Understanding the COVID19 infection curves – finding the right numbers

Dragos-Victor Anghel (✉ dragos@theory.nipne.ro)

Horia Hulubei National Institute for R&D in Physics and Nuclear Engineering <https://orcid.org/0000-0003-4809-0482>

Ioan Tudor Alexandru Anghel

Colegiul National de Informatica Tudor Vianu

Research Article

Keywords: COVID19, epidemic models, compartmental models, SIR model

Posted Date: June 3rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-32516/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Understanding the COVID19 infection curves – finding the right numbers

Dragos-Victor Anghel* and Ioan Tudor Alexandru Anghel†

June 1, 2020

Abstract

We analyze the evolution of the COVID19 pandemics and show that the basic compartmental SIR model cannot explain the data, some characteristic time series being by more than an order of magnitude different from the fit function over significant parts of the documented time interval. To correct this large discrepancy, we amend the SIR model by assuming that there is a relatively large population that was infected but was not tested and confirmed. This assumption qualitatively changes the fitting possibilities of the model and, despite its simplicity, in most cases, all the time series can be quite well reproduced. Nevertheless, in some cases (i.e., countries or regions) the estimated susceptible population decreases too fast. In such a case, the *observed* dynamic is only due to the transitions between the two infected compartments—the *confirmed infected* and the *unconfirmed infected*—and the rate of closing the cases (by recovery or death) in the *confirmed infected* compartment. Our analysis proves that the number of infected people is significantly larger than the one recorded and we provide a method to estimate it. We also discuss some relevant extensions of this model, to improve the interpretation and the fitting of the data.

1 Introduction

In the absence of a vaccine and efficient antiviral drugs, the containment of the COVID19 pandemics relies mainly on non-pharmaceutical interventions. Such interventions depend upon the correct determination of the parameters that characterize the spread of the virus (like the transmission rate, recovery rate, death rate, etc.), which are fed into mathematical models to predict the time evolution of the pandemics. But obtaining the correct numbers is difficult while the pandemics unfolds, because of incomplete data and uncertainties in their interpretation. One such error, together with its solution, is addressed in this paper.

The mathematical modeling of the spreading of infectious diseases is centuries old [1] and eventually started with the work of Daniel Bernoulli on the variola spreading [2]. These models may be grouped in two categories: stochastic and deterministic. Whereas the stochastic models, based mainly on simulations, may be more suitable for smaller communities, the deterministic models are usually employed for large communities, like countries or cities, where the fluctuations around averages are small and the time evolution is accurately determined by a small number of parameters. Because the analyzed population is divided into compartments, the deterministic models are also called *compartmental models* [3]. The most simple such model is the SIR model, composed of three compartments: *Susceptible*, *Infected*, and *Recovered*, which was introduced in 1927 by Kermack, McKendrick, and Walker [3, 4]. The *Susceptible* compartment includes the healthy persons that are not immune and may get infected by contact with the sick people from the *Infected* compartment. The persons in the *Infected* compartment are transferred into the *Recovered* compartment as they recover or die (we shall call this also the *Closed cases* compartment). In this model it is also assumed that the recovered people are immune and cannot get sick again (therefore they do not return to the *Susceptible* compartment upon recovery).

If we denote by S , I , and R the number of people in the compartments *Susceptible*, *Infected*, and *Recovered*, respectively, then, in the simplest case one may assume that the total number of people, $N = S + I + R$ is constant. It is more simple to work with the normalized values, so we introduce the notations $s = S/N$, $y = I/N$, and

*Institutul Național de Cercetare-Dezvoltare pentru Fizică și Inginerie Nucleară Horia Hulubei, Măgurele, Romania, Email: dragos@theory.nipne.ro

†Colegiul Național de Informatică “Tudor Vianu”

$r = R/N$. Then, the time evolution of s , y , and r is given by the set of equations (see the Supplemental Material and Refs. [4, 5, 1])

$$\dot{s} \equiv \frac{ds}{dt} = -ps(t)y(t), \quad (1a)$$

$$\dot{y} \equiv \frac{dy}{dt} = [ps(t) - \lambda]y(t), \quad (1b)$$

$$\dot{r} \equiv \frac{dr}{dt} = \lambda y(t), \quad (1c)$$

where p and λ will be called the *infection rate* and the *recovery rate*, respectively. In general, p and λ may also be functions of t (for example, p is influenced by regulations imposed to the population, such as social distancing, whereas λ is influenced by the response of the public health system and the time passed since the infection), but for the moment we shall assume that they are constants. From Eqs. (1), only the first two are independent, the third one being a consequence of the normalization condition $s + y + r = 1$, which yields $\dot{r} = -\dot{s} - \dot{y}$. Equations (1) may be solved with the initial boundary conditions $s(0) = s_0$ and $y(0) = y_0$ (see the Supplementary Material for more details).

2 Results

2.1 Method 1 (M1): fitting the time series of the *active cases*

With the formalism presented above, one may try to describe the evolution of the COVID19 pandemics. The official numbers of infected, dead, and recovered persons may be found on several databases. We use the one from GitHub (<https://github.com/CSSEGISandData/COVID-19>). In Fig. 1 we plot $y(t)$ (*Active cases*) for Romania, Hubei (China), Germany, and South Korea, together with the fit of the SIR model (minimum square deviation), with constant p and λ (or a). This simplified model does not fit the data very well and more complex models should be employed.

Although the details of fitting the time series of the *Active cases* are important, here we shall focus on a more conspicuous problem. In Fig. 2 we show the time series of the *Closed cases* of the same countries and on the same periods as in Fig. 1. We observe that the SIR results for the same parameters as in Fig. 1 differ from the recorded data by orders of magnitude. Such discrepancies cannot be mended by small adjustments of the parameters p and λ or by assuming that they vary with time.

We did such fits for several countries and the situation was the same for all of them.

2.2 Method 2 (M2): fitting the time series of the *infections*

One may try another approach. The time variation of s , and therefore of y , is determined by the number of infectious persons, which may not be accurately described by the number of *active cases*, as it is assumed in the SIR model. This happens because there is no medical procedure that could eliminate a patient from the list of *active cases* immediately when it becomes noninfectious. Rather, one is eliminated from this list after a certain interval of time, when it tests negative for the COVID19 disease. This may allow for an overestimation of the number of *active cases* at any moment in time. For this reason, we may assume that the (total) number of infected people, which is $y(t) + r(t) = 1 - s(t)$, is more appropriate for the fitting procedure. In this way we obtain the plots of Fig. 3. Obviously, this fits the time series of the infectious population much better than Method 1, but it fails to fit (by orders of magnitude) the time series of the *Active cases*, as one can see in Fig. 4. Furthermore, for longer time intervals, the fitting functions for Romania and Germany of Fig. 3 are likely to differ significantly from the data.

Therefore, we observed in this section that the SIR model cannot simultaneously fit the time series of the active cases and the total number of infections or, equivalently, the active cases and the closed cases. We may expect that by changing the constant p and λ parameters with functions of time we may be able to better fit the *active cases* or the *closed cases*, but not both of them at the same time, because of the huge differences observed between the predictions and the time series. This prompts us to make some substantial changes in the model.

Figure 1: The *Active cases* $y(t)$ of COVID19 for four examples of countries (blue circles) and the SIR model fit (red curve). The parameters are indicated in each plot. In each case, we use the data from the day in which the number of confirmed cases exceeded 100, until the date indicated.

3 Methods

To explain the discrepancies observed in the previous section, we extend the SIR model by assuming that the number of infected people is significantly different from the one we know. For simplicity, we shall refer to this as the *SIR2* model. We assume that beside the known infected population I_1 , there are I_2 infected persons that we do not know about, such that the total infected population is $I = I_1 + I_2$. Similarly, the known and unknown recovered populations are denoted by R_1 and R_2 , respectively (see Fig. 5 for the schematic representation of this model). If the total population is again denoted by N , then the susceptible population is $S = N - I_1 - I_2 - R_1 - R_2$. If we introduce the densities $s \equiv S/N$, $s_1 \equiv S_1/N$, $y_1 \equiv I_1/N$, $y_2 \equiv I_2/N$, $r_1 \equiv R_1/N$, $r_2 \equiv R_2/N$, then the system of coupled differential equations becomes

$$\dot{s} = -(p_1 y_1 + p_2 y_2) s(t), \quad (2a)$$

$$\dot{y}_1 = \xi y_2 - \lambda_1 y_1, \quad (2b)$$

$$\dot{y}_2 = (p_1 y_1 + p_2 y_2) s(t) - \xi y_2 - \lambda_2 y_2, \quad (2c)$$

$$\dot{r}_1 = \lambda_1 y_1(t) + \zeta r_2, \quad (2d)$$

$$\dot{r}_2 = \lambda_2 y_2(t) - \zeta r_2. \quad (2e)$$

In the system (2) we assumed that the rates of contamination due to the infected populations I_1 and I_2 , denoted by p_1 and p_2 , respectively, are different, because the population I_1 is much better monitored and isolated from the healthy population (one may safely assume that $p_1 \ll p_2$). Furthermore, a contaminated person first enters the I_2 population and it has the rate ξ (assumed constant) to enter the *known* infected population I_1 . The rates by which

Figure 2: The *Closed cases* $r(t)$. The model is fitted to the time series of the *Active cases*, as shown in Fig. 1.

the cases are closed (by recovery or death) in the populations I_1 and I_2 , denoted by λ_1 and λ_2 , respectively, may be different because of the following two (competing) reasons: (1) the population I_1 gets better health care and (2) the symptoms of I_1 may be (on average) more severe than the ones of the population I_2 , because they may have led to the observation of COVID19.

The rate ζ of transferring population from R_2 to R_1 is mostly due to the postmortem detection of the SARS-Cov-2 virus. In practice, this may be very small and therefore it may be ignored (we set $\zeta = 0$ in numerical calculations).

If we (wrongly) assume that all the infected cases are known (as it was the case in the simple SIR model), $N = S_1 + I_1 + R_1$, we obtain the *fake* number conservation condition $\dot{s}_1 + \dot{y}_1 + \dot{r}_1 = 0$, which, in this case, may lead to discrepancies.

The initial conditions are $s_0 \equiv s(0)$, $y_{1,0} \equiv y_1(0)$, $y_{2,0} \equiv y_2(0)$, $r_{1,0} \equiv r_1(0)$, and $r_{2,0} \equiv r_2(0)$. Since not all these parameters are independent ($s_0 + y_{1,0} + y_{2,0} + r_{1,0} + r_{2,0} = 1$), we may choose, for example, $r_{2,0}$ to be determined by the other ones. To fit the data, we choose $p_1 = p_2/10$ or $p_2/20$ (the final differences are insignificant), $\lambda_1 = \lambda_2 = ap_2$, $\zeta = 0$ (as stated above). In this way we are left with the fitting parameters p_2 , a , and ξ . The results of the fit are shown in Figs. 6 - 9.

We observe that the SIR2 model fits the European countries Romania and Germany (considering the simplicity of the model), but the data for Hubei and South Korea are not fitted well—although it is much better than the simple SIR model. These discrepancies may happen for several reasons, from the biases that exist in the data to the time dependence of the parameters, as discussed above. Nevertheless, the dramatic differences between the fitting of different time series, observed in the previous section, do not appear in SIR2.

Similar results are obtained for other countries, as we show in the Supplemental Material.

Figure 3: The number of confirmed cases $1 - s(t) = y(t) + r(t)$ and the fit with the SIR model in four populations: Romania, Hubei (China), Germany, and South Korea. The fit parameters and the time interval from which the data is taken are shown in the plots.

4 Discussion

Whereas the SIR model with constant p and λ (infection and healing rates) completely fails to describe the data because of the large discrepancies between the fits of different time series (if we fit one time series, another one may differ from the fitting function by more than one order of magnitude in significant parts of the fitting range), the SIR2 model fits quite well the available data for several countries (e.g. Figs. 6 and 7 and the Supplemental Material), whereas for other countries some discrepancies may appear (e.g. Figs. 8 and 9).

We also notice that s decreases fast. According to our assumptions, s cannot be directly calculated from the available data, since we do not know y_2 and r_2 (or, equivalently, I_2 and R_2). We estimate it in the model. If s becomes too small, then the time variations of y_1 and r_1 are not determined by the actual infections, but by the rate of people transfer from compartment I_2 into compartment I_1 (COVID19 diagnosis at some time after the infection) and from I_1 to R_1 (recovery or death of people confirmed with COVID19).

The discrepancies between the data and the SIR2 estimates may be due to the fact that the parameters that we use in the models should be replaced with functions of time, according to the time evolution of the pandemic (people awareness, health system improvement, governmental measures, time dependence of the recovery rates, etc.) and the testing of the population. Another important aspect is that the spread of the disease is not geographically homogeneous and one should split countries (or regions) more localized containers in which the infections are more uniformly distributed. Furthermore, even in the same geographical area, some people are more exposed to the virus than others, which would lead to further splitting of the population into containers. These aspects are not incorporated into our model, but our description, while it remains simple, is reliable.

Nevertheless, maybe the most important conclusion of this work is that there is a significant part of the population

Figure 4: The time series of the *Active cases* and the SIR predictions, for the parameters obtained by fitting the *Confirmed cases* with the SIR model as shown in Fig. 3.

which was infected and which is not detected. This implies that at the time of this writing, at least in Europe, the susceptible population has decreased sufficiently in most countries, so that, if the virus causes immunity, a second wave of infections may be much smaller or it may not be possible if restrictions are lifted. If an increase of the number of case would appear upon relaxing the restrictions, this would be most likely due to the unequal exposure of the population to the virus. Furthermore, in some countries the present time variation of the COVID19 cases may be due to the diagnosis of infections that took place some time ago, rather due to fresh infections.

Acknowledgements

We are grateful to Prof. Dan Pîrjol for drawing our attention to this subject and for introducing us to basic SIR calculations.

Authors' contributions

DVA formulated the concept and the SIR2 model, contributed to the numerical calculations, to the interpretation of the results, and wrote the first draft of the paper. ITAA discussed the models, contributed to the numerical calculations (he wrote the first C++ codes for this), to the interpretation of the results, and to the writing of the paper.

Figure 5: The schematic representation of the extension of the SIR model, denoted here as SIR2 model.

Figure 6: The fitting of the SIR2 model to the data in Romania, for the indicated period. The obtained parameters are shown in the plots.

References

- [1] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599, 2000.
- [2] D. Bernoulli. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Mémoires de Mathématiques et de Physique, Académie Royale des Sciences, Paris*, 1760:1, 1760.
- [3] Compartmental models in epidemiology. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology.
- [4] W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115:700, 1927.

Figure 7: The fitting of the SIR2 model to the data in Germany, for the indicated period. The obtained parameters are shown in the plots.

Figure 8: The fitting of the SIR2 model to the data in Hubei, for the indicated period. The obtained parameters are shown in the plots.

- [5] T. Harko, F. S. N. Lobo, and M. K. M.K. Mak. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Applied Mathematics and Computation*, 236:184 – 194, 2014.
- [6] Dan Pîrjol. Solutie analitica pentru modelul SIR. private communication.

Figure 9: The fitting of the SIR2 model to the data in South Korea, for the indicated period. The obtained parameters are shown in the plots. The fluctuations observed in the last plot (*susceptible* population) are numerical error.

Understanding the COVID19 infection curves – finding the right numbers

Supplemental material

Dragos-Victor Anghel* and Ioan Tudor Alexandru Anghel†

June 1, 2020

1 The SIR model

As explained in the main text of the article, the SIR model consists of three compartments, *Susceptible*, *Infected*, and *Recovered*, each of them being populated by S , I , and R persons. The time evolution of these populations is given by the equation [1, 2, 3]

$$\dot{S} \equiv \frac{dS}{dt} = -pI(t) \frac{S(t)}{N}, \quad (1a)$$

$$\dot{I} \equiv \frac{dI}{dt} = \left[p \frac{S(t)}{N} - \lambda \right] I(t), \quad (1b)$$

$$\dot{R} \equiv \frac{dR}{dt} = \lambda I(t), \quad (1c)$$

where p and λ are called the *infection rate* and the *recovery rate*, respectively (see also the main text). By dividing the system (1) by N , we obtain the system (1) of the main text, namely

$$\dot{s} = -ps(t)y(t), \quad (2a)$$

$$\dot{y} = [ps(t) - \lambda]y(t), \quad (2b)$$

$$\dot{r} = \lambda y(t). \quad (2c)$$

If we divide the Eqs. (2) by p and denote $x \equiv pt$ and $a \equiv \lambda/p$, Eqs. (2) get the more general form [4]

$$s'(x) \equiv \frac{ds}{dx} = -s(x)y(x), \quad (3a)$$

$$y'(x) \equiv \frac{dy}{dx} = [s(x) - a]y(x), \quad (3b)$$

$$r'(x) \equiv \frac{dr}{dx} = ar(x), \quad (3c)$$

with a being the only parameter. From Eqs. (3) we see that a sets the shape of the curve whereas p sets the scale of the time interval or the evolution speed of the epidemic. In what follows, we shall omit to include explicitly the dependence of the functions on the t or x variables, unless we consider it necessary.

Other characteristics of the time evolution of s may be readily obtained from (3). For example, the first two equations may be combined into a single equation for s ,

$$s'' = s'^2/s + (s - a)s', \quad (4)$$

which may be further simplified if we introduce the notation $w(x) = \ln[s(x)]$:

$$w'' = (e^w - a)w' \quad (5)$$

*Institutul Național de Cercetare-Dezvoltare pentru Fizică și Inginerie Nucleară Horia Hulubei, Măgurele, Romania, Email: dragos@theory.nipne.ro

†Colegiul Național de Informatică “Tudor Vianu”, Bucharest, Romania

Figure 1: (S1) The function $f_a(w)$ (continuous lines) and the constants $a \ln(1 - y_0)$ (dashed lines) from the equation that gives the asymptotic value of w , when $x \rightarrow \infty$ (we assume $r_0 = 0$).

(by $\cdot''$ we denote the second derivative d^2/dx^2). Starting from the initial conditions $w(0) \equiv w_0 = \ln(s_0) = \ln(1 - y_0 - r_0) < 0$ and $w'(0) = -y_0 < 0$, one may obtain an analytical expression for $w(t)$. For this, we integrate both sides of Eq. (5) to obtain

$$w'(x) = a \ln(1 - y_0 - r_0) - \left[1 - r_0 + aw(x) - e^{w(x)} \right] \equiv -D(w, a, y_0, r_0). \quad (6)$$

In Eq. (6) we may separate the variables and obtain

$$x = \int_w^{\ln(1-y_0-r_0)} \frac{dw'}{D(w', a, y_0, r_0)}. \quad (7)$$

In the equations above, one can neglect r_0 , since this is in general either zero or very small.

We see that $D(w, a, y_0, r_0) \geq 0$. Furthermore, the integral diverges if and only if the denominator D becomes zero at the lower integration limit. If we denote this limit by w_∞ , then $D(w_\infty, a, y_0, r_0) = 0$ and $\lim_{x \rightarrow \infty} w(x) = w_\infty$. Therefore, w_∞ may be determined from the equation

$$1 - r_0 + aw_\infty - e^{w_\infty} = a \ln(1 - y_0 - r_0) \equiv aw_0. \quad (8)$$

Since $s' < 0$ for any x , then $w' < 0$, so $w_\infty < w_0 < 0$.

To find the solution of Eq. (8), we denote $f_a(w) \equiv 1 + aw - e^w$ and we observe that $f_a(\pm\infty) = -\infty$ and $f_a(0) = 0$. The derivative $f'_a(w) = a - e^w$ is monotonically decreasing, from $\lim_{w \rightarrow -\infty} f'_a(w) = a$ to $\lim_{w \rightarrow \infty} f'_a(w) = -\infty$. This implies that $f_a(w) = 0$ has two solutions, one of them being $w = 0$ and the other one being smaller or bigger than zero, depending on a . The function $f_a(w)$ has a maximum at w_m , which satisfies the condition $0 = f'_a(w_m) = a - e^{w_m}$, that is, $w_m = \ln a$. Therefore, if $a < 1$, $w_m < 0$, so the second solution is negative, whereas if $a > 1$, $w_m > 0$ and the second solution is positive. If $a = 1$, then both solutions coincide with $w_m = 0$. These aspects may be seen in Fig. S1.

From Eq. (8) and Fig. S1 we can see that if $a \geq 1$, then $\lim_{y_0+r_0 \rightarrow 0} w_\infty = 0$, whereas if $a < 1$, then $\lim_{y_0+r_0 \rightarrow 0} w_\infty < 0$. In epidemics $a < 1$, that is, a small initially infected population may spread the disease in a finite fraction of a large population. In this case, since at the beginning $y_0 + r_0 \ll 1$, then $s_0 \lesssim 1$ and $y'(x=0)$ of Eq. (3b) is positive. On the other hand, if $a \geq 1$, then $y'(x) < 0$ and the number of active cases decreases monotonically.

1.1 Fitting the data with SIR

To find the approximation of w for $x \ll 1$, we may start either from Eq. (6) or from Eq. (7). From Eq. (6), for example, using $w'(x \ll 1) \approx w'(0) = -y_0$, we obtain directly

$$w(x) \approx w(0) + w'(0)x = \ln(1 - y_0 - r_0) - u_0 x \quad \text{and} \quad s(x) \approx (1 - y_0 - r_0)e^{-y_0 x}. \quad (9a)$$

From Eq. (9a) we obtain the other quantities,

$$r(x) \approx ay_0x \quad \text{and} \quad y(x) \approx y_0e^{(1-y_0-r_0-a)x} \quad (9b)$$

As expected, in the short time limit, both s and y vary exponentially with time (or with $x = pt$). This exponential variation is a hallmark of the epidemic spreading at $x \ll 1$.

In the beginning of an epidemic, fitting the data with Eqs. (9) on a logarithmic scale, one can extract the model parameters a and u_0 :

$$\ln(y) = \ln(y_0) + (1 - y_0 - r_0 - a)pt \quad \text{and} \quad \ln(s) = \ln(1 - y_0 - r_0) - y_0pt. \quad (10)$$

We assume that $y_0 \ll 1 - a < 1$, so we can use the approximations $1 - y_0 - r_0 - a \approx 1 - a$ and $\ln(1 - y_0 - r_0) \approx 0$.

If an epidemic had evolved for some time and is far from the beginning, as is the case of the COVID19 pandemic, one should fit the whole time series. The COVID19 started spreading in different countries or regions at different times, so we fit each country or region separately. As explained in the main text, we take the data from <https://github.com/CSSEGISandData/COVID-19> and form the vectors of data S_d , I_d , and R_d , corresponding to the compartments S, I, and R, respectively. The unit of time is *1 day* and we choose $t = 0$ the day when the number of infections exceeds 100. We choose this starting point because, on one hand, we want to eliminate the large fluctuations that may occur in the beginning of the epidemic spreading and, on the other hand, to keep as many data points as possible. If N is the total population of the country or region analyzed, we calculate the densities $s_d = S_d/N$, $y_d = I_d/N$, and $r_d = R_d/N$. On the other hand, we calculate the time evolution $s(t)$, $y(t)$, and $r(t)$, using Eqs. (2) and the initial conditions $s_0 = s_d(1)$, $y_0 = y_d(1)$, and $r_0 = r_d(1)$. The functions $s(t)$, $y(t)$, and $r(t)$ depend on the parameters p and λ , which we determine by fitting the data using the *least squares method*.

The fitting of the data may be done in several ways. We present here the two methods that we used.

Method 1 (M1) Eventually the most important time series is y_d , which represents the time evolution of the number of *active cases*. Therefore we first determine the parameters p and λ by fitting $y(t)$ to y_d (with the initial conditions as specified above). The results are plotted in Fig. 1 of the main text.

We notice that while some of the details may not be fitted by this method, the calculated functions reproduce the general behavior of the time series. The details may be adjusted by assuming that p depends on time (because of the restrictions imposed and the reaction of the population), whereas λ depends both on the time of the infection and on the time elapsed from the infection (because of the time variation of the health care system efficiency and of the human body response). Using such adjustments, one may expect to be able to fit the time series $y(t)$, but the real problem of this method is that the same parameters should be used to calculate the estimation of the other time series, namely of $s(t)$ and $r(t)$. It is shown in Fig. 2 of the main text that the difference between $r(t)$ and the data $y_d(t)$ is very big—in relevant time intervals, the estimation is more than one (or even two) orders of magnitude bigger than the reported values. Therefore, by adjusting the parameters to fit the y_d time series in more detail is not expected to make such a huge difference in the values of $r(t)$ to compensate for the differences observed in Fig. 2.

We have to find another solution.

Method 2 (M2) As a second method, we try to fit the time series of the *confirmed cases* $y_d + r_d$. We applied again the least squares method and we obtained the results presented in Fig. 3 in the main text. Even though the calculations $y(t) + r(t)$ are similar to the data $y_d + r_d$, this time the active cases $y(t)$ differ by orders of magnitude from the time series y_d , as one can see in Fig. 4 (main text). As in the case of M1, the fine adjustments of p and λ to fit the time series of *confirmed cases* cannot compensate for the big differences obtained for the *active cases*.

We also tried to fit simultaneously y_d and r_d , by applying the least square method to both sets, but by doing so, in the best case we could fit only one of the time series. Visually, in general this method led to worse results for both, y_d and r_d , whereas quantitatively, the time series of bigger numbers was better fit than the time series consisting of the smaller numbers.

2 Fitting with the modified SIR model—SIR2

In order to fit all the observed time series, we had to make qualitatively important adjustments to the model, as explained in Section 3 of the main text. We split the compartment I of the SIR model into I_1 and I_2 , where I_1

represents the number of confirmed active cases, whereas I_2 is the number of infected people which are not known and therefore are missed from the statistics. Similarly, R is split into R_1 and R_2 , R_1 representing the known closed cases (*recovered persons* plus *deaths*) and R_2 being the number of closed cases which are missing from the statistics. As in the SIR model, we define the densities $s \equiv S/N$, $y_1 = I_1/N$, $y_2 = I_2/N$, $r_1 = R_1/N$, and $r_2 = R_2/N$, whereas the rest of the parameters are defined in the main text. This model allowed us to fit quite well all the time series for most of the countries, as one can see in the main text and in Fig. S2 of this Supplemental Material, but there are also some countries and regions that cannot be fitted so well, like Hubei (China), South Korea (see main text), and the Netherlands (Fig. S2).

We also notice that the susceptible population decreases faster in some of the cases, which means that the dynamics of y_1 and r_1 in the time interval when $s(t) \ll 1$ is mainly due to the crossing from I_2 to I_1 (at rate ξ) and recovery from I_1 to R_1 (at rate λ_1). The oscillations and discontinuities of s , when $s \ll 1$, are numerical artifacts, since $s \geq 0$ and according to Eq. (2a) of the main text $\dot{s} < 0$ for any t . These artifacts are irrelevant for the time evolution of the other quantities because $s \ll 1$.

References

- [1] W. O. Kermack, A. G. McKendrick, and G. T. Walker. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115:700, 1927.
- [2] T. Harko, F. S. N. Lobo, and M. K. M.K. Mak. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Applied Mathematics and Computation*, 236:184 – 194, 2014.
- [3] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42:599, 2000.
- [4] Dan Pîrjol. Solutie analitica pentru modelul SIR. private communication.

Declarations:

The authors declare no competing interests.

Figure 2: (S2) Examples of time series of COVID19 infections fitted with the modified model SIR2. As the total populations, we used the values found on Wikipedia: Italy - $N = 60.36 \times 10^6$, Spain - $N = 46.94 \times 10^6$, the Netherlands - $N = 5.518 \times 10^6$, the United States - $N = 328 \times 10^6$, and Brazil - $N = 209.5 \times 10^6$. We see that in most cases, the fit is quite good (eventually with the exception of the Netherlands).

Figures

Figure 1

The Active cases $y(t)$ of COVID19 for four examples of countries (blue circles) and the SIR model fit (red curve). The parameters are indicated in each plot. In each case, we use the data from the day in which the number of confirmed cases exceeded 100, until the date indicated.

Figure 2

The Closed cases $r(t)$. The model is fitted to the time series of the Active cases, as shown in Fig. 1.

Figure 3

The number of confirmed cases $s(t) = y(t) + r(t)$ and the fit with the SIR model in four populations: Romania, Hubei (China), Germany, and South Korea. The fit parameters and the time interval from which the data is taken are shown in the plots.

Figure 4

The time series of the Active cases and the SIR predictions, for the parameters obtained by fitting the Confirmed cases with the SIR model as shown in Fig. 3.

Figure 5

The schematic representation of the extension of the SIR model, denoted here as SIR2 model.

Figure 6

The fitting of the SIR2 model to the data in Romania, for the indicated period. The obtained parameters are shown in the plots.

Figure 7

The fitting of the SIR2 model to the data in Germany, for the indicated period. The obtained parameters are shown in the plots.

Figure 8

The fitting of the SIR2 model to the data in Hubei, for the indicated period. The obtained parameters are shown in the plots.

Figure 9

The fitting of the SIR2 model to the data in South Korea, for the indicated period. The obtained parameters are shown in the plots. The fluctuations observed in the last plot (susceptible population) are numerical error.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [COVID19evolutionSM.pdf](#)