



Time series analysis of the number of Covid-19 deaths in Iraq

Sarab D. Shukur^{a,*}, Tasnim Hasan Kadhim^a

^aDepartment of Mathematics, College of Science, University of Baghdad, Iraq

(Communicated by Madjid Eshaghi Gordji)

Abstract

In this paper, the time series data for the number of deaths from Coronavirus (COVID-19) patients in Iraq were analyzed for the period from 4/3/2020 to 18/2/2021. ARCH, GARCH, and TGARCH models were applied due to the changing volatility of the series leading to a heteroscedastic variance. The appropriate models for the series were diagnosed and the best model was chosen and used for forecasting by the exponential smoothing methods. The comparison criterion used was the Root Mean Squared Error and the Sum of Squared Residuals. The most appropriate model for modeling and forecasting the Coronavirus deaths series in Iraq was diagnosed as TGARCH (1,1). Finally, the method of Holt-winter-additive forecasting was the best method among the exponential smoothing methods.

Keywords: COVID-19, Time Series, volatility, heteroscedastic variance, GARCH, ARCH, TGARCH.

1. Introduction

The coronavirus disease (COVID-19) was first appeared in Wuhan, China, at the end of December 2019. To better manage the spread of COVID-19 and deaths, it is critical to monitor and forecast their spread. Time series models are crucial for predicting the COVID-19's impact. In Iraq, the coronavirus pandemic spread started from February 24, 2020, in the city of Najaf. Then the disease spread along all over the country within a short period of time and has never stopped yet. The total number of confirmed cases in Iraq reached 657,453 cases, including 13,220 deaths as of February 18, 2021, according to the official statistics of the Ministry of Health in Iraq.

*Corresponding author

Email addresses: sarab.shukur1203@sc.uobaghdad.edu.iq (Sarab D. Shukur),
tasnim.h@sc.uobaghdad.edu.iq (Tasnim Hasan Kadhim)

Received: March 2021 *Accepted:* July 2021

Time series is a statistical method used in analyzing data that is based on predicting the values of the phenomenon in the future based on previous data series. Some researchers focus on time series topics because of their importance in studying the behavior of different phenomena across specific time periods by analyzing and interpreting them and building a model to predict them. Time series topics include extremely wide fields, medical, environmental, economic, financial, . . . etc. Time series analysis aims to obtain an accurate description of the features of the phenomenon that results in the time series, building a model to explain the behavior of that phenomenon, and predicting future observations of the studied phenomenon based on what happens in the past. The model for the time series is built on several stages, the first stage of model building is the model identification, the second stage is the estimation in which the parameters of the model are estimated according to the common methods of estimation, then examining the accuracy and suitability of the model through the third stage the diagnostic checking using statistical tests, and finally, the forecasting stage.

There are many methods used in estimating parameters in time-series models, including the Least Square method, denoted by the symbol (OLS) when the distribution of errors is unknown, and the Maximum Likelihood Method, denoted by the symbol (MLE). The resulting estimators of these methods must have the same characteristics as the good estimators in the normal case or close to them, and they may be efficient, consistent, and appropriate in the case of the availability of time series conditions. In our study, we will use the least squares method.

The problem of volatility in time series was addressed as Autoregressive Conditional Heteroscedastic (ARCH) models by (Engle) for the first time in 1982. It is an autoregressive time series conditioned by the non-stationarity of the homogeneity of error variances (volatility) with autocorrelation in this series and is symbolized by the symbol ARCH(p). Engle (1982)[6] carried out an analysis different from the traditional analyses, by studying the variance of error conditioned by the information of the past, which is variable with time and the models are better matched to those data (non-stationary in the variance). These models are called Generalized Autoregressive Conditional Heteroscedastic (GARCH Models). GARCH models are often used for financial and economic time series. In this paper, we will use them in a completely different topic, which is the spread of a disease over time. Vasilis Papastefanopoulos, Pantelis Linardatos, and Sotiris Kotsiantis (2020) presented a study on Coronavirus pandemic; the objective of the study was to examine time-series approaches for forecasting the percentage of active COVID-19 cases in the general population. They also employed development and estimation for six different methodologies, including ARIMA , the Holt–Winters additive model (HWAAS) , TBAT, Facebook’s Prophet , DeepAR, and N-Beats, are described in detail[10]. Another study made by Zeynep Ceylan to investigate the COVID-19 prevalence in Italy, Spain, and France, where ARIMA (0,2,1), ARIMA (1,2,0), and ARIMA (0,2,1) are chosen as the best models for Italy, Spain, and France respectively[3]

2. Data and Methodology

The data was taken from the World Health Organization website represent the daily deaths of the Coronavirus disease (COVID-19) pandemic in Iraq for the period from 4/3/2020 to 18/2/2021 and consisting of 352 observations,
[<https://data.humdata.org/m/dataset/novel-coronavirus-2019-ncov-cases>]

3. Methodology

A time series is a sequence of data points, typically consisting of successive measurements or observations on the quantifiable variable(s), made over a time interval, these observations are dependent

and organized according to time and taken at regular intervals (days, months, years, . . . etc.), but the sampling could also be irregular [12, 5]. Time series can be classified to two types: stationary and non-stationary time series. The word stationary refers to the absence of growth in the data meaning that the data fluctuate around a constant level without any increasing or decreasing trend [13] The time series consists of two variables: the first one is an independent variable and the other is the dependent variable according to the phenomenon studied. It can be expressed mathematically as [9]

$$Y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-n}) + e_t$$

Where y_{t-1} is the value of Y for the previous observation, y_{t-2} is the value of two previous observations ago, . . . etc., and e_t represents noise that does not follow a predictable pattern (this is called a random shock). Values of variables occurring prior to the current observation are called lag values [5]

3.1. Autoregressive Conditional Heteroscedasticity (ARCH) Model

The Autoregressive Conditional Heteroscedastic ARCH(p) model was proposed first by Robert Engle in 1982. The ARCH model can be described as a time series with conditional mean and conditional variance. The conditional average of the time series $\mu(t)$ is constant. As for the conditional variance of the time series, it is in the form of a model containing an error term and an equation of instability. The equations for the ARCH model can be described as follows [6, 7]

$$y_t = \mu + x_t, \quad x_t = \sigma_t \varepsilon_t$$

ε_t independent and identically distributed (*iid*) $N(0,1)$

$$\sigma_t^2 = \Omega + \alpha_1 x_{t-1}^2 + \alpha_2 x_{t-2}^2 + \dots + \alpha_p x_{t-p}^2$$

This equation is called the equation of volatility and non-stationarity, which can be reformulated as follows [11]

$$\sigma_t^2 = \Omega + \sum_{j=1}^p \alpha_j x_{t-j}^2$$

where

μ : The conditional average.

σ_t The conditional standard deviation

ε_t Random error.

$\Omega > 0, \alpha_j \geq 0, j = 1, 2, \dots, p$ And Ω, α_j , represent the model parameters.

And when $p = 1$, we have an ARCH (1) model of the first order and it will become σ_t^2 , shown by the following equation:

$$\sigma_t^2 = \Omega + \alpha_1 x_{t-1}^2$$

The process is in steady state if and only if the sum of the autoregressive parameters is positive and less than one [7, 2].

3.2. Generalized Autoregressive Conditional Heteroscedastic GARCH (p,q) Model

Bollerslev (1986) developed the ARCH model and proposed a more general model which he called Generalized Autoregressive Conditional Heteroscedastic model, the conditional autoregressive model can be defined by the non-homogeneity of the generalized variance of degree $p \geq 1$ and $q \geq 1$, which is denoted by the symbol GARCH according to the following equations [7, 8, 4]:

$$y_t = \mu + x_t^2, \quad x_t = \sigma_t \varepsilon_t, \text{ where } \varepsilon_t \text{ is } iidN(0,1)$$

$$\sigma_t^2 = \Omega + \alpha_1 x_{t-1}^2 + \alpha_2 x_{t-2}^2 + \dots + \alpha_p x_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \beta_2 \sigma_{t-2}^2 + \dots + \beta_q \sigma_{t-q}^2$$

Since

- y_t : The series of returns, which is a stationary and uncorrelated series
- μ : The average of the series
- ε_t : Is a series of independent and symmetric distribution errors and follows the standard normal distribution with mean of 0 and variance 1.

The equation of volatility, σ_t^2 can be rewritten as follows:

$$\sigma_t^2 = \Omega + \sum_{j=1}^p \alpha_j x_{t-j}^2 + \sum_{i=1}^q \beta_i \sigma_{t-i}^2$$

where $\Omega > 0, \alpha_j \geq 0, j = 1, 2, \dots, p, \beta_i \geq 0, i = 1, 2, \dots, q.$ Ω, α_j and β_i represent the parameters of the model.

when $p = 1, q = 1$, we have a GARCH (1,1) model of the first order.

$$\sigma_t^2 = \Omega + \alpha_1 x_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

And that what is meant by volatility is the fluctuations in the time series data, which leads to the non-stationarity in the variation of the time series over time, and that this change is called heteroscedastic variance and it occurs in time series with high frequencies. The values of the parameters α and β determines the short-term dynamics of time series fluctuations. If the sum of the coefficients is equal to one, then any shock will lead to a permanent change in all future values [7, 11].

3.3. Threshold Generalized Autoregressive Conditional Heteroscedastic Models (TGARCH)

The idea behind TGARCH models lies in the fact that it is better to capture negative shock movements as they have a greater effect on volatility than positive shocks. Therefore, to capture these movements, a model must be studied that determines the conditional standard deviation by referring to the previous lagging values. TGARCH allows obtaining functions with different fluctuations, depending on the signal and the value of the shock.

TGARCH models of degree (p,q), $p \geq 1$ and $q \geq 1$, can be defined as [7, 1]

$$y_t = \mu + x_t,$$

$$x_t = \sigma_t \varepsilon_t \quad \varepsilon_t \sim iid N(0,1)$$

$$\sigma_t^2 = \Omega + \sum_{i=1}^p (\alpha_i + \varpi_i d_{t-i}) x_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2$$

4. Application and Results

The Coronavirus deaths data series was chosen for its accuracy compared to recorded cases of infection and recovery cases because many of the infection cases are not recorded in the statistics since many of the people resort to treatment at home, and therefore it is linked to cases of recovery, so the statistics are less accurate compared to deaths. A time series plot of the numbers of deaths with Coronavirus (COVID-19) in Iraq were plotted for the period from 4/3/2020 to 18/2/2021 to determine the behavior of the series as shown in Figure (1).

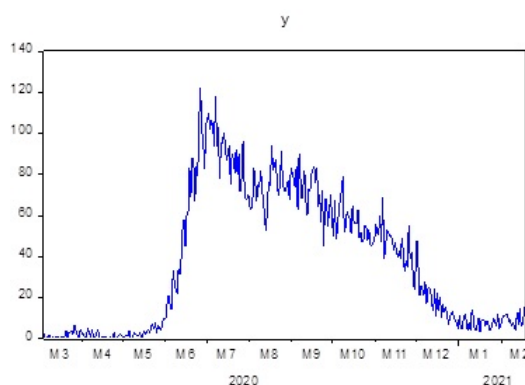


Figure 1: Coronavirus deaths data in Iraq for the period from 4/3/2020 to 18/2/2021

It can be seen from the figure the fluctuation of the data and the non-stationarity of the time series in the mean and variance because, it takes a general trend. For more accuracy, we draw the autocorrelation function (ACF) and the partial autocorrelation (PACF) respectively according to Figure (2).

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
1	1.000	0.966	0.966	331.06	0.000
2	0.966	0.960	0.402	658.96	0.000
3	0.931	0.950	0.138	981.35	0.000
4	0.896	0.945	0.109	1300.9	0.000
5	0.861	0.939	0.069	1617.3	0.000
6	0.826	0.931	0.014	1929.8	0.000
7	0.791	0.924	0.001	2238.4	0.000
8	0.756	0.910	-0.136	2538.4	0.000
9	0.721	0.900	-0.054	2832.9	0.000
10	0.686	0.889	-0.042	3121.0	0.000
11	0.651	0.878	-0.049	3402.5	0.000
12	0.616	0.866	-0.041	3677.4	0.000
13	0.581	0.853	-0.048	3944.7	0.000
14	0.546	0.834	-0.134	4201.4	0.000
15	0.511	0.820	-0.036	4450.2	0.000
16	0.476	0.806	-0.005	4691.4	0.000
17	0.441	0.791	-0.036	4923.9	0.000
18	0.406	0.773	-0.068	5146.7	0.000
19	0.371	0.764	0.125	5365.2	0.000
20	0.336	0.743	-0.073	5572.5	0.000
21	0.301	0.728	-0.001	5772.2	0.000
22	0.266	0.710	-0.038	5962.5	0.000
23	0.231	0.694	-0.009	6144.9	0.000
24	0.196	0.679	0.038	6320.0	0.000
25	0.161	0.662	-0.003	6487.0	0.000
26	0.126	0.647	0.001	6647.0	0.000
27	0.091	0.628	-0.018	6798.2	0.000
28	0.056	0.615	0.047	6943.5	0.000
29	0.021	0.597	-0.007	7081.2	0.000
30	-0.014	0.578	-0.077	7210.6	0.000
31	-0.029	0.567	0.083	7335.4	0.000
32	-0.044	0.549	-0.047	7452.5	0.000
33	-0.059	0.530	-0.047	7562.5	0.000

Figure 2: The ACF and PACF for the Original Series

We note from Figure (2) that the ACF coefficients are positive and are decreasing exponentially, and that displacement (1) is outside and all of them are outside the confidence limits. This is an indication of the lack of stationarity in the series.

To overcome the data no stationarity, we will take the first difference to the original series. The graph of the resulting series became as shown in Figure (3).

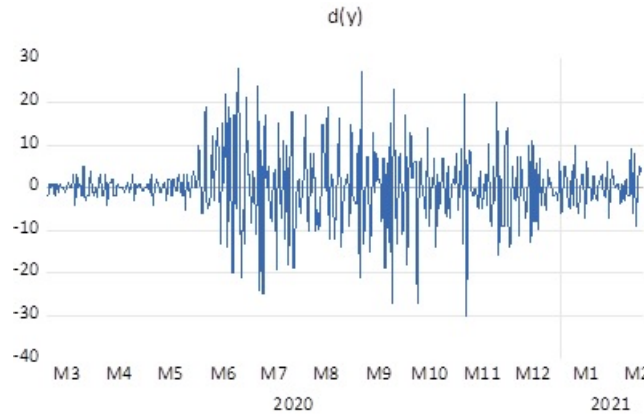


Figure 3: Coronavirus deaths series after taking the first difference

The expanded Augmented Dickey-Fuller test (ADF) was conducted to ascertain the stationarity of the series of Coronavirus (Covid-19) deaths in Iraq, without the constant, the trend and constant, and with secular trend and the constant, respectively and the test results were as shown in Figure 4.

1- With constant		
	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-30.15604	0.0000
Test critical values:		
1% level	-3.448835	
5% level	-2.869581	
10% level	-2.571122	

2- Without the constant		
	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-30.19816	0.0000
Test critical values:		
1% level	-2.571586	
5% level	-1.941732	
10% level	-1.616093	

3- With secular trend and the constant		
	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-30.17832	0.0000
Test critical values:		
1% level	-3.984496	
5% level	-3.422716	
10% level	-3.134249	

Figure 4: Dickey – Fuller test results

We tested the hypothesis $H_0 : \phi_1 = 0$. The series has a unit root versus $H_1 : \phi_1 \neq 0$ It has no unit root. we note that the absolute tabular values are greater than the values calculated under the specified P-value which is less than the level of significance (0.01) and it is located inside the unit circle. That is; the series is stationary in the mean, so that the null hypothesis that the series

is non-stationary is rejected and the alternative hypothesis that the series is stationary and has no unit root is accepted under the level of significance 1%.

Several models were suggested to represent the effect of ARCH. We will to choose the most efficient model according to the lowest value of the information criteria AIC, SIC, H-Q.

Table 1: Suggested GARCH Models of the Coronavirus Deaths Series

Model	AIC	SIC	H-Q	Notes
ARCH(1,0)	6.930414	6.97441	6.947924	
GARCH(1,1)	6.513456	6.579452	6.534722	
TGARCH(1,1)	6.499396	6.565393	6.525662	Best model

From Table (1) the it can be noted that the best model between the suggested GARCH models is the TGARCH (1, 1). The TGARCH(1, 1) model parameters was estimated by the least-squares method. The estimation results are shown in Table (2). The calculated z-values and the level of significance shows that most of the parameters of the model were significant when the significance level 5% which indicates the suitability of the model to the series data.

Table 2: TGARCH(1,1) Model Estimation Results

Coefficient	Std.Error	z-statistic	p-value	
constant	0.372809	0.205818	1.811359	0.0701
Ω	0.249275	0.104897	2.376374	0.0175
α	0.231600	0.036034	6.427295	0.0000
β_1	0.919989	0.018241	50.43567	0.0000
γ_2	-0.317612	0.057602	-5.513871	0.0000

The efficiency of the TGARCH (1, 1) model will be tested by the residual independence test and the residuals normal test. The hypothesis of independence test is H_0 : There is no serial correlation among residuals versus, H_1 : There is serial correlation among residuals. The autocorrelation and the partial autocorrelation functions of the residuals (errors) of the estimated model were extracted and plotted. We conclude from Figure (5) that the values of the autocorrelation coefficients and the values of the partial autocorrelation coefficients of the residuals are within the confidence limits (95%). This means that the null hypothesis is not rejected, that is, the absence of correlations, the values of each of them were greater than the level of significance (0.05) which indicates the independence of errors. This means that the series of residuals represent random variables, and that the estimated model is good, efficient, and suitable.

Fore the residuals normal test, we use this test to show whether the residuals are normally distributed by the hypothesis: H_0 : Residual are normally distributed versus H_1 : Residual are not normally distributed.

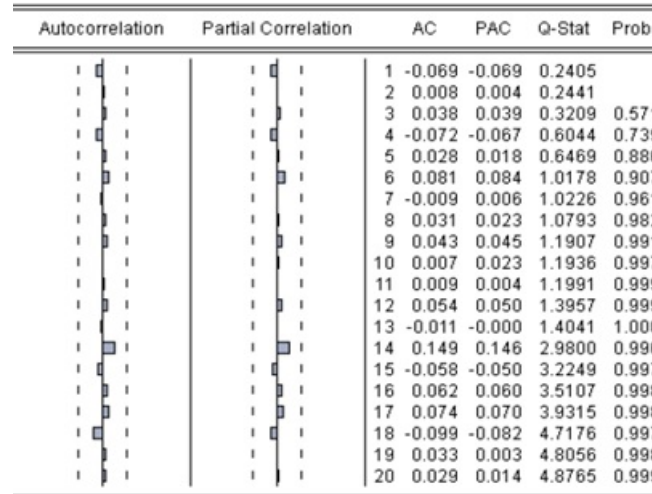


Figure 5: ACF and PACF functions for the residuals of the estimated model

It appears from the results of the test statistic (Jarque-Bero) from which it is shown that the test statistic reached (38.38830) with a probability value (0.0625690) and since the probabilistic value is greater than the level of significance (0.05), we must accept the null hypothesis and reject the alternative. If two of the residual tests are achieved, then we consider the results of the study acceptable, and this is the best result that can be reached.

Table 3: Summary Statistics of the Coronavirus deaths series

Indices	Values
Mean	0.067259
Median	-0.033769
Maximum	4.131822
Minimum	-2.278806
Std.dev	0.995520
Skewness	0.620093
Kurtosis	4.042486
Jarque-Bero	38.38830
Probability	0.0625690

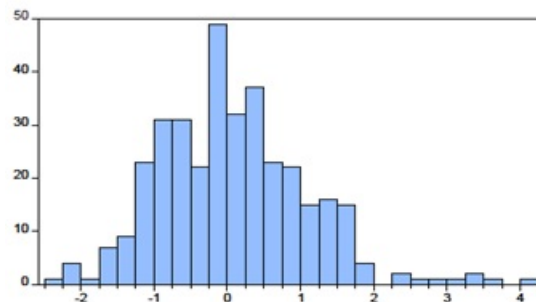


Figure 6: Normal distribution residuals test results

The final step of the time-series selected model is to predict the value of the studied phenomenon. We will carry out the forecasting of the estimated model TGARCH (1,1) inside the time series for the last six days from 13/2/2021 to 18/2/2021 using the exponential smoothing methods to choose the best method for forecasting, and then we predict it for next two weeks of the period for the Coronavirus deaths. After comparing the different measures RSS and RMSE, it turns out that the lowest value achieved by the fourth method, the Holt-winters-additive method among the exponential smoothing methods as shown in the table (4), and it was closer to the real results of the deaths of Coronavirus patients as shown in the table (5):

Table 4: RSS and RMSE measures for the exponential smoothing forecasting methods

Method forecasting	RSS	RMSE
Single Exponential	21257.03	7.771055
Double Exponential	19826.99	7.505110
Holt-winters-no seasonal	18974.34	7.341960
Holt-winters-additive	18664.83	7.281833

Table 5: Forecasting results for the Holt-winters-additive method

Date	real values	Forecasting values
13/2/2021	7	9
14/2/2021	15	13
15/2/2021	6	8
16/2/2021	7	7
17/2/2021	12	11
18/2/2021	16	15

The plot of the forecasting inside the series for (Y_1, Y_2, Y_3, Y_4) is shown in the figure (7)

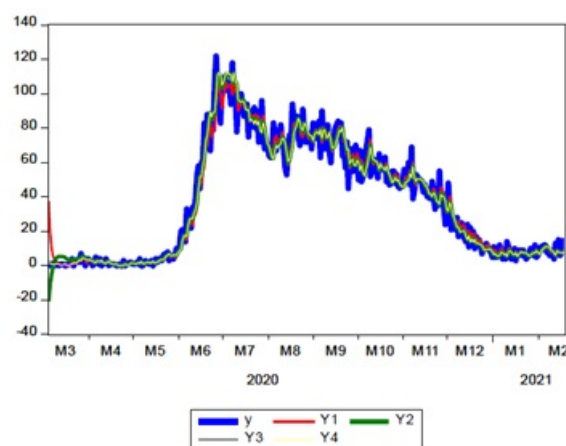


Figure 7: Forecasting results for series (Y_1, Y_2, Y_3, Y_4) .

Now we will carry out forecasting outside the series for two weeks from 19/2/2021 to 4/3/2021 by the Holt-winters-additive method that achieved the best results and compare them with the real deaths from Coronavirus in Iraq as shown in the tables (6) and the plot of Figure (8).

Table 6: Comparison of the real values and the values for forecasting outside the series.

Date	real values	Forecasting values for method Holt-winters-additive
19/2/2021	12	12
20/2/2021	13	13
21/2/2021	27	26
22/2/2021	23	24
23/2/2021	16	14
24/2/2021	13	13
25/2/2021	27	27
26/2/2021	14	15
27/2/2021	18	18
28/2/2021	23	23
1/3/2021	22	22
2/3/2021	30	32
3/3/2021	25	25
4/3/2021	24	24

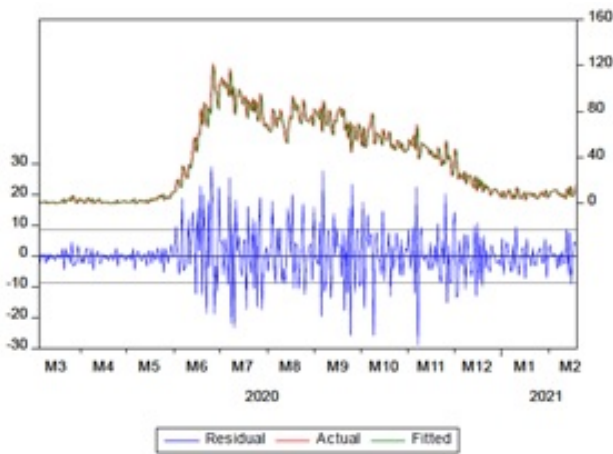


Figure 8: Forecasting outside the time series.

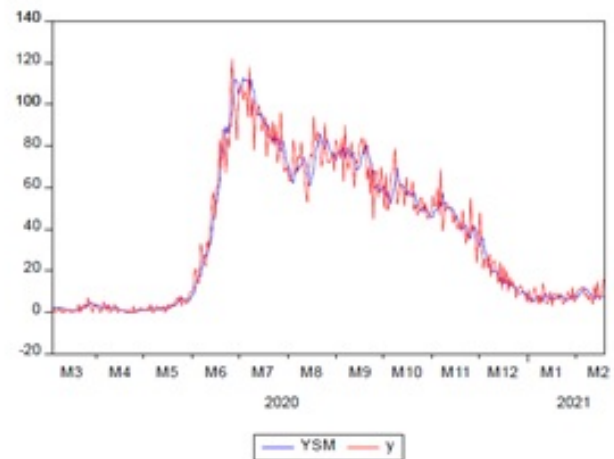


Figure 9: The original series with the forecasting series

Moreover, Figure (9) shows the original series with the forecasting or expected from the use of TGARCH (1, 1) model. By noting the predicted values, which are close to the original values the expected model matches the original model or series by 94%, and the model used is TGARCH (1, 1, 1) using the holt-winters-additive exponential smoothing method. The forecasted values of deaths from Coronavirus in Iraq were obtained accordingly.

5. Conclusions

Based on the results that have been reached, the series of deaths of Coronavirus patients witnessed wiggle and great volatility. The series **was heading** towards a general trend towards instability and after taking the first difference the series became stationary.

The developed Dickey Fuller test showed that the series was unstable in the mean. The presence of the ARCH effect advocated the use of GARCH models, which are known to represent the time series that contains the characteristic of Heteroscedasticity.

Several candidate models from the ARCH family were compared using information criteria AIC, SIC and H-Q. It was found that the most appropriate model for modeling and forecasting of the Coronavirus deaths in Iraq was the TGARCH (1,1) model. The residual independence test and the normally distribution test proved the absence of autocorrelation and the residuals are normally distributed which confirms the efficiency of the selected model and its ability to predict Coronavirus patients deaths.

The Holt-winter-additive method was the best method among the exponential smoothing forecasting methods, according to the statistical criteria's the RMSE and the RSS.

Finally, the graph of the original series and the forecasting series showed that there is a great convergence between them, whether in increasing or decreasing, and this confirms the validity of the presented statistical approach.

References

- [1] G. Ali, *EGARCH, GJR-GARCH, TGARCH, AVGARCH, NGARCH, IGARCH and APARCH models for pathogens at marine recreational sites*, J. Stat. Econ. Meth. 2 (2013) 57–73.
- [2] P. Catani and N. Ahlgren, *Combined Lagrange multiplier test for ARCH in vector autoregressive models*, Econ. Stat. 1 (2017) 62–84.
- [3] Z. Ceylan, *Estimation of COVID-19 prevalence in Italy, Spain, and France*, Sci. Total Environ. 729 (2020) 1–8.
- [4] P. Cheteni, *Stock market volatility using GARCH models: Evidence from South Africa and China stock markets*, MPRA, 77355 (2016) 1–13.
- [5] J.H. Cochrane, *Time series for macroeconomics and finance*, Manuscript, University of Chicago, Chicago, 2005.
- [6] R. Engle, *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation*, J. Econ. 50 (1982) 987–1007.
- [7] C. Francq and J.M. Zakoian, *GARCH Models: Structure, Statistical Inference and Financial Applications*, John Wiley & Sons, 2019.
- [8] A. Grek, *Forecasting accuracy for ARCH models and GARCH (1, 1) family: Which model does best capture the volatility of the Swedish stock markets*, Örebro University, Advanced Level Thesis, 2014.
- [9] T. C. Mills, *Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting*, Academic Press, London, 2019.
- [10] V.P. Papastefanopoulos, Linardatos and S. Kotsiantis, *COVID-19: a comparison of time series methods to forecast percentage of active cases per population*, Appl. Sci. 10 (2020) 38–80.
- [11] E. Rossi, *Lecture Notes on GARCH Models*, University of Pavia, 2004.
- [12] R.H. Shumway, D. S. Stoffer and D.S. Stoffer, *Time Series Analysis and its Applications*, Springer, London, 2015.
- [13] W.W. Wei, *Time Series Analysis*, Oxford Handbook of Quantitative Methods in Psychology, 2013.