parameters (which correspond to the amount and quality of data publicly available now), the confinement may slow down the epidemics. The measurements are nearly included in the obtained intervals validating the prediction (the value of δ was selected to ensure this property).

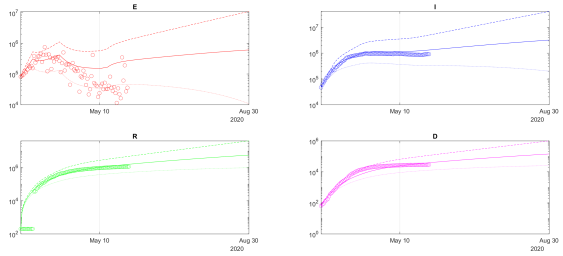

Fig. 6. The results of simulation of (5) for France under $\pm 7.5\%$ variation of all parameters

There are two variants of epidemic development demonstrated in these results: optimistic, which corresponds to the lower bounds of I and E, and pessimistic presented by the respective upper bounds. Under the current level of uncertainty, in the optimistic setting, the maximum of infected people has been already reached approximately mid-April.

To check the accuracy of the prediction we can select a part of the data for identification, and another part for verification of prediction reliability. The results of such a validation are shown in Fig. 7, where the interval prediction for the infectious population I is presented with a deviation of all parameters. As previous, blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the case is given in Fig. 6, the bold lines are calculated using previous day initial conditions), the blue circles and squares are the measured information used for identification and validation, and the red line is the average behavior. In the top plot, all data (*i.e.*, for $t = 0, 1, \ldots, J$) is used for identification as described above, then the interval predictor is launched from the last day initial conditions (bold dashed and dotted blue lines and the red one). In the bottom plot only the data points for $t = 0, 1, \ldots, J-40$ are utilized, they are shown by circles, and the interval predictor is initiated with the data for $t = J-41$, then square data points (which were not taken into account during identification for $t = J-40, \ldots, J$) can be compared with the predictor trajectories (bold dashed and dotted blue lines and the red one). As we can see, the points marked by squares are well included in the predicted interval, which confirms the reliability of (5) at least for 40 days, then returning to the top plot we can similarly use it for prediction of future development of COVID-19 for $30-50$ days.

In general, a further precision of the model and the parameters is needed, but as a recommendation after these preliminary simulations: the *preservation* of the quarantine rules is desirable. The model shows a rather low decreasing in the number of infected individuals, then prolongation of isolation for a fragile part of the population and social distancing is reasonable (it is worth noting that the value of p_R is selected *ad-hoc* and probably too high).

V. Other countries

In this section we do not repeat all the considerations and analysis previously presented for France. For the other countries, we introduce the used data together with the selected parameters, if they differ from the ones accepted above,

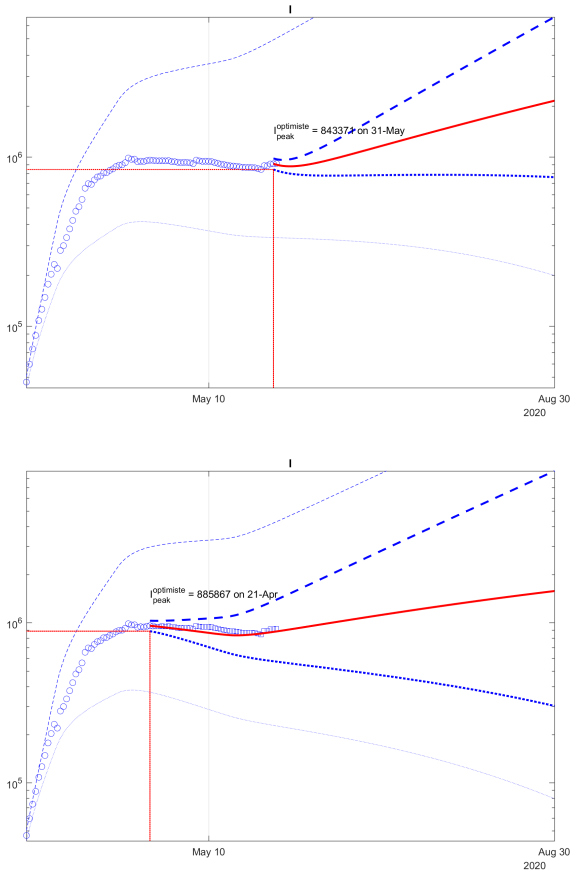

Fig. 7. Validation of prediction of I for France with J and $J-40$ points of data under deviations of values of all parameters

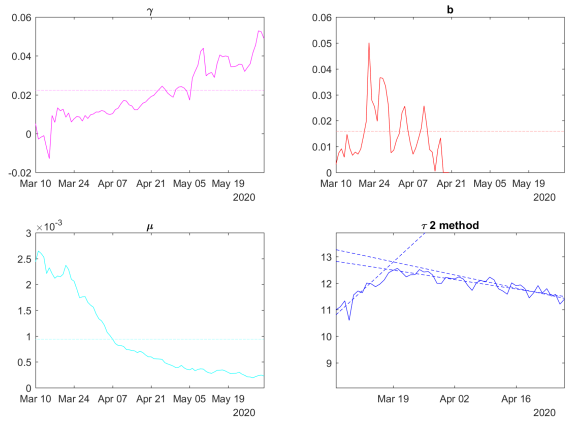

Fig. 8. The identified parameters for Italy

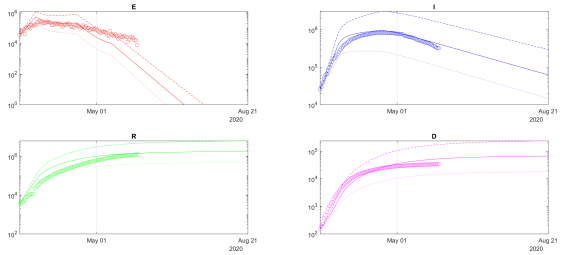

Fig. 9. The results of simulation of (5) for Italy under $\pm 7.5\%$ variation of all parameters

identify the parameters (as in Fig. 3) and simulate the interval predictor (as in Fig. 6 together with the plots of validation Fig. 7). The common parameters assigned to all countries (to simplify the analysis) are:

$$\sigma = \frac{1}{7},\ \kappa = 0.1,$$

for chosen values of p_Q, p_N, p_R, p_M, please, check the code in Github. Of course, adjusting these values for each country definitely improves the forecast precision, but our goal here is mainly illustrate a wide applicability of the proposed method for virus propagation interval prediction.

A. Italy

The current population in Italy is $N = 60359546$[7],

$$\alpha_1 = 7.7,\ \tau_r = 10,\ \tau_p = 30.$$

The data available from public sources for the time period between March 5th and June 2nd is provided in Github[8].

Applying the proposed procedure to parameter identification leads to

$$\mu = 9.3987 \times 10^{-4},\ \gamma = 0.0223,\ b = 0.0159$$

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 8. France and Italy have the recovery rate γ approximately close, but the virus infection rate b in Italy is lower compared to France.

The simulation results of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 9 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows well the measured statistics for Italy.

A restricted fit of the model to the data is demonstrated in Fig. 10 (blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the magenta circles and squares are the measured information used for identification and validation), where the square points belong to the middle of the predicted interval in the bottom plot, while the top plot can be used for prediction of the number of infectives in $10-30$ days. Unfortunately, we selected model cannot follow such a quick decay of the infected individuals.

B. Spain

The current population in Spain is $N = 46600396$[9],

$$\alpha_1 = 6.7,\ \tau_r = 8,\ \tau_p = 30.$$

[7] https://en.wikipedia.org/wiki/Demographics_of_Italy.
[8] Sources: Eficiens, Johns Hopkins University.
[9] https://en.wikipedia.org/wiki/Demographics_of_Spain.

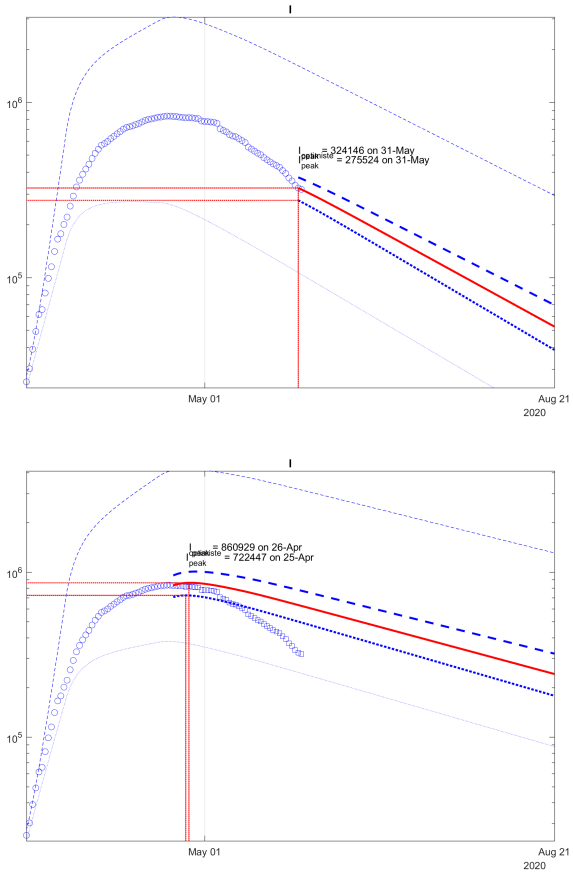

Fig. 10. Validation of prediction of I for Italy with J and $J-40$ points of data under deviations of values of all parameters

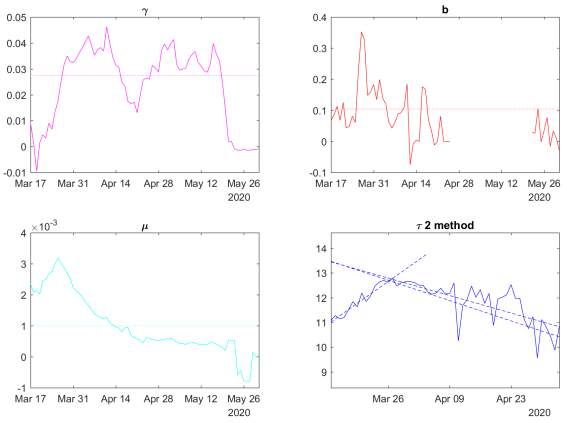

Fig. 11. The identified parameters for Spain

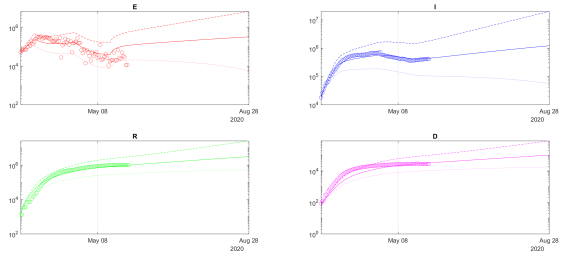

Fig. 12. The results of simulation of (5) for Spain under $\pm 7.5\%$ variation of all parameters

The data available from public sources for the time period between March 12th and June 2nd is provided in Github[10].

Applying the proposed procedure to parameter identification leads to

$$\mu = 10.00 \times 10^{-4},\ \gamma = 0.0275,\ b = 0.1041$$

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 11. Thus, the recovery rate γ and the infection rate b in Spain belong to the same range of values as in France (having slightly higher values in Spain, which implies a faster virus propagation in the country), similarly for the mortality rate μ.

The simulation results of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 12 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows well the measured statistics for Spain.

A good fit of the model to the data is demonstrated in Fig. 13 (blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the blue circles and squares are the measured information used for identification and validation), where the square points lie close to the middle of the predicted interval except at the end (there is an issue with Spain data during last weeks) in the bottom plot, while the top plot can be used for prediction of the number of infectives in $30-50$ days.

C. Germany

The current population in Germany is $N = 46600396$[11],

$$\alpha_1 = 3.3,\ \tau_r = 3,\ \tau_p = 21.$$

The data available from public sources for the time period between March 12th and June 2nd is provided in Github[12].

Applying the proposed procedure to parameter identification leads to

$$\mu = 8.9617 \times 10^{-4},\ \gamma = 0.0693,\ b = 0.1152$$

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 14. Thus, the recovery rate γ in Germany is much higher, while the mortality μ and the infection b rates belong to the same range of values as in France (b is slightly higher). The method of delay estimation can also be well illustrated here since the decreasing phase is active during several weeks.

The simulation results of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 15 (the dashed and dotted lines represent, respectively, upper and lower interval bounds,

[10]Sources: Statista, Johns Hopkins University.

[11]https://en.wikipedia.org/wiki/Demographics_of_Germany.

[12]Sources: www.rki.de, Johns Hopkins University.

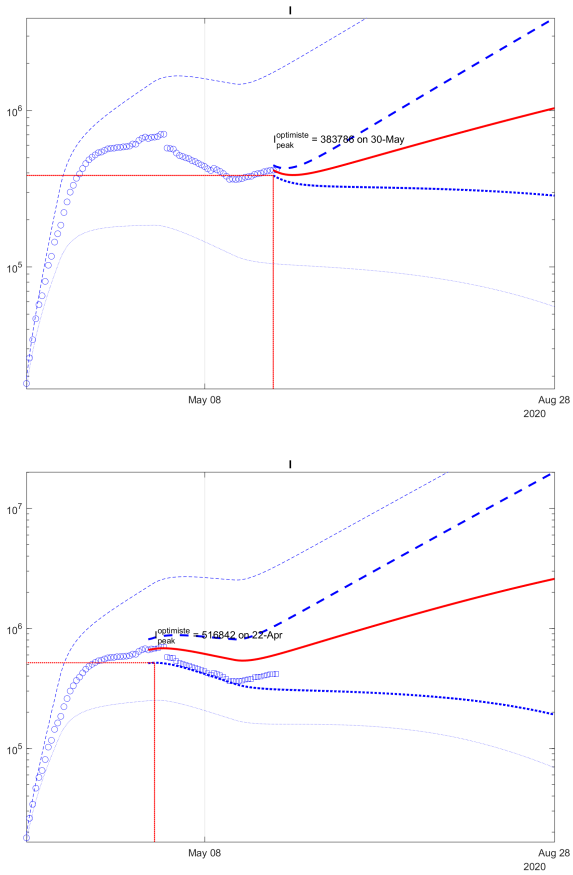

Fig. 13. Validation of prediction of I for Spain with J and $J-40$ points of data under deviations of values of all parameters

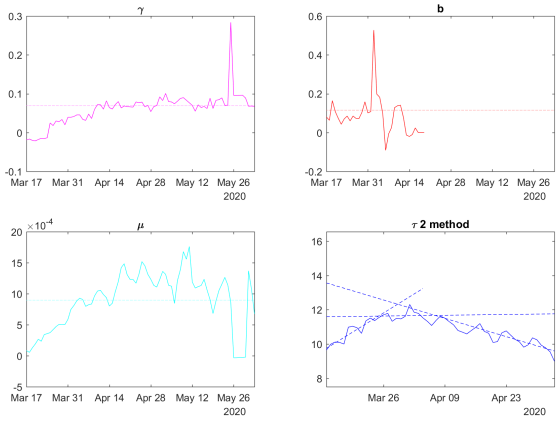

Fig. 14. The identified parameters for Germany

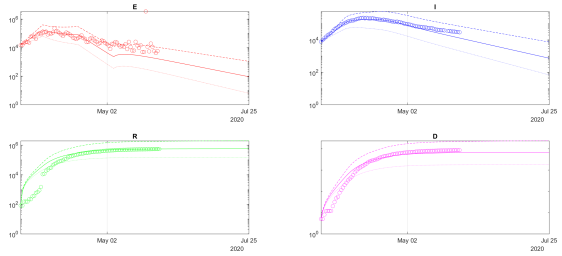

Fig. 15. The results of simulation of (5) for Germany under $\pm 7.5\%$ variation of all parameters

the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows well the measured statistics for Germany. The country infection peak seems to be already passed in the middle of April.

The predictor validation is demonstrated in Fig. 16 (blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the blue circles and squares are the measured information used for identification and validation), where the square points are included again in the predicted interval in the bottom plot, while the top plot can be used for prediction of the number of infectives in $30-50$ days.

D. Brazil

The current population in Brazil is $N = 212559417$[13],

$$\alpha_1 = 6.7,\ \tau_r = 3,\ \tau_p = 35.$$

The data available from public sources for the time period between March 12th and April 27th is provided in Github[14].

Applying the proposed procedure to parameter identification leads to

$$\mu = 12.00 \times 10^{-4},\ \gamma = 0.0579,\ b = 0.1473$$

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 17. The mortality μ and infection b rates are slightly higher than the same quantities for France, Italy, Spain and Germany, but the value of the recovery rate γ has a lot of variation, which can be related with the fact that the epidemics in Brazil is more important, or that different protocols are used for registration.

The simulation results of the interval predictor (5) with $\delta = 12\%$ are presented in Fig. 18 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows the measured statistics for Brazil, but with higher variations.

[13] https://en.wikipedia.org/wiki/Demographics_of_Brazil.
[14] Sources: Johns Hopkins University.

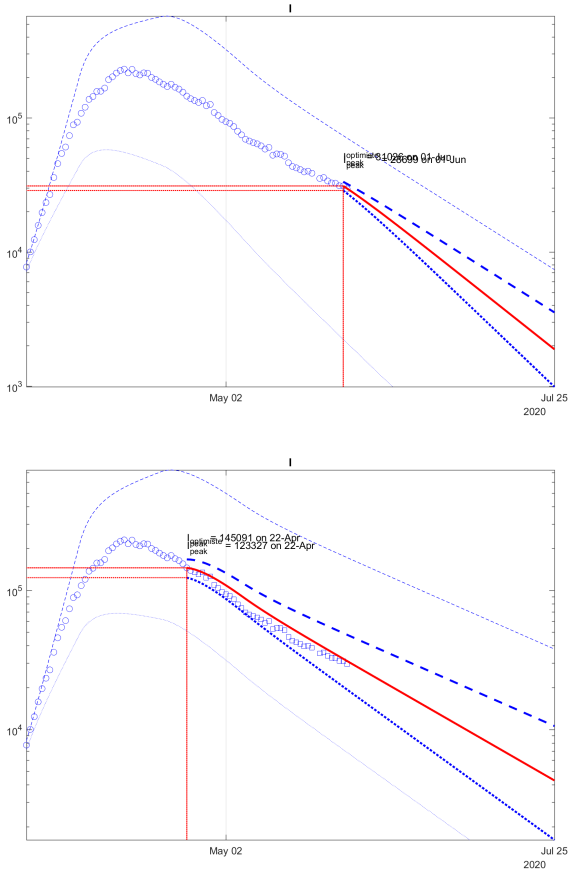

Fig. 16. Validation of prediction of I for Germany with J and $J-40$ points of data under deviations of values of all parameters

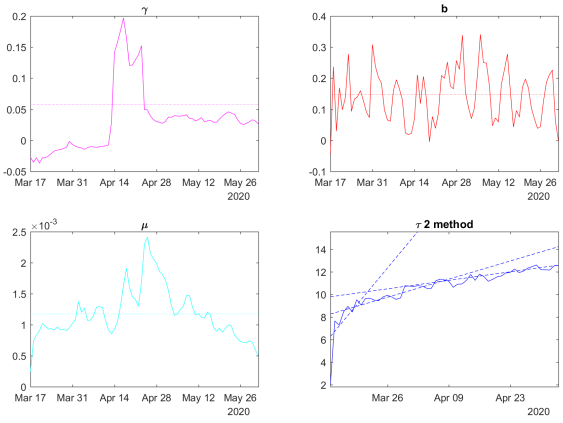

Fig. 17. The identified parameters for Brazil

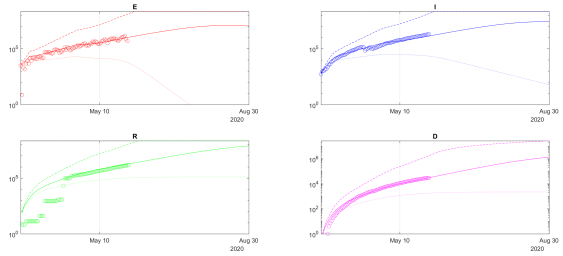

Fig. 18. The results of simulation of (5) for Brazil under $\pm 12\%$ variation of all parameters

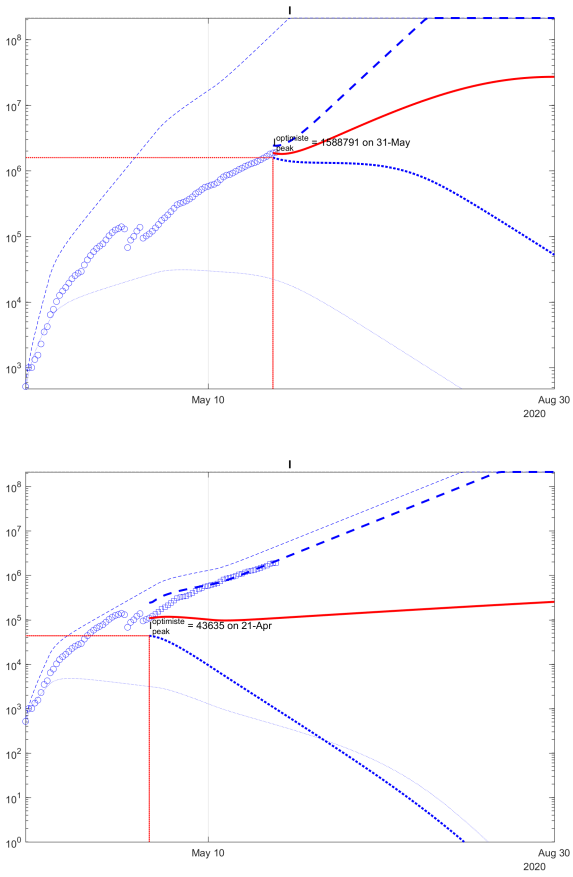

Fig. 19. Validation of prediction of I for Brazil with J and $J-40$ points of data under deviations of values of all parameters

Verification of accuracy of the predictor using a part of data is demonstrated in Fig. 19 (blue dashed and dotted lines correspond do the upper $\bar{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the blue circles and squares are the measured information used for identification and validation), where the square points belong to the predicted interval in the bottom plot.

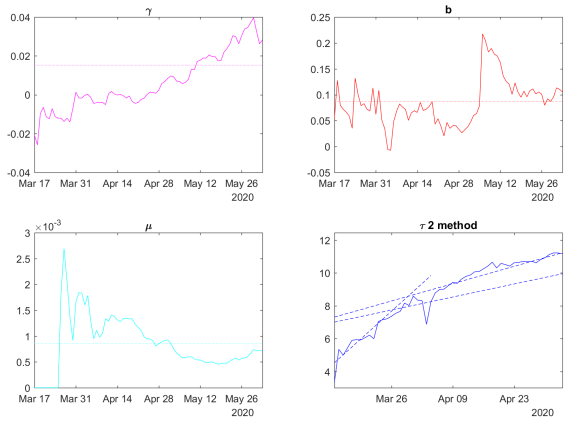

Fig. 20. The identified parameters for Russia

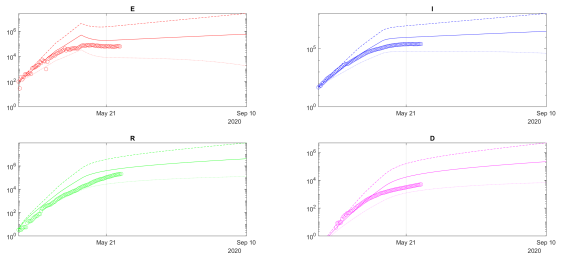

Fig. 21. The results of simulation of (5) for Russia under ±7.5% variation of all parameters

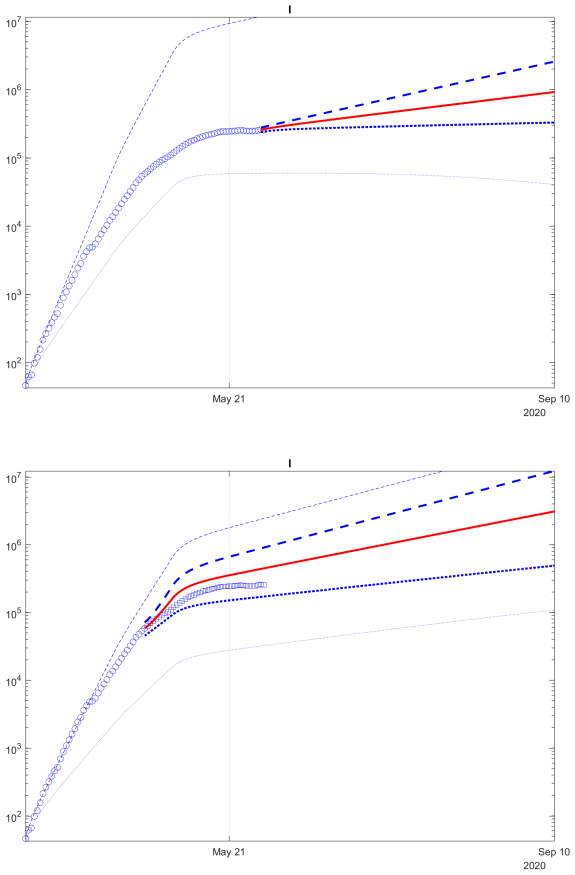

Fig. 22. Validation of prediction of I for Russia with J and $J-40$ points of data under deviations of values of all parameters

E. Russia

The current population in Russia is $N = 146745098$[15],

$$\alpha_1 = 1.1,\ \tau_r = 15,\ \tau_p = 20.$$

The data available from public sources for the time period between March 12th and June 2nd is provided in Github[16].

Applying the proposed procedure to parameter identification leads to

$$\mu = 8.5619 \times 10^{-4},\ \gamma = 0.0152,\ b = 0.0870$$

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 20. In this case the values of the parameters μ and b are close to those for European countries, but the recovery rate γ is rather low.

The results of simulation of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 21 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows the measured statistics for Russia pretty well.

An adequate fit of the model to the data is demonstrated in Fig. 22 (blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the blue circles and squares are the measured information used for identification and validation), where the square points belong to the lower part of the predicted interval in the bottom plot, while the top plot can be used for prediction of the number of infectives in $30-50$ days (in these cases, in the optimistic scenario the model predicts that the peak will be passed in June).

F. New York State (US)

Since US is a big and strongly heterogeneous country, we decided just to test our approach on one state, New York. The current population in New York State is $N = 19453561$[17],

$$\alpha_1 = 4,\ \alpha_2 = 5\alpha_1,\ \tau_r = 5,\ \tau_p = 20,$$

where we make a correction for α_2 since this state performs an intensive testing of the population that changes the measurement patterns. The data available from public sources for the time period between March 16th and June 2nd is provided in Github[18].

Applying the proposed procedure to parameter identification leads to

$$\mu = 6.5199 \times 10^{-4},\ \gamma = 0.0271,\ b = 0.0815$$

[15] www.en.wikipedia.org/wiki/Demographics_of_Russia.
[16] Sources: Wikipedia, /www.rbc.ru.
[17] www.en.wikipedia.org/wiki/New_York_(state).
[18] Sources: Wikipedia, Johns Hopkins University.

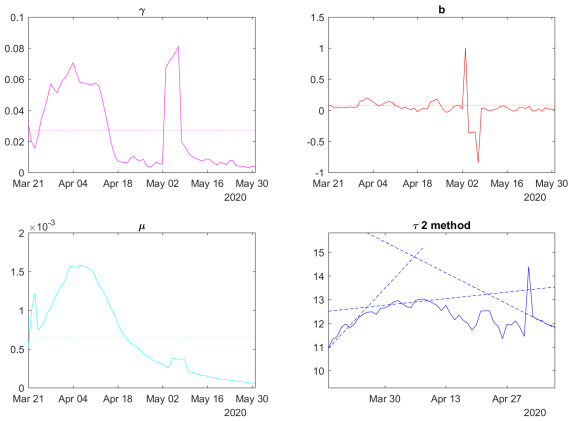

Fig. 23. The identified parameters for NY State

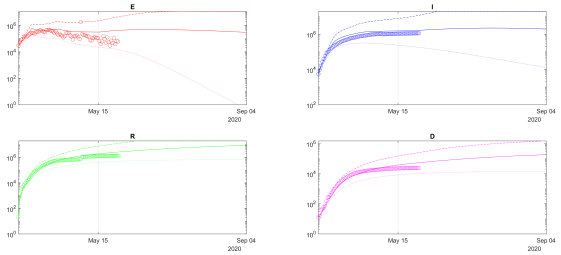

Fig. 24. The results of simulation of (5) for NY State under ±7.5% variation of all parameters

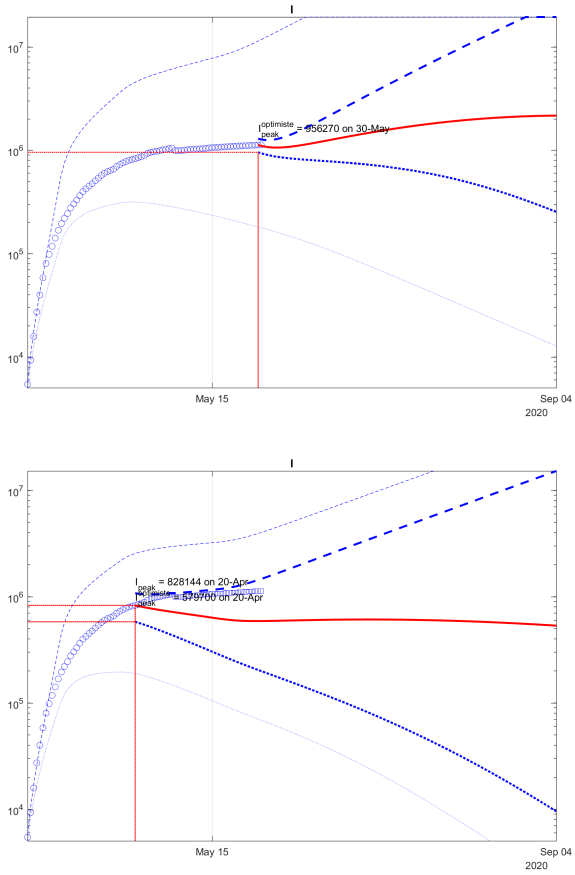

Fig. 25. Validation of prediction of I for NY State with J and $J-40$ points of data under deviations of values of all parameters

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 23. In this case, the values of all parameters are close to those for European countries.

The results of simulation of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 24 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows the measured statistics for New York State pretty well.

A limited fit of the model to the data is demonstrated in Fig. 25 (blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the blue circles and squares are the measured information used for identification and validation), where the square points belong to the top of the predicted interval in the bottom plot, while the top plot can be used for prediction of the number of infectives in $30-50$ days (in both cases, in the optimistic scenario the model predicts that the peak is passed in May).

G. China

The current population in China is $N = 1438070898$[19],

$$\alpha_1 = 1,\ \tau_r = 1,\ \tau_p = 15,$$

and despite big dimensions of the country, it is very homogeneous, following well the regulations, which allows us to consider it at whole. The data available from public sources for the time period between January 16th and June 2nd is provided in Github[20].

Applying the proposed procedure to parameter identification leads to

$$\mu = 10.40 \times 10^{-4},\ \gamma = 0.0760,\ b = 0.0238$$

with the values γ_k, b_k, μ_k, $\ln(E_t)$ given in Fig. 26. In this case the value of the parameter b is close to those for other countries, but the recovery γ and the mortality μ rates are rather high. The delay identification procedure is also well illustrated in this figure.

The results of simulation of the interval predictor (5) with $\delta = 7.5\%$ are presented in Fig. 27 (the dashed and dotted lines represent, respectively, upper and lower interval bounds, the solid lines correspond to the average behavior, the circles depict measured and reconstructed data points used for identification). As we can conclude from these plots, the obtained model follows the measured statistics for China rather well (the exclusion is E, but these points are not measured, they are constructed).

[19] www.worldometers.info/world-population/china-population/.

[20] Sources: Wikipedia, Johns Hopkins University.

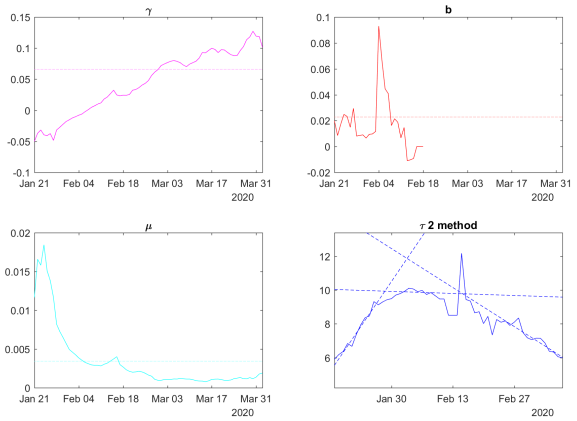

Fig. 26. The identified parameters for China

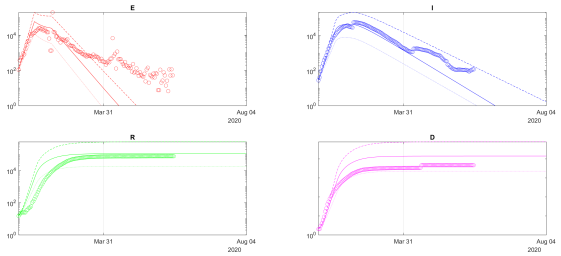

Fig. 27. The results of simulation of (5) for China under ±7.5% variation of all parameters

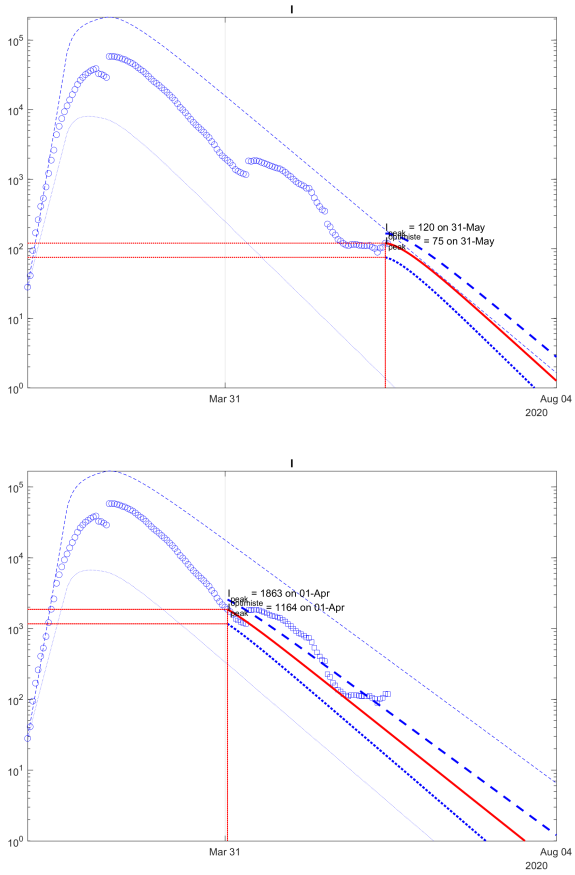

Fig. 28. Validation of prediction of I for China with J and $J-60$ points of data under deviations of values of all parameters

A reasonable fit of the model to the data is demonstrated in Fig. 28 (blue dashed and dotted lines correspond do the upper $\overline{I}$ and the lower bounds $\underline{I}$ (the bold lines are calculated using the last day included in the identification data), the red line is the average, the blue circles and squares are the measured information used for identification and validation), where the square points mainly belong to the predicted interval in the bottom plot, while the top plot can be used for prediction of the number of infectives in $50-70$ days.

VI. Conclusion

A simple discrete-time SEIR epidemic model was identified and used to predict the influence of the quarantine on the SARS-CoV-2 virus propagation in France, Italy, Spain, Germany, Brazil, Russia, New York State and China. To enlarge the model prediction performance, an interval predictor method was also used to analyze the COVID-19 course. It was demonstrated that the reliability of the interval prediction for $30-50$ days is rather good, even by such a simple model. The prediction showed that a longer confinement may be a bit more efficient, but under the current level of uncertainty, a more strict as possible quarantine seems to be advisable. The obtained results show that by selecting different profiles of contacts between the compartments in the countries it is possible to predict the outbreak development with a good accuracy (at least for a month).

The eight considered countries can be divided into two groups: four European states (France, Italy, Spain, and Germany) and China, where the virus presence is already well developed with several weeks of quarantine, and two BRICS countries (Brazil and Russia) with the US, where the epidemics started later and somewhat general confinement has been imposed more recently. The identified models for these groups of countries have common patterns (*e.g.*, a big variation of the recovery rate γ for Brazil and Russia). Our prediction showed that in European countries the peak of infections will occur in April-May in the optimistic scenario. An increased severity of the confinement could significantly decrease the amplitude of the peak discharging the load on health services.

Machine learning tools can be further used to identify and optimize the time profile for the confinement. Another possible direction of improvement of the proposed approach is to consider a SEIR model with population separation either by age or by region (or by both), but this implies an increasing number of parameters to be identified (that can be impossible) and also needs a special structured data to be available.

References

[1] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008. I

[2] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, "Statistical physics of vaccination," *Physics Reports*, vol. 664, pp. 1–113, 2016. I

[3] S. Das, P. Ghosh, B. Sen, and I. Mukhopadhyay, "Critical community size for COVID-19 – a model based approach to provide a rationale behind the lockdown," 2020. I

[4] R. Dandekar and G. Barbastathis, "Neural network aided quarantine control model estimation of global covid-19 spread," 2020. I

[5] L. R. Lopez and X. Rodo, "A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: simulating control scenarios and multi-scale epidemics," *medRxiv*, 2020. I

[6] B. Nussbaumer-Streit, V. Mayr, A. Dobrescu, A. Chapman, E. Persad, I. Klerings, G. Wagner, U. Siebert, C. Christof, C. Zachariah, and G. Gartlehner, "Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review," *Cochrane Database of Systematic Reviews*, no. 4, 2020. I

[7] Z. Yang, Z. Zeng, K. Wang, S.-S. Wong, W. Liang, M. Zanin, P. Liu, X. Cao, Z. Gao, Z. Mai, J. Liang, X. Liu, S. Li, Y. Li, F. Ye, W. Guan, Y. Yang, F. Li, S. Luo, Y. Xie, B. Liu, Z. Wang, S. Zhang, Y. Wang, N. Zhong, and J. He, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *Journal of Thoracic Disease*, vol. 12, no. 3, 2020. I-A, I-A, I-C2, 3, I-E

[8] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. C. Okell, S. van Elsland, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. G. Walker, C. Walters, P. Winskill, C. Whittaker, C. A. Donnelly, S. Riley, and A. C. Ghani, "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand," WHO Collaborating Centre for Infectious Disease Modelling, MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics Imperial College London, COVID-19 reports, 2020. I-A, 3, II-C

[9] B. F. Maier and D. Brockmann, "Effective containment explains subexponential growth in confirmed cases of recent COVID-19 outbreak in Mainland China," *medRxiv*, 2020. I-A, 3

[10] J. Lourenco, R. Paton, M. Ghafari, M. Kraemer, C. Thompson, P. Simmonds, P. Klenerman, and S. Gupta, "Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic," *medRxiv*, 2020. I-A, 3, II-A

[11] L. Peng, W. Yang, D. Zhang, C. Zhuge, and L. Hong, "Epidemic analysis of COVID-19 in China by dynamical modeling," 2020. I-A, 3

[12] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application," *Annals of Internal Medicine*, 2020. I-C2

[13] A. d'Onofrio, P. Manfredi, and P. Poletti, "The interplay of public intervention and private choices in determining the outcome of vaccination programmes," *PLOS ONE*, vol. 7, no. 10, p. e45653, 2012. I-C2

[14] B. Cantó, C. Coll, and E. Sánchez, "Estimation of parameters in a structured SIR model," *Advances in Difference Equations*, vol. 2017, no. 1, p. 33, 2017. [Online]. Available: https://doi.org/10.1186/s13662-017-1078-5 I-C2

[15] P.-A. Bliman, D. Efimov, and R. Ushirobira, "A class of nonlinear adaptive observers for SIR epidemic model," in *Proceedings of ECC'18, the 16th annual European Control Conference*, June 2018. I-C2

[16] P. Magal and G. Webb, "The parameter identification problem for SIR epidemic models: identifying unreported cases," *J. Math. Biol.*, vol. 77, pp. 1629–1648, 2018. I-C2

[17] R. Ushirobira, D. Efimov, and P. Bliman, "Estimating the infection rate of a SIR epidemic model via differential elimination," in *2019 18th European Control Conference (ECC)*, June 2019, pp. 1170–1175. I-C2

[18] J. Gouzé, A. Rapaport, and M. Hadj-Sadok, "Interval observers for uncertain biological systems," *Ecological Modelling*, vol. 133, pp. 46–56, 2000. I-D

[19] F. Mazenc and O. Bernard, "Interval observers for linear time-invariant systems with disturbances," *Automatica*, vol. 47, no. 1, pp. 140–147, 2011. I-D

[20] T. Raïssi, D. Efimov, and A. Zolghadri, "Interval state estimation for a class of nonlinear systems," *IEEE Trans. Automatic Control*, vol. 57, no. 1, pp. 260–265, 2012. I-D

[21] F. Mazenc, T. N. Dinh, and S. I. Niculescu, "Interval observers for discrete-time systems," *International Journal of Robust and Nonlinear Control*, vol. 24, pp. 2867–2890, 2014. I-D, 9

[22] D. Efimov and T. Raïssi, "Design of interval observers for uncertain dynamical systems," *Autom. Remote Control*, vol. 77, no. 2, pp. 191–225, 2016. I-D, 3, IV, IV-A

[23] K. H. Degue, D. Efimov, and A. Iggidr, "Interval estimation of sequestered infected erythrocytes in malaria patients," in *2016 European Control Conference (ECC)*, June 2016, pp. 1141–1145. I-D

[24] M. S. Aronna and P.-A. Bliman, "Interval observer for uncertain time-varying SIR-SI epidemiological model of vector-borne disease," in *2018 16th European Control Conference (ECC)*, Limassol, 2018. I-D

[25] K. H. Degue and J. Le Ny, "An interval observer for discrete-time SEIR epidemic models," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 5934–5939. I-D

[26] Z. Hu, Q. Ge, S. Li, E. Boerwincle, L. Jin, and M. Xiong, "Forecasting and evaluating intervention of Covid-19 in the World," *arXiv e-prints*, p. arXiv:2003.09800, Mar. 2020. 3

[27] P. Magal and G. Webb, "Predicting the number of reported and unreported cases for the covid-19 epidemic in South Korea, Italy, France and Germany," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/24/2020.03.21.20040154 6

[28] E. Leurent, D. Efimov, T. Raïssi, and W. Perruquetti, "Interval prediction for continuous-time systems with parametric uncertainties," in *Proc. IEEE Conference on Decision and Control (CDC)*, Nice, 2019. IV-A, 9