Research of COVID-19 infection waves using mathematical models at different levels

Sergey Misyurin[1*], *Andrey* Nelubin[1], *Alexander* Trofimov[2], *Anna* Nozdracheva[3], *Natalia* Nosova[1], and *Nikolay* Khokhlov[2]

[1]Blagonravov Mechanical Engineering Research Institute RAS, 4, Malyi Kharitonievsky Per., Moscow, 101990, Russia
[2]National Research University MEPhI, 31, Kashirskoe Shosse, Moscow, 115409, Russia
[3]National Research Center for Epidemiology and Microbiology named after the honorary academician N. F. Gamaleya, 18, Gamaleya St., Moscow, 123098, Russia

Abstract. The work is devoted to the problem of predicting the increase in the incidence of viral infections such as COVID-19 using mathematical models of different levels. The world continues to study the entire accumulated set of data on the fight against the new coronavirus infection, including morbidity statistics, using various analytical methods. One of these methods is the construction of mathematical models of the COVID-19 epidemic process, which is based on the "susceptible-infected-recovered" (SIR) model proposed more than a century ago. The work shows that the complication of mathematical models that take into account the change in genovariant of SARS-CoV-2 can lead to incorrect results and erroneous conclusions, both for short and long periods. At the same time, the use of a fairly simple SIR model for each period of dominance of a certain variant of the pathogen gives an acceptable forecast result for a short period.

1 Introduction

COVID-19, which caused one of the largest epidemics in modern history and claimed several million lives, continues to spread across the planet. There are historical examples in the world of combating such large-scale epidemics (for example, those caused by influenza viruses), but the coronavirus pandemic is by far the most characterized in terms of the achievements of modern medicine, and measures to combat it include a wide range of methods and means. At the same time, the biological threat of the spread of another variant of a known or new pathogen with the emergence of a pandemic remains. This makes the accumulated experience in combating the spread of COVID-19 the most valuable and significant. One of the main methods for studying and forecasting the spread of morbidity is mathematical modeling of viral activity and the search for patterns that influence the nature of changes in the solution of the mathematical model. In this regard, the large number of virus spread models that have emerged recently are important for determining optimal approaches. The study of the entire accumulated set of data, including morbidity statistics,

* Corresponding author: nelubin@gmail.com

continues using various analytical methods. One of these methods is the construction of mathematical models of the epidemic process of coronavirus infection, which are based on the "susceptible-infected-recovered" (SIR) model proposed about a century ago [1], and used in our time, including with variable coefficients in works [2-4]. By "susceptible", we mean persons who have not encountered the pathogen and do not have immunity to it. The first attempt to use mathematical apparatus to study the mechanisms of disease spread was made by D. Bernoulli. Continuing his work, British scientists A. Kermack and W. McKendrick developed the SIR model that is widely used today [1]. Within the framework of this model, the dynamics of the spread of an infectious disease is described using systems of differential equations (in the general case). This system of differential equations has three variables: the number of susceptible individuals, the number of infected individuals and the number of recovered individuals; two coefficients characterizing the intensity of contacts between individuals with subsequent infection; recovery rate of infected individuals; constant – the total population size.

Subsequently, various researchers proposed more expanded systems of differential equations that took into account such quantities as the incubation period during infection, mortality among patients, different degrees of severity of the disease, the degree of contact between different groups of the population and many other factors influencing the nature of the spread of viral infection.

With the advent of the first statistical data on the infection and course of the disease COVID-19, many works appeared on the analysis of the nature of the course of the infection based on the use of various mathematical models, which are based on the same SIR model, such as SEIR, SIRS, SAIR [5-8] etc. It became clear that it is necessary to take into account not only the infected and recovered, but such variables/parameters as mortality among infected people, the incubation period during infection, the presence of immunity in the population, as well as many other parameters. One of the key epidemiological parameters is the incubation period, which for COVID-19, according to WHO estimates, varies from 3 to 7 days. In works [9, 10], the optimal value of the incubation period for mathematical models was 5.2 days. The need to take this parameter into account in SEIR models was demonstrated by Chinese scientists, for example, in [11-13]. The article [14] proposes a QSEIR model that substantiates the need to take into account quarantine measures, proposes a corresponding mathematical model, and makes a forecast of the development of the epidemic in China. More complex mathematical models that take into account various clarifying factors are given in [15-17]. The more complex the model, the more unknown parameters it contains, which are often difficult to estimate. Based on the analysis of data on the number of infected people in the first days of the increase in incidence, it is possible, by solving the inverse problem, to obtain all the uncertain parameters, but due to the inaccuracy of measurements, their reliability is difficult to determine. The work [18] noted that, according to various estimates, the statistical error in monitoring COVID-19 ranges from 20 to 80%. Finding unknown model parameters using inaccurate statistical data is an ill-posed task.

A mathematical model that takes into account such components as hospitalized patients with severe illness; are in critical condition; connected to a ventilator (SEIR-HCD), was first proposed by French scientists in [19], and is based on a system of seven nonlinear ordinary differential equations. In [20], this model was used as a forecast of the COVID-19 epidemic situation using the example of the Moscow and Novosibirsk regions, and an analysis of the identifiability of the mathematical model was carried out.

As the analysis of the above publications shows, as the variables increase, the model becomes more accurate. But the more accurate the model, the more it shows an incorrect result in the case of unaccounted for the impact of both objective and subjective influences on the situation. Restrictive measures to influence the epidemiological situation cannot give

an unambiguous response to statistical indicators at a large stage of observation. Here both the human factor and various situations of natural and environmental impact come into force. Restrictive measures are influenced by seasonal conditions, social fatigue, climatic conditions, and social problems of various kinds. This all leads to incorrect results from mathematical models with a large number of variables.

In statistical analysis using machine learning, this effect is called overfitting of the model [21, 22]. More accurate models achieve a smaller error on the training data set, but at the same time lose generalization ability and give a worse prediction on the test data set. To achieve the best generalization ability, it is necessary to seek a balance between models that are too simple and too complex.

The goal of this work is to show that the complication of mathematical models, which take into account, among other things, the change in genovariant of SARS-CoV-2, can lead to incorrect results and erroneous conclusions, both for short and long periods. At the same time, the use of a fairly simple SIR model for each period of dominance of a certain variant of the pathogen gives an acceptable forecast result for a short period. The results of this work may be useful for assessing the spread of SARS-CoV-2 and other pathogens with similar properties for a short period according to the first harmonic, which will allow the correct decision to be made on vaccinating the population in the initial period of the spread of infection, as well as the nature of the spread during subsequent waves of increasing incidence**.**

2 Mathematical models of SIR and SEIR-HCD

The **SIR** system of equations is as follows:

$$\begin{cases} \frac{dS(t)}{dt} = -\frac{\beta \cdot I(t) \cdot S(t)}{N} \\ \frac{dI(t)}{dt} = \frac{\beta \cdot I(t) \cdot S(t)}{N} - \gamma \cdot I(t) \\ \frac{dR(t)}{dt} = \gamma \cdot I(t) \end{cases} \quad (1)$$

$$N = S(t) + I(t) + R(t),$$

here:

• **S(t)**: The number of people susceptible to the disease. When a susceptible person and an infectious person come into “infectious contact”, the susceptible person becomes infected;

• **I(t):** The number of infectious people. These are those who have become infected and can transmit the disease to susceptible people;

• **R(t)**: The number of people recovered (or died). These are those who became infected and either recovered, becoming immune, or died. The number of deaths is believed to be insignificant compared to the general population. This category may also be called “resistant”;

β – Coefficient of intensity of contacts of individuals with subsequent infection;

γ – Recovery rate of infected individuals;

N – Total population size.

The first equation of system (1) means that the change in the number of healthy individuals (and at the same time susceptible to the disease) decreases over time in proportion to the number of contacts with infected people. After contact, infection occurs and the susceptible person becomes infected. The second equation of the system (1) shows that the rate of increase in the number of infected people increases in proportion to the

number of contacts between healthy and infected people and decreases as the latter recover. The third equation of system (1) demonstrates that the number of recovered people per unit time is proportional to the number of infected people. In other words, everyone who gets sick should recover after some time.

This model is good at predicting the course of infectious diseases transmitted from person to person, in which recovery provides long-term immunity, such as measles, mumps and rubella.

The variables (*S*, *I* and *R*) represent the number of people in each category at a particular point in time. To reflect that the number of susceptible, infectious, and recovered individuals may change over time (even if the total population remains constant), we make the exact numbers a function of time t: S(t), I(t), and R(t). For a specific disease in a specific population, these functions can be calculated to predict possible outbreaks and control them.

The model is dynamic, which means the number of people in each category can fluctuate over time. During an epidemic, the number of susceptible people drops rapidly as many of them become infected and thus move into the infectious and immune categories. The disease cannot break out again until the number of susceptible people is restored, for example by the birth of new susceptible people.

Each member of a population typically goes from susceptible to infectious to recovered. This can be shown in the form of a flowchart, where the boxes represent the different categories and the arrows represent the transitions between them (Figure 1).

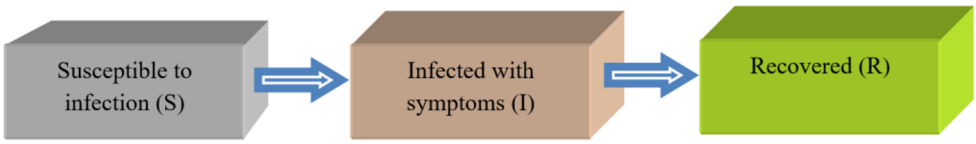

Fig. 1. Block diagram of the SIR model.

As noted above, the most detailed model used in the analysis of COVID-19 incidence data is the SEIR-HCD model, given in [19, 20]. Here, in contrast to the SIR model, the following variables are added: infected or in the incubation period; hospitalized; are in critical condition; and the dead. The system has seven equations (2) and 14 parameters.

In [20], the SEIR-D model was also studied, which, unlike the previous one, is simpler and does not take into account hospitalized and critically ill patients. A comparison was made of the results obtained from two mathematical models such as SEIR-HCD and SEIR-D, the parameters of which were refined for Moscow and the Novosibirsk region using optimization methods. It was noted that the SEIR-HCD model also performs well in predicting identified cases over a short period. However, in some cases, the simpler SEIR-D model performed better. It was concluded that the use of the roughest mathematical model (with a smaller number of homogeneous groups) is justified in the case of a sufficiently large amount of statistical data and a short forecast period.

The **SEIR-HCD** system of equations is as follows:

$$\begin{cases} \frac{dS(t)}{dt} = -\frac{5-a(t-\tau)}{5}\left(\frac{\beta_I \cdot I(t) \cdot S(t)}{N} + \frac{\beta_E \cdot I(t) \cdot S(t)}{N}\right) + \delta R(t) \\ \frac{dE(t)}{dt} = -\frac{5-a(t-\tau)}{5}\left(\frac{\beta_I \cdot I(t) \cdot S(t)}{N} + \frac{\beta_E \cdot I(t) \cdot S(t)}{N}\right) + (\kappa - \rho)E(t) \\ \frac{dI(t)}{dt} = \kappa \cdot E(t) - (\gamma + \upsilon) \cdot I(t) \\ \frac{dR(t)}{dt} = \gamma \cdot I(t) - \sigma \cdot R(t) + \rho E(t) + \varepsilon_{HR} H(t) \\ \frac{dH}{dt} = \upsilon I(t) + \varepsilon_{CH} C(t) - (\varepsilon_{HR} + \varepsilon_{HC})H(t) \\ \frac{dC}{dt} = \varepsilon_{CH} H(t) - (\varepsilon_{CH} + \mu)C(t) \\ \frac{dD}{dt} = \mu C(t) \end{cases} \quad (2)$$

Here

E(t) – Infected or incubating individuals;

H(t) – Hospitalized, i.e. with severe illness;

C(t) – In critical condition, connected to a ventilator;

D(t) – Dead

a(t) – Self-isolation index according to open data;

β_I – A parameter of infection between infected and susceptible populations that is related to the contagiousness of the virus;

β_E – Infection parameter between asymptomatic and susceptible populations ($\beta_E >> \beta_I$);

κ – Rate of symptom onset in open cases, resulting in transition from an asymptomatic to an infected population;

ρ – Recovery rate of detected cases (cases that are detected but recover without any symptoms);

γ – Recovery rate of infected cases;

δ – Reinfection rate. This parameter is the reciprocal of the virus's immunity level (0 – stable immunity, 0.001 – probability of re-infection);

υ – Proportion of hospitalized cases with severe disease;

ε_{HR} – Probability of recovery for individuals in serious condition;

ε_{HC} – Proportion of hospitalized cases in critical condition requiring ventilator support;

ε_{CH} – Probability of turning off the patient's ventilator;

μ – Mortality due to COVID-19.

The model can be represented schematically as shown in Figure 2. This model began to be actively used when it became clear that seriously ill patients were being connected to mechanical ventilation, while this procedure did not help in all cases; there was a possibility of re-infection due to the fact that the presence of antibodies does not guarantee protection against recurrent disease; the disease may be asymptomatic and/or other factors. Of course, this model is very important, and with its help you can understand many of the features and patterns of this infection. But as it turned out, the virus mutates quite quickly, several stamps can be active at once, the nature of restrictive measures has an ambiguous effect on the overall situation and it is problematic to obtain a long-term prognosis. Even new waves of the pandemic are not similar in nature to previous ones. Over long periods of modeling, it is necessary to take into account the dynamics of the measures introduced to combat coronavirus infection (mandatory use of masks indoors, restriction of population movement in the regions, border closures, etc.), which affect real data, quantitative and qualitative indicators of system parameters, pandemic development scenarios.

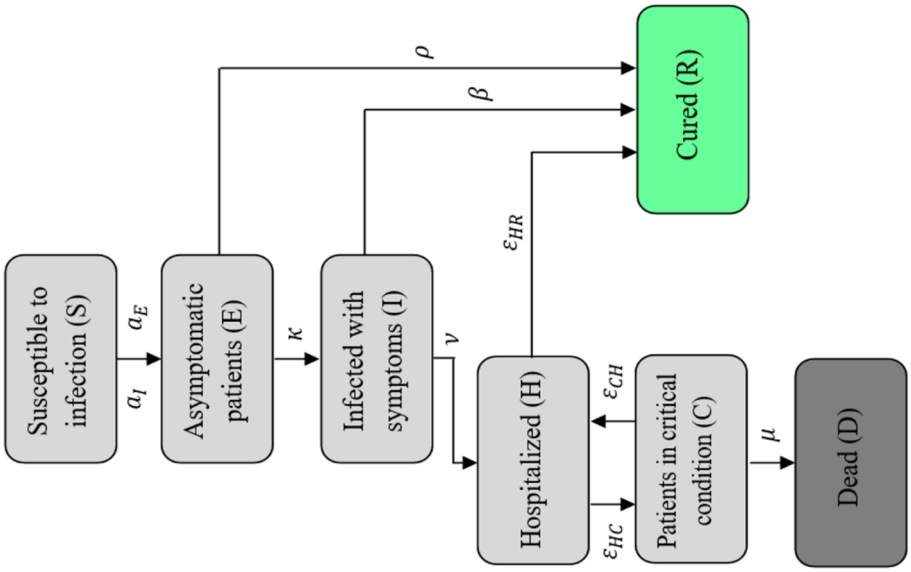

Fig. 2. Block diagram of the SEIR-HCD model.

In this case, it is advisable to consider simpler models, for example, SIR, and adjust the result with daily addition of data. This paper will give an example of applying the SIR model over a short period.

An analysis of the presented data on modeling the spread of COVID-19 showed that the high-order mathematical models used, which take into account many factors influencing the epidemic, are not able to predict the nature of the incidence for a long period. This certainly does not mean that higher order models with more variables are not worth considering. More complex models can show non-obvious dependencies and correlations between variables and independent constants, between introduced restrictive measures and the model's response to these measures, the dependence of parameters when the dominant strain of the virus changes, etc.

3 Statement of the inverse problem

In section 1, we looked at various model options for describing and modeling epidemics, such as SIR, SEIR-HCD, SEIR-D and others. All of these systems contain internal parameters, such as infection and recovery rates, that determine the dynamics of the spread of the disease. These parameters are usually not known in advance and must be estimated based on observed data. Setting and solving such an inverse problem allows one to accurately model and predict the behavior of an epidemic, which is important for the effective management of infectious disease outbreaks. Let us have a model $A(t;\theta)$, depending on the parameters θ, and a sample of real observations X_{train}, $T = |X_{train}|$ - size of the training sample. The inverse problem is to minimize some error function $Loss(\theta; X_{train}, A, metric)$, where metric is a measure of the quality of the model. Formally, the problem looks like this:

$$\min_{\theta} Loss(\theta;\ X_{train}, A, metric) = \frac{1}{T}\sum_{t=1}^{T} metric\ (A(t;\theta), X_{train}(t)) \qquad (3)$$

here

• A(t; θ) – A mathematical model that describes the dynamics of the system and is a mapping $R \rightarrow R^n$, i.e. a model that makes predictions for day t.

• $\theta \in R^m$ – Vector of model parameters that need to be estimated;

• $X_{train} \in R^{T \times n}$ – Observational data, which is a set of measurements $\{X(t_1), X(t_2), \ldots, X(t_T)\}$;

• $metric$ – A function that measures the discrepancy between model predictions and observations.

Let's consider this using the SIR model (1) as an example. The initial conditions are set as follows:

$$S(0) = S_0 I(0) = I_0 R(0) = R_0.$$

The goal of the inverse problem is to find such parameter values $\theta = (\beta, \gamma, S_0, I_0, R_0)$ that minimize the difference between the model solutions and the observed data. It can be noted that the parameters S_0, I_0, R_0 are the initial conditions for the Cauchy problem, and this is the first point from which we build a forecast. But since the measurement has a probabilistic error, the initial values are optimized within the limits of the measurement error.

Let $I_{obs}(t)$ be the observed data on the number of infectious people at time t, and $I_{model}(t; \theta)$ be the model solution for given θ at time t. Then the mean square error (MSE) of the model is determined as follows:

$$Loss^{MSE}(\theta) = \frac{1}{T}\sum_{t=1}^{T}(I_{molel}(t;\theta) - I_{obs}(t))^2 \quad (4)$$

And the mean absolute error (MAE) of the model is defined as follows:

$$Loss^{MAE}(\theta) = \frac{1}{T}\sum_{t=1}^{T}|I_{molel}(t;\theta) - I_{obs}(t)| \quad (5)$$

In this work, both metrics, (4) and (5), were studied. To solve the inverse problem, the following methods were used: least squares, stochastic gradient descent, genetic algorithm, differential evolution, particle swarm optimization, simulated annealing, as well as direct enumeration of solutions at grid nodes. To numerically solve optimization problems, a Python program was written using publicly available libraries with algorithms for the listed methods.

After selecting the parameters on the training data set, the quality of the model was assessed on the test data. The results of numerical modeling were analyzed for different sizes of training and test data samples. Evolutionary algorithms (genetic algorithm, differential evolution) and random search methods (particle swarm optimization, double annealing) showed stable results on test data and resistance to local minima. The direct enumeration method turned out to be less effective compared to other optimization methods, especially when the size of the parameter space increases, due to its high computational complexity.

4 Results of solving the inverse problem for the SIR model

Figure 3 shows an example of disease statistics for Moscow in the period from 01.2020 to 01.2024.

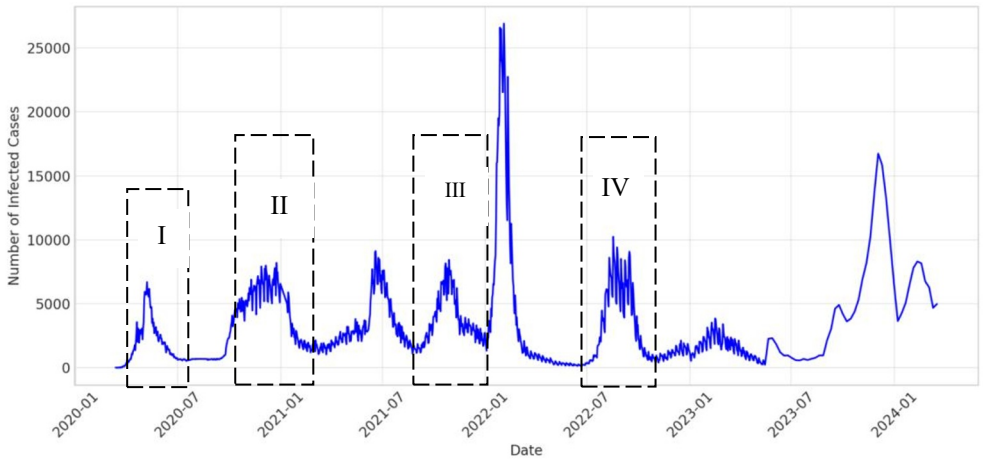

Fig. 3. Dynamics of COVID-19 incidence for the period from 01/2020 to 01/2024. according to data (data source). Roman numerals indicate the intervals at which the modeling of the COVID-19 epidemic process was carried out.

In Figure 3 we see waves of rises and falls in the incidence of COVID-19, which vary both in amplitude (number of cases) and duration. Let us consider the results of numerical simulation over short intervals. Figure 3 shows 6 intervals characterized by the beginning, maximum value and decline of incidence. Figures 4-9 show the results of the forecast (numerical modeling) of morbidity in the intervals (I, II, III, IV) of Figure 3. Green, both solid and dotted, indicates statistical data. Solid green is the data used to solve the inverse problem. Blue color – model forecast. Vertically the number of cases, horizontally the number of days.

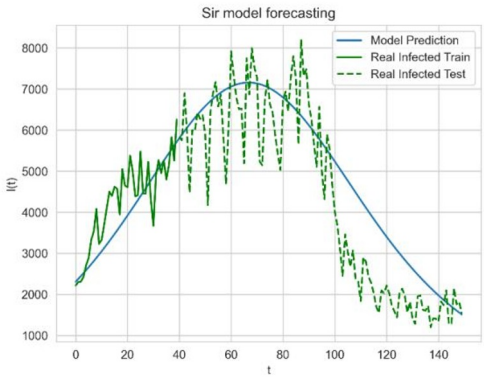

Fig. 4. The results of the forecast (numerical modeling) of morbidity in the interval (I) of Figure 3.

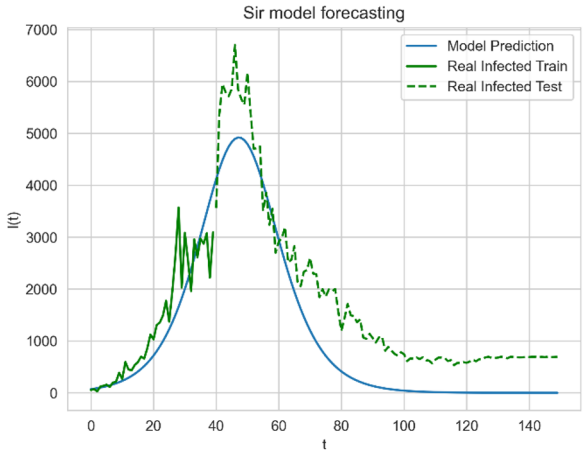

Fig. 5. The results of the forecast (numerical modeling) of morbidity in the interval (II) of Figure 3.

Evaluating the model's forecast results can be done in different ways. One of the common approaches is to use the error functions (4) or (5) on the test data set. In addition, in the applied problem under consideration, the substantive meaning and greatest practical importance is the forecast of the peak of morbidity, namely, the highest level of morbidity and the duration of its achievement and subsequent decline.

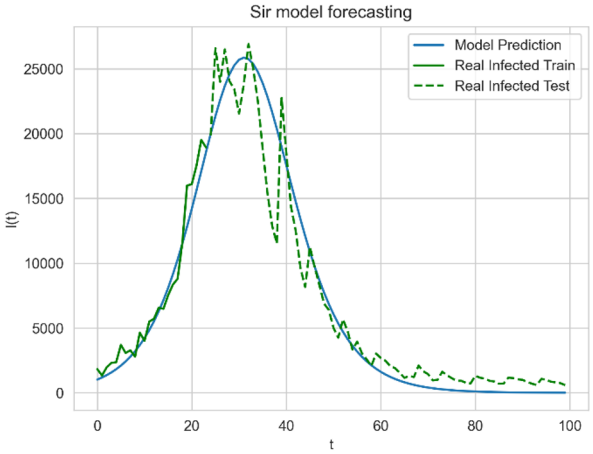

Fig. 6. The interval (III) of Figure 3, but with a different sample according to the time of observation of the increase in incidence to determine the vector θ, 20 days.

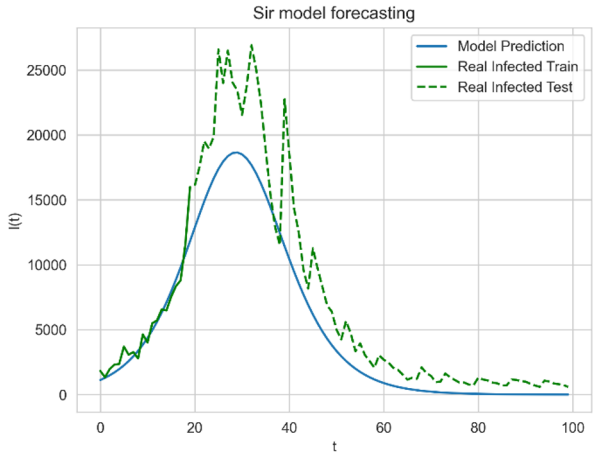

Fig. 7. The interval (III) of Figure 3, but with a different sample according to the time of observation of the increase in incidence to determine the vector θ, 25 days.

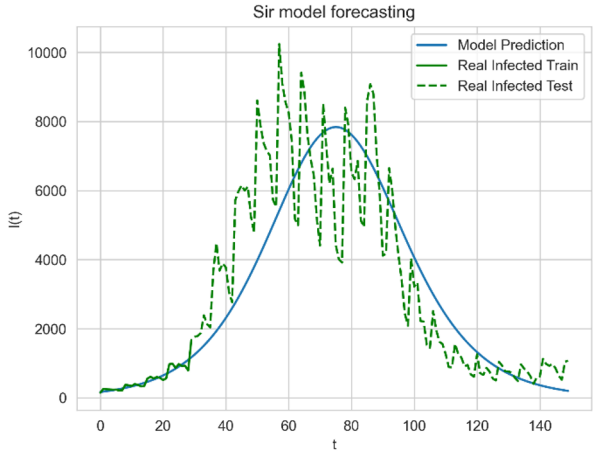

Fig. 8. The interval (III) of Figure 3, but with a different sample according to the time of observation of the increase in incidence to determine the vector θ, 30 days.

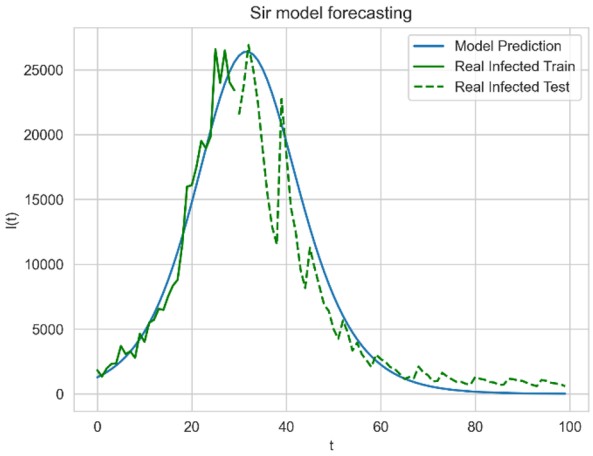

Fig. 9. The increasing incidence of interval (IV) of Figure 3.

Figures 4 and 5 were obtained on the basis of a mathematical model, where 40 days after the start of the increase in incidence were used to determine the vector θ. The graphs reflect time intervals (I) and (II) of Figure 3, respectively. There is some error, but qualitatively the forecast and observation graphs coincide quite well. Figures 6, 7, 8 show the same interval (III) of Figure 3, but with a different sample according to the time of observation of the increase in incidence to determine the vector θ, 20, 25 and 30 days, respectively. Qualitatively, all three graphs show good results, but with increasing days of observation, the accuracy of the forecast increases. Those. With daily observation, you can obtain a clarifying result to correct measures to counter the spread of the epidemic. Figure 9 shows the increasing incidence of interval (IV). This example shows a forecast graph with an observation period of 20 days. As we increase the observation days, just as in the previous example, we will obtain a more accurate forecast schedule.

5 Conclusion

This work analyzes the use of mathematical models of varying complexity on the spread of COVID-19 infection over a long period of time. It has been shown that long-term forecasting of the development of the epidemic using mathematical models is difficult due to the following reasons:

- Statistics on the spread of the epidemic often contain significant errors.
- It is difficult to obtain reliable values of coefficients used in mathematical models with many variables.
- It is problematic to take into account a number of factors influencing the development of the epidemic. These are primarily: seasonal factors; restrictive measures in society; society's fatigue from being under stress during a long epidemic; virus mutation and a number of other reasons.

The computational experiments carried out made it possible to determine which optimizers are most effective for specific epidemic models and which of them provide the best quality of predictions on test data for different sizes of training and test data samples. Evolutionary algorithms (genetic algorithm, differential evolution) and random search methods (particle swarm optimization, double annealing) showed stable results on test data and resistance to local minima. The direct enumeration method turned out to be less

effective compared to other optimization methods, especially when the size of the parameter space increases, due to its high computational complexity.

At the same time, the use of the classical SIR system with constant coefficients for a short period of one "wave" gives an acceptable result, which qualitatively shows the passage of the epidemic's peak of growth and the nature of the decrease in morbidity (the number of infected).

References

1. W. O. Kermack, A. G. McKendrick, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **115(772)**, 700–721 (1927)
2. S. Cauchemez, A.-J. Valleron, P.-Y. Boëlle, et al., Nature **452(7188)**, 750–754 (2008)
3. J. M. Read, J. R. E. Bridgen, D. A. T. Cummings, et al., Philosophical Transactions of the Royal Society B: Biological Sciences **376(1829)**, 1–9 (Article No. 20200265) (2021)
4. B. F. Maier, D. Brockmann, Science **368(6492)**, 742–746 (2020)
5. T. T. Marinov, R. S. Marinova, Chaos, Solitons & Fractals **X5**, 1–15 (Article No. 100041) (2020)
6. M. Y. Li, J. S. Muldowney, Mathematical Biosciences **125(2)**, 155–164 (1995)
7. T. Odagaki, Infectious Disease Modelling **5**, 691 – 698 (2020)
8. S. Contreras, H A. Villavicencio, D. Medina-Ortiz, et al., Chaos, Solitons & Fractals **136**, 1 – 5 (Article No. 109925) (2020)
9. J. A. Backer, D. Don Klinkenberg, J. Wallinga, Euro Surveill **25(5)**, 2000062 (2020)
10. World Health Organization. Novel Coronavirus (2019-nCoV). Situation Report – 10., 1–7 (2020)
11. A. Zlojutro, D. Rey, L. GardnerScientific Reports **9**, 1–14 (Article No. 2216) (2019)
12. B. Tang, X. Wang, Q. Li, et al., Journal of Clinical Medicine **9(2)**, 1–13), (Article N0. 462) (2020)
13. R.-G. Fan, Y.-B. Wang, M. Luo, et al., Journal of University of Electronic Science and Technology of China **49(3)**, 369–374 (2020)
14. X. Liu, G. Hewings, Sh. Wang, et al., Modelling the situation of COVID-19 and effects of different containment strategies in China with dynamic differential equations and parameters estimation. medRxiv, 1–31 (2020)
15. A. M. Ramos, M. R. Ferrández, M. Vela-Pérez, et al., Physica D: Nonlinear Phenomena **421**, 1–22 (Article No. 132839) (2021)
16. U. A. P. de León, A. G. C. Pérez, E. Avila-Vales, Chaos, Solitons & Fractals **140**, 1–23 (Article No. 110165) (2020)
17. M. Higazy, Chaos Solitons Fractals **138**, 1–19 (Article No. 110007) (2020)
18. E. S. Kurkina, E. M. Koltsova, *Mathematical modeling and forecasting of the spread of the COVID-19 coronavirus epidemic*, in Proceedings of the 4th International Conference. Designing the future. Problems of digital reality, pp. 178–192, Moscow, IPM im. M.V. Keldysh (2021)
19. E. Unlu, H. Léger, O. Motornyi, et al., Epidemic analysis of COVID-19 Outbreak and Counter-Measures in France. medRxiv, 1–10 (2020)

20. O. I. Krivorot'ko, S. I. Kabanikhin, N. Y. Zyat'kov, et al., Numerical Analysis and Applications **13 (4)**, 332–348 (2020)
21. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd edn. (Springer, New York, NY, 2009)
22. D. M. Hawkins, Journal of Chemical Information and Computer Sciences **44 (1)**, 1–12 (2004)