Modeling the COVID-19 pandemic - parameter identification and reliability of predictions

Preprint, April 07, 2020

Klaus Hackl

Institute of Mechanics of Materials, Ruhr-Universität Bochum, Germany

Address all correspondence to: klaus.hackl@rub.de

Abstract

In this paper, we try to identify the parameters in an elementary epidemic model, the so-called SI-model, via non-linear regression using data of the COVID-19 pandemic. This is done based on the data for the number of total infections and daily infections, respectively. Studying the convergence behavior of the two parameter sets obtained this way, we attempt to estimate the reliability of predictions made concerning the future course of the epidemic. We validate this procedure using data for the case numbers in China and South Korea. Then we apply it in order to find predictions for Germany, Italy and the United States. The results are encouraging, but no final judgment on the validity of the procedure can yet be made.

Keywords: COVID-19, modeling, parameter identification, predictions

1 Introduction

In most countries, social distancing measures are in effect now in order to fight the covid-19 pandemic. Considering the serious effects of these measures on the affected societies and the ensuing political discussions on their intensity and duration, it would be highly desirable to be able to make modeling based predictions on the future timeline of the epidemic, as long as the measures are upheld. Of course, many attempts are made in this direction. However, most of them require very detailed data that are laborious and time-consuming to generate.

In this work, we try to study the possibility to base predictions on data sets readily available, namely the number of reported infections. We are aware, that these numbers depend strongly on the intensity of testing done in the various countries and the reliability of the reported numbers. In this work we presume that there is a factor, country-specific,

but constant in time, between the reported and the actual number of cases. If this assumption were valid, the total number of infected individuals would be off by this very factor. However, other parameters, like the point in time when the peak in the numbers of daily infections would occur, or the following rate of decay of these numbers, would not be affected.

Finally, we would like to stress, that we intend this work to be the starting point of a discussion and maybe further research. By no means, having a background in engineering and not in virology or epidemiology, we are claiming any medical expertise. The paper should be rather seen as a general exercise in modeling and interpretation of data.

2 An elementary model

Our aim is to model a situation where social distancing measures are in effect, as currently is the case in most countries. This means, that only a small portion of the population is affected, which is well but not completely isolated from the rest. As starting point, we refer to the compartmental model by Kermack, McKendrick and Walker, [3]. It is defined by the differential equations

$$\dot{S} = -\alpha SI, \qquad \dot{I} = \alpha SI - \gamma I. \tag{1}$$

Here $I(t)$ is the number of individuals in the infectious population and $S(t)$ denotes the number of individuals in the susceptible population, in our case those who can get infected because they are not protected by social distancing. This formulation is also called the SIR-model, where the dependent variable $R(t)$ stands for the removed (by recovery of death) population, and we have $\dot{R} = \gamma S$. The parameter α is related to the basic reproduction number by

$$R_0 = NT_{\text{inf}}\,\alpha, \tag{2}$$

where N is the initially susceptible population and $T_{\text{inf}} = 1/\gamma$ is the time period during which an individual is infectious. For Sars-Cov-2, no definite value for T_{inf} has yet been reported. The parameter γ denotes the rate at which individuals are removed from the infected population because of an outcome (recovery or death).

In our study, we are going to neglect the term γI, which will have an effect only in late stages of the epidemic when $S \approx \gamma/\alpha$. As a result, we obtain the so-called SI-model, [4]. Later, we will see, that it is basically impossible to identify the parameter γ from the data available unless the epidemic is in a very late stage. So, for our purposes, this simplification amounts to a necessity.

Employing the assumption above, Eqs. (1) become equivatent to the so-called logistic

differential equation having the closed form solution

$$S(t) = \frac{I_{\max}\, e^{I_{\max}\alpha t_{\max}}}{e^{I_{\max}\alpha t_{\max}} + e^{I_{\max}\alpha t}}, \qquad I(t) = \frac{I_{\max}\, e^{I_{\max}\alpha t}}{e^{I_{\max}\alpha t_{\max}} + e^{I_{\max}\alpha t}}, \tag{3}$$

where

$$I_{\max} = I(\infty) = S_{\max} = S(-\infty) = N, \tag{4}$$

and $t_{\max}$, marking the peak of the epidemic, is defined by

$$S(t_{\max}) = I(t_{\max}) = \frac{1}{2}\, I_{\max}. \tag{5}$$

3 Parameter identification

In order to achieve a more robust parameter identification, we precondition our solution by introducing new parameters a, b given by

$$a = \alpha\, I_{\max}^2, \qquad b = \alpha\, I_{\max}. \tag{6}$$

Note, that Eq. (6) implies

$$I_{\max} = \frac{a}{b}, \qquad R_0 = T_{\inf} b. \tag{7}$$

After substitution of Eqs. (6) into Eqs. (3), the number of total infections is then given as

$$I_{a,b,t_{\max}}(t) = \frac{a}{b}\, \frac{e^{bt}}{e^{bt_{\max}} + e^{bt}}, \tag{8}$$

and the rate of daily infections becomes

$$\Delta I_{a,b,t_{\max}}(t) = \frac{\mathrm{d}}{\mathrm{d}t} I_{a,b,t_{\max}}(t) = a\, \frac{e^{b(t+t_{\max})}}{(e^{bt_{\max}} + e^{bt})^2}. \tag{9}$$

We determine the three parameters $\{a, b, t_{\max}\}$ of our model via non-linear regression. The data taken from the worldometer web page, [1], which essentially uses the data from the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE). For the parameter identification done in this paper, we have used the available data up to including Apr. 4, 2020. The data are provided in form of lists $\{(t_1, I_1), \ldots, (t_{N_{\text{data}}}, I_{N_{\text{data}}})\}$ for the total number of infections up to day t_i, and $\{(t_1, \Delta I_1), \ldots, (t_{N_{\text{data}}}, \Delta I_{N_{\text{data}}})\}$ for the number of daily infections. Time is measured in days, starting on Jan. 1, 2020. Hence, $t = 1\,\mathrm{d}$ corresponds to Jan. 1, $t = 32\,\mathrm{d}$ to Feb. 1, $t = 61\,\mathrm{d}$ to Mar. 1, 2020, and so on. Obviously, we have

$$I_i = \sum_{j=1}^{i} \Delta I_j. \tag{10}$$

Let us define errors $e_0(a,b,t_{\max})$ with respect to the total cases and $e_1(a,b,t_{\max})$ with respect to the daily cases by

$$e_0(a,b,t_{\max})^2 = \sum_{i=1}^{N_{\text{data}}} \left(I_{a,b,t_{\max}}(t_i) - I_i\right)^2, \quad e_1(a,b,t_{\max})^2 = \sum_{i=1}^{N_{\text{data}}} \left(\Delta I_{a,b,t_{\max}}(t_i) - \Delta I_i\right)^2. \tag{11}$$

In order to judge the accuracy of the modeling, let us define the data norms

$$n_0^2 = \sum_{i=1}^{N_{\text{data}}} I_i^2, \quad n_1^2 = \sum_{i=1}^{N_{\text{data}}} \Delta I_i^2, \tag{12}$$

and the relative errors

$$e_{0,\text{rel}}(a,b,t_{\max}) = \frac{e_0(a,b,t_{\max})}{n_0}, \quad e_{1,\text{rel}}(a,b,t_{\max}) = \frac{e_1(a,b,t_{\max})}{n_1}. \tag{13}$$

Finally, we find the parameters $a, b, t_{\max}$ by minimizing the errors:

$$\{a^0, b^0, t^0_{\max}\} = \text{argmin}\left\{ e_0(a,b,t_{\max})^2 \,\middle|\, a,b,t_{\max} \right\}, \tag{14}$$

$$\{a^1, b^1, t^1_{\max}\} = \text{argmin}\left\{ e_1(a,b,t_{\max})^2 \,\middle|\, a,b,t_{\max} \right\}. \tag{15}$$

Minimization is done using the computer algebra system *Mathematica*, [2]. For our purposes, the *simulated annealing* global minimization algorithm works best. Attention has to be given, though, to choosing appropriate initial intervals for the parameters in order to achieve convergence.

4 Results

In Figs. 1 to 5, the numbers of daily cases (left) and total cases (right) are plotted versus time in days. In order to get an estimation of the variability of the predictions, we use both parameter identification schemes defined in Eqs, (14) and (15). The results obtained by fitting the number of total cases according to Eq. (14) are shown in red color and those obtained obtained by fitting the number of daily cases according to Eq. (15) are shown in magenta. The corresponding data are shown in blue color.

In Fig. 1 and Fig. 2 the data for China and South Korea are displayed. Both countries can be considered to be in a late stage of the epidemic and the data are matched well by the model. Especially for China, the predictions obtained by both parameter identification schemes are close together. The pronounced spike in the number of daily cases is due to a change of the procedure how infections are counted, and is averaged out by the model. It is apparent that the model cannot fit the remaining almost constant level of daily infections around 100 and the corresponding ongoing rise in the total cases in the South Korea data. This causes the predictions generated by both procedures to lie a little

further apart.

In Figs. 3, 4 and 5, the numbers of daily new cases are plotted for Germany, Italy and the United States. These countries can be considered to be in earlier stages of the epidemic. For Germany and Italy, the predictions generated by both procedures agree closely in the time range where data are already available and are divergent for later points in time. This divergence is especially strong in case of the United States data, indicating a very dynamic epidemic process of exponential growth taking place there.

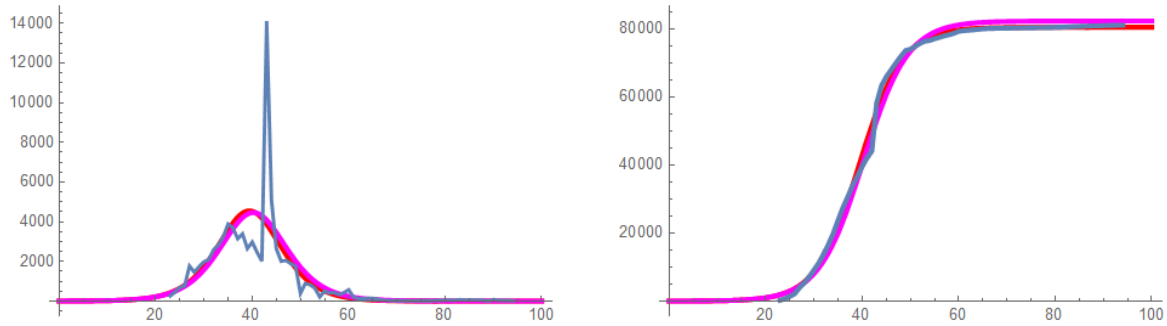

Figure 1: China, left: daily cases ($\Delta I_{a^0,b^0,t^0_{\max}}(t)$ red, $\Delta I_{a^1,b^1,t^1_{\max}}(t)$ magenta), right: total cases ($I_{a^0,b^0,t^0_{\max}}(t)$ in red, $I_{a^1,b^1,t^1_{\max}}(t)$ in magenta), data in blue

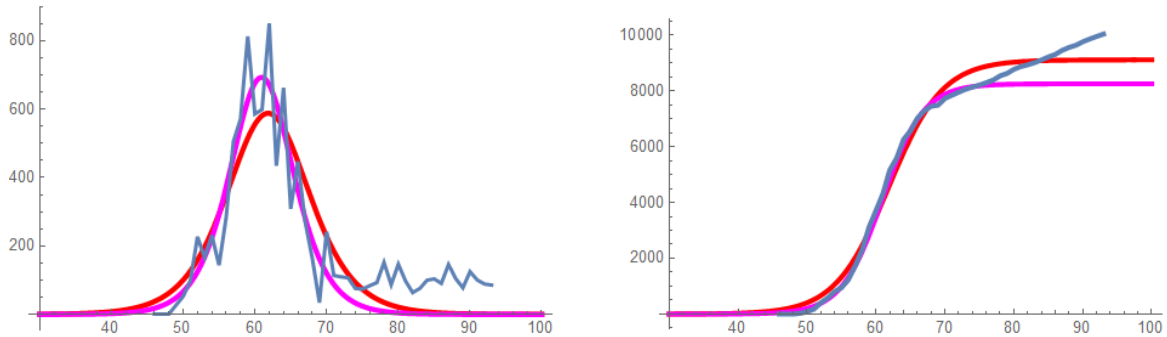

Figure 2: South Korea, left: daily cases ($\Delta I_{a^0,b^0,t^0_{\max}}(t)$ red, $\Delta I_{a^1,b^1,t^1_{\max}}(t)$ magenta), right: total cases ($I_{a^0,b^0,t^0_{\max}}(t)$ in red, $I_{a^1,b^1,t^1_{\max}}(t)$ in magenta), data in blue

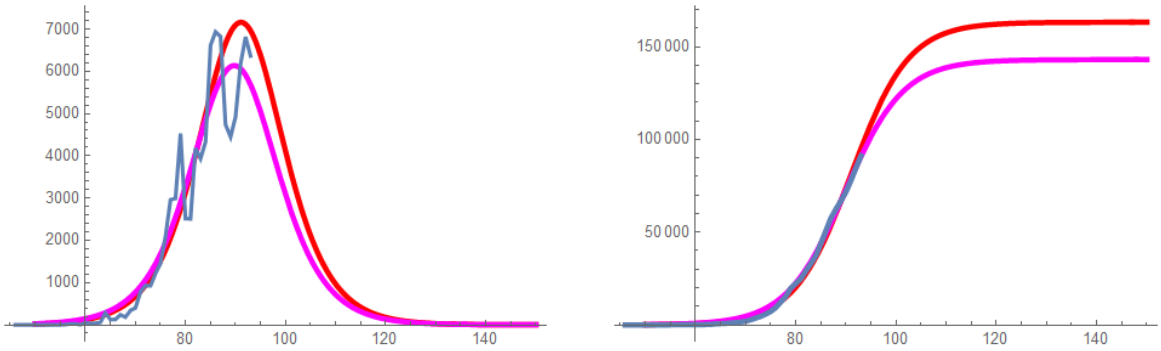

Figure 3: Germany, left: left: daily cases ($\Delta I_{a^0,b^0,t^0_{\max}}(t)$ red, $\Delta I_{a^1,b^1,t^1_{\max}}(t)$ magenta), right: total cases ($I_{a^0,b^0,t^0_{\max}}(t)$ in red, $I_{a^1,b^1,t^1_{\max}}(t)$ in magenta), data in blue

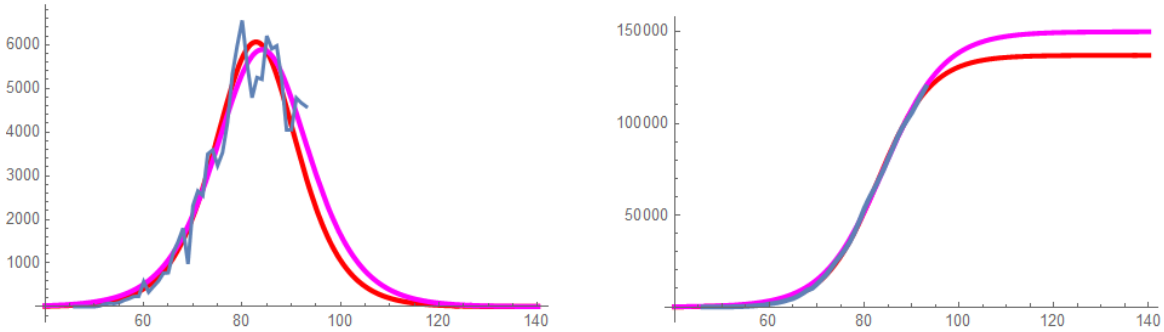

Figure 4: Italy, left: left: daily cases ($\Delta I_{a^0,b^0,t^0_{\max}}(t)$ red, $\Delta I_{a^1,b^1,t^1_{\max}}(t)$ magenta), right: total cases ($I_{a^0,b^0,t^0_{\max}}(t)$ in red, $I_{a^1,b^1,t^1_{\max}}(t)$ in magenta), data in blue

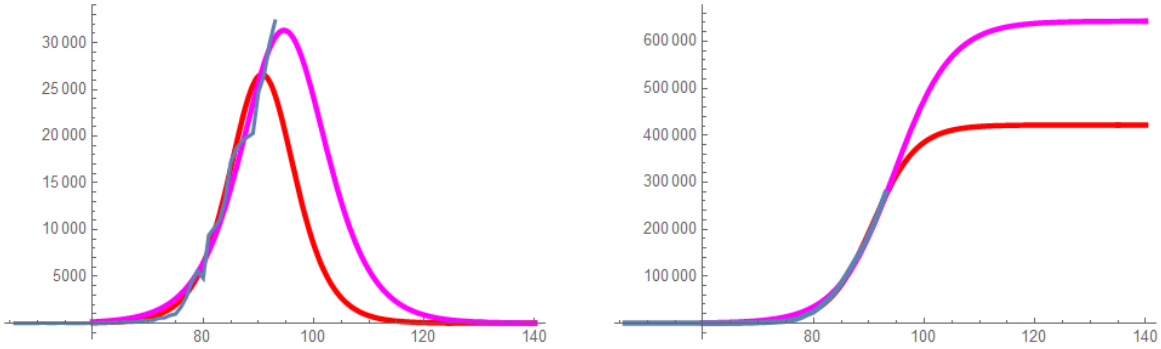

Figure 5: United States, left: left: daily cases ($\Delta I_{a^0,b^0,t^0_{\max}}(t)$ red, $\Delta I_{a^1,b^1,t^1_{\max}}(t)$ magenta), right: total cases ($I_{a^0,b^0,t^0_{\max}}(t)$ in red, $I_{a^1,b^1,t^1_{\max}}(t)$ in magenta), data in blue

Some key data provided by the model are given in Table 1. They are defined as follows:

$I^0_{\max} = a^0/b^0$, $I^1_{\max} = a^1/b^1$: This is the total number of population getting infected, if the social distancing measures taken are upheld until the epidemic has completely subsided.

Note, that for China and South Korea,these numbers correspond closely to the total number of cases reported.

T^0_{100}, T^1_{100}: This is the time predicted when the number of daily new cases will drop below 100 using both regression procedures, indicating a point in time when social distancing measures might be loosened. For Germany, Italy and the United States the values are in the range from mid of April to mid of May.

T_{data}: The dates, when the number of daily new cases dropped below 100 in China and South Korea, agreeing very well with the model data. (For South Korea, the number of daily new cases is fluctuating strongly. So we took the date, when the number dropped below 100 the second time.)

$e_{0,\text{rel}}, e_{1,\text{rel}}$: The is relative errors produced by both parameter identification procedures, as defined in Eq. (13). Note, that due to the fluctuating nature of the number of daily cases, $e_{1,\text{rel}}$ is much larger than $e_{0,\text{rel}}$.

	$I^0_{\max}$	$I^1_{\max}$	T^0_{100}	T^1_{100}	T_{data}	$e_{0,\text{rel}}$	$e_{1,\text{rel}}$
China	80442	82273	62	64	66	0.023	0.58
South Korea	9110	8251	73	70	74	0.059	0.31
Germany	163215	143074	123	121		0.052	0.21
Italy	137044	149882	113	118		0.016	0.13
United States	421846	642700	118	131		0.027	0.09

Table 1: values of key parameters on Apr. 4, 2020

5 Reliability of predictions

In the initial stages of an epidemic, the number of cases grows exponentially. This is easy to see by considering the limit $t \to -\infty$ in Eqs. (8) and (9), giving

$$I^{\text{ini}}_{a,b,t_{\max}}(t) = \frac{a}{b}\,\frac{e^{bt}}{e^{bt_{\max}}}, \qquad \Delta I^{\text{ini}}_{a,b,t_{\max}}(t) = a\,\frac{e^{bt}}{e^{bt_{\max}}}. \tag{16}$$

From Eq. (16), we see that during an early stage, it is impossible to identify the parameters a and $t_{\max}$ independently, because the occur via the common factor $a/e^{bt_{\max}}$. Hence, there are infinitely many pairs of a and $t_{\max}$ giving the same behavior and thus being minimizers in Eqs. (14) and (15). By virtue, it is even harder to fit all parameters in the SIR-model or one of the many existing extensions of it. Only past this phase of exponential growth, it is possible to identify all three parameters and thus arrive at viable predictions. But how to identify this point in time from the available data?

Our suggestion for a solution of this problem is to monitor the two different parameter

sets $\{a^0, b^0, t^0_{\max}\}$ and $\{a^1, b^1, t^1_{\max}\}$ defined in Eqs. (14) and (15). Theoretically, the values should be close to each other. However, during a phase of exponential growth, the mentioned ill-posedness of the minimization problems given by Eqs. (14) and (15) will give results lying substantially apart.

Lets test this hypothesis: In Figs. 6 to 10, we display the parameters $I^0_{\max}$ (in red) and $I^1_{\max}$ (in magenta) to the left and b^0 (in red) and b^1 (in magenta) to the right, respectively, versus time in days.

Looking at the timeline for China, we can state stable behavior starting between day 50 and 60, agreeing with the converged behavior in Fig. 1. For South Korea, we have convergence around day 66. Afterwards, the graphs diverge slightly again, which is likely due to the constant number of daily infections occurring in the later stage of the epidemic, already mentioned above. From this observation we can deduce with some caution, that the predictions based on the present model will be reliable to a certain extent as soon as the parameters identified by the two procedures stated in Eqs. (14) and (15) approach each other.

Applying this reasoning to the data for Italy, Fig. 9, we can assume reliable predictions starting on day 83. This is supported by the close values for T_{100} in Table 1. Less confidence can be put into the predictions for Germany. In Fig. 8, the graphs for $I^0_{\max}$ and $I^1_{\max}$ seem to have converged, but for b^0 and b^1, this cannot be stated with certainty. We will have to wait for the development during the upcoming few days. No convergence can be observed up to now for the United States data, Fig. 10, where we are likely still in a phase of rapid growths of the case numbers.

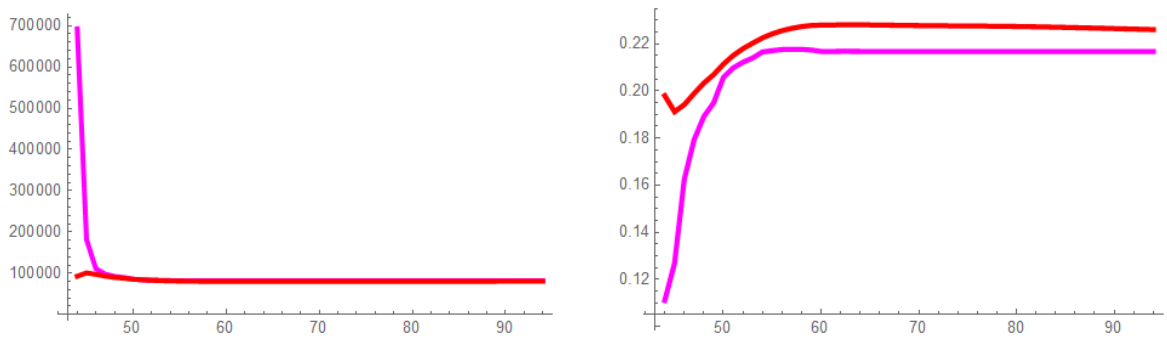

Figure 6: China, left: $I^0_{\max}$ (in red) and $I^1_{\max}$ (in magenta), right: b^0 (in red) and b^1 (in magenta)

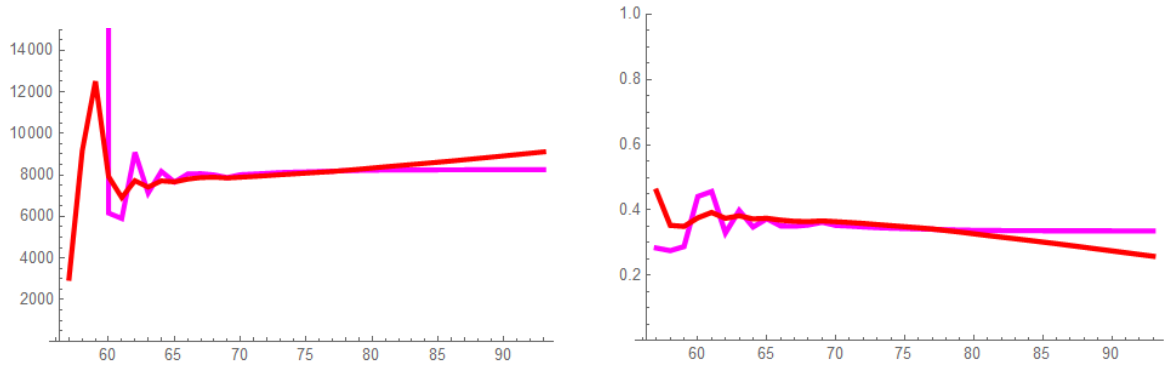

Figure 7: South Korea, left: $I^0_{\max}$ (in red) and $I^1_{\max}$ (in magenta), right: b^0 (in red) and b^1 (in magenta)

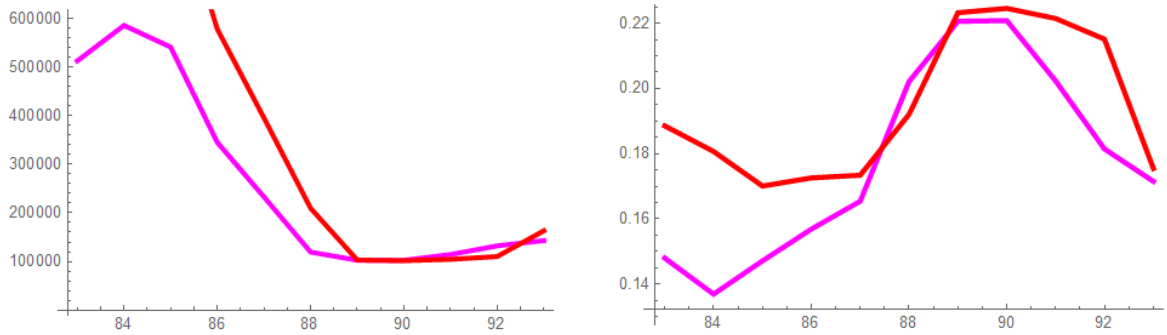

Figure 8: Germany, left: $I^0_{\max}$ (in red) and $I^1_{\max}$ (in magenta), right: b^0 (in red) and b^1 (in magenta)

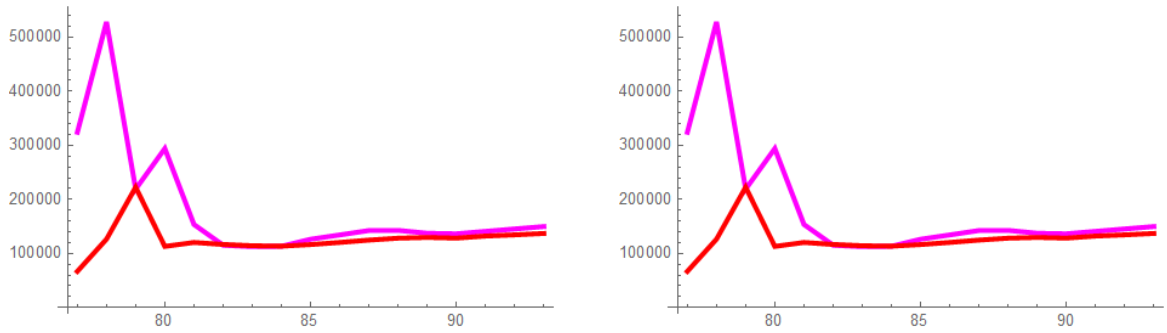

Figure 9: Italy, left: $I^0_{\max}$ (in red) and $I^1_{\max}$ (in magenta), right: b^0 (in red) and b^1 (in magenta)

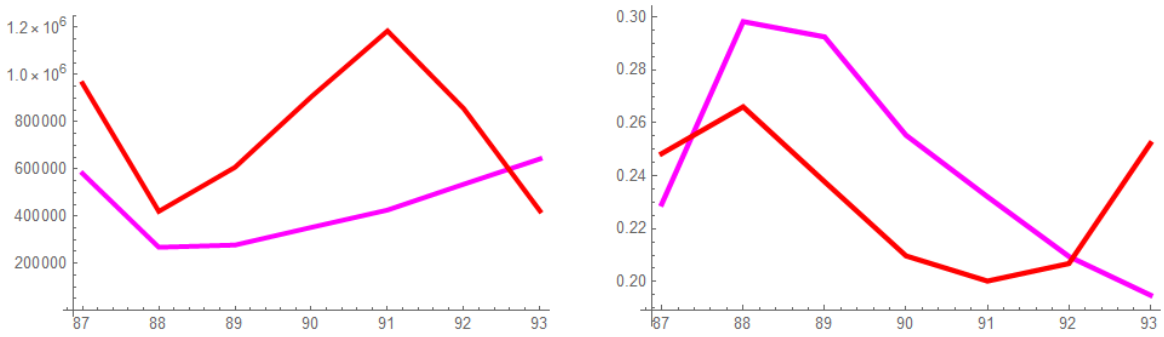

Figure 10: United States, left: $I^0_{\max}$ (in red) and $I^1_{\max}$ (in magenta), right: b^0 (in red) and b^1 (in magenta)

6 Conclusion

We have identified the parameters in an elementary epidemic model via non-linear regression using data of the covid-19 pandemic. Furthermore, we have attempted to get an insight into the reliability of predictions based on this procedure by observing the timeline of the parameters calculated by two different schemes. Our results indicate, that this approach might work. However, more detailed studies will be necessary in order to establish this method as valid. So caution is required when interpreting the results stated here. In the future, it would be desirable, too, to identify more complex models. It is uncertain, though, if this will be possible at all without more detailed data available.

References

[1] worldometer web page. `https://www.worldometers.info/coronavirus/`.

[2] Wolfram Research, Inc. Mathematica, Version 12.1. Champaign, IL, 2020.

[3] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927.

[4] James D. Murray. *Mathematical Biology*. Springer, 2002.