Holt's Linear Model of COVID-19 Morbidity Forecasting in Ukraine

Darina Kapusta, Serhii Krivtsov, Dmytro Chumachenko

National Aerospace University "Kharkiv Aviation Institute", Chkalow str., 17, Kharkiv, Ukraine

Abstract

The use of methods for predicting epidemic processes and mathematical modeling of the dynamics of morbidity allows the development and implementation of scientifically based methods for the prevention and containment of the epidemic spread of an infectious disease. The study focuses on the development and implementation of the linear Holt model for predicting the incidence of COVID-19 in Ukraine. The advantage of the method is its high accuracy for short-term forecasting for 10 days. The disadvantage of this method is the impossibility of identifying factors that affect the behavior of the epidemic process.

Keywords 1

Epidemic process simulation, machine learning, Holt's linear model, COVID-19 simulation, epidemic model.

1. Introduction

In the 21st century, infectious diseases still remain a significant threat to humanity. From an evolutionary point of view, the emergence and persistence of pathogens of infectious diseases is inevitable, and any cellular form of life, including humans, is a target for parasitic microorganisms [1]. In response to bacterial or viral (bacteriophage) threats, eukaryotic and prokaryotic organisms have developed defense mechanisms to counteract pathogenic infectious agents, while the latter, in turn, reacted by developing counter-protection mechanisms to overcome host defense systems [2]. This constant evolutionary "arms race" continues today, and therefore, unpredictable outbreaks of infectious diseases remain a constant challenge for humanity [3]. The outbreak of coronavirus infection (COVID-19) is the largest recent challenge for health systems [4]. The rapid spread of the disease around the world is due to many factors, and globalization plays a key role in this process [5].

After the first reports in December 2019 of an outbreak of pneumonia in China, the new coronavirus quickly spread across all countries and continents [6]. By the beginning of May 2021, the number of registered cases of SARS-CoV-2 infection amounted to more than 163 million, almost 3.5 million people died due to coronavirus-related causes [7]. In Ukraine, according to official statistics, by the beginning of May 2021, more than 2 million SARS-CoV-2 infected were detected and 48 thousand deaths were recorded [8]. However, the reported number of infected persons is only a fraction of the actual number [9].

Ukraine became embroiled in the current COVID-19 pandemic later than other countries outside of China. The first confirmed case was reported on March 3, 2020 in a man who returned from Italy on February 26. The first death was recorded on March 13 in a woman who returned from Poland. By March 24, 2020, the number of cases exceeded 100, by April 3, 2020 – 1000 [10]. On April 20, 2020, 5710 cases and 151 deaths were recorded. The peculiarity of the outbreak in Ukraine was that after the announcement of the closure and blocking in many EU countries, primarily in Italy, a large number of citizens who were employed in Italy, Poland and other EU countries began to return to

MoMLeT+DS 2021: 3rd International Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Lviv-Shatsk, Ukraine

EMAIL kapusta.darina.yurievna@gmail.com (D. Kapusta); krivtsovpro@gmail.com (S. Krivtsov); dichumachenko@gmail.com (D. Chumachenko).

ORCID: 0000-0002-8168-8411 (D. Kapusta); 0000-0001- 5214-0927 (S. Krivtsov); 0000- 0003-2623-3294 (D. Chumachenko).

Ukraine [11]. Thus, a stream of potential sources of infection poured into Ukraine from countries where the epidemic was already raging.

On March 12, the first restrictive measures were introduced in the country. All educational institutions were closed, all public events were prohibited. On March 17, only 10 passengers were allowed in any public transport, and the metro was banned. Catering establishments and shopping centers were closed. On March 18, intercity train and bus traffic within the country was stopped. Since April 6, citizens are required to cover their faces with masks or respirators in public places; visits to parks and recreation areas have been prohibited. From April 7, 2020, it became possible to cross the Ukrainian border only by car at 19 checkpoints. However, the increase in the incidence of COVID-19 continued.

The process of detecting infected people in Ukraine continues to demonstrate not just, but a significantly growing trend, which shows the daily number of confirmed cases of COVID-19 from March 24th, 2020 – the day of detection of the first 100 cases according to the Ministry of Health of Ukraine [12].

The COVID-19 pandemic in Ukraine is characterized by territorial unevenness, which is associated with different population densities, differences in communications, religious traditions, uneven migration flows, regional characteristics of the fight against COVID-19, and the like. At the beginning of the pandemic, an increase in incidence was observed in the western regions.

The COVID-19 pandemic is an unpredictable threat to public health, the national economy and social development that can have uncontrollable consequences and cascading consequences for all aspects of society [13]. To successfully cope with the threat that has arisen, to reduce the impact on health systems and the economies of countries, to overcome the unforeseen consequences of public health measures taken to combat the COVID-19 pandemic, it is necessary to respond in a timely manner to the changing epidemic situation by introducing effective, scientifically based, adequate to the current situation. rational measures [14].

The use of an adequate epidemiological model makes it possible to restore data on the COVID-19 incidence and, on their basis, to carry out computational experiments on a computer for various variants of the development of the epidemic situation [15]. In this regard, it seems relevant to conduct predictive and analytical studies based on the developed mathematical model of the spread of COVID-19.

Mathematical modeling is an effective tool for predicting the dynamics of the development of the epidemic process [16]. With the help of modeling epidemic processes, it is possible to solve the problems of predicting the time series of morbidity [17], identifying factors affecting the dynamics of morbidity [18], calculating the required proportion of vaccination of the population [19], diagnostics [20], medical images analysis [21], etc.

The aim of the study is to develop a linear Holt model for predicting the incidence of COVID-19 in Ukraine, and to assess the possibility of its use.

2. Setting the epidemic process forecasting problem

The epidemic process is the only form of existence of infectious diseases. The living organism of a person or animal, that is, the natural habitat of the pathogen, where the pathogen lives, multiplies and is released in one way or another into the environment, was named the source of the causative agent of the infection (hereinafter the source of infection), and the way the pathogen changes from one host to another – the mechanism of transmission of the pathogen [22]. When the mechanism of transmission of the pathogen from the organism of the source of infection is implemented, it enters a susceptible organism, in which the pathogen multiplies, accumulates and begins to be released into the environment, turning into a source of infection. The epidemic process ensures the continuity of successive generations of the pathogen with a constant change of hosts and its preservation as a biological species. The consequence of this process is the epidemiological state of the population [23].

Thus, the introduction of the category “epidemic process” makes it possible to clearly understand the principles of the spread of diseases in society, provides a tool for understanding the concept of “elimination of infections” and outlines ways to combat infectious diseases.

Methods for predicting the spread of infectious diseases have been actively developing since the beginning of the 20th century and have proven to be a powerful tool for studying the patterns of epidemic processes and predicting their development in time and space. Recently, there has been an improvement in information systems for epidemiological surveillance, large volumes of statistical data available for analysis have appeared, new mathematical approaches have been developed for their processing, the power of electronic computers has increased, which has provided new opportunities for researching epidemiological data for predicting the infectious morbidity of the population.

The direct driving forces of the epidemic process are the source of infection, the transmission mechanism and the susceptible human body, which create a chain of successive infections with infections. Without these links, the existence of the epidemic process is impossible. The biological hazard of the population is determined by the possibility of the emergence and spread of biological agents that are pathogenic for humans. In this case, the process of spreading pathogens in space and time, the distribution of cases among different groups of the population and the dependence of these events on various phenomena and processes occurring in nature and society are important. Therefore, no less significant are the secondary driving forces of the epidemic process - social and natural factors that affect the intensity and manifestation of the epidemic process, slowing down or accelerating its development.

The main tasks of practical epidemiology are to assess the existing epidemic situation, identify the causal relationships due to which it has developed, and analyze risk factors, that is, factors whose action on the epidemic situation determines the likelihood of its complication. In order to rationally control the epidemic process, one should take into account the direction of evolution of the epidemic process and evaluate the most influential factors affecting the incidence of the population.

Let the values of the time series be available at discrete times $t = 1,2, ..., T$. Let the time series be $Z(t) = Z(1), Z(2), \dots, Z(T)$. At the moment of time T, it is necessary to determine the value of the process $Z(t)$ at the moments of time $T + 1, \dots, T + P$. The moment of time T is called the moment of forecast, and the value of P is the time of warning.

To calculate the values of the time series at future points in time, it is necessary to determine the functional dependence, reflecting the relationship between the past and future values of this series

$$Z(t) = F(Z(t-1), Z(t-2), Z(t-3), \dots) + \varepsilon_t. \quad (1)$$

This relationship is called the forecasting model. It is necessary to create a forecasting model for which the average absolute deviation of the true value from the predicted value tends to the minimum for a given P

$$\bar{E} = \frac{1}{p}\sum_{t=T+1}^{T+P} |\varepsilon_t| \to min. \quad (2)$$

In addition to obtaining future values of $\hat{Z}(T+1), \dots, \hat{Z}(T+P)$, it is necessary to determine the confidence interval of possible deviations of these values.

3. Holt's linear model

Forecasting models are divided into statistical models and structural models. In statistical models, the functional relationship between the future and actual values of the time series, as well as external factors, is set analytically. Statistical models include the following groups:

- regressive models;
- autoregressive models;
- exponential smoothing models.

Exponential smoothing is a method for smoothing time series, the computational procedure of which includes the processing of all previous observations, taking into account the aging of information as it moves away from the forecast period [24]. In other words, the "older" the observation, the less it should influence the value of the forecast estimate. The idea behind exponential smoothing is that as we "age", corresponding observations are given decreasing weights. This forecasting method is considered to be quite effective and reliable. The main advantages of the method are the ability to take into account the weights of the initial information, the simplicity of computational operations, and the flexibility of describing various dynamics of processes. The exponential smoothing method makes it possible to obtain an estimate of the trend parameters that characterize not the average level of the process, but the trend that has developed by the time of the last observation. The method has found the greatest application for the implementation of medium-term forecasts. For the exponential smoothing method, the main point is the choice of the smoothing parameter (smoothing constants) and initial conditions.

Identifying and analyzing the trend of a time series is often done by flattening or smoothing it [25]. Exponential smoothing is one of the simplest and most common techniques for aligning a series. Exponential smoothing can be represented as a filter, the input of which sequentially receives the members of the original series, and the output forms the current values of the exponential average.

Let $X = \{x_1, \dots, x_T\}$ is a time series. Exponential smoothing of the series is carried out according to the recurrent formula (3):

$$S_t = \alpha x_t + (1-\alpha)S_{t-1},\ \alpha \in (0,1) \tag{3}$$

The smaller α, the more filtered and suppressed the oscillations of the original row and noise. If we consistently use this recurrent relationship, then the exponential average S_t can be expressed through the values of the time series X.

$$\begin{aligned} S_t &= \alpha x_t + (1-\alpha)(\alpha x_{t-1} + (1-\alpha)S_{t-2}) = \cdots \\ &= \alpha \sum_{t=0}^{t-1} (1-\alpha)^2 x_{t-2} + (1-\alpha)^t S_0 \end{aligned} \tag{4}$$

If earlier data exists by the time smoothing starts, then the arithmetic average of all available data or some part of it can be used as the initial value S_0.

If earlier data exists by the time smoothing starts, then the arithmetic average of all available data or some part of it can be used as the initial value S_0.

When making forecasts using trend models, it must be borne in mind that the model is rigidly anchored, and all time series data equally affect the forecasts. If there is a need to give more weight to the new data, exponential smoothing can be used. However, we noted earlier that the exponential smoothing method does not give satisfactory results if the data is monotonically increasing or decreasing, that is, it contains a trend. In such cases, the trend-based exponential smoothing method (Holt's method, or two-parameter exponential smoothing method [26]) can be applied.

Simple exponential smoothing of time series containing a trend leads to a systematic error associated with the lag of the smoothed values from the actual levels of the time series. To take into account the trend in non-stationary series, a special two-parameter linear exponential smoothing is used. Unlike simple exponential smoothing with one smoothing constant (parameter), this procedure smooths both random disturbances and a trend using two different constants (parameters).

Holt's model consists of three equations.

The first is the data smoothing equation:

$$a_t = \alpha y_t + (1-\alpha)(a_{t-1} + b_{t-1}). \tag{5}$$

The second is the trend smoothing equation:

$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}. \quad (6)$$

And the last is the forecast equation for the period $t = k$:

$$y^*_{t+k} = a_t + b_t k. \quad (7)$$

where a_t is the smoothed value of the predicted indicator for the period t,

b_t is the estimate of the growth trend,

α is smoothing parameter ($0 \leq \alpha \leq 1$),

β is smoothing parameter ($0 \leq \beta \leq 0$),

k is the number of time periods for which the forecast is made.

The smoothing parameters α and β are selected subjectively, that is, by the forecaster based on the experience of previous forecasts, or by minimizing the forecast error. At larger values of the parameters, there will be a faster response to changes that occur, since the larger the parameter, the more smoothing the data will be. and vice versa, if the smoothing parameters are small, which tend to zero, then the model's response to changes in the data will be weaker, and the structure of the smoothed values will be less even.

4. Results

For the experiments, data on the COVID-19 new cases, provided by the Center for Public Health of the Ministry of Health of Ukraine, were used. Holt's model for predicting the incidence of COVID-19 was implemented as a software product in the Python programming language.

One of the most important stages in the creation of any information system is the design of a tool that can implement all the tasks set at the beginning of the project. A diagram of work execution, information exchange, workflow is graphically presented, visualizes a model of a business process.

To model business processes, the IDEF0 and DFD methodologies were used. Within the framework of the IDEF0 (Integration Definition for Function Modeling) methodology, a business process is represented as a set of function elements that interact with each other, and also show information, human and production resources consumed by each function. The IDEF0 methodology prescribes the construction of a hierarchical system of diagrams - single descriptions of system fragments. The functional model of the system is shown in Figure 1.

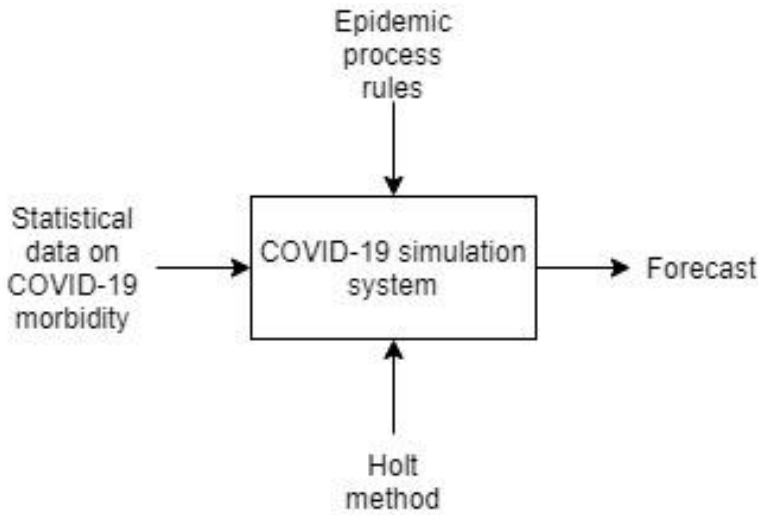

Figure 1: Functional model of system.

The model is based on the concepts of an external entity, process, data storage (storage) and data flow.

An external entity is a material object or individual acting as sources or receivers of information, for example, customers, personnel, suppliers, bank customers, and the like.

Process is converting input data streams to output in accordance with a certain algorithm. Each process in the system has its own number and is associated with the executor who performs this transformation. As in the case of functional diagrams, the physical transformation can be carried out by computers, manually or by special devices. At the upper levels of the hierarchy, when the processes have not yet been defined, instead of the concept of "process", the concepts of "system" and "subsystem" are used, which respectively denote the system as a whole or its functionally complete part.

A data warehouse is an abstract device for storing information. The type of device and methods of placement, removal and storage for such a device are not detailed. Physically, it can be a database, a file, a table in RAM, a card file on paper, and the like.

Data flow is the process of transferring some information from a source to a receiver. Physically, the process of transferring information can occur through cables under the control of a program or software system, or manually with the participation of devices or people outside the designed system.

Functional diagram was decomposed and decomposition is presented in figure 2.

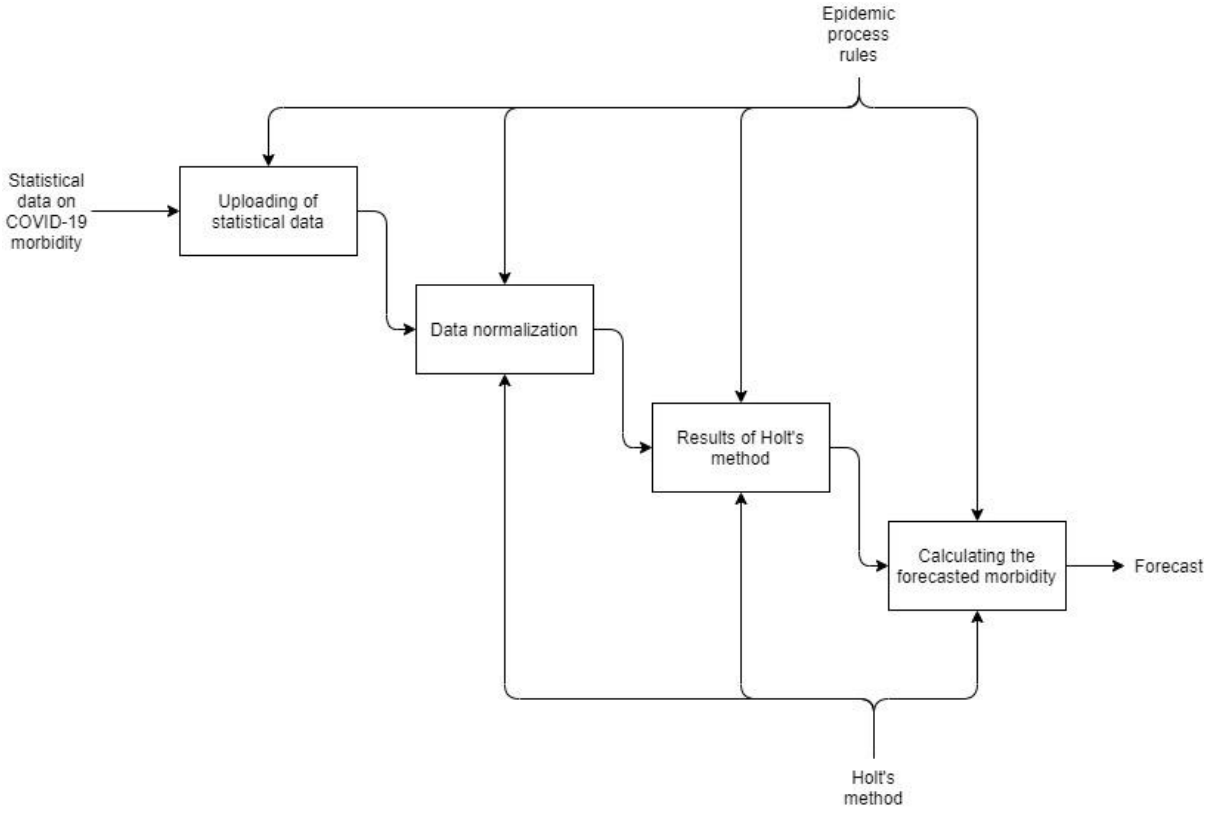

Figure 2: Decomposition of functional model.

The development plan is depicted in the IDEF3 diagram. The IDEF3 method is designed to develop models that describe any activity as an ordered sequence of events and objects that participate in this activity. The diagram is presented on Figure 3.

The application includes both data on COVID-19 morbidity in the world, which are automatically loaded from the John Hopkins University database, and more detailed data on Ukraine provided by the Center for Public Health of the Ministry of Health of Ukraine (fig. 4).

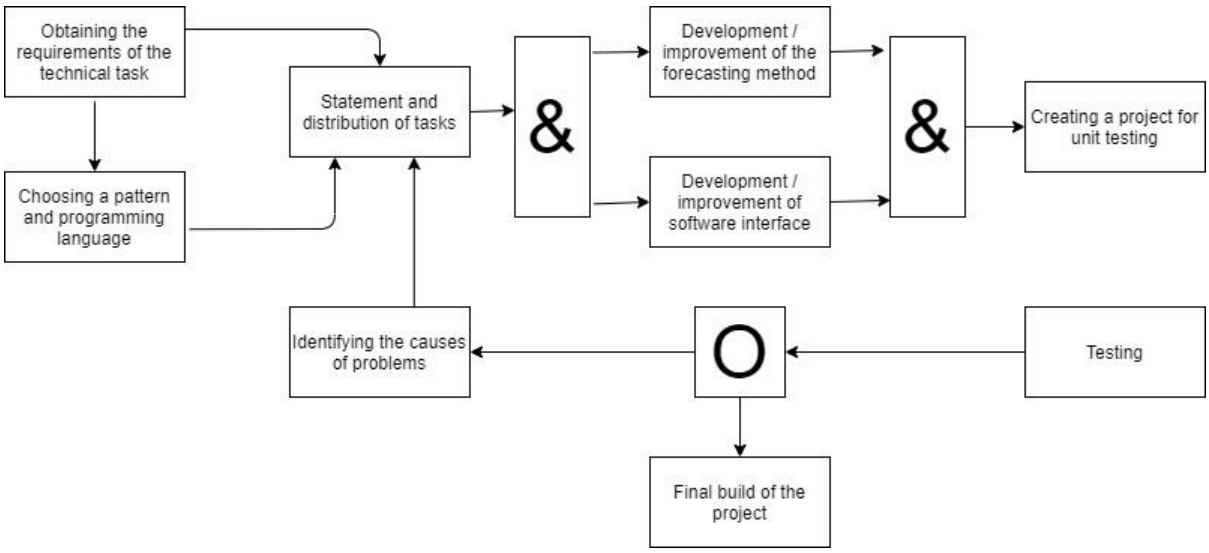

Figure 3: IDEF3 model of the application.

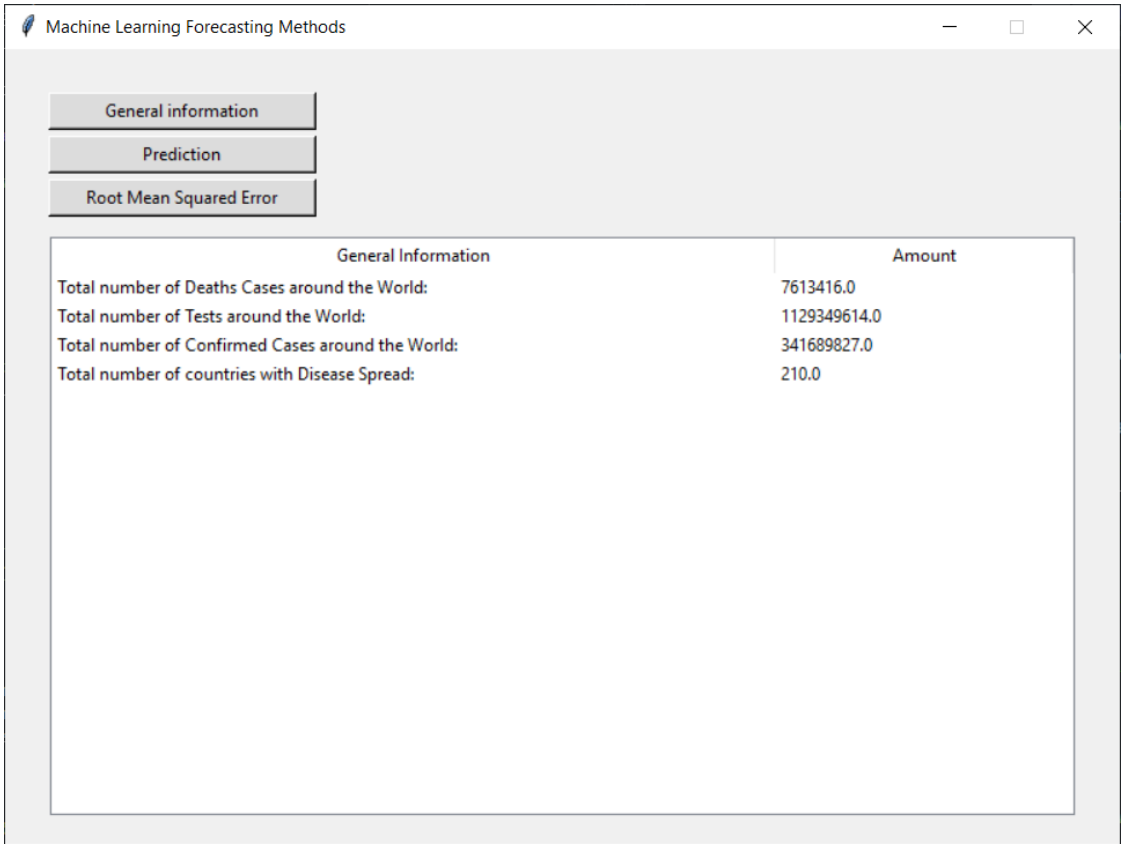

Figure 4: Interface of program product.

Results of COVID-19 morbidity in Ukraine using Holt's linear model is shown in Figure 5.

The results of the forecasting of incidence for 10, 20 and 30 days show that short-term forecasting for 10 days is the most accurate. Nevertheless, for use in practice, you can also use the results for 20 days, since such a forecast shows what will happen to the population after the 14-day incubation period.

Figure 5: Results of experimental study.

5. Conclusions

In this work, as a result of the analysis of various models and methods for predicting machine learning in order to predict the incidence rate and compare the results of calculations of each of the models, the Holt method was studied. As part of the study, a linear Holt model was developed, the quality of the models was shown in the form of the size of the forecast error, and a graph was built on which you can clearly see the calculation of the predicted incidence of COVID-19 for 10, 20 and 30 days.

On the basis of the developed model, an information system has been implemented that allows analyzing the dynamics of the incidence of COVID-19 in the regions of Ukraine.

The results of the forecasting of incidence for 10, 20 and 30 days show that short-term forecasting for 10 days is the most accurate. Nevertheless, for use in practice, you can also use the results for 20 days, since such a forecast shows what will happen to the population after the 14-day incubation period.

6. Acknowledgements

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” [27].

7. References

[1] D. Butov, et. al., Treatment effectiveness and outcome in patients with a relapse and newly diagnosed multidrug-resistant pulmonary tuberculosis, Medicinski Glasnik 17 (2) (2020) 356–362. doi: 10.17392/1179-20

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

[3] V. Arefiev, et. al. Complete genome sequence of salmonella enterica subsp. enterica serovar Kottbus strain Kharkiv, isolated from a commercial pork production facility in Ukraine, Microbiology Resource Announcements 9 (49) (2020) e01171-20. doi: 10.1128/MRA.01171-20

[4] I. Steffens, A hundred days into the coronavirus disease (COVID-19) pandemic, EuroSurveillance 25 (14) (2020) 2000550. doi: 10.2807/1560-7917.ES.2020.25.14.2000550

[5] J. V. Lazarus, et. al. A global survey of potential acceptance of a COVID-19 vaccine, National Medicine (2020) 1–4. doi: 10.1038/s41591-020-1124-9

[6] Q. Li, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia, New England Journal of Medicine 382 (13) (2020) 1199–1207. doi: 10.1056/NEJMoa2001316

[7] T. Khan, et. al., COVID-19: a worldwide, zoonotic, pandemic outbreak, Alternative therapies in health and medicine 26(S2) (2020) 56-64.

[8] A. Mohammadi, et. al., Compartment model of COVID-19 epidemic process in Ukraine, CEUR Workshop Proceedings 2824 (2021) 100-109.

[9] E.J. Favaloro, J. Thachil, Reporting of D-dimer data in COVID-19: some confusion and potential for misinformation, Clinical chemistry and laboratory medicine 58(8) (2020) 1191–1199. doi:10.1515/cclm-2020-0573

[10] R. Mazzucchelli, A. Agudo Dieguez, E.M. Dieguez Costa, N. Crespi Villarias, Democracy and Covid-19 mortality in Europe, Revista espanola de salud publica 94 (2020) e202006073.

[11] H. Li, Z. Liu, J. Ge, Scientific research progress of COVID-19/SARS-CoV-2 in the first five months, Journal of cellular and molecular medicine 24(12) (2020) 6558-6570. doi: 10.1111/jcmm.15364

[12] A. Meteliuk, et. al., Rapid transitional response to the COVID-19 pandemic by opioid agonist treatment programs in Ukraine, Journal of substance abuse treatment 121 (2021) 108164. doi:10.1016/j.jsat.2020.108164

[13] S.K Mishra, T. Tripathi, One year update on the COVID-19 pandemic: Where are we now?, Acta tropica 214 (2021) 105778. doi:10.1016/j.actatropica.2020.105778

[14] M. Iyer, et. al., COVID-19: an update on diagnostic and therapeutic approaches, BMB reports 53(4) (2020) 191–205. doi:10.5483/BMBRep.2020.53.4.080

[15] S. N. Gerasin, et. al., Set coverings and tolerance relations, Cybernetics and Systems Analysis 44 (3) (2008) 333-340.

[16] V. P. Mashtalir, S. V. Yakovlev, Point-set methods of clusterization of standard information, Cybernetics and Systems Analysis 37 (3) (2001) 295-307.

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

[18] V. Yesina, et. al., 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) 171–176. doi: 10.1109/infocommst.2018.8631897

[19] A.V. Bondarenko, et. al., Anaplasmosis: Experimental immunodeficient state model, Wiadomosci Lekarskie 72 (9-2) (2019) 1761-1764.

[20] T. Dudkina, I. Meniailov, K. Bazilevych, S. Krivtsov, A. Tkachenko, Classification and prediction of diabetes disease using decision tree method, CEUR Workshop Proceedings 2824 (2021) 163–172.

[21] I. Sova, I. Sidenko, Y. Kondratenko, Machine learning technology for neoplasm segmentation on brain MRI scans, CEUR Workshop Proceedings 2791 (2020) 50–59.
[22] V. Arefiev, et. al., Complete genome sequence of Salmonella Enterica Subsp. Enterica Serovar Kottbus Strain Kharkiv, isolated from a commercial pork production facility in Ukraine, Microbiology Resource Announcements, 9 (49) (2020) e01171-20. doi: 10.1128/MRA.01171-20
[23] T. Chaychenko, et. al., Nutritional practices as an obesogenic predictor in school-age children from Eastern and Western regions of Ukraine, Problemi Endokrinnoi Patologii 1 (2021) 75–83. doi: 10.21856/j-PEP.2021.1.10
[24] K. Ugryumova, I. Meniailov, I. Trofymova, M. Ugryumov, A. Myenyaylov, Synthesis of Robust Optimal Control Program for Axial Flow Compressor Turning Guide Vanes, International Journal of Computing 19(3) (2020) 347–354.
[25] N. Bakumenko, et. al. Synthesis Method of Robust Neural Network Models of Systems and Processes, Lecture Notes in Networks and Systems 188 (2021) 3–16.
[26] S. Maurya, S. Singh, Time Series Analysis of the Covid-19 Datasets, 2020 IEEE International Conference for Innovation in Technology (INOCON) (2020) 1-6. doi:10.1109/INOCON50539.2020.9298390.
[27] S. Yakovlev, et. al., The concept of developing a decision support system for the epidemic morbidity control, CEUR Workshop Proceedings 2753 (2020) 265–274.