FORECASTING COVID-19 CONFIRMED CASES BY COUNTRY

Author:
Stepanenko Julia,
11th-grade student
Lyceum №100 "Podil",
Kyiv,
Ukraine

Supervisor:
Efimova Tetyana Leonidivna,
Head of the Mathematical Modeling Section of the Kyiv Junior Academy of Sciences,
Ph.D. in Physical and Mathematical Sciences

2021

Abstract

Coronavirus is a disease that affects the lives of millions of people around the world. Predicting its development is very important because it allows you to plan the load on the health care system, the required number of medical staff, medical and laboratory equipment, the number of free beds in hospitals, and the introduction of restrictive measures such as a lockdown.

This work aims to study the methods of predicting the development of the epidemic and the application of the Holt-Winters method to make a forecast with the least deviation from the statistics.

The paper considers the classical SIR model for forecasting the development of epidemics and found that this model in its classical form is suitable only for long-term forecasting of the general direction of the epidemic, provided there are sufficient reliable statistics. After Holt-Winters. Successful choice of seasonality and coefficients of the method allowed to achieve a small deviation from the statistics. A web portal has been created to dynamically display forecasting results.

Forecasting provided for the USA, Ukraine, Russia, Poland and can be extended to other countries.

Keywords: coronavirus disease, SIR model for predicting the development of epidemics, time series analysis, Holt-Winters method.

TABLE OF CONTENTS

INTRODUCTION

The goal of this research is to predict the development of coronavirus disease in the USA, Ukraine, and other countries.

The purpose of this work is to study methods for predicting the development of the epidemic, the use of the Holt-Winters method to make a forecast with the least deviation from the statistics.

Achieving this goal involves the following **tasks**:

1) Consider the classical model for predicting the development of epidemics;
2) Consider the method of trend analysis of time series using exponential smoothing by Holt-Winters;
3) Build forecasts based on the above methods;
4) Create a web portal to dynamically display forecasting results.

Research methods:

1) Application of the classic SIR model.
2) Application of the method of time series analysis using exponential Holt-Winters smoothing.
3) Development of an algorithm for dynamic adjustment of the model taking into account the influence of external factors.

The scientific novelty of the work is as follows:

1) the author **independently** conducted a numerical experiment for different ratios of the parameters of the basic reproductive number (Ro) in the classical SIR model using statistical data of the country's officials;
2) the author **independently** selected the coefficients for the Holt-Winters method;
3) the author **independently** conducted a numerical experiment and determined the magnitude of the error for forecasting using the Holt-Winters method using statistics of the country's officials;
4) the author **independently** investigated the effectiveness of using the classical SIR model for epidemics for short-term forecasting and the Holt-Winters method of statistical analysis;
5) **for the first time,** a web portal was created with a visual display of the forecast of the number of patients and the magnitude of the error.

1.BACKGROUND RESEARCH ON FORECASTING EPIDEMICS

1.1. The idea of forecasting and the classical SIR model

Today the state of epidemic modeling can be described as actively evolving due to the spread of COVID-19. However, since such active development has only recently begun, most scientists are building on already known models and methodologies that have been tested using theoretical and statistical data on previous epidemics, such as the Hong Kong flu or the plague pandemic.

Consider the main approaches to predicting the development of epidemics on the example of the classical model of SIR. They provide an understanding of the complex dynamics of epidemics and their features. Such a model was first proposed by O. Kermak and Anderson Gray McCandrick [1].

In the simplest case, the population is divided into two groups: people who are potentially susceptible to the disease (denoted as S - from Susceptible), and already infected persons (denoted as I - from Infected). Conventional differential equations (which are deterministic) are used to study these models. Later, these models began to take into account the recovered persons (denoted as R - from Recovered).

This work used a standard model, without taking into account additional factors such as latent (incubation) period of the disease, mortality from the disease, immunity acquired through vaccination, life cycle (demography: birth/death), asymptomatic cases, and others. There are adaptations of the model, which in one way or another take into account these factors of influence, for example, SEIR, E - from Exposed, taking into account the incubation period, SIRD, D - from Deceased, etc. These models, although not significant, give slightly different results.

The model we are considering is described by the following system of differential equations:

$$\frac{dS}{dt} = -\frac{\beta IS}{N},$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - I\gamma,$$

$$\frac{dR}{dt} = \gamma I,$$

where

N - the total population of the study area;
S - the number of "vulnerable" on the day of the study;
I - the number of infected on the day of the study;
R - the number of those who recovered on the day of the study;
β - the number of contacts, that led to the infection;
γ - the period of "contagion" in days;
β/γ - the basic reproductive number (Ro).

It is possible to draw a rather obvious conclusion: the sum of infected, "suspected" and recovered is a constant and equal to the total population of the study group, because any person, regardless of whether he is ill now, can be assigned to one of the three categories described above.

Also, the dynamics of the spread of the disease very seriously depend on the rate of infection Ro. This is the ratio of the number of people that an infected person can "infect" during the period of his illness to the "period of contagion" of the same person, i.e. how many days a person is a carrier of the infection.

1.2. Variables in the classical model

As mentioned earlier, the classical model can be used in both an extended and a "narrower" sense. However, certain variables will be used regardless of which subspecies of the model are chosen.

S — Susceptible . These are people who "have a chance of getting infected." That is, they have not yet become ill, have not recovered, are not infected, but are only people who have not yet been in contact with the disease at all.

I — Infected. These are people who are currently carrying the disease, are infected and contagious.

R — Recovered. This is a group of people who have already suffered from the disease and are no longer carriers of it. Although some people may continue to struggle with the complications or side effects of the disease, they still fall into this category as soon as they are no longer infected with the disease under study.

In this model, the sum of all three components is a constant (N) and is equal to the population of the area to which this model applies

$N = S + I + R$

1.3. Parameters in the classical model

Parameters are constant values whose values can change under certain conditions and affect the overall result.

One of the parameters of the classical model of SIR is the basic reproductive number (Ro) - the average number of people directly infected with the patient during the entire infectious period of the disease, provided the patient enters a completely uninfected population. it is possible not only to draw conclusions about the "activity" of the spread of the disease but also to "adjust" the rate of its spread.

Ronald Ross, Alfred Lotki, and others were among the first scientists to study and introduce the concept of reproductive number [2]. However, it was first used by George MacDonald, who built models for the spread of malaria in 1952. [3]

The reproductive number consists of two more components and is actually their ratio:

β is the number of contacts or the number of people infected with the vector per unit time, and

γ - the period of "contagion" of the same person, the period when a person can infect others

If an individual in the infectious period of the disease makes β contacts per unit time, producing new infections with an average infectious period of $1 / \gamma$, then the baseline reproductive number is calculated as follows:

$$Ro = \beta / \gamma$$

Reproductive numbers are not altered by vaccinations or other factors that alter the population's susceptibility to the disease. However, it changes quite rapidly in the case of the introduction of such precautionary measures as social distance, "lockdown" and others.

Take, for example, Severe Acute Respiratory Syndrome (SARS) and measles. For the first disease, the reproductive number is 0.19–1.08, and for the second - 12–18. This suggests that SARS spreads more slowly, is less rapid and active than measles, which is extremely rapid in the spread.

From the above information we can draw the following conclusions:

- if the reproductive number is less than 1, the infection is already at the stage of extinction and after some predicted time will almost completely end its existence in the study area.
- if the reproductive number is more than 1, the infection will actively spread among the population and will continue to exist. The higher this number, the more difficult it is to control the prevalence of the disease.

Conclusions to the Section 1

After analyzing the theoretical information, we can conclude that the modeling of epidemics is well studied on the example of previous epidemics. However, the prediction of COVID-19 has not been studied so well. Therefore, this topic is new and interesting for research.

2.CLASSIC SIR MODEL FOR THE CORONAVIRUS DISEASE

2.1. Usage of the classical model for the coronavirus disease and analysis of the results

We form the Cauchy problem for the system of differential equations (1.1). To do this, take the data, for example, from November 7:

Using additional information: the population of the USA - 329916099 people (according to the census.gov), we can calculate the following initial values for the forecast for November 8:

S - 315672690,
I - 10266151,
R - 3851465.

Besides, we need such parameters as the basic reproductive number, the "period of contagion" and the number of "infected" people infected.

According to Wikipedia for the fall of 2020, it is estimated that the reproductive number can vary in the range from 1.5 to 5.7, the incubation period can be from 2 to 15 days, on average 5 days.

Based on the system of equations (1.1), we will perform calculations and construct possible graphs of the epidemic. Calculations for all initial data given above will be performed with different parameters. In Table 2.2, some fractions were not simplified to better understand the content of the parameters. The calculations were performed using the Euler method and the Runge-Kut method of the fourth-order. Calculations were performed both manually and using Microsoft Excel and Maple.

Table 2.1

Settings for modeling			Settings explanations: will get infected...	Result (forecast): number of infected **per day**				
β	Ro	γ		11/08	11/09	11/10	11/11	11/12
2/15	2	1/15	2 people in 15 days	**625836**	658220	691535	725708	760649
2/5	2	1/5	2 people in 5 days	**1877508**	2163000	2468621	2786178	3103278
1/10	1,5	1/15	1.5 person in 15 days	**298275**	303794	309209	314504	319662
3/10	1,5	1/5	1.5 person in 5 days	**894824**	942899	987611	1027620	1061474
Statistics by the country's officials				**109130**	124222	150486	149177	165871

Figures 2.1 and 2.2 show graphs of morbidity. The Y-axis shows the percentage of the population corresponding to the suspectible (blue line), infected (red line), and those who have recovered (green line) for better visualization.

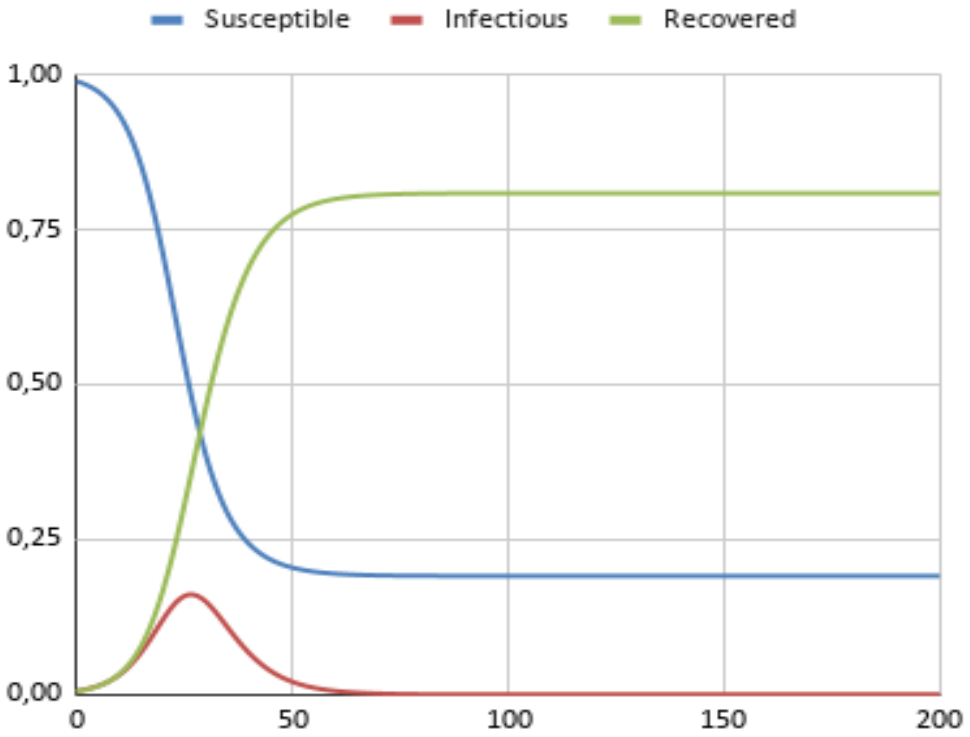


Рис 2.1 - Graph for the following values β=2/5, Ro=2.

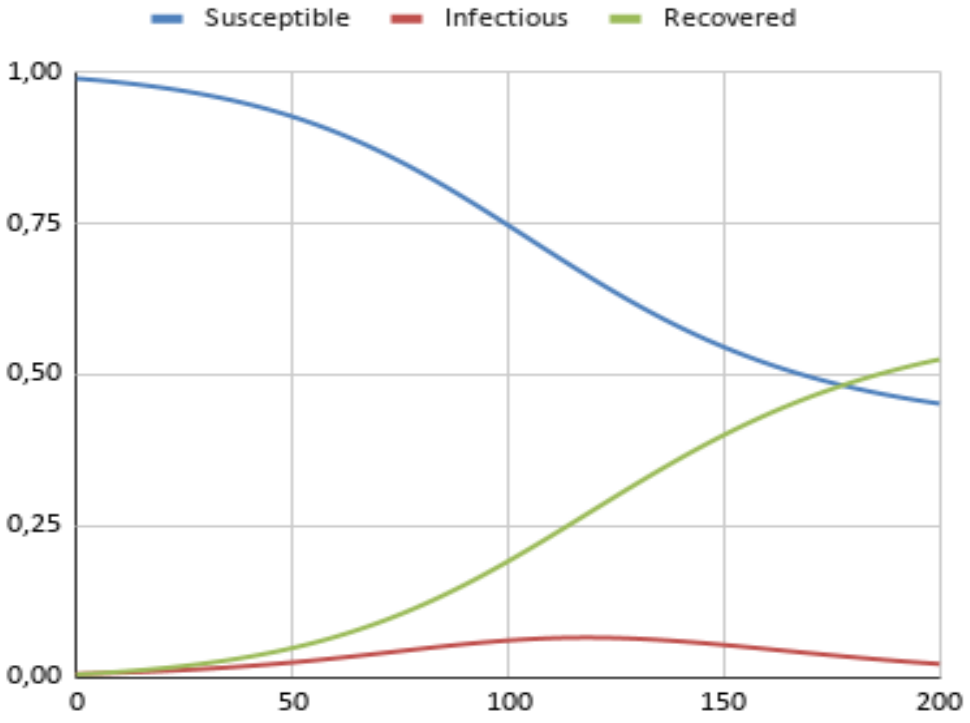


Рис. 2.2 Graph for the following values β=1/10, Ro=1.5.

The analysis of the results shows that the obtained data differ from the statistical one by 3 times and more. For some Ro, the difference will be even greater. The application of the SIR model leads to graphs that change smoothly, and real data change cyclically and in waves.

Conclusions to the Section 2

According to the data obtained, the following conclusion was made:

The basic SIR model in direct form cannot give the desired results. It can be used for long-term prediction and analysis of the general trend of the disease, however, without the use of additional modifications will not give a sufficiently accurate and reliable forecast for a short period - for several days or weeks.

Perhaps this model gives a forecast close to the real statistics, but since the real number of patients is unknown, taking into account asymptomatic, unregistered, and other cases, we are not able to verify the results.

3.FORECASTING WITH THE HOLT-WINTERS METHOD

3.1. Extrapolation and exponential smoothing

Extrapolation is a method of scientific research, which is based on the spread of past and present trends, patterns, relationships for the future development of the object of forecasting. The methods of extrapolation include the method of moving average, the method of exponential smoothing, the method of the least-squares.

Exponential smoothing is a method of mathematical transformations in predicting time series. Its essence is that for each repeated application of a mathematical operation with updated data, in order to approach the most accurate/correct result, all previous values of this series are taken into account, but the degree of consideration decreases exponentially.

It is believed that the exponential function in the 17th century proposed to use Poisson as a continuation of the method of numerical analysis, and later it began to be used for signal processing in the 1940s. Exponential smoothing was first proposed in the statistical literature by Robert Brown in 1956 [4], who used it to predict the amount of inventory to be stored in warehouses.

The method of simple exponential smoothing is the most effective for developing short-term forecasts. It can be used for forecasting only for one period. Its main advantages: simplicity of calculations and the ability to detect errors and deviations in the statistics used in forecasting.

The formula of the exponential smoothing method:

$$U_{t+1} = \alpha \times y_t + (1 - \alpha) \times U_t ,$$

where

t - the period preceding the forecast;

t + 1 - forecast period;

Ut + 1— the forecasted indicator;

α is the smoothing parameter α (0, 1);

Yt - the actual value of the studied indicator for the period preceding the forecast;

Ut is the exponentially weighted average for the period preceding the forecast.

When forecasting using this method, there are two problems:

selection of the value of the smoothing parameter α;

determining the initial value of Uo.

The value of α determines how quickly the weight of the influence of previous observations decreases. The greater α, the less the influence of previous data. If the value of α is close to unity, this leads to taking into account when forecasting mainly the impact of recent observations. If the value of α is close to zero, all (or almost all) past observations are taken into account during the prediction.

Thus, if it is certain that the initial conditions based on which the forecast is developed are reliable, a small value of the smoothing parameter ($\alpha \to 0$) should be used. When the smoothing parameter is small, the investigated function behaves like the average of numerous past levels. If there is not enough confidence in the initial forecasting conditions, then a large value of α should be used, which will take into account when forecasting mainly the impact of recent observations.

There is no exact method for selecting the optimal value of the smoothing parameter α. In some cases, Brown proposed to determine the value of α based on the length of the smoothing interval. In this case, α is calculated by the formula:

$$\alpha = \frac{2}{n+1},$$

where n is the number of observations included in the smoothing interval.

The problem of choosing Uo (exponentially weighted mean initial) is solved in the following ways:

if there is data on the development of the phenomenon in the past, you can use the arithmetic mean and equate it to Uo;

if such information is not available, then Uo uses the original first value of the forecast base. You can also use expert assessments.

3.2. Holt-Winters time series analysis

Improvements in Brown's method have been proposed by the American scientists Holt (1957) and Winters (1960). Holt proposed an upgraded dual model of exponential smoothing that took into account the trend in the time series. Winters proposed an even more advanced model of triple exponential smoothing, which in addition to the trend also takes into account seasonality. It consists of the following parts:

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1})$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1},$$

$$I_t = \beta \frac{y_t}{S_t} + (1-\beta)I_{t-L},$$

Final forecast:

$$F_{t+m} = (S_t + mb_t)I_{t-L+m}$$

where Yt is the value of the time series in the period t,
St - smoothed values in period t, level component,
β is a factor/component of the trend,
I - seasonal index/seasonality component,
F - forecast for m periods ahead,
t is the index denoting the calculation period,
L is the length of the season in periods
m is the desired forecast interval,
α, β and γ (0, 1) - smoothing coefficients, constants to be determined.

Seasonality is defined as the tendency of time series data to exhibit behavior that is repeated every period. The term "season" is used to denote the period before the behavior begins to repeat.

3.3. Usage of the Holt-Winters method for forecasting the number of infectious people per day and analysis of the results

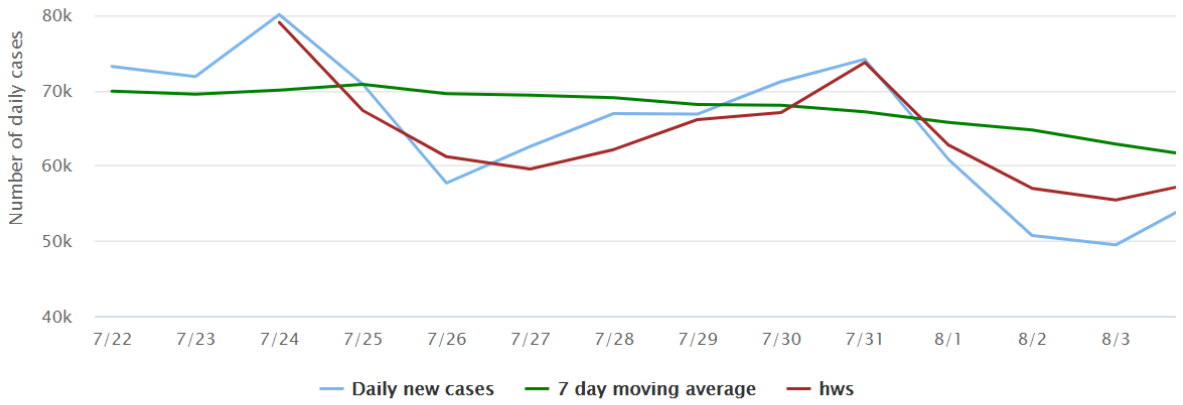
We use coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).

After analyzing the data, it was concluded that they have a seven-day period of fluctuations associated with working days and weekends of the week, which corresponds to the setting of the season length of 7 days.

For prediction, data for a maximum of the last 6 weeks are taken into account, which corresponds to the length of the incubation period (up to 2 weeks) and the average disease duration (up to 4 weeks).

The values of coefficients $\alpha = 0.67$, $\beta = 0.3$ and $\gamma = 0.3$ were chosen for the calculations.

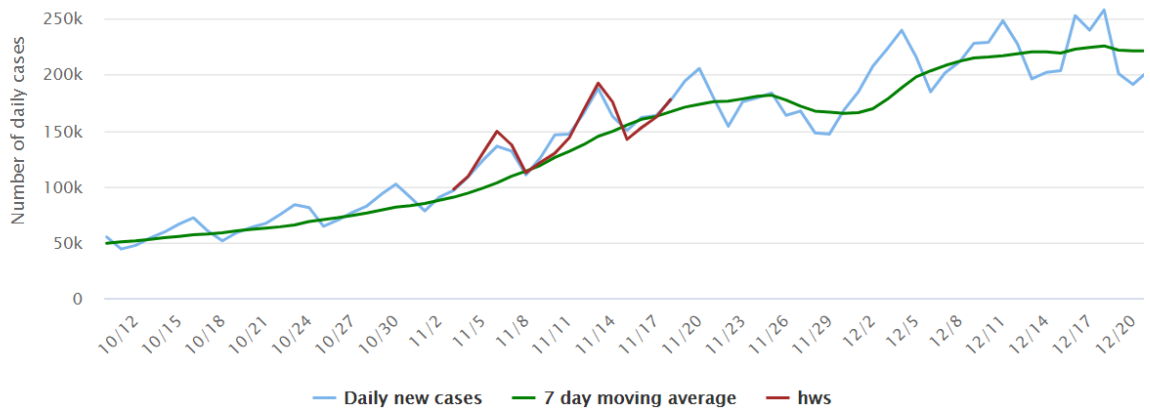
Since at the time of writing there is sufficient data on the incidence that has accumulated since the beginning of the epidemic in February 2020, it is possible to create a forecast for any period and check the effectiveness of the method (see section 4). To illustrate, below, Fig. 3.1, 3.2, 3.3, 3.4 show some forecasting results for different dates of 2020 and 2021.



Date	7/24	7/25	7/26	7/27	7/28	7/29	7/30	7/31	8/1	8/2	8/3	8/4	8/5	8/6	8/7	8/8	8/9
Forecast	79024	67314	61172	59540	62127	66115	67044	73712	62745	56979	55419	57784	61447	62262	68400	58177	52787
Real	80079	70807	57685	62548	66913	66824	71155	74123	60836	50706	49462	55441	56318	60503	64495	59007	49276
Error (%)	1.32	4.93	5.7	4.81	7.15	1.06	5.78	0.55	3.04	11.01	10.75	4.05	8.35	2.83	5.71	1.41	6.65

Forecast (days) 14 Back days (data use) 42 Auto ☑ Reset RMSE(N) 0.11589

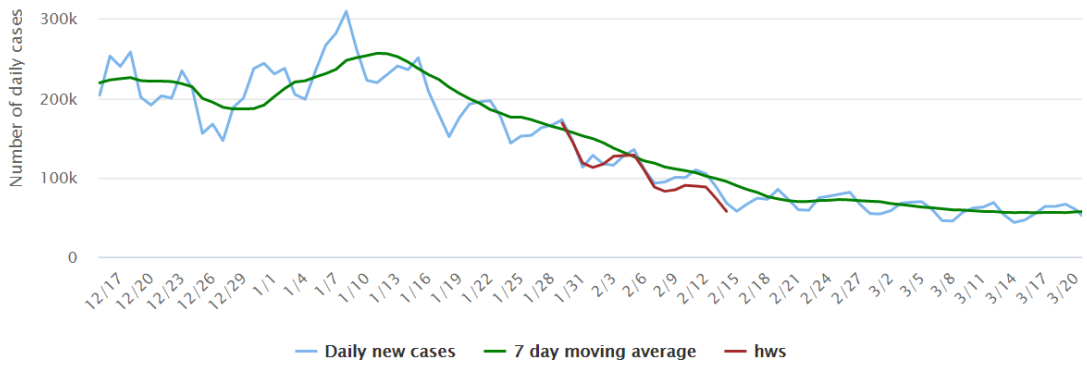
Fig. 3.1 - Forecast for July 24th, 2020 p. for the USA using Holt-Winters method.



Date	11/3	11/4	11/5	11/6	11/7	11/8	11/9	11/10	11/11	11/12	11/13	11/14	11/15	11/16	11/17	11/18
Forecast	97495	109081	129376	149280	137351	112161	121437	129582	143369	168296	192337	175400	142048	152610	161669	177657
Real	96311	108430	123022	135945	131603	110342	125560	146066	146785	165450	187563	162553	150095	161669	163357	176540
Error (%)	1.21	0.6	4.91	8.93	4.18	1.62	3.28	11.29	2.33	1.69	2.48	7.32	5.36	5.6	1.03	0.63

Forecast (days) 14 Back days (data use) 38 Auto ☑ Reset RMSE(N) 0.08235

Fig.3.2 - Forecast for November 3rd, 2020 for the USA using Holt-Winters method.



Date	1/29	1/30	1/31	2/1	2/2	2/3	2/4	2/5	2/6	2/7	2/8	2/9	2/10	2/11	2/12	2/13	2/14
Forecast	168485	145073	117939	112492	116724	126542	127461	128188	109149	87663	82515	84388	90042	89115	87892	73225	57387
Real	172525	147106	113037	127813	117037	115323	127210	135056	110227	92744	94311	100091	99931	109461	104714	88016	67791
Error (%)	2.34	1.38	4.16	11.99	0.27	8.87	0.2	5.09	0.98	5.48	12.51	15.69	9.9	18.59	16.06	16.8	15.35

Forecast (days) 14 Back days (data use) 35 Auto Reset RMSE(N) 0.10322

Fig. 3.3 - Forecast for January 29th, 2021 for the USA using Holt-Winters method.

Conclusions to the Section 3

From the analysis of the obtained results, we see that the error of the forecast for the period (normalized standard error/deviation), in some cases, does not exceed 0.15, and the error for a particular day may not exceed 1%.

The Holt-Winters triple exponential smoothing method, taking into account the seasonal component, shows fairly accurate statistical forecast results under a short-term forecast and a steady trend. Therefore, this method of forecasting can be called effective. Additional settings may be required for the period of a trend reversal.

4.WEBSITE FOR DYNAMIC FORECASTING

4.1. Development of the website

The work created a web portal that serves for:

- visual demonstration of the ability to predict the number of infected people over a period of time to a wide range of Internet users;
- analysis of forecasting quality, based on preliminary data at any time.

The web portal is implemented on the GitHub platform and is located at https://covid19-info.github.io/covid/

4.2. Capabilities of the website

- The portal contains information on the number of people infected with coronavirus in the USA and several other countries since the beginning of the epidemic. When choosing countries, it is possible to do:
- short-term forecast of the number of people who fell ill in the future, starting from the current date;
- forecast from any previous date, to compare the accuracy of the forecast with real data. The percentage error is displayed for each forecast date. The normalized root-mean-square error/deviation for the selected data analysis period is also displayed;
- settings of the model for the forecast, which allow you to choose the period based on which the forecast will be created, you can also choose the number of days for which you should make a forecast;
- enable automatic adjustment mode, which will display the most accurate forecast among those that will be created by the Holt-Winters method, based on different intervals of past data. Based on the analysis of deviations from real data in the past, the forecast with the smallest deviations will be selected, which ensures accuracy.

Conclusions to the Section 4

With the creation of a web portal to demonstrate the results of the study, it became possible to effectively and quickly find out the prognosis of the disease. The portal works on the link https://covid19-info.github.io/covid/ in real-time. Forecasting provided for the USA, Ukraine, Russia, and Poland.

CONCLUSIONS

Author **independently**:

1) an experiment was performed for different ratios of the parameters of the basic reproductive number (Ro) in the classical SIR model using statistical data from COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU);

2) the coefficients for the Holt-Winters method are selected;

3) an experiment was performed and the amount of error for forecasting was determined using the Holt-Winters method using statistical data;

4) the efficiency of using the classical SIR model for epidemics and the method of statistical analysis of Holt-Winters for short-term forecasting is investigated;

5) **For the first time,** a web portal was created with a visual display of the forecast of the number of patients and the magnitude of the error. The portal works at the link https://covid19-info.github.io/covid/ in real-time with up-to-date statistics from the COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). The **reliability** of the obtained results is ensured by comparing the forecasting results with the available statistical data from Johns Hopkins University;

The **theoretical significance** of the obtained results lies in the study of methods for predicting epidemics and choosing the optimal method.

The **practical significance** of the obtained results is that the obtained forecast allows forecasting the results of restrictive measures such as lockdown, planning future measures in quarantine conditions, choosing the optimal dates, and so on.

References

1. Kermack, W. O. and McKendrick, A. G. A Contribution to the Mathematical Theory of Epidemics // Proc. Roy. Soc. Lond. — 1927.
2. "Forecasting: Principles And Practice (2Nd Ed)". Otexts.Com, 2021, https://otexts.com/fpp2/. Accessed 15 Apr 2021.
3. G, MACDONALD. "The Analysis Of Equilibrium In Malaria". Tropical Diseases Bulletin, vol 49, no. 9, 1952, p. ., https://pubmed.ncbi.nlm.nih.gov/12995455/. Accessed 15 Apr 2021.
4. Brown, Robert G. (1956). Exponential Smoothing for Predicting Demand. Cambridge, Massachusetts: Arthur D. LittleInc. 15c.
5. Holt, Charles C. "Forecasting Seasonals And Trends By Exponentially Weighted Moving Averages". International Journal Of Forecasting, vol 20, no. 1, 2004, pp. 5-10. Elsevier BV, doi:10.1016/j.ijforecast.2003.09.015. Accessed 15 Apr 2021.
6. Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. Management Science, 6, C. 324–342.
7. "The SIR Model For Spread Of Disease | Mathematical Association Of America". Maa.Org, 2021, https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease. Accessed 15 Apr 2021.
8. NIST/SEMATECH e-Handbook of Statistical Methods, April, 2012 (DOI) URL:https://doi.org/10.18434/M32189 (дата звернення: 20.12.2020).
9. Забродская, Лана. "Прогнозирование На Основе Метода Экспоненциального Сглаживания". Ekonomika-St.Ru, 2021, http://www.ekonomika-st.ru/drugie/metodi/metodi-prognoz-1-4.html. Accessed 15 Apr 2021.
10. "Compartmental Models In Epidemiology - Wikipedia". En.Wikipedia.Org, 2021, https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology. Accessed 15 Apr 2021.
11. J. Stepanenko Forecasting COVID-19 confirmed cases in Ukraine // Proceeding of XIX International Scientific – Practical Conference «Shevchenkivska Vesna – 2021» April 2021, Kyiv, Ukraine