Seminar of I. Vekua Institute of Applied Mathematics REPORTS, Vol. 46, 2020

PROBLEM OF PREDICTING THE SPREAD OF CORONAVIRUS (COVID-19)

Gabelaia A.

Abstract. Prediction problem of coronavirus (Covid-19) is discussed. The possibilities for using models such as logistic, trendy and auto-regression and moving average are shown for this purpose. The correction capabilities of these models are shown and the accuracy of our prognostic estimates is analyzed.

Keywords and phrases: Coronavirus, prediction, prediction models, predictive validity.

AMS subject classification (2010): 91B02.

1. Recently, the spread of coronavirus COVID-19 has become a vital problem for the whole world. Suffice it to say that it was declared a pandemic by the World Health Organization on March 11 this year. Therefore, it is understandable what importance can be attached to the problem of predicting its prevalence. The results and experience of our research in this regard will be described below. For certain, it should be noted that in terms of our prediction, we considered the main indicators of the spread of coronavirus, such as the total number of cases of infection (in the world) for the current time (**total cases**), which is described below with the variable infic and the number of active cases (i.e. infected with this virus) for the current moment (**active cases**), which is denoted below by the variable ac. (Understandably, the difference between these two values is the total number of patients who have recovered and died (worldwide).)

Initially we tried to predict the values of the infic variable in the database for the period 22 / 01-13 / 02 2020 (in days), for the period 13 / 02-29 / 02. Therefore, it is clear that the data of February 13 were used by us in the so-called ex post forecasting or for selecting a forecasting model. It is interesting to note that based on the characteristics of these models (at the first stage) and the data of February 13, we considered the most reliable model of forecasting the Ferhulsts logistic growth model ([1]), whose general appearance looks like this

$$P(t) = \frac{KP_0}{(K - P_0)e^{-rt}) + P_0}, \tag{1}$$

where P_0 denotes the number of population (in this case infected) at the initial moment, and K is the maximum number of population (in this case infected). It should be noted that during this period the virus was mainly spread in China and the data were mainly consistent with its spread in China.

In this case, taking into account the data of February 13, we took the value of the K parameter equal to 85000 (which in this case turned out to be quite accurate, considering that this figure was 85214 as of September 15!) (I.e. within 7 months forecasting Error with real value, only 0.25% came out !!)

So the magnitude of the maximum rate of increase, r, in this case we were looking for from the following equation of regression (it should be noted that all our basic calculations were performed on the basis of the well-known computer program EViews-10 ([2])):

INFIC=(85000*580)/((85000-580)*EXP(-C(1)*@TREND)+580).

The value of the r parameter obtained as a result of this regression was r = 0.279355. However, the determination coefficient of this regression was in the order of 0.93, with a very high (equal to 59.2) t-statistic, only the Durbin-Watson statistic came out very small, indicating that the regression may be characterized by systematic errors. In addition, a diagram of its error gives some idea of the accuracy of this model (see Fig. 1)).

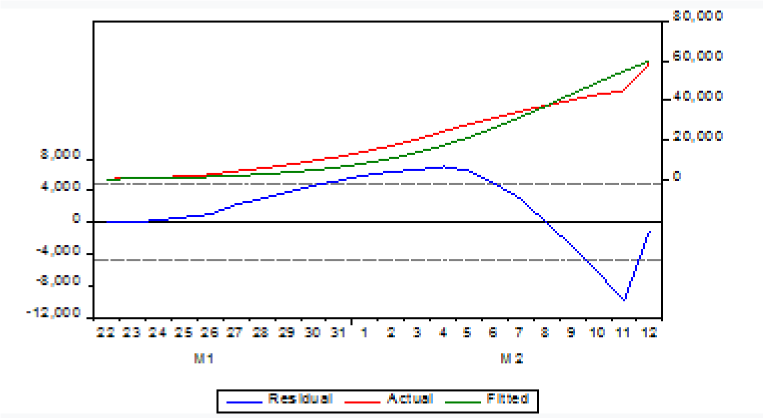


Fig. 1. Error diagram of the logistics model of infection

As for the predictive estimates of infic magnitude based on this model and their ratio to the actual values of this magnitude for a given period, is shown in Figs. 2, where infic and inficf1, respectively, indicate the actual magnitude of the number of infected and its (pre-) forecast.

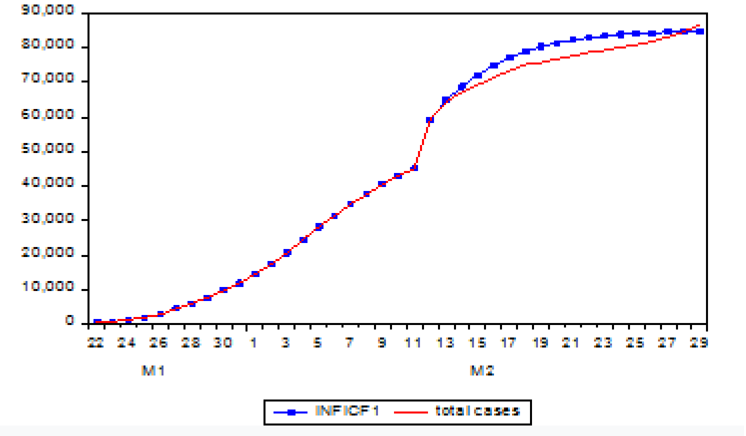
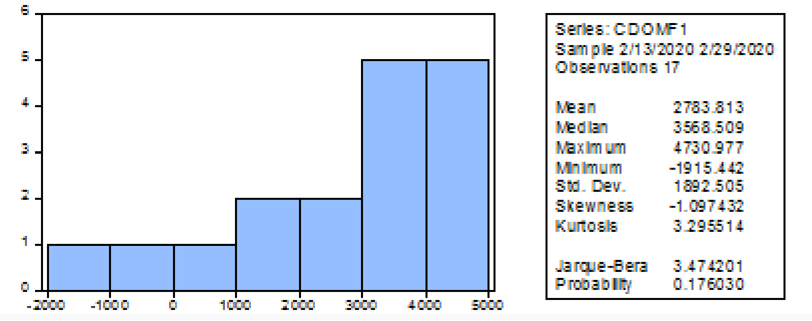


Fig. 2. Dynamics of prognostic and real indicators of infection for the period

13/02 – 29/02 2020

A more accurate representation of the error of the obtained predictive estimates is given by the cdomf1=inficf1-infic histogram of its error and the statistical characteristics shown in Figs. At 3 p.m.



Series: CDOMF1	
Sample 2/13/2020 2/29/2020	
Observations 17	
Mean	2783.813
Median	3568.509
Maximum	4730.977
Minimum	-1915.442
Std. Dev.	1892.505
Skewness	-1.097432
Kurtosis	3.295514
Jarque-Bera	3.474201
Probability	0.176030

Fig. 3. Histogram and statistical characteristics of error in predictive estimates of infection

As we can see, the forecasting results at this **(first) stage** are not so bad (the mean error and standard deviation for the forecasting period were 2784 and 1593 units, respectively).

We can be sure that the so-called mean error of the approximation (percentage of the modulus of error relative to the real values), which in this case is calculated by the formula

$$A = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{\inf icfI_i - \inf ic_i}{\inf ic_i} \right| \right) * 100\%$$

(where n denotes the length of the forecast horizon) amounted to 3.3%!

It should be noted that gradually, after the virus has crossed the borders of one country, the dynamics of its spread has significantly deviated from the form of the logistic curve used above (which, as mentioned above, allows for relatively long-term forecasting!) and demanded more complex, e.g. Use of ARIMA (Integrated Models of Autoregression and moving average (which is visually well visible on the actual virus distribution graphs). However, the disadvantage of these models is that they usually give short-term forecasts!

In particular, **in the second stage**, we tried to predict the total number of infected people for March, based on the data already in January-February (more precisely, 22 / 01-29 / 02 2020), based on the already mentioned ARIMA type models.

The trend model of the infic variable, with an autoregressive member, for this period took the form:

$$INFIC = 2419.82056382 * @TREND$$

$$+[AR(1) = 0.934518885427, UNCOND, ESTSMPL = "1/23/20202/29/2020"]. \qquad (2)$$

It should be noted that the coefficient of determination of this regression came out very high (in the order of 0.99), the t-statistics of the parameters were equal to 6.4, 10.8 and 5.8, respectively, only the Durbin-Watson statistics came out a little low (in the order of 1.26), which suggests that regression is not insured against systematic errors. Clearly, the diagram of its error also gives some idea of the accuracy of this model (see Fig. 4)).

As for the predictive estimates of infic magnitude based on this model and their ratio to the actual values of this magnitude based on the March 1-15 data shown in Figs. 5, where infic and inficf2, respectively, indicate the actual number of infected and its (pre-) forecast.

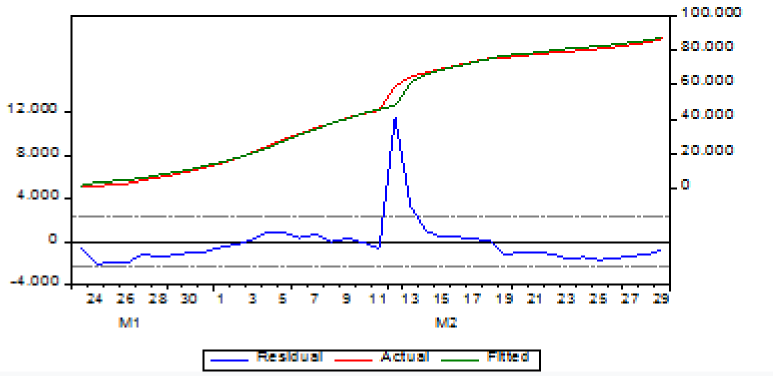


Fig. 4. Infection (2) model error diagram

As we can see, the accuracy of the prediction deteriorated rapidly from March 11, when the virus spread to virtually the entire world. Naturally, these dramatic changes, in the area and conditions of the spread of the virus, greatly complicated the problem of its prediction, especially since we here usually use passive prediction methods based on the assumption that the future should look like the past. This explains the circumstance that below we were forced to do the so-called optimistic and pessimistic forecast estimates that differed significantly from each other.

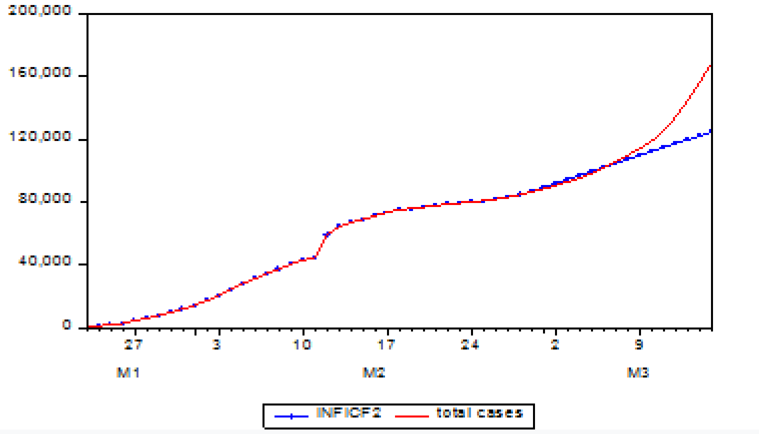


Fig. 5. Dynamics of prognostic and real indicators of infection
For the period 1 / 03-15 / 03 2020

Therefore, we made the decision to adjust our forecast estimates from March 11 this year. As for the accuracy of the forecast estimates we received for the first decade of March, the average error of its approximation was 2% !

It should be noted that the average error in the approximation of the predicted estimates of the infic variable obtained from similar models above was 2.8% for the second decade of March!

Based on all the above (already in the fourth stage!), we made forecasts for the third decade of March, for which the average approximation error was 1.75%!

In addition, as the relevant calculations show, for the period 13 February-31 March, the weighted average error of the approximation of our forecasts was 2.57%!

In the fifth phase, the average error in the approximation of the prognostic assessment of infection for the first half of April was 2.14%.

In the sixth phase, i.e. in the second half of April, the average error in the approximation of forecast estimates was 1.7%.

In the seventh stage, i.e. in the first half of May, the average error in the approximation of our forecast estimates was 0.54%.

In the eighth phase, for the period May 16-June 15, the average error in the approximation of our forecast estimates was 0.9%.

In the ninth phase, i.e. for the second half of June, the average error in the approximation of our forecast estimates was 2.73%.

In the tenth phase, the average error in the approximation of our forecast estimates for the first half of July was 2.23%.

At the eleventh stage, i.e. in the second half of July, the average error in the approximation of our forecast estimates was 1.2%.

In the twelfth stage, the average error in the approximation of the forecast estimates was 2.15%.

Finally, the dynamics of the main characteristics of the spread of the virus to the world (in terms of days) between January and August 2020 looked like this (see Figure 6-9):

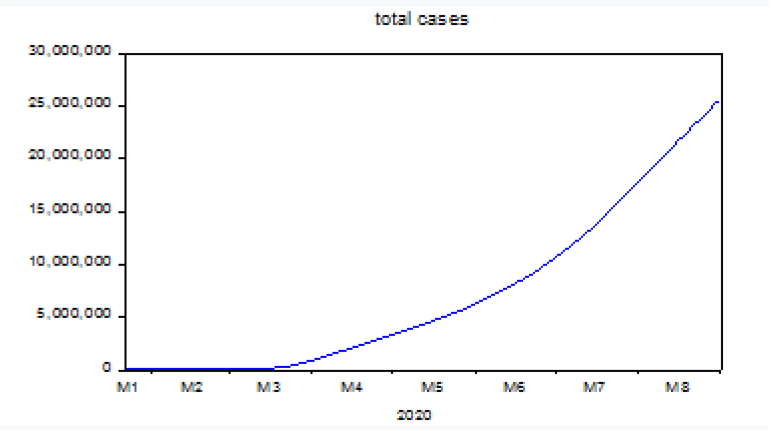


Fig. 6. Dynamics of the number of infected (infic) in the period January-August 2020 (in terms of days)

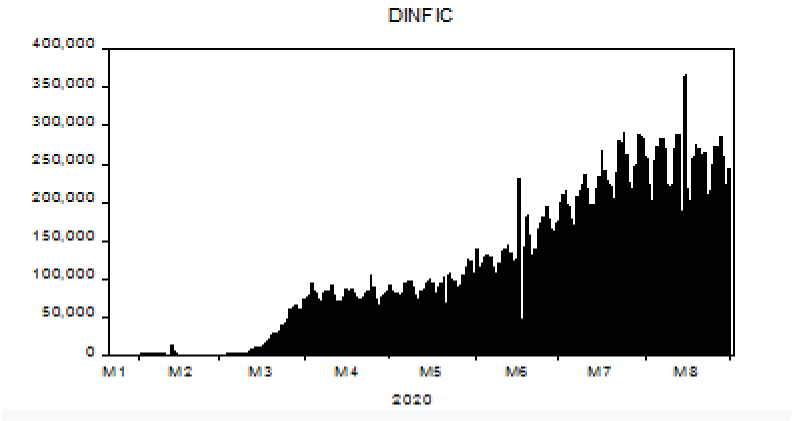


Fig. 7. Dynamics of the growth rate of the number of infected people (d(infic)) in the period January-August 2020 (in terms of days)

As we can see, the forecasting models we use (which are known to work well in the short run!) Show really high enough accuracy for a maximum of a month (then their accuracy drops!). On the other hand, the virus is “not going to stop” in the near future, which calls into question the prospects of using the above prediction methods. Therefore (in order to increase the forecasting horizon!), It may make sense to consider a new indicator such as e.g. “Average daily increase in the number of infected people per month”, which will allow us to make a forecast of this figure for a horizon of several months, which we are going to do in the future!

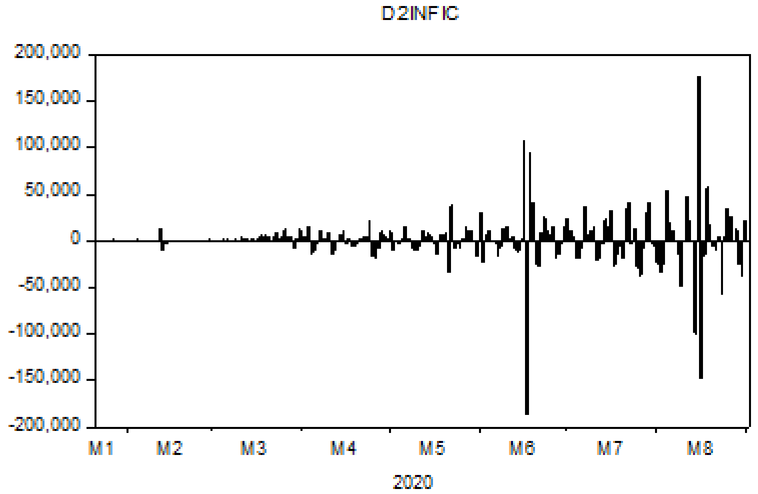


Fig. 8. Dynamics of “acceleration” of the growth of the number of infected people (d(infic, 2)) in the period January-August 2020 (in terms of days)

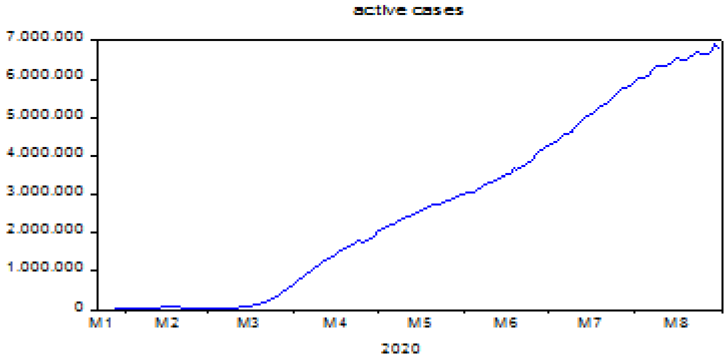


Fig. 9. Dynamics of the number of patients with the virus (ac) in the period January-August 2020 (in terms of days)

Therefore (in order to increase the forecasting horizon!), it may make sense to consider a new indicator such as e.g. "Average daily increase in the number of infected people per month", which will allow us to make a forecast of this figure for a horizon of several months, which we are going to do in the future! Moreover, according to the central limit theorem, the distribution of this indicator should be close to normal, which will simplify the task of finding reliable predictive estimates for it.

In particular, the dynamics of this indicator for the world in January-October 2020 looked like this:

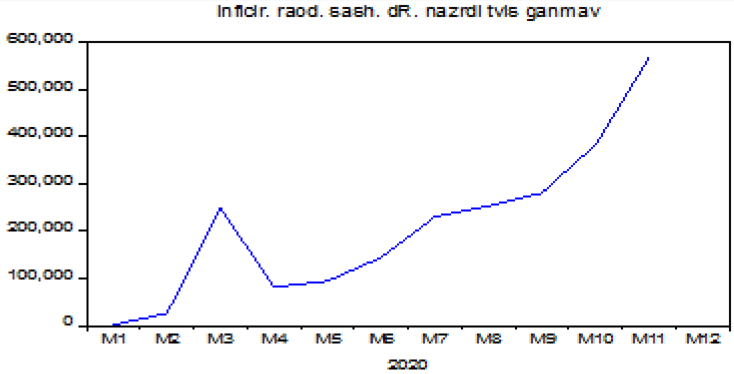


Fig. 10. Average daily increase in the number of infected dynamics in the world with January-October 2020 data

As we can see, in the given period this indicator (if we do not count the local maximum in March) was developing in an ascending line, i.e. the spread of the virus in the world had a permanently upward dynamics, although this growth was not as avalanche-like in Georgia as it has been in the last two months.

It should be noted that as for the mortality rate caused by this virus (percentage of deaths to the total number of survivors and deaths), it initially increased rapidly enough to 21%, but then slowly dropped to 3% (as of the end of October!). However, we must bear in mind that above we talked about the cumulative (accumulated) value of this indicator! As for its current importance (since “critical” make up only 1% of the current number of patients), we must assume that the current mortality rate for this virus is even lower. However, unfortunately, it is characterized by a very high speed of propagation.

2. Consider now the problem of predicting the spread of coronavirus for Georgia. More specifically, note the number of coronavirus infections in Georgia with tcge and discuss the problem of its prognosis. Given that this virus was first detected in Georgia on February 26, we initially considered the problem of its prediction on the basis of the February 26 - April 6 database.

The ARMA type model of the given indicator, with the addition of the trend member, looked like this:

$$TCGE = 4.40633779642 * @TREND$$

$$+[AR(1) = 0.9663677571, MA(1) = 0.528597598922, UNCOND, ESTSMPL$$

$$= "2/26/2020\ 4/06/2020"].$$

The coefficient of determination of this regression was very high (in the order of 0.99), the t-statistics of the parameters were quite high, and the Durbin-Watson statistic was in the order of 2.03, which indicates a very high accuracy of the model. Clearly, a diagram of its error also gives some idea of the accuracy of the model (see Fig. 11).

The dynamics of the forecast estimates of the tcge variable derived from this model for the period 7-30 April are shown in Figs. 12 in the form of the tcgef1 graph, from which it appears that these predicted estimates can be considered as optimistic estimates.

As for the pessimistic estimates of this magnitude (which are represented by the tcgef2 graph in Fig. 12), they were found from the autoregressive model of LOG (TCGE), with the addition of a trend member, which in this case looked like this:

$$LOG(TCGE) = 0.142281868034 * @TREND$$

$$+[AR(1) = 0.914201964397, UNCOND, ESTSMPL = "2/26/20204/06/2020"].$$

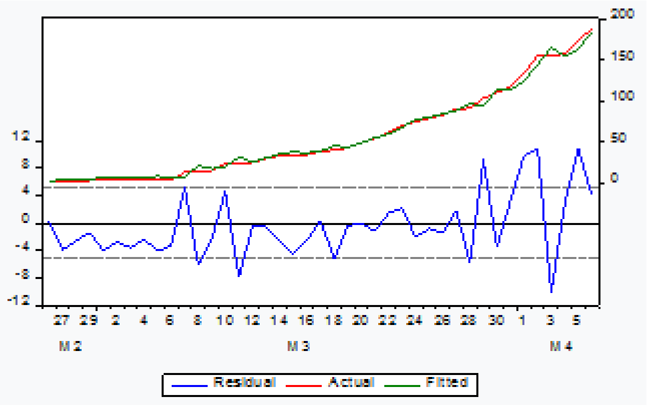


Fig. 11. Error diagram of the model (3) of infected with the virus in Georgia

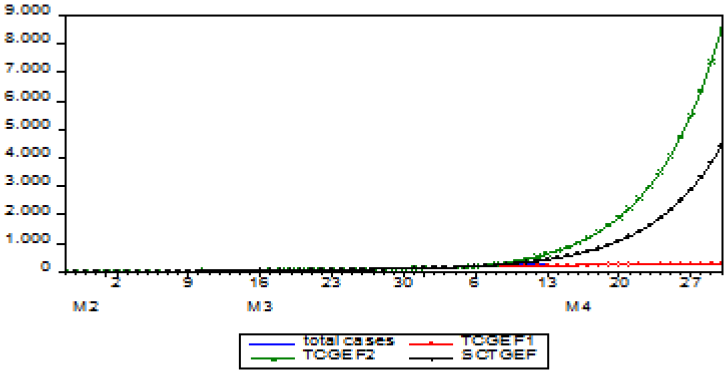


Fig. 12. Prognostic estimates of the number of infected (tcge) in Georgia for the period April 7-30

However, the coefficient of determination of this regression was quite high (in the order of 0.98), the t-statistics of the parameters were quite high, and the Durbin-Watson statistic was in the order of 2.16, which indicates a very high accuracy of the model.

It should be noted that Figs. 12 indicates the average forecast graph obtained by $stcgef = 0.5 * (tcgef1 + tcgef2)$ from $stcgef$.

As for the accuracy of the forecast estimates obtained above, our optimistic estimates (tcgef1) came out closest to reality in the given period! However, the average error in the approximation of these forecast estimates was 14.2%.

Subsequently, in line with the above, we carried out the work of predicting the spread of coronavirus in Georgia in seven stages. In particular, in the seventh stage, for August, the average error in the approximation of our forecast estimates was 2.96%, which should be considered a good enough result, given the instability of the dynamics of the spread of the virus during this period!

At the same time, the dynamics of virus spread in Georgia (in terms of days) in February-August 2020 looked like this (see Figure 13):

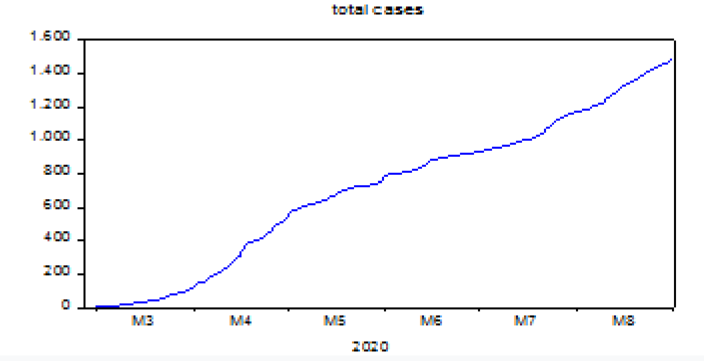


Fig. 13. TCGE dynamics of the number of infected people in Georgia in February-August
2020 (in terms of days)

Finally, as in the case of the world (in order to increase the forecast horizon!) it may have a sense to consider a new indicator such as e.g. "Average daily increase in the number of infected people per month", which will allow us to make a forecast of this indicator for a horizon containing several months. In particular, the dynamics of this indicator in Georgia in February-November is shown in Figs. 14.

As it is known, this indicator developed in an ascending line in February-April, then until June we had a slightly descending line, which was followed by an avalanche growth trend in the last three months. As a result, the average daily growth rate in October reached 1056 (while in August this figure was equal to 10), and in November it was already 3222.

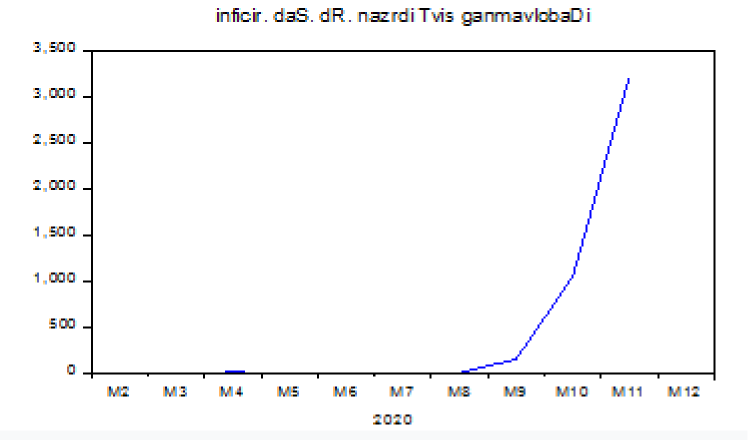


Fig. 14. Average daily increase in the number of infected Dynamics in Georgia in February-November 2020

Which, obviously, is an extremely alarming figure (caused by the overabundance of many negative factors!).

REFERENCES

1. Verhulst P. F. Recherches Mathematiques sur La Loi D'Accroissement de la Population, *Nouveaux Memoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, 18, Art. 1, 1-45, 1845 (Mathematical Researches into the Law of Population Growth Increase).

2. Gabelaia A., Gabelaia L. Econometrical analyzes and Forecasting package EViews fundamentals (Georgian). *Tbilisi*, 2017.

Received 17.09.2020; revised 30.09.2020; accepted 05.12.2020

Author's address:

A. Gabelaia
Georgian Technical University
77, M. Kostava St. Tbilisi 0175
Georgia
E-mail: agabelaia@gtu.ge; agabelaia@gmail.ru