HYBRIDIZATION MODEL FOR CAPTURING LONG MEMORY AND VOLATILITY OF BRENT CRUDE OIL PRICE DATA

REMAL SHAHER HUSSIEN AL-GOUNMEEIN

UNIVERSITI SAINS MALAYSIA

2022

HYBRIDIZATION MODEL FOR CAPTURING LONG MEMORY AND VOLATILITY OF BRENT CRUDE OIL PRICE DATA

by

REMAL SHAHER HUSSIEN AL-GOUNMEEIN

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

July 2022

ACKNOWLEDGEMENT

Praise to Allah, God of the World, for making it possible for me to study at Universiti Sains Malaysia. I would like to express my sincere gratitude to my supervisor, Associate Professor Mohd Tahir Ismail, for his unrestrained support and encouragement during my research and the writing of this thesis. I have benefited from his experience, knowledge, and expertise in the field of statistics and publication. Many thanks to him for supporting and encouraging me to publish all my researches in prestigious international journals.

I would also like to take this opportunity to thank the Dean of the School of Mathematical Sciences, Universiti Sains Malaysia, Professor Hailiza Kamarulhaili, the lecturers, and the staff of the department for their support, which has helped me to complete my Ph.D. thesis. I would like to thank all the postgraduate students at Universiti Sains Malaysia who encouraged me during my studies, particularly the students at the School of Mathematical Sciences.

I would like to express my sincere appreciation to my mother, Amneh Ahmad Alemami, for encouraging and supporting me throughout the study. I would also like to thank her for the sacrifices she has made and the care she has continually granted. She has taught me that even the most daunting task can be perfectly accomplished if it is done step by step. Therefore, I would like to dedicate this work to her.

Lastly, I thank my friends and classmates who embodied the spirit of cooperation and were not stingy with advice. I thank everyone who provided me with help, advice, and encouragement, as well as those who wished me success. Many thanks again to my friends and work colleagues in Al-Hussein Bin Talal University-Ma'an-Jordan in my home country who have continually encouraged me to complete this research. I am grateful to everybody for their helpful and constructive comments in the writing of this thesis.

Remal Shaher Al-Gounmeein

2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	Autocorrelation function
ADF	Augmented Dickey-Fuller test
AHW	Additive Holt-Winters
AI	Artificial Intelligence
AIC	Akaike information criterion
AICc	Corrected Akaike information criterion
ANN	Artificial neural network
APARCH	Asymmetric power autoregressive conditional heteroscedasticity
AQL	Asymptotic quasi-likelihood
ARFIMA	Autoregressive fractionally integrated moving average
ARIMA	Autoregressive integrated moving average
AVGARCH	Absolute value generalized autoregressive conditional heteroscedasticity
BIC	Schwarz Bayesian information criterion
Cov	Autocovariance function
csGARCH	Component standard generalized autoregressive conditional heteroscedasticity
EGARCH	Exponential generalized autoregressive conditional heteroscedasticity
FFMNNs	Feed-forward multilayer neural networks
fGARCH	Functional generalized autoregressive conditional heteroscedasticity
Fracdiff	Fractionally-Differenced
GARCH	Generalized autoregressive conditional heteroscedasticity
GJRGARCH	Glosten, Jagannathan, and Runkle generalized autoregressive conditional heteroscedasticity

GPH’s	Geweke and Porter–Hudak’s Estimator
HQ	Hannan-Quinn
HW	Holt-Winters
IGARCH	Integrated generalized autoregressive conditional heteroscedasticity
KF	Kalman filter
LM	Lagrange Multiplier test
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MASE	Mean absolute scaled error
ME	Mean error
MHW	Multiplicative Holt-Winters
MLP	Multilayer Perceptron
MPE	Mean percentage error
NAGARCH	Nonlinear asymmetric generalized autoregressive conditional heteroscedasticity
NGARCH	Nonlinear generalized autoregressive conditional heteroscedasticity
PACF	Partial autocorrelation function
PP	Phillips–Perron test
QLR	Quandt likelihood ratio test
RMSE	Root mean square error
R/S	Range over the standard deviation
SARIMA	Seasonal autoregressive integrated moving average
sGARCH	Standard generalized autoregressive conditional heteroscedasticity
Sperio	Smoothed Periodogram

SS	State space
TGARCH	Threshold generalized autoregressive conditional heteroscedasticity
Var	Variance

LIST OF SYMBOLS

$AR(p)$	Autoregression for the order p
B	Backward shift operators
b_t	Slope (trend) of the series at time t
d	Fractional difference
∇	Nonseasonal difference operators
ϵ_t	White noise process at time t
ε_t	Random error values at time t
$f(.)$	Activation function of the hidden layer
f_{t+m}	Forecast values for m periods ahead at time t
γ_t	State equation at time t
H_a	Alternative hypothesis
H_0	Null hypothesis
iid	Independently and identically distributed
l	Maximum likelihood for the model
L_t	Level of the series at time t
$MA(q)$	Moving average of the order q
μ	Population mean
n	Sample size
$\phi_p(B)$	Polynomial of the autoregression for the order p at the backward shift operators B
$\rho(h)$	Autocorrelation function at lag(h)
S_{t-1}^{-}	Sign bias variable (or dummy variable)
$S_{t-1}^{-}\,\hat{\varepsilon}_{t-1}^{2}$	The negative size bias variable

$S_{t-1}^{+}\ \hat{\varepsilon}_{t-1}^{2}$	The positive size bias variable
s_t	Seasonal component of the series at time t
σ_t^2	Conditional variation at time t
σ	Standard deviation
t	Continuous time
$\theta_q(B)$	Polynomial for the moving average of the order q at the backward shift operators B
χ^2	Chi-square statistic
x_t	Original time series data at time t
$\bar{X}_t$	Sample mean of x series at time t
Y_t	The growth rate of x_t series at time t
$Z_{t(di)}$	Time series after calculating the fractional difference d_i at time t

MODEL HIBRID BAGI MERAKAM MEMORI JANGKA PANJANG DAN KEMERUAPAN DATA HARGA MINYAK MENTAH BRENT

ABSTRAK

Indeks harga minyak mentah Brent biasanya bertingkah laku tidak linear, tidak pegun, dan tidak normal dengan ingatan panjang dan heteroskedastisiti yang tinggi; oleh itu, menghurai dan mengawal perubahannya adalah sukar. Seterusnya, fenomena ini melemahkan kesahan dan ketepatan hasil kaedah ramalan. Oleh itu, kajian ini memfokuskan pada kaedah hibridisasi untuk menangkap tingkah laku ingatan panjang dan heteroskedastisiti dalam set data dan meningkatkan ketepatan ramalan harga minyak mentah Brent. Kebelakangan ini, kaedah hibridisasi untuk model autoregresi purata bergerak pecahan bersepadu (ARFIMA) telah diperkenalkan sebagai teknik yang berkesan untuk mengatasi tingkah laku tidak linear, tidak pegun, dan tidak normal dengan heteroskedastisiti tinggi dalam set data siri masa. Kaedah hibridisasi ARFIMA menunjukkan beberapa ciri yang tidak dimiliki oleh kaedah tradisional lain. Maka tesis ini mencadangkan tiga model baharu dan 12 teknik yang berbeza berdasarkan gabungan dan hibridisasi model ARFIMA dengan teknik ramalan tradisional untuk meramalkan harga minyak mentah Brent. Ketiga-tiga model baharu tersebut adalah ARFIMA dengan autoregresi bersyarat heteroskedastisiti kuasa asimetri (ARFIMA-APARCH), ARFIMA dengan autoregresi bersyarat heteroskedastisiti teritlak Glosten, Jagannathan dan Runkle (ARFIMA-GJRGARCH) dan ARFIMA dengan GARCH komponen terpiawai (ARFIMA-csGARCH) telah dicadangkan. Cadangan ini bertujuan untuk mendapatkan hasil ramalan yang lebih baik dan menyelesaikan masalah ketidaktepatan ramalan dalam siri harga minyak. Di

samping itu, dengan menggunakan set data yang sama, 15 teknik yang berbeza ini dibandingkan dengan dua kaedah ramalan individu tradisional. Perbandingan ini didasarkan pada satu fungsi pengukuran ralat (iaitu, ralat purata punca kuasa dua [RMSE]) dan satu ujian linear (iaitu, ujian Ljung-Box) untuk menilai hasil kaedah ramalan yang dicadangkan. Dapatan menunjukkan bahawa menggunakan kaedah pemodelan yang berbeza dalam satu model hibrid berdasarkan model ARFIMA mengintegrasikan kekuatan model masing-masing dan menghasilkan model hibrid yang lebih tepat dan efisien untuk meramalkan harga minyak mentah Brent. Kajian ini menunjukkan bahawa model ARFIMA simetri (2,0.3589,2) -IGARCH (1,1) di bawah taburan normal dapat digunakan untuk memodelkan dan meramalkan turun naik harga minyak mentah Brent dalam jangka pendek. Tambahan pula, model hibrid ini adalah yang terbaik di antara 15 model yang dicadangkan berdasarkan nilai RMSE, ARCH-LM, dan ujian reja Ljung-Box untuk masalah pemodelan taburan reja tidak normal. Oleh itu tiga kaedah yang telah dicadangkan adalah sumbangan penting dalam literatur kajian tidak linear, tidak pegun, kelakuan tidak normal dengan ingatan panjang dan heteroskedastisiti tinggi dalam siri harga minyak mentah Brent.

HYBRIDIZATION MODEL FOR CAPTURING LONG MEMORY AND VOLATILITY OF BRENT CRUDE OIL PRICE DATA

ABSTRACT

The Brent crude oil price indices are typically nonlinear, nonstationary, and non-normal behavior with a long memory and high heteroscedasticity; hence, capturing the controlling properties of their changes is difficult. Subsequently, these phenomena weaken the validity and the accuracy of the result of the forecasting methods. Therefore, this study focuses on the hybridization method to capture long memory behavior and heteroscedasticity in the dataset and improve Brent crude oil price forecasting accuracy. Recently, the hybridization method for the autoregressive fractionally integrated moving average (ARFIMA) model has been introduced as an effective technique for overcoming the nonlinear, nonstationary, and non-normal behavior with high heteroscedasticity in a time series dataset. ARFIMA hybridization method presents several characteristics that other traditional methods do not have. Thus, this thesis proposed three new models and employed 12 different techniques based on combining and hybridizing the ARFIMA model with traditional forecasting techniques to forecast the Brent crude oil price. The three new models, namely, ARFIMA with the asymmetric power autoregressive conditional heteroscedasticity (ARFIMA-APARCH), ARFIMA with the Glosten, Jagannathan, and Runkle generalized autoregressive conditional heteroscedasticity (ARFIMA-GJRGARCH), and ARFIMA with the component standard GARCH (ARFIMA-csGARCH) are proposed. This proposal aims to obtain improved forecasting results and solve the forecasting inaccuracy problem in oil price series. Moreover, using the same dataset,

these 15 different techniques are compared with two traditional individual forecasting methods. This comparison is based on one error measurement function (namely, root-mean-square error [RMSE]) and one linear test (namely, Ljung-Box test) to evaluate the results of those proposed and employed forecasting methods. The results demonstrate that using different modeling methods in one hybrid model based on the ARFIMA model integrates the strength of the individual models and produces a more accurate and efficient hybrid model for forecasting the Brent crude oil price. This study indicates that the symmetric ARFIMA(2,0.3589,2)-IGARCH(1,1) model under normal distribution can be used to model and forecast Brent crude oil price volatility in the short term. Furthermore, this hybrid model is the best one among all the 15 considered models based on the RMSE value, the ARCH-LM, and the Ljung-Box tests of the residuals for modeling the problem of non-normal residual distribution. Therefore, those three proposed techniques are the main contribution to the literature of studying the nonlinear, nonstationary, and non-normal behavior with a long memory and high heteroscedasticity in the Brent crude oil price data set.

CHAPTER 1
INTRODUCTION

1.1 General Introduction

A time series is a collection of numerical observations arranged in regular order. These observations are associated with a particular instant or interval of time (i.e., taken at either discrete or continuous times) whose observations change over time (Bloomfield, 2000). The analysis of time series data is currently an essential topic in various research fields, such as economics, agricultural economics, econometrics, business, psychology, engineering, and social sciences, especially in forecasting. Time series forecasting is vital in developing and extending a model and describing the primary relationship of a dataset to study its future movement. Although modeling is a useful and important approach when the general sequence and pattern of the dataset are unknown, it cannot describe the current and future patterns. Nevertheless, in recent decades, numerous attempts have been made to develop and improve time series forecasting models in numerous fields and describe data through illustrative, satisfactory, and accurate mathematical rules by describing current and future patterns.

The use of forecasts in the economic and financial fields is hugely significant at all levels, particularly at the national, regional, and international levels. This is because forecasts can help investors reduce financial risks and increase profits despite the global economic volatility, especially in crude oil prices. Consequently, oil prices and its volatility have remained an important study for researchers in economic trends because of their importance in managing risk and increasing investment in financial and industrial markets. Therefore, an untraditional and accurate statistical technique must be used to describe the changes in these prices in terms of increasing and decreasing trends.

Besides volatility, the long memory is another feature the getting attention from researchers of oil prices.

In recent year, researchers in finance and economic literature realized the importance of long memory in analyzing time series data (Mostafaei and Sakhabakhsh, 2012). In modeling long memory behavior for any time series, the operation can usually be performed accurately by relying on autoregressive fractionally integrated moving average (ARFIMA) models compared with ARIMA models. Thus, the ARFIMA is long memory models mostly used in time series research (Karia et al., 2013). Meanwhile, volatility is an important consideration for any time series. Volatility can exist in some time series, especially in crude oil prices (Lee and Huh, 2017). Therefore, studying volatility is necessary.

Different methods and approaches have been applied to improve modeling and forecasting. Furthermore, modeling and forecasting methods are being increasingly used by researchers. Common examples of these methods include the ARIMA, ARFIMA, the generalized autoregressive conditional heteroscedasticity (GARCH), artificial neural network (ANN), Kalman filter (KF), and Holt-Winters (HW) methods. The individual model is also a standard forecasting method and a common approach used in several previous studies. An additional strategy for obtaining accurate forecasts is using a hybrid method (i.e., based on more than one model) to obtain future data forecasts and overcome the disadvantages of individual models, such as those that manage non-normal residuals. These hybrid methods can help solve problems with linear and nonlinear structures.

Based on the above, ARIMA (p,d,q) models are a popular method in modeling time series data by assuming the differencing parameter (d) as an integer value. In the event that this model is extended, assuming the differential parameter (d) value has a fractional value between $-0.5 < d < 0.5$, this kind of model with long memory behavior can be classified as an ARFIMA (p,d,q) model. ARFIMA models are linear time series models, but they are unsuitable for time series containing nonlinear structures. The ARFIMA model has been used and implemented in different areas, such as in prices. This model is also used to fit time series data to understand data or forecast the future. The GARCH-type models are considered the most prominent tools to capture the mentioned changes (Arachchi, 2018) in terms of symmetric and asymmetric effects.

Meanwhile, ANNs have been applied to various disciplines, such as system identification and control, decision making, pattern recognition, medical diagnosis, finance, data mining, and visualization, among others (Chen et al., 2018). These models can model any time series regardless of the structure of the series, and they are known to yield good forecasting results. Another method, KF, can be used to obtain optimal and high-accuracy forecasts (Xu et al., 2017). The HW method is a simple, fast, and inexpensive procedure that is widely used in forecasting as it can cope with trends and seasonal variations (Chatfield, 1978; Gamberini et al., 2010; Tratar, 2013). All the above mention methods have contributed a lot to the forecasting field.

As such, this present study focuses on employing ARFIMA to aid in forecasting Brent crude oil prices by integrating ARFIMA with existing forecasting methods. Subsequently, 15 forecasting models are employed. The experimental results of the 15 models show that each has had its own strengths and weaknesses in terms of accuracy forecasting measure and test, as will be discussed in later sections. Furthermore, ARFIMA has attracted researchers attention, especially when hybridizing nonstationary and nonnormality time series in several fields.

1.2 Problem Statement

Time series forecasting methods have effectively solved most of the forecasting problems in various fields, especially in the financial and economic time series. However, four major problems related to Brent crude oil prices dataset currently exist, which include the following:

1. The first problem is that it assumes the normality of the dataset and the normal distribution for the residuals in the modeling phase. The studied forecasting methods assumed that a normal relationship exists between the time series observations and the residuals. Thus, the normality assumption in real-life time series data (such as Brent crude oil prices data) is not always right.
2. The second problem of time series dataset forecasting methods is that it assumes the stationary on the dataset. In another meaning, the existed forecasting methods assumed that the properties of the time series (i.e., mean, variance, and autocorrelation) do not depend on the time where the time series is observed. Therefore, the studied methods used the transformation between data to overcome this problem. Nonstationary problem was confirmed by Ismail and Awajan, 2017; Xu et al., 2017.
3. The third problem of the time series forecasting methods is that it assumes the linearity of the dataset. The studied forecasting methods assumed that a linear relationship exists between the time series observations. Thus, the linearity assumption in real time series data (such as Brent crude oil prices data) is not always right. This problem was also reported by Zhang, 2003;

Ebrahim et al., 2013; Montgomery et al., 2015; Ismail and Awajan, 2017; Tadesse and Dinka, 2017; Al-Gounmeein and Ismail, 2020.

4. The Brent crude oil prices dataset does not follow statistical time series assumptions, such as data stability. Moreover, the changes in financial and economic market conditions, such as the supply, demand, and industry environment, can cause significant volatilities that mainly affect financial market data and stock exchange. Subsequently, volatility is an essential consideration in any time series, especially that of oil prices, given that the modality of these data grows exponentially, nonstationary, and are volatile. Moreover, Brent crude oil prices expectations remain highly important for investors and researchers. They pose a challenging problem to them due to the unique characteristics of these prices and their remarkable impact on various economic global sectors, particularly in the current COVID-19 pandemic. The volatility problem was also notified by Ebrahim et al., 2013; Jibrin et al., 2015; Lee and Huh, 2017; Ambach and Ambach, 2018; Bukhari et al., 2020; Dhliwayo et al., 2020.

These problems motivated this study to develop three new forecasting hybridization models and compare 15 forecasting methods, including individual and hybrid methods, which are highly efficient. The 15 forecasting methods in this study will overcome the problems of existing forecasting models (i.e., the non normality, the non stationarity, the nonlinearity, and the heteroscedasticity in time series dataset).

1.3 Research Questions

Based on the problems previously discussed, the research questions of this thesis can be derived as follows:

1. What are the necessary conditions to ensure the effectiveness and accuracy of an individual or hybrid ARFIMA models when modeling real time series dataset, especially Brent crude oil prices dataset?
2. Can we find the best solutions and the best ways in spite of the existing problems?
3. Despite the numerous studies on the volatility of Brent crude oil prices, this question remains unanswered: which volatility GARCH-type model (symmetric or asymmetric type) is most appropriate for modeling the problem of non-normal residual distribution for the individual ARFIMA model?
4. What is the best methodology for ANN as an alternative to the GARCH and ARCH models when dealing with the hybridization approaches in the ARFIMA models?
5. Despite the numerous studies on Brent crude oil prices, the following question remains unanswered: which hybridization approach is the most appropriate for modeling the problem of non-normal residual distribution for the individual ARFIMA model?

To answer these questions, 15 prediction methods have been considered in this study.

1.4 Research Hypothesis

Based on the sections above, the following hypothesis has been formulated: the model with the best Akaike information criterion (AIC) value does not necessarily produce the most outstanding forecast model for the dataset. Thus, this study dealt with the three models that have the smallest value for this criterion and compared them with one another to test this hypothesis.

This hypothesis proposed in this section is unique, given the nature of the studies results obtained and the methods used, which relates to this thesis, as will be discussed in Chapter 5.

1.5 Research Objectives

To answer these research questions and the research hypothesis, the thesis is centered on the forecasting performance of nonlinear, nonstationary, and non-normal Brent crude oil prices data with high heteroscedasticity based on the ARFIMA method with the following objectives:

1. to verify the presence of long memory feature in the real dataset by using numerous different methods.
2. to determine the best appropriate hybridization approach for modeling the problem of non-normal residual distribution for the individual ARFIMA models and to subject the model to various statistical testings.
3. to identify all conditions that necessary to ensure the effectiveness and accuracy of an individual or hybrid ARFIMA models when modeling the real dataset.
4. to compare all models based on AIC value to test the effectiveness of this criterion.
5. to close the gap in the literature by solving the problems for the individual ARFIMA models (i.e., the nonnormality, the nonstationarity, the

nonlinearity, and the heteroscedasticity in the time series dataset) by using the hybrid technique.

1.6 Scope of the Study

This thesis focuses on solving the problem of forecasting accuracy in the nonlinear, nonstationary, and non-normal Brent crude oil prices series data with high heteroscedasticity by considering 15 forecasting models based on ARFIMA. It also focuses on using hybrid models to overcome the disadvantages of individual models, such as those that manage non-normal residuals. The monthly Brent crude oil prices series data are used in this study, as will be discussed in Section 4.1.

1.7 Significance of the Study

The forecasting accuracy of the Brent crude oil prices is important and must be studied by researchers, investors, economists, stock market regulators, decision makers worldwide. Brent crude oil price data are nonlinear, nonstationary, and non-normal behavior with high heteroscedasticity; hence, capturing the controlling properties of their changes is difficult. Subsequently, these phenomena weaken the forecasting execution of most time series forecasting methods. This study intends to address these problems by employing new three techniques and compared with 12 traditional forecasting methods on the basis of the combination and hybridization of ARFIMA, which is an efficient technique in dealing with nonlinear, nonstationary, and non-normal time series data, and several traditional models. The three new models combine ARFIMA with APARCH, GJRGARCH, and csGARCH, respectively. The remaining 12 models combine ARFIMA with sGARCH, fGARCH, EGARCH, TGARCH,

IGARCH, AVGARCH, NGARCH, NAGARCH, APARCH, GJRGARCH, csGARCH, MLP, KF, MHW, and AHW, respectively. This study proves that the forecasting of the Brent crude oil prices series by the 15 proposed and employed models have their own strengths and weaknesses based on accuracy forecasting measure and test. In addition, this thesis used better methods than the traditional methods in obtaining forecasting results with more accuracy, which will be discussed in Chapter 5. Thus, these results contribute to the body of knowledge on time series analysis in terms of forecasting economic and financial time series data.

1.8 Thesis Organization

The remaining structure of the thesis is organized as follows: Chapter 2 presents a review of related literature and the background of ARFIMA and introduces the long memory in literature with the forecasting methods that are used in this study. Chapter 3 presents the proposed and the employed forecasting methods (i.e., hybrid of ARFIMA-sGARCH, ARFIMA-fGARCH, ARFIMA-EGARCH, ARFIMA-TGARCH, ARFIMA-IGARCH, ARFIMA-AVGARCH, ARFIMA-NGARCH, ARFIMA-NAGARCH, ARFIMA-APARCH, ARFIMA-GJRGARCH, ARFIMA-csGARCH, ARFIMA-MLP, ARFIMA-KF, ARFIMA-MHW, and ARFIMA-AHW methods). The statistical information criteria, the root mean square error (RMSE), and time series testings are also presented in Chapter 3. Chapter 4 presents an analysis of the monthly Brent crude oil prices data used in this study. Chapter 5 discusses the results obtained from the application of the proposed methodologies. Finally, Chapter 6 concludes the research work by summarizing the findings and future work.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter contains fifteen sections. The second section presents a brief description of heteroscedasticity definition. The third section reports nonstationary time series properties. The fourth section describes the nonnormality in the time series and how to deal with it. The fifth section discusses the existence of nonlinear behavior in any time series. The sixth section describes and explains the various time series volatility techniques that are either used or compared with the forecasting methods employed in this thesis, which are GARCH-type models in terms of symmetric and asymmetric effects. The seventh section provides a brief theoretical explanation of several graphs and statistical tests that are used to verify the long memory feature in time series and some methods and functions that are used to estimate the long memory parameter. This section also presents the ARFIMA model. The eighth section discusses ANN, particularly MLP, as explained in the literature. The ninth section reviews a description of the KF technique and explains SS modeling. The tenth section describes two types of the HW method, namely, AHW and MHW, as presented in the literature. The eleventh section discusses the employed hybrid method. The twelfth section provides an overview of developments in the field of ARFIMA methodology and its application in different areas. The thirteenth section lists recent studies that applied ARFIMA to forecasting time series. An overview of oil price changes and their volatility is discussed in the fourteenth section. The last section summarizes this chapter.

2.2 Heteroscedasticity

In statistics, heteroscedasticity (also known as heteroskedasticity) occurs when the variance or any other statistical dispersion measure (deviation or variation) of a variable in the time series dataset is nonconstant. Conditional heteroskedasticity identifies the nonconstant volatility related to prior periods of volatility (e.g., daily or monthly), which means that any new values depend on others. Another definition for this concept is as follows: if heteroscedasticity exists, then the model residuals cannot be independent. Thus, a good model yields residuals similar to a white noise process, and it requires the following: mean zero, constant variance, and no autocorrelation (i.e., independent residuals). Meanwhile, the opposite of heteroscedasticity is homoscedasticity (also known as homoskedasticity). Homoscedasticity refers to a condition in which the residuals variance is constant, which is one of the assumptions of linear regression modeling. Moreover, heteroscedasticity can be tested using the Lagrange multiplier test (David et al., 2016), as will be discussed in Section 3.4.4.

2.3 Non Stationarity

A time series is a strictly stationary time series if its statistical properties are not affected by a change in time. It is often characterized as a constant probability distribution in time (Montgomery et al., 2015). Stationary time series implies a type of statistical stability in the data in terms of the mean, variance, and covariance, as reported by Montgomery et al. (2015) as follows:

i. The expected value of the time series x_t is fixed, finite, and does not depend on time t, which is defined as: $\mu_x = \mathrm{E}(x_t)$.

   Consequently, the time series has a constant mean.

ii. The variance of the time series x_t is fixed and finite and does not depend on time t, which is defined as follows: $\sigma_x^2 = Var(x) = E(x^2) - \mu^2$.

Consequently, the time series has constant variance.

iii. The autocovariance function defined as $Cov(x_t, x_{t+k})$ for any lag(k) is dependent only on a function of k (where k is the number of periods) and not dependent on time t, which is defined as follows: $Cov(x_t, x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)] = \gamma_k$.

The MA(q) process (Section 2.7.4) is always stationary regardless of the values of the weights. Based on the above, these features are the characteristics of the stationary time series. Otherwise, the time series is nonstationary, as will be discussed in Section 3.4.1.

2.4 Non Normality

In all fields of time series, it is necessary to apply statistical methods in a correct way. The most commonly used between them is the normal assumption. The statistical framework is based on the assumption of normality in the time series and if this assumption is violated, the inference breaks down (Das and Imon, 2016). In that case, outliers can be dealt with by transforming the dataset to correct the nonnormality problems (Field, 2009). So, it is essential to test this assumption before any statistical analysis of the dataset (Das and Imon, 2016). In statistics, the normality assumption can be checked by graphical and statistical tests (Öztuna et al., 2006), as will be discussed in Section 3.4.2.

2.5 Nonlinearity

The time series $x(t)$ is a nonlinear time series if it cannot be modeled by the linear model, and a nonlinear structure cannot be found in linear systems (Theiler et al., 1992; Zhang, 2003). Suppose the errors are normally distributed, as is commonly assumed, then a linear model results in a normally distributed process. In that case, meanwhile, the predictive distributions of nonlinear models are generally non-normal and often difficult (Cryer and Chan, 2008). Accordingly, the residuals of the models always provide information about the existence of linearity or not. Thus, several tests have been proposed to test the linearity characteristic when analyzing time series data in the literature, as will be discussed in Section 3.4.3.

2.6 Volatility

Interest in modeling the volatility of time series has remained high in recent years. Volatility is an essential consideration in any time series, as evidenced by various studies related to finance, economics, tourism, and other areas where data are widely scattered (Tendai and Chikobvu, 2017; Akter and Nobi, 2018). Apparent volatilities can appear in several time series types, particularly that of crude oil prices (Lee and Huh, 2017). Consequently, volatilities should be examined to avoid inaccuracies in developing plans and strategies, either for critical current decisions or future forecasts. Moreover, the impact of volatilities when forecasting should be determined to avoid financial risks that may cause losses to investors considering that the forecasting of financial time series data is only one of the challenging tasks caused by nonstationarity and nonlinearity (Ismail and Awajan, 2017). These phenomena are popular features of different data. Ramzan et al. (2012) reported that GARCH-type models had been confirmed successful in forecasting volatility in many cases. GARCH-type models are currently considered the most prominent tools for capturing the previously mentioned changes (Arachchi, 2018). Thus, our thesis intends to explain many GARCH-type models in terms of symmetric and asymmetric effects, as discussed in the next sections.

Bollerslev (1986) expanded the ARCH model with order q, which was previously developed by Engle (1982) as the GARCH model with order (p, q) (Francq and Zakoian, 2019). The first model was dependent on uncorrelated random error ε_t values, but the GARCH model relies on conditional variation σ_t^2 . Hence, the general form of the $GARCH(p, q)$ model was specified by Francq and Zakoian (2019) as follows:

$$\varepsilon_t = \eta_t \sigma_t \quad , \qquad with \quad \eta_t \overset{iid}{\sim} N(0,1) \tag{2.1}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \quad , \tag{2.2}$$

where $\omega > 0$, $\beta_j \geq 0$, and $\alpha_i \geq 0$ are constants; $i = 1,2, \dots q$; $j = 1,2, \dots p$; and $t \in \mathbb{Z}$. If $\beta_j = 0$, then Equation (2.2) is called $ARCH(q)$. If $p = q = 0$, then Equation (2.2) is a white noise. If the conditional variance of the process is unknown, then the asymptotic quasi-likelihood (AQL) method, which integrates the kernel procedure, is used to estimate the parameter of the GARCH model (Alzghool, 2017). Accordingly, some symmetric (i.e., univariate models) and asymmetric (i.e., nonlinear models) GARCH-type models are described. The specific extensions of the GARCH model are elucidated in the succeeding sections.

2.6.1 Symmetric Generalized Autoregressive Conditional Heteroscedasticity (GARCH)-type Models

2.6.1(a) Standard GARCH (sGARCH)

The conditional variance σ_t^2 at time t is expressed by the symmetric effects of the sGARCH model (Singh et al., 2016). The formula of this model is similar to that of Equation (2.2), in which ε_t is considered the residual returns of Equation (2.1), as mentioned earlier (Miah and Rahman, 2016).

2.6.1(b) Integrated GARCH (IGARCH)

The IGARCH model is a unit root GARCH model (Tsay, 2010). That is, the IGARCH (p, q) formula is given by $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j = 1$ (Francq and Zakoian, 2019). Singh et al. (2016) stated that the IGARCH model belongs to the family of symmetric GARCH models.

2.6.1(c) Component Standard GARCH (csGARCH)

Ding and Granger (1996) introduced the csGARCH model. Engle and Lee (1999) subsequently developed this model and concluded that volatilities exhibit long-term and short-term movements. The specific formula of the csGARCH (1,1) model given by Engle and Lee (1999) is as follows:

$$\sigma_t^2 = q_t + \alpha_1(\varepsilon_{t-1}^2 - q_{t-1}) + \beta_1(\sigma_{t-1}^2 - q_{t-1})\,, \tag{2.3}$$

where

$$q_t = \omega + \rho q_{t-1} + \phi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2)\,, \tag{2.4}$$

where $\omega > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, and $\phi \geq 0$. If $\rho < 1$ and $\alpha_1 + \beta_1 < 1$, then weak stationarity holds (Chu et al., 2017). The general formula of the csGARCH (p,q) model in accordance with Hemanth and Patil (2017) is presented as follows:

$$\sigma_t^2 = q_t + \sum_{i=1}^{q} \alpha_i \left(\varepsilon_{t-i}^2 - q_{t-i}\right) + \sum_{j=1}^{p} \beta_j \left(\sigma_{t-j}^2 - q_{t-j}\right), \tag{2.5}$$

where

$$q_t = \omega + \rho q_{t-1} + \phi\left(\varepsilon_{t-1}^2 - \sigma_{t-1}^2\right). \tag{2.6}$$

The aforementioned authors indicated that this model can also analyze conditional variance as two components: permanent component of volatility q_t and transitory component of volatility $(\sigma_{t-j}^2 - q_{t-j})$. In addition, Zhang et al. (2018) explained that the q_t component can measure systematic risk, where ρ is utilized as a coefficient to measure the persistence of the permanent component. The term $(\varepsilon_{t-1}^2 - \sigma_{t-1}^2)$ is the prediction error that renders the long-term volatility variant, and the term $(\sigma_{t-j}^2 - q_{t-j})$ measures non-systematic risk. These authors also mentioned that coefficients ϕ and α_i can measure the short-term external shock effects of the permanent and transitory components, respectively.

2.6.2 Asymmetric Generalized Autoregressive Conditional Heteroscedasticity (GARCH)-type Models

2.6.2(a) Functional GARCH (fGARCH)

Considering the urgent need to describe the high-frequency volatilities that abound in financial data, an appropriate rational description of the problem may be considered for the function (Francq and Zakoian, 2019). Hörmann et al. (2013) previously suggested the functional approach of the ARCH model. Then, Aue et al.

(2017) expanded the approach as discussed in Francq and Zakoian (2019). In particular, Aue et al. (2017) focused on the fGARCH (1,1) model and defined it as follows:

$$\sigma_i^2 = \delta + \alpha\varepsilon_{i-1}^2 + \beta\sigma_{i-1}^2 \quad , \tag{2.7}$$

where ε_t is a sequence of random functions that satisfy Equation (2.1), i.e., $\delta \geq 0$, $\alpha \geq 0$, $\beta \geq 0$, and $i \in \mathbb{Z}$. Furthermore, $t \in [0,1]$ and x are arbitrary elements of the Hilbert space $\mathcal{H} = L^2[0,1]$, and the integral operators α and β are defined by $(\alpha x)_t = \int_0^1 \alpha(t,s)x(s)ds$ and $(\beta x)_t = \int_0^1 \beta(t,s)x(s)ds$, respectively. In addition, the integral kernel functions $\alpha(t,s)$ and $\beta(t,s)$ are elements of $L^2[0,1]^2$. Researchers naturally question how the asymmetric fGARCH models are created and their relationship and respective differences with classic asymmetric GARCH-type models (Sun and Yu, 2019). As mentioned previously, the method depends on a daily division of data (Francq and Zakoian, 2019), with the possibility of using other time units (Aue et al., 2017), e.g., monthly.

2.6.2(b) Exponential GARCH (EGARCH)

Nelson (1991) proposed that the asymmetric EGARCH model can solve various defects, such as nonnegative constraints and leverage effects, caused by the ARCH and GARCH models (Arachchi, 2018). Black (1976) studied the leverage effect, which includes the asymmetric effect of past positive and negative values on the recent volatility (Francq and Zakoian, 2019). The description of the $EGARCH(p,q)$ model provided by Francq and Zakoian (2019) is as follows:

$$\varepsilon_t = \eta_t\sigma_t \, , \qquad with \qquad \eta_t \overset{iid}{\sim} N(0,1) \tag{2.8}$$

$$ln\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \, g(\eta_{t-i}) + \sum_{j=1}^{p} \beta_j \, ln\sigma_{t-j}^2 \quad , \tag{2.9}$$

where

$$g(\eta_{t-i}) = \theta\eta_{t-i} + \xi(|\eta_{t-i}| - E|\eta_{t-i}|)\ , \tag{2.10}$$

where $\omega, \alpha_i, \beta_j, \theta,$ and ξ are real numbers; and $\sigma_t > 0$.

2.6.2(c) Threshold GARCH (TGARCH)

Another volatility model that is typically applied to address the leverage effect is the TGARCH model, as proposed by Glosten et al. (Tsay, 2010). The TGARCH model focuses on the asymmetric effects of good or bad news (Arachchi, 2018). The formula of the TGARCH (p, q) model provided by Tsay (2010) is as follows:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q}(\alpha_i + \gamma_i N_{t-i})\varepsilon_{t-i}^2 + \sum_{j=1}^{p}\beta_j\,\sigma_{t-j}^2\ , \tag{2.11}$$

where

$$N_{t-i} = \begin{Bmatrix} 1 & if \quad \varepsilon_{t-i} < 0 \\ 0 & if \quad \varepsilon_{t-i} \geq 0 \end{Bmatrix}, \tag{2.12}$$

where $\omega > 0, \alpha_i \geq 0, \gamma_i \geq 0,$ and $\beta_j \geq 0$. In Equation (2.11), $\varepsilon_{t-i} \geq 0$ indicates that bad news contributes $\alpha_i\varepsilon_{t-i}^2$ to σ_t^2. By contrast, $\varepsilon_{t-i} < 0$ indicates that good news entails considerably more substantial impact $(\alpha_i + \gamma_i)\varepsilon_{t-i}^2$ with $\gamma_i \geq 0$ (Tsay, 2010; Arachchi, 2018).

2.6.2(d) Absolute Value GARCH (AVGARCH)

Taylor (1986) and Schwert (1989) introduced the AVGARCH model, as reported by Francq and Zakoian (2019). This asymmetric model describes the conditional standard deviation as a linear combination of the absolute value of the shock and the lagged conditional standard deviation (Hentschel, 1995). The formula of the asymmetric AVGARCH (p,q) model is given by Zhang et al. (2017) as follows:

$$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}[|\varepsilon_{t-i} - b| - c(\varepsilon_{t-i} - b)] + \sum_{j=1}^{p} \beta_j \sigma_{t-j} \ , \tag{2.13}$$

where $\omega > 0, \alpha_i \geq 0, \beta_j \geq 0,$ and $|c| \leq 1.$ Moreover, parameters b and c are called the shift factor and rotation factor, respectively, which impose shift and rotation properties into the $\varepsilon_{t-i} - (|\varepsilon_{t-i} - b| - c(\varepsilon_{t-i} - b))$ space, as detailed in Zhang et al. (2017).

2.6.2(e) Nonlinear GARCH (NGARCH)

The NGARCH model (Higgins and Bera, 1992; Hsieh and Ritchken, 2005; Duan et al., 2006) is a symmetric-type model (Rezitis and Stavropoulos, 2010), and it exhibits the leverage effect. This model demonstrates the desirable feature of stock return (Emenogu et al., 2018). The formula of the NGARCH (p,q) model in accordance with Emenogu et al. (2018) is as follows:

$$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \, \varepsilon_{t-i}^2 + \sum_{i=1}^{q} \gamma_i \, \varepsilon_{t-i} + \sum_{j=1}^{p} \beta_j \, \sigma_{t-j} \, , \tag{2.14}$$

where all the parameters in Equation (2.14) satisfy the requirements of the GARCH model.

2.6.2(f) Nonlinear Asymmetric GARCH (NAGARCH)

The NAGARCH model was suggested by Engle and Ng (1993). The NAGARCH (p,q) model determines the shocks and volatilities of the past period. Its general equation provided by Rezitis and Stavropoulos (2010) is as follows:

$$\sigma_t = \omega + \sum_{i=1}^{q} \alpha_i \, (\varepsilon_{t-i} + \nu \sqrt{\sigma_{t-i}})^2 + \sum_{j=1}^{p} \beta_j \, \sigma_{t-j} \, . \tag{2.15}$$

The parameters in Equation (2.15) satisfy the aforementioned requirements. In this formula, ν is called the asymmetry parameter. If $\nu \neq 0$, then asymmetry occurs (Rezitis and Stavropoulos, 2010).

2.6.2(g) Asymmetric Power ARCH (APARCH)

The generalized version of the ARCH model as proposed by Ding et al. (1993) is known as the $\mathrm{APARCH(p, q)}$ model, and the following formula represents it:

$$\sigma_t^{\delta} = \omega + \sum_{i=1}^{q} \alpha_i \left(|\varepsilon_{t-i}| - \gamma_i\, \varepsilon_{t-i}\right)^{\delta} + \sum_{j=1}^{p} \beta_j\, \sigma_{t-j}^{\delta}\,, \tag{2.16}$$

where $\omega > 0, \alpha_i \geq 0, \beta_j \geq 0, \delta \geq 0,$ and $|\gamma_i| < 1.$ Hentschel (1995) proposed the family of GARCH models as a comprehensive model that combines a variety of other popular symmetric and asymmetric GARCH models, including the APARCH model. Moreover, the APARCH model also belongs to the family of asymmetric GARCH models (Singh et al., 2016). Notably, the apARCH model is a variance model in the rugarch package of R software. By contrast, the APARCH model is a submodel of the fGARCH model in the same package.

2.6.2(h) Glosten, Jagannathan, and Runkle GARCH (GJRGARCH)

Another extension of the GARCH model is the GJRGARCH model proposed by Glosten et al. (1993). This model exhibits an asymmetric effect (Engle and Ng, 1993). It also assumes a particular parametric form for conditional heteroscedasticity in a zero-mean white noise series (Emenogu et al., 2018). The $\mathrm{GJRGARCH\ (p, q)}$ model is given by the following general formula (Rezitis and Stavropoulos, 2010; Emenogu et al., 2018; Francq and Zakoian, 2019):

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i\, \varepsilon_{t-i}^2 + \sum_{i=1}^{q} \gamma_i\, S_{t-i}^{+}\, \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j\, \sigma_{t-j}^2 \ , \qquad (2.17)$$

where

$$S_{t-i}^{+} = \begin{Bmatrix} 1 & if \quad \varepsilon_{t-i} > 0 \\ 0 & if \quad \varepsilon_{t-i} \leq 0 \end{Bmatrix}, \qquad (2.18)$$

where $\omega > 0, \alpha_i \geq 0,\ \beta_j \geq 0,$ and $\sum \alpha_i + \sum \beta_j < 1$; while $|\gamma_i| \leq 1$ (Emenogu et al., 2018) is called an asymmetry parameter. If $\gamma_i \neq 0$, then asymmetry occurs; if $0 < \gamma_i \leq 1$, then a positive shock causes more volatility than a negative shock of the same size (Rezitis and Stavropoulos, 2010). Notably, the gjrGARCH model corresponds to the variance model in the rugarch package of R software. By contrast, the GJRGARCH model is a submodel of the fGARCH model in the same package.

2.7 Long Memory

Long memory is a phenomenon that can be observed in a time series; it manifests when the distance between two points is increased (Bahar et al., 2017), and it considerably affects the financial field in term of various transformations of stock index returns under the sub-categories; credit quantity aggregates, price indexes, stock prices, exchange rates and interest rates (Bhardwaj and Swanson, 2006). Experimental research on long memory processes dates back to Hurst (1951), who studied the hydrological properties of the Nile Basin. However, interest in using long memory models for economic data series was elicited when Granger and Joyeux (1980) observed that many such series are nonstationary in terms of the mean value.

Here, we explain the concept of long memory in a time series as presented by Palma (2007) and Hassler (2019). Suppose that $\rho(h)$ is the autocovariance function at lag(h) of a stationary process $y_t : t \in \mathbb{Z}$. Then, y_t exhibits long memory if the autocorrelation sequence decays extremely slowly such that it is not absolutely summable, as follows:

$$\sum_{h=-\infty}^{\infty} |\rho(h)| = \infty \, , \tag{2.19}$$

where $$\rho(h) = \mathrm{E}(y_t, y_{t+h}) \, . \tag{2.20}$$

Otherwise, y_t exhibits short memory if the following formula is verified:

$$\sum_{h=-\infty}^{\infty} |\rho(h)| < \infty \, . \tag{2.21}$$

However, several graphs and statistical tests are used to verify the long memory feature. These graphs and tests are as follows.

2.7.1 Verifying the Long memory Feature by Using Graphs

Many graphs provide an indication of the existence of the long memory feature. They include the autocorrelation function (ACF), range over the standard deviation (R/S), variance, variogram, spectral density function, and Higuchi plots.

2.7.1(a) Autocorrelation Function Plot

A long memory phenomenon can be specified when ACF decays more slowly than exponential decay (Bahar et al., 2017), as reported by Palma (2007), in accordance with the following formula:

$$\rho(h) = \frac{\Gamma(1-d)\,\Gamma(\mathrm{h+d})}{\Gamma(d)\,\Gamma(1+h-d)} \, . \tag{2.22}$$

It can also be written using another formula, as follows:

$$\rho(h) \sim \frac{\Gamma(1-d)}{\Gamma(d)}\ h^{2d-1}, \tag{2.23}$$

where $-0.5\ < d < 0.5$ and $h \to \infty$.

ACF shows the correlation between observations for different periods. In applied work, an ACF diagram is frequently used as a primary diagnostic tool when studying time series. It is considerably important in highlighting some of the important characteristics of a time series, particularly in verifying the presence of long memory in a time series. We select this type of graphics for the discussion and analysis chapters of this thesis because it is more commonly used.

2.7.1(b) R/S Plot

The R/S graph was described by Beran (1994) to have the following steps.

i. $\mathrm{Q} = \mathrm{R/S}$ is calculated for all possible values of time t and lag(k).
ii. $\mathrm{Log(Q)}$ versus log(k) is plotted.
iii. A straight line $\mathrm{y} = \mathrm{a} \pm \mathrm{b}$ log(k) that corresponds to the ultimate behavior of the data is drawn. Coefficients a and b can be estimated, e.g., via the least squares method.

The slope of this straight line is considered a measure for distinguishing between short and long memory processes. In particular, the slope of this straight line is greater than 0.5 for operations involving long memory and tends to be 0.5 for most short memory operations.

2.7.1(c) Variance Plot

The variance of the sample mean of a long memory process based on m observations was explained by Palma (2007) by calculating the following formula:

$$Var(\bar{y}_m) \sim c\ m^{2d-1}\ , \tag{2.24}$$

where $$Var(\bar{y}_m) = \frac{1}{m}\ [\ 2\sum_{j=1}^{m-1}\left(1-\frac{j}{m}\right)\ \gamma(j) + \gamma(0)\ ]\ , \tag{2.25}$$

and c is a positive constant. Consequently, by dividing a sample with size n into k blocks with size m, we obtain

$$\log\ (\ Var(\bar{y}_j)\ ) \sim c + (2\mathrm{d}-1)\log(\ \mathrm{j}\ ) \tag{2.26}$$

for $\mathrm{j} = 1, 2, \dots, \mathrm{k}$; and $(\bar{y}_j)$ is the mean of the jth block. That is,

$$\bar{y}_j = \frac{1}{m}\sum_{t=(j-1)\times m+1}^{j\times m} y_t\ . \tag{2.27}$$

Thus, for a long memory process, the slope of the line described by Equation (2.26) should be greater than −1. By contrast, the slope of the line should be −1 for a short memory process.

2.7.1(d) Variogram Plot

Journel and Huijbregts (1976) defined the variogram for the lag distance k formula as follows:

$$V(k) = \frac{1}{2}\ E[(X_t - X_{t-k})^2]\ , \tag{2.28}$$

where t denotes all possible locations.

Thus, the presence of long memory in data can be inferred through the behavior of the variogram in terms of the slow ascent and non-zigzagging of the plot in accordance