ECONOMETRICS
EKONOMETRIA
Advances in Applied Data Analysis

Econometrics. Ekonometria.
Advances in Applied Data Analysis
Year 2025, Vol. 29, No. 4
ISSN 2449-9994
journals.ue.wroc.pl/eada

Variability in Times of Disease. Application of ARMA-GARCH in Modelling and Predicting Volatility of S&P500 Index Return Rates in COVID-19

Damian Wiśniewski

University of Warmia and Mazury in Olsztyn

e-mail: dawi@matman.uwm.edu.pl

ORCID: 0000-0001-7909-6903

Quote as: Wiśniewski, D. (2025). Variability in Times of Disease. Application of ARMA-GARCH in Modelling and Predicting Volatility of S&P500 Index Return Rates in COVID-19. *Econometrics. Ekonometria. Advances in Applied Data Analysis*, *29*(4), 18-32.

DOI: 10.15611/eada.2025.4.02

JEL: C32, G01

Abstract

Aim: The article considers the time series case of the closing prices of the S&P500 index over the period from January 2020 to April 2021. The author selected the best ARMA(p,q)-GARCH(1,1) models with different forms of probability density functions. The errors of the forecasts generated both in terms of logarithmic returns and their variability were compared.

Methodology: The study followed the Box-Jenkins procedure. Applying the information criterion the study considered the best among these models with normal, skewed Student's t, generalised error and generalised hyperbolic distribution.

Results: The author obtained the following representations: ARMA(2,0)-GARCH(1,1) and ARMA(0,2)-GARCH(1,1), with normal, skewed Student's t and generalised error distribution. The assessment of forecast accuracy showed that in the case of conditional variance forecasts, the ARMA(2,0)GARCH(1,1) models with a normal distribution and a generalised error distribution were the best. The largest errors of conditional variance forecasts were generated by models with a skewed Student's t-distribution.

Implications and recommendations: It is worth extended the study to models based on the range of fluctuations (such as Range GARCH-RGARCH or Conditional Autoregressive Range Model-CARR).

Originality/value: The author considered models with various probability density functions, showing that such diversity was important when looking for the best models in times of high volatility.

Keywords: the COVID-19 pandemic, ARMA-GARCH, time series, S&P500 index, stationarity

1. Introduction

The COVID-19 pandemic has significantly affected global economies and caused uncertainty in global stock markets. There are studies (e.g. Baker et al., 2020) which showed that no previous epidemic had disrupted stock market trading as significantly as the pandemic did. In recent years there have been numerous papers modelling the impact of the COVID-19 pandemic on the volatility of financial markets from different parts of the world, using models from the GARCH type family. For example, its impact on the stock market crash risk in China (Liu et al., 2021) was investigated using the GARCH-S model, whilst the impact of the two waves of COVID-19 infections on the return and volatility of the stock market indices of the euro area countries (Duttilo et al., 2021) was examined applying the GARCH(1,1)-in-Mean model, and the short-term reaction of the Visegrad countries' financial markets to the COVID-19 pandemic was assessed using the TGARCH model (Czech et al., 2020). There are also studies in which the conditional mean and variance equation was modified by including a dummy variable for the pandemic period (Chaudhary et al., 2020).

What distinguishes this article is the consideration of the ARMA-GARCH models in the pandemic period with different forms of the probability density function. A common assumption in the analysis of returns is that they are characterised by a distribution with time-invariant parameters and, additionally, that this distribution is normal. The assumption of normality of distributions has been questioned based on empirical research (see: Mandelbrot, 1963 and Fama, 1965), which documented the fact that most return distributions are leptokurtic distributions with thicker tails than a normal distribution. Hence it is also reasonable to consider distributions that allow for the effect of thick tails, such as the Student's t-distribution, the generalised error distribution or the generalised hyperbolic distribution. Furthermore, in empirical time series, negative or positive rates of return occur more frequently than rates of return with the opposite sign, therefore one should take into account the skewness of the distributions of rates of return (see: Kon, 1984; Alles & Kling, 1994). Thus, the article considers models with not only the normal, but also skewed Student's t, generalised error and generalised hyperbolic distribution. The research goal was to investigate whether such models perform better (in terms of forecast errors) with the unprecedented market turbulence caused by the COVID-19 pandemic.

The author employed the daily quotations of the S&P500 index from 2 January 2020 to 30 April 2021 with 335 observations. The stock prices were taken from https://stooq.pl/. The designations successively used in the figures and frequently in the text are: SP500 (quoted closing price of the S&P500) and r_t (logarithmic returns of the SP500). The selected time period dates back to the beginning of the general knowledge of the pandemic. As early as 31 December 2019, the WHO branch in the People's Republic of China noted a statement on the Wuhan Municipal Health Commission's website about cases of 'viral pneumonia'. On the same day, a report by the International Society of Infectious Diseases referring to the same group of cases of 'pneumonia of unknown cause' was also published. On 1 January 2020, WHO requested information from the Chinese authorities on unusual cases of pneumonia in Wuhan and on 4 January WHO's first tweet stated that a group of pneumonia cases had occurred in Wuhan, with no deaths, and that investigations were underway to identify the cause.

The end of April 2021 marked a turning point in the history of the COVID-19 pandemic, as it was then that the first clear evidence from real-world studies emerged that mRNA vaccines were not only effective in reducing the risk of infection, but above all protected against the most severe effects of the disease – hospitalisation and death. The CDC's (Center for Disease Control and Prevention – the national public health agency of the United States) indication that the Pfizer-BioNTech and Moderna vaccines reduce the risk of hospitalisation in older people by as much as 94%, thus providing a real tool to control the course of the pandemic. This shifted the perspective from uncertainty and reactive crisis management to a conscious strategy of returning to normality. As a result, the end of April 2021 became a moment when the fight against COVID-19 began to be based not only on restrictions, but on scientific evidence confirming the effective protection of entire populations.

The methodology of the study followed the Box-Jenkins procedure. In a first step, the degree of integration of the SP500 series was investigated by visual inspection of the time series and correlogram, as well as using the ADF and KPSS tests. Then, arbitrarily guided by the significance of the coefficients of the ACF and PACF functions, the maximum lag values p_{max} and q_{max} were determined. Next, the coefficients of all the ARMA(p,q) models for which $0 \leq p \leq p_{max}, 0 \leq q \leq q_{max}$ were estimated using the Maximum Reliability Method. Only those models for which the conditions are simultaneously satisfied were left for further consideration: the parameters at the lagged variables are significant at the significance level of 0.05 and the roots of the characteristic equations are, as far as the modulus is concerned, greater than 1. For the models selected in this way, the best were indicated in light of the information criteria. The author acknowledges that these criteria were not, however, the main reason to reject particular models from further analysis. The parameters at the lagged variables were significant at the significance level of 0.05, and the roots of the characteristic equations, as far as the modulus was concerned, were greater than 1. For the models selected in this way, an indication of the absence of autocorrelation of the residuals from the model (verified by the Ljung-Box test) were made. In addition, tests were carried out for the normality of the distribution of the residuals (using the Doornik-Hansen test) and the ARCH effect (the Engle test).

2. Determining the Best ARMA-GARCH Models

2.1. Order of Integration

A visual inspection of Figure 1a, showing the SP500 time series, indicates that it is a process that rarely goes through its mean, which allowed to presume its non-stationarity. The Doornik-Hansen test with a statistic of χ^2 = 5.28 and p = 0.0714 suggested that there were no grounds to reject the null hypothesis of normality of the distribution at the assumed significance level, as shown in Figure 2a.

The situation is different when plotting the time series of the S&P500 logarithmic returns. On the basis of Figure 1b, one can speculate that the r_t series is stationary as it passes through its mean relatively frequently. This shows how great an impact the beginning of the general awareness of the pandemic and its spread had on the level of volatility of logarithmic returns. As a result of the significant I – but mostly uncertain – information emerging regarding the virus, fears about its spread, contagiousness, the spiralling speculation and market panic, the declines of late February and March 2020 were accompanied by very high levels of volatility in the rates. Moreover, in April 2020, with a certain degree of stabilisation of the threat, a greater awareness and understanding of the situation, there were also rises indicating the hope that everything, perhaps slowly but surely, would return to normal.

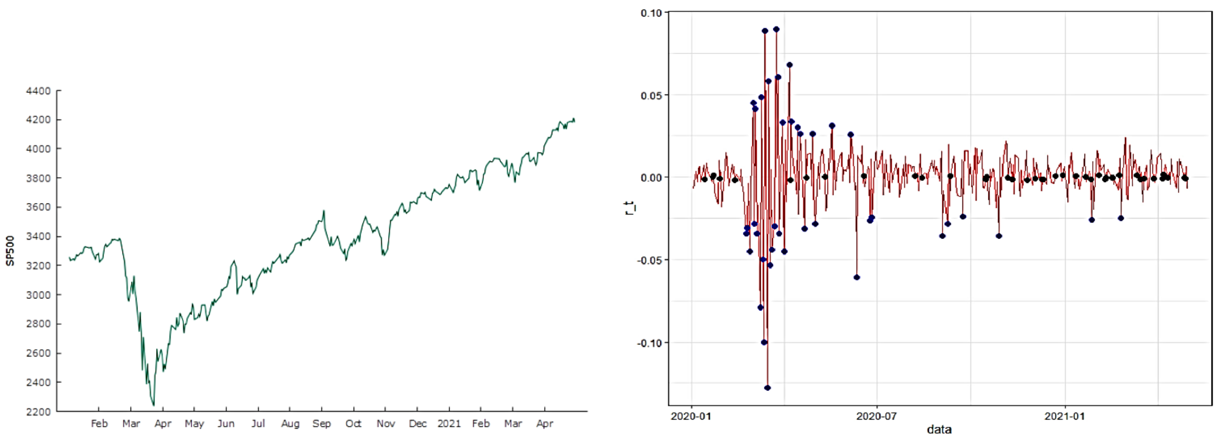

Fig. 1. a. S&P 500 Index closing price quotations, b. Time series of logarithmic S&P 500 Index returns

Note: The 30 observations with the largest logarithmic return values are shown in blue, while the 30 observations with the smallest values are shown in black.

Source: own elaboration.

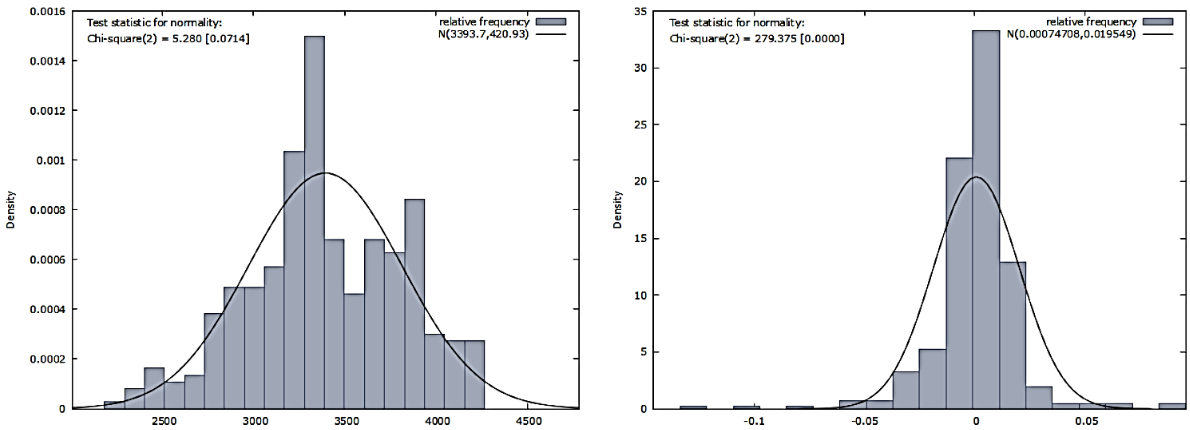

Fig. 2. a. Density plot for S&P 500 Index closing price quotations, b. Density plot for logarithmic returns from S&P 500 Index

Source: own elaboration.

In the case of logarithmic returns, according to the Doornik-Hansen test, the null hypothesis should be rejected in favour of the alternative hypothesis, i.e. the series *rt* does not have a normal distribution. The distribution was found to be leptokurtic, with the values of the series clustering around the mean more than in a normal distribution, and at the same time there was a relatively high probability of outliers. As one can see from Figure 2b this was indeed the case – in the central part, as well as for the outliers, there were relatively many more observations than in a normal distribution. Thus, this is a distribution with fat tails, in accordance with the typical properties of returns (Bollerslev et al., 1994).

This can also be seen in Figure 3 of the quantile-quantile type, where the quantiles of the normal distribution were placed on the horizontal axis and the quantiles of the observed values on the vertical axis. For a sample from a normal distribution, the graph should (almost) coincide with the straight line $y = x$ (Johnson & Wichern, 2002). One can see that for the logarithmic returns of the S&P500 index, the curve highlighted in red deviates systematically from the straight line, particularly at the extremes, in the typical case of leptokurticism.

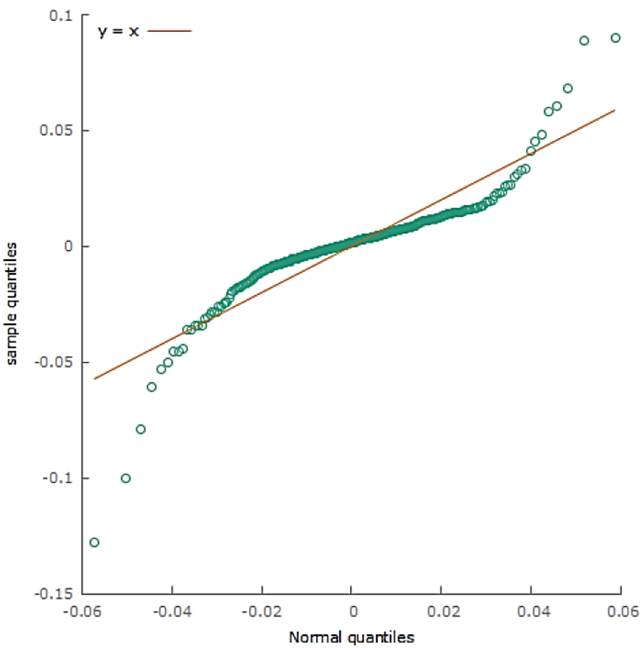

Fig. 3. Q-Q plot for logarithmic returns of S&P 500 Index closing price

Source: own elaboration.

A similar assumption about both series can be made by analysing the autocorrelation (ACF) and partial autocorrelation (PACF) functions from the sample shown in Figure 4. The bars, which are a graphical representation of the autocorrelation function values from the sample, converge very slowly to zero, hence the SP500 variable was most likely non-stationary. For the r_t process, both the autocorrelation function and the partial autocorrelation function are infinite, having the shape of damped sinusoids.

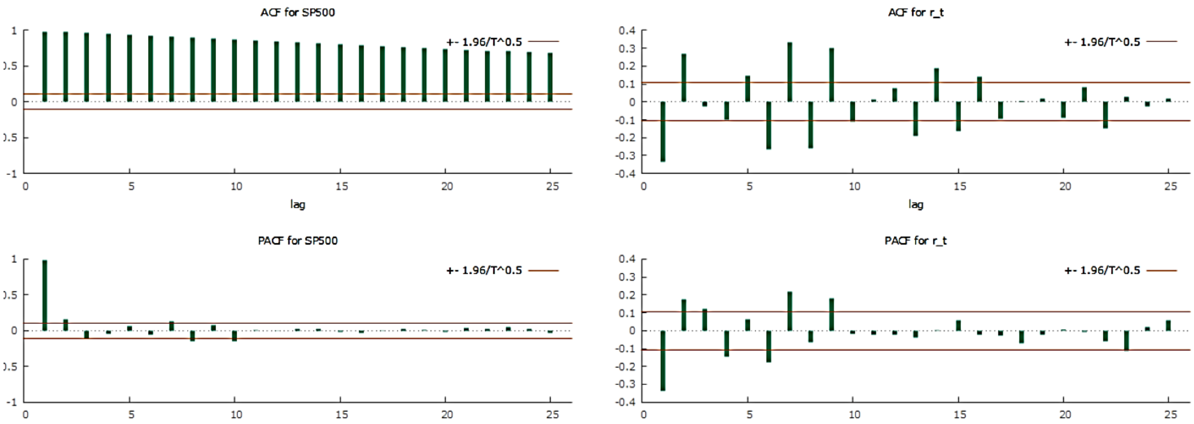

Fig. 4. Plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) for quotes of S&P 500 Index closing prices and logarithmic returns of S&P 500

Source: own elaboration.

Furthermore, based on the values of the Ljung-Box statistics (as shown in Table 1) with p value at 0, the author rejected the null hypothesis of the white noise of the r_t process, as well as the SP500 time series itself, in favour of the alternative hypothesis.

Table 1. Values of the ACF and PACF functions and the Ljung-Box statistic for the S&P500 closing price quotes (columns 2-5) and logarithmic returns of the S&P500 (columns 6-9) (***, **, * indicate significance at the 1%, 5%, 10% levels)

LAG	ACF	PACF	Q-stat.	p-value	ACF	PACF	Q-stat.	p-value
1	-0.3352 ***	-0.3352 ***	37.8717	[0.000]	0.9858 ***	0.9858 ***	328.4492	[0.000]
2	0.2683 ***	0.1757 **	62.2076	[0.000]	0.9761 ***	0.1552 ***	651.4658	[0.000]
3	-0.0265	0.1238 **	62.4458	[0.000]	0.9627 ***	-0.1156 **	966.5957	[0.000]
4	-0.1016 *	-0.1467 **	65.9539	[0.000]	0.9494 ***	-0.0388	1274.0248	[0.000]
5	0.1447 ***	0.0627	73.0994	[0.000]	0.9376 ***	0.0574	1574.7508	[0.000]
6	-0.2665 ***	-0.1769 ***	97.4045	[0.000]	0.9240 ***	-0.0504	1867.7093	[0.000]
7	0.3317 ***	0.2175 ***	135.1754	[0.000]	0.9149 ***	0.1300 **	2155.8361	[0.000]
8	-0.2590 ***	-0.0653	158.2614	[0.000]	0.9005 ***	-0.1498 ***	2435.8169	[0.000]
9	0.3001 ***	0.1830 ***	189.3670	[0.000]	0.8904 ***	0.0759	2710.3727	[0.000]
10	-0.1104 **	-0.0188	193.5850	[0.000]	0.8749 ***	-0.1508 ***	2976.2440	[0.000]
11	0.0104	-0.0204	193.6224	[0.000]	0.8609 ***	0.0072	3234.4634	[0.000]
12	0.0780	-0.0228	195.7408	[0.000]	0.8467 ***	-0.0020	3485.0423	[0.000]
13	-0.1928 ***	-0.0380	208.7400	[0.000]	0.8322 ***	0.0229	3727.8493	[0.000]
14	0.1874 ***	0.0026	221.0522	[0.000]	0.8203 ***	0.0246	3964.5154	[0.000]
15	-0.1617 ***	0.0583	230.2513	[0.000]	0.8059 ***	-0.0225	4193.6553	[0.000]
16	0.1426 ***	-0.0203	237.4278	[0.000]	0.7937 ***	-0.0298	4416.5834	[0.000]
17	-0.0948 *	-0.0266	240.6096	[0.000]	0.7795 ***	0.0046	4632.3232	[0.000]
18	0.0046	-0.0721	240.6170	[0.000]	0.7674 ***	0.0241	4842.0697	[0.000]
19	0.0167	-0.0231	240.7158	[0.000]	0.7547 ***	0.0088	5045.5310	[0.000]
20	-0.0911 *	0.0048	243.6823	[0.000]	0.7416 ***	-0.0183	5242.6334	[0.000]
21	0.0818	-0.0086	246.0824	[0.000]	0.7305 ***	0.0319	5434.4705	[0.000]
22	-0.1498 ***	-0.0621	254.1531	[0.000]	0.7187 ***	0.0175	5620.7847	[0.000]
23	0.0286	-0.1107 **	254.4492	[0.000]	0.7099 ***	0.0543	5803.1168	[0.000]
24	-0.0242	0.0193	254.6620	[0.000]	0.7003 ***	0.0196	5981.1385	[0.000]
25	0.0188	0.0561	254.7906	[0.000]	0.6912 ***	-0.0251	6155.1809	[0.000]

Source: own elaboration.

In order to formally verify the assumptions formulated, the study used the ADF unit root test (augmented Dickey-Fuller test) and KPSS stationarity test (Kwiatkowski-Philips-Schmidt-Shin test).

In the ADF test, the hypothesis was verified of the presence of a unit root for a given series x_t (Dickey & Fuller, 1979). From the data provided in Table 2, there was no basis for rejecting the null hypothesis of non-stationarity for the SP500 series, as evidenced by test statistic values greater than the asymptotic critical values, together with empirical probabilities greater than 0.05. Yet, for the r_t series, the value of the test statistics was in each case smaller than the corresponding asymptotic critical values, with an empirical probability close to zero, therefore the null hypothesis was rejected in favour of stationarity of the logarithmic returns.

Table 2. ADF test results

		ADF test without constant	ADF test with constant	ADF test with constant and trend
SP500	test statistic	0.720	-0.634	-3.179
	asymptotic p-value	0.871	0.861	0.089
	AIC	3486.1	3487.55	3478.41
r_t	test statistic	**-4.788**	**-4.825**	**-4.908**
	asymptotic p-value	0.000	0.000	0.000
	AIC	-1710.46	-1708.87	-1707.71

Note: Bold font was used to highlight the evaluations of the test statistics on the basis of which the null hypothesis of non-stationarity was rejected in favour of process stationarity.

Source: own elaboration.

Let us now consider the KPSS stationarity test, where the hypothesis arrangement was the opposite of the ADF test (Kwiatkowski et al., 1992). The results are presented in Table 3.

Table 3. KPSS test results

	KPSS test without trend	KPSS test with trend
SP500	4.490	0.483
r_t	**0.154**	**0.058**

Note: Bold font was used to write the evaluations of the test statistic, based on which there were no grounds to reject the null hypothesis of stationarity at the $\alpha = 0.05$ significance level

Source: own elaboration.

The critical value for KPSS test without trend was 0.462, and with trend, 0.148. Thus, the null hypothesis of stationarity of the S&P500 index closing quotation series was rejected in favour of its non-stationarity, while for the logarithmic return series there were no grounds to reject the null hypothesis of stationarity of the process.

2.2. The Best ARMA(p,q) Models

Both tests considered in the previous subsection and a visual inspection of the time series graphs indicated that the SP500 series was integrated of order one and r_t was integrated of order 0, thus the author proceeded to construct the ARMA(p, q) (*Autoregressive Moving Average*) model defined as a combination of the AR(p) autoregressive model and the MA(q) moving average:

$$r_t = c + \sum_{i=1}^{p} \delta_i r_{t-i} + \sum_{j=1}^{q} \eta_j \varepsilon_{t-j} + \varepsilon_t,$$

where δ_i, η_j are the parameters appearing in the autoregressive and moving average parts respectively, c is some constant, and $\varepsilon_{t-j}, t = 0,1, \dots, q$.

In line with the Box-Jenkins procedure, the maximum values of p and q in the ARMA(p, q) model were 9 and 22, respectively. Based on the results obtained in Gretl, the only models for which the conditions were satisfied simultaneously where the parameters at lagged variables were significant at the 0.05 significance level, and the roots of the characteristic equations as to modulus greater than 1 (which is a guarantee of stationarity and reversibility of the process) were those presented in Table 4.

Table 4. Values of the information criteria AIC, BIC, H-Q

ARMA(p,q)	AIC	BIC	H-Q
(1, 0)	-1715.34	-1703.91	-1710.78
(0, 1)	-1702.75	-1691.32	-1698.20
(1, 1)	-1718.46	-1703.22	-1712.39
(2, 0)	-1723.78	-1708.53	-1717.70
(0, 2)	-1729.49	**-1714.25**	**-1723.42**
(3, 0)	-1726.90	-1707.84	-1719.30
(0, 6)	**-1732.31**	-1701.98	-1720.31

Note: The best models within each information criterion are in bold.

Source: own elaboration.

In light of the Schwartz Bayesian information criterion and the Hannan-Quinn information criterion, the best among the models considered was ARMA(0,2), whilst the ARMA(0,6) model was the best in terms of the Akaike criterion, yet it is worth noting that this criterion favours the more elaborate models. Due to the compatibility of the two criteria and the much simpler structure (compared to ARMA(0,6)), ARMA(0,2) was found to be the best-fitting model: $r_t = 0.00075 - 0.27051\varepsilon_{t-1} + 0.3074\varepsilon_{t-2} + \varepsilon_t$.

Next, the goodness of fit of all the above models was verified by assessing the normality of the residuals distribution and the absence of autocorrelation, as well as whether there was an ARCH effect (as shown in Table 5). Regarding the Doornik-Hansen test, the high values of the test statistic χ^2 with $p = 0$ indicated that the null hypothesis of normality of the distribution of the residuals should be rejected in any case. This was of no further consequence due to the fact that that the study used the method of the greatest reliability and not the least squares method.

Table 5. Results of Doornik-Hansen, Ljung-Box and Engle tests for selected ARMA models

ARMA(p,q)	Doornik - Hansen test		Ljung - Box test		Engle test	
	test statistic	p-value	test statistic	p-value	test statistic	p-value
(1,0)	144.83	0.00	17.89	0.00	**82.44**	0.00
(0,1)	153.46	0.00	29.98	0.00	**109.33**	0.00
(1,1)	170.66	0.00	11.74	0.01	**69.42**	0.00
(2,0)	227.39	0.00	**6.12**	0.11	**59.08**	0.00
(0,2)	238.52	0.00	**3.07**	0.38	**59.18**	0.00
(3,0)	233.25	0.00	**4.16**	0.12	**56.56**	0.00
(0,6)	110.89	0.00	21.22	0.00	**110.07**	0.00

Note: Bold values of test statistics indicating the absence of autocorrelation of the residuals (Ljung-Box test) and the presence of the ARCH effect (Engle test) at 0.05 significance level. In addition, models satisfying both conditions are in bold.

Source: own elaboration.

The Ljung-Box test is based on the statistic $Q'(m) = T(T+2)\sum_{i=1}^{m}\frac{\hat{\varrho}_i^2}{T-i}$, where T is the sample size (the number of observations), $\hat{\varrho}_i$ is the autocorrelation coefficient from the sample, and m is the number of lags tested [1]. The results obtained in this study indicated that there were no grounds to reject the null hypothesis of no autocorrelation of the random components up to and including order 5 for the ARMA(2,0), ARMA(0,2), ARMA(3,0) models due to the value of the test statistic Q' of 6.12, 3.07, 4.16 with p values of 0.11, 0.38 and 0.12, respectively. In conclusion, the correct lag distributions (p, q) were used, namely $(2,0)$, $(0,2)$, $(3,0)$.

The Engle test checked whether the random component of the model had a constant conditional variance, by considering regression $\varepsilon_t^2 = \alpha_0 + \sum_{i=1}^{S} \alpha_i \varepsilon_{t-i}^2 + u_t$, where S is the length of the series of residuals. In this case, placing $S = 5$ for values of all LM test statistics obtained contained between 56 and 111 along with p statistically equal to 0, indicated that we reject the null hypothesis that there was an ARCH effect in the rest of the models considered. Thus, there was variability in the conditional variance in each case. A similar conclusion was obtained by following the McLeod-Li test, i.e. by examining the squares of the model residuals (McLeod & Li, 1983). The values of the Ljung-Box statistic implied that the null hypothesis of the quadratic squares of the residuals were white noise and should be rejected in favour of the alternative hypothesis, which in turn again suggested that there was an ARCH effect in the residuals.

2.3. ARMA(p,q)-GARCH(1,1)

Due to the presence of the ARCH effect, ARMA models were not sufficient to describe the evolution of SP500 returns. It was necessary to specify an ARMA-GARCH type model in which the linear component and the variability of the conditional variance were estimated simultaneously. Based on the results of (Hansen and Lunde, 2005) (see also Doman & Doman, 2009, p.81), GARCH(1,1) was chosen for further study, therefore considering models of the following form:

$$\text{ARMA(p,q)-GARCH(1,1)}: \begin{cases} r_t = c + \sum_{i=1}^{p} \delta_i r_{t-i} + \sum_{j=1}^{q} \eta_j \varepsilon_{t-j} + \varepsilon_t \\ h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1} \\ \varepsilon_t = z_t \sqrt{h_t}, \alpha_0, \alpha_1, \beta_1 > 0. \end{cases}$$

Based on the tests performed, the models worthy of further study were: ARMA(2,0)-GARCH(1,1), ARMA(0,2)-GARCH(1,1) and ARMA(3,0)-GARCH(1,1). As mentioned in the introduction, the author examined not only normal distribution ("norm" in the graphs and tables) but also the skewed Student's t-distribution ("sstd"), the generalised error distribution (GED) ("ged") and the generalised hyperbolic distribution ("ghyp") (Piontek, 2002). Their use led to better, in terms of information criteria, ARMA(p,q)-GARCH(1,1) models shown in Figure 5.

The best, according to both information criteria, were the models with a skewed Student's t-distribution, followed by those with a hyperbolic distribution, with a generalised error distribution, and finally a normal distribution. However, it should be noted that few models had statistically significant coefficients (at the 0.05 significance level). The coefficients of the ARMA(2,0)-GARCH(1,1) and ARMA(0,2)-GARCH(1,1) models for each of the three distributions considered are shown in Table 6.

[1] An interesting issue is the choice of maximum lag in this test. Reviewing chronologically the most important propositions we have: $m = 5$ (Ljung, 1986), $m = ln\, T$ (Tsay, 2005), $m = 20$ (Shumway & Stoffer, 2017), or $m = \min\left(10, \frac{T}{5}\right)$ (Hyndman & Athanasopoulos, 2018). Thus, this is an unsolved problem. Moreover, the conclusions of a recent study (Hassani & Yeganegi, 2020) indicated that, for tests with a power of at least 0.9, all the cited propositions are flawed in general, as they are either underestimated for large (T around 1000) trials ($m = 5$, $m * ln(T)$), or overestimated for smaller (T around 500) trials ($m = 20, m = \min\left(10, \frac{T}{5}\right)$). Furthermore, the authors pointed out that for large samples, the optimal lag level depends not only on the length of the time series but also on the adopted level of significance.

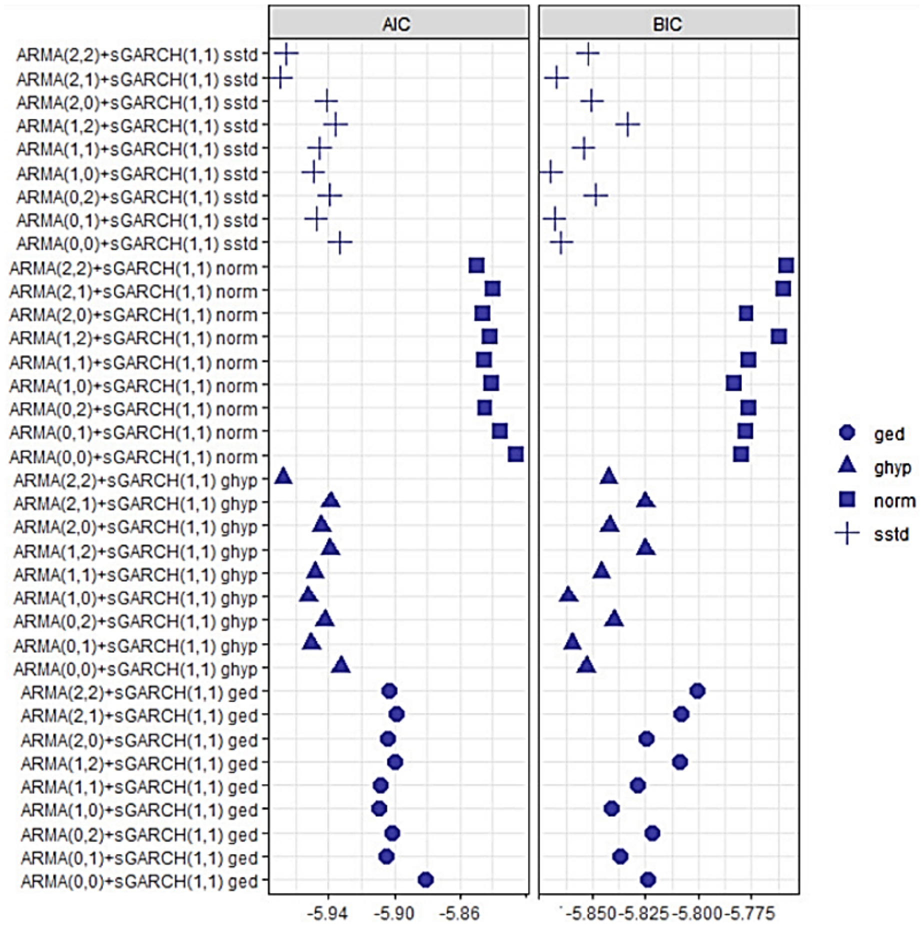

Fig. 5. Comparison of information criteria values for selected ARMA(p,q)-GARCH(1,1) models using different types of distributions

Note: ARMA(3,0)-GARCH(1,1) models have higher values for both information criteria in each case compared to the ARMA(2,0)-GARCH(1,1) models, while their coefficients do not gain significance, hence they were dropped in the later part of this section. Furthermore, models with a generalised hyperbolic distribution were also discarded from further analysis, as in both cases the number of statistically insignificant parameters was relatively too large.

Source: own elaboration.

Table 6. Values of ARMA(2,0)-GARCH(1,1) and ARMA(0,2)-GARCH(1,1) model coefficients using normal distribution, Student's skewed t-distribution and generalised error distribution

	ARMA(2,0)-GARCH(1,1)			ARMA(0,2)-GARCH(1,1)		
	norm	sstd	ged	norm	sstd	ged
coefficient	coefficient value					
c	**0.001489 (±0.000523)**	**0.001312 (±0.000493)**	**0.001873 (±0.000364)**	**0.001461 (±0.000536)**	**0.001293 (±0.000496)**	**0.001876 (±0.000473)**
δ_1	**-0.133016 (±0.062857)**	**-0.151144 (±0.058069)**	**-0.190370 (±0.045580)**	x x	x x	x x
δ_2	**0.129070 (±0.055491)**	0.038525 (±0.053268)	0.075754 (±0.043965)	x x	x x	x x
η_1	x x	x x	X x	-0.119330 (±0.062577)	**-0.145117 (±0.058431)**	**-0.177130 (±0.059011)**
η_2	x x	x x	x x	**0.151450 (±0.054586)**	0.045706 (±0.062291)	**0.089197 (±0.037460)**
α_0	**0.000012 (±0.0000003)**	0.000009 (±0.0000007)	0.000011 (±0.0000009)	**0.000012 (±0.0000003)**	0.000009 (±0.0000007)	0.000011 (±0.0000008)
α_1	**0.335486 (±0.061584)**	**0.260102 (±0.083764)**	**0.305899 (±0.045499)**	**0.338080 (±0.063051)**	**0.260734 (±0.084750)**	**0.307485 (±0.057297)**

	ARMA(2,0)-GARCH(1,1)			ARMA(0,2)-GARCH(1,1)		
	norm	sstd	ged	norm	sstd	ged
coefficient	coefficient value					
β_1	**0.654432 (±0.039606)**	**0.728569 (±0.072414)**	**0.680818 (±0.033669)**	**0.651451 (±0.040575)**	**0.727960 (±0.073572)**	**0.678796 (±0.022611)**
skew	x x	**0.771665 (±0.068597)**	x x	x x	**0.768136 (±0.069952)**	x x
Shape	x x	**5.536307 (±2.487446)**	**1.258264 (±0.032009)**	x x	**5.516976 (±2.537111)**	**1.261202 (±0.024676)**

Note: The values of significant (at the 0.05 significance level) coefficients are in bold. The standard error of the estimate is given in brackets.

Source: own elaboration.

Note that the typical properties for the conditional variance equation occurred here. All the coefficients were positive (a necessary condition to ensure the positivity of the conditional variance), and the sum of coefficients α_1 and β_1 was close to 1. Finally, to determine whether the conditional mean equation was correctly specified, it was necessary to examine whether the series of standardised residuals from the ARMA-GARCH model were white noise. In turn, a similar property for the squares of the standardised residuals meant that the conditional variance equation was correctly specified. In the rugarch package available in R software, both issues were tested by the so-called weighted Ljung-Box test with statistic $\tilde{Q}_W(m)$:

$$\tilde{Q}_W(m) = T(T+2)\sum_{i=1}^{m} \frac{m-(i-1)}{m} * \frac{\hat{\varrho}_i^2}{T-i},$$

where fraction $\frac{m-i+1}{m} = \frac{m-(i-1)}{m}$ was the weight (Fisher and Gallagher, 2012). One can see that, according to this definition, smaller lags are given more weight, since for a fixed m the weight is a decreasing function of variable i. The null hypothesis in the test accordingly assumed that there was no autocorrelation in the standardised residuals or their squares, indicating the white noise of the time series. The test results are presented in Table 7.

Table 7. Ljung-Box weighted test results for standardised residuals and their squares for ARMA(2,0)-GARCH(1,1) and ARMA(0,2)-GARCH(1,1) models using the normal distribution, Student's skewed t-distribution and generalised error distribution

model	lags	weighted test LB for residuals		weighted test LB for squares	
		$\tilde{Q}_W(m)$	p-value	$\tilde{Q}_W(m)$	p-value
ARMA(2,0)-GARCH(1,1) norm	1	0.227	0.63	0.136	0.71
	5	0.581	1.00	0.354	0.98
	9	4.495	0.57	0.801	0.99
ARMA(2,0)-GARCH(1,1) sstd	1	0.712	0.40	0.000	1.00
	5	1.783	0.99	0.331	0.98
	9	5.555	0.34	0.858	0.99
ARMA(2,0)-GARCH(1,1) ged	1	1.654	0.20	0.023	0.88
	5	2.116	0.93	0.214	0.99
	9	5.937	0.27	0.739	0.99
ARMA(0,2)-GARCH(1,1) norm	1	0.091	0.76	0.180	0.67
	5	0.459	1.00	0.348	0.98
	9	4.507	0.57	0.777	0.99
ARMA(0,2)-GARCH(1,1) sstd	1	0.561	0.45	0.004	0.95
	5	2.039	0.95	0.339	0.98
	9	5.920	0.27	0.866	0.99
ARMA(0,2)-GARCH(1,1) ged	1	1.230	0.27	0.056	0.81
	5	2.002	0.96	0.233	0.99
	9	5.997	0.26	0.755	0.99

Source: own elaboration.

3. Evaluation of Forecast Accuracy

3.1. Forecast of Returns

On the basis of the obtained model parameter estimates, the author generated daily forecasts of logarithmic returns over the period under consideration, and then, comparing them with the actual values of the time series r_t, assessed the accuracy of the forecasts within the sample. Here, classical measures of forecast error were applied, such as: mean absolute error MAE, root mean square error RMSE and average relative percentage error MAPE.

As one can see in Table 8, the smallest MAE forecast errors were generated by the ARMA(2,0)-GARCH(1,1) model with a generalised error distribution. At the same time, it is the second largest model in terms of RMSE error, second only to the forecasts of the ARMA(0,2)-GARCH(1,1) model with the normal distribution used. The MAPE relative measure, on the other hand, indicated the ARMA(0,2)-GARCH(1,1) model with a skewed Student's t-distribution, therefore it was difficult to make a clear assessment of which models performed better in predicting logarithmic returns during such a turbulent period. However, regardless of whether ARMA was actually an autoregressive or moving average model, the largest mean absolute and mean squared errors fell on models with a skewed Student's t-distribution, yet the models with such a distribution were also by far the best in terms of the MAPE measure. At this point, it must be also acknowledged that the mean relative percentage error was very high in each case (however, the poor resilience of this measure to near-zero values of the observed quantity must be kept in mind).

Table 8. Values of MAE, RMSE and MAPE of return forecast errors

model	MAE		RMSE		MAPE	
	score	ranking	score	ranking	score	ranking
ARMA(2,0)-GARCH(1,1) norm	0.011370	3	0.018403	3	133.6%	5
ARMA(2,0)-GARCH(1,1) sstd	0.011414	5	0.018592	5	121.4%	2
ARMA(1,1)-GARCH(1,1) ged	**0.011299**	1	0.018388	2	135.9%	6
ARMA(0,2)-GARCH(1,1) norm	0.011397	4	**0.018385**	1	131.7%	3
ARMA(0,2)-GARCH(1,1) sstd	0.011449	6	0.018665	6	**119.9%**	0.011449
ARMA(0,2)-GARCH(1,1) ged	0.011350	2	0.018463	4	133.4%	4

Note: The best models within each measure are in bold.

Source: own elaboration.

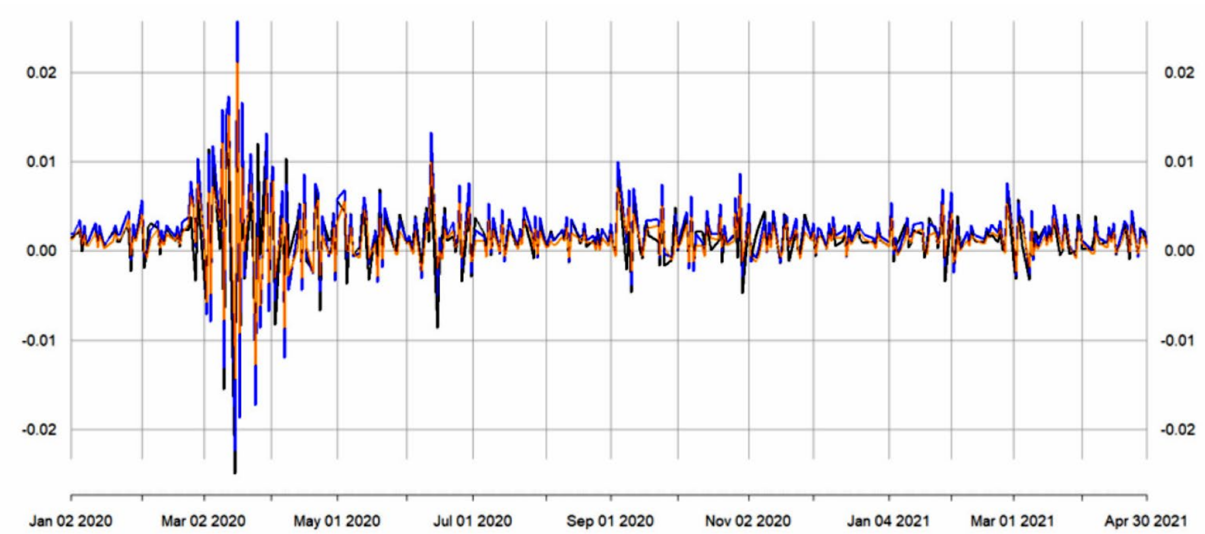

Fig. 6. Forecasts of logarithmic rates of return, January 2020-April 2021. The following models were used: ARMA(2,0)-GARCH(1,1) ged (blue colour), ARMA(0,2)-GARCH(1,1) norm (black colour), and ARMA(0,2)-GARCH(1,1) sstd (orange colour)

Source: own elaboration.

Figure 6 shows the forecasts of logarithmic returns of the S&P500 for the best models according to the measures considered. It is clear that the least fluctuation was by the model marked in orange, i.e. ARMA(0,2)-GARCH(1,1) with a skewed Student's t-distribution. On the other hand, the ARMA(2,0)-GARCH(1,1) model with a generalised error distribution was the one most likely to reach local maxima. This may explain the model's relatively good performance in both the MAE and RMSE classifications due to the fact that the actual logarithmic returns quoted during the period were much larger, in absolute value, than those obtained in the forecast models (see Figure 7).

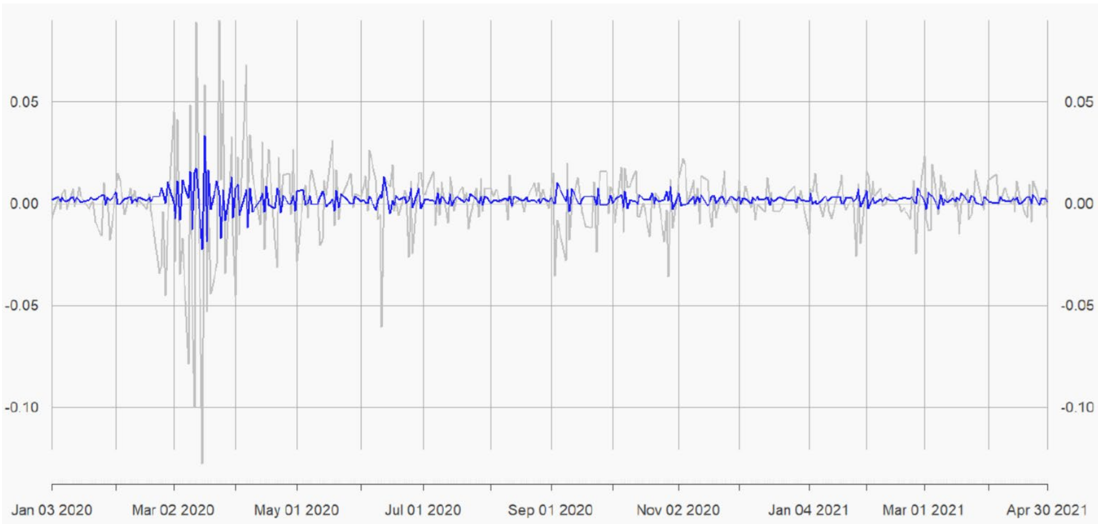

Fig. 7. Logarithmic rates of return versus their forecast, January 2020-April 2021. The grey colour shows the curve illustrating the course of logarithmic returns, while the blue colour is used to indicate the forecasts obtained from the ARMA(2,0)-GARCH(1,1) ged model

Source: own elaboration.

3.2. Forecasts of Conditional Variances

Let us now turn to the conditional variance forecasts. In this paper, the square of the logarithmic returns (more specifically, the arithmetic mean of the squares of the logarithmic returns) was taken as the empirical value of the observed variance, as practiced in numerous academic studies (e.g. Skoczylas, 2013). The MAPE measure was omitted due to its low robustness to the observed values close to 0, using instead the QLIKE measure. The choice of the RMSE and QLIKE measures was not random, as it had been found that only these functions indicated the optimal forecast as the correct one regarding any estimator of the variance (Patton, 2011). Moreover, they are also homogeneous, meaning that scaling the data (forecasts/observed values) does not change the ranking of the models. Table 9 shows the ratings and rankings of forecast errors over the full pandemic period.

Table 9. Values of MAE, RMSE and QLIKE measures of conditional variance forecast errors, using return squares as observed volatility approximation, January 2020-April 2021

model	MAE		RMSE		MAPE	
	score	ranking	score	ranking	score	ranking
ARMA(2,0)-GARCH(1,1) norm	0.0003494	2	**0.001040**	1	-7.709	2
ARMA(2,0)-GARCH(1,1) sstd	0.000356	5	0.001071	5	-7.697	5
ARMA(1,1)-GARCH(1,1) ged	**0.0003485**	1	0.001048	3	**-7.712**	1
ARMA(0,2)-GARCH(1,1) norm	0.000351	3	0.001044	2	-7.706	4
ARMA(0,2)-GARCH(1,1) sstd	0.001044	6	0.001076	6	-7.694	6
ARMA(0,2)-GARCH(1,1) ged	0.000352	4	0.001057	4	-7.708	3

Note: The best models within each measure are in bold.

Source: own elaboration.

The ARMA(2,0)-GARCH(1,1) model with a generalised error distribution took the lowest MAE and QLIKE error values, whereas the ARMA(2,0)-GARCH(1,1) model with a normal distribution should be considered the best in terms of the conditional variance forecast error measure. The largest errors occurred for the models with a skewed Student's t-distribution, as in the case of logarithmic returns. The above best models generated very similar conditional variance forecasts (see Figure 8).

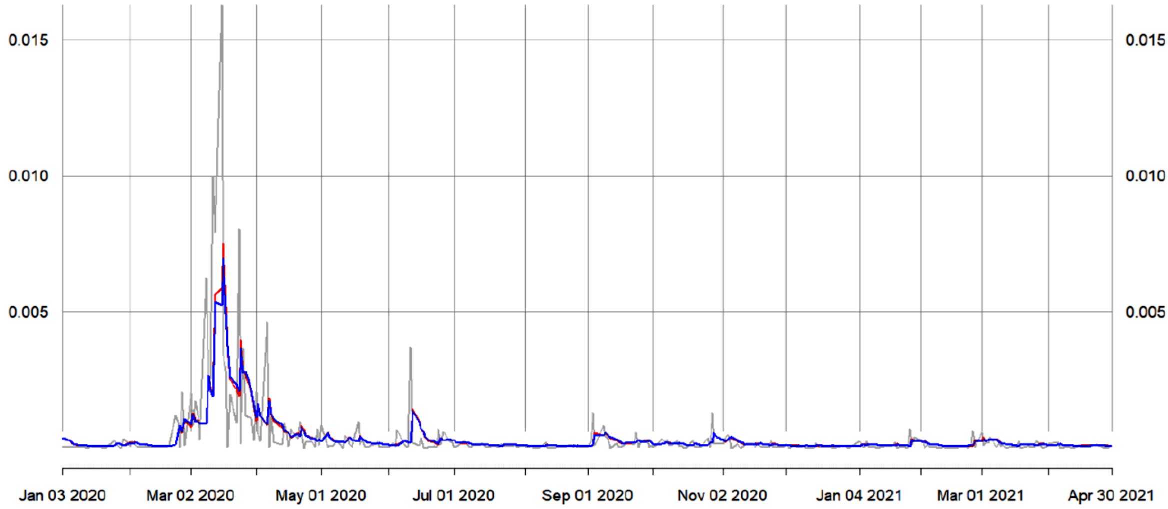

Fig. 8. Variance estimator in the form of logarithmic squares of returns versus conditional variance forecasts, January 2020-April 2021. The grey colour is used for the curve illustrating the logarithmic squares of returns, and forecasts based on the ARMA(2,0)-GARCH(1,1) norm (red) and ARMA(2,0)-GARCH(1,1) ged (blue)

Source: own elaboration.

4. Conclusions

In this paper the author studied ARMA-GARCH models with normal, skewed Student’s t, generalised error and generalised hyperbolic distribution. In practice, financial data often exhibit characteristics such as asymmetry, leptokurticity or the presence of outliers, which are not well described by the standard normal distribution. By choosing different error distributions it was possible to better reflect these characteristics, leading to more realistic models, and as a result the best models were identified for the pandemic period under consideration. The following representation was obtained: ARMA(2,0)-GARCH(1,1) and ARMA(0,2)-GARCH(1,1), both with three types of distributions – normal, skewed Student's t and generalised error distribution. The assessment of forecast accuracy showed that it was difficult to identify the model that generated the lowest logarithmic return forecast errors, while in the case of conditional variance forecasts, the ARMA(2,0)-GARCH(1,1) models with a normal distribution and a generalised error distribution were the best. The largest errors of conditional variance forecasts were generated by models with a skewed Student's t-distribution, although such models achieved the best results in terms of MAPE for the forecast of returns.

References

Alles, L., & Kling, J. (1994). Regularities in the Variation of Skewness in Asset Returns. *Journal of Financial Research*, 17, 427-438.

Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., & Viratyosin, T. (2020). The Unprecedented Stock Market Reaction to COVID-19. *The Review of Asset Pricing Studies, 10*(4), 742-758.

Bollerslev, T., Engle, R. F., & Nelson, D. (1994). ARCH Models. *Handbook of Econometrics*, *4*, 2963-2967.

Chaudhary, R., Bakhshi, P., & Gupta, H. (2020). Volatility in International Stock Markets: An Empirical Study during COVID-19. *Journal of Risk and Financial Management*, *13*(9), 208. https://doi.org/10.3390/jrfm13090208

Czech, K., Wielechowski, M., Kotyza, P., Benešová, I., & Laputková A. (2020). Shaking Stability: COVID-19 Impact on the Visegrad Group Countries' Financial Markets. *Sustainability, 12*(15), 6282. https://doi.org/10.3390/su12156282

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427-431.

Doman, M., & Doman, R. (2009). *Modelowanie zmienności i ryzyka: metody ekonometrii finansowej.* Wolters Kluwer.

Duttilo, P., Gattone, S. A., & Di Battista, T. (2021). Volatility Modeling: An Overview of Equity Markets in the Euro Area during COVID-19 Pandemic. *Mathematics*, *9*(11), 1212. https://doi.org/10.3390/math9111212

Fama, E. (1965). The Behaviour of Stock Market Prices. *Journal of Business*, *38*, 34-105.

Fisher, T. J., & Gallagher, C. M. (2012). New Weighted Portmanteau Statistics for Time Series Goodness of Fit Testing. *Journal of the American Statistical Association*, *107*(498), 777-787. https://doi.org/10.1080/01621459.2012.688465

Fiszeder, P. (2009). *Modele klasy GARCH w empirycznych badaniach finansowych*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Gherghina, Ș. C., Armeanu, D. Ș., & Joldeș, C. C. (2021). COVID-19 Pandemic and Romanian Stock Market Volatility: A GARCH Approach. *Journal of Risk and Financial Management*, *14*(8), 341. https://doi.org/10.3390/jrfm14080341

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH (1,1)? *Journal of applied econometrics*, *20*(7), 873-889. https://doi.org/10.1002/jae.800

Hassani, H., & Yeganegi, M. R. (2020). Selecting optimal lag order in Ljung–Box test. *Physica A: Statistical Mechanics and its Applications*, *541*, 123700. https://doi.org/10.1016/j.physa.2019.123700

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis.* Prentice Hall.

Kon, S. (1984). Models of Stock Returns – A Comparison. *Journal of Finance*, *39*, 147-165.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1), 159-178. https://doi.org/10.1016/0304-4076(92)90104-Y

Liu, Z., Huynh, T. L. D., & Dai, P.-F. (2021). The impact of COVID-19 on the stock market crash risk in China. *Research in International Business and Finance, 57*, 101419. https://doi.org/10.1016/j.ribaf.2021.101419

Ljung, G. M. (1986). Diagnostic Testing of Univariate Time Series Models. *Biometrika*, *73*(3), 725-730. https://doi.org/10.2307/2336539

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, *36*, 394-419.

McLeod, A. I., & Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of time series analysis*, *4*(4), 269-273. https://doi.org/10.1111/j.1467-9892.1983.tb00373.x

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246-256. https://doi.org/10.1016/j.jeconom.2010.03.034

Piontek, K. (2002). *Modelowanie i prognozowanie zmienności instrumentów finansowych*. Akademia Ekonomiczna we Wrocławiu (PhD thesis).

Schwert, G. W. (2002). Tests for Unit Roots: A Monte Carlo Investigation. *Journal of Business & Economic Statistics*, *20*(1), 5-17. http://www.jstor.org/stable/1392146

Shumway, R. H., & Stoffer, S. D. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer International Publishing, Cham, Springer Texts in Statistics. http://link.springer.com/10.1007/978-3-319-52452-8

Skoczylas, T. (2013). Modelowanie i prognozowanie zmienności przy użyciu modeli opartych o zakres wahań. *Ekonomia. Rynek, Gospodarka, Społeczeństwo*, *35*, 65-81.

Tsay, R. S. (2005). *Analysis of financial time series*. John Wiley & Sons.

Zmienność w czasach zarazy, czyli o zastosowaniu modeli ARMA-GARCH w modelowaniu i prognozowaniu zmienności stóp zwrotu z indeksu S&P500 w dobie pandemii COVID-19

Streszczenie

Cel: W artykule przeanalizowano szereg czasowy cen zamknięcia indeksu S&P500 w okresie od stycznia 2020 r. do kwietnia 2021 r. Dokonano wyboru najlepszych modeli ARMA(p,q)-GARCH(1,1) z różnymi postaciami funkcji gęstości prawdopodobieństwa. Porównano błędy generowanych prognoz zarówno w zakresie logarytmicznych stóp zwrotu, jak i warunkowych wariancji.

Metodyka: W badaniu zastosowano procedurę Boxa–Jenkinsa. Uwzględniając kryteria informacyjne, rozpatrzono najlepsze spośród modeli z rozkładem normalnym, skośnym t-Studenta, uogólnionym rozkładem błędu i uogólnionym rozkładem hiperbolicznym.

Wyniki: Otrzymano następującą reprezentację: ARMA(2,0)-GARCH(1,1) i ARMA(0,2)-GARCH(1,1), z rozkładami: normalnym, skośnym t-Studenta i uogólnionym rozkładem błędu. Ocena trafności prognoz wykazała, że w przypadku prognoz warunkowych wariancji najlepsze były modele ARMA(2,0)-GARCH(1,1) z rozkładem normalnym i uogólnionym rozkładem błędu. Największe błędy prognoz warunkowych wariancji generowały modele ze skośnym rozkładem t-Studenta.

Implikacje i rekomendacje: Warto rozważyć rozszerzenie zawartego w pracy badania na przypadek modeli opartych o zakres wahań (takich jak Range GARCH-RGARCH lub Conditional Autoregressive Range Model-CARR).

Oryginalność/wartość: W pracy rozważono modele z różnorodnymi funkcjami gęstości prawdopodobieństwa, pokazując, że taka różnorodność miała znaczenie przy szukaniu najlepszych modeli w czasach dużej zmienności.

Słowa kluczowe: COVID-19, ARMA-GARCH, szereg czasowy, S&P500 index, stacjonarność