Supervisor
Krzysztof Podgorski

LUND UNIVERSITY

ARMA and GARCH models for silver, nickel and copper price returns

Authors
Ola Andersson
Olle Holmberg
Mats Hansson

Abstract

This thesis compares Auto Regressive Moving Average (ARMA) and Generalized Auto Regressive Conditional Heteroscedacity (GARCH) models for three metal commodities. ARMA models have an unconditionally non-random and constant variance, which typically serves well in effectively representing homoscedastic data. The GARCH models feature variable variance that is non-random when conditioning on the past. Thus these models are often used to represent heteroscedastic data. It is documented that financial data, including metal commodities frequently exhibit heteroscedacity. This thesis investigates if this heteroscedacity in the observed historical data is shown in the quality of its ARMA and GARCH fits. The data used for comparison involve three time series of logarithmic price return for silver, nickel and copper. In the hypothesis it is assumed that GARCH is more efficient than ARMA. The efficient market hypothesis is also tested.

The logarithmic price returns are stationary which is confirmed by statistical tests. Thereby, it is appropriate to fit ARMA and GARCH models. The ARMA and GARCH models with the lowest Akaike's Information Criterion (AIC) are selected from each series. The models forecasted values and running standard deviations are cross-validated with the observed historical data using three measures. These measures are Mean Absolute Scaled Error (MASE), symmetric Mean Absolute Percentage Error (sMAPE) and correct pairs of sign which all provide different assessment of magnitude of error in estimation of the observed historical records. The correct pairs of sign are then tested against the efficient market hypothesis.

The error in estimation for forecast values does not yield a difference between ARMA and GARCH models by MASE. For the running standard deviation, both measures MASE and sMAPE are applied. The GARCH model is then more efficient than ARMA. In this sense, the thesis confirms the increased efficiency of using GARCH models for metal commodities.

According to correct pairs of sign measure, nickel has no arbitrage opportunities for logarithmic price return. This is expected according to the efficient market hypothesis. However, the test indicates it is possible to predict correct sign of logarithmic price return for copper and silver, which indicates that the efficient market hypothesis does not always apply.

Keywords: ARMA, GARCH, MASE, sMAPE, Heteroscedasticity, Stationarity, Ljung-Box test, McLeod-Li test, Running Standard Deviation, Forecast value.

Content

1 Introduction

Metal commodities account for a substantial part of trading and commerce around the world. Therefore, analyses of their price returns are important. A method for modeling price return is through time series analysis. For homoscedastic data Autoregressive Moving Average (ARMA) models are used and for heteroscedastic data, Generalized Autoregressive Conditional Heteroscedasticity (GARCH) models are used.

Time series analysis has developed considerably since the publication of. Box and Jenkins classic *Time Series Analysis: Forecasting and Control* (1974). For example, times series analysis has also been developed to fit financial data. Financial data have been observed to have heteroscedastic properties, meaning variable volatility according to Cryer and Chan (1998). Therefore, a new model called Auto Regressive Conditional Heteroscedastic (ARCH) was introduced by Engle (1982). The ARCH model estimates variance more efficiently.

The purpose of this thesis is to compare effectiveness of ARMA and GARCH models. ARMA models assuming homoscedastic properties, meaning constant variance. Also, the efficiency of the market for these metals is investigated through observing if fitted models can forecast future prices.

1.1 Hypothesis

Commodities including metals are often assumed to have heteroscedastic data for price returns. Due to this the GARCH model should be more effective than ARMA. Also, the efficient market hypothesis, with no arbitrage opportunities, states that future prices should not be possible to predict.

The aim of this thesis is to verify this hypothesis and in the process answer these questions:

- Which of ARMA or GARCH models is the most accurate for time series of silver, nickel and copper price return?
- Are models accounting for heteroscedastic data needed for these types of time series?
- Does the efficient market hypothesis apply for the logarithmic price returns of silver, nickel and copper?

1.2 Framework

This thesis focuses on statistical analysis. In addition, the economic theory of the *efficient market hypothesis* is interpreted in the context of the fitted model. The central aim of this thesis is the comparison of performance between ARMA and GARCH models. Only one type of GARCH model, is applied, which is viewed to be sufficient for the ordinary comparison. In short, considering other types of GARCH models is not needed to answer the main question of model adequacy.

Three time series are used to facilitate this comparison. The times series sample sizes (2,609 for both copper and nickel 8,871 for silver) are considered adequate for statistical fitting of the model's. The comparison of model accuracy is conducted only between the models one step ahead forecasts, longer steps of predictions are not discussed. While there are many cross-validation measures for model accuracy, they are measuring similar features and therefore only three have been selected: Mean Absolute Scaled Error (MASE), symmetric Mean Absolute Percentage Error (sMAPE) and correct pairs of sign.

2 Data

The data set used to facilitate the comparison of the models consist of three time series of silver, nickel and copper prices. The time series of silver price starts on 1980-01-01 and ends on 2013-12-31, with 8,871 observations. The time series of nickel and copper price starts on 2004-06-04 and ends on 2014-06-04, with 2,609 observations each. The prices have been observed and registered at the London Metal Exchange. They are the closing prices in US cents per troy ounce, for silver. The closing prices for nickel and copper, from each trading day, is in US dollar per metric ton. (Thomson Reuters 2014). Below, these data are plotted in Figure 1.

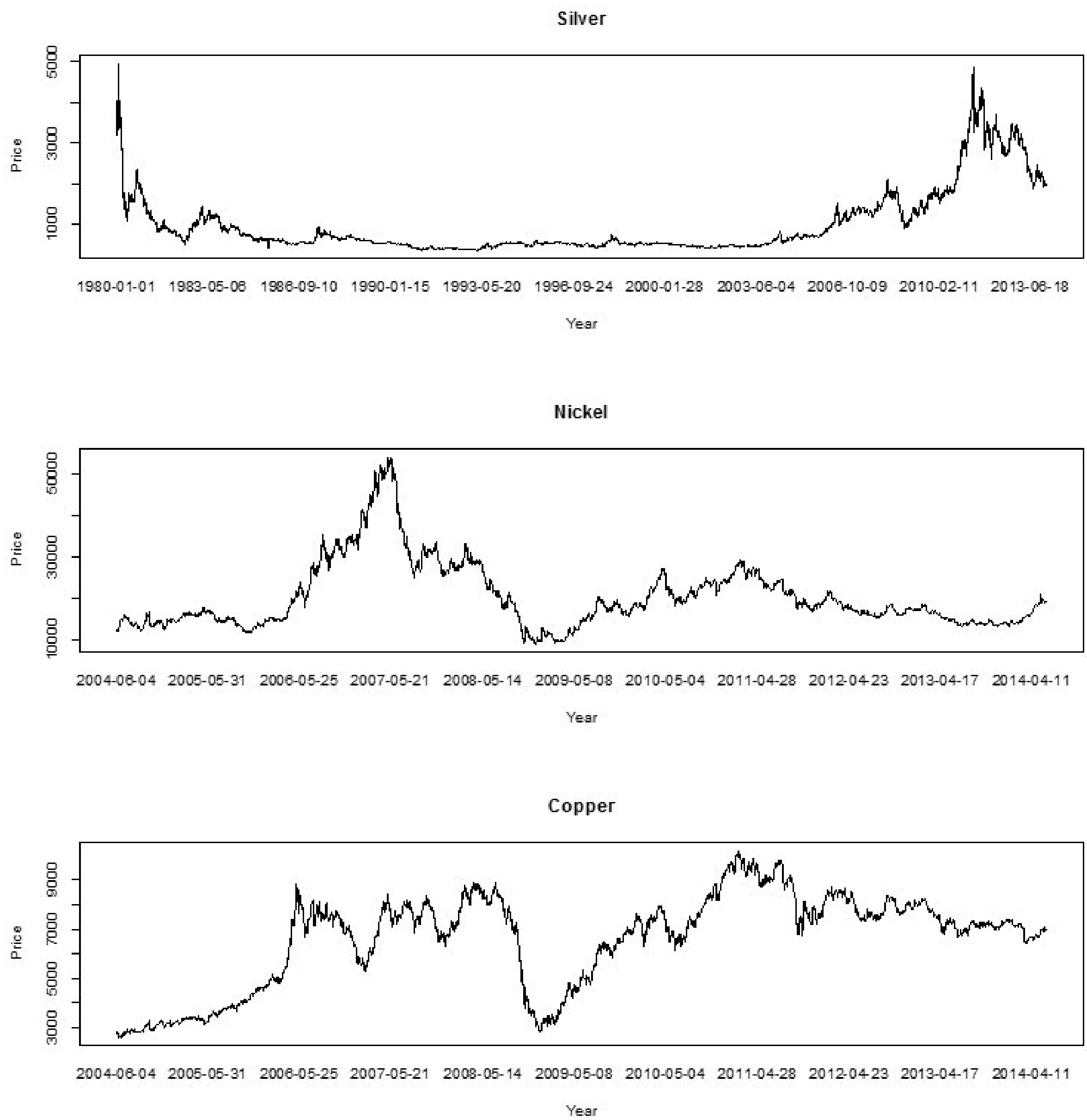


Figure 1. Original time series for silver price, nickel price and copper price

2.1.1 Transforming the commodity prices

From Figure 1, one can see that metal prices can be difficult to model directly. The long lasting trends are far from stationary and difficult to model stochastically with models such as ARMA and GARCH. It is, therefore, a standard procedure to model the logarithmic price return rather than prices themselves. Here, the procedure is briefly discussed.

If X_t is the price on the day t, then the return on the price for the next day is X_{t+1}/X_t . This can be viewed as the return on one dollar invested the previous day. The logarithmic return is defined as

$$Y_t = \log\left(\frac{X_{t+1}}{X_t}\right)$$

The plotted logarithmic price returns are shown in Figure 2.

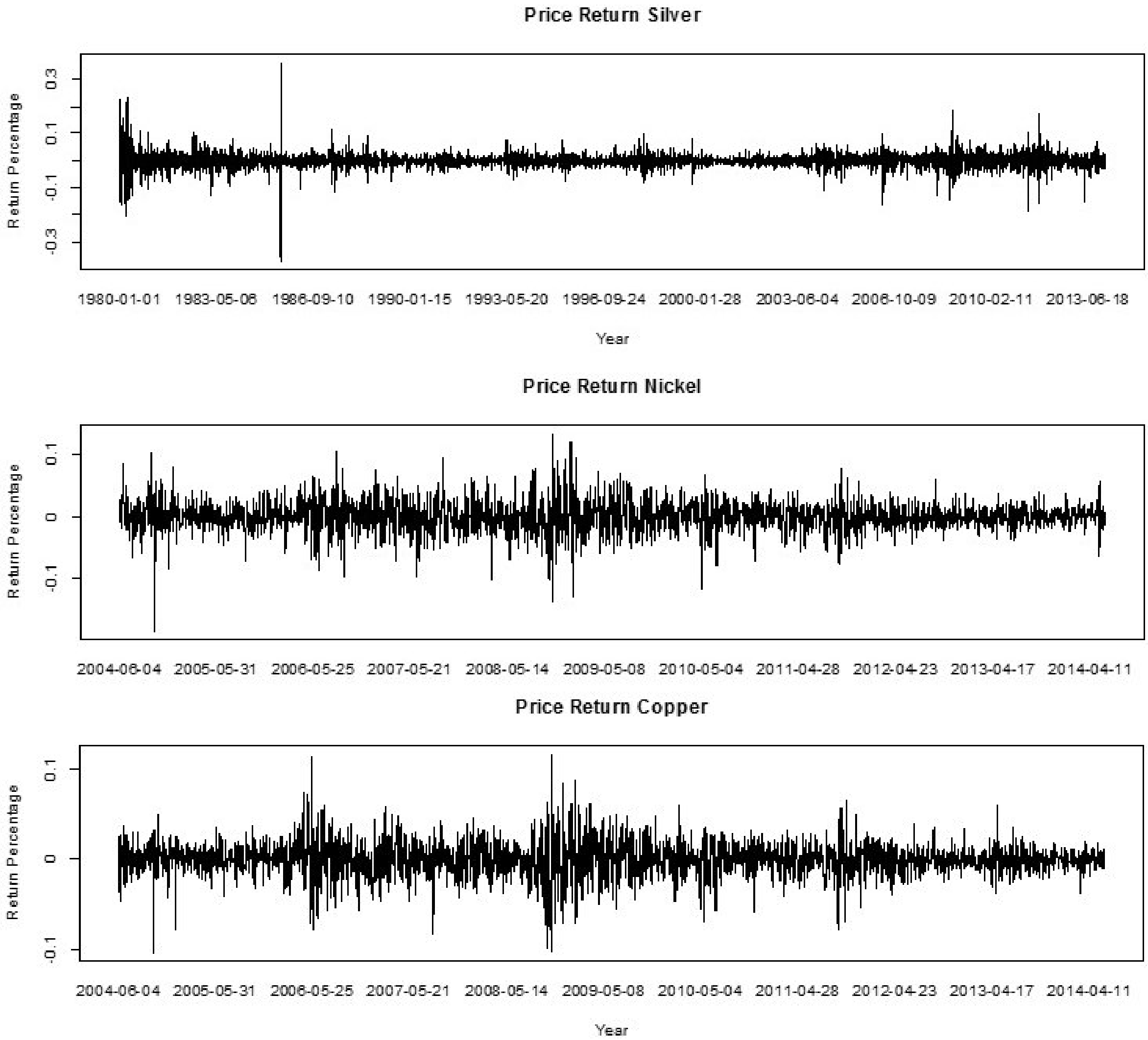


Figure 2. Transformed time series for price return for silver, nickel and copper.

2.1.2 Efficient Market Hypothesis

The transformed data, shown in Figure 2, can be modeled by a stationary process. In fact, this connects with the issue of arbitrage opportunities.

The arbitrage can shortly be described as a possibility of making gains on the market by predicting future prices. The efficient market hypothesis states that the market prices should not allow for arbitrage. There is an age-long debate about this in the theory of finance. Here, it is briefly discussed in the context of transforming the data in order to obtain a stationary process.

If the efficient market hypothesis applies, the price series can be described as random walk processes. This was shown to be true for daily price return of copper and nickel between 1989 and 2007, see Otto (2010).

Frequently, the price is replaced by the logarithmic price and the price return by the logarithmic price return:

$$\mathrm{W_t = \log(X_t)}, \quad \textbf{(1)}$$

$$\mathrm{W_t - W_{t-1} = Y_t = \log\left(\frac{X_t}{X_{t-1}}\right)}. \quad \textbf{(2)}$$

It can be demonstrated that the efficient market hypothesis can be reduced to the random walk assumption on the logarithmic prices, i.e. random walk process W_t.

$W_t = c + W_{t-1} + e_t$, where c is a drift for a random walk process.

When a transformation to logarithmic price return is conducted as above, the new transformed time series is a stationary white noise process:

$$c + \mathrm{e_t = W_t - W_{t-1}} \quad \textbf{(3)}$$

which satisfies,

$$c = E[(e_t + c)], \quad \textbf{(4)}$$

$$\sigma_e^2 = Var[(e_t + c)]. \quad \textbf{(5)}$$

This means that the model cannot forecast future values any better than guessing, implying that there are no arbitrage opportunities.

From this brief argument it follows that the logarithmic price return process should have a form of white noise, which is a stationary process. However, very often one does not initially restrict to this special class of stationary processes and sometimes dependence is often considered, for example through adding an autoregressive part to the white noise. It is then required to check if the autoregressive part indeed brings some benefits in modeling real data.

2.1.3 Dickey-Fuller test

In order to confirm whether the logarithmic price return is stationary and to be able to continue with model fitting of ARMA and GARCH, the Dickey-Fuller test is used. For the case of an AR (1)-process, the autoregressive parameter should be smaller than one in its absolute value, The Dickey-Fuller test subtracts y_{t-1} on both sides of the AR (1)-process and investigates the unit root from the characteristic equation. If there is a unit root, then the time series is non-stationary. (Enders 2009, 221-225).

H_0: Unit root
H_1: No unit root

Table 1. Test For Stationarity

Time Series	Dickey-Fuller (τ)	P-value
Silver	-19.581	0.01
Nickel	-13.241	0.01
Copper	-11.909	0.01

The actual p value is smaller than 0.01 but the R package, `tseries' that is used for performing this test does not report smaller values (in the package the function `adf.test()' is used on the data). Since the null hypothesis are rejected for all series, it implies that the logarithmic price returns are stationary which makes it reasonable to model with ARMA and GARCH.

3 Methodology

Selecting specific models to compare is the next step. First, the orders that will be examined for ARMA (p,q) and GARCH (p,q,u,v) are chosen. The different combinations of orders leads to 23 different for ARMA and 168 of GARCH. All these are then fitted, and the ARMA and GARCH with the lowest Aikaike's Information Criterion (AIC) value are selected from each time series. The ranges of orders that are examined are sufficient to answer the thesis questions. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are also regarded as an indication of order. However, the AIC criterion is prioritized.

The standardized residuals of the selected models are analyzed to determine if they are considered as white noise by the Ljung-Box test and if they are homoscedastic or heteroscedastic by the McLeod-Li test. These tests imply which model, ARMA or GARCH that is most appropriate.

The selected ARMA and GARCH models are used to estimate one-step ahead forecast value and running standard deviation. The estimated values are cross-validated with the observed historical values using three measures: sMAPE, MASE and correct pairs of signs.

3.1 Models

This thesis focuses on two classes of models: ARMA and GARCH. This part will provide a short explanation of what these models and their respective properties are. For these definitions of general models and later in chapter 3, when explaining the general functions of the applied tests and measures the $\{x_t, x_{t-1}, x_{t-2} \ldots x_n\}$ realization of the generic time series x_t is used. It should not be confused with $\mathrm{Y_t}$, which is defined as the logarithmic price return.

3.1.1 ARMA

An ARMA model is an Autoregessive process (AR) and a Moving Average process (MA). The ARMA model contains both parts. When a time series requires both autoregressive and moving average components, an ARMA (p,q) model is used. Its general form is:

$$x_t = \mu + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q}. \tag{6}$$

For stationarity, invertibility for the MA (q) part and stationarity for the AR (p) part are both required.

In the MA (q)-process, the q stands for the order and describes how many time lags back in time the model considers. The order also equals the number of parameters in the MA (q)-process:

$$x_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}, \tag{7}$$

so that the generic value y_t is a linear function of a constant value and q stochastic variables (Cryer and Chan. 2008, 57-65). In the above equation, e_t represent random innovation, i.e. a random noise for which:

$$E[e_t] = 0, \tag{8}$$

$$Var(e_t) = \sigma_e^2, \tag{9}$$

where σ_e^2 is the variance of the innovation. The MA (q)-process is stationary when the process is invertible. An MA (q)-process is invertible when it can be rewritten as an AR (p)-process. In order for that to occur, its roots of the MA (q)-process characteristic equation must exceed one in absolute value (Cryer and Chan. 2008, 57). The characteristic equation for the MA (q)-process is:

$$1 - \theta_1 x - \theta_2 x^2 - \theta_3 x^3 - \cdots - \theta_q x^q = 0. \tag{10}$$

Similarly, in the AR (p)-process, the p is the order and also describes how many time lags back in time the model considers. The order also equals the number of parameters in the AR (p)-process,

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \cdots + \phi_p x_{t-p} + e_t, \tag{11}$$

is a linear combination of its p last values of itself (Cryer and Chan. 2008, 66). In the above equation, e_t represent random innovation, i.e. a random noise for which:

$$E[e_t] = 0, \tag{12}$$

$$Var(e_t) = \sigma_e^2. \tag{13}$$

The condition for stationarity for the general AR (p)-process is the same as for an MA (q)-process, namely, roots of the characteristic equation must exceed one in modulus (Cryer and Chan. 2008, 76). The characteristic equation of the AR (p)-process is:

$$1 - \phi_1 x - \phi_2 x^2 - \phi_3 x^3 - \cdots - \phi_p x^p = 0. \tag{14}$$

3.1.2 GARCH

In the above models the variance of innovations (often referred to as volatility) is constant over time (homoscedastic variance). This often proves to be too restrictive of an assumption for real data. Under such an assumption, features like volatility clustering cannot be modelled. For this reason an ARCH model has been proposed by Engle (1982) to account for heteroscedastic variance of a time series applied on an AR (p)-process. The GARCH that combines with an ARMA (p, q)-process was introduced by Bollerslev (1986), where the conditional variance σ_t is used. A GARCH (p, q, u, v)-process for an ARMA (p, q)-process is defined as,

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + e_t + \Theta_1 e_{t-1} + \cdots + \Theta_q e_{t-q}, \quad \textbf{(15)}$$

where,

$$e_t = \sigma_t \varepsilon_t, \quad \textbf{(16)}$$

where ε_t is defined as the standardized residuals and,

$$\sigma_t^2 = \omega + a_1 e_{t-1}^2 + \cdots + a_u e_{t-u}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_v \sigma_{t-v}^2. \quad \textbf{(17)}$$

The notation for the order is u and v, which represents the number of parameters α and β. It is important to note that the innovation term e_t is not defined same as in the ARMA model (Cryer and Chan. 2008, 285-289). The most important property of the GARCH model is that the stochastic and variable volatility σ_t, is not random given the past of the process x_t, which can be seen from (12).

3.2 Model fitting

For all models the maximum likelihood method is used for parameter estimation. The likelihood estimation is defined on the joint probability density function and has for the conditional estimation a variance dependent on time, $L(\phi, \theta, \mu, \sigma_t^2 | x_1, x_2, \ldots, x_t)$, which the unconditional estimation has not $L(\phi, \theta, \mu, \sigma^2)$. It is based on the normal distribution of ε_t in (16). Hamilton (1994, 117-148). An advantage with this method is that it can estimate the AIC-value and provides efficient parameter estimation methods for both ARMA and GARCH models.

3.2.1 Model diagnostics

Once the models have been fitted there are several methods to select the most suitable one. In this thesis the AIC is used as a comparable measure of the models suitability. The models with the lowest values are the most suitable. The AIC is calculated as,

AIC = -2log (maximum likelihood) + 2k, **(17)**

where k is the number of parameters, p is the order of the AR (p)-process and q is the order of the MA(q)-process. (Akaike. 1973). It is a commonly used method for selecting the model order.

Further, the standardized residuals of each selected model is tested with the Ljung-Box and McLeod-Li test to examine whether the model is appropriate to use for the time series. The standardized residuals are in these large samples approximated as normal distributed due to the central limit theorem. To investigate the suitability of this approximation, the Jarque-Bera test is used. The standardized residuals are also illustrated with histograms and quantile-quantile plot (QQ-plot). Below, the Jarque-Bera and Ljung-Box tests are briefly described.

3.2.2 Jarque-Bera test

The Jarque-Bera test suggests if the standardized residuals are normally distributed or not. It takes in account that normal distribution has zero skewness and excess kurtosis, assuming that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is independently and identically distributed standardized residuals. g_1 is the sample skewness, g_2 is the sample excess kurtosis and n is the sample size (Cryer and Chan. 1998, 284-285). The test is based on the following statistic:

$$JB = \frac{ng_1^2}{6} + \frac{ng_2^2}{24} \mathrel{\widetilde{\in}} \chi^2(2) \tag{18}$$

If the above statistic is large as measured by the upper quantiles of the chi-square distribution, the following H_0 hypothesis is rejected in the favor of H_1.

H_0: Skewness and excess kurtosis are zero – normal distribution

H_1: Skewness or excess kurtosis are not zero – non-normal distribution

3.2.3 Ljung-Box test

The Ljung-Box testis used to test if the residuals of the model have autocorrelation. The squared autocorrelation of a samples residuals is used as the base for the test. Its autocorrelation function is defined:

$$\widehat{\rho_\iota}(k) = \frac{\sum_{t=k+1}^{n}(x_t-\bar{x})(x_{t-k}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}, k = 1,2 \ \ldots \ K, \tag{19}$$

where $\bar{x}$ is the sample mean and K = maximum lag. The Ljung-Box test is then based on the test statistic,

$$Q_{LB} = n(n+2)\left(\frac{\widehat{\rho_1^2}}{n-1} + \frac{\widehat{\rho_2^2}}{n-2} + \ldots + \frac{\widehat{\rho_K^2}}{n-K}\right) \mathrel{\widetilde{\in}} \chi^2(K-p-q) \tag{20}$$

If the Q_{LB} statistic is large as measured by the by the upper quantiles of the chi-square distribution the following H_0 hypothesis is rejected in the favor of H_1.

$H_0: Q_{LB} \leq \chi^2_{k-p-q}$ Residuals constitute a white noise

$H_1: Q_{LB} > \chi^2_{k-p-q}$ Residuals show some dependence – they are not a white noise.

If the null hypothesis is not rejected, the residuals of the autocorrelation function is considered as white noise and if it is rejected it indicates there is some dependence. (Cryer and Chan.1998, 183-184).

3.2.4 Homoscedasticity and heteroscedasticity

A homoscedastic time series has a constant variance independent from time, which is assumed in ARMA models. A heteroscedastic time series has a variable variance dependent on time, which is taken in account by the GARCH model. Heteroscedasticity is present in many financial data sets, including price for metal commodities. Therefore, GARCH are assumed to be more efficient than ARMA models (Cryer and Chan. 1998, 277).

3.2.5 McLeod-Li test

The McLeod Li- test is based on the Ljung-Box test. The difference between these two tests is that the McLeod-Li test tests the residuals autocorrelation of a squared series x_t^2, thereby easier detecting volatility clustering. First, the autocorrelation function of the squared series is defined:

$$\hat{\rho}_M(k) = \frac{\sum_{t=k+1}^{n}(x_t^2-\bar{x}^2)(x_{t-k}^2-\bar{x}^2)}{\sum_{t=1}^{n}(x_t^2-\bar{x}^2)^2}, k = 1,2 \ldots K, \quad \textbf{(21)}$$

where $\overline{x}^2$ is the squared sample mean and K = maximum lag. The McLeod-Li test statistic is then based on the same as for the Ljung-Box test:

$$Q_{ML} = n(n+2)\left(\frac{\widehat{\rho_1^2}}{n-1} + \frac{\widehat{\rho_2^2}}{n-2} + \ldots + \frac{\widehat{\rho_K^2}}{n-K}\right) \approx \chi^2(K-p-q). \quad \textbf{(22)}$$

After the Q_{ML} statistic is calculated the decision rule is the same as for the Ljung-Box and the Jarque-Bera test. If the Q_{ML} statistic is large as measured by the upper quantiles of the chi-square distribution the following H_0 hypothesis is rejected in the favor of H_1.

H_0: $Q_{ML} \leq \chi^2_{k-p-q}$ Still white noise - Homoscedastic lag

H_1: $Q_{ML} > \chi^2_{k-p-q}$ Volatility clustering - Heteroscedastic lag

This test visualizes if the standardized residuals are homoscedastic or heteroscedastic. (Chen. 2002). If the null hypothesis is not rejected the standardized residuals are assumed to be homoscedastic. If for some lags the result is significant then there is indication for heteroscedastic residuals. (Cryer and Chan. 1998, 282-283.)

3.3 Forecasting

3.3.1 Forecasting methods

Forecasting with ARMA and GARCH models is conducted through:

$$\hat{x}_{(l)} = \phi_1\hat{x}_{(l-1)} + \phi_2\hat{x}_{(l-2)} + \cdots + \phi_p\hat{x}_{(l-p)} + \Theta_0 - \Theta_1 E(e_{t+l-1}|x_1, x_2, \ldots, x_t) - \Theta_2 E(e_{t+l-2}|x_1, x_2, \ldots, x_t) - \cdots - \Theta_q E(e_{t+l-q}|x_1, x_2, \ldots, x_t) \quad \textbf{(23)}$$

where,

$$E(e_{t+l-q}|x_1, x_2, \ldots, x_t) = \begin{cases} 0 \; for \;\; j > 0 \\ e_{t+j} \; for \; j \leq 0 \end{cases} \quad \textbf{(24)}$$

The forecasting value is based on the estimated parameters in the model. Through those parameters a value for the next coming one and two up to (p,q) time lags in the future can be predicted (Cryer and Chan. 1998, 199-201). Both forecast values and running standard deviations are estimated in order to cross-validate and evaluate models rather than forecast future price return. The running standard deviation is used in order to compare the models efficiency in estimating the volatility of the series. The running standard deviations is repeatedly computed standard deviations for pairs of observations. This computation is repeated throughout the whole length of time series, the running standard deviation is defined as the standard deviation for x_i and x_{i+1}.

3.3.2 Model accuracy

To compare ARMA with GARCH, the accuracy of the estimated models is compared. The three measures used for this are MASE, sMAPE and correct pairs of sign. MASE is applied on both forecast values and running standard deviation. It is a scaled error measurement that compares the difference of forecasted values of the logarithmic price return, defined as $f_{t,}$ and observed historical values x_t with the residuals obtained from naïve forecasting. The same notation is used for sMAPE. When MASE is <1, the model gives on average, smaller error than the naïve method (Hyndman and Koehler. 2006).

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|x_t - f_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|x_t - x_{t-1}|}\right) \qquad \textbf{(25)}$$

The second measure utilized in this thesis is sMAPE, which shows the average error in percentage with a range from -200 % up to 200 %. sMAPE has a slightly different formula depending on the author. In this thesis, a variant of sMAPE defined by Makridakis (2000) is applied, written with the same notation as above. sMAPE has a disadvantage in that it does not show which direction the forecasting error occurs. Furthermore, it is not an applicable measure for negative forecast values. Due to this, it will only be applied for running standard deviation.

$$sMAPE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|x_t - f_t|}{(x_t + f_t)/2}\right) \times 100 \qquad \textbf{(26)}$$

The final measure will count the pairs with similar signs of observed historical value and forecast value. The measure also calculates the percentage of pairs with similar signs. If the efficient market hypothesis applies the probability for a correct result should equal the probability for an incorrect result. To test if the measured percentage is significantly separated from p = 0.5, a Z-test will be conducted, where $\hat{p} \in \text{Bin}(\text{n}, p)$. The numbers of observation can approximately be normally distributed under the central limit theorem. The Z-test is as follows:

$$Z = \frac{\hat{p}-p}{\sigma} \widetilde{\in} N(0, 1). \qquad (27)$$

According to the efficient market hypothesis:

$H_0: \hat{p} \leq \frac{1}{2}$,
$H_1: \hat{p} > \frac{1}{2}$.

If the null hypothesis is not rejected it means that market efficiency cannot be ruled out. If, however, the null hypotheses is rejected, it suggests there are arbitrage opportunities in predicting future prices.

4 Results

Presented below are the results from model fitting, diagnostics and, finally, measures of forecast accuracy. First, the ACF and PACF results are interpreted as indicators for model selection.

4.1 ACF and PACF

The silver logarithmic price return has a significant lag at 1 and 9 for both ACF and PACF. Other lags are barely significant. The similar results for ACF and PACF indicate that silver logarithmic price return should have the same order of p and q. Nickel logarithmic price return barely has any significant lags for either ACF or PACF. This indicates that it can be described as a white noise process. Copper's logarithmic price return has significant lags at 1 and 4 for the ACF. The PACF is significant at lag 1, 3 and barely at 4. This indicates that it ought to have an order of 4 for p and 3 or 4 for q. For further details, see appendix D.

4.2 ARMA Models

ARMA models are selected from 23 different combinations. The orders of q in the MA (q)-process range from 0-3, and the orders of p, in the AR (p)-process range from 0-5. When minimizing the AIC-value, these models are selected.

Table 2. Selected models of ARMA

	Silver		Nickel		Copper	
	Estimate	**Std. Error**	**Estimate**	**Std. Error**	**Estimate**	**Std. Error**
ϕ_1	-0.4182	0.1069	-0.5089	0.1504	0.4117	0.1761
ϕ_2	-0.8171	0.0744	-0.7469	0.1401	-0.0470	0.4538
ϕ_3	0.1507	0.0899	N/A	N/A	0.4735	0.4236
ϕ_4	N/A	N/A	N/A	N/A	0.1021	0.0266
θ_1	0.2899	0.1057	0.5319	0.1516	-0.4874	0.1765
θ_2	0.7678	0.0662	0.7456	0.1496	0.0743	0.4602
θ_3	-0.2417	0.0856	-0.0272	0.0246	-0.5055	0.4429
σ	0.02322	N/A	0.02511	N/A	0.01950	N/A

The silver ARMA (3, 3) and the nickel ARMA (2, 3) have standard deviations lower than its parameters. The copper ARMA (4, 3) has, on the other hand, standard deviations that, in some cases, are larger than its parameters which indicates that these parameter estimations are not efficient. All three model orders are coherent with their ACF and PACF results, where nickel is modeling white noise. Both models for silver and nickel have a zero-mean which is expected for a market with no arbitrage opportunities. Copper, on the other hand, has a mean of 0.003. For further details, see Appendix A.

The Jarque-Bera test suggests that the standardized residuals (ε_t) are not normally distributed, for all three models. Despite that the test suggests non-normality, the method of estimating the parameters, which is based on the normal likelihood, is still valid. This is due to the so-called *quasi-likelihood method*. This method is similar to the regression, where the normal likelihood yields the least square estimate that are also valid for non-normal models. The distribution is also visualized in the QQ-plot, where all the models render approximately the same results. Histograms are visually asymptotic normally distributed. For QQ-plots and Histograms see Appendix B.

Table 3. Standardized residuals tests for Silver

Test	Residuals		Statistic	P-value
Jarque-Bera Test	ε_t	Chi^2	327531.6	< 2.2e-16
Ljung-Box Test	ε_t	Q(10)	55.8875	2.155e-08
Ljung-Box Test	ε_t	Q(15)	74.7729	6.223e-10
Ljung-Box Test	ε_t	Q(20)	90.278	6.628e-11

The Ljung-Box test null hypothesis is rejected at significance level <1 % for Q (10), Q (15) and Q (20). This indicates that the standardized residuals are not considered as white noise. This means that the ARMA model does not explain the times series efficiently.

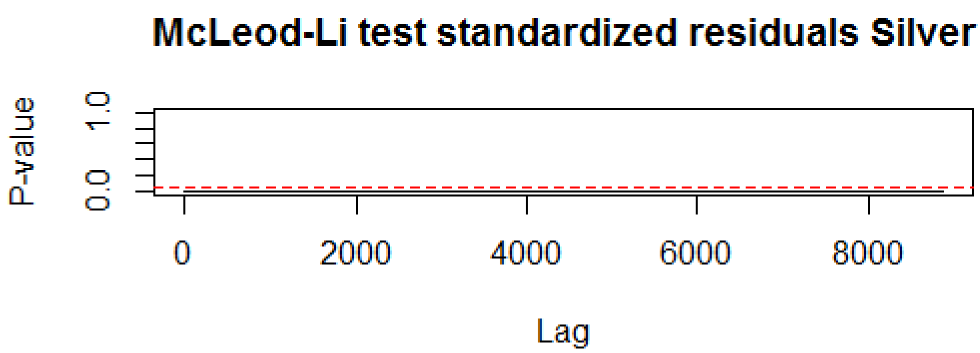


Figure 3. McLeod-Li test for standardized residuals of ARMA (3, 3) for silver price return.

The McLeod-Li test indicates that the standardized residuals are heteroscedastic. The null hypothesis is rejected for all lags at significance level <5 %. This confirms that the ARMA model does not account for heteroscedastic data.

Table 4. Standardized residuals tests for Nickel

Test	Residuals		Statistic	P-value
Jarque-Bera Test	ε_t	Chi^2	1316.235	< 2.2e-16
Ljung-Box Test	ε_t	Q(10)	2.4172	0.992
Ljung-Box Test	ε_t	Q(15)	6.6753	0.966
Ljung-Box Test	ε_t	Q(20)	7.9319	0.9923

The Ljung-Box test null hypothesis is accepted for Q (10), Q (15) and Q (20). This indicates that the standardized residuals are considered as white noise. This means that the ARMA model explains the times series efficiently.

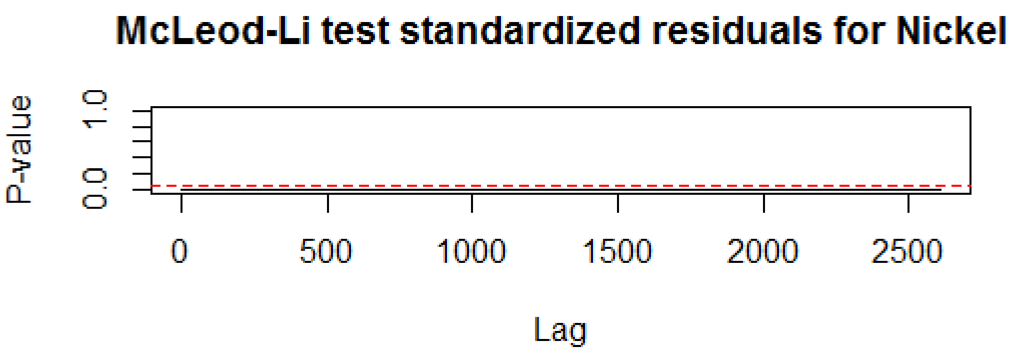


Figure 4. McLeod-Li test for standardized residuals of ARMA (2, 3) for nickel price return.

The McLeod-Li test indicates that the standardized residuals are heteroscedastic. The null hypothesis is rejected for all lags at significance level <5 %. This implies, again that the ARMA model does not account for heteroscedastic data.

Table 5. Standardized residuals tests for Copper

Test	Residuals		Statistic	P-value
Jarque-Bera Test	ε_t	Chi^2	1225.115	< 2.2e-16
Ljung-Box Test	ε_t	Q(10)	2.8704	0.9843
Ljung-Box Test	ε_t	Q(15)	9.1196	0.8712
Ljung-Box Test	ε_t	Q(20)	12.6074	0.8936

The Ljung-Box test null hypothesis is accepted for Q (10), Q (15) and Q (20). This indicates that the standardized residuals are considered as white noise. This means that the ARMA model explains the times series efficiently.

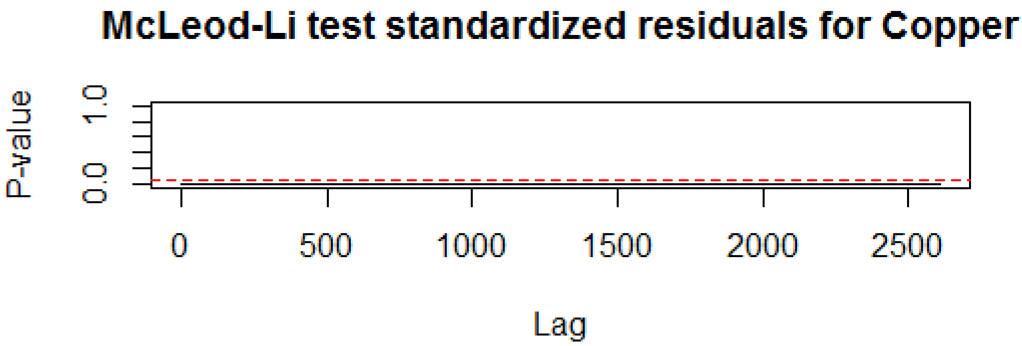


Figure 5. McLeod-Li test for standardized residuals of ARMA (4, 3) for copper price return.

The McLeod-Li test indicates that the standardized residuals are heteroscedastic. The null hypothesis is rejected for all lags at significance level <5 %. Again, the ARMA does not account for heteroscedastic data.

4.3 GARCH Models

GARCH models are selected from 168 combinations. The GARCH (p, q, u, v)-process ranges from 0-5 for both p, q and 0-2 for both u and v, where its conditional distribution is set to normal distribution. The five models with the lowest AIC-value for each time series are presented in Appendix A

Table 6. Selected GARCH model for Silver

	Estimate	Std. Error	t-value	Pr(>|t|)
μ	-4.309e-04	4.750e-04	-0.907	0.364347
ϕ_1	-0.2062	9.486e-02	-2.174	0.029688
ϕ_2	-0.4626	2.975e-02	-15.547	< 2e-16
ϕ_3	0.5464	5.317e-02	-10.276	< 2e-16
ϕ_4	-0.5935	6.939e-02	--8.553	< 2e-16
ϕ_5	-0.2567	6.718e-02	-3.821	0.000133
θ_1	0.1190	9.728e-02	1.224	0.221086
θ_2	0.4710	2.964e-02	15.891	< 2e-16
θ_3	0.5062	5.946e-02	8.515	< 2e-16
θ_4	0.5747	7.203e-02	7.978	1.55e-15
θ_5	0.2095	6.839e-02	3.064	0.002188
ω	1.061e-05	1.331e-06	7.971	1.55e-15
α_1	0.1777	1.049e-02	16.943	< 2e-16
β_1	0.1926	3.398e-02	5.669	1.44e-08
β_2	0.6243	3.454e-02	18.075	< 2e-16

All parameters for the ARMA (5, 5)-GARCH (1, 2) are significant at 5 % with exception of θ_1. The parameters that are not significant are kept because omitting orders of the MA (q)-process results in a higher AIC-value. The parameters of GARCH (u, v) are all significant which is important because they model the conditional variance and facilitates the comparison between ARMA and GARCH.

Table 7. Selected GARCH model for Nickel

	Estimate	Std. Error	t-value	Pr(>|t|)
μ	1.430e-04	2.440e-04	0.586	0.558018
ϕ_1	-5.698e-01	5.844e-02	-9.751	< 2e-16
ϕ_2	8.373e-02	1.131e-01	0.740	0.459224
ϕ_3	-1.614e-01	1.086e-01	-1.486	0.137212
ϕ_4	2.896e-01	8.768e-02	3.303	0.000957
ϕ_5	8.150e-01	8.194e-02	9.947	< 2e-16
θ_1	5.799e-01	6.199e-02	9.355	< 2e-16
θ_2	-8.181e-02	1.188e-01	-0.689	0.491056
θ_3	1.414e-01	1.201e-01	1.177	0.239153
θ_4	-2.817e-01	9.165e-02	-3.074	0.002110
θ_5	-7.909e-01	9.028e-02	-8.760	< 2e-16
ω	5.496e-06	1.878e-06	2.926	0.003431
α_1	6.493e-02	9.780e-03	6.639	3.15e-11
β_1	9.275e-01	1.095e-02	84.739	< 2e-16

Four parameters $\phi_2, \phi_3, \theta_2, \theta_3$ are not significant at 5 %. All other parameters are significant. However excluding these parameters results in a higher AIC-value. Due to this, the model is selected. The parameters for conditional heteroscedasticity are significant, which again, is important for the comparison between ARMA and GARCH.

Table 8. Selected GARCH model for Copper

	Estimate	Std. Error	t-value	Pr(>|t|)
μ	1.803e-03	3.765e-07	4788.570	< 2e-16
ϕ_1	-4.092e-01	1.520e-05	-26921.552	< 2e-16
ϕ_2	-8.343e-01	1.642e-05	-50806.678	< 2e-16
ϕ_3	1.463e-01	1.651e-05	8859.930	< 2e-16
ϕ_4	5.034e-02	1.555e-05	3236.865	< 2e-16
ϕ_5	-4.421e-02	1.553e-05	-2846.798	< 2e-16
θ_1	3.578e-01	1.668e-05	21445.319	< 2e-16
θ_2	8.280e-01	1.790e-05	46259.102	< 2e-16
θ_3	-2.475e-01	1.685e-05	-14685.392	< 2e-16
θ_4	-6.563e-02	1.652e-05	-3973.318	< 2e-16
ω	2.664e-06	9.698e-07	2.747	0.00602
α_1	6.258e-02	1.627e-02	84.739	0.00012
β_1	9.309e-01	3.228e-01	2.884	0.00393
β_2	1.000e-08	3.083e-01	0.000	1.00000

All parameters are significant, with the exception of β_2. The p-value of 1 for β_2 is due to that the maximum numbers of iterations was reached before convergence of the probability density function. The R package `fGarch' was used and the number of iterations was increased (in the package the function `garchFit' is used to fit models) with no effect. When fitting other models the AIC was increased, due to this the model was selected.

The Jarque-Bera test suggests that the standardized residuals (ε_t) are not normally distributed for all GARCH models. It appears to be a small improvement in the QQ-plots. The Histograms are the same as for ARMA. The quasi-likelihood method for parameter estimation still applies.

Table 9. Standardized residuals tests for Silver

Test	Residuals		Statistic	P-value
Jarque-Bera Test	ε_t	Chi^2	216651.7	0
Ljung-Box Test	ε_t	Q(10)	13.08905	0.2187354
Ljung-Box Test	ε_t	Q(15)	20.44664	0.1554603
Ljung-Box Test	ε_t	Q(20)	29.19832	0.0839272

The Ljung-Box test null hypothesis is accepted for Q (10), Q (15) and Q (20). This indicates that the standardized residuals are considered as white noise which means that the model is sufficient. This result is different from the selected ARMA model for silver, where the null hypothesis is rejected.

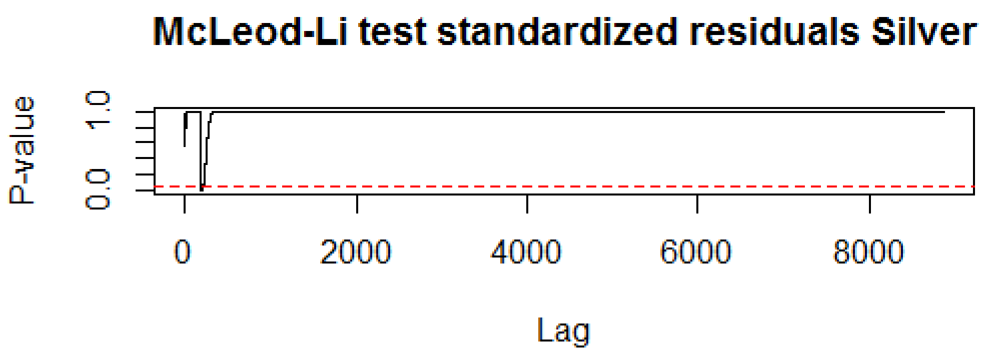


Figure 6. McLeod-Li test for standardized residuals of ARMA (5, 4), GARCH (2, 1) for silver price return.

The McLeod-Li test indicates that the standardized residuals are homoscedastic, accepting the null hypothesis for almost all lags. This is different from the selected ARMA model where all standardized residuals reject the null hypothesis. It is clear that the GARCH model has a better fit than ARMA.

Table 10. Standardized residuals tests for Nickel

Test	Residuals		Statistic	P-value
Jarque-Bera Test	ε_t	Chi^2	151.9799	0
Ljung-Box Test	ε_t	Q(10)	4.022492	0.9463266
Ljung-Box Test	ε_t	Q(15)	5.95384	0.9805127
Ljung-Box Test	ε_t	Q(20)	9.200908	0.9804613

The Ljung-Box test null hypothesis is accepted for Q (10), Q (15) and Q (20). This indicates that the standardized residuals are considered as white noise. This means that the model is sufficient.

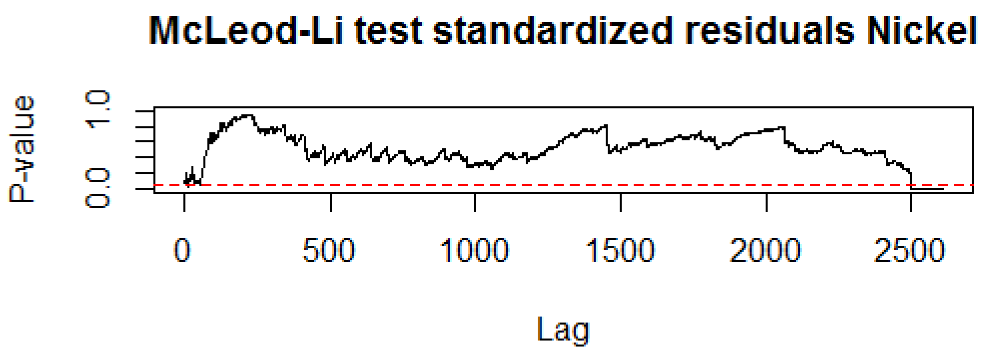


Figure 7. McLeod-Li test for standardized residuals of ARMA (5, 5), GARCH (1, 1) for nickel price return.

The McLeod-Li test indicates that the standardized residuals are homoscedastic, accepting the null hypothesis for almost all lags. This is different from the selected ARMA model where all standardized residuals reject the null hypothesis. This model is therefore more suitable.

Table 11. Standardized residuals tests for Copper

Test	Residuals		Statistic	P-value
Jarque-Bera Test	ε_t	Chi^2	430.0639	0
Ljung-Box Test	ε_t	Q(10)	9.757793	0.4619939
Ljung-Box Test	ε_t	Q(15)	14.01829	0.5241428
Ljung-Box Test	ε_t	Q(20)	18.60587	0.5475616

The Ljung-Box test null hypothesis is accepted for Q (10), Q (15) and Q (20). This indicates that the standardized residuals are considered as white noise. This means that this model is also sufficient.

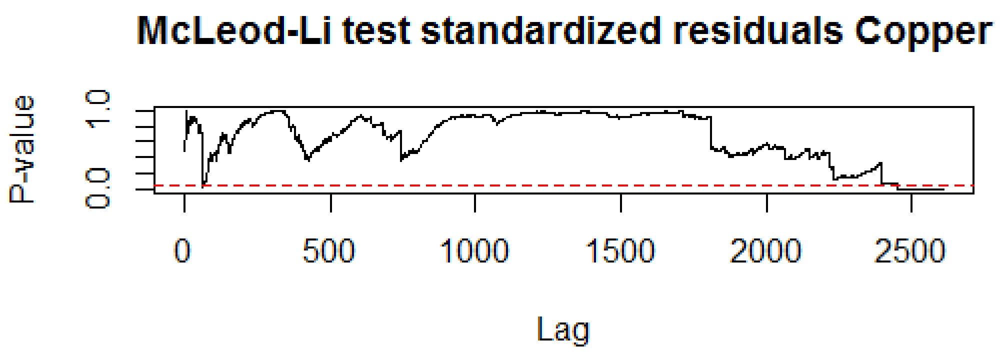


Figure 8. McLeod-Li test for standardized residuals of ARMA (5, 4), GARCH (1, 2) for copper price return.

The McLeod-Li test indicates that the standardized residuals are homoscedastic, accepting the null hypothesis for almost all standardized residuals. This is different from the selected ARMA model where all standardized residuals reject the null hypothesis. It is clear that this model is preferred for this time series.

For visualizations of the QQ-plots and Histograms for the standardized residuals of the GARCH models, see Appendix B.

4.4 Model forecasting accuracy

Table 12.	**Forecast values(f_t)**
Model	**MASE** t_{+1}
ARMA Silver	0.65247
GARCH Silver	0.65093
ARMA Nickel	0.7027278
GARCH Nickel	0.6989116
ARMA Copper	0.65763
GARCH Copper	0.651093

All models are more effective than the naïve method. ARMA and GARCH models have approximately the same result for one-step ahead forecast values.

Table 13.	**Running standard deviation(s_r)**	
Model	**MASE**	**sMAPE**
ARMA Silver	1.180883	136.5816
GARCH Silver	0.9976456	73.98603
ARMA Nickel	1.233946	164.1894
GARCH Nickel	0.9633961	49.60723
ARMA Copper	1.292256	150.6513
GARCH Copper	0.9047982	67.93425

The Running standard deviation has a lower value for each GARCH model compared to ARMA, for both MASE and sMAPE. This result strongly suggests that GARCH models are more efficient than ARMA models for these time series. For graphic illustration for both forecast values and running standard deviation, see Appendix C.

Table 14.	**Correct pairs of observed signs**	
Model	Correct pairs	z-value
ARMA Silver	0.51206	2.271638*
GARCH Silver	0.49932	-0.1280857
ARMA Nickel	0.48313	-1.723054
GARCH Nickel	0.50383	-0.06945327
ARMA Copper	0.47124	-2.937465*
GARCH Copper	0.551	5.208995**

*significant at 5 %
** significant at 1 %

The results for the correct pairs of signs are varying. There are only three models which are significantly better at predicting positive and negative values correct. The ARMA model for silver is significant at 5 % level, with 51.206 % and the GARCH model for copper is significant at 1% level with 55.1 % correct predictions. If the ARMA model for copper is interpreted in reverse, it has correct prediction at 52.876 % and is significant at 5 % level.

5 Conclusion

For forecast values ARMA and GARCH are equal in accuracy. However, the GARCH is more efficient at forecasting the running standard deviation. If only forecast values are demanded, it is not necessary to use heteroscedastic models. However, for financial data, efficiency is often seen as crucial and then models accounting for heteroscedasticity are needed. The efficient market hypothesis is not rejected for nickel and silver. Copper, on the other hand, seems to reject this hypothesis.

The Ljung-Box test gives a false reassurance when suggesting that the ARMA models are sufficient for nickel and copper. This is contrary to all other test results and measures. The McLeod-Li test strongly suggests that GARCH should be used for all three time series. When measuring the forecast accuracy of the running standard deviations the GARCH is more effective than ARMA, in all three time series, for both MASE and sMAPE. For forecast values, ARMA and GARCH are equally accurate according to MASE. It can be concluded that GARCH models are more accurate than ARMA and that models accounting for heteroscedastic data are necessary for metal prices. Also, that the relevance of the Ljung-Box test as an indicator of most efficient model is questionable.

The efficient market hypothesis seems to apply for nickel, where correct pair of sign does not yield a significant result. The ACF and PACF also suggests that nickel's logarithmic price return should be considered as white noise. This means that the logarithmic price of nickel can be described as a random walk process with no arbitrage opportunities. For silver, the pairs of sign is significant only for ARMA. However, with such a large sample size, rejecting the null hypothesis becomes more probable. The ACF and PACF indicates that the silver logarithmic price return is not white noise. These results are inconclusive to reject the efficient market hypothesis. For copper, the test using correct pair of sign is significant for both ARMA and GARCH, even if the results of the ARMA model needs to be interpreted in reverse. The ACF and PACF indicates that logarithmic price return is not white noise. Copper's logarithmic price return suggests arbitrage opportunities and, thereby, rejects the efficient market hypothesis.

The selected models are the ones which minimize AIC, but the difference between the AIC values for the models are only marginal for all metals. As an example, the difference between the highest and lowest AIC, for the ARMA models for silver, is about 0.005 %. The order of ARMA and GARCH does not, in terms of AIC seem to make a crucial difference. It is possible that other criteria for selecting models would render a more thorough analysis.

Another approach to compare efficiency of ARMA and GARCH models would be to conduct a quantitative simulation study. In this thesis only three empirical time series are used. One can apply our methodology on a larger number of different metals, which would yield a more general result. Another angle of incidence would be to further analyze and assume another distribution of the residuals when fitting models. By using the likelihood method, based on normally distributed residuals, for estimating parameters, the parameters estimates are correct but not necessarily effective when the residuals are not normally distributed. For a more thorough test of the efficient market hypothesis other distributions should be tried to achieve more efficient models. Also, only simple GARCH models are examined. For further research, it would be relevant to compare several types of GARCH models such as the EGARCH, QGARCH and GJR-GARCH.

6 Reference

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving-average models. *Biometrika,* 60 (2).

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics.* 31.

Box, G. and Jenkins, G. (1974). *Time series analysis: Forecasting and Control*. San Francisco: Holden-Day.

Chen, Y-T. (2002). On the Robustness of Ljung−Box and McLeod−Li Q Tests: A Simulation Study. *Economics Bulletin*. 3 (17).

Cryer, D. J. and Chan, K-S. (1998). *Time Series Analysis: with applications in R.* 2nd Edition. Springer

Enders, W. (2009). *Applied Econometric Time Series. 3e Edition*. Iowa State University. John Wiley Song.

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

Hamilton, D. J. (1994). *Time series analysis*. New Jersey: Princeton University Press.

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

Makridakis, S. and Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting* 16 (4).

Otto, S. (2010): Does the London Metal Exchange follow a random walk? Evidence from the predictability of futures prices. *The Open Economics Journal*. 3.

Thomson Reuters (2014). *Commodities information.* Available: http://extranet.datastream.com/data/commodities/Top30Benchmarkseries.htm (Last accessed: 2014-05-27)

Appendix A

ARMA AIC values

Table 15. Silver

AR/MA (AIC)	**0**	**1**	**2**	**3**
0	-41394.42	-41536.51	-41534.64	-41537.62
1	-41532,2	-41534.58	-41533.15	-41536.51
2	-41536.18	-41550.11	-41536.83	-41563.8
3	-41538.45	-41548.12	-41546.13	-41565.37
4	-41536.94	-41546.13	-41544.22	-41542.64
5	-41535.1	-41533.01	-41531.08	-41559.6

Table 16. Nickel

AR/MA (AIC)	**0**	**1**	**2**	**3**
0	-11799.6	-11798.99	-11798.54	-11799.64
1	-11798.92	-11797.37	-11797.27	-11798.33
2	-11798.84	-11797.56	-11800.87	-11802.47
3	-11799.56	-11798.48	-11802.4	-11800.8
4	-11800.22	-11798.71	-11798.59	-11798.55
5	-11799.63	-11798.31	-11799.46	-11794.55

Table 17. Copper

AR/MA (AIC)	**0**	**1**	**2**	**3**
0	-13099.94	-13112.16	-13110.26	-13110.93
1	-13111.96	-13110.37	-13108.19	-13112.46
2	-13110.09	-13108.42	-13106.51	-13114.71
3	-13112.1	-13112.83	-13113.48	-13109.51
4	-13114.71	-13112.71	-13111.81	-13118.54
5	-13112.71	-13110.71	-13110	-13116.45

GARCH AIC values

Table 18. Silver

ARMA(p,q)-GARCH(u,v)	**AIC- value**
ARMA(5,5)-GARCH(1,2)	-5.148399
ARMA(4,5)-GARCH(1,2)	-5.148085
ARMA(1,5)-GARCH(2,2)	-5.148010
ARMA(5,1)-GARCH(2,2)	-5.147965
ARMA(1,5)-GARCH(1,2)	-5.147936

Table 19. Nickel

ARMA(p,q)-GARCH(u,v)	AIC- value
ARMA(5,5)-GARCH(1,1)	-4.723885
ARMA(5,5)-GARCH(1,2)	-4.723457
ARMA(5,5)-GARCH(2,1)	-4.723080
ARMA(5,5)-GARCH(2,2)	-4.722690
ARMA(4,2)-GARCH(1,1)	-4.721061

Tabl1 20. Copper

ARMA(p,q)-GARCH(u,v)	AIC- value
ARMA(5,4)-GARCH(1,2)	-5.3065
ARMA(5,5)-GARCH(1,2)	-5,3026
ARMA(4,5)-GARCH(2,1)	-5,2909
ARMA(2,5)-GARCH(2,1)	-5,2907
ARMA(5,5)-GARCH(2,1)	-5,2904

Appendix B

ARMA QQ-plot and Histogram

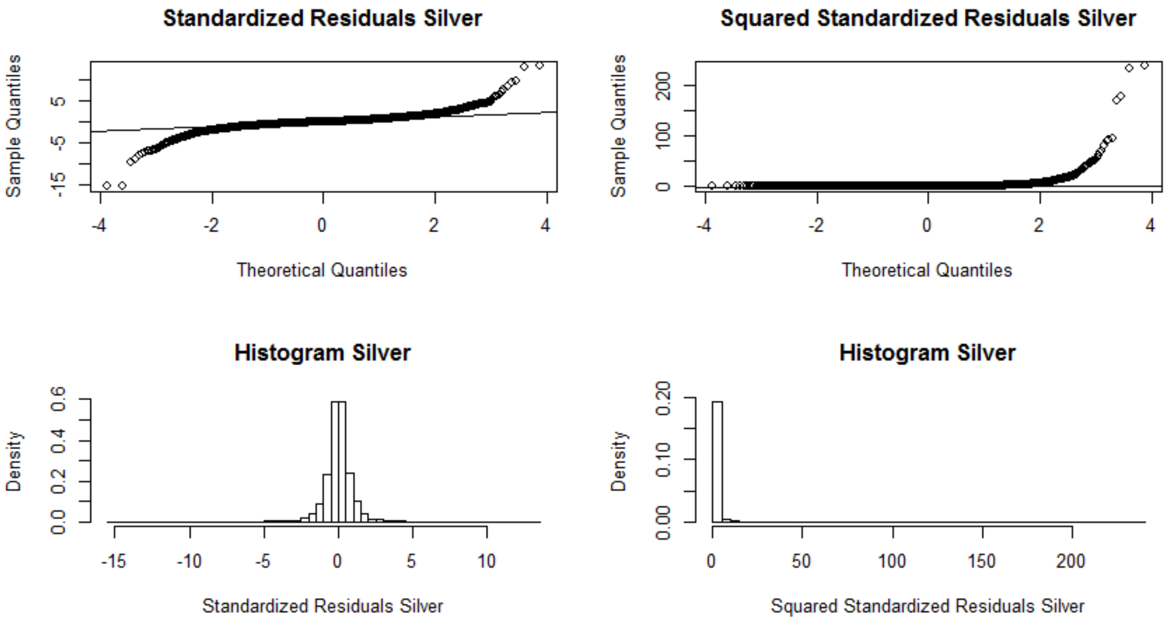


Figure 9. QQ-plot and Histogram for standardized residuals of ARMA (3, 3) for silver price return.

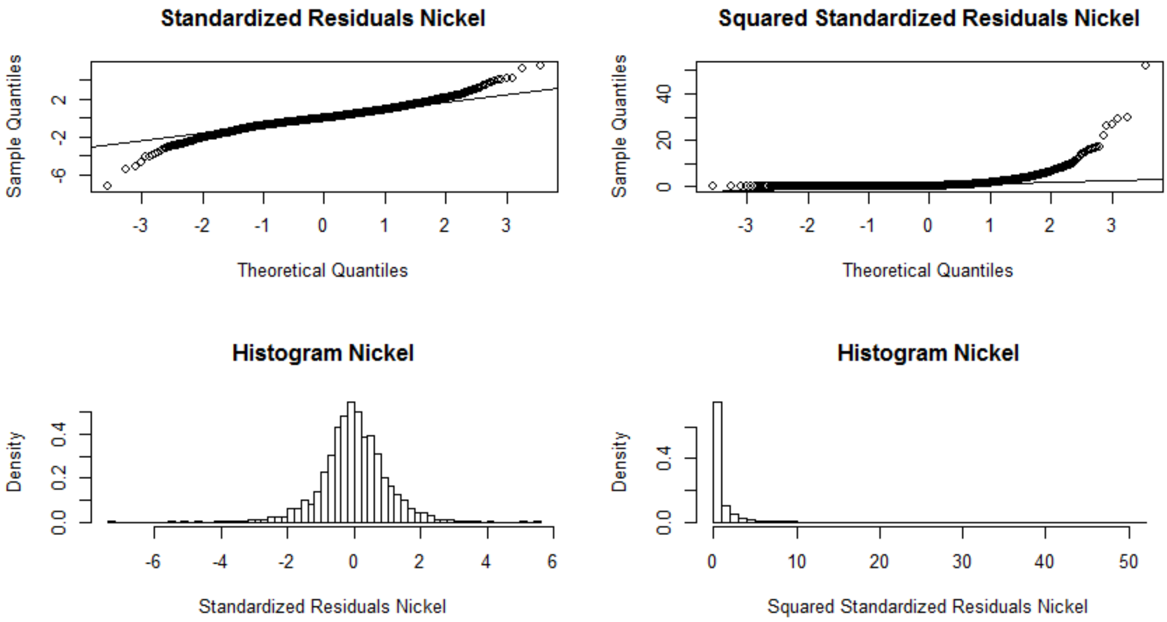


Figure 10. QQ-plot and Histogram for standardized residuals of ARMA (2, 3) for nickel price return.

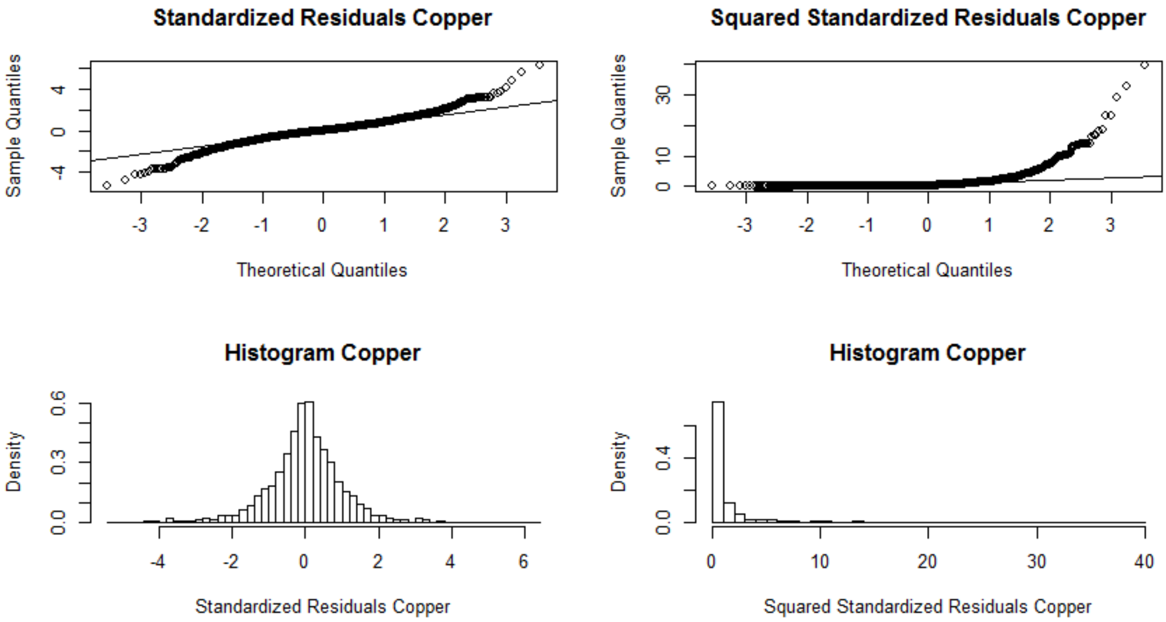


Figure 11. QQ-plot and Histogram for standardized residuals of ARMA (4, 3) for copper price return.

GARCH QQ-plot and Histogram

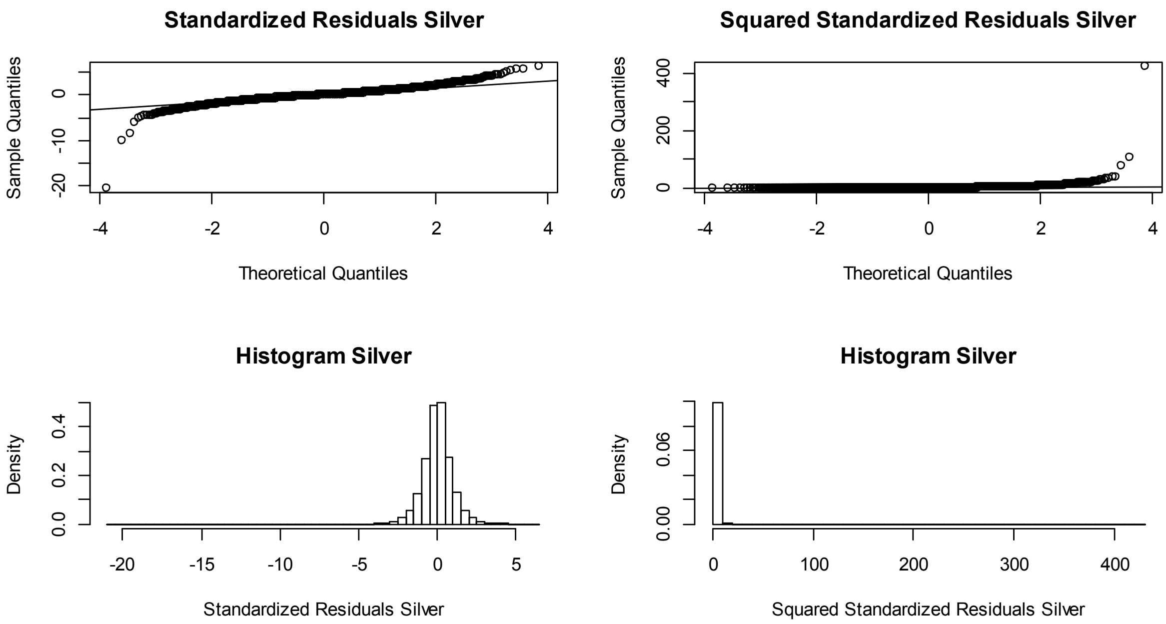


Figure 12. QQ-plot and Histogram for standardized residuals of ARMA (5, 4), GARCH (2, 1) for Silver price return.

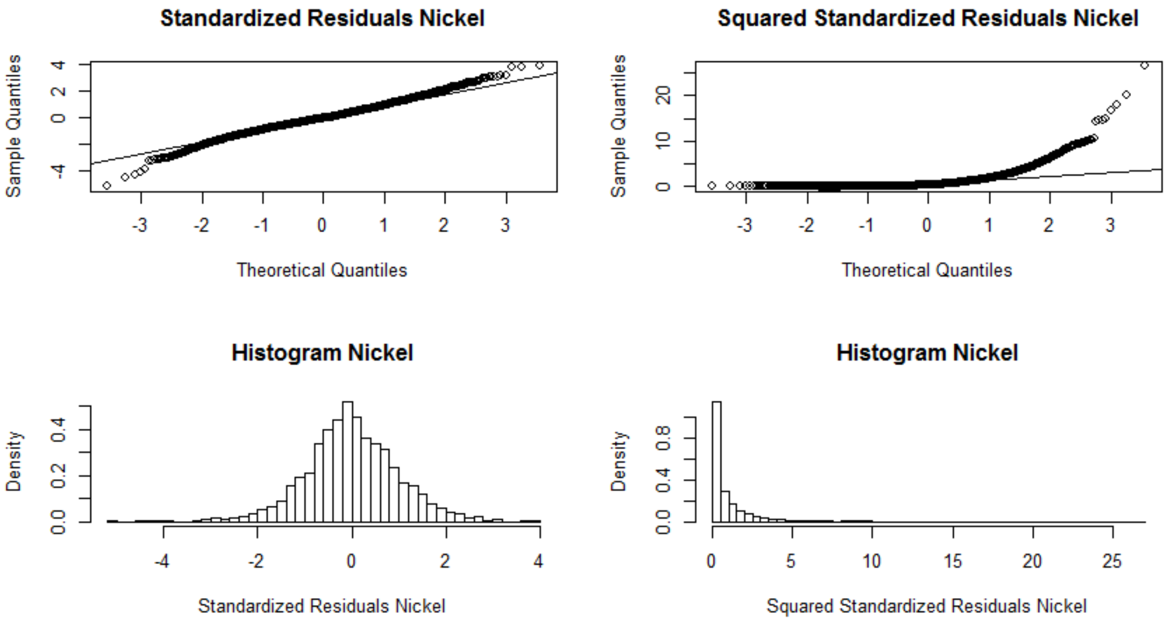


Figure 13. QQ-plot and Histogram for standardized residuals of ARMA (5, 5), GARCH (1, 1) for nickel price return.

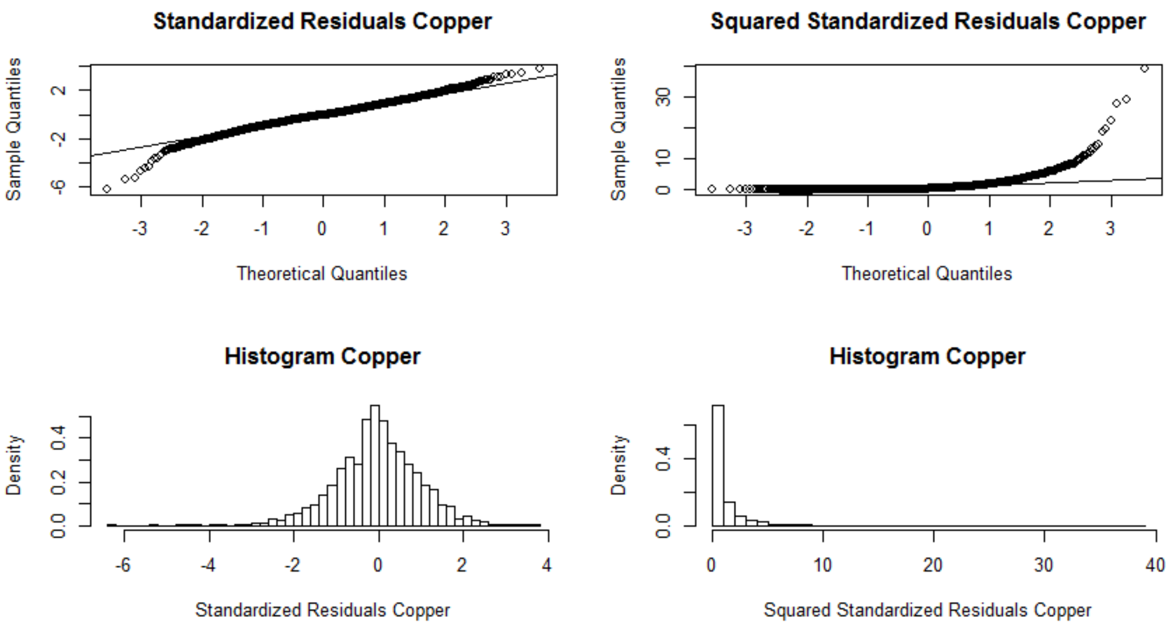
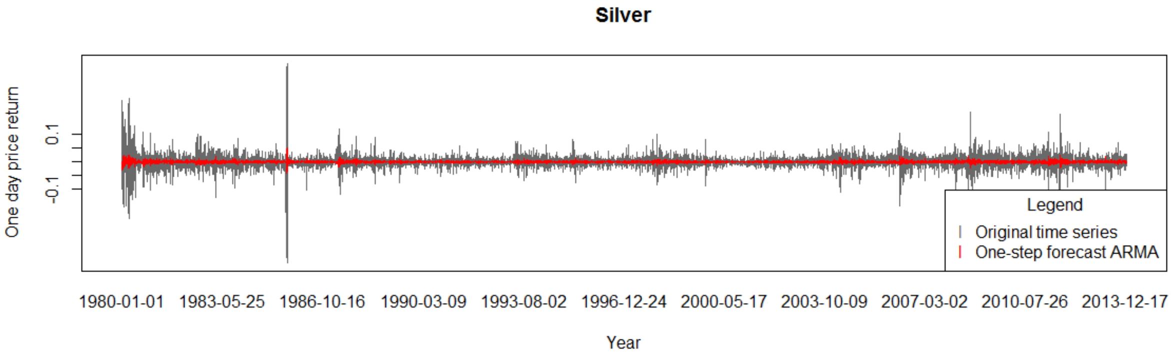
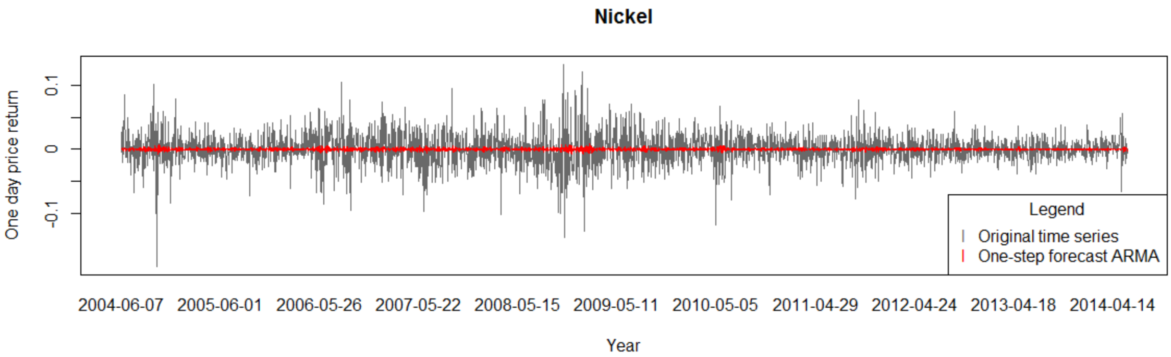


Figure 14. QQ-plot and Histogram for standardized residuals of ARMA (5, 4), GARCH (1, 2) for copper price return.

Appendix C

Visualization of one-step ahead forecast





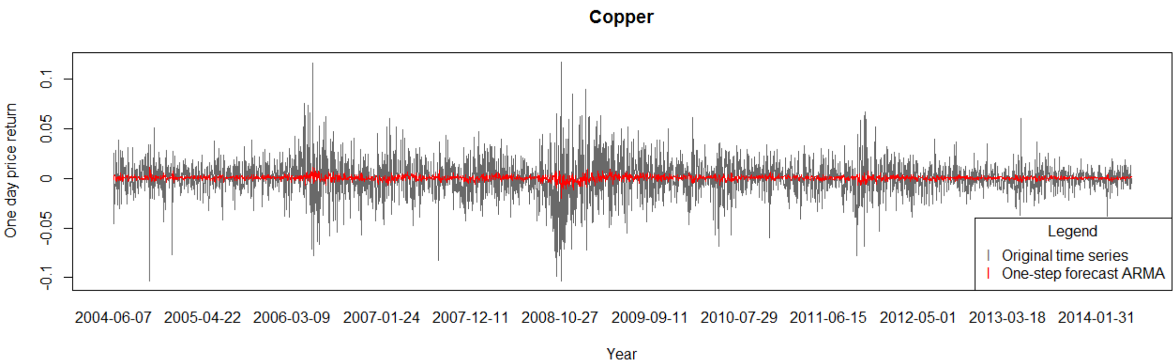


Figure 15. Logarithmic price return and one-step ahead point estimation for silver ARMA (3, 3), nickel ARMA (2, 3) and copper ARMA (4, 3).

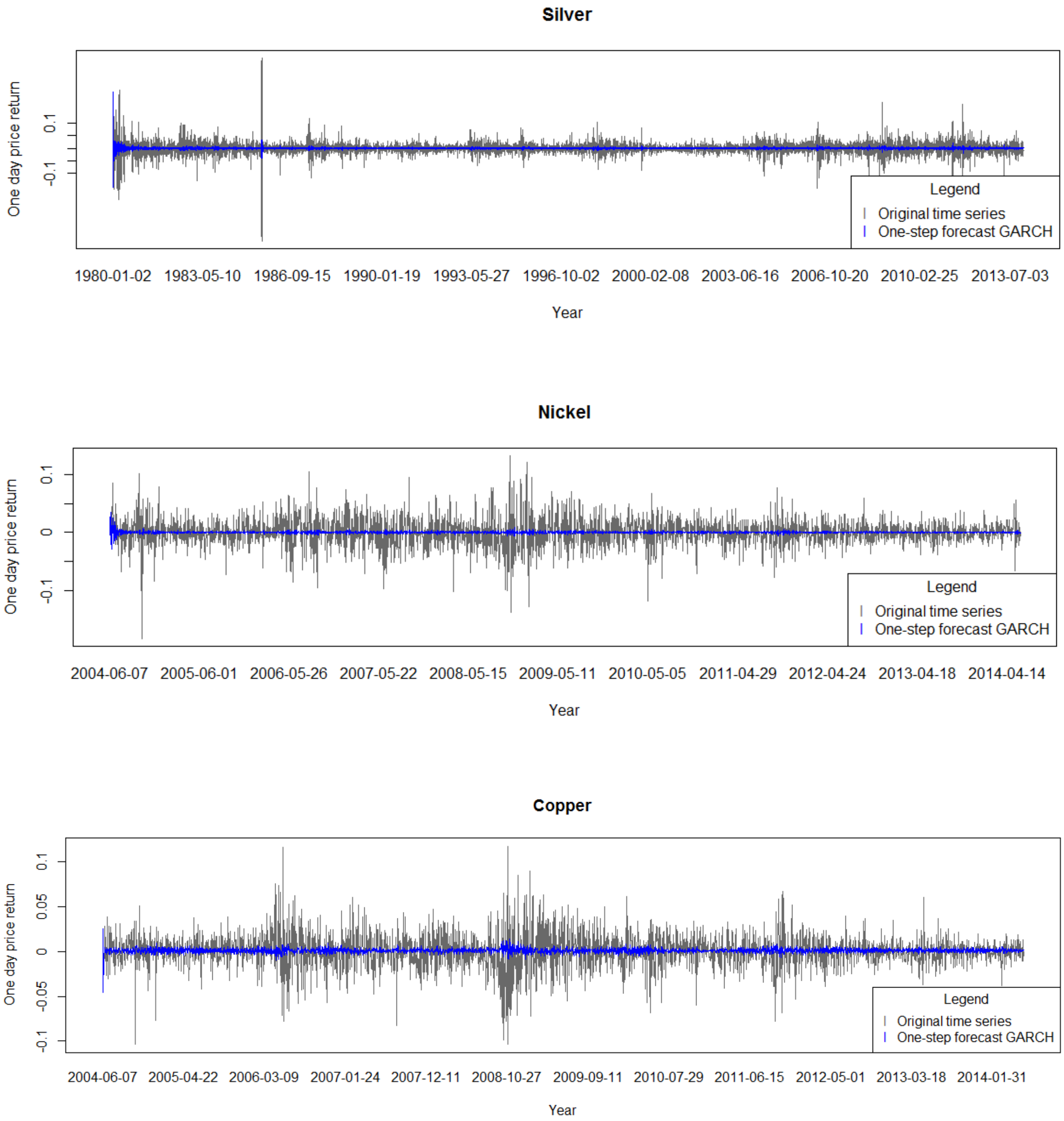


Figure 16. Logarithmic price return and one-step ahead point estimation for silver ARMA (5, 4) GARCH (2, 1), nickel ARMA (5, 5) GARCH (1, 1) and copper ARMA (5, 4) GARCH (1, 2).

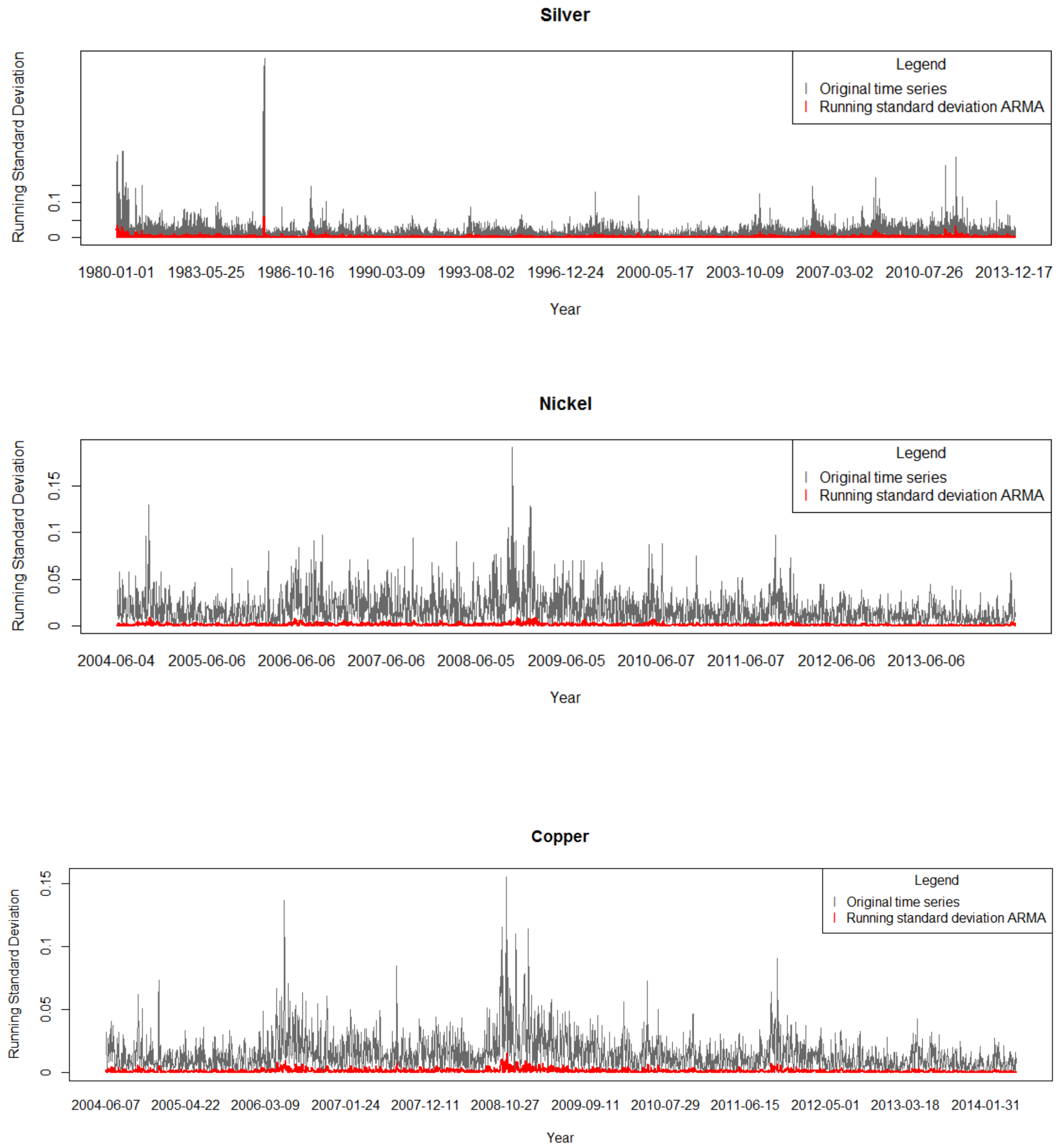
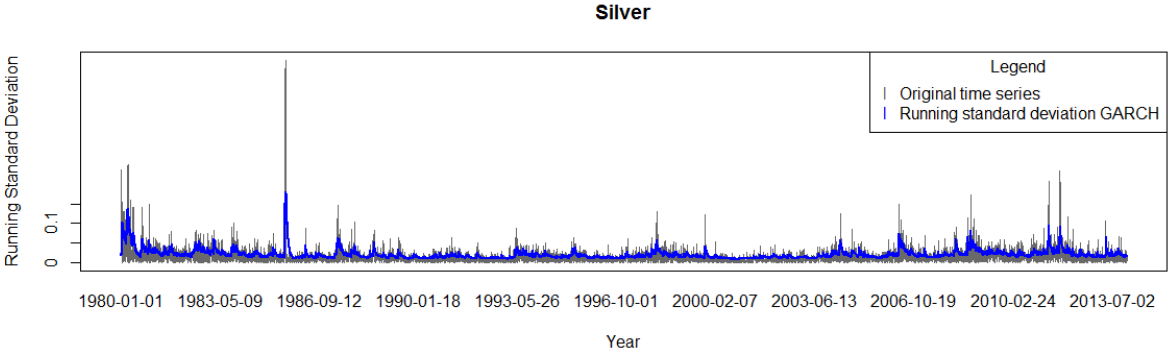
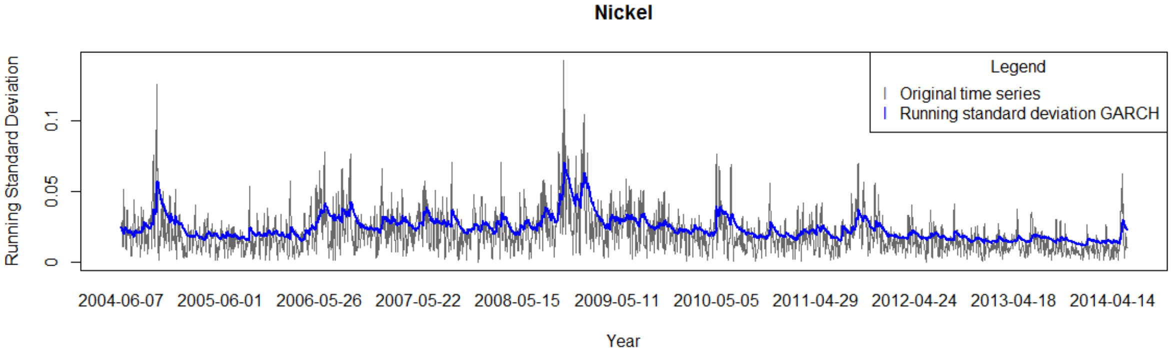


Figure 17. Running standard deviation for copper and one-step ahead forecasts for silver ARMA (5, 4) GARCH (2, 1), nickel ARMA (5, 5) GARCH (1, 1) and copper ARMA (5, 4) GARCH (1, 2).





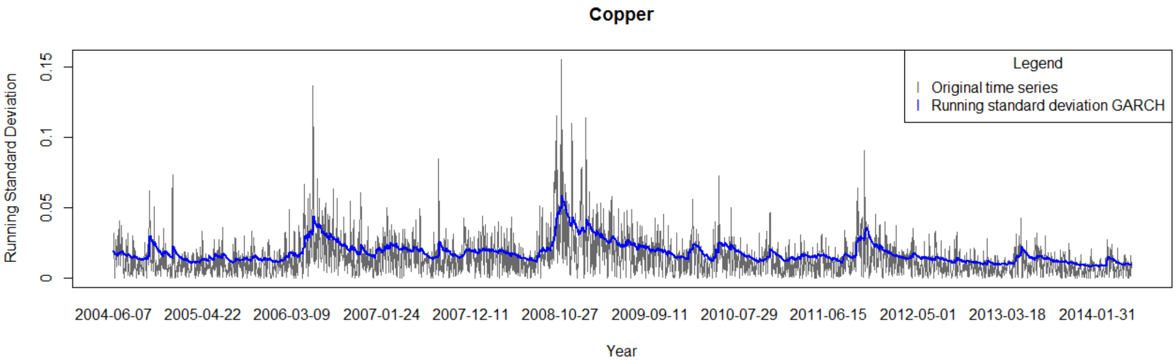


Figure 18. Running standard deviation for nickel and one-step ahead forecasts for silver ARMA (5, 4) GARCH (2, 1), nickel ARMA (5, 5) GARCH (1, 1) and copper ARMA (5, 4) GARCH (1, 2).

Appendix D Autocorrelation and Partial Autocorrelation for Price Return

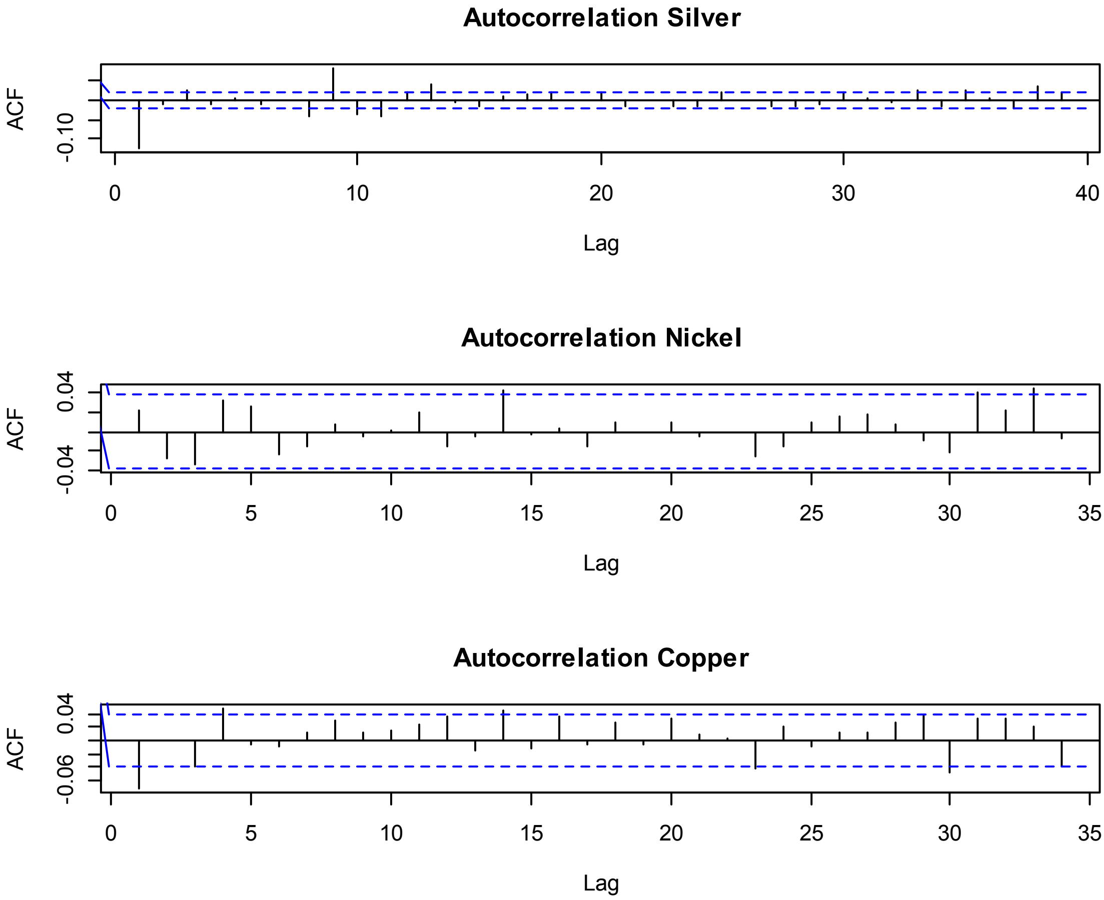


Figure 19. Autocorrelation function for price return.

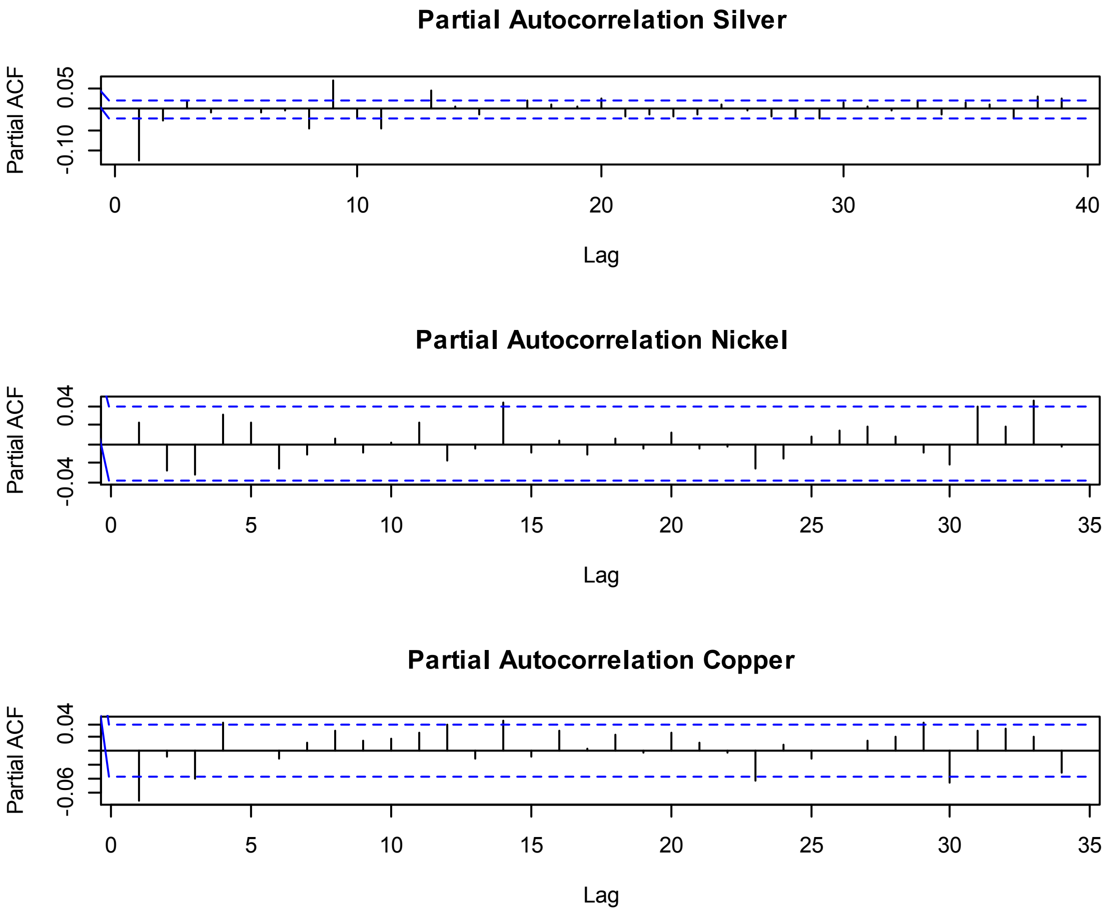


Figure 20. Partial autocorrelation function for price return.

Appendix E The simulated ARMA and GARCH

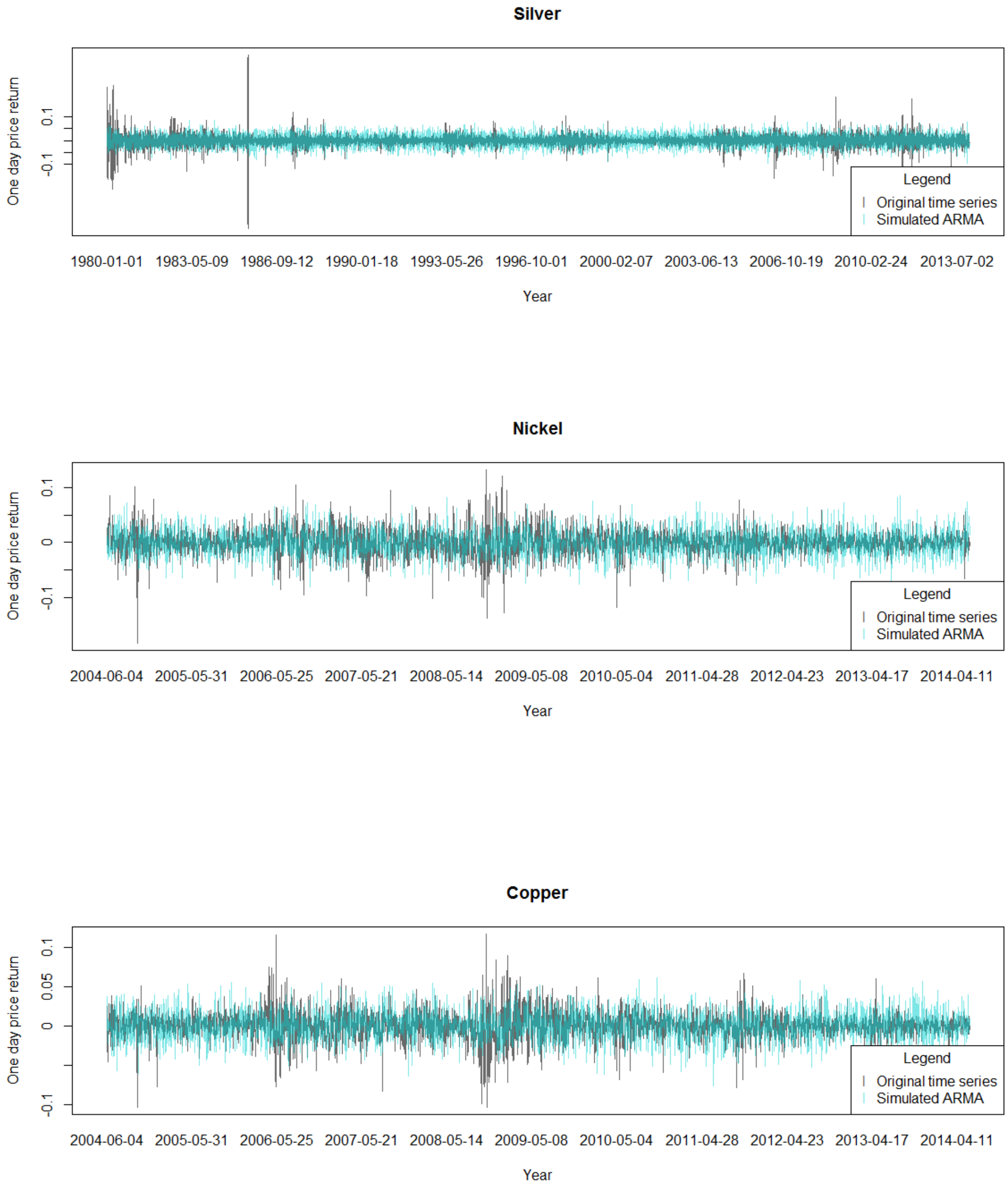


Figure 21. The simulated ARMA for silver, nickel and copper.

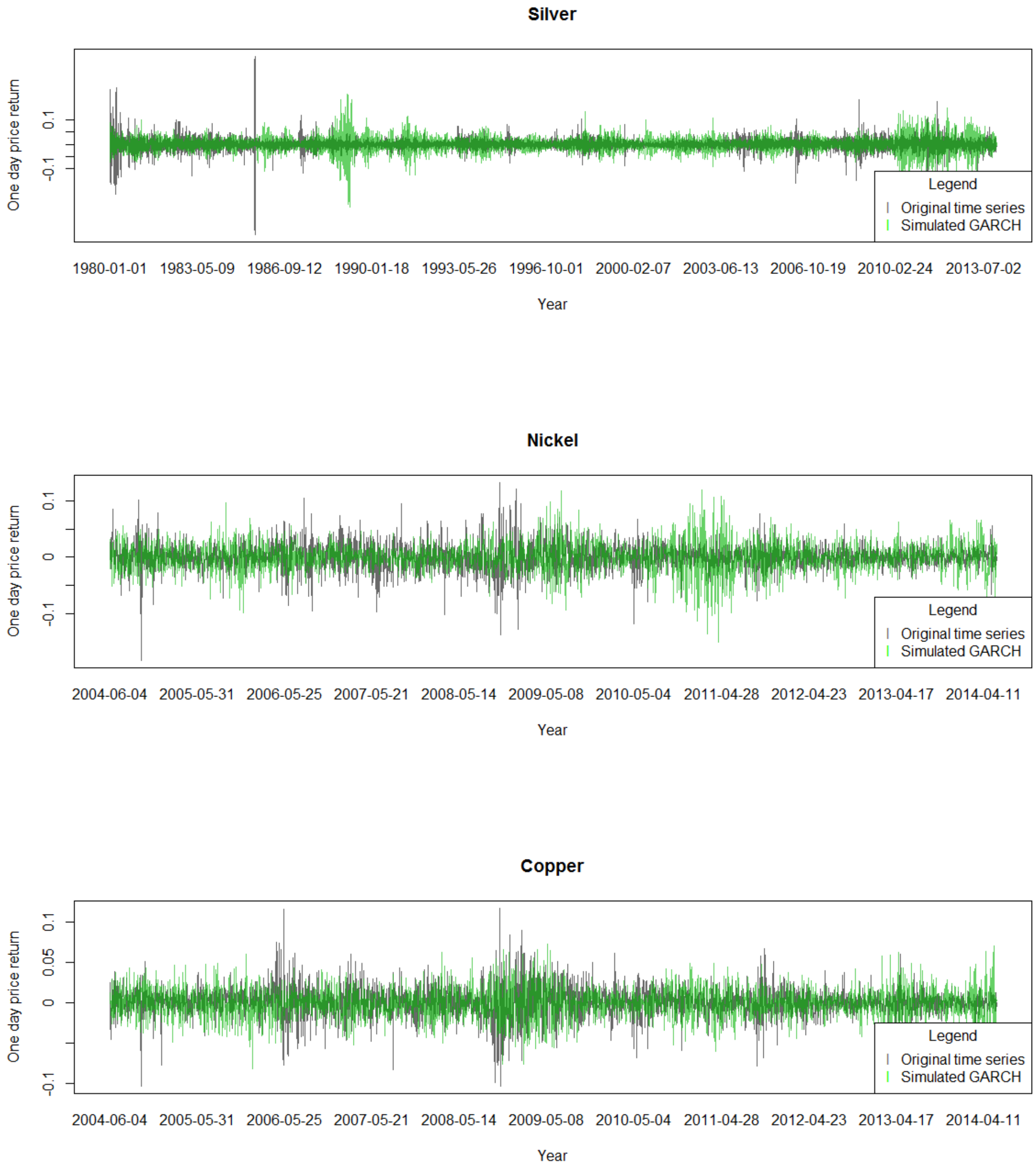


Figure 22. The simulated GARCH for silver, nickel and copper.