Stochastic Simulation of the Initial Phase of the COVID-19 Epidemic in Slovakia (version 2)

Radoslav Harman

Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava, Slovakia

April 6, 2020

1 Introduction

We propose a simulation model to roughly estimate plausible values of some directly unobservable characteristics of the spread of the infection of the COVID-19 epidemic in Slovakia. At the moment of this writing we can only use the data series of the numbers of all tests and all confirmed cases, without a single death, and very few confirmed recovered individuals.

This manuscript reflects the current state of a work in progress. The main purpose of this work is to support some modelling decisions of the governmental Institute of Health Policies of the Slovak Republic and other teams, for instance with respect to the calibration of more complex SIR-type models.

An important requirement of this model is that it is easy to implement and understand such that it can undergo scrutiny from a broader community of professionals and produce useful results at the early stages of the epidemic.

2 The numbers J_t of newly-symptomatic individuals

Define $\delta_1 = 1$, $\delta_{t+1} = \kappa_1 \delta_t$ for $t = 1, \ldots, t_0$ and $\delta_{t+1} = \kappa_2 \delta_t$ for $t = t_0 + 1, \ldots, t_{\max}$. Here, t_0 is the day of the state-wide mobility restriction and safe physical contact guidelines[1], and $t_{\max}$ is the current day (both t_0 and $t_{\max}$ are counted from the hypothetical initial day of the outbreak). The number J_t of individuals who experience the onset of the covid-induced symptoms at day t is assumed to be Poisson-distributed with expectation δ_t. Naturally, in a later phase of the epidemic the model for the number of new symptomatic individuals will

[1] In the actual simulations, we chose t_0 to correspond to 12 March 2020.

be more complex, but the piecewise-exponential trend of the initial phase seems to fit the data well.

Based on the very early data on the number of confirmed cases in Slovakia and the early number of deaths in Italy, France and Spain we roughly estimated the value of κ_1 to be 1.25. Note that in Slovakia the restriction measures have been introduced relatively early, therefore we do not expect the phase $1, \ldots, t_0$ to be very long and a precise specification of κ_1 may not be crucial. Moreover, an under- or over-estimated κ_1 can be compensated by a slight shift of the effective date of the epidemic outbreak.

The second exponential parameter, κ_2 and the time $t_{\max}$ are at the moment considered to be free simulation parameters to be calibrated to fit the actual data. Note that t_0 is completely determined by $t_{\max}$, the known current date, and the known date of the state-wide restrictions.

3 The time course of the covid-like symptoms

A second general model assumption is that each covid-infected symptomatic individual exhibits symptoms that (greatly) vary in time and intensity. For each newly-infected individual i we first generate its age-decade ($0-10$ years, ..., $80-89$ years) from the known age distribution in Slovakia[2]. Second, the probability $d^{(i)}$ of covid-induced death is based on the age of i and the estimated death rates observed in the countries with earlier onsets of the epidemic[3].

At the moment of this writing, there are only rough quantitative results on the time course of the COVID-19 symptoms; searching the early literature we decided to model the course as a piece-wise linear triangular curve, non-zero at times $t_b^{(i)}, \ldots, t_e^{(i)}$ and culminating in the middle of the interval. The magnitude $m^{(i)}$ of the peak was modeled using the beta distribution $B(1, b^{(i)})$, where $b^{(i)}$ has been computed such that $P[B(1, b^{(i)}) > 0.5] = d^{(i)}$. Note that the beta distribution with the first shape parameter equal to 1 assigns a large proportion of the probability to values close to 0, and the proportion increases with increasing $b^{(i)}$.[4]

Based on the observations in other countries, the number of days elapsed from the onset of symptoms of i to the culmination has been modeled as $28m^{(i)} + U^{(i)}$, where $U^{(i)}$ is uniformly distributed on the interval $(0, 14)$. This means that a very mild peak symptoms lead, in general, to an earlier recovery (roughly within 14 days from the day of the infection), but the possible death occurs usually between 14 and 28 days. Recovery of critical cases (peak magnitude of symptoms close to 0.5) may take even much longer according to the model.

[2]The age of i is generated independently of the ages of other infected individuals; this model takes the overall age structure into account, but not age-related differences in interactions between people.

[3]Note that in Slovakia, these rates may be different and need to be adjusted once we have more data.

[4]The model does not take the pre-symptomatic phase of the infection into account, because it models only the numbers of symptomatic individuals, and does not attempt to model interactions between people.

Of course, this crude approximation has the potential to be improved based on further knowledge about the disease.

4 The number of detected cases of COVID-19

The numbers C_t of confirmed cases strongly depend on the (known) numbers T_t of the tests performed at days $t = 1, \ldots, t_{\max}$. However, another crucial (and unknown) parameter is how efficiently does the system select people for testing. Unfortunately, people often report symptoms resembling those of COVID-19 even in the case of no covid infection, for instance in the case of flu.

In our model, we assume that the covid-like symptoms in the non-covid population occur according to the beta distribution $B(1, b_0)$, where b_0 is the third free parameter of the simulation, effectively measuring the overall quality of the selection process for testing[5] The actual number of detected cases at day t is then computed from pooling the magnitudes of symptoms of the entire Slovak population (excluding the individuals that have already been positively tested), and selecting those T_t individuals for testing that exhibit the strongest covid-like symptoms[6]).

5 Calibrating parameters $\kappa_2, t_{\max}, b_0$

Each fixed triple of parameters $\kappa_2, t_{\max}, b_0$ can be used to simulate the time series of confirmed cases. The fit with the actual confirmed cases can then be used to estimate the value of these parameters, and consequently other directly unobservable quantities such as the real overall number of symptomatic covid-infected individuals. We measured the error of the fit via the expectation of the log-likelihood

$$-\sum_{t=1}^{t_{\max}} \ln P[Z_t(\kappa_2, t_{\max}, b_0), C_t],$$

where $Z_t(\kappa_2, t_{\max}, b_0)$ is the random variable representing the simulated cumulative number of detected cases until day t and C_t is the real cumulative number of detected cases. For two nonnegative integer numbers z and c, the probability is taken to be

$$P[z, c] = \exp(-2\lambda)\frac{\lambda^{2\lambda}}{z!c!},$$

where $\lambda = \frac{z+c}{2}$. That is, $P[z, c]$ is the probability that two independent random variables distributed according to $Po(\lambda)$ attain values z and c, respectively.

[5]That is, the greater b_0, the better the efficiency of the system in restricting non-covid individual from testing, the smaller b_0, the more non-covid individuals interfere with real covid infections.

[6]In the code, this process is not simulated, but directly computed using the distribution function of $B(1, b_0)$. The actual simulation of the covid-like symptoms for the entire Slovak population would be computationally too demanding.

For the actual computations, note that the logarithm of the factorial can be very well approximated using the truncated Stirling series

$$\ln(n!) \approx n\ln(n) - n + \frac{1}{2}\ln(2\pi n)$$

Note that it is also possible to use a "normalized" version

$$-\sum_{t=1}^{t_{\max}} \ln \frac{P[Z_t(\kappa_2, t_{\max}, b_0), C_t]}{P[\lambda_t(\kappa_2, t_{\max}, b_0), \lambda_t(\kappa_2, t_{\max}, b_0)]} =$$

$$\sum_{t=1}^{t_{\max}} \Big[\ln(Z_t(\kappa_2, t_{\max}, b_0)!) + \ln(C_t!) - 2\ln(\Gamma(\lambda_t(\kappa_2, t_{\max}, b_0) + 1)) \Big],$$

where $\lambda_t(\kappa_2, t_{\max}, b_0)$ is the arithmetic mean of $Z_t(\kappa_2, t_{\max}, b_0)$ and C_t[7]. The advantage is that it provides the error value of 0 if and only if the two time series $\{Z_t(\kappa_2, t_{\max}, b_0)\}_t$ and $\{C_t\}_t$ coincide.

We observed[8] that for each fixed κ_2, the simulated curves Z_t exhibit the best fit with the observed C_t along a convex curve of values $b_0, t_{\max}$; see the top-left panels of Figures 1-3 corresponding to $\kappa_2 = 0.95, 1.00$, and 1.05 respectively. We remark that for $\kappa_2 < 0.93$ and $\kappa_2 > 1.07$, the fit significantly deteriorates (for all values of $b_0, t_{\max}$). Since the data is scarce, the implied range of plausible values of all parameters is relatively wide.

For the parameters that fit the data reasonably well, the current numbers of all symptomatic individuals infected with COVID-19 (including a few recovered individuals) are estimated to be in the range 800 to 1500, although the larger number is only compatible with the number of deaths that is much greater than reported.

6 Interface with other models

The crucial problem of more complex models of the epidemic spread is their calibration. This simple simulation model can be used to support the calibration as follows:

1. Find a set of plausible parameters of the simulation model and the most likely scenarios that they produce. Try to capture the entire space of plausible parameter values.[9]

2. From the scenarios, extract the trajectories of the time series $(I_n)_n$ of (real, unobserved) symptomatic infected individuals. Use these trajectories to calibrate the more the advanced model(s).

[7] The factorial of a possibly non-integer positive number λ id replaced by $\Gamma(\lambda + 1)$.

[8] We used the times series of the numbers of tests and confirmed cases available until 27 March 2020. The codes are temporarily available at `http://www.iam.fmph.uniba.sk/ospm/Harman/programs/corona_sim.txt`.

[9] Of course, it is also possible to use the Bayesian approach and estimate the posterior distribution on the parameters. This is, however, unrealistic at the moment.

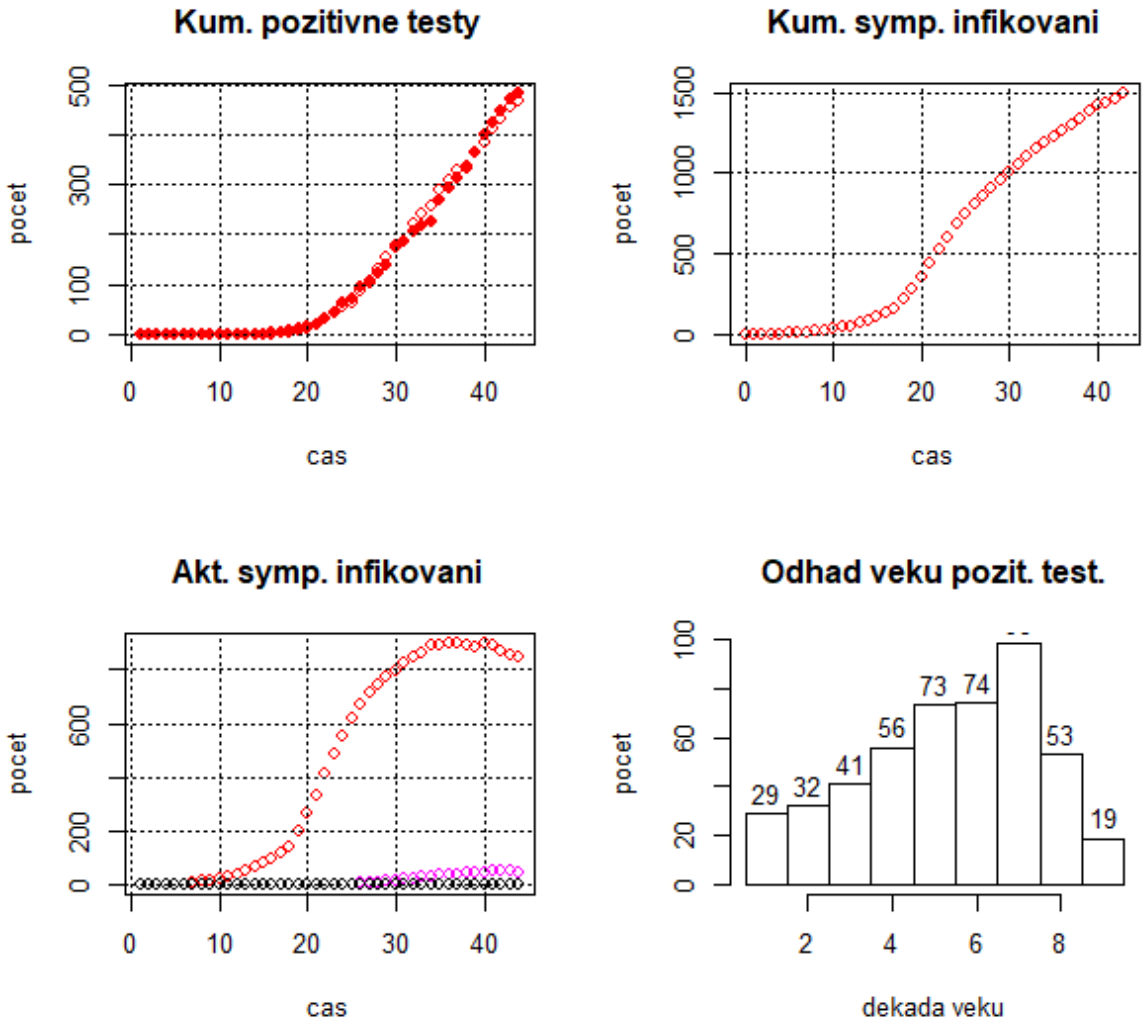

Figure 1: Results of simulation runs for $\kappa_2 = 0.95$, $b_0 = 80$, $t_{\max} = 43$. In this possible scenario, the total number of symptomatic COVID-19 infected individuals is estimated to be 1500 (top-right panel) and the actual number of infected individuals (bottom-left panel) is slowly decreasing. Most confirmed infected individuals are between the ages 60 and 70.

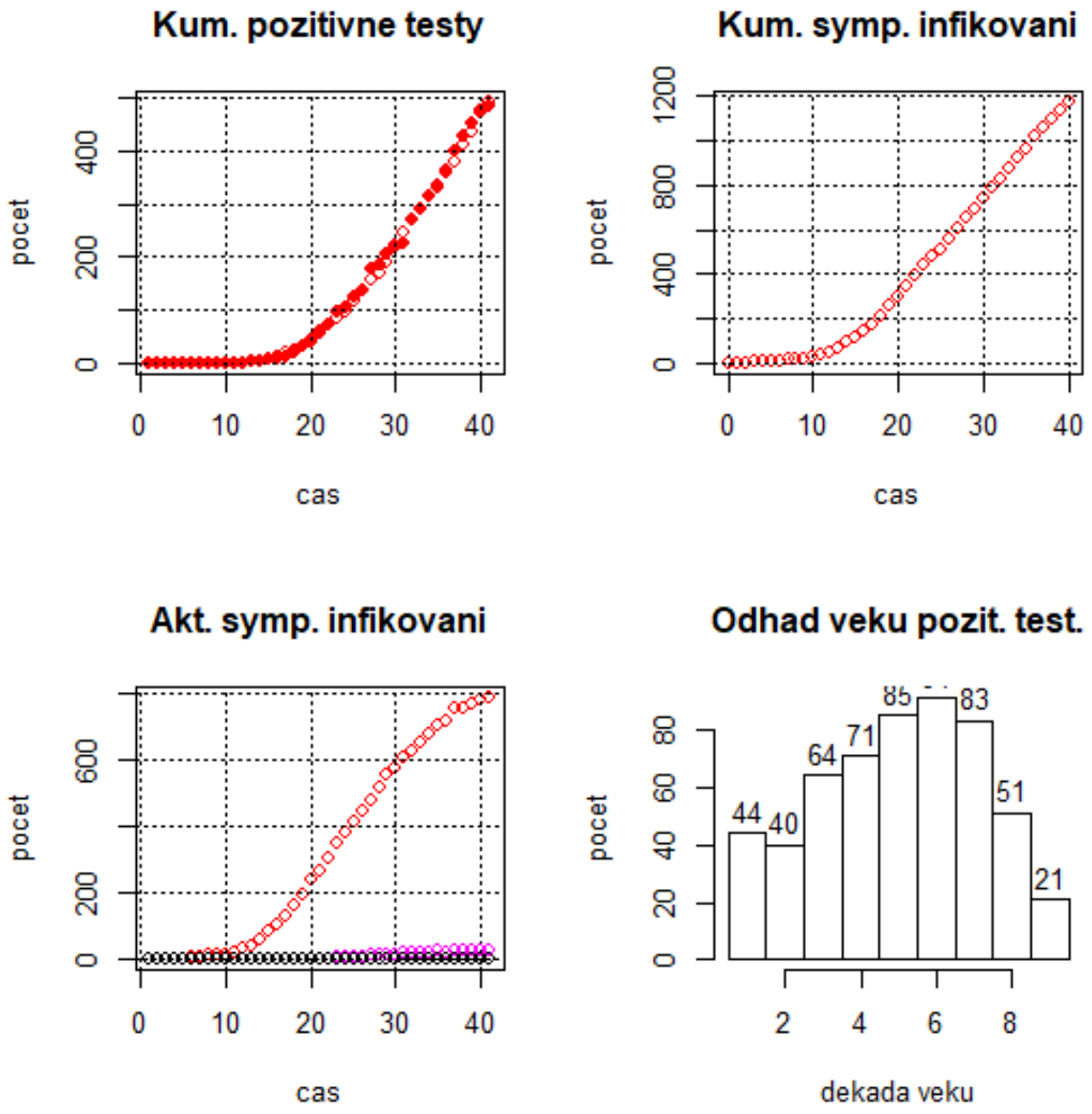

Figure 2: Results of simulation runs for $\kappa_2 = 1.00$, $b_0 = 120$, $t_{\max} = 41$. In this possible scenario, the total number of symptomatic COVID-19 infected individuals is estimated to be 1200 (top-right panel) and the actual number of infected individuals (bottom-left panel) is approaching the culmination point. Most confirmed infected individuals are between the ages 50 and 60.

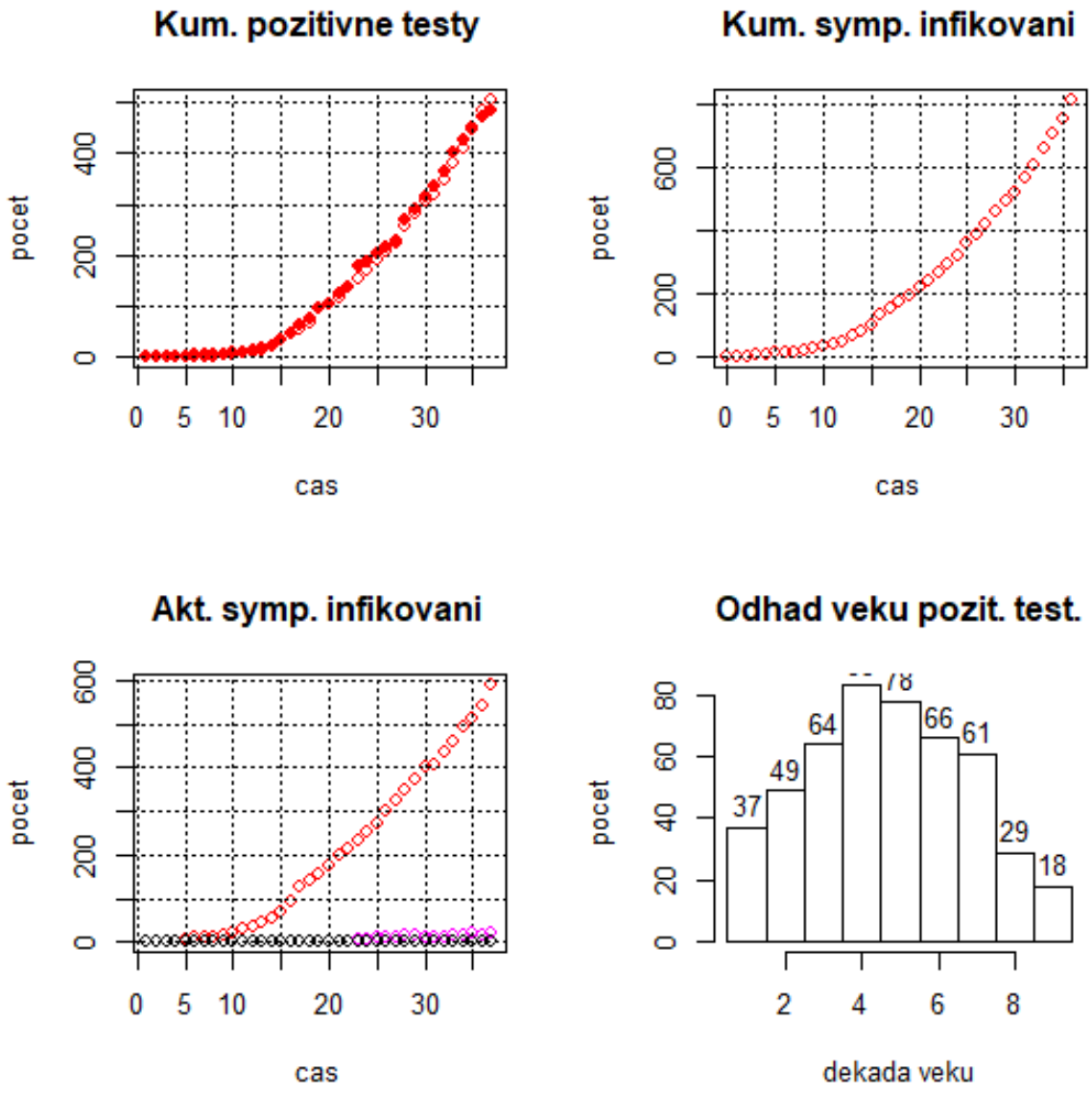

Figure 3: Results of simulation runs for $\kappa_2 = 1.05$, $b_0 = 280$, $t_{\max} = 37$. In this possible scenario, the total number of symptomatic COVID-19 infected individuals is estimated to be 800 (top-right panel) and the actual number of infected individuals (bottom-left panel) is still significantly growing. Most confirmed infected individuals are between the ages 30 and 50.

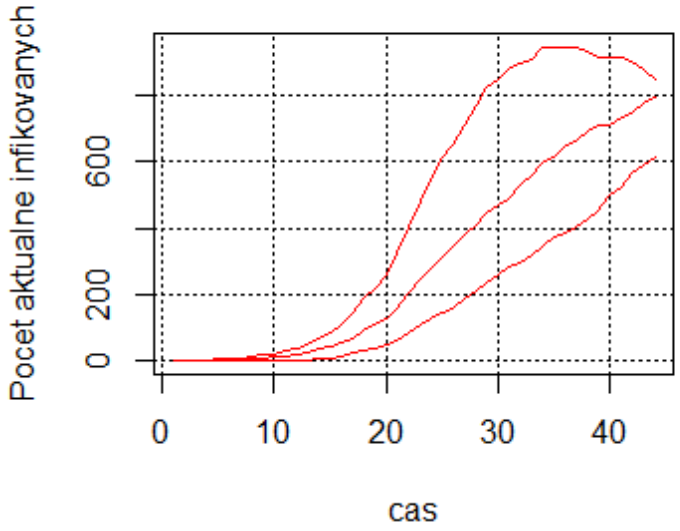

Figure 4: An example of 3 possible simulated trajectories of the unobserved time series $(I_t)_t$. The top trajectory corresponds to $\kappa = 0.95$ and the corresponding most likely parameters $b0$ and $t_{\max}$. The bottom trajectory corresponds to $\kappa = 1.05$ and the corresponding most likely parameters $b0$ and $t_{\max}$.

An example of three possible trajectories (the most likely trajectory for two "boundary" plausible sets of parameters and for an "intermediate" set of parameters) are depicted in Figure 4.

7 Limitations

The model has many limitations and obvious suggestions for improvement; a few of them are listed below.

7.1 Hard-to-overcome limitations

The model does not utilize the geographic heterogeneity of Slovakia and does not take the structure of contacts between people into account (at least not explicitly). For instance, super-spreaders may unexpectedly render the model for J_t invalid. The model cannot capture the individuals that simultaneously 1) do not seek testing, and 2) do not significantly contribute to the spread of the disease. The model assumes that the selection process of symptomatic individuals for testing is primarily based on the magnitude of symptoms, not (for instance) on the history of contacts with a detected case. The model is largely in-

validated if some group of people is systematically denied testing despite the presence of symptoms. The model does not directly operate with the standard composite statistical measure R_0 and the application of the simulated results to more standard epidemic models may not be completely straightforward.

7.2 What can be improved with more data and knowledge

The onset of the epidemic may be better modeled using a different parametrized curve[10]. The estimates and potential predictions based on calibrating the model on the data may be sensitive with respect to the assumed model on the real curve of J_t. The impact of the restrictions may be gradual (the model assumes a sudden change). In Slovakia, the age-depended lethality may be different than what those observed in other countries (that the model currently uses). The model does not provide an estimate of the number of infected people requiring the ICU and how many of them require ventilators. The time-course and the magnitude of the COVID-19 symptoms is at the moment modelled only roughly. The model assumes that the tests can reliably determine the presence or the absence of the COVID-19 infection. The model assumes that the number of infected individuals is very small compared to the number of non-infected and still susceptible.

Despite of the numerous limitations, the model exhibits a very good fit with the data (for some triples of parameters). This may make it a reasonable choice for a rough estimation of the underlying reality of the epidemic.

8 Acknowledgments

RNDr. Zuzana Chladná, Dr., Mgr. Lenka Filová, PhD., doc. RNDr. Iveta Waczulíková, PhD., Mgr. Lukás Polácek, Mgr. Samuel Rosa, PhD., Mgr. Samuel Hudec, PhD.

9 References

To be added.

[10]It will definitely need to be corrected at later stages.