The Implications of the Size of Estimation Windows on The Accuracy of Volatility Forecasting in Foreign Exchange Markets

Master Thesis Financial Economics

Name: Connor Mason

Student Number: 511732

Supervisor: Dr. Sebastian Gryglewicz

Date of Final Version: 26/06/2019

Abstract: The effectiveness of alternative volatility forecasting models is a thoroughly explored area. This study explores a less saturated topic; the effectiveness of the size of the estimation window on the accuracy of one-day and 126-day forecasts in foreign exchange markets. This is done using a rolling window methodology and forecasting 126-days ahead of our in-sample period with GARCH(1,1) and ARIMA(1,1,1) models. We use exchange rate data of three currency pairs; USD/EURO, USD/GBP and EURO/GBP over the period 21/01/2009 to 29/03/2019. This paper finds that generally more data increases model forecast accuracy, although this statement is not universally found to be true.

Acknowledgment

I would like to make a dedication to those who have helped me this past year. Firstly, I would like to thank my supervisor, Dr. Sebastian Gryglewicz whose assistance and guidance has been invaluable to this piece of work. I would like to thank my parents, whose support the past year, and all my life has been essential to being where I am. Finally, I would like to thank the Gibney family who took me into their own home when I needed help. I hope that I can help someone in the future as much as they did for me.

Table of Contents

Table of Figures

1. Introduction

Volatility prediction is an essential element of risk management. Following subsequent financial crises, and the aftermath of the global financial crisis international and domestic regulatory requirements have become evermore pressing ensuring risk management is as critical as ever. Having effective procedures that accurately predict risk allows financial institutions to ensure they can meet internal and regulatory requirements. Volatility and variance are a critical component to the calculation of modern risk measures, such as Value-at-Risk. By accurately modelling measures such as VaR financial institutions can ensure not just that they are able to meet liquidity and operational requirements, but to ensure that they can remain solvent in periods of economic stress. The minimum liquidity requirements financial institutions are mandated to meet have increased in strain and complexity under the progressive Basel Accords (I, II, III) and so institutions must employ the most effective modelling techniques available to meet the demands of the modern financial world.

There has been much investment of resources in computational finance from academic and professional sources and with it there have been many advances in technology and methods. Calculations that once took days can now be done in minutes from standard computers allowing for much more sophisticated methods to be employed on a commercially viable basis. As a result, there has been a great deal of research into the effectiveness of alternative volatility models, and derivations of those models. There has however been less research into the implications of other parameters involved in the modelling process. And so, this study aims to shed more light on an area that has received less attention to date, namely, the implications of how much data to use in volatility forecasting; the size of the estimation window. Obtaining sufficient data can often be costly, or not viable; indeed, during this study we found ourselves limited on data availability when calculating realized volatility and so a proxy was required. Given that there are often real-world strains on data availability, it is important to know the implications of limited data obtainability. How large are the consequences of limited data? This study aims to examine the implications of the size of the estimation window on the accuracy of volatility forecasts using two relatively basic forecasting models; GARCH(1,1) and ARIMA(1,1,1). We model next day, and 126 days into the future to test the implications on the short and median term. It is expected that longer estimation windows will result in more accurate forecasts, with the implications being much larger for the 126-day forecasts and the one-day forecasts. It would also be expected that the implications of more data are more beneficial to GARCH than ARIMA.

The paper is structured as follows; after this introduction, the second section provides the background of the paper through an examination of the surrounding literature. An examination of the surrounding academia proves that there is a great deal of research on the effectiveness of differing volatility models, however, the size of estimations windows is comparatively neglected. The literature provides

techniques that we shall employ in this study; section three provides discussion of the theoretical background which provides justification of the internal workings of the models. This is then followed by section four which states the research hypothesises this piece of work will investigate. Next is section five and six which discusses the data and methodology employed within in this paper respectively, which provide insight into how we will address our hypotheses. We use GARCH(1,1) and ARIMA(1,1,1) models which are evaluated using RMSE and MAPE performance measures.

Following this, we have an examination of our results in section seven; this paper finds that generally more data improves the accuracy of volatility forecasts, however this is not a unanimous result and the benefits are not consistent across estimation window size, models or forecast horizons. The benefits of more data are greater for GARCH than ARIMA and somewhat more important for 126-day forecasts than 1-day forecasts. The scope of this paper is limited, leaving much room for subsequent research discussed in section 8. The field would benefit from an investigation over a larger range of forecasting horizons such as 10 and 20 days due to their common reference in regulatory requirements. An investigation with different iterations of models would also benefit the field. The paper is then finished with a conclusion found in section 9.

2. Literature review

Volatility prediction

The literature surrounding volatility prediction in financial markets is rich and has evolved greatly over the preceding decades. Academic and professional innovations and advances in computational power have allowed more sophisticated methods to be pioneered and adopted more widely, and so this field has continued to grow. Poon and Granger (2003) provide a summary of the literature on financial markets volatility forecasting and discuss how the field has changed over time. As they note, the issues found within financial data have become well known within the field, such as fat tail distributions of risky asset returns, volatility clustering, asymmetry and mean reversion. The former two of these issues, fat tails, and volatility clustering are closely related. Many models employed for volatility forecasting rely on the normal distribution whereas actual data is more clustered, often experiencing kurtosis, that is a higher probability of extreme returns than the normal distribution predicts. This phenomenon is well documented, such as the 1987 ‘Black Monday' and the 2007 global financial crisis when financial markets became extremely volatile. It is also common to observe skew, a difference in the third order moment of the distribution which occurs when data is not symmetrically distributed (Duffie and Pan, 1997). Although, the nature of skew is often dependent on the specific asset, or even instrument. The story of volatility prediction is that of evolving methodologies to more accurately model data.

GARCH

In 1982 as a means of modelling UK inflation Engle developed Autoregressive Conditional Heteroscedastic (ARCH) modelling. This innovation was then built upon further by Bollerslev (1986) extending to Generalised Autoregressive Conditional Heteroscedastic (GARCH). These models capture the long-run average variance which is considered heteroscedastic and they describe the probability distribution of returns conditionally. The GARCH model includes dependency on lags of past conditional variances providing greater weight to more recent data, which allows the incorporation of clustering found in financial data. The weights decrease as data moves back further in time; however, it never reaches a weight of zero. As noted by Alexander and Sheedy (2008) a drawback of GARCH is that the process requires a large amount of data to work effectively.

The success of GARCH has led to many iterations of the model, which raises the question, which GARCH is best? In an attempt to answer this question Hansen and Lunde (2005) performed an analysis of daily movements in the Deutsche Mark-Dollar exchange rate and IBM stock returns. Through the use of 330 GARCH type models, they do not find evidence to support the use of more elaborate GARCH models for volatility estimation in the Deutsche Mark-Dollar exchange rate. However, through testing out-of-sample results for IBM returns, Hansen and Lunde find that asymmetric models generate predictions that are significantly more accurate. This is likely due to their ability to cope with the leverage effect, a reason for Nelson (1991) developing EGARCH, a methodology that adjusts weights based on if returns are positive or negative.

ARIMA

ARIMA or Auto-Regressive Integrated Moving Average was first popularised by Box and Jenkins in their 1970 seminal textbook "Time Series Analysis: Forecasting and Control". The method has since become a staple of time series forecasting. The story of ARIMA begins in 1926 when Yule first introduced Auto-Regressive (AR) models, that is, that the output variable depends linearly on its own preceding values and on a stochastic term (an imperfectly expected term). Another key development occurred in 1937 when Slutsky (1937) introduced Moving Averages (MA) to consider the changes in trends over time. These two elements were combined in 1938 by Wold who created the ARMA model, demonstrating that the process can be used to model stationary time series data provided the appropriate order (p) and the number of MA (q) terms are specified. These models would then go on to be developed, by the inclusion of an ability for ARMA models to cope with stationarity by differencing (I) the data by Box and Jenkins thus creating ARIMA.

Due to its relative simplicity, ARIMA has become a very popular model in time-series data. Adebayo and Sivasamy (2014) conduct a study using ARIMA to forecast stock market returns. They find that the functional form of the appropriate ARIMA model is not consistent across stock markets in different countries.

Performance Testing

Performance testing is an essential aspect of any investigation into the accuracy of forecasting. There is a very broad selection of performance testing methods available, one of the most popular of which is RMSE, or root mean square error. RMSE is a performance measure that aggregates the magnitude of the errors, that is the difference between predicted and realized values, into a single measure of predictive power (Hyndman and Koehler, 2006). RMSE is the square root of the average squared errors which means that the effect of errors on the RMSE is proportional to the size of the squared error; thus, higher magnitude errors have a disproportionally large impact on the RMSE metric (Willmott and Matsuura, 2005).

Willmott and Matsuura (2006) raised an additional concern that RMSE is inferior to Mean Absolute Error (MAE) due to its inability to describe the average error. MAE is simply the average absolute difference between the predicted and realized value. Indeed, although RMSE is not complicated, interpretation of MAE is much more intuitive. Despite these concerns, Chai & Draxler (2014) find that the RMSE is a sufficient measure to compare model predictions hence why the measure is used in so many studies (Taylor, 1987; Adebayo and Sivasmy, 2014; Ganbold et al, 2017). Chen et al (2017) discuss the limitations of both RMSE and MAE, with a primary concern being scale dependence. With regards to this issue, the two-performance metrics are not able to consider the relative standing of their errors. Without this context, it is difficult to know what the size of the metrics really means. A solution to this problem is the use of percentage-based measures, such as Mean Absolute Percentage Error (MAPE). MAPE is very similar to MAE, but with the inclusive of scale context by converting MAE into a percentage of the magnitude of the error relative to the observed variable.

Within the literature, it is very common to use multiple performance testing criteria. A 2014 paper by Miswan et al compares the performance of ARIMA and GARCH models for forecasting the volatilities of the Malaysian Property and Equities markets from July 1997 to July 2012. Using Akaike's Information Criterion (AIC), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) as performance measures they conclude that GARCH models do not outperform ARIMA. A rationale they provide for these findings is that there is not sufficient kurtosis in the returns for GARCH to offer increased performance over ARIMA. Although not mentioned in their paper, the perhaps inappropriate time period used in their data set could have distorting effects; the study starts in July

1997, the exact time of the 1997 Asian financial crisis, in which Malaysia was involved. Adebayo and Sivasamy (2014) using a plethora of performance measures, i.e., AIC, BIC, HQC, RMSE, and MAE find that ARIMA (3, 1, 1) and ARIMA (1, 1, 4) models are best at forecasting models for Botswana and Nigeria stock markets respectively.

Horizons and Estimation periods

Whilst the accuracy of alternative forecasting models has been thoroughly explored, other parameters in the estimation process have been comparatively neglected, including the estimation window and its relationship with forecasting horizons. A notable exception to this is a 2012 paper by Brownless et al who investigate the type of model, the amount of data to use in the estimation, the frequency of estimation update, and the relevance of heavy-tailed likelihoods for volatility forecasting. This study is conducted across a variety of assets, including equity and foreign exchange where ten exchange rates are examined. How much data to use in estimation becomes an important issue if parameters are unstable, as data that is excessively aged can bias estimates and cloud forecasts. The authors' estimates reveal slowly varying movements in model parameters and the results show that using the longest possible estimation window provides the most accurate forecast results. However, even when using long data histories, the paper finds that models should be re-estimated at least once per week to mitigate the effects of parameter drift. These results hold across the 5 GARCH based models used in the paper. The Brownless et al paper uses a quasi-likelihood loss (QL) and the mean square error (MSE) performance measures.

These findings are consistent with the earlier work on estimation periods. Ng and Lam (2006) who conduct a study on the implications of the size of the estimation window on GARCH, MEM-GARCH model effectiveness for the NASDAQ composite index. Their estimation window sample sizes consist of 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1500, 2000 and 3000. Through an investigation of the correlation of conditional variances at the different sample sizes, they recommend 1000 samples for GARCH and 800 for MEM-GARCH. Hwang and Pereira (2006) study the implications of small sample sizes for GARCH and ARCH models. Using S&P500 index daily returns and Monte Carlo simulations, the authors demonstrate that a high level of persistence in GARCH(1,1) models obtained from the use of a large number of observations have autocorrelations lower than these estimations suggest in smaller samples. This is consistent with negative biases in small samples in the maximum likely-hood estimates of ARCH and GARCH parameters. Therefore, in small samples, there may be significant convergence errors from the negatively biased ARCH and GARCH estimates. Considering the size of biases and convergence errors, it is proposed that at least 250 observations are needed for ARCH(1,1,1) models and 500 observations for GARCH(1,1) models.

The implications of forecasting horizons on volatility prediction accuracy is an important area for risk management. A comprehensive piece of work in this field was conducted in 2010 by Chan et al. The paper empirically tests how and to what extent the choice of the sampling frequency, realized volatility measure, forecasting horizon, and time-series model affect the quality of volatility forecast. The study finds that the quality of prediction is significantly affected by the forecasting horizons, that is, there is forecasting accuracy decay over longer horizons. The finding is intuitive; with longer forecasting horizons there is greater potential for forecasting errors. Similar findings are echoed by Yıldıran and Fettahoğlu (2017) who find that when predicting US treasury rates, short term ARIMA models outperform longer-term models over the years 2005-2017.

Volatility forecasting in foreign exchange markets

Volatility prediction in foreign exchange markets is an extensively explored field, although there remains much to be discovered. In a 2017 paper, Lahmiri uses various volatility models, including GARCH, EGARCH, and ANN (artificial neural networks) methodology to forecast US/Canada and US/Euro exchange rates volatilities. The paper concludes that ANN with technical analysis indicators outperformed conventional GARCH type models in terms of mean absolute error, mean of squared errors, and the Theil index. In an effort to assess the efficiency of foreign exchange markets at incorporating data, Pilbeam and Langeland (2014) assess the relative effectiveness of GARCH and implied volatility models at forecasting exchange rate volatility. Over the period 2002-2012 the authors find that implied volatility forecasts outperform GARCH based models in periods of high and low volatility. From this, the authors conclude that the results suggest markets efficiently price in future volatility.

Amongst GARCH models there is no consensus on which models perform the best. Vilasuso (2002) finds that the FIGARCH model is more able to capture the salient features of exchange rate volatility than are the more commonly used GARCH and IGARCH models. And perhaps more importantly, the FIGARCH model generates superior out-of-sample volatility forecasts, and the gains in forecast accuracy using MAE and MSR (Mean square error) performance assessments are substantial. Three years later Hansen and Lunde (2005) find that when assessing 330 GARCH type models they do not find evidence to support the use of more elaborate GARCH based models for volatility estimation in the Deutsche Mark-Dollar exchange rate. This demonstrates the lack of agreement on the topic. Further, Dunis and Williams (2003) examine the medium-term forecasting ability of several alternative models of currency volatility, including GARCH and Auto-regressive models with respect to 8 currency pairs. They find that no specific volatility model outperforms in forecasting volatility for the period 1991-99. Ganbold et al (2017) study volatility forecasting in the Turkish Lira exchange rate, which makes for an interesting study due to the consistent growth in political uncertainties causing volatility in financial

markets. The authors forecast the volatility of the TL/USD with three models, ARIMA, SARIMA, and SVAR. The forecast comparison by RMSE and MAE finds that the SARIMA model forecasts are more accurate than the other models.

Realized volatility

An important element of studies that evaluate volatility forecasting is what the forecasts are being compared to; the benchmark. Taylor and Xu (1997) and Andersen and Bollerslev (1998) show that measurement errors in the estimation of realized volatility can have profound detrimental implications on the informational content of volatility forecasts. By having an inaccurate benchmark which results are compared to, the validity of a study is easily questionable. An effective method of calculating realized volatility is through the use of intraday returns, a methodology employed by Hansen and Lunde (2005). This method is powerful in that it uses a relatively large amount of data for a specific day, making it a precise measure. The importance of realized volatility is emphasised by Chan et al (2010) who assess four measures of realized volatility; intraday, total, scaled total, and close-to-close volatility measures. The authors find that the impact of forecasting horizon does not only depend on forecasting models but also on the choice of ex-post volatility. In general, the choice of realized volatility measure is at least as important as the choice of forecasting models.

Unfortunately, obtaining intraday returns can be a challenging endeavour. Andrade et al encounter this problem in a 2004 conference proceeding assessing volatility in the Brazilian Real exchange rate against the US Dollar. As a solution, the authors aim to improve the quality of their measures of realized volatility by using the Parkinson (1980) estimator, which improves the efficiency of realized volatility measures by using information embedded in daily high and low prices. Garman and Klass (1980) proved that this is an unbiased estimator of volatility, which is around five times more efficient than the sample standard deviation.

3. Theoretical background

This next section will provide theoretical information on the models and concepts used in this study. The most basic of which is our definition of return which is defined as the natural logarithm of the spot exchange rate at time t divided by the spot exchange rate of the previous period outlined below in equation 1.0.

$$R_{t+1} = \ln(\frac{S_{t+1}}{S_t}) \quad \textbf{(1.0)}$$

GARCH(p,q)

The GARCH model is specified by two terms p and q. These terms specify the number of lagged squared returns (p) and lagged variances (q) to make to make a next day (t+1) variance forecast. The model consists of autoregression (AR), conditional variance, and heteroskedasticity. The first of these elements, autoregression, specifies that tomorrows variance is a regressed-on today's variance, hence it regresses on itself. The second element, conditional variance, means that tomorrows variance is dependent on today's variance. The final term, heteroskedasticity, specifies that variance varies over time; variance is not constant. Variance, as seen in equation 2.0 in the next period denoted by σ_{t+1}^2 is a function of a constant (ω) plus squared returns at time t multiplied by a coefficient (αr_t^2) plus the squared volatility at time t multiplied by a coefficient ($\beta\sigma_t^2$). The weights denoted α & β are calculated to determine the relevance of the square returns and variance of the previous period on the next period variance, along with a constant denoted ω. The weights of denoted α & β must be less than 1 (2.1).

$$\sigma_{t+1}^2 = \omega + \alpha r_t^2 + \beta\sigma_t^2 \quad \textbf{(2.0)}$$

$$\text{Where } \alpha + \beta < 1 \quad \textbf{(2.1)}$$

ARIMA

The ARIMA(p,d,q) model is characterised by three elements; p, d and q. The autoregressive term (p) allows for the incorporation of past values into the model by regressing the variable of interest on its own prior values, hence autoregressive. The moving average term (q) denotes the number of lagged forecast errors in the prediction equation. The third term (d) specifies the number of differences represented by the "I" term in ARIMA for "integrated". The number of differences is the difference between the current value and previous values which is used to remove non-stationarity from the forecasting equation.

The Parkinson estimator realized volatility measure

The 1980 Parkinson estimator incorporates the natural logarithm of the highest and lowest spot exchange rate on day t, denoted by H_t and L_t respectively. The equation for calculation can be found below in equation 3.0.

$$\boldsymbol{PK_t = \sqrt{\frac{1}{4ln(2)}\frac{1}{T_t}\sum_{k=1}^{T_t}(H_{t+k} - L_{t+k})^2}} \quad \textbf{(3.0)}$$

RMSE

The RMSE performance measure takes the square root of the average squared prediction error, with the predicted value denoted P and the observed O. The equation can be found at 4.0.

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}} \quad \textbf{(4.0)}$$

MAPE

The MAPE measure takes an average of the absolute percentage prediction error, that is, the difference between the predicted and observed value denoted P and O respectively. By dividing the absolute difference between the predicted and observed by the observed value, the scale, i.e the relative size of the error is accounted for. This is demonstrated in equation 5.0 below.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{[O_i - P_i]}{O_i} \quad \textbf{(5.0)}$$

4. Research Hypotheses

Based on the findings of the surrounding literature we arrive at three formal hypotheses regarding our volatility forecasts in foreign exchange markets that we research in this study:

Hypothesis one: Longer estimation periods will result in more accurate volatility forecasts

Hypothesis two: The benefits of longer estimation periods in terms of accuracy will be greater for forecasts that are further into the future

Hypothesis three: The benefits of longer estimation periods in terms of accuracy will be greater for GARCH forecasts than ARIMA forecasts

5. Data

This study investigates the currency asset class as this area had studies that were somewhat similar, but a significant literature gap remained that this piece of work aims to contribute to whilst being complementary to other research. We use exchange rate data of three currency pairs; USD/EURO, USD/GBP and EURO/GBP over the period 21/01/2009 to 29/03/2019. These currency pairs were chosen in an attempt to mitigate external factors that may interfere with our research; these pairs have large flow volumes and similar economic development of the countries they represent. The time period was chosen to avoid the unpredictability of including the global financial crises whilst being up-to date to remain relevant and having a sufficiently large dataset. Starting in 2009 we do not completely avoid the abnormal impacts of the global financial crisis; however, this was a trade-off between avoiding this element and incorporating more data into the study. We use closing price at 16:00 GMT for the currency spot rate combined with daily high and low spot rates for our calculation of realized volatility consistent with the Parkinson Estimator (equation 3.0). The data is obtained from Thomas Reuters Datastream.

Summary statistics

Figure 1 below shows the descriptive statistics for the currency pairs used in this study. The EURGBP currency pair has a much lower standard deviation than either of these currencies have with the dollar, which could indicate that the dollar is a more volatile currency, or that the Euro and GBP have a greater correlation with each other, which is very possible as the UK and Eurozone have a high degree of economic integration despite independent monetary policies. The EURGBP currency pair is the only pair to exhibit positive skew and kurtosis. This means that extreme returns in the right tail of the distribution i.e extreme appreciation in the exchange rate is more likely than the normal distribution would predict over our sample period.

Figure 1: Descriptive Statistics

Descriptive Statistics

Measure	USDEUR	USDGBP	EURGBP
Mean	1.25509	1.497627	1.196157
Standard deviation	0.118571	0.133422	0.078843
Minimum	1.0387	1.20485	1.059
Maximum	1.514	1.71655	1.44145
Skew	-0.00781	-0.57333	1.156213
Kurtosis	-1.18963	-0.97126	0.756156

6. Methodology

This study assesses the implications of the of the size of the estimation window in model forecasting accuracy. To achieve this, in its most basic form, this study assesses the accuracy of our exchange rate volatility predictions by comparing our forecasted results generated from alternative parameter models from the actual observed results in the foreign exchange markets. To assess the implications of changing the size of the estimation window parameter we use windows of different sizes. These include windows of two years, one year, half a year, and a quarter of a year. For this we require window sizes of 504, 252, 126, 63 observations respectively to compensate for the number of trading days in a year. We use one-day and 126-day forecast horizons to consider the impact on forecast accuracy on short-term and longer-term predictions. For our one-day predictions, we use a rolling window methodology and so the size of the estimation window is equal to the size of the rolling window. The window moves forward by one for each iteration of the model with the model re-estimated for each forward roll. Our half year volatility forecasts predict from the end of our sample period to 126 days ahead of the in-sample period, which brings us to the date 29/03/2019. Most of our empirical analysis is conducted in MATLAB, with minor usage of excel.

To achieve our estimations, we use GARCH (1,1) and ARIMA (1,1,1) models. The literature surrounding which volatility models are best employed is vast, and so as the purpose of this paper is not to investigate volatility model accuracy. We simplify the controversial topic and employ two of the most basic and common models available whilst using two models for the purpose of checking robustness. Once our models have been specified, we use Monte Carlo simulations to model the probability of different outcomes in a process that uses random variables. The use of this process incorporates the risk arising from the random element of our forecasting procedure. We run 1,000 iterations for each estimate.

To determine the accuracy of our model forecast we employ the RMSE and MAPE measures, both of which are commonly found in the literature. These measures compare the predicted results with the observed results, that is, realized volatility. To provide additional support we investigate our 126-day forecast errors to determine if we can see any trends that may provide further support for our findings. The results of our findings can be found in figure 2 with the forecast errors of the 126-day predictions plotted in figures 3-9; the findings of these are discussed and explained in section 7. For all measures, smaller results indicate more accurate volatility forecasting as our predictions are on average closer to the realized volatility measure. Ideally, we would have used intra-day returns to calculate daily volatility, however we were unable to obtain this data. To compensate, as a proxy for realized daily volatility we use the Parkinson (1980) estimator, which incorporates daily high and low exchange rate data to make a prediction on volatility. This allows us to incorporate data on the nature of daily movements, rather than simply the closing price, a very simple methodology. The Parkinson estimator

is said by Duque & Paxson (2019) to be 10 times more efficient compare with the classical volatility estimator, and it is crucial that our realized volatility estimate is accurate as it serves as our benchmark from which we compare our model estimations.

7. Results

In this section we discuss the findings of our empirical analysis. First, we will examine the GARCH forecast results, then followed by the ARIMA results. After these two elements have been discussed we will compare the two models results to gain a deeper understanding of the implications of the size of the estimation period. The GARCH results can be found in figure 2 whilst the ARIMA results can be found in figure 6; both of which specify RMSE and MAPE values over different estimation windows at one and 126-day forecasts. This is complimented by an inspection of the model 126-day forecast errors found in figures 3-5 for GARCH and figures 7-9 for ARIMA. The lower the RMSE and MAPE meaures in figures 2 and 6 the more accurate the forecast in terms of being closer to our observed volatility. Figures 3-5 for GARCH and figures 7-9 for ARIMA show a plot of the 126-day forecast errors, that is, the absolute difference between our forecasted results and the benchmark observed volatility for each of the three curreney pairs. The X-axis represents time in days whereas the Y-axis demonstrates the size of the absolute forecast errror; smaller absoute forecast errors represent more accurate results.

GARCH

Firstly, we will investigate the results from our GARCH estimation model, which can be found in figure.2. With regards to the one-day forecasts, we can see that the USDGBP currency pair is the only pair of the three to have consistently more accurate forecasts in terms of both RMSE and MAPE as the length of the estimation period increases. The EURGBP currency pair demonstrates more accurate forecasts for longer estimation periods for MAPE, however there is a marginal increase from 0.000916 to 0.000948 between estimation periods 252 and 504 for the RMSE measure. For the USDEUR currency pair performance measures improve up to and including the 252-estimation period, with a noticeable decrease in performance at the 504-estimation period, although performance is still much greater than the 126-estimation period for both RMSE and MAPE. For all measures we can suggest that longer estimation periods increase accuracy up until the 252-estimation period at which point forecasts become less accurate, or the marginal rate of improvement decreases. This could indicate that there is a point at which too much data becomes detrimental, or less useful to forecast one-day ahead volatility. This evidence somewhat supports hypothesis one; a larger estimation window does increase one-day forecasts, but only up to a point, after which it can become detrimental or the marginal benefit decreases. We will now move the investigation to the 126-day GARCH forecasts. The trend in these results is

much more clear cut than the one-day forecasts; other than an increase in the USBGBP MAPE measure from estimation period 63 to 126, for the rest of the performance measures across all currency pairs there is persistent increase in performance with longer estimation periods for both RMSE and MAPE. The results indicate that longer estimation periods almost always result in more accurate 126-day forecasts providing significant support to hypothesis one.

We can obtain additional support for this statement from an inspection of the 126-day absolute forecast errors plotted in figures 3-5. In these graphs we see that beyond the forecast period of 10 days the 504 day estimation period forecasts are always at least as good as the other estimation period forecasts, having substantially lower errors than the 63-day and 126-day forecasts in all currency pairs, and the 252-day estimation period forecasts in the USDEUR and EURGBP currency pairs, with the USDGBP 252-day and 504-day difference appearing quite marginal. From comparison with the other currency pairs, it appears this is due to the USDGBP 504-day estimation period forecasts doing relatively poorer rather than the 252-day estimation period results delivering higher performance. Generally, the plots show that longer estimation periods result in lower errors, although after roughly the 60-day forecast the 63-day and 126-day estimation period errors converge in all currency pairs.

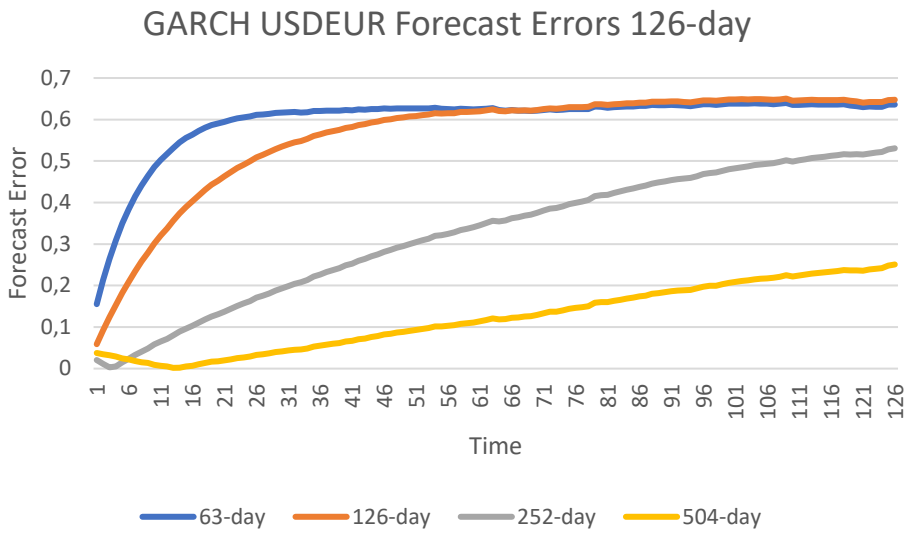
The results in figure 2 appear to support hypothesis two; the trend in the 126-day forecasts is clear that longer estimation windows result in more accurate forecasts in the RMSE and MAPE measures whereas this trend is only partial, that is, up to a certain size in the one-day forecasts.

Figure 2: GARCH Performance

GARCH(1,1) Forecasting Results Performance

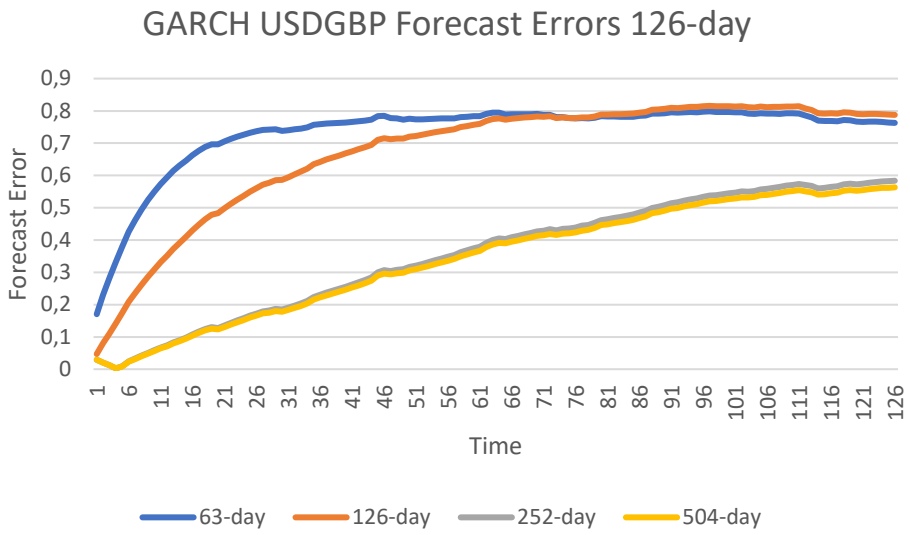
		USDEUR		USDGBP		EURGBP	
GARCH(1,1)	Estimation period	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
One-Day	63	0.009496	191.6627	0.018773	279.3506	0.009936	279.3506
	126	0.00242	96.59999	0.005149	141.2539	0.002587	141.2539
	252	0.000804	53.82761	0.001545	74.46388	0.000916	74.46388
	504	0.000964	66.95659	0.000787	51.57906	0.000948	51.57906
126-Day	63	0.182974	1387.38	0.276036	1209.109	0.176119	1107.294
	126	0.166394	1305.059	0.243568	1567.858	0.159403	1016.601
	252	0.064846	772.2888	0.07975	584.1096	0.05995	560.1089
	504	0.010395	293.6536	0.074416	564.1979	0.008425	193.2002

Figure 3: GARCH 126-Day Forecast Errors USDEUR



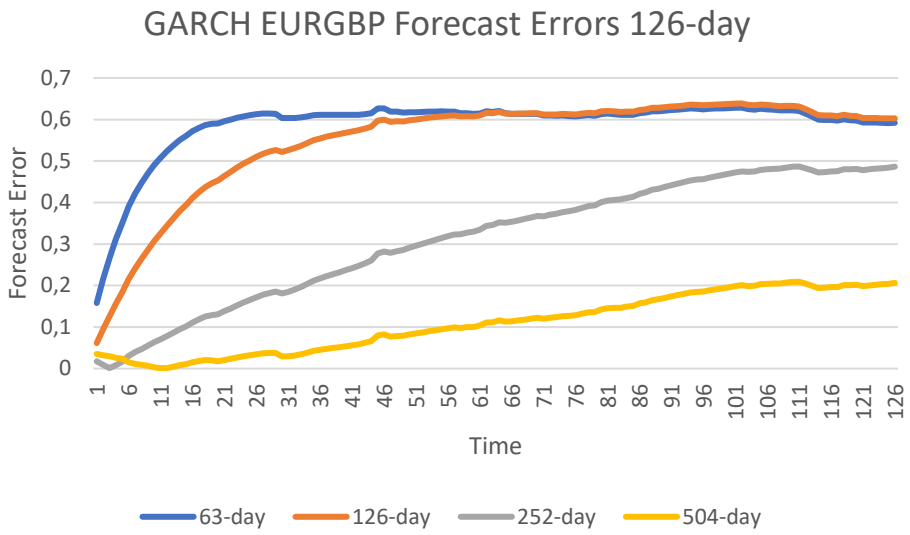
The X- axis signifies time in days and the Y-axis shows the absolute forecast error (Absolute difference between forecasted and realzied volatility)

Figure 4: GARCH 126-Day Forecast Errors USDGBP



The X- axis signifies time in days and the Y-axis shows the absolute forecast error (Absolute difference between forecasted and realzied volatility)

Figure 5: GARCH 126-Day Forecast Errors EURGBP



The X- axis signifies time in days and the Y-axis shows the absolute forecast error (Absolute difference between forecasted and realzied volatility)

ARIMA

We will now investigate the results of our ARIMA forecast, the results of our RMSE and MAPE performance measures can be found in figure 6. The one-day forecasts show improvements in performance measures up until the 252 estimation period, after which point the RMSE and MAPE measures degrade in performance across all currnecy pairs. This provides somewhat weak support for hypothesis one; a larger estimation window does increase one-day forecasts, but only up to a point, after which it becomes detrimental to the ARIMA forecasts. Discussing the 126-day forecasts, across all currency pairs there is an increase in performance as the estimation period increases with an exception; the 126 estimation period performs worse than the 63 and 252 estimations. This phenomenom can also be seen in figures 7-9 which demonstatres a plot of the absolute forecast errors. Past a forecast of roughly 30-days the 63-day estimation window forecasts outperform the 126-day estimation window results across all three currnecy pairs. Figures 7-9 also show that aside from the unexpected performance of the 63-day estimation, or indeed the 126-day estimation window forecasts that more data improves the accuracy of ARIMA forecasts. Although in general it does appear that more data results in more accurate 126-day forecasts, however we can only say that the results provide some support. It is particuarly interesting that the 126-day esimtation results are unexpected considering this is the size of the forecast horizon. With regards to the unexpected 63-day vs 126 day result, and we are unaware of any potential rationale that sufficiently explains the phenomenon.

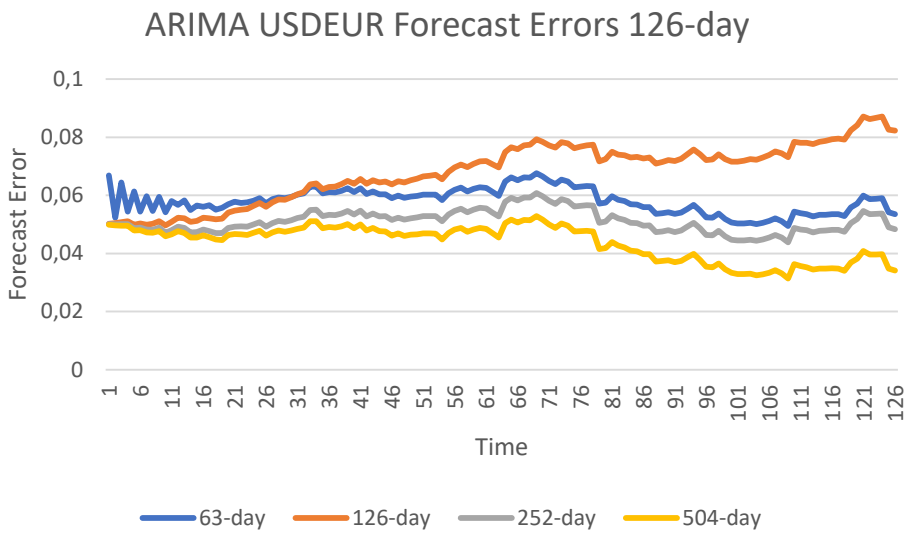
The results of figure 6 do not sufficiently support hypothesis two. The one-day results increase performance up to an estimation period of 252, with results becoming detrimental thereafter. However, this is only one estimation window (504) out of the four. The 126-day results increase performance with the size of the estimation window in all estimation windows apart from the 126 estimation window, which is also one of the four.

Figure 6: ARIMA Performance

ARIMA(1,1,1) Forecasting Results Performance

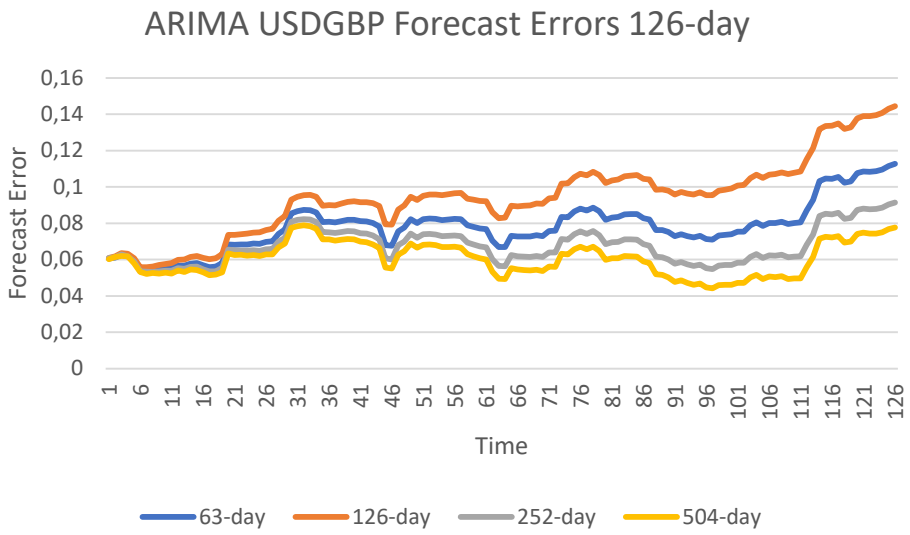
		USDEUR		USDGBP		EURGBP	
ARIMA(1,1,1)	Estimation period	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
One-Day	63	0.002799	113.0531	0.002743	115.9144	0.002549	115.9144
	126	0.002331	106.6223	0.002309	107.2673	0.002095	107.2673
	252	0.002262	106.2641	0.002278	107.0738	0.002045	107.0738
	504	0.002269	106.6189	0.002334	108.0017	0.002055	108.0017
126-Day	63	0.001699	133.0715	0.003134	126.5622	0.002454	126.7674
	126	0.002367	157.8379	0.004645	219.4329	0.00334	142.9347
	252	0.001312	116.9852	0.002331	109.5315	0.002001	113.0799
	504	0.000976	99.44977	0.001876	98.12698	0.001545	100.2056

Figure 7: ARIMA 126-Day Forecast Errors USDEUR



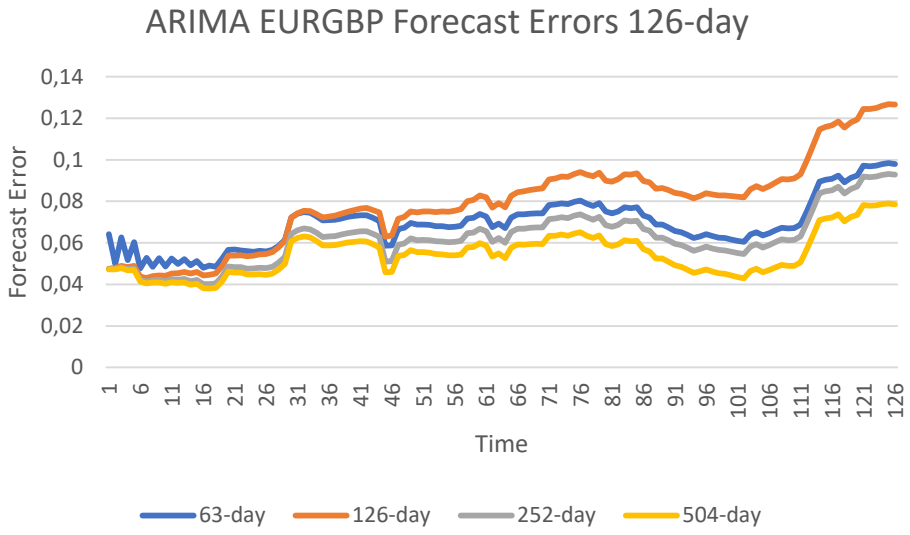
The X- axis signifies time in days and the Y-axis shows the absolute forecast error (Absolute difference between forecasted and realzied volatility)

Figure 8: ARIMA 126-Day Forecast Errors USDGBP



The X- axis signifies time in days and the Y-axis shows the absolute forecast error (Absolute difference between forecasted and realzied volatility)

Figure 9: ARIMA 126-Day Forecast Errors EURGBP



The X- axis signifies time in days and the Y-axis shows the absolute forecast error (Absolute difference between forecasted and realzied volatility)

Comparison of the GARCH and ARIMA forecasts

For the GARCH and ARIMA forecasts a general trend can be seen across both models; more data results in more accurate forecasts, although there are many caveats to this statement. Writing generally, for both models the one-day forecasts witness increased performance until a point, after which there is a nominal or marginal decrease in performance as the estimation period increases. This provides some but not universal support for hypothesis one. With the 126-day forecasts the GARCH results are almost universal with an increase in performance with longer estimation periods, this applies to the ARIMA forecasts although with an exception for the 126 day forecasts; this is the same period in which the GARCH 126-day forecasts notice a decrease in the MAPE performance measure in the 126-day USDGBP currency pair. Perhaps the rationale for this occcurance in the two models is linked, although this would require further investiagtion. Again, the 126-day results provides some but not universal support for hypothesis one.

It appears that longer estimation periods are more critical to the GARCH one-day and 126-day forecasts than those of the ARIMA models, supporting hypothesis three. The GARCH forecasts performance measures start out very poorly at the 63 day estimation periods and generally improve drastically with longer estimation periods for both measures. The increase in performance is large in both nominal and relative terms. With the ARIMA forecasts, there is comparatively much less improvement from the 63-day forecasts to any of the longer estimation period results. This can be observed noticabily in forecast error plot seen in figures 3-5 for GARCH and 7-9 for ARIMA. The benefits of longer estimation windows can be seen as substiantial for GARCH but comparitively marginal for ARIMA; interestingly, the differences in performance are apparent early on in the GARCH forecasts but they only become more noticable towards the end of the ARIMA forecasts. With the one-day forecasts, the improvement in performance of the ARIMA forecasts is indeed very small; a statement that also somewhat applies to the early ARIMA 126-day forecast errors. This is as would be expected; the literature notes that it is critical GARCH estimates have sufficient data to make acccurate forecasts, which our findings echo.

8. Further research

The scope of this study is limited, and so the field could benefit from further investigation of elements this article does not sufficiently, or at all address. A primary aspect that warrants further investigation is the implications of estimation windows over different forecast horizons on a rolling basis. This study investigates rolling one-day forecasts and 126-days from the end of the in-sample period; methodology that investigates rolling estimation windows over 10-day and 20-day forecasts would be an important development for the private sector due to the importance of 10-day and 20-day forecasts in regulatory capital requirements. Further investigation into the size of estimation windows is also warranted; as we only conducted our research into four window sizes we were unable to support hypothesis two in the ARIMA forecasts as both the one-day and 126-day forecasts had one estimation window size that delivered less accurate forecasts than the window before it. Perhaps a methodology that assesses estimation windows on a continuous basis would be helpful, although this would be computationally very straining. Another subtopic that deserves attention is the implications of the estimation window on different forecasting models. The performance of different models is a saturated topic and so this study investigated two of the most basic; GARCH(1,1) and ARIMA(1,1,1). A justified twist would be the performance of these models under different parameters, including the size of the estimation window.

9. Conclusion

Overall, this study finds that generally more data leads to more accurate forecasts in foreign exchange rate volatility providing some support to hypothesis one. The results are not unanimous, with some unexpected results. The one-day forecasts accuracy increases with more data up to a point, after which more data becomes detrimental, or the marginal increase decreases for both GARCH and ARIMA. More data is more important for the 126-day forecasts than the one-day forecasts for GARCH however we cannot say the same for ARIMA. Overall, this provides some but not complete support for hypothesis two. Having more data is more important for GARCH estimations than for those of ARIMA, supporting hypothesis three, which is consistent with other research. And so, if data is limited the detrimental impact is not so great with one-day forecasts, or when ARIMA(1,1,1) forecast models are used. Agents should consider which model to use when faced with limited data, or endeavour to acquire more data.

10. References

Adebayo, F., & Sivasamy, R. (2014). Forecasting stock market series with ARIMA model. *Statistical and Econometric Methods*, (3), 65–77.

Alexander, C., & Sheedy, E. (2008). Developing a stress testing framework based on market risk models q. *Journal of Banking and Finance*, *32*, 2220–2236.

Alkhazaleh, M. M. H. (2018). Forecasting Banking Volatility in Amman Stock Exchange by Using ARIMA Model. *British Journal of Management*, *29*(3), 1–9.

Anderson, T., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review*, *39*(4), 885–905.

Andrade, S. C. De, Tabak, B. M., & Chan, E. (2004). Tracking Brazilian Exchange Rate Volatility. In *Econometric Society 2004 Far Eastern Meetings 487*. Econometric Society.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Economics*, *31*(3), 307–327.

Box, G., & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco.: Holden-Day.

Brownlees, C., Engle, R., & Kelly, B. (2012). A practical guide to volatility forecasting through calm and storm. *The Journal of Risk*, *14*(2).

Chai, T., & Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev*, *6*, 1247–1250.

Chan, W., Cheng, X., & Fung, J. (2010). Forecasting volatility: Roles of sampling frequency and forecasting horizon. *The Journal of Futures Markets*, *30*(12), 1167–1191.

Chen, C., Twycross, J., & Garibaldi, J. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLOS ONE*, *12*(3).

Christoffersen, P. (2019). Evaluating Interval Forecasts. *International Economic Review*, *39*(4), 841–862.

Duffie, D., & Pan, J. (1997). An Overview of Value at Risk. *The Journal of Derivatives*, *4*(3), 1–85.

Dunis, C., & Williams, M. (2003). *Applications of Advanced Regression Analysis for Trading and Investment*.

Duque, J., & Paxson, D. (2019). Empirical evidence on volatility estimators.

Ederington, L., & Guan, W. (2019). Longer-Term Time-Series Volatility Forecasts. *The Journal of Financial and Quantitative Analysis*, *45*(4), 1055–1076.

Engle, R. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, *50*(4), 987–1007.

Ganbold, B., Akram, I., & Lubis, R. (2017). *Exchange rate volatility: A forecasting approach of using the ARCH family along with ARIMA SARIMA and semi-structural-SVAR in Turkey*. (No. 84447).

Garman, M., & Klass, M. (1980). On the Estimation of Security Price Volatilities from Historical Data. *The Journal of Business*, *53*(1), 67–78.

Hansen, P., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, *20*, 873–889.

Hwang, S., & Pereira, P. (2006). Small sample properties of GARCH estimates and persistence Small Sample Properties of GARCH Estimates and Persistence. *The European Journal of Finance*, 473–494.

Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Lahmiri, S. (2017). Modeling and predicting historical volatility in exchange rate markets. *Physica A: Statistical Mechanics and Its Applications*, *471*(1), 387–395.

Miswan, N., Ngatiman, N., & Hamzah, K. (2014). Comparative Performance of ARIMA and GARCH Models in Modelling and Forecasting Volatility of Malaysia Market Properties and Shares. *Applied Mathematical Sciences*, *8*(140), 7001–7012.

Nelson, D. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica*, *59*(2), 347–370.

Ng, H., & Lam, K. (2006). *How Does the Sample Size Affect GARCH Model?*

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, *53*(1), 61–65.

Pilbeam, K., & Langeland, K. (2014). Forecasting exchange rate volatility: GARCH models versus implied volatility forecasts. *International Economics and Economic Policy*.

Poon, S.-H., & Granger, C. W. J. (2003). Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature*, *41*(2), 478–539.

Slutzky, E. (1937). The Summation of Random Causes as the Source of Cyclic Processes. *Econometrica*, *5*(2), 105–146.

Taylor, S. (1987). Forecasting the volatility of currency exchange rates. *International Journal of Forecasting*, *3*(1), 159–170.

Taylor, S., & Xu, X. (1997). The incremental volatility information in one million foreign exchange quotations. *Journal of Empirical Finance*, *4*(4), 317–340.

Vilasuso, J. (2002). Forecasting exchange rate volatility. *Economics Letters*, *76*(1), 59–64.

Willmott, J., & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research*, *30*(79).

Willmott, J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science*, *20*(1), 89–102.

Wold, H. (1938). *A Study in The Analysis of Stationary Time Series*. Almqvist & Wiksells.

Yıldıran, C. U., & Fettahoğlu, A. (2017). Forecasting USDTRY rate by ARIMA method. *Cogent Economics & Finance*, *5*(1), 1–11.

Yule, G. (1926). Why do we Sometimes get Nonsense-Correlations between Time-Series?--A Study in Sampling and the Nature of Time-Series. *Journal of the Royal Statistical Society*, *89*(1), 1–63.