**Seminar of I. Vekua Institute
of Applied Mathematics
REPORTS, Vol. 47, 2021**

THE PROBLEM OF MODIFYING MODELS FOR PREDICTING THE SPREAD OF CORONAVIRUS (COVID-19)

Gabelaia A.

Abstract. The possibility of modifying the developed models of coronavirus (Covid-19) prediction, which aims to increase the prediction horizon, is discussed. Therefore, it may make sense to consider a new indicator such as “average daily increase in the number of infected during the month” and “total number of infected by the end of the period (in this case, the month)”. The presented paper is dedicated to the study of these possibilities.

Keywords and phrases: Coronavirus, prediction, prediction models, prediction horizon.

AMS subject classification (2010): 91B02.

The problem of predicting the spread of coronavirus (COVID-19) in the world, we have been discussing since February 13, 2020 [1]. However, given that the virus was mainly spread only in China during this period, we used a logistic function for forecasting [2]. This gave us a pretty good result, considering that the number of infected people in China as of September 15 was 85214! (i.e. within 7 months of forecasting, the forecast error with respect to the real value came out to be only 0.25%!).

However, later, when the virus spread all over the world and the dynamics of its spread became very complicated, we had to use shorter-term forecasting models, namely ARIMA (Integrated Models of Autoregression and moving average) models [3] (with the addition of trending components) and periodically correct our forecast estimations.

For the sake of clarity, it should be noted that in the past, we have considered in terms of prognosis such key indicators of the spread of the coronavirus as the number of total cases of infection and the number of active cases at the present time (to date).

However, as practice has shown, the models we used showed high enough accuracy over a maximum of one month (then their accuracy dropped). On the other hand, given that the virus is “not going to stop” in the near future, the problem of increasing the forecast horizon is on the agenda.

Therefore, it may make sense to consider a new indicator such as “average daily increase in the number of infected people per month”. This makes it possible to predict this indicator for a horizon containing several months, especially since according to the central limit theorem of probability theory, the distribution of this indicator should be close to normal, which somewhat simplifies the task of making reliable forecast estimates for it.

In addition, it is possible to increase the forecast horizon on the basis of an indicator such as the total number of infected by the end of the period (in this case, the month).

Clearly, however, this in itself implies that the accuracy of such predictions should increase with the accumulation of relevant (over a period of months) information. The presented paper is dedicated to the study of these possibilities.

In the interest of increasing the forecast horizon, we initially addressed the problem of forecasting a rate such as the “average daily increase in the number of infected people per month” (SDTC).

We tried to forecast this figure for the first quarter of 2021 (although in this case the statistics were very small and therefore great accuracy of forecasting was not expected!).

As for the quantitative characteristics of these forecasts, the average error in the approximation of our optimistic forecasts for the first quarter of 2021 was 35.6%, with this error being only 5% in January (again emphasizing the short-term effectiveness of the models used). (It is interesting to note that the average relative error was within 31% when using more accurate forecast errors (compared to the actual figure for the previous period!).

In addition, we tried to forecast this figure for the second quarter of 2021 (obviously based on the 2020 and 2021 first quarter databases).

In particular, the trend model of this indicator, with the addition of a moving average, looked like this:

SDTC = 42753.8683336*@TREND
+ [MA(3)=-1,UNCOND,ESTSMPL="2020M01 2021M03"].

However, the coefficient of determination of this regression came out very high (in the order of 0.9), the t-statistic of the trend parameter was quite high, only the Durbin-Watson statistic came out low (in the order of 1.5), indicating systematic error possibilities in the model.

As for the second estimate of this indicator (which was initially optimistic and then became pessimistic!), its equation looked like this:

LOG(SDTC) = -0.0227969509625*(@TREND)∧2
+ 0.564327466074*@TREND + 9.7369576181
+ [AR(2)=-0.757862290481,MA(2)
=-0.662599126347,UNCOND,ESTSMPL="2020M01 2021M03"].

However, the coefficient of determination of this regression was 0.93, the t-statistics of the parameter were quite high, and the Durbin-Watson statistic was 1.76 (indicating the possibility of some systematic errors in the model).

As for the quantitative characteristic of the accuracy of these forecast estimates, the average error of the approximation of the average forecast estimates obtained by us for the second quarter of 2021 was 17%, which was twice better than the same indicator of the previous quarter!

In addition, given that the dynamics of the average daily increase in the number of infected people turns out to be quite complicated, we switched to forecasting such an indicator as the total number of infected people at the end of the period (in this case, the month) denoted by tcp. This, in our opinion, is justified even because the basic dynamics of this indicator is much simpler (due to its monotonicity (non-deficiency!)), than the dynamics of the average daily increase of the total number of infected (in terms of months).

Since we started to find the forecast estimates of this magnitude in the first days of May of the current year, we made its forecast for the second quarter of 2021 (January 2020-March 2021 database), i.e. we made an ex post forecast for April (based on the relevant model), to be able to compare with obtained our forecast estimations.

In particular, the trend equation for this indicator (TCP), with the addition of the corresponding member of the moving average, was as follows:

TCP = -1703180.96828*@TREND + 807507.387172*(@TREND)∧2
+ [MA(1)=0.919288114933,UNCOND,ESTSMPL="2020M01 2021M03"].

However, the determination coefficient of this regression came out very high (in the order of 0.998). The t-statistics of the parameters were also quite high, only the Durbin-Watson statistics came out low (of the order of 1.6).

As for the optimistic forecast estimates of this indicator, the corresponding autoregressive equation with the trend component had the form:

$$\begin{gathered}\text{TCP} = 1.36442167005{*}\text{TCP}(\text{-}1) \text{ - } 0.480876472847{*}\text{TCP}(\text{-}2)\\ + 1830732.9788{*}@\text{TREND} \text{ - } 4339333.23072.\end{gathered}$$

The coefficient of determination of this regression was very high (in the order of 0.997), the t-statistics of the basic parameters were also satisfactory, while the Durbin-Watson statistic was in the order of 1.96, which indicates the high accuracy of the model.

As for the quantitative characteristic of the accuracy of these forecast estimates, the average error in the approximation of the average forecast estimates obtained by us for the second quarter of 2021 was 2.1%, which is clearly a much better indicator than the above for the average daily increase in the number of infected.

Note. As we can see, it is relatively easy to predict the total number of infected by the end of the period (month), even compared to predicting the average daily increase in the number of infected. However, on the other hand, this figure is less visible in terms of pandemic prevalence analysis. For this purpose, we can use the mean daily increase in the number of infected people, or (more easily calculated!) the increase in the total number of infected people at the end of the period (growth “speed”) d (tcp), which can be easily calculated for the forecasted estimates (e.g. For the average forecast estimates). In addition, for a more complete analysis of the dynamic characteristics of the process (as we did before!) We can consider the rate of “acceleration” of the total number of infected people in the period d (tcp, 2), which is calculated (using a computer software package like Eviews). very easy, including for predictive estimates.

Finally, we found TCP forecast estimates for the second half of this year.

The autoregressive equation of this index with the trend component, had the form:

$$\begin{gathered}\text{TCP} = 1.19821695705{*}\text{TCP}(\text{-}1) \text{ - } 0.333747850998{*}\text{TCP}(\text{-}2)\\ + 2304432.80237{*}@\text{TREND} \text{ - } 5845770.21444.\end{gathered}$$

The coefficient of determination of this regression came out very high in the order of 0.998, the t-statistics of the basic parameters were also satisfactory, only the Durbin-Watson statistics came out a little lower, in the order of 1.7, indicating the possible existence of systematic errors in the model.

As for the pessimistic forecast estimates of this indicator, its equation in this case was as follows:

$$\begin{gathered}\text{TCP} = 641496.493664{*}(@\text{TREND})\wedge 2 + [\text{AR}(1){=}0.565166679309,\text{MA}(1)\\ {=}0.733733540997,\text{UNCOND},\text{ESTSMPL}{=}\text{”2020M01 2021M06”}].\end{gathered}$$

However, the coefficient of determination of this regression came out very high (in the order of 0.998), the t-statistics of the parameters were also quite high, only the Durbin-Watson statistics came out low (in the order of 1.3), indicating systematic errors in the model.

As for the accuracy of the forecast estimates obtained for this indicator, the error in the approximation of the average forecast estimate for this period was 0.75%, which should undoubtedly be considered a good result.

Finally, it should be noted that, in parallel with this work, we discussed the problem of predicting the spread of the coronavirus for Georgia. However, the accuracy of the forecast estimates obtained by us in the latter case would usually fall short of similar figures obtained for the world, which in turn requires some analysis.

REFERENCES

1. Gabelaia A. Problem of predicting the spread of Coronavirus (COVID-19). *Seminar of I. Vekua Institute of Applied Mathematics REPORTS*, **46** (2020), 17-26.

2. Verhulst P. F. Recherches Mathmatiques sur La Loi DAccroissement de la Population, Nouveaux Mmoires de lAcadmie Royale des Sciences et Belles-Lettres de Bruxelles, 18, Art. 1, 1-45, 1845 (Mathematical Researches into the Law of Population Growth Increase).

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

Received 22.09.2021; revised 30.09.2021; accepted 06.10.2021

Author's address:

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