UDC 004.942:614.4 **doi: 10.32620/reks.2021.3.01**

**A. MOHAMMADI, I. MENIAILOV, K. BAZILEVYCH,
S. YAKOVLEV, D. CHUMACHENKO**

National Aerospace University "Kharkiv Aviation Institute", Ukraine

COMPARATIVE STUDY OF LINEAR REGRESSION AND SIR MODELS OF COVID-19 PROPAGATION IN UKRAINE BEFORE VACCINATION

The global COVID-19 pandemic began in December 2019 and spread rapidly around the world. Worldwide, more than 230 million people fell ill, 4.75 million cases were fatal. In addition to the threat to health, the pandemic resulted in social problems, an economic crisis and the transition of an ordinary life to a "new reality". Mathematical modeling is an effective tool for controlling the epidemic process of COVID-19 in specified territories. Modeling makes it possible to predict the future dynamics of the epidemic process and to identify the factors that affect the increase in incidence in the greatest way. The simulation results enable public health professionals to take effective evidence-based responses to contain the epidemic. The study aims to develop machine learning and compartment models of COVID-19 epidemic process and to investigate experimental results of simulation. The ***object of research*** *is COVID-19 epidemic process and its dynamics in territory of Ukraine. The* ***research subjects*** *are methods and models of epidemic process simulation, which include machine learning methods and compartment models. To achieve this aim of the research, we have used* ***machine learning forecasting methods*** *and have built COVID-19 epidemic process linear regression model and COVID-19 epidemic process compartment model. Because of experiments with the developed models, the predictive dynamics of the epidemic process of COVID-19 for 30 days were obtained for confirmed cases, recovered and death. For 'Confirmed', 'Recovered' and 'Death' cases mean errors have almost 1.15, 0.037 and 1.39 percent deviant, respectively, with a linear regression model. For 'Confirmed', 'Recovered' and 'Death' cases mean errors have almost 3.29, 1.08, and 0.71 percent deviant, respectively, for the SIR model.* ***Conclusions.*** *At this stage in the development of the epidemic process of COVID-19, it is more expedient to use a linear model to predict the incidence rate, which has shown higher accuracy and efficiency, the reason for that lies on the fact that the used linear regression model for this research was implemented on merely 30 days (from fifteen days before 2nd of March) and not the whole dataset of COVID-19. Also, it is expected that if we try to forecast in longer time ranges, the linear regression model will lose precision. Alternatively, since SIR model is more comprised in including more factors, the model is expected to perform better in fore-casting longer time ranges.*

Keywords: *epidemic model; epidemic process; epidemic simulation; simulation; linear regression; SIR model; COVID-19.*

Introduction

The new coronavirus, first identified at the end of 2019, spread rapidly around the world, and on April 4, 2021, the number of infected was 131,717,907 and the number of deaths was 2,863,227. The ease of transmission of the virus by aerosol from person to person contributed to the high rate of spread of the pathogen in different countries and created difficulties in the fight against infection. With the aerosol mechanism of infection, the most effective measure to prevent the circulation of the virus is to create a high level of herd immunity. A person acquires immunity either naturally – as a result of an illness, or artificially – as a result of vaccination. In the world, vaccination of the population began in mid-December 2020, starting with vaccinations for risk groups, primarily medical workers. In Ukraine, vaccination began on February 24, 2021, and as of April 3, 2021, 290,566 people were vaccinated in the country, which is not enough to limit the circulation of the virus. The dynamics of the epidemic process and its patterns differ from country to country. Vaccination rates and immunization coverage also differ. For a correct understanding of the development of the epidemic in the specific socio-economic conditions of the country and forecasting the epidemic situation to make optimal management decisions to mitigate the consequences of the epidemic (outbreak), it is necessary to develop a mathematical apparatus that can most accurately predict the dynamics of morbidity in a specific period in a specific territory.

Two aspects typically determine a model fitting to a data set [1]. Model complexity is obtained by the variables number and parameters are the first aspect. Generally, more complicated models might give premier fits to data. Nevertheless, less complicated models appear more lucid and provide more valuable and influential insights in illustrating the trends. Interchanging between bias and variance plays an important role in choosing the complexity optimal level [2]. Relevance between the parameters is the second aspect. It is substantial to understand that the basis of such relativity does not need to be completely precise for the model to be feasible. Structural stability plays an important role here, which refers to the possibility of making small changes in the model assumptions that might cause substantial forecasting changes [3].

For this research, we have chosen two models. Firstly, Linear Regression is handy to execute and utilize for analyzing the COVID-19 trend. Based on limited factors, the results executed by Linear Regression can be accountable for investigating the trend. Secondly, Susceptible Infective Recovered or SIR includes more parameters and complicated calculations yet easy to execute and operate on the desired dataset and also analyzes the epidemic main drivers.

In the first part, each model will be explained. In the next part, models will be run on the COVID-19 dataset and the result of both implemented models will be represented and investigated for a limited time range. The final part will include the comparison among both model performances in forecasting the trend of COVID-19 and actual data which is the purpose of this research. All the investigation was done on Ukraine recorded data provided by the Center of Public Health of Ministry of Health of Ukraine.

The paper aims to develop machine learning and compartment models of the COVID-19 epidemic process and to investigate the experimental results of the simulation. The object of research is the COVID-19 epidemic process and its dynamics in the territory of Ukraine. The subjects of research are methods and models of epidemic process simulation, which include machine learning methods and compartment models.

To achieve the aim of the research following tasks have been formulated:

1. Machine learning model of the COVID-19 epidemic process based on the linear regression method should be developed.

2. Compartment model of the COVID-19 epidemic process should be developed.

3. Experimental study of the Linear regression model of the COVID-19 epidemic process should be provided.

4. Experimental study of SIR model of COVID-19 epidemic process should be provided.

5. Results obtained during the experimental studies should be analyzed and compared.

The respective contribution of this study is threefold. Firstly, the development of models based on regression methods will allow estimating the accuracy of simple machine learning methods applied to epidemic process simulation. Secondly, the development of a compartment model will allow to estimation application of classical approaches to novel coronavirus disease simulation. And, finally, a comparison study of two different approaches to novel emergence disease epidemic process simulation will contribute to the normative and empirical evaluation of given models' application advisability to develop effective evidence-based anti-epidemic and control measures.

In this paper, section 1, namely materials and methods, provides a brief overview of linear regression, multiple linear regression and compartment approach, and the development of epidemic process models based on proposed methods. Section 2 constitutes the results and findings based on linear regression and SIR compartment models' of the COVID-19 epidemic process in Ukraine. Comparison results of given models' performance are described in Section 2.3. Conclusions describe outcomes of the proposed methodology.

1. Materials and Methods

1.1. Linear Regression

Linear regression is considered a statistical test that is applicable to a set of data and quantifies and defines the relevance among the dependent and independent variables. Linear regression is impressively mighty in analyzing data and gives the researcher the allowance to control the confounders' effects in realizing the relevance among two variables [4].

In clinical research, the researcher intends to figure out the correlation among two or more independent variables as inputs and therefore to find out a dependent variable as an output. Thus, this might be comprehended as how the independent variables are considered for the forecasting of the disease occurrence chance [5]. The regression model forecasts a dependent variable value regarding at least one independent variable value and represents the chance of understanding the "independent variables (risk variables) – dependent variables (e.g., disease)" relationship and describes it mathematically [6].

The linear regression model usage is considered for some main reasons which are being descriptive which support analyzing the association strength among the dependent variable as the output and the inputs as independent variables and adjustment which optimizes covariates effects or the confounders [7]. Also, it supports estimating the major independent factors that influence the dependent variable and analyzes the influence on the dependent variable caused by changing the independent variable per one unit. Finally forecasting the new cases is another feature of the linear regression model [8].

The main point to consider is linear regression analysis forecast does not imply causation. Accordingly, a researcher is not able to conclude that an independent variable causes a dependent variable based on given cross-sectional survey data. As an example, let's assume that a person's income and conspicuous consumption are relevant [9]. Thus, to forecast the person's income level within a margin of error, the person's purchasing behavior information can be utilized. Nonetheless, it would not be correct to conclude that income is caused by. It is important to indicate that causation assessment requires a proper design with cause and influence temporally isolation and spuriousness prevention [10]. As a result, linear regression with the mentioned design can be utilized to assess causal hypotheses.

To achieve validated results of linear regression usage, there are assumptions that the basic data needs to meet [11]. These assumptions are considered the same for different types of regression such as simple linear regression, multiple regression, and hierarchical regression.

These assumptions are as follows:

1. The researcher sets the independent variable values x.

2. Measuring the independent variable x must be done with no experimental error.

3. For each independent variable x value, a normally distributed subset of variables y is there up and down the Y-axis and the subset of variables y differences are uniformly distributed.

4. The subsets of variables y mean values perches on a straight line, which implies the assumption of the existence of a linear correlation among the dependent and independent variables.

5. All the y values are dependent on x and independent individuals [8].

The Determination Coefficient is the total variation part in the dependent variable that is described by independent variable variation. When R2 is bigger than one, that means there is a perfect linear relevance among x and y, or in other words, the variation in y is explained by variation in x precisely. When 0 < R2 < 1, that means there is a less precise linear relevance among x and y, thus the y variation is explained by x variation incompletely [12].

Simple linear regression is a model including a single regressor x that has relativity with a response y which is a straight line.

The R2 is a best-fit straight-line slope, which perches as close as possible to the data points collection on an x-y scatter plot, where the x-axis indicates independent variable values and the y-axis indicates dependent variable values. The best-fit line with a non-zero y-intercept can be utilized to forecast the dependent variable values in connection with the slope. The simple linear regression formula is:

$$\hat{y} = B_0 + B_1 x + \varepsilon, \qquad (1)$$

where $\hat{y}$ is forecasted value;

B_0 is the slope;

B_1 is the value of x and they are called regression coefficients;

ε is an error term that reconciles variances among actual values and forecasted values and corresponds to $y - \hat{y}$ [13].

Simplicity and usefulness interpretation are characteristics of these coefficients. The change in the distribution mean of y occurred by a unit x change is defined by slope B_1. For x = 0 in the data range on x, the intercept B_0 is the distribution mean of the response y. For data range on x without including zero, B_0 does not have practical interpretation [14].

Let's assume that we have a set of n samples of paired observations (x_i, y_i) (i =1, 2, ..., n). These observations are considered to fit the simple linear regression model; thus, we have the equation as follows:

$$y_i = B_0 + B_1 x_i + \varepsilon_i, \ (i = 1, 2, \ldots, n). \qquad (2)$$

The least squares basically estimate B_0 and B_1 parameters by minimizing the summation of the squares of the variance among the observations and the scatter diagram line. This might be observed from different perspectives. Direct regression is defined as when the vertical difference among the observations and the scatter diagram line is considered, and its squares summation is minimized to obtain B_0 and B_1 estimation. The ordinary least square's estimation is another name for this method [15].

The summation of the squares is minimized in direct regression method.

$$S(B_0, B_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - B_0 - B_i x_i)^2, \qquad (3)$$

for B_0 and B_1 respectively. B_0 and B_1 solutions are calculable through the formula as follows:

$$\delta S(B_0, B_1) / \delta B_0 = 0, \qquad (4)$$

$$\delta S(B_0, B_1) / \delta B_1 = 0. \qquad (5)$$

Direct regression estimators or ordinary least squares (OLS) estimators of B_0 and B_1 are the names given to solutions of these two equations.

Alternatively, the reverse regression method is defined as when the summation of the squares of the variance among the observations and the scatter diagram line in the horizontal direction can be minimized to obtain B_0 and B_1 estimation [16].

The reverse regression equation is as follows:

$$x_i = B^*_0 + B^*_1 y_i + \delta_i \text{ ,}(i = 1, 2, \ldots, n), \qquad (6)$$

where δ_i are the random error associated components.

The reverse regression estimates $\hat{B}_{OR}$ of B^*_0 and $\hat{B}_{1R}$ of B^*_1 which are reached by interchanging the x and y in the B_0 and B_1 direct regression estimators. The estimates formulas are as follows:

$$\hat{B}_{OR} = \bar{x} - \hat{B}_{1R}\, \bar{y}, \qquad (7)$$

$$\hat{B}_{1R} = S_{yy} / S_{xy} \qquad (8)$$

for B_0 and B_1 respectively. The residual squares summation in this case is

$$SS^*_{res} = S_{xx} - S^2_{xy} / S_{yy}, \qquad (9)$$

where b_1 is the slope parameter direct regression estimator and the correlation coefficient among x and y is r_{xy}.

Subsequently, if r^2_{xy} goes close to 1, the two regression lines will get closer to one another. Reverse regression method is feasible to be utilized in solving the calibration problem [17].

1.2. Multiple Linear Regression

Practically, a meaningful model includes more than one variable. Simultaneously, one regression principle is to use as few variables as possible and merely comprising the most important explanatory variables [18].

The general multiple linear regression model formula with K explanatory variables is as follows:

$$y = B_1x_1 + B_2x_2 + \ldots + B_kx_k + \varepsilon, \qquad (10)$$

where there are n observations of the outcome y, and for each one there are the corresponding explanatory variables values $(x_1, \ldots, x_K)$. The x_{ij} indicates the variable value x_i corresponding to the j-th observation.

The estimates basically minimize the squared residuals summation

$$\sum_{i=1}^{n} \varepsilon_i^2, \qquad (11)$$

where

$$\varepsilon_i = y_i - \hat{y}_i, \qquad (12)$$

$$\hat{y}_i = \hat{B}_0 + \hat{B}_1x_{1i} + \ldots + \hat{B}_kx_{ki}. \qquad (13)$$

The basic least squares assumptions are similar to simple regression except it has one more assumption which prohibits the redundancy possibility among the explanatory variables [19]. For instance, it's not possible to have two variables that contain different units but totally the same information.

Multicollinearity is defined as the redundancy possibility among the explanatory variables. In its presence, it degrades the precision of the regression model. Also, one important point is to know that the estimated slopes $\hat{B}_i$ depend on which including variables. Adding and deleting variables changes the other $\hat{B}_i$ [20].

1.3. SIR Model

The Susceptible Infected Recovered or SIR model was developed in the twentieth century by Ronald Ross, William Hamer, and others. This model includes a three-coupled nonlinear ordinary differential equations system. Kermack and McKendrinck's theoretical papers from 1927 to 1933 have had an impressive impact in modeling infectious diseases mathematically [21]. Increasingly utilizing mathematical models since then has led to clarifying several diseases transmission. Studying SIR models is crucial in enhancing the fundamental knowledge of spreading out infectious diseases [22]. Although these models might seem simple, they evaluate the control program's potential impact in decreasing mortality and morbidity. In the past few years, and enhancement in mathematical models' representation trend has been seen which shows the interdisciplinary importance. The SIR model is a tool which is at first gives a comprehensive understanding of what occurs rapidly; then, based on enhanced knowledge, enriching the model by adding more details in the formulation is possible [23].

The SIR model is a mathematical representation of outbreaking an infection throughout a population over time in a simpler way. Population division into pieces of compartments Forms epidemic models. The SIR model includes three main compartments as follows:

- Susceptible (S): Individuals who are susceptible to infection; possibly including the ones who lose their immunity or immune once. Also, more commonly, in case of newborn infants whose mother has not passed on any immunity because she has never been infected;
- Infected (I): The parasite level is impressively large and potentially there is the chance of infection transition to other individuals;
- Recovered or Resistant (R): All individuals who have recovered after infection.

The acute infection dynamics are captured by this epidemiological model. That confers immunity permanently after recovery. Some diseases for which the SIR model might be implemented are measles, smallpox, chickenpox, mumps, typhoid fever, and diphtheria. As an assumption, the total size of the population is constant, i.e., N = S + I + R. Then the demographic factors inclusion or exclusion studies and distinguishes two cases.

The SIR model basically includes ordinary differential equations (ODEs) and infers to a deterministic model which doesn’t involve the randomness with consecutive time. Similar to the reaction kinetics principles, the assumption is that confrontations between susceptible and infected individuals happen at an adequate rate to their population respective numbers [24]. The new infection rate can thus be described as BSI, where B is the infectivity parameter. Infected individuals are considered to recover anytime with a steady probability, which interprets into a rate of constant per capita recovery that is denoted here with r, and (γI) is the recovery overall rate. Based on the mentioned assumptions, the differential equation form of SIR models is as follows:

$$\frac{dS}{dt} = -B*S*I, \quad \frac{dI}{dt} = B*S*I - \gamma*I, \quad \frac{dR}{dt} = \gamma*I. \quad (14)$$

Another assumption is that population size (S+I+R) is considered as constant and equal to the initial size of it, which is denoted with the N parameter [25].

Typically, three threshold values are utilized in epidemiology. The first value is called the basic reproduction number or basic reproduction ratio or basic reproductive rate, which is denoted by R_0 and considered as the most important value. It is defined as the secondary infections average number that happens when one infective is defined throughout a wholly susceptible population and represents a borderline between persistence and a disease death [26].

σ is called the contact number, and defined as adequate contacts average number of a common infective within a period of infectiousness. If the individual who contacts with the susceptible is infective, the definition of an adequate contact will be formed which is the one who is sufficient for transmission. As an assumption, for the whole infectious period the infected individual is inside it and mingles as same as a native with the host population.

R is called the replacement number and defined as the secondary infections average number turned out by a common infective within the whole infectiousness period. It performs as a function of time (t) by disease evolution after the initial invasion.

These three threshold values are all equal at the inception of an infectious disease breakout [27].

R is the secondary cases actual number from a common infective so that it is always less than the R_0 after everyone is infected. Furthermore, after spreading out the infection, the susceptible fraction becomes less than one, so that all adequate contacts result partially in a new case and therefore R becomes less than σ [28].

The result is as follows:

$$R_0 \geq \sigma \geq R. \quad (15)$$

An epidemic will occur if an individual infected found in the population and if dI / dt > 0. Replacing S with N in Equation (15) results as BN / γ > 1. The equation will be as follows [16]:

$$R_0 = \frac{BN}{\gamma}. \quad (16)$$

2. Results

2.1. Experimental Study of Linear Regression Model

The Ministry Public Health Center of Ukraine provided data, reports the cumulative recorded death numbers and medical tests since the start of the COVID-19 pandemic in Ukraine respectively. To illustrate the COVID-19 daily confirmed cases increasing number, performed tests, recovered cases and deaths, a Regression model was built and the outputs were saved into two new fields. Next, a list of all the dates was created and converted from string to date-time format. The daily confirmed cases increasing number recovered cases, death cases, and performed tests are shown below respectively. The Linear Regression model was built in a Python environment and all implementations were done inside it. The obtained results are shown in figures 1-3.

Figures 1-3 show an ascending trend by the time which is somehow expected. By the time has passed the amount of confirmed, recovered, death, and performed tests have increased. The trends don’t suffer from considerable bias. Thus, there is no need to normalize them.

A range of 30 days from 30/01/2021 until 01/03/2021 was considered for our investigation and executing linear regression to forecast the first 15 days of March (from 02/03 until 15/03). obtained determination coefficient for ‘Confirmed’, ‘Recovered’ and ‘Death cases’ are (0.98889), (0.98877), (0.98718) respectively. The plot for each variable is shown in figures 4-6. Figures 4-6 show an ascending trend which was expected.

The upward trend in morbidity and mortality indicates the natural course of the COVID-19 epidemic process in Ukraine. From the upward trend, it can be concluded that the anti-epidemic measures taken within the simulated time frame are ineffective. On the other hand, the upward trend of the recovered, which has fewer dynamics than the predicted values of the sick, indicates a further increase in the total number of patients in the simulated territory.

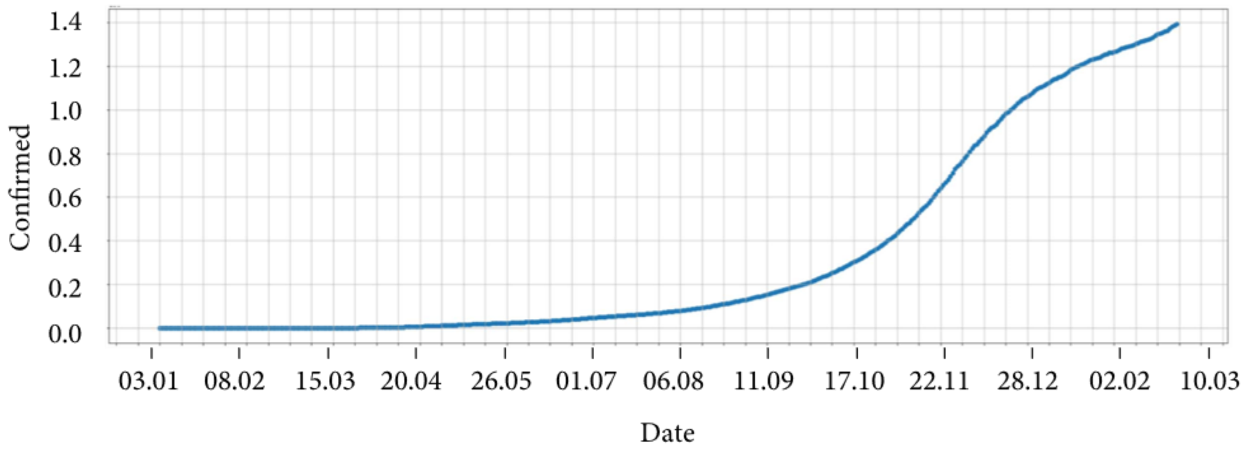


Fig. 1. Confirmed cases trend based on COVID-19 actual dataset until 01/03/2021

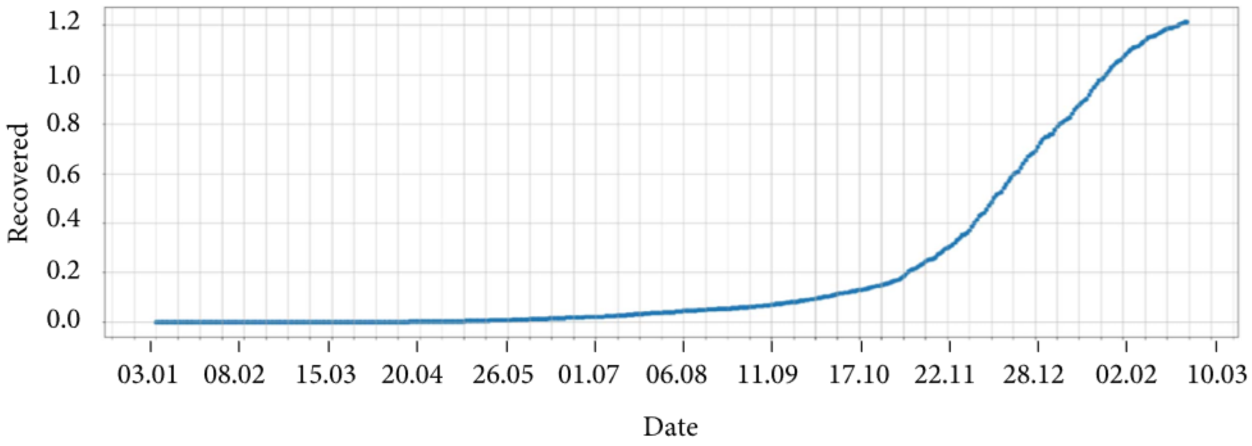


Fig. 2. Recovered cases trend based on COVID-19 actual dataset until 01/03/2021

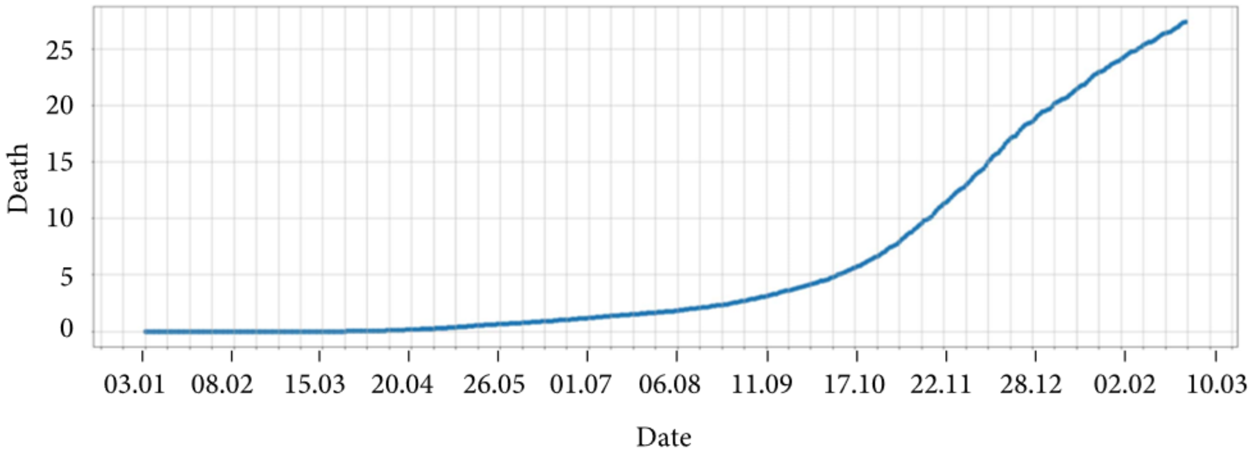


Fig. 3. Death cases trend based on COVID-19 actual dataset until 01/03/2021

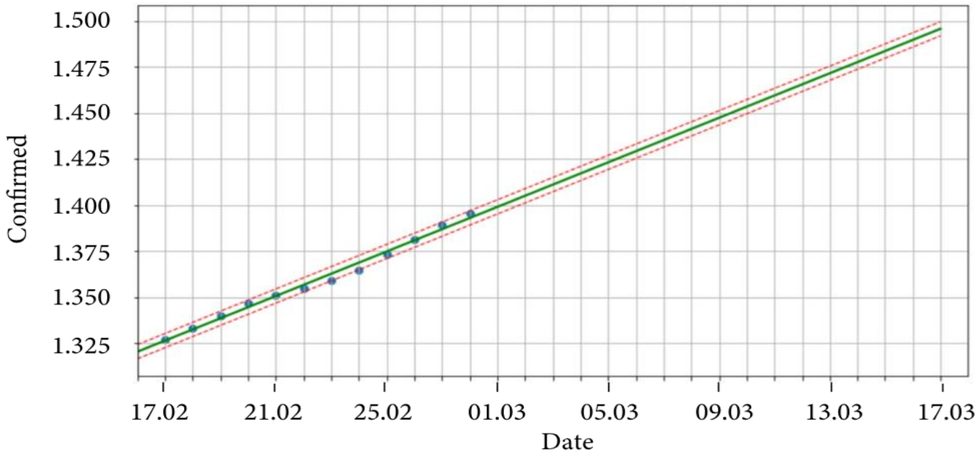


Fig. 4. “Confirmed cases” trend in 30 days; dots are actual data

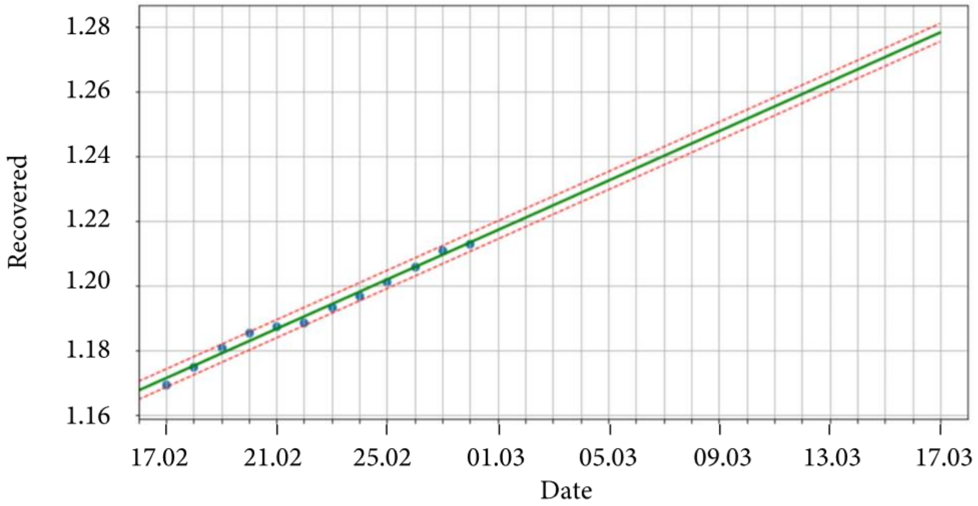


Fig. 5. “Recovered cases” trend in 30 days; dots are actual data

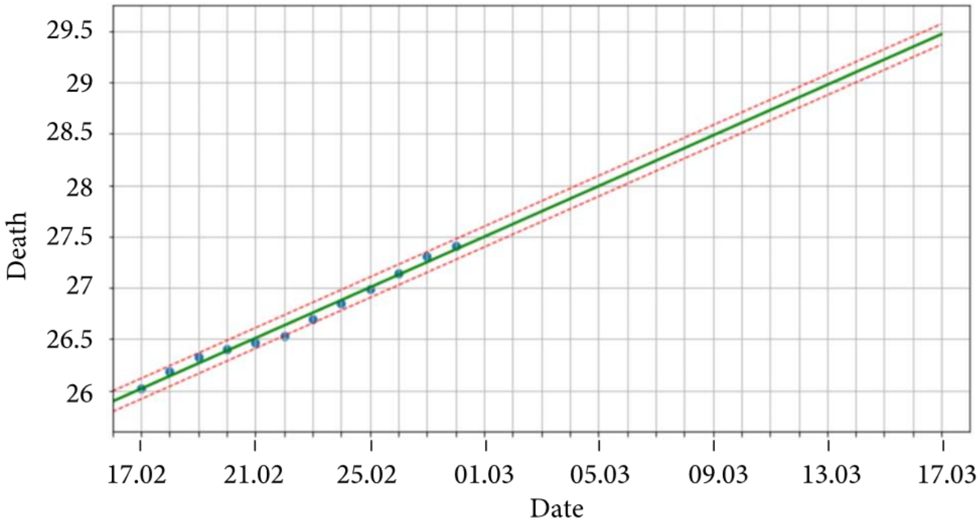


Fig. 6. “Death cases” trend in 30 days; dots are actual data

Now let’s have a look at the obtained forecasted numbers for each variable based on linear regression model in a table 1.

Table 1

Linear regression forecasted results

Date	Confirmed	Recovered	Deaths
2-Mar	1404965	1220933	27618
3-Mar	1411007	1224742	27742
4-Mar	1417049	1228551	27865
5-Mar	1423091	1232360	27989
6-Mar	1429134	1236170	28112
7-Mar	1435176	1239979	28235
8-Mar	1441218	1243788	28359
9-Mar	1447260	1247597	28482
10-Mar	1453303	1251406	28606
11-Mar	1459345	1255215	28729
12-Mar	1465387	1259024	28852
13-Mar	1471429	1262833	28976
14-Mar	1477472	1266642	29099
15-Mar	1483514	1270451	29223
16-Mar	1489556	1274260	29346

2.2. Experimental Study of SIR Model

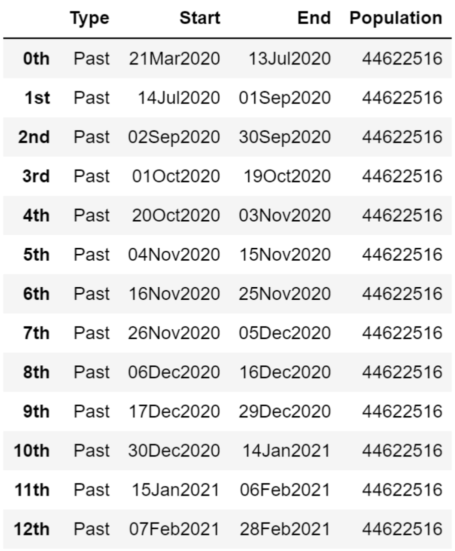
The COVID-19 epidemic process dynamics dependence and interepidemic countermeasures were investigated by our SIR developed model and the influence of each factor on the dynamics was determined. As a matter of using a more developed model and thus having the capability of including more complexity and factors, the SIR-F was used for this research instead of the classic SIR model. ‘F’ represents the ‘Fatal with confirmation’ in the SIR-F model.

Figure 7 illustrates the COVID-19 epidemic process in different phases in Ukraine based on the implementation of the SIR-F model on COVID-19 dynamics changes started from the date each stage.

The final step is to illustrate the forecasting results of COVID-19 by SIR models. Figure 8 shows the trend of COVID-19 ‘Infected’, ’Fatal’ and ‘Recovered’ cases. It is characterized by increasing “Infected” and “Recovered” by December 2020 and by increasing “Recovered” with decreasing “Infected” after December 2020.

Based on these separated phases we can compare the results of SIR-F model parameters estimation. The results are shown in figure 9.

Table 2 shows the forecasted numbers in each case.



	Type	Start	End	Population
0th	Past	21Mar2020	13Jul2020	44622516
1st	Past	14Jul2020	01Sep2020	44622516
2nd	Past	02Sep2020	30Sep2020	44622516
3rd	Past	01Oct2020	19Oct2020	44622516
4th	Past	20Oct2020	03Nov2020	44622516
5th	Past	04Nov2020	15Nov2020	44622516
6th	Past	16Nov2020	25Nov2020	44622516
7th	Past	26Nov2020	05Dec2020	44622516
8th	Past	06Dec2020	16Dec2020	44622516
9th	Past	17Dec2020	29Dec2020	44622516
10th	Past	30Dec2020	14Jan2021	44622516
11th	Past	15Jan2021	06Feb2021	44622516
12th	Past	07Feb2021	28Feb2021	44622516

Fig. 7. Different phases in Ukraine.

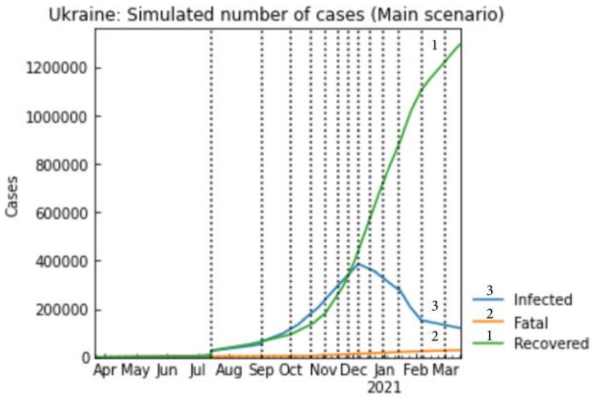
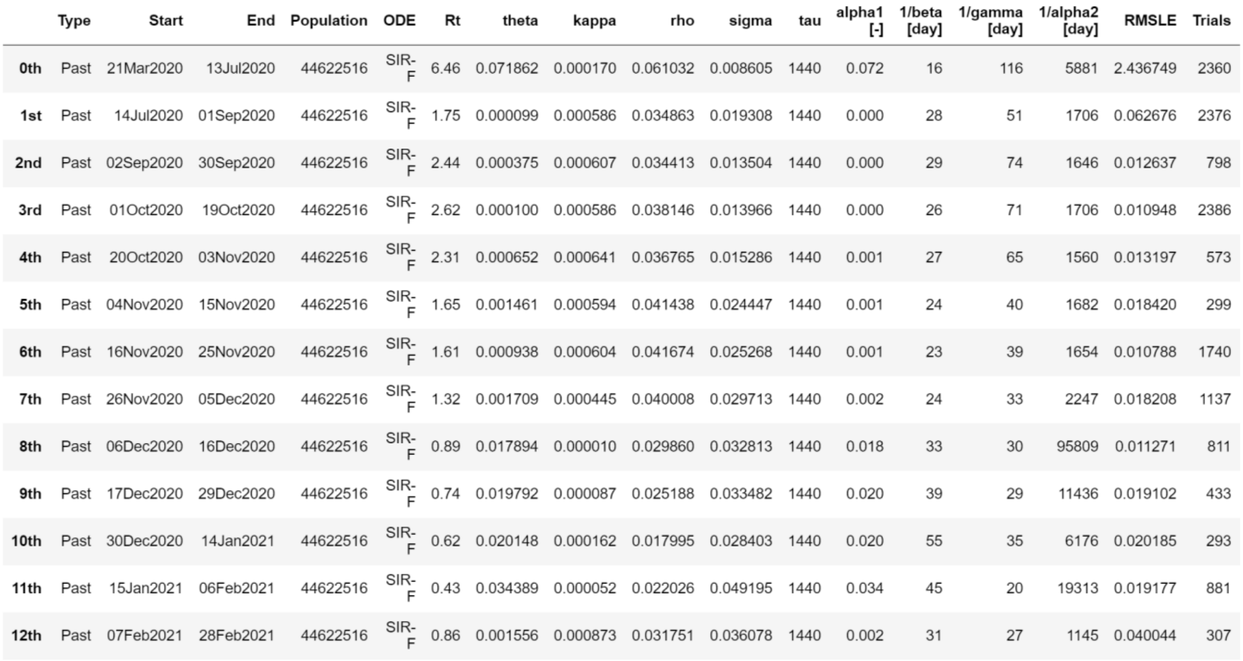


Fig. 8. The illustration of COVID-19 trends of ‘Infected’, ‘Fatal’, ‘Recovered’ cases in Ukraine until 16/03/2021

Table 2

SIR model forecasted results

Date	Confirmed	Recovered	Deaths
2-Mar	1389629	1229196	27846
3-Mar	1393695	1233965	27968
4-Mar	1397735	1238704	28089
5-Mar	1401749	1243413	28209
6-Mar	1405739	1248093	28329
7-Mar	1409702	1252744	28447
8-Mar	1413641	1257366	28565
9-Mar	1417557	1261960	28683
10-Mar	1421445	1266524	28799
11-Mar	1425310	1271060	28915
12-Mar	1429151	1275568	29030
13-Mar	1432965	1280047	29144
14-Mar	1436758	1284499	29258
15-Mar	1440526	1288923	29371
16-Mar	1444269	1293319	29483



	Type	Start	End	Population	ODE	Rt	theta	kappa	rho	sigma	tau	alpha1 [-]	1/beta [day]	1/gamma [day]	1/alpha2 [day]	RMSLE	Trials
0th	Past	21Mar2020	13Jul2020	44622516	SIR-F	6.46	0.071862	0.000170	0.061032	0.008605	1440	0.072	16	116	5881	2.436749	2360
1st	Past	14Jul2020	01Sep2020	44622516	SIR-F	1.75	0.000099	0.000586	0.034863	0.019308	1440	0.000	28	51	1706	0.062676	2376
2nd	Past	02Sep2020	30Sep2020	44622516	SIR-F	2.44	0.000375	0.000607	0.034413	0.013504	1440	0.000	29	74	1646	0.012637	798
3rd	Past	01Oct2020	19Oct2020	44622516	SIR-F	2.62	0.000100	0.000586	0.038146	0.013966	1440	0.000	26	71	1706	0.010948	2386
4th	Past	20Oct2020	03Nov2020	44622516	SIR-F	2.31	0.000652	0.000641	0.036765	0.015286	1440	0.001	27	65	1560	0.013197	573
5th	Past	04Nov2020	15Nov2020	44622516	SIR-F	1.65	0.001461	0.000594	0.041438	0.024447	1440	0.001	24	40	1682	0.018420	299
6th	Past	16Nov2020	25Nov2020	44622516	SIR-F	1.61	0.000938	0.000604	0.041674	0.025268	1440	0.001	23	39	1654	0.010788	1740
7th	Past	26Nov2020	05Dec2020	44622516	SIR-F	1.32	0.001709	0.000445	0.040008	0.029713	1440	0.002	24	33	2247	0.018208	1137
8th	Past	06Dec2020	16Dec2020	44622516	SIR-F	0.89	0.017894	0.000010	0.029860	0.032813	1440	0.018	33	30	95809	0.011271	811
9th	Past	17Dec2020	29Dec2020	44622516	SIR-F	0.74	0.019792	0.000087	0.025188	0.033482	1440	0.020	39	29	11436	0.019102	433
10th	Past	30Dec2020	14Jan2021	44622516	SIR-F	0.62	0.020148	0.000162	0.017995	0.028403	1440	0.020	55	35	6176	0.020185	293
11th	Past	15Jan2021	06Feb2021	44622516	SIR-F	0.43	0.034389	0.000052	0.022026	0.049195	1440	0.034	45	20	19313	0.019177	881
12th	Past	07Feb2021	28Feb2021	44622516	SIR-F	0.86	0.001556	0.000873	0.031751	0.036078	1440	0.002	31	27	1145	0.040044	307

Fig. 9. Experimental study of SIR model parameters based on separation to phases

2.3. Models Performance Comparison

In this part we will compare the obtained results by each model with the actual data in a range of fifteen days (from 02/03/2021 to 16/03/2021). Table 3 shows the percent error and mean error caused by comparing linear regression model results and actual data.

Table 3

Error obtained by linear regression model forecast

Date	Confirmed	Recovered	Deaths
2-Mar	0.0305346	-0.074124	0.115866
3-Mar	0.1233871	0.057237	0.367674
4-Mar	0.4148057	0.162468	0.64956
5-Mar	0.7103551	0.201159	0.850334
6-Mar	0.9322429	0.174005	0.996016
7-Mar	1.0148581	0.045485	0.945635
8-Mar	0.9854859	-0.11875	0.906238
9-Mar	0.7963324	-0.341617	0.758374
10-Mar	0.8230906	-0.167572	1.115151
11-Mar	1.0351904	-0.02406	1.622054
12-Mar	1.5088164	0.06092	2.017191
13-Mar	2.001048	0.165738	2.443401
14-Mar	2.2004478	0.087791	2.529297
15-Mar	2.2481082	0.065489	2.552784
16-Mar	2.4865799	0.262348	3.043004
mean error	1.1540855	0.037101	1.394172

For ‘Confirmed’, ‘Recovered’ and ‘Death’ cases mean errors have almost 1.15, 0.037 and 1.39 percent deviant respectively. This means the linear regression model forecasted impressively precise. Table 4 shows the percent error and mean error caused by comparing SIR model results and actual data.

Table 4 shows that or ‘Confirmed’, ‘Recovered’ and ‘Death’ cases mean errors have almost 3.29, 1.08, and 0.71 percent deviant respectively. This means although the SIR model forecasted impressively precise, the error deviation has a higher number in ‘Confirmed’ and ‘Recovered’ cases forecasted by the linear regression model. Thus, we can acclaim that linear regression model has forecasted more precisely in a short time range (fifteen days) than the SIR model. The linear regression model performed reasonably acceptable in short time range close enough to actual data.

Table 4

Error obtained by SIR model forecast

Date	Confirmed	Recovered	Deaths
2-Mar	1.1344755	-0.745853	-0.70387
3-Mar	1.3670853	-0.690619	-0.44336
4-Mar	1.8023445	-0.658511	-0.15308
5-Mar	2.243697	-0.689554	0.063809
6-Mar	2.6120069	-0.782954	0.222387
7-Mar	2.8402457	-0.973942	0.193342
8-Mar	2.9554887	-1.197344	0.17854
9-Mar	2.9083839	-1.475879	0.052296
10-Mar	3.0827784	-1.359232	0.437515
11-Mar	3.4478114	-1.270357	0.968356
12-Mar	4.0825637	-1.236861	1.391664
13-Mar	4.7389853	-1.181285	1.852869
14-Mar	5.0965437	-1.303621	1.97211
15-Mar	5.2993837	-1.368584	2.036022
16-Mar	5.7001847	-1.215168	2.56419
mean error	3.2874652	-1.076651	0.708852

Conclusions

The paper describes experimental research on two approaches to epidemic process simulation, which are based on the linear regression method and compartment modeling approach. Models are verified and investigated on COVID-19 morbidity and mortality data in Ukraine provided by the Center for Public Health of Ministry of Health of Ukraine.

The novelty of the research is the development of epidemic process models based on state-of-art methods and approaches applied to novel emergence disease COVID-19. The main difference of the proposed study is that the epidemic process of COVID-19 has not been investigated, appeared suddenly, and spread rapidly across the planet. This dictates the development of new methods for modeling epidemic processes that investigate diseases for which there is no sufficient amount of data.

The disadvantages of the SIR model may be because the coronavirus mutates and new strains appear, immunity after a previous illness is not always long-lasting, as evidenced by cases of repeated COVID-19 disease, one dose of the vaccine does not lead to the development of long-term and intense immunity, a very small proportion the population of Ukraine has post-vaccination immunity.

At this stage in the development of the epidemic process of COVID-19, it is more expedient to use a linear model to predict the incidence rate, which has shown higher accuracy and efficiency.

Generally, the reason for that lies in the fact that the utilized linear regression model for this research was implemented on merely 30 days (from fifteen days before the 2nd of March) and not the whole dataset of COVID-19.

Acknowledgment. The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management” [29].

Future research directions

It is expected that if we try to forecast in longer time ranges, the linear regression model will lose precision. Alternatively, since the SIR model is more comprised in including more factors, the model is expected to perform better in forecasting longer time ranges. These assumptions can be investigated in detail in further researches.

Also, further researches are aimed to develop an ensemble methodology of epidemic process investigation which will combine both compartment and machine learning methods. It is necessary to increase the accuracy with machine learning methods and make it possible to investigate factors influencing the epidemic process with compartment models.

References (GOST 7.1:2006)

1. Gorbenko, A. Exploring timeout as a performance and availability factor of distributed replicated database systems [Text] / A. Gorbenko, O. Tarasyuk // Radioelectronic and Computer systems. – 2020. – No. 4 (96). – P. 98-105. DOI: 10.32620/reks.2020.4.09

2. Wawrzynski, T. Artificial intelligence and cyberculture [Text] / T. Wawrzynski // Radioelectronic and Computer systems. – 2020. – Vol. 3, iss. 95. – P. 20-26. DOI: 10.32620/reks.2020.3.02.

3. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method [Text] / I. Izonin, R. Tkachenko, I. Dronyuk, P. Tkachenko, M. Gregus, M. Rashkevych // Mathematical Biosciences and Engineering. – 2021. – Vol. 18, iss. 3. – P. 2599-2613. DOI: 10.3934/mbe.2021132.

4. Liang, J. Multivariate linear regression method based on SPSS analysis of influencing factors of CPI during epidemic situation [Text] / J. Liang // 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME). – 2020. – P. 294-297. DOI: 10.1109/ICEMME51517.2020.00062.

5. Li, J. Construction of Big Data Epidemic Forecast and Propagation Model and Analysis of Risk Visualization Trend [Text] / J. Li // 2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI). – 2020. – P. 21-25. DOI: 10.1109/ICAACI50733.2020.00009.

6. Treatment effectiveness and outcome in patients with a relapse and newly diagnosed multidrug-resistant pulmonary tuberculosis [Text] / D. Butov, V. Myasoedov, M. Gumeniuk, G. Gumeniuk, O. Choporova, A. Tkachenko, O. Akymenko, O. Borysova, O. Goptsii, Y. Vorobiov, T. Butova // Medicinski Glasnik. – 2020. – Vol. 17, iss. 2. – P. 356–362. DOI: 10.17392/1179-20.

7. Anaplasmosis: Experimental immunodeficient state model [Text] / A. V. Bondarenko, S. I. Pokhil, M. V. Lytvynenko, T. V. Bocharova, V. V. Gargin // Wiadomosci Lekarskie. – 2019. – Vol. 72, iss. 9-2. – P. 1761-1764.

8. Kumari, K. Linear regression analysis study [Text] / K. Kumari, S. Yadav // Journal of the Practice of Cardiovascular Sciences. – 2018. – Vol. 4, iss. 1. – P. 33-36. DOI: 10.4103/jpcs.jpcs_8_18.

9. Method of Data Openness Estimation Based on User-Experience in Infocommunication Systems of Municipal Enterprises [Text] / V. Yesina, N. Matveeva, I. Chumachenko, N. Manakova // 2018 International Scientific-Practical Conference on Problems of Info-

communications Science and Technology, PIC S and T 2018 – Proceedings. – 2019. – P. 171–176. DOI: 10.1109/INFOCOMMST.2018.8631897.

10. Hussein, B. A. A Modeling and Simulation Approach to Analyze and Control Transition States in Epidemic Models [Text] / B. A. Hussein, S. T. Hasson // 2019 2nd International Conference on Engineering Technology and its Applications (IICETA). – 2019. – P. 94-98. DOI: 10.1109/IICETA47481.2019.9012976.

11. Dhaka, A. Comparative Analysis of Epidemic Alert System using Machine Learning for Dengue and Chikungunya [Text] / A. Dhaka, P. Singh // 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). – 2020. – P. 798-804. DOI: 10.1109/Confluence47617.2020.9058048.

12. Research on the relationship between Covid-19 epidemic and gold price trend based on Linear Regression Model [Text] / Y. Jianyi, W. Chenyang, H. Yupeng, L. Zicheng // 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). – 2020. – P. 1796-1798. DOI: 10.1109/ITAIC49862.2020.9338828.

13. Using TensorFlow to Establish multivariable linear regression model to Predict Gestational Diabetes [Text] / Y. Zou, X. Gong, P. Miao, Y. Liu // 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). – 2020. – P. 1695-1698. DOI: 10.1109/ITNEC48623.2020.9084664.

14. Sharma, A. Linear Regression Model for Agile Software Development Effort Estimation [Text] / A. Sharma, N. Chaudhary // 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE). – 2020. – P. 1-4. DOI: 10.1109/ICRAIE51050.2020.9358309.

15. Fedushko, S. Operational Intelligence Software Concepts for Continuous Healthcare Monitoring and Consolidated Data Storage Ecosystem [Text] / S. Fedushko, T. Ustyianovych // Advances in Intelligent Systems and Computing. – 2021. – Vol. 1247. – P. 545–557. DOI: 10.1007/978-3-030-55506-1_49.

16. Liu, T. U.S. Pandemic Prediction Using Regression and Neural Network Models [Text] / T. Liu // 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI). – 2020. – P. 351-354. DOI: 10.1109/ICHCI51889.2020.00080.

17. Mandayam, A. U. Prediction of Covid-19 pandemic based on Regression [Text] / A. U. Mandayam, S. Siddesha, S. K. Niranjan // 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). – 2020. – P. 1-5. DOI: 10.1109/ICRCICN50933.2020.9296175.

18. Research on Time Series Problem Model Based on Dynamic Network NAR and Multiple Regression [Text] / Z. Liu, J. Zuo, R. Lv, Y. Sun, H. Kang // 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE). – 2020. – P. 416-419.

19. Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network [Text] / H. Xue, Y. Bai, H. Hu, H. Liang // IEEE Access. – 2018. – Vol. 6. – P. 563-575. DOI: 10.1109/ACCESS.2017.2771798.

20. Green computing and communications in critical application domains: Challenges and solutions [Text] / V. Kharchenko, A. Gorbenko, V. Sklyar, C. Phillips // International Conference on Digital Technologies. – 2013. – P. 191-197. DOI: 10.1109/DT.2013.6566310.

21. Akman, C. Covid-19 SEIQR Spread Mathematical Model [Text] / C. Akman, O. Demir, T. Sönmez // 2021 29th Signal Processing and Communications Applications Conference (SIU). – 2021. – P. 1-4, DOI: 10.1109/SIU53274.2021.9477975.

22. Sano, H. State Estimation of Kermack-McKendrick PDE Model With Latent Period and Observation Delay [Text] / H. Sano, M. Wakaiki // IEEE Transactions on Automatic Control. – 2020. – Vol. 66, No. 10. – P. 4982-4989. DOI: 10.1109/TAC.2020.3047360.

23. Guo, Y. Global stability analysis of a class of SIRS models with nonlinear incidence [Text] / Y. Guo, N. Liu, H. Jiao // 2020 International Conference on Public Health and Data Science (ICPHDS). – 2020. – P. 269-272. DOI: 10.1109/ICPHDS51617.2020.00059.

24. Machine Learning Algorithms for Binary Classification of Liver Disease [Text] / A. Sokoliuk, G. Kondratenko, I. Sidenko, Y. Kondratenko, A. Khomchenko, I. Atamanyuk // 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T). – 2020. – P. 417-421. DOI: 10.1109/PICST51311.2020.9468051.

25. Nan, X. Dynamic Crowd Aggregation Simulation Using SIR Model Based Emotion Contagion [Text] / X. Nan, Z. Zehong, P. Zhigeng // 2017 International Conference on Virtual Reality and Visualization (ICVRV). – 2017. – P. 352-353. DOI: 10.1109/ICVRV.2017.00080.

26. Rodrigues, H. S. Application of SIR epidemiological model: new trends [Text] / H. S. Rodrigues // International Journal of Applied Mathematics and Informatics. – 2016. – Vol. 10. – P. 92–97.

27. Yeling, L. SIR Infectious Disease Model Based on Age Structure and Constant Migration Rate and its Dynamics Properties [Text] / L. Yeling, W. Jing // 2020 International Conference on Public Health and Data Science (ICPHDS). – 2020. – P. 158-165. DOI: 10.1109/ICPHDS51617.2020.00039.

28. Yang, Y. Mathematical Models and Control Methods of Infectious Diseases [Text] / Y. Yang, H. Zhang // 2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE). – 2020. – P. 383-388. DOI: 10.1109/CACRE50138.2020.9230170.

29. The concept of developing a decision support system for the epidemic morbidity control [Text] / S. Yakovlev, K. Bazilevych, D. Chumachenko, T. Chumachenko, L. Hulianytskyi, I. Meniailov, A. Tkachenko // CEUR Workshop Proceedings. – 2020. – Vol. 2753. – P. 265–274.

References (BSI)

1. Gorbenko, A., Tarasyuk, O. Exploring timeout as a performance and availability factor of distributed replicated database systems. *Radioelectronic and Computer systems*, 2020, no. 4 (96), pp. 98-105. DOI: 10.32620/reks.2020.4.09.

2. Wawrzynski, T. Artificial intelligence and cyberculture. *Radioelectronic and Computer systems*, 2020, vol. 3, iss. 95, pp. 20-26. DOI: 10.32620/reks.2020.3.02

3. Izonin, I., Tkachenko, R., Dronyuk, I., Tkachenko, P., Gregus, M., Rashkevych, M. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. *Mathematical Biosciences and Engineering*, 2021, vol. 18, iss. 3, pp. 2599-2613. DOI: 10.3934/mbe.2021132.

4. Liang, J. Multivariate linear regression method based on SPSS analysis of influencing factors of CPI during epidemic situation. *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, 2020, pp. 294-297, DOI: 10.1109/ICEMME51517.2020.00062.

5. Li, J. Construction of Big Data Epidemic Forecast and Propagation Model and Analysis of Risk Visualization Trend. *2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI)*, 2020, pp. 21-25, DOI: 10.1109/ICAACI50733.2020.00009.

6. Butov, D., Myasoedov, V., Gumeniuk, M., Gumeniuk, G., Choporova, O., Tkachenko, A., Akymenko, O., Borysova, O., Goptsii, O., Vorobiov, Y., Butova, T. Treatment effectiveness and outcome in patients with a relapse and newly diagnosed multidrug-resistant pulmonary tuberculosis. *Medicinski Glasnik*, 2020, vol. 17, iss. 2, pp. 356-362. DOI: 10.17392/1179-20.

7. Bondarenko, A. V., Pokhil, S. I., Lytvynenko, M. V., Bocharova, T. V., Gargin, V. V. Anaplasmosis: Experimental immunodeficient state model, *Wiadomosci Lekarskie*, 2019, vol. 72, iss. 9-2, pp. 1761-1764.

8. Kumari, K., Yadav, S. Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, 2018, vol. 4, iss. 1, pp. 33-36. DOI: 10.4103/jpcs.jpcs_8_18.

9. Yesina, V., Matveeva, N., Chumachenko, I., Manakova, N. Method of Data Openness Estimation Based on User-Experience in Infocommunication Systems of Municipal Enterprises. *2018 International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 – Proceedings*, 2019, pp. 171–176. DOI: 10.1109/INFOCOMMST.2018.8631897.

10. Hussein, B. A., Hasson, S. T. A Modeling and Simulation Approach to Analyze and Control Transition States in Epidemic Models. *2019 2nd International Conference on Engineering Technology and its Applications (IICETA)*, 2019, pp. 94-98, DOI: 10.1109/IICETA47481.2019.9012976.

11. Dhaka, A., Singh, P. Comparative Analysis of Epidemic Alert System using Machine Learning for Dengue and Chikungunya. *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020, pp. 798-804, DOI: 10.1109/Confluence47617.2020.9058048.

12. Jianyi, Y., Chenyang, W., Yupeng, H., Zicheng, L. Research on the relationship between Covid-19 epidemic and gold price trend based on Linear Regression Model. *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 2020, pp. 1796-1798. DOI: 10.1109/ITAIC49862.2020.9338828.

13. Zou, Y., Gong, X., Miao, P., Liu, Y. Using TensorFlow to Establish multivariable linear regression model to Predict Gestational Diabetes. *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2020, pp. 1695-1698. DOI: 10.1109/ITNEC48623.2020.9084664.

14. Sharma, A., Chaudhary, N. Linear Regression Model for Agile Software Development Effort Estimation. *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 2020, pp. 1-4. DOI: 10.1109/ICRAIE51050.2020.9358309.

15. Fedushko, S., Ustyianovych, T. Operational Intelligence Software Concepts for Continuous Healthcare Monitoring and Consolidated Data Storage Ecosystem. *Advances in Intelligent Systems and Computing*, 2021, vol. 1247, pp. 545-557. DOI: 10.1007/978-3-030-55506-1_49.

16. Liu, T. U.S. Pandemic Prediction Using Regression and Neural Network Models. *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, 2020, pp. 351-354, DOI: 10.1109/ICHCI51889.2020.00080.

17. Mandayam, A. U., Siddesha, S., Niranjan, S. K. Prediction of Covid-19 pandemic based on Regression. *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 2020, pp. 1-5. DOI: 10.1109/ICRCICN50933.2020.9296175.

18. Liu, Z., Zuo, J., Lv, R., Sun, Y., Kang, H. Research on Time Series Problem Model Based on Dynamic Network NAR and Multiple Regression. *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, 2020, pp. 416-419. DOI: 10.1109/ICAICE51518.2020.00088.

19. Xue, H., Bai, Y., Hu, H., Liang, H. Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network. *IEEE Access*, 2018, vol. 6, pp. 563-575. DOI: 10.1109/ACCESS.2017.2771798.

20. Kharchenko, V., Gorbenko, A., Sklyar, V., Phillips, C. Green computing and communications in critical application domains: Challenges and solutions. *International Conference on Digital Technologies*, 2013, pp. 191-197. DOI: 10.1109/DT.2013.6566310.

21. Akman, C., Demir, O., Sönmez, T. Covid-19 SEIQR Spread Mathematical Model. *2021 29th Signal Processing and Communications Applications Conference (SIU)*, 2021, pp. 1-4, DOI: 10.1109/SIU53274.2021.9477975.

22. Sano, H., Wakaiki, M. State Estimation of Kermack-McKendrick PDE Model With Latent Period and Observation Delay. *IEEE Transactions on Automatic Control*, 2020, vol. 66, no. 10, pp. 4982-4989.

DOI: 10.1109/TAC.2020.3047360.

23. Guo, Y., Liu, N., Jiao, H. Global stability analysis of a class of SIRS models with nonlinear incidence. *2020 International Conference on Public Health and Data Science (ICPHDS),* 2020, pp. 269-272. DOI: 10.1109/ICPHDS51617.2020.00059.

24. Sokoliuk, A., Kondratenko, G., Sidenko, I., Kondratenko, Y., Khomchenko, A., Atamanyuk, I. Machine Learning Algorithms for Binary Classification of Liver Disease. *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T),* 2020, pp. 417-421. DOI: 10.1109/PICST51311.2020.9468051.

25. Nan, X., Zehong, Z., Zhigeng, P. Dynamic Crowd Aggregation Simulation Using SIR Model Based Emotion Contagion. *2017 International Conference on Virtual Reality and Visualization (ICVRV),* 2017, pp. 352-353. DOI: 10.1109/ICVRV.2017.00080

26. Rodrigues, H. S. Application of SIR epidemiological model: new trends. *International Journal of Applied Mathematics and Informatics*, 2016, vol. 10, pp. 92-97.

27. Yeling, L., Jing, W. SIR Infectious Disease Model Based on Age Structure and Constant Migration Rate and its Dynamics Properties. *2020 International Conference on Public Health and Data Science (ICPHDS),* 2020, pp. 158-165. DOI: 10.1109/ ICPHDS51617.2020.00039.

28. Yang, Y., Zhang, H. Mathematical Models and Control Methods of Infectious Diseases. *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE),* 2020, pp. 383-388. DOI: 10.1109/CACRE50138.2020.9230170.

29. Yakovlev, S., Bazilevych, K., Chumachenko, D., Chumachenko, T., Hulianytskyi, L., Meniailov, I., Tkachenko, A. The concept of developing a decision support system for the epidemic morbidity control. *CEUR Workshop Proceedings*, 2020, vol. 2753, pp. 265–274.

Надійшла до редакції 4.08.2021, розглянута на редколегії 23.09.2021

ПОРІВНЯЛЬНЕ ДОСЛІДЖЕННЯ МОДЕЛЕЙ ЛІНІЙНОЇ РЕГРЕСІЇ І SIR ПОШИРЕННЯ COVID-19 В УКРАЇНІ ДО ВАКЦИНАЦІЇ

А. Мохаммаді, Є. С. Меняйлов, К.О. Базілевич, С. В. Яковлев, Д. І. Чумаченко

Глобальна пандемія COVID-19 почалася в грудні 2019 року і стрімко поширилася по всьому світу. У всьому світі захворіли понад 230 мільйонів людей, 4,75 мільйона випадків закінчились летальним результатом. Крім загрози здоров'ю наслідком пандемії стали соціальні проблеми, економічна криза і перехід звичного життя в «нову реальність». Математичне моделювання є ефективним інструментом для контролю епідемічного процесу COVID-19 на заданих територіях. Моделювання дозволяє спрогнозувати майбутню динаміку епідемічного процесу і виявити фактори, які впливають на підвищення захворюваності найбільшим чином. Результати моделювання дозволяють фахівцям громадської охорони здоров'я вживати ефективні науково обґрунтовані заходи щодо стримування епідемії. **Метою** статті є розробка моделей машинного навчання і компартментних моделей епідемічного процесу COVID-19, а також дослідження експериментальних результатів моделювання. **Об'єкт дослідження** – епідемічний процес COVID-19 і його динаміка на території України. **Предметом дослідження** є моделі і методи моделювання епідемічних процесів, в тому числі методи машинного навчання і компартментні моделі. Для досягнення мети дослідження ми використовували **методи** прогнозування машинного навчання і побудували модель лінійної регресії епідемічного процесу COVID-19 і модель SIR епідемічного процесу COVID-19. В результаті експериментів з розробленими моделями була отримана прогнозна динаміка епідемічного процесу COVID-19 на 30 днів для підтверджених випадків, тих, що видужали і летальних. Для випадків «Підтверджений», «Той, що видужав» і «летальний» середні помилки мають відхилення 1,15, 0,037 і 1,39 відсотка відповідно в результаті моделі лінійної регресії. Для випадків «Підтверджений», «Той, що видужав» і «Летальний» середні помилки мають відхилення 3,29, 1,08 і 0,71 відсотка відповідно для моделі SIR. **Висновки.** На даному етапі розвитку епідемічного процесу COVID-19 для прогнозування захворюваності доцільніше використовувати модель лінійної регресії, яка показала більш високу точність і ефективність. Як правило, причина цього полягає в тому, що використана модель лінійної регресії для цього дослідження була реалізована лише за 30 днів (за 15 днів до 2 березня), а не з використанням всього набору даних COVID-19. Крім того, очікується, що якщо ми спробуємо прогнозувати в більш тривалих часових діапазонах, модель лінійної регресії втратить точність. Як альтернатива, оскільки модель SIR включає більшу кількість факторів, очікується, що модель буде краще працювати при прогнозуванні на більш тривалі часові діапазони.

Ключові слова: модель епідемії; епідемічний процес; моделювання епідемії; імітаційне моделювання; лінійна регресія; модель SIR; COVID-19.

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МОДЕЛЕЙ ЛИНЕЙНОЙ РЕГРЕССИИ И SIR РАСПРОСТРАНЕНИЯ COVID-19 В УКРАИНЕ ДО ВАКЦИНАЦИИ

А. Мохаммади, Е. С. Меняйлов, К. А. Базилевич, С. В. Яковлев, Д. И. Чумаченко

Глобальная пандемия COVID-19 началась в декабре 2019 года и стремительно распространилась по всему миру. Во всем мире заболели более 230 миллионов людей, 4,75 миллиона случаев закончилось летальным исходом. Кроме угрозы здоровью следствием пандемии стали социальные проблемы, экономический кризис и переход привычной жизни в «новую реальность». Математическое моделирование является эффективным инструментом для контроля эпидемического процесса COVID-19 на заданных территориях. Моделирование позволяет спрогнозировать будущую динамику эпидемического процесса и выявить факторы, которые влияют на повышение заболеваемости наибольшим образом. Результаты моделирования позволяют специалистам общественного здравоохранения принимать эффективные научно обоснованные меры по сдерживанию эпидемии. **Целью** статьи является разработка моделей машинного обучения и компартментных моделей эпидемического процесса COVID-19, а также исследование экспериментальных результатов моделирования. **Объект исследования** – эпидемический процесс COVID-19 и его динамика на территории Украины. **Предметом исследования** являются модели и методы моделирования эпидемических процессов, в том числе методы машинного обучения и компартментные модели. Для достижения цели исследования мы использовали методы прогнозирования машинного обучения и построили модель линейной регрессии эпидемического процесса COVID-19 и модель SIR эпидемического процесса COVID-19. В **результате** экспериментов с разработанными моделями была получена прогнозная динамика эпидемического процесса COVID-19 на 30 дней для подтвержденных случаев, выздоровевших и летальных. Для случаев «Подтвержденный», «Выздоровевших» и «Летальный» средние ошибки имеют отклонение 1,15, 0,037 и 1,39 процента соответственно в результате модели линейной регрессии. Для случаев «Подтвержденный», «Выздоровевший» и «Летальный» средние ошибки имеют отклонение 3,29, 1,08 и 0,71 процента соответственно для модели SIR. **Выводы.** На данном этапе развития эпидемического процесса COVID-19 для прогнозирования заболеваемости целесообразнее использовать модель линейной регрессии, показавшую более высокую точность и эффективность. Как правило, причина этого заключается в том, что использованная модель линейной регрессии для этого исследования была реализована всего за 30 дней (за 15 дней до 2 марта), а не с использованием всего набора данных COVID-19. Кроме того, ожидается, что если мы попытаемся прогнозировать в более длительных временных диапазонах, модель линейной регрессии потеряет точность. В качестве альтернативы, поскольку модель SIR включает большее количество факторов, ожидается, что модель будет лучше работать при прогнозировании на более длительные временные диапазоны.

Ключевые слова: модель эпидемии; эпидемический процесс; моделирование эпидемии; имитационное моделирование; линейная регрессия; модель SIR; COVID-19.

Аліреза Мохаммаді – асп. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Меняйлов Євген Сергійович – канд. техн. наук, ст. викл. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Базілевич Ксенія Олексіївна – канд. техн. наук, доц. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Яковлев Сергій Всеволодович – д-р фіз.-мат. наук, проф., проф. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Чумаченко Дмитро Ігорович – канд. техн. наук, доц., доц. каф. математичного моделювання та штучного інтелекту, Національний аерокосмічний університет ім. М. Є. Жуковського «Харківський авіаційний інститут», Харків, Україна.

Alireza Mohammadi – PhD Student of Department of Mathematical Modelling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: alireza.mohammadi9207@gmail.com, ORCID: 0000-0002-4964-4494.

Ievgen Meniailov – PhD in Mathematical Modelling and Optimization Methods, Senior Lecturer of Department of Mathematical Modelling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: evgenii.menyailov@gmail.com, ORCID: 0000-0002-9440-8378.

Kseniia Bazilevych – PhD in Information Technologies, Associate Professor of Department of Mathematical Modelling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: ksenia.bazilevich@gmail.com, ORCID: 0000-0001-5332-9545.

Sergey Yakovlev – Dr. Sc. in Physics and Mathematics, Professor of Department of Mathematical Modelling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: svsyak7@gmail.com, ORCID: 0000-0003-1707- 843X.

Dmytro Chumachenko – PhD in Artificial Intelligence, Associate Professor of Department of Mathematical Modelling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine,
e-mail: dichumachenko@gmail.com, ORCID 0000- 0003-2623-3294.