Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Modelling and forecasting of growth rate of new COVID-19 cases in top nine affected countries: Considering conditional variance and asymmetric effect

Aykut Ekinci
Samsun University, Department of Economics and Finance, Samsun, Turkey

1. Introduction

The recent COVID-19 pandemic was first reported in Wuhan, China in December 2019 and has spread out to the whole world in a short period of time. Italy became a new epicenter of the pandemic in March, 2020 in Europe and United Kingdom (UK), France and Germany followed Italy. United States of America (USA) and Brazil were the next epicenters in the world after the first wave in Europe. Unfortunately, nowadays, India is the pandemic's new epicenter of the world. Even after over a year, world is still fighting against COVID-19. The confirmed deaths are over 3 million and total confirmed cases are over 155 million as of April, 2021.

COVID-19 has created devastating effects not just on economies, but also on supply chains, resulting in loss of jobs, increase in poverty and social unrest, causing learning loss in education, worsening human psychology and so on. Such environment has made forecasting new COVID-19 cases or deaths an urgent and essential research area. Many researchers from different disciplines such as health, math, engineering, finance, economics have started to employ different forecasting techniques to get a better insight into spread of pandemic.

Shinde et al. [31] have categorized papers on COVID-19 forecasting into four categories by data and forecasting techniques: (i) big data (ii) social media/other communication media data (iii) stochastic theory/mathematical models and (iv) data science/machine learning techniques.

(i) Papers using big data: This type of papers use mathematical equations or machine learning techniques by using the big data such as temperature, wind speed, humidity, spread rate, disease control interventions, traffic restrictions (see [3,7,20,32,40]) (ii) Papers using social media/other communication media data: These use the correlation between the social media and web searches with the number of daily COVID-19 using datasets from Google, Baidu search engines, mobile phones, newspapers and various websites like Github (see [1,2,9,20,36,41]) (iii) Stochastic theory/mathematical models: This type of models mostly use the susceptible infected-recovered (SIR) model or similar pandemic infection models to estimate the mortality and the spread rate of COVID-19 (see [5,9,10,13,14,21,23,35,37]). (iv) Data science/machine learning techniques: most of the studies are in this category, using time series and/or machine learning techniques. Stated as Shinde et al. [31], the main challenge of this type of studies is that time period for training or estimation period of the model is too short (see [12,16,17]).
Autoregressive Integrated Moving Average (ARIMA) model is well-known time series forecasting methods especially in finance and economics. To forecast the spread of COVID-19, most of the researchers from different disciplines used ARIMA models because of their simplicity, systematic structure and acceptable forecasting performance [38]. As an early study, Ceylan [8] used ARIMA models to forecast COVID-19 prevalence in Italy, Spain, and France between the period of 21/02/2020 and 15/04/2020. Since all series became stationary after the second difference, ARIMA (0,2,1), ARIMA (1,2,0), and ARIMA (0,2,1) were chosen by minimum MAPE value as the best models for Italy, Spain, and France, respectively. The estimation period covers the all data and out-of-sample period for the next 10-day was applied to forecast confirmed cases of COVID-19 in these countries.

Petropoulos et al. [28] attempted to predict the cumulative number of confirmed cases and cumulative number of deaths, using basic time series models, i.e. non-seasonal multiplicative error and multiplicative trend exponential smoothing model. They produced 12 rounds of 10-day-ahead non-overlapping forecasts, covering a four-month period from February to May 2020. Authors claim that their model results offer competitive forecast accuracy. Duan and Zhang [42] predict the daily new confirmed cases in Japan and South Korea from 20/01/2020 until 26/05/2020 using ARIMA. ARIMA(6,1,7) and ARIMA(2,1,3) models were selected as best models for Japan and South Korea, respectively and the daily new confirmed cases for the 7-day period from April 27, 2020 to May 3, 2020 were predicted.

Ribeiro et al. [29] studied on short-term forecasting of COVID-19 cumulative confirmed cases in ten Brazilian states. ARIMA, cubist regression (CUBIST), random forest (RF), ridge regression (RIDGE), support vector regression (SVR) and stacking-ensemble learning were used for one, three, and six-days ahead forecasting of the COVID-19 cumulative confirmed cases. They ranked the models according to their forecasting ability based on sMAPE in all scenarios from the best to the worst, SVR, stacking-ensemble learning, ARIMA, CUBIST, RIDGE, and RF models, respectively.

Maleki et al. [24] modeled the total number of confirmed and recovered COVID-19 cases in the world by the proposed autoregressive time series models based on two-piece scale mixture normal distributions, called TP–SMN–AR models. TP–SMN–AR(7) model was selected as the best model for the estimation period from 22/01/2020 to 30/04/2020 for the total confirmed COVID-19 cases, and TP–SMN–AR(2) model was selected as the best model under the estimation period from 02/02/2020 to 30/04/2020 for the total recovered COVID-19 cases. The validation period was the last 10 days of the confirmed and recovered cases, i.e. 21/04/2020-30.04.2020. Maleki et al. [24] claimed that proposed models improve the forecasting accuracy for confirmed and recovered COVID-19 cases in the world.

Malki et al. [25] investigated the end of COVID-19 pandemic and the risk of second rebound using Seasonal ARIMA (SARIMA) models, selected by AIC. Authors scaled the data using the min-max scalar function to stabilize the value of variance and improved the forecast ability of the model. The estimation period covers 70% of all data, i.e. from 22/01/2020 to 15/06/2020 and the validation period used to forecast for the next 60 days, i.e. from 15/06/2020 until 12/08/2020. Out-of-sample period starts from 13/08/2020 until 11/09/2020. Then Malki et al. [25] try to estimate the expected deadline of COVID-19 for some countries in the first and second rebounds using the forecasted models based on normal distribution. Lastly, authors have compared the forecast performance of ARIMA, Machine learning (Random Forest) and deep learning model (LSTM) and concluded that SARIMA model can be extended and used to predict confirmed cases of other countries.

Taliki et al. [33] forecast the number of confirmed and death caused COVID-19 in Iran using nine different time series models, i.e. NNETAR, ARIMA, Hybrid, Holt-Winter, BSTS, TBATS, Prophet, MLP, and ELM network models. The daily confirmed and death
cases data covers from 20/02/2020 until 15/08/2020 and spined train (70%) and testing part (30%), then the forecasting accuracy of the models is evaluated by RMSE, MAE, and MAPE. The authors selected the ARIMA(1,0,0), i.e. AR(1), and ARIMA(1,0,1), i.e. AR(1), for confirmed and death cases, respectively. Out-of-sample period starts from 16/08/2020 until 14/09/2020 for 30 days. Empirical findings of the paper offer the MLP network model for confirmed cases forecasting and Holt- Winter model for forecasting death cases.

Sahai et al. [30] forecast daily confirmed cases for five most affected countries. ARIMA models estimated for the period of 15/02/2020-30/06/2020 and India (4,2,4), Brazil (3,1,2), Russia (3,0,0), Spain (4,2,4) and US (1,2,1) models selected by AIC. Actual data from 1st July to 18th July was used as validation period and the forecast efficiency was determined by the mean absolute deviation (MAD) and the mean absolute percentage error (MAPE). MAPE was the lowest for Russia and Spain at 1.09% and 0.83%, and 3.70%, 1.84% and 2.88% for India, Brazil and US, respectively.

Kirbas et al. [18] modeled confirmed COVID-19 cases in Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland and Turkey, using three different estimation techniques, i.e. ARIMA, Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM) approaches. Cumulative confirmed case data was used with different baseline date based on the first case was seen. They concluded that LSTM and ARIMA models had the best forecast performance. Katris [43] investigated the estimation of the spread and the reversion of the pandemic from the beginning of the reported cases until August 14 of 2020. The Newbolt/Granger combination scheme of TES, ARIMA and ANN models were built to forecast the daily cases in Greece. Katris built alternative scenarios for the evolution of the pandemic using the model results and the log-normal distribution, which is the best-fit distribution to the data.

However, these studies reviewed above, which use the ARIMA model to forecast COVID-19 confirmed cases or deaths, have two limitations. The first limitation is about data, and the second is about modelling. The paper has three data stages: Data period, data transformation and data cleaning. ARIMA model, which is further explained in the econometric model section, is a univariate linear regression model trying to find right coefficients for the lags of the dependent variable or residual terms based upon the historical returns or growth rates. Longer time period means better forecast ability for a model. Accordingly, the estimation period set in this paper is one-year period, from March 2020 to March 2021. Validation periods are one-day ahead forecast of the April, 2021 and three waves of COVID-19 pandemic.

Time series should be stationary, as explained in the data section below, but as can be seen in the literature, some of the papers use non-stationary new COVID-19 cases data and some use taking the first difference or the second difference of new COVID-19 cases. Taking the first difference is a proper method of making a series stationary, but in case of COVID-19 pandemic, this would lead misspecification of the model. It is obvious that the number of daily new cases were low on the onset of the pandemic, but the new cases reached over 200,000 in USA and India at the late stage of the pandemic or during the strong waves of the pandemic. The growth rate of the daily new COVID-19 cases (7 day smoothed) was used by taking the log difference of the series following the economics or finance literature.

Data cleaning is an important part of time series modelling as most of the time series might have outlier, missing or dirty data. Thus finding the best-fitted ARIMA model is the key part of the modelling, but when the data has outliers, these potential large deviations from mean can make parameter estimation or forecasting inappropriate. New COVID-19 cases have two outliers in USA and one outlier in Germany, so adjusted new cases are used for these outliers. On the other hand, Turkey, one of the fifth most affected country in the world, is excluded from the sample since new COVID-19 cases in Turkey has two structural breaks, which indicates that COVID-19 cases do not follow the natural trajectory of the pandemic spread. So the previous and/or future studies should take outliers in COVID-19 data and time path of the new cases into account.

The second limitation is about modelling. In the literature which aim to forecast of new cases or deaths by using ARIMA, a mean equation is modelled under the assumption of homoscedasticity. ARIMA or mean equation indicates that the current value of the daily confirmed COVID-19 cases is a linear function of its past values as well as current and previous residual series. However, financial returns show volatility clustering, as first noted by Mandelbrot [26] stating that “large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes”. Generalized autoregressive conditional heteroscedasticity (GARCH) model usually has a better performance than traditional statistical methods like ARMA by capturing the non-linearities and heterogeneity of the time series. Yet this paper differs from the literature, using ARMA-GARCH model. Moreover, the asymmetric characteristics in the volatility of new COVID-19 cases were modelled with two most common asymmetric GARCH models – the Threshold GARCH (TARCH) Model and the Exponential GARCH (EGARCH).

In this paper, ARMA, ARMA-GARCH, ARMA-TGARCH and ARMA-EGARCH models are employed for one-day ahead forecasting performance for April, 2021 and three waves of COVID-19 pandemic for nine most affected countries, USA, India, Brazil, France, Russia, UK, Italy, Spain and Germany. The Section 2 introduces the data and econometric model, the Section 3 presents the empirical results and the Section 4 concludes.

2. Data and econometric model

2.1. Data

The number of the daily new confirmed COVID-19 cases for the ten most effected countries was taken from the World Healthy Organization database. The most effected countries were selected according to which countries have the most cumulative COVID-19 cases as of 4th May 2021, which are USA (33.2 million cases), India (20.4 million cases), Brazil (14.7 billion cases), France (5.6 million cases), Turkey (4.9 million cases), Russia (4.8 million cases), UK (4.4 million cases), Italy (4.0 million cases), Spain (3.5 million cases) and Germany (3.4 million cases).

Data cleaning is an important task before modelling the time series. First, for the USA and Germany daily new cases showed some significant outliers. The number of new cases in USA was 229,915 on 12/19/2020 and jumped to 402,270 on 12/20/2020 then decreased to 200,257 on 12/21/2020. The same situation is observed in another date; the number of new cases in USA was 41,486 on 3/9/2021, jumped to 126,229 on 3/10/2021 then decreased to 52,732 on 3/11/2021. For Germany, the number of the new cases was 19,185 on 4/18/2021, 0 on 4/19/2021 and 21,046 on 4/20/2021. To get rid of such outliers in this paper, the outlier days are taken as the average value of the day before and after the outlier day. Accordingly, the adjusted number of the new cases will be 215,086 on 12/20/2020 and 47,109 on 3/10/2021 for USA and 20,115.5 on 4/19/2021 for Germany. The second issue is about the path of time series of Turkey. The new cases in Turkey have two structural breaks, one on 7/29/2020 and one on 11/27/2020. Therefore, Turkey was excluded from the sample (see the attached Excel file for data preprocessing steps).

Reducing the unconditional variance is an important issue when modelling the pandemic spread since the new cases have
an exponential growth characteristic in the early stage of the pandemic. As a first step, the 7-day smoothed series were used by taking 7 days moving average of the daily new confirmed cases. Some countries have included weekend lockdowns or national holiday lockdowns, consequently the cases show an irregular pattern for weekends or the days when there were legal COVID-19 restrictions. Using 7-day moving average helps to avoid this type of changes in the number of the COVID-19 cases. In other words, the forecasted value would be 7-day average of new COVID-19 cases. From now on when referring to models and estimation in this paper, “new cases” means daily new confirmed cases of COVID-19 (7-day smoothed) for the country i, if otherwise noted. As the second step, starting date is determined as the first day when the number of the new cases is over 100. This way it allowed us to get rid of the exponential growth rates of the early days in new cases to some extent.

Our dependent variable, \( Y_t \), should be stationary according the Box-Jenkins procedure. However, the daily confirmed new cases (smoothed) for all countries have a time-dependent trend and the series are non-stationary based on Augmented Dickey Fuller (ADF) and Phillips-Perron (PP) tests\(^1\). The empirical results may be spurious (see [44]), if the series are non-stationary. Taking the first difference, i.e. \( \Delta Y_t = Y_t - Y_{t-1} \), will remove the linear trend and \( \Delta Y_t \) becomes stationary and denotes as \( Y_t \), integrated of order one. In economics and finance literature, stationary condition is often met by using variables growth rate or return, such as real GDP growth or stock returns obtained by taking the log difference of the series, i.e. \( r_t = \ln p_t - \ln p_{t-1} \). Taking logs stabilizes the variance of a series (see [22]). ARMA model is an expected mean ARMA(p,q) model and tries to select the best-fitted model by looking at the past residuals. Then creating a pattern, say selecting the appropriate ARMA(p,q) model from the past growth rates of new COVID-19 cases, becomes easier than obtaining a model from the first difference of the COVID-19 cases, which brings scaling problem and higher variance with it in return.

The dependent variable is daily growth rate, i.e. the log difference of daily new confirmed cases of COVID-19 (7-day smoothed) Figure 1. Table 1 displays the descriptive statistics of growth rate of the new cases in nine most affected countries. The first country, which passed the 100 new cases, is Italy on March 1th, 2020 and the last one is India, which is now the second in the World in regards to total COVID-19 cases. Daily average of the growth rates is fluctuating between 1% ~ 1.5%, the highest mean growth rate is 2.06% in India and the lowest one is 0.74% in UK thanks to vaccination. The maximum daily growth rate of confirmed COVID-19 cases (7 day smoothed) is 107.94% in France and the minimum is again in France with ~78.25%. Standard deviation is maximum in France with 0.1135 and minimum in Russia with 0.0453, and it provides a priori knowledge about considering heterogeneity in forecast models of this study.

The kurtosis of the normal distribution is 3. Then all series have high kurtosis, the distribution is leptokurtic, that is, more peaked than a normal distribution with longer tails, especially for USA. It is an outcome of wild pandemic cycle periods, i.e. the effect of uncontrolled period and controlled period on growth rates of COVID-19 cases. Skewness is a measure of asymmetry and it is expected as zero under the normal distribution. All series have positive skewness implying that the distribution has a long right tail due to high growth rates of COVID-19 cases. Skewness is maximum in USA with 6.97 and minimum in France 0.65. The Jarque-Bera test statistic is calculated based on skewness and kurtosis values of the distribution, as expected it rejects the null hypothesis of a normal distribution for all series.

High standard deviation and skewness&kurtosis values of the series indicate that our forecasting models should take non-linearities and heterogeneity of the time series into account. It is known that ARMA models have constant variance assumption, which means volatility in new COVID-19 cases does not any effect on forecasted new cases. If data have heavy-tailed probability distribution, then GARCH procedure ensures modelling both the conditional heteroscedasticity and the heavy-tailed distributions of the daily growth rate of the new confirmed cases.

2.2. Econometric model

2.2.1. Mean equation or ARIMA

ARIMA model is a well-known forecasting models in finance and economics and first introduced by Box and Jenkins [6] in their seminal text-book– Time Series Analysis: Forecasting and Control. ARIMA model is a type of data generating process, which attempts to find right patterns in the historical returns or growth rates of financial or economic variables.

If \( Y_t \) shows the daily growth rate of the new COVID-19 cases for the country i at time t, It can be modelled as univariate time series as in Eq. (1).

\[
Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots + \theta_p Y_{t-p} + \varepsilon_t \tag{1}
\]

where \( \theta_0 \) is constant term, \( p \) determines the appropriate lag of the autoregressive process of the AR(p) model, and \( \varepsilon_t \) shows the residual term for the country i.

\( Y_t \) can also be modelled as its own past residuals as shown in Eq. (2)

\[
Y_t = \varphi_0 + \varphi_1 \varepsilon_{t-1} + \varphi_2 \varepsilon_{t-2} + \ldots + \varphi_q \varepsilon_{t-q} + \varepsilon_t \tag{2}
\]

where \( \varphi_0 \) indicates the mean of the series, \( q \) shows the appropriate lag of the model and \( \varepsilon_t \) is the white noise term.

\(^1\) The test results can be obtained from author upon request.
The autoregressive component (AR(p)) and moving average component (MA(q)) can be integrated as the autoregressive moving average (ARMA(p,q)) model as follows:

\[ Y_t = \mu + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots + \theta_p Y_{t-p} + \varepsilon_t \varepsilon_{t-1} + \varepsilon_{t-2} + \ldots + \varepsilon_q \varepsilon_{t-q} + \varepsilon_t \]  

(3)

where \( \varepsilon_t \) is independently and identically distributed error term with a mean of zero and a constant variance of \( \sigma^2 \).

2.2.2. Variance equation or GARCH

The main assumption behind the ARMA model is that the expected variance of all error terms is the same at any given point, i.e. \( \varepsilon_t \) is independently and identically distributed error term with a mean of zero and a constant variance of \( \sigma^2 \). This assumption is called homoskedasticity and it is an implausible assumption by Engle [11], who introduced autoregressive conditional heteroskedasticity (ARCH) model as a new class of stochastic process. Bollerslev [4] and Taylor [34] generalized the model as GARCH (1, 1) model shown in Eq. (4):

\[ Y_t = \mu + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_q \varepsilon_{t-q} + \varepsilon_t \]  

(4)

\[ \sigma^2_t = \omega + \alpha \varepsilon^2_{t-1} + \beta \sigma^2_{t-1} \]  

(5)

in which the mean equation is presented in Eq. (4). Eq. (5) is called as the conditional variance equation, where \( \sigma^2_t \) shows the one period ahead forecast variance based on past information. Conditional variance of the growth rate of new cases for country \( i \), \( \sigma^2_i \), is a function of constant, \( \omega \), the ARCH term, \( \varepsilon^2_{t-1} \), and the GARCH term, \( \sigma^2_{t-1} \). ARCH term is the previous day effect or new information effect, measured as the lag of the squared error term from the mean equation. GARCH term indicates long term volatility, measured as last period’s forecast variance. High order GARCH models can be estimated by choosing different p or q, but it is well known that the finance literature GARCH(1,1) model performs better in most cases.

2.2.3. Variance equation with asymmetric effect or the asymmetric GARCH models

The conditional variance of the daily growth rate of new COVID-19 cases is a function of past values of squared error term, without any correlation with the sign of the error term. If the question is about inquiring the differential effects of negative or positive shocks on conditional variance, then asymmetry coefficient should be added up to the variance equation. The Threshold GARCH (TARCH) Model and the Exponential GARCH (EGARCH) are two most known asymmetric GARCH Models.

2.2.3.1. TARCH model. TARCH model was introduced by Glosten, Jaganathan, and Runkle [15] and Zakoian [39]. The conditional variance in the TARCH(1, 1) model:

\[ \sigma^2_t = \omega + \alpha \varepsilon^2_{t-1} + \gamma \varepsilon^{-}_{t-1} + \beta \sigma^2_{t-1} \]  

(6)

where \( \gamma \) is only if \( \varepsilon_t < 0 \) and \( \varepsilon_t = 0 \) otherwise. If the negative shocks have a significant effect on conditional variance on the contrary of positive shocks, then negative shock has an impact of \( \alpha + \gamma \).

2.2.3.2. EGARCH model. EGARCH model was developed by Nelson [27] and conditional variance equation of an EGARCH model is defined as follows:

\[ \log(\sigma^2_t) = \omega + \alpha \varepsilon^2_{t-1} + \delta \varepsilon^{-}_{t-1} + \beta \log(\sigma^2_{t-1}) \]  

(7)

where \( \log(\sigma^2_t) \) is the log of the conditional variance, which provides positive conditional variance even if the parameters are negative. The asymmetric effect can be tested by the hypothesis that \( \delta \neq 0 \). If the asymmetry coefficient, \( \delta \), is statistically insignificant, then there is no asymmetric effect. If \( \delta > 0 \), it implies that positive shocks generate higher volatility more than negative shocks. The parameter \( \beta \) captures the persistence in conditional volatility.

3. Empirical results

3.1. Forecasting the last month, April 2021

As explained in the econometric model section, mean equation shows the ARMA(p,q) model under the assumption constant variance in Eq. (3), considering the time varying conditional variance.
with the GARCH model as presented in Eqs. (4) and (5), and lastly adding the asymmetric effect into the variance equation called as TGARCH model in Eq. (6) and EGARCH model in Eq. (7).

The one-day ahead forecasting performance of the models for April, 2021 is presented in Table 2. The root mean squared error (RMSE), the square root of the mean of the square of all of the error or basically standard deviations of the residuals, is used for the selecting the best model among the four models for the country i, i.e. ARMA(p,q), ARMA(p,q)-GARCH(1,1), ARMA(p,q)-TGARCH(1,1) and ARMA(p,q)-EGARCH (1,1). The main conclusion is that modelling the daily growth rate of COVID-19 new cases with ARMA model means loss of forecasting power. ARMA(p,q)-GARCH(1,1) models for all nine countries have lower RMSE in comparison with ARMA models.

Adding asymmetric effect to the conditional variance equation has positive effect on forecast accuracy for USA, Brazil and UK for EGARCH models, and for all countries except Russia and Spain for T Gavin models. However, this improvement on forecasting accuracy, i.e. lower RMSE, is weak for most of the countries.

As a next step, ARMA(4,4) and ARMA(4,4)-GARCH(1,1) model forecasting results are compared for USA to provide a better insight on difference between the forecasted and actual series. Fig. 2 shows the daily growth rate of the new cases and the forecasted growth rates. It is obvious that ARMA(4,4) model has larger fluctuations on actual growth rate of new cases in USA.

Forecasted value of the new confirmed cases (7 day smoothed) can be calculated using the forecasted growth rates obtained from two main models, i.e. ARMA(4,4) and ARMA(4,4)-GARCH(1,1) models (see Fig. 3). It can be easily seen in Fig. 3 that ignoring the conditional variance caused weak forecast power.

Fig. 4 presents the deviations from actual new cases, the blue bar shows the deviations from actual new cases from forecasted ones by ARMA(1,1) and red bar shows the deviations from actual new cases from forecasted ones by ARMA(1,1)- GARCH(1,1). For example, first bars show the deviations on 1st April. The new cases (7 day smoothed) on April 1 in USA was 63,167, it is forecasted as 61,713 by ARMA(4,4) model and 63,268 by ARMA(1,1)-GARCH(1,1); then the deviation is so small for GARCH model as seen Fig. 4.

Table 2 shows the statistical properties of deviations presented in Fig. 4. The mean of the deviations for 30 observations is 152 new cases for GARCH model but 471 new cases for ARMA model. It is clear that the maximum (3,504 for ARMA and 1,504 for GARCH) and the minimum (-3,522 for ARMA and -1,260 for GARCH) deviations from the actual ones are so high for ARMA model, which causes a high standard deviation for ARMA model. Under the null hypothesis of a normal distribution, Jarque-Bera test statistics confirm that both series have normal distribution, which also confirms that specifications of the models are built successfully.

3.2. Forecasting three waves of new COVID-19 cases or forecasting turning points

While a model has higher forecasting accuracy for normal periods, it is harder to forecast non-normal periods such as financial crisis, financial stress periods or disease outbreaks such as COVID-19. Most of the countries faced three COVID-19 waves; first wave was at the onset of the COVID-19 pandemic, second wave was seen during the last months of the 2020 and third one was around March, 2021. However, pandemic has its own growing path for most of the countries since its spread rate depends on so many variables such as age structure, health expenditure per capita, GDP per capita, population density, public measures and so on.

Fig. 5 presents COVID-19 waves in the nine most affected countries. The waves were defined based on before and after 15 days
of the local maximum of the new cases (smoothed). As seen in Fig. 5 and Table 3, the time periods of the waves have changed between countries; India has single wave while Russia has two waves.

Table 3 shows the one-step ahead forecast model results in terms of RMSE for the first COVID-19 wave. The model results confirm the forecast results of April, 2021; all countries have lower RMSE when using the ARMA(p,q)-GARCH(1,1) model. Taking into account heteroscedasticity is an essential part for forecasting models. However, asymmetric effect has shown some mixed results as it did in baseline validation period, April 2021 (see Table 2). Some of the TARCH or EGARCH models are better than GARCH models, for example while TARCH works better for India and for Brazil, France, Russia, UK for both models work well; on the other hand, Italy, Spain and Germany have higher RMSE values, even higher than ARMA models in some cases.

Table 4 shows the second COVID-19 wave for eight countries. ARMA(p,q)-GARCH(1,1) models have still lower RMSE but not for all countries, France and UK have slightly lower RMSE under the ARMA models. Asymmetric GARCH Models have still mixed performance results.

The model results for the third COVID-19 wave is presented in Table 5. GARCH models have better performance accuracy for USA, Brazil and Germany in terms of RMSE, the loss of forecasting power is small for France, UK, Italy and strong for Spain. Including

---

3 EGARCH model for Spain is an exception with the lowest RMSE value.
sumption.

Table 5
One-step ahead forecast for the third COVID-19 wave.

| Country | Third COVID-19 Wave | Mean Equation | Variance Eq. | Asymmetric GARCH Models |
|---------|---------------------|---------------|--------------|------------------------|
|         | Period              | ARMA(p,q)     | RMSE         | ARMA(p,q) -GARCH(1,1)  | ARMA(p,q) -TGARCH(1,1) | ARMA(p,q) -EGARCH(1,1) |
| USA     | 12/26/2020 1/25/2021| ARMA(4,4)     | 0.02984      | 0.0247                | 0.0255                 | 0.0255                 |
| Brazil  | 1/4/2021 2/3/2021   | ARMA(4,4)     | 0.04089      | 0.0310                | 0.0308                 | 0.0340                 |
| France  | 3/23/2021 4/22/2021 | ARMA(4,3)     | 0.06897      | 0.0694                | 0.0712                 | 0.0676                 |
| UK      | 12/22/2020 1/21/2021| ARMA(4,2)     | 0.04405      | 0.0455                | 0.0442                 | 0.0458                 |
| Italy   | 3/8/2021 4/7/2021   | ARMA(4,3)     | 0.02319      | 0.0244                | 0.0237                 | 0.0261                 |
| Spain   | 1/8/2021 2/7/2021   | ARMA(4,4)     | 0.04058      | 0.0470                | 0.0471                 | 0.0468                 |
| Germany | 4/16/2021 4/30/2021 | ARMA(4,3)     | 0.03242      | 0.0303                | 0.0296                 | 0.0303                 |

Fig. 5. COVID-19 waves (shaded areas) in nine most affected countries.

asymmetric effect into the variance equation creates higher RMSE for most of the countries.

4. Conclusion

COVID-19 pandemic has been in our lives for more than a year and there is a growing literature focusing on forecasting the parameters of pandemic spread such as total cases, new cases, mortality rates, spread rate and so on. Since COVID-19 pandemic is a multidisciplinary research area, many researchers from different disciplines have contributed to the literature using mathematical models or machine learning techniques. ARIMA models are well-known time series models to forecast univariate variables such as new COVID-19 cases. On the other hand, the main assumption behind the ARIMA models are that the expected variance of all error terms is the same at any given point, i.e. homoskedasticity assumption. However, COVID-19 pandemic has strong waves, which leads heavy-tailed distributions. GARCH models are able to capture the non-linearities and heterogeneity of the time series. Asymmetric effect is another feature of GARCH-type models, i.e. TGARCH and EGARCH, by measuring the differential effects of negative or positive shocks on conditional variance.

When modelling COVID-19 new cases, the growth rate of daily COVID-19 new cases (7 day smoothed) is used by taking into account the outlier effects on the data and last one-year period data can be used for the estimation period. Validation periods are one-day ahead forecast of the April, 2021 and three waves of COVID-19 pandemic. ARMA, ARMA-GARCH, ARMA-TGARCH and ARMA-EGARCH models are employed for one-day ahead forecasting performance for April, 2021 and three waves of the COVID-19 for the nine most affected countries– USA, India, Brazil, France, Russia, UK, Italy, Spain and Germany. ARMA(p,q)-GARCH(1,1) models have lower RMSE for the period of April, 2021 and first wave of COVID-19 pandemic for all countries. It is same for the second wave of
pandemic for all countries except France and Russia but for the third waves of the pandemic USA, Brazil and Germany have lower RMSE. As shown in the paper, ARMA-TGARCH and ARMA-EGARCH models have mixed results in terms of achieving the lower RMSE.

The paper has some suggestions for future research. (i) Distribution of new cases of COVID-19 has heavy-tailed distributions with high variance. Therefore, GARCH modelling is essential to capture non-linearities in data and rising forecast accuracy. (ii) Most of the machine learning models such as artificial neural networks, support vector machines do not have restrictions or data transformation as did in time series models. Using packages in Python, R or similar programs also support this type of analysis, without noticing the data cleaning and data transformation. Before the modelling stage, we should make sure that input data is appropriate for this kind of modelling. (iii) Seasonal ARIMA models can be estimated for capturing the seasonality effect on the data. (iv) Three or seven days ahead forecasting can be employed to see longer forecasting performance of the models.

Declaration of Competing Interest

The author declares that I have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.chaos.2021.112227.

CRediT authorship contribution statement

Aykut Ekinci: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing.

References

[1] Anastassopoulos C, Russo L, Tsalikis A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS ONE 2020;15(3): e0230465.
[2] Bayham, J., & Fenichel, E.P. (2020). The impact of school closure for COVID-19 on the US healthcare workforce and the net mortality effects. Available at SSRN 3555259.
[3] Bhattacharjee, S. (2020). Statistical investigation of relationship between spread of coronavirus disease (COVID-19) and environmental factors based on study of four mostly affected places of China and five mostly affected places of Italy. arXiv preprint arXiv:2003.11277.
[4] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. J Econom 1986;31(3):307–27.
[5] Botha, A.E. & Dedman, W. (2020). A simple iterative map forecast of the COVID-19 pandemic. arXiv preprint arXiv:2003.10532.
[6] Box G, Jenkins G. Time series analysis, forecasting and control. San Francisco, CA: Holden-Day: 1970.
[7] Caccavo, D. (2020). Chinese and Italian COVID-19 outbreaks can be correctly described by a modified SIRD model. medRxiv.
[8] Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020;729:138817.
[9] Coelho, F.C., Lana, R.M., Cruz, O.G., Cedeço, C.T., Vilela, D., Bastos, L.S., & Gomes, M.F. (2020). Assessing the potential impacts of COVID-19 in Brazil: mobility, morbidity and impact to the health system. medRxiv.
[10] Demongeot J, Greitese Q, Magal P. SI epidemic model applied to COVID-19 data in mainland China. R Soc Open Sci 2020;7(12):201878.
[11] Engel RF. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 1982;50:987–1007.
[12] Feng SJ, Li G, Dey N, Crespo RG, Herrera-Viedma E. Composite Monte Carlo decision making under high uncertainty of novel coronavirus epidemic using hybridised deep learning and fuzzy rule induction. Appl Soft Comput 2020;93:106282.