Epidemiological modeling based on coronavirus data

Monika Szymura [1], Martyna Horst [1] and Anna Altmann [1]

[1] *Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44100 Gliwice, POLAND*

Abstract

The article presents selected compartmental epidemiological models. The SI model is discussed, which is considered to be the most basic epidemiological compartmental model. The SVIR model was also analyzed, which additionally takes into account e.g. use of vaccinations. In addition to theoretical elaboration, the article also includes simulations based on real coronavirus pandemic data in selected countries. The obtained results show that on the basis of real data from a certain period of time, it is possible to predict the trend of the further course of infection in a given population.

Keywords

SARS-Cov-2, epidemic, mathematical model, differential equation, epidemiological model

1. Introduction

Mathematics is omnipresent in our lives. Moreover, most problems with a natural basis are based on mathematical models. It is important to realize how significant a role both mathematics and informatics play in medicine, for example, in the process of diagnosing patients [4,19]. Nowadays, we are facing the problem of the coronavirus epidemic (COVID-19), that is an infectious disease of the respiratory system caused by infection with the SARS-Cov-2 virus [10].

Mathematical models play an extremely important role in many fields. They are created for various needs, as can be seen in [8, 9, 18, 20]. Also for the purpose of fighting the epidemic, many mathematical models and simulations have been created to help the authorities of countries around the world defeat the invisible enemy [16]. Models created to fight various epidemics are called epidemiological models. However, the field of epidemiology was developing much earlier. The publication written by W.O. Kermack and A.G. McKendrick [12] from 1927 is considered a breakthrough in the field of epidemiological modeling. However, this science continued to develop. This is evidenced, among others, by the publications of Dieckmann-Heesterbeek [5] or Murray [14, 15].

In epidemiological modeling, we distinguish deterministic and stochastic models [1]. Very often we can come across models based on ordinary differential equations that describe the dynamics of epidemic development [6, 7, 16]. There are many different methods for solving equations and systems of differential equations [20], including, for example, the argument shift method [3, 17]. Of course, mathematics is closely related to physics. Differential equations also come from this field. They are related to the concept of the derivative of a function and the derivative is related to speed. This allows us to describe the changes of a certain quantity $x(t)$ depending on the time $[t, t + \Delta t]$ [6]. As we will see later, this is extremely important when modeling the course of various diseases.

In our article, we will focus on a few selected epidemiological models. We will also present examples of simulations conducted on the basis of statistical COVID-19 data [13, 22, 23] in selected European countries.

28th Conference on Information Society and University Studies (IVUS'2023), May 12, 2023, Kaunas, Lithuania
EMAIL: moniszy832@student.polsl.pl (M. Szymura); marthor323@student.polsl.pl (M. Horst); annawal057@student.polsl.pl (A. Altmann)

2. Compartmental epidemiological models

The models analyzed in this article are called group or compartmental models, because a given population is divided into groups (subpopulations) due to the state of health, the possibility of infecting other individuals or the level of resistance to a given disease [6]. In the simplest compartmental model, called SI, we assume that there are only healthy (i.e. susceptible to infection) and sick individuals. We also assume that the population is of constant size and not too small, and that the individuals are well mixed (evenly distributed). These assumptions are necessary because the size of groups versus time will be described by differential equations. Before the first case of the disease occurs, the entire population is included in the group of individuals at risk of infection, and each subsequent case of infection causes the size of respective groups to change [6, 7].

2.1. SI model

As already mentioned, the simplest compartmental epidemiological model is the SI model. We will therefore begin our considerations with its discussion.

In the SI model [7], we analyze a closed population of N evenly distributed individuals. There are two separable subpopulations:

- a subpopulation of individuals susceptible to infection (denoted as S),
- a subpopulation of infected individuals (denoted as I).

Figure 1 presents a diagram illustrating the possibilities of individuals to move within groups in the analyzed population. Note that individuals can only move between these groups in one direction. If an individual from subpopulation S gets sick, it automatically goes to the group of infected individuals and it is no longer possible to return to subpopulation S.

It should be noted here that the discussed SI model can be modified, taking into account the possibility of returning infected individuals to group S. Thus, an infected individual returns to subpopulation S after the disease has passed and becomes susceptible to infection again. Such a modified model can be found in the literature under the name of SIS [7].

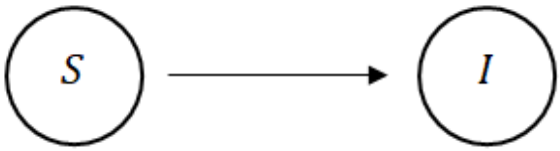

Figure 1: Scheme of movement of individuals in the SI model

Due to the fact that we are considering a population with a constant number of individuals, we must remember that it is not possible to include births or deaths of individuals from the analyzed population in this model. We must also disregard any immigration or emigration. Assuming, for example, that the initial population is the population of a given country, then in our model a resident cannot leave this population by moving to another country.

Now let's introduce the functions $S(t)$ and $I(t)$ which for a given time $t \geq 0$ express the size of subpopulations S and I, respectively. Bearing in mind that the total size of the population does not change over time, we get for any $t \geq 0$ relation:

$$S(t) + I(t) = N. \tag{1}$$

Differentiating equation (1) we get

$$\frac{dS(t)}{dt} + \frac{dI(t)}{dt} = 0. \tag{2}$$

We will take a day as the unit of time t. We will now use the previously introduced functions to describe the considered model using a system of differential equations:

$$\begin{cases} \frac{dS(t)}{dt} = -\alpha S(t)I(t), \\ \frac{dI(t)}{dt} = \alpha S(t)I(t), \end{cases} \tag{3}$$

where $S(0) > 0$, $I(0) > 0$. The $\alpha > 0$ parameter is called the infection rate. An individual from the S subpopulation can only become ill as a result of contact with an infected person. Hence, in a unit of time (i.e. in one day) one sick individual can infect $\alpha S(t)$ of susceptible individuals [6]. Based on the system of equations (3), it is easy to conclude that the equality (2) is satisfied.

From (3) we also deduce that $I(t)$ is an increasing function, while $S(t)$ is a decreasing function. This is of course due to the fact that

$$\frac{dI(t)}{dt} = \alpha S(t) I(t) \geq 0$$

and

$$\frac{dS(t)}{dt} = -\alpha S(t) I(t) \leq 0.$$

2.1.1. Example

Figure 2 shows the graph of the functions $I(t)$ and $S(t)$ over the next 100 days from the diagnosis of the first three patients in a population of 1000 individuals. We assume that the course of infection can be represented by the SI model with the parameter $\alpha = 0.00015$.

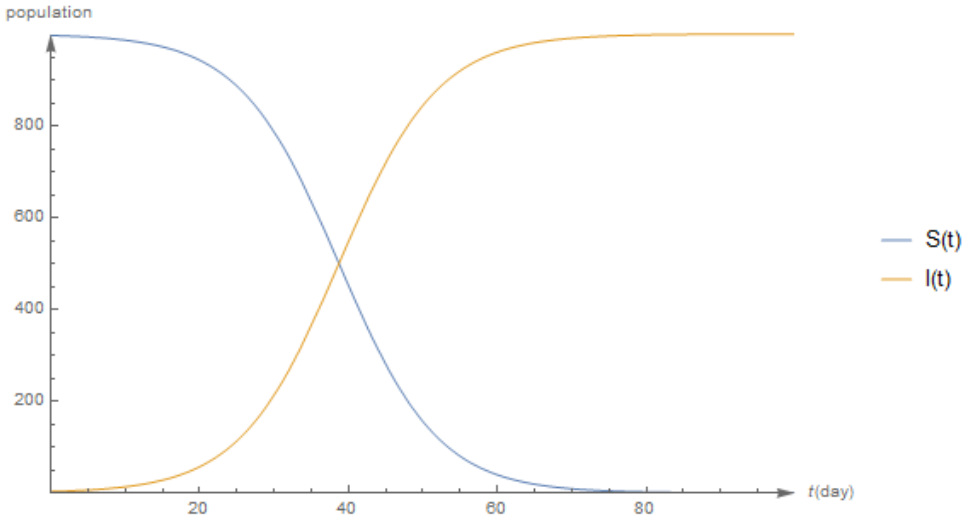

Figure 2: Epidemic simulation

2.2. SVIR model

In the SVIR model, we also consider a closed population of N individuals. The analyzed group of evenly distributed individuals is divided into the following subgroups:

- a subpopulation of individuals susceptible to infection (denoted as S),
- a subpopulation of infected individuals (denoted as I),
- a subpopulation of individuals who have already suffered from the disease (denoted as R),
- a subpopulation of vaccinated individuals (denoted as V).

It should be noted here that, as in the SI model, the above subpopulations are disjoint, that is a given individual may belong to only one of them at a given time.

It is also extremely important that due to the assumption of a constant size of the analyzed population, we omit the birth of new individuals, death or any kind of migration in our considerations.

Figure 3 shows a diagram illustrating how individuals can move between subpopulations. It is easy to see that a susceptible individual can become vaccinated and thus end up in the V subpopulation. Another possibility is that the individual becomes ill as a result of contact with an infected person. In this case, the individual is transferred to subpopulation I. Then, after recovering from the disease, the individual is transferred to subpopulation R.

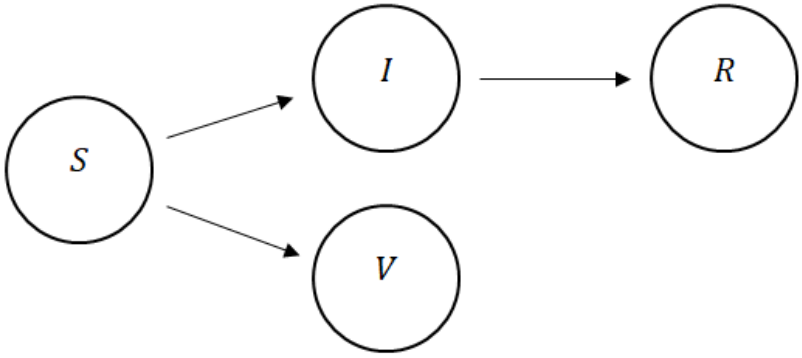

Figure 3: Scheme of movement of individuals in the SVIR model

Let us now proceed to the mathematical description of the presented model. The functions $S(t)$, $V(t)$, $I(t)$ and $R(t)$ for a given time t express the size of respective subpopulations. Thus, for example, the function $S(t)$ determines how many individuals are in subpopulation S at time t, where the time unit t is a day. According to the assumption of a constant size of the analyzed population, for any $t \geq 0$ the following equality holds:

$$S(t) + V(t) + I(t) + R(t) = N. \tag{4}$$

The SVIR model based on the assumptions presented above can be represented by a system of differential equations

$$\begin{cases} \dfrac{dS(t)}{dt} = -\alpha S(t)I(t) - \delta S(t), \\ \dfrac{dV(t)}{dt} = \delta S(t), \\ \dfrac{dI(t)}{dt} = \alpha S(t)I(t) - \gamma I(t), \\ \dfrac{dR(t)}{dt} = \gamma I(t), \end{cases} \tag{5}$$

where $S(0) > 0$, $V(0) \geq 0$, $I(0) > 0$, $R(0) \geq 0$. The $\alpha > 0$ parameter defines the infection rate. Whereas $\gamma > 0$ is the recovery rate (i.e. the frequency of leaving the I subpopulation), and $\delta > 0$ is an indicator determining the intensity of vaccination. $\frac{1}{\gamma}$ represents the average duration of infection for a single individual [6, 7].

Adding the sides of the equation of the system (5), we get

$$\frac{dS(t)}{dt} + \frac{dV(t)}{dt} + \frac{dI(t)}{dt} + \frac{dR(t)}{dt} = 0.$$

Hence and from the equality (4) we conclude that, according to the initial assumption, N is a constant quantity for each $t \geq 0$.

2.2.1. Example

Let's consider a closed population of 1000 individuals. We know that there are 10 individuals infected with a certain disease. So far, 1 person has recovered from the infection. Moreover, no one has yet been vaccinated to make the body immune to the virus that causes the disease. Let's assume that the infection follows the SVIR model for infection, recovery and vaccination rates of $\alpha = 0.0003$, $\gamma = 0.07$, $\delta = 0.01$, respectively. Figure 4 illustrates the course of the disease in the analyzed population during the next 100 days from the moment of diagnosis of ten initial patients and one recovered.

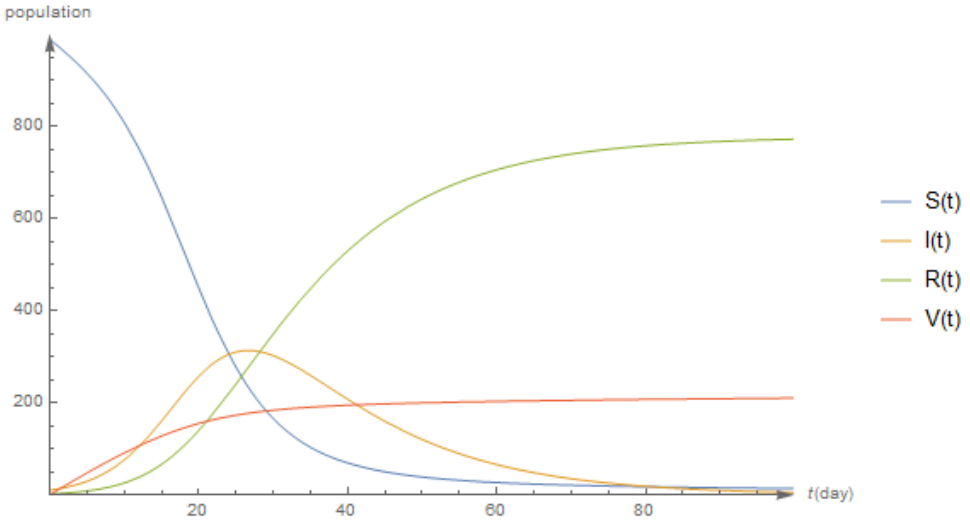

Figure 4: Epidemic simulation

3. Theil coefficient

In the further part of the article, we will use, among others, Theil's coefficient to assess the quality of the model. So let's recall the necessary information about it.

The Theil coefficient [21] is used to calculate the relative prediction error to determine the quality of the model. We use the following formula to calculate this coefficient

$$T^2 = \frac{\sum_{t=1}^{n}(y_t - y_t^p)^2}{\sum_{t=1}^{n} y_t^2}$$

where y_t is the real value a t, y_t^p is the predicted value at time t, and n is the length of the testing period. Most often we give the result as $T = \sqrt{T^2} \cdot 100\%$. How should the Theil coefficient be interpreted? The higher the value of the coefficient, the lower the quality of the model. If the coefficient is equal to 0, then the model can be said to be very well defined.

4. Simulation of the course of the coronavirus pandemic

We will run simulations within certain periods of infection in selected countries. For the purposes of the simulation, we assume that N (the size of the population in which the epidemic broke out) is equal to the total number of all confirmed cases of infection from the beginning of the outbreak to the last day of prediction (end of the considered infection period). In the case of the SVIR model, N is also affected by the number of vaccinations performed to date.

Due to the fact that the SVIR model does not take into account the phenomenon of death as a result of disease, we will assume that the deceased are included in subpopulation R (as people who acquired immunity as a result of infection) [6].

Real pandemic data comes from [13, 22, 23]. To simulate the course of the coronavirus pandemic, we will use the Mathematica software.

4.1. Simulation of the course of the coronavirus pandemic in Croatia using SI model

Let's now take a closer look at the course of the coronavirus pandemic in Croatia. We will consider the period from October 11, 2020 to February 8, 2021. We assume that $N = 235473$, because 235473 of coronavirus cases were diagnosed by February 8, 2021. We will try to select the α parameter in the SI model so that the model reflects the actual development of the epidemic in the first 31 days of the

analyzed time period as accurately as possible. For this purpose, we will use the method of least squares by minimizing the expression:

$$\sum_{t=1}^{31}(I_{real}(t) - I_{model}(t))^2 \tag{6}$$

where $I_{real}(t)$ expresses the actual number of sick people on particular days. Whereas $I_{model}(t)$ expresses the data obtained from the considered model. After minimizing, we get the information that the expression (6) takes the smallest value when $\alpha = 2.11836 \cdot 10^{-7}$.

Next, we will do ex post prediction to see if the model with the determined parameter α correctly reflects the trend of disease development during the next 90 days. For this purpose, we calculate the mean absolute percentage error $MAPE$ and Theil's coefficient T^2:

$$MAPE = \frac{1}{90}\sum_{t=32}^{121}\left|\frac{I_{real}(t) - I_{model}(t)}{I_{real}(t)}\right| \cdot 100\% \approx 3.74\%,$$

$$T^2 = \frac{\sum_{t=32}^{121}(I_{real}(t) - I_{model}(t))^2}{\sum_{t=32}^{121}(I_{real}(t))^2} \approx 0.002,$$

$$T = \sqrt{T^2} \cdot 100\% \approx 4.56\%.$$

Thus, the data on infected people obtained from the model differ from the real data by about 3.74% -- 4.56%on average. It can therefore be concluded that the SI model with the set parameter α well reflects the actual development of the epidemic in Croatia, and the error, which does not even exceed 5%, is relatively small.

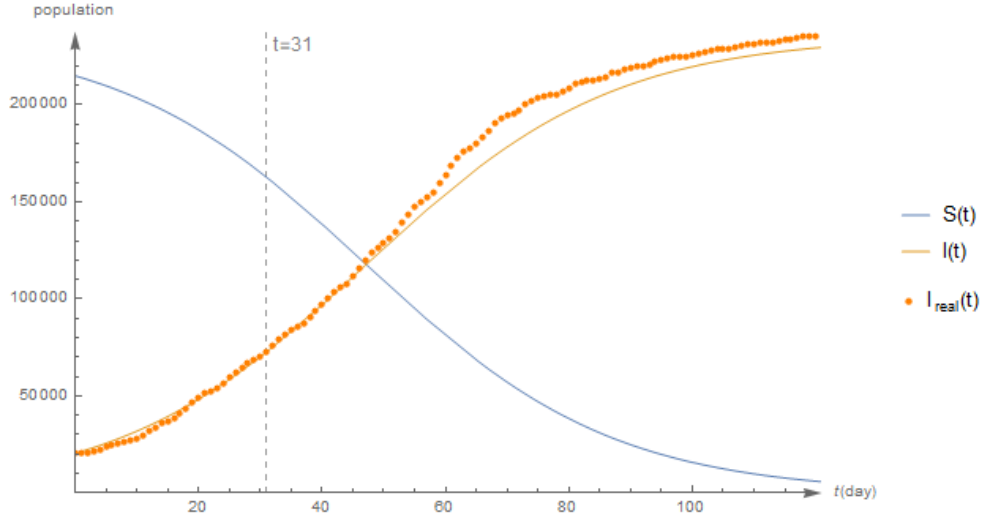

Figure 5: Epidemic simulation

A graphic illustration of the discussed simulation of the course of coronavirus infection in Croatia is presented in Figure 5.

4.2. Simulation of the course of the coronavirus pandemic in Bulgaria using SI model

In the case of Bulgaria, let's consider the time period from October 21, 2020 to January 19, 2021. We assume $N \approx 213000$. First, we will fit the model parameters to the initial 31 days. Using the least squares method, it turns out that the best fit of the real values to the model during the initial 31 days of the analyzed period occurs for the parameter $\alpha = 2.91903 \cdot 10^{-7}$.

Subsequently, we assess the quality of mapping real data by the model over the next 60 days. We get that

$$MAPE = \frac{1}{60}\sum_{t=32}^{91}\left|\frac{I_{real}(t) - I_{model}(t)}{I_{real}(t)}\right| \cdot 100\% \approx 1.18\%,$$

$$T^2 = \frac{\sum_{t=32}^{91}(I_{real}(t) - I_{model}(t))^2}{\sum_{t=32}^{121}(I_{real}(t))^2} \approx 0.00018,$$
$$T = \sqrt{T^2} \cdot 100\% \approx 1.36\%.$$

So we can see that the average error of the predictions obtained from the model does not even exceed 1.5%. Therefore, we assume that the SI model with the determined parameter α reflects very well the tendency of the disease development in Bulgaria in the considered period.

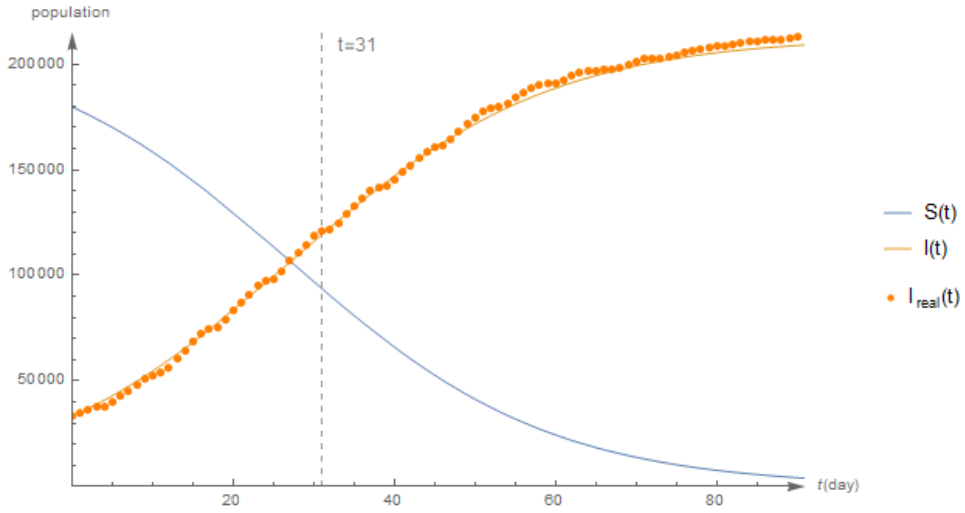

Figure 6: Epidemic simulation

4.3. Simulation of the course of the coronavirus pandemic in Denmark using the SVIR model

Using statistical data related to COVID-19 in Denmark from February 26, 2021 to April 16, 2021, we will present a simulation of the course of the pandemic in this country. We will adjust the parameters of the model to the initial 20 days, more precisely from February 26, 2021 to March 17, 2021. We assume $N \approx 1600000$, because by April 16, 2021, this is the total number of SARS-Cov-2 infections and vaccinations in Denmark. After minimizing the expression

$$\sum_{t=1}^{20}((V_{real}(t) - V_{model}(t))^2 + (I_{real}(t) - I_{model}(t))^2 + ((R_{real}(t) - R_{model}(t))^2)$$

we get information that the infection rate $\alpha = 1.72152 \cdot 10^{-7}$, the recovery rate $\gamma = 0.0748762$ and the vaccination rate $\delta = 0.0192519$.

Figure 7 presents a simulation of the course of SARS-Cov-2 virus infection for the next 30 days, that is until April 16, 2021.

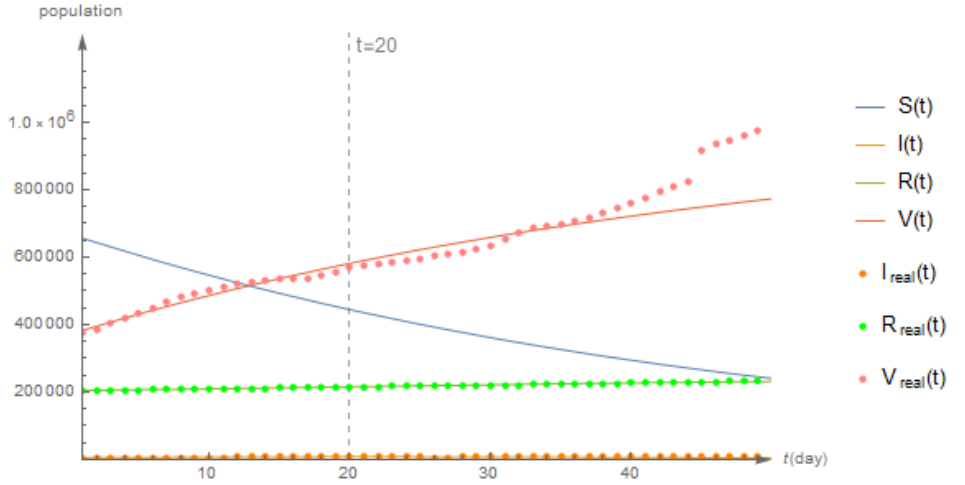

Figure 7: Epidemic simulation

The mean absolute percentage error for the infected population is 15.8515%. Theil's coefficient for $I(t)$ is equal to 0.0325189, so $T = 18.033\%$. The data we get from the simulation differ about 15.85 - 18% from the real data. In this situation, we can conclude that the model presents the course of infections in Denmark with a significant error.

The mean absolute percentage error for recovered people, $R(t)$, is 0.472322%. Theil's coefficient is equal to 0.000029394, so $T = 0.542162\%$. Thus, we can see that when comparing the predicted data with the real data, they differ by about 0.47 - 0.54%. So the difference between them is small.

The mean absolute percentage error for $V(t)$, i.e. vaccinated people, is 6.46959% and Theil's coefficient is equal to 0.0127645, so $T = 11.298\%$. We can say that in this case the simulation data differs by about 6.47 - 11.30% from the real data.

Let's also note that, according to the assumptions of the model, the average duration of illness per person is $\frac{1}{\gamma} \approx 13$. It turns out that this information is consistent with the facts reported in the literature, because the average duration of SARS-Cov-2 infection is considered to be two weeks [2].

4.4. Simulation of the course of the coronavirus pandemic in the Czech Republic using the SVIR model

To create a simulation of the course of the coronavirus pandemic in the Czech Republic, we will use data for the period from February 15, 2021 to April 15, 2021. We adjust the parameters of the model to the first 30 days of the considered period. In the Czech Republic, until April 15, 2021, a total of 3.16 million coronavirus infections and vaccinations were recorded, therefore we assume $N \approx 3160000$. Minimizing the expression

$$\sum_{t=1}^{30}((V_{real}(t) - V_{model}(t))^2 + (I_{real}(t) - I_{model}(t))^2 + ((R_{real}(t) - R_{model}(t))^2)$$

we find out that the parameters of the model take the following values:

$$\alpha = 5.56285 \cdot 10^{-8},$$
$$\gamma = 0.0613324,$$
$$\delta = 0.0108898.$$

Then, for $I(t)$, $R(t)$ and $V(t)$, we do ex post prediction for the next 30 days.

The mean absolute percentage error for the $I(t)$ infected population is 12.3121%, while Theil's coefficient is 0.0259539. Hence $T = 16.1102\%$. This means that the model differs by about 12.31 - - 16.11% from the actual course of coronavirus infection in the Czech Republic.

The mean absolute percentage error for $R(t)$, i.e. for recovered, is 1.24234%, and the Theil coefficient in this case is equal to 0.000189218. Then $T = 1.37556\%$. The difference between the data from the proposed model and the real data is about 1.24 - 1.38%.

The mean absolute percentage error for vaccinated people is 27.2049%. Theil's coefficient for $V(t)$ is equal to 0.0822293, so $T = 28.6756\%$. We can see that the difference between the predicted and actual number of people vaccinated is significant, as it is around 27.20 - 28.68%. From the Figure 8 you can see that from a certain point the actual number of people vaccinated is much higher than the one obtained from the model. The reason for this phenomenon may be the fact that people - aware of the threat posed by the coronavirus in previous periods - were more willing to be vaccinated.

An illustration of the data on the discussed simulation of the course of coronavirus infection in the Czech Republic is shown in Figure 8.

Let's take a closer look at the interpretation of the γ parameter. Turns out $\frac{1}{\gamma} \approx 16$ days. However, according to the literature, the disease caused by the SARS-Cov-2 virus lasts an average of 14 days [2]. Thus, also in this case, it can be assumed that the $\frac{1}{\gamma}$ value correctly reflects the average duration of the disease of any individual.

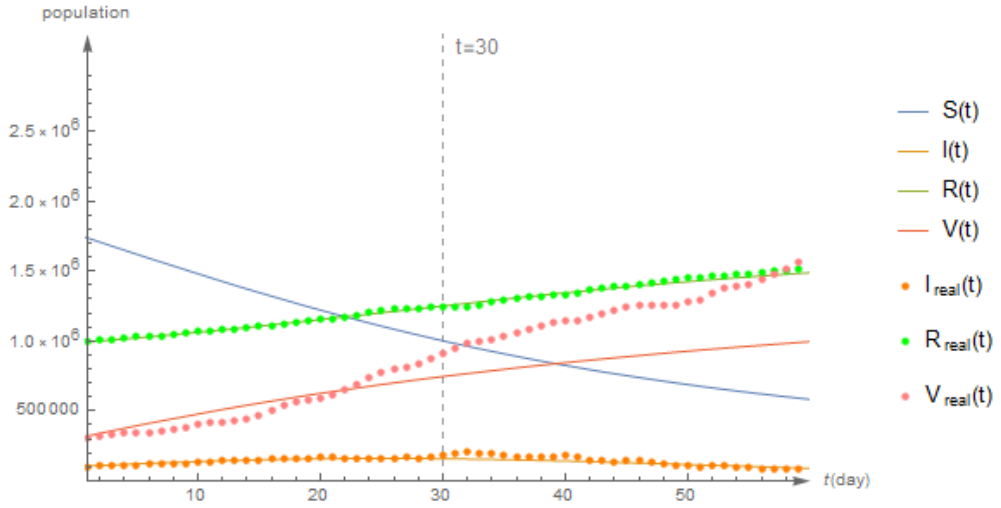

Figure 8: *Epidemic simulation*

5. Summary

It turns out that even the simplest SI model can correctly reflect the actual trend of coronavirus development in selected countries. Moreover, this model can be modified relatively easily. Thanks to this, it is possible to include more factors influencing the development of the disease in the considerations. We may consider, for example, vaccination of the analyzed population against the virus causing the infection. The results obtained from the performed simulations reflect the real data with a certain error. It can be assumed that this is due to the fact that the model does not take into account all the factors that affect the development and spread of the disease.

6. References

[1] L.J.S. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis, Infect Dis Model, 2017 Mar 11;2(2):128-142, DOI: 10.1016/j.idm.2017.03.001

[2] H.R. Bhapkar, P.N. Mahalle, N. Dey, K.C. Santosh, Revisited COVID-19 Mortality and Recovery Rates: Are we Missing Recovery Time Period?, Journal of Medical Systems 44, 202, 2020.

[3] R. Brociek, M. Pleszczyński, A. Zielonka, A. Wajda, S. Coco, G. Lo Sciuto, C. Napoli, Application of heuristics algorithms in the tomography problem for pre-mining anomaly detection in coal seams, Sensors, 2022, vol.22, nr 19, s.1-22, DOI: 10.33906/s22197297.

[4] J. Chaki, M. Woźniak, Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review, Biomedical Signal Processing and Control, Volume 80, Part 1, 2023, 104223, DOI: 10.1016/j.bspc.2022.104223.

[5] O. Diekmann, J.A.P. Heesterbeek, Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation, John Wiley, New York, 2000.

[6] U. Foryś, A. Dzwonkowska, J. Krawczyk, MATHEMATICS IN EPIDEMIOLOGY. HOW TO MODEL COVID-19 (in polish), Kosmos. Problemy nauk biologicznych, 2021, no 3, 475-484.

[7] E. Gudowska-Nowak, K. Oleś, B. Dybiec, A bit about epidemic modeling: SI, SIS, SIR and R_0 (in polish), Institute of Theoretical Physics, Jagielonnian University, 2020.

[8] E. Hetmaniok, M. Pleszczyński, Comparison of the selected methods used for solving the ordinary differential equations and their systems, Mathematics, 2022, vol.10, nr 3, s.1-15, DOI: 10.3390/math11030306.

[9] E. Hetmaniok, M. Pleszczyński, K. Yasir, Solving the integral differential equations with delayed argument by using the DTM method, Sensors, 2022, vol.22, nr 11, s.1-21, DOI: 10.3390/s22114124.

[10] B. Hu, H. Guo, P. Zhou, Z.-L. Shi, Characteristics of SARS-CoV-2 and COVID-19, Nature Reviews Microbiology 19, 141–154 (2021), DOI: 10.1038/s41579-020-00459-7.

[11] J. Jäschke, M. Ehrhardt, M. Günther, B. Jacob, A Two-Dimensional port-Hamiltonian Model for Coupled Heat Transfer, Mathematics 2022, 10, 4635, DOI: 10.3390/math10244635.
[12] W. O. Kermack, A.G. McKendrick, G.T. 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, Vol. 115, No. 772. (Aug. 1, 1927), pp. 700-721. (1927).
[13] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, M. Roser, Coronavirus Pandemic (COVID-19), 2020. URL: https://ourworldindata.org/coronavirus.
[14] J.D. Murray, Mathematical Biology: I. An Introduction, Springer, New York, NY, 2002.
[15] J.D. Murray, Mathematical Biology: II Spatial Models and Biomedical Applications, 3rd Edition, Springer, New York, 2003.
[16] N. Parolini, L. Dede', P.F. Antonietti, G. Ardenghi, A. Manzoni, E. Miglio, A. Pugliese, M. Verani, A. Quarteroni, SUIHTER: a new mathematical model for COVID-19. Application to the analysis of the second epidemic outbreak in Italy, Proceedings of the Royal Society, A477: 20210027 (2021).
[17] M. Pleszczyński, A. Zielonka, M. Woźniak, Application of nature-inspired algorithms computed tomography with incomplete data, Symmetry-Basel, 2022, vol.14, nr 11, s.1-15, DOI: 10.3390/sym14112256.
[18] M.H. Trinh, Q.D. Pham, V.N. Giap, Optimal Lyapunov-Based Sliding Mode Control for Slotless-Self Bearing Motor System, Applied System Innovation, 2023, 6(1):2, DOI: 10.3390/asi6010002.
[19] M. Woźniak, M. Wieczorek, J. Siłka, BiLSTM deep neural network model for imbalanced medical data of IoT systems, Future Generation Computer Systems, vol. 141, pages 489-499, 2023, DOI: 10.1016/j.future.2022.12.004.
[20] Z. Zhang, Y. Cai and D. Zhang, Solving Ordinary Differential Equations With Adaptive Differential Evolution, IEEE Access, vol. 8, pp. 128908-128922, 2020, DOI: 10.1109/ACCESS.2020.3008823.
[21] VisualMonsters.cba.pl, Theil's coefficient, 2022. URL: https://visualmonsters.cba.pl/prognozowanie/wspolczynnik-theila/.
[22] data on the course of the COVID-19 pandemic in selected countries. URL: https://covid.observer.
[23] Worldometers.info, data on the course of the COVID-19 pandemic in Croatia and Bulgaria, 2022. URL: https://www.worldometers.info/coronavirus/.