UNIVERSITY OF GOTHENBURG

SCHOOL OF BUSINESS, ECONOMICS AND LAW

Forecasting monthly LME Copper returns

Nils Lervik and Philip Thorsell
Supervisor: Marcin Zamojski

Master's thesis in Finance, 30 hec
Spring 2022
Graduate School, School of Business, Economics and Law, University of Gothenburg, Sweden

Abstract

We evaluate if monthly LOCADY returns on the London Metal Exchange can be accurately predicted one, two and three months ahead. In total ten models are constructed using time-varying parameters and bandwidth optimization. The models are evaluated against one another using the following pseudo-out-of sample test statistics: Diebold and Mariano (1995), Clark and West (2006), Giacomini and White (2006) and the Campbell and Thompson (2008) out-of-sample R^2. The test statistics generated are inconsistent. A few models are able to generate positive out-of-sample R^2 values for one and two month predictions. No model significantly outperforms a random walk for the three step ahead prediction.

Acknowledgments

We would like to thank our thesis supervisor Marcin Zamojski, as well as Erik Lehman, Jacob Mannheimer and the entire team at Prognos MKA AB, for their guidance and support throughout this project.

Contents

1 Introduction

Forecasting, as it pertains to commodity pricing is considered highly relevant for different types of market participants. In this thesis we evaluate the copper market as a proxy for metal commodity markets. The objective of the evaluation is to determine what models are able to forecast, and which model is best suited to forecast LME Copper returns. We implement various models such as AR(p), MA(q), ARMA(p,q) and VAR(p) to forecast future copper prices. We evaluate a range of the aforementioned forecasting model's ability to accurately predict out-of-sample observations in the short-, mid-, and long-term horizon. We define short term as one month ahead, mid-term as two months ahead, and long-term as three months ahead. The models are compared based on their ability to forecast commodity prices, specifically monthly spot copper prices from the London Metal Exchange, LME. Each model allows for time-varying parameters which we operationalize by estimating them in moving windows with an optimized bandwidth. We also create model ensembles and evaluate their performance. The models forecast returns rather than prices to accommodate the assumptions of stationarity related to the models and tests in question.

Forecasting of metal commodity prices is a frequently addressed topic in research. Copper has been the focus of a range of sophisticated forecasting models such as Liu et al. (2017) who apply a decision tree learning model, DTLM. They find the DTLM to reliably forecast the price of copper deviating less then 4% for horizons ranging from days to years. Further examples are Kriechbaumer et al. (2014) who suggest their wavelet-ARIMA model approach to be a promising technique for forecasting metal prices with high accuracy. Notably, Buncic and Moretto (2015) concludes that their dynamic model averaging and selection (DMA/DMS) approach yields an out-of-sample R^2 of 18.5% when forecasting monthly returns on LME copper prices. Furthermore, Buncic and Moretto (2015) found that a simple OLS method yields close to 10% out-of-sample R^2. The implications of a simple model having significant, and surprisingly large, explanatory power for what is assumed to be a highly liquid market with interest from a considerable amount of financial actors inspired us to attempt to evaluate simple models with less information than the 18 variables used by Buncic and Moretto (2015).

We contribute to existing literature through evaluating which type of model provides the better forecasts for commodity returns. Considering our range of more basic forecasting models the question arises as to whether one model consistently outperforms the rest. Subsequently, the applicability of the models is evaluated against a benchmark model. The Benchmark model follows a random walk to forecast future prices. We compare the performance of the model ensembles with the performance of the individual models. The aforementioned models and model ensembles are evaluated through the lens of four different tests: Clark and West (2006), Diebold and Mariano (1995), Giacomini and

White (2006) and Campbell and Thompson (2008) out-of-sample R^2.

Our thesis results are somewhat consistent with the findings in Buncic and Moretto (2015) as our limited models, using less information, are found to produce positive out-of-sample R^2 values, similarly to their simple OLS model. Overall we find that the ARMA(1,6) and one of our model ensembles outperforms the historical average in the short- and mid-term. In the long-term none of the models and model combinations outperform our benchmark model. The only model indicated to outperform the random walk in two out of four tests is the AR(1) model. However, the AR(1) model does not outperform the benchmark model in the short-term. Thus, none of the models consistently beats the random walk across all three horizons. The internal ranking of the models suggest the AR(1), ARMA(1,6), and TOP3 perform well according to two out of the four tests. The definite ranking between the three are however not consistent between tests. Subsequently, it is not clear that an ensemble consistently performs any worse or better than the individual models. The Diebold and Mariano (1995) test deviates from the other test statistics with far more conservative outputs finding very few significant differences in forecasting accuracy.

The remainder composition of our thesis is as follows. Section 2 outlines the theoretical framework upon which the thesis relays. Following the theoretical framework, Section 3 provides an explanation of how we implement the theoretical framework. It outlines the selection of data, models and tests as as well as how we have adapted relevant theory to construct our samples, models and tests. Section 4 analyses the test results, considering the performance of models, the discrepancies between the tests, and the relation to previous literature. The analysis and findings are summarized in Section 5 and suggestions to further research is detailed in Section 6.

2 Theory

In this section we begin by introducing return predictability as we build forecast models for asset returns, specifically monthly copper returns. We then proceed with introducing loss functions which are the foundation for how the forecast models are evaluated and as such also affects the construction of them. As the forecasting model is constructed there are a number of challenges that need to be dealt with. After introducing the challenges we include additional concepts utilized in the construction of our forecast models such as, window selection and stationarity. We proceed to introduce the different types of time-series models that are implemented in the thesis. Lastly, the tests we execute to evaluate, and compare, the models with one-another are outlined.

2.1 Return Predictability

Timmermann (2018) states that the most common prediction model for returns used in empirical studies is a simple linear model where r_{t+1} is equal to the excess return from holding an asset from t to $t+1$, i.e., one period

$$r_{t+1} = \mu + \beta x_t + \mu_{t+1}, \tag{1}$$

where $x_t \in \Omega_t$ is a prediction variable that is known at time t. This linear model in equation 1 can be derived from the general first order form in equation 2 under a few conditions.

$$E_t[m_{t+1} r_{t+1}] = 0, \tag{2}$$

where m_{t+1} is a pricing kernel. The pricing kernel that Cochrane (2009) use is a positively valued stochastic discounting factor. Conditional expectations given information at t are denoted by $E_t[\cdot] = E[\cdot|\Omega_t]$. The assumptions that need to hold for it to be possible to derive Equation (1) from Equation (2) as Cochrane (2009) does. The assumptions are that cash flows are formed as linear combinations of a finite-dimensional, stationary VAR, vector auto-regression, no arbitrage and no transaction costs. This derivation is shown in detail by Timmermann (2018) who achieves the same results as Cochrane (2009) but instead utilizes a log linear asset pricing model. This suggest, generally, that linear models are applicable for forecasting returns.

2.2 Loss Function

Constructing a prediction model that correctly forecasts the value at all times is impossible. Therefore, all prediction models face estimation errors and need a method to deal with them in different decision making processes. Loss functions have been developed to formulate the trade-off, or rather, the cost associated with a particular prediction error. A loss function is denoted as $L(\cdot)$ by Elliott and Timmermann (2016) and describes how costly an imperfect forecast, f, is in relation to an outcome, Y, and other data, Z .

It is pivotal that the trade-offs between different prediction errors are reflected accurately in the loss function. The choice of the loss function affects the parameter estimation and the evaluation of different forecast models. Typical loss functions are symmetric and one-tailed loss functions. A symmetric loss function penalizes positive and negative forecast errors equally whereas a one-tailed loss function only associates a loss with either a positive or negative forecast errors. It is important to choose a loss function that approximately depicts the most essential trade-off for the forecasting problem if not a specific loss function is constructed for the problem at hand (Elliott and Timmermann, 2016; Lee, 2008). An example of when a one-tailed loss function is suitable is when you consider if you should enter a call option contract by forecasting the spot price at the exercise date of the option. If the predicted spot price is below a certain level the payoff does not cover the cost of the contract and you would incur a loss. You are not as concerned with how large a profit you might make but rather that you make a profit. In this case you would want to construct the loss function so that a overestimation would incur a loss and a under prediction does not. Why does a under prediction not incur a direct cost? Since you would not enter the contract in the first place if your forecast did not generate a profit. However this level of reasoning fails to take opportunity costs in to account. A more suited loss function in this case would be a two-tailed function and depending on how you as an individual value realized losses against loss of opportunity a weighting scheme is fitted to the deviations.

It is assumed that the loss function is minimized when the forecast equals the outcome, $min_f L(f, Y, z) = L(y, y, z)$. Y denotes a continuous random variable, y is the set of all potential outcomes and f denotes a point forecast. The intuition behind that assumption is that the loss function is minimized when the forecast is perfect. A perfect forecast implies that there is no forecast error. When the loss function is not dependant on Z, which gives $L(f, Y, Z) = L(f, Y)$. Then the loss can be normalized giving the minimum value of the loss function at 0 (Elliott and Timmermann, 2016; Lee, 2007). Loss functions that only depend on the forecast errors take the form of $L(e)$ where $e = y - f$. These loss functions can be summarized in three requirements as suggested by Granger (1999):

$$L(0) = 0 \tag{3}$$

$$L(e) \geq 0 \quad \forall \quad e \tag{4}$$

$$L(e_1) \leq L(e_2) \quad if \quad e_2 < e_1 < 0 \quad and \quad L(e_1) \leq L(e_2) \quad if \quad e_2 > e_1 > 0 \tag{5}$$

Requirement one, Equation (3), is interpreted as; if the there is no forecast error the true value the loss function equals zero, i.e. no loss occurs. The second requirement in Equation (4) implies that there cannot be any negative loss. The second assumption is in place because profiting from having a deviation in the forecast would be incongruous to the goal of generating accurate predictions. The third requirement in Equation (5) states that a greater forecast error must incur a greater loss.

Loss functions can also take into account other properties such as symmetry, homogeneity, boundedness, or differentiability amongst others. We are interested in symmetry as our loss is not dependent on whether the forecast over- or underestimates the outcome, just how large the deviation is in absolute terms. We utilize a symmetrical loss function as we do not want to weight deviations differently from one-another and introduce a bias towards a certain type of estimation error in the model selection and creation. If a one-tailed loss function where to be implemented a one percent underestimation may be less costly than a two percent overestimation. This would result in a model, and model selection process that favors a model that rarely under estimates the return but more often overestimates the return. Symmetry of forecast errors is defined as in Equation (6) by Elliott and Timmermann (2016); Fildes and Makridakis (1988)

$$L(-e) = L(e). \tag{6}$$

2.3 Challenges to Forecasting

There are a number of challenges that arise when constructing a forecasting model. The following paragraphs cover the problems that arise due to weak predictors, persistent predictors, model instability and over-fitting.

Weak Predictors are explanatory variables with low forecasting power. A variable has low forecasting power if the signal to noise ratio is low. Weak predictors are common when forecasting financial asset returns as markets are highly competitive. If there is some explanatory power in a variable and it can be used for forecasting the price development, financial actors will use the explanatory power to trade until there is no profit to be made. Timmermann (2018) suggest that if there are weak predictors in the model the estimation errors are approximately of the same magnitude as the signals the variables omit. Timmermann (2018) continues to argue that tests such as Diebold and Mariano (1995) will be ineffective when evaluating predictive performance of returns in such cases. The aforementioned uncertainty associated with weak predictors creates difficulty in variable selection which may be difficult for conventional selection methods to detect.

For variable selection which in most of our models is equivalent with lag selection we use one out of two different information criteria. Heinze et al. (2018) suggest using this method which involves selecting a model from a set of potential models. They then proceed to introduce two different information criterions, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The AIC estimates the information loss under some distribution against the assumed true data generating process. AIC uses a maximum likelihood approach and rewards descriptive accuracy. However, Wagenmakers and Farrell (2004) present two drawbacks of the AIC. Firstly, the AIC can lead to overly optimistic assessments when the likelihood values are not highly concentrated around the maximum value. In other terms, the variability of the parameter is neglected. Secondly, the probability of AIC recovering a true low-dimensional model does not approach unity when the number of observations is very large. That is, consistency does not hold for AIC (Bozdogan, 1987). An alternative method to AIC is the BIC method. BIC allows for variability in the parameter and Wagenmakers and Farrell (2004) states that it is consistent as sample size approaches infinity. The main difference in assumptions is that BIC assumes that the true model exists in the set of models whereas the AIC does not (Wagenmakers and Farrell, 2004). Often AIC and BIC suggest similar lag orders. In this thesis we base our lag orders on BIC. We do so because we want to allow for variability in the parameter.

Persistent Predictors When forecasting for example stock returns forecasters face persistent predictors. A persistent predictor is one for which a small shock has a permanent effect on the parameter estimate. This affects future predictions beyond the point of relevance of the shock. In the case of copper prices, the relationship between different variables may change over time. If these relationships change, the model may suffer as it includes parameter estimates based on outdated information. In which case, the estimated model would be less suited to predict the future.

Stambaugh (1999) points out that such persistent predictors could lead to biases in the slope coefficients, but only if the innovation of the predictors have strong correlations between shocks and returns (Timmermann, 2018). Assuming that $E(u_t|x_s, x_w) \neq 0, s < t$, so that the residuals, u_t, in equation (1) are correlated with past or future values of x, $\sigma_{uv} = E[u_t v_t] \neq 0$. Using the aforementioned conditions, Timmermann (2018) shows that the finite-sample bias in $\hat{\beta}$ can be computed as:

$$E[\hat{\beta} - \beta] = (\sigma_{u\nu}/\sigma_{\nu}^2)E[\hat{\rho} - \rho] = (-\sigma_{u\nu}/\sigma_{\nu}^2)((1 + 3\rho)/T) + O(T^{-2}). \quad (7)$$

There is no finite-sample bias in $\hat{\beta}$ if u_t and v_t are uncorrelated, but if $\sigma_{u\nu} \neq 0$ Stambaugh (1999) shows that there can be a large finite-sample bias.

Model Instability Utilizing the same predictor variables and estimators over an extended period of time might according to Timmermann (2018) not generate an optimal model. The selected set of predictor variables that are utilized in a model may change over time. Therefore, there may be no model that is superior at all times and forecast horizons. An example of when a model needs to change variables is when the underlying pricing dynamics of the asset change. For example, the model for copper returns might have included some industrial production index to indicate the direction of demand but as technological advances where made the main driver of copper demand now stems from electronics. Then we want to replace with the industrial production variable for some index for electronics production. When new information is made public it will affect the forecasts made by an asset managers. If the new price forecast is shared by the broader market the new set of information will be incorporated into the market price through trading (Timmermann, 2018).

According to Timmermann (2018) conventional time-series regressions have difficulty handling breaks. It is challenging to detect a break with only a few post-break observations. As such building a model that quickly adjusts for breaks is difficult. The difficulty lies within the parameter estimations on the few data points that are available after the break. Timmermann (2018) says that if this is not managed the end result could be erratic forecasts with very large estimation errors rendering the model useless.

Over-fitting A model is over-fitted when the model uses too many parameters such that the model includes noise and thus performs poorly when estimating out-of-sample data. Koehrsen (2018) states that an over-fitted model can be viewed as too flexible. By that he means that the model adjust more to the observed data points than the true model. The opposite of this is an under-fitted model that instead is too rigid. To avoid using an over-fitted or under-fitted model to produce forecasts, the models are evaluated using a test sample, pseudo out-of-sample evaluation (Timmermann, 2018). It is important to control for over-fitting as an over-fitted model would perform worse when generating accurate out-of-sample forecasts.

2.4 Bandwidth selection for estimation of time-varying parameters

In this section we introduce two methods for the bandwidth selection utilized when estimating the parameters of the models. The first method is to implement an expanding window implying that as t increases there are more observations available for the parameter estimation. We then proceed to introduce a rolling window method which utilizes the same size of window, but moves the window ahead as t increases so that only the latest set of observations are used to estimate the parameter values. However, the first step in sampling is to divide the full sample size into a training set and out-of-sample test set. The reasoning behind implementing an in- and out-of-sample split as argued by Hansen and Timmermann (2015) is that a pseudo out-of-sample set is superior to an in-sample test. Using an in-sample test increases the risk of selecting an over-fitted model compared to if an out-of-sample test is used.

Expanding Data Window An expanding data window increases its bandwidth with each observation. For instance, given that a training set consist of 100 observations the expanding window would vary. When the model estimates the parameter values to generate the 50th prediction, it uses the previous 49 observations available. As the model estimates the parameters to predict the 100th observation it utilizes all 99 previous observations. In other words, the number of observations used in the window increases with each observation and estimation. Hansen and Timmermann (2012) suggest that in a stationary and stable environment recursive estimation based on an expanding data window makes most efficient use of the data. Their suggestion is stated to be applicable specifically when dealing with linear models. However, the expanding window encounters issues when applied to large time series data if the time-series is not stable. The relation between two data-points may have drastically changed over a period of ten years due to changes in the underlying market dynamics. Therefore, it does not make sense to include data that is too old as it might generate less accurate parameter estimations if the time-series is unstable.

Rolling Window Alternatively, Elliott and Timmermann (2005) suggest that a commonly used method for addressing slowly moving data is to use a rolling window sampling method. Elliott and Timmermann (2005) raise a concern in regards to how the method removes data in an arbitrary fashion without basing this decision on tests for breaks. For a model that is not believed to contain constant parameters Pesaran and Timmermann (2002) suggest that a fixed rolling window should be implemented. A rolling window is also implemented by Fama and MacBeth (1973) to estimate security betas.

2.5 Stationarity

Stationarity implies that the properties of the time series are constant over time. This increases the power of the models used in this thesis. Some of the tests and models that we introduce and use assume that stationarity holds. A time series is stationary when the moments do not change over two different consecutive time series of the same variable if the two time series are of the same size (Kwiatkowski et al., 1992). Stationarity can take two forms, weak and strong stationarity. Strong stationarity implies that the distribution of a stochastic process remains consistent over all time periods. Weak stationarity implies that the mean in the two consecutive time series are identical but that the moments depend on the size of the period evaluated. Palachy (2019) presents several types of stationarity, some of them are: first order stationarity, cyclostationarity, trend stationarity, joint stationarity.

The linear models used in this thesis require stationarity for the asymptotic properties of the models to hold. Stationarity can be tested for using the Dickey Fuller or the Augmented Dickey Fuller test. We use the Augmented Dickey Fuller test for stationarity.

Augmented Dickey Fuller(ADF) test can be utilized when dealing with larger and more complicated time series data since the augmentation allows for a higher order of the regressive process according to Cheung and Lai (1995). The null hypothesis of the Augmented Dickey Fuller test is that the time-series has a unit root. If a time-series has a unit root it is equivalent to being non-stationary. Thus, the alternative hypothesis is that the time-series is stationary. The ADF test evaluates stationarity by fitting the following ordinary least squares, OLS, regression.

$$\Delta y_t = \rho y_{t-1} + \sum_{j=1}^{p} \delta_{j,p} \Delta y_{t-j} + e_{tp} \tag{8}$$

where e_{tp} is assumed to be an i.i.d. random variable with zero mean. Let $\hat{\sigma}(\hat{\rho}_n)$ denote the standard deviation of the above OLS estimator, $\hat{\rho}$, of ρ. The ADF test statistic is computed as (Dickey and Fuller, 1981):

$$DF = \frac{\hat{\rho}_n - 1}{\hat{\sigma}(\hat{\rho}_n)}. \tag{9}$$

2.6 Forecasting models

In this section we briefly introduce the forecasting models that are evaluated against one another in this thesis.

Autoregressive Models of order p ,AR(p), uses a time-series of the past values of the response variable. An AR(p) is defined as:

$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \tag{10}$$

where y_t represents the variable being estimated, ϕ_0 are the intercept, p denotes the lag order of the model, ϕ_i's are the the coefficients. The error term is denoted as ε_t and it is typically assumed that it follows a normal distribution (Clark and Ravazzolo, 2012; Elliott and Timmermann, 2016).

Moving Average models of order q, MA(q), are given as:

$$y_t = \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{11}$$

where ε_t is the intercept and θ_i denotes the coefficients value associated with each respective lag order i (Elliott and Timmermann, 2016).

ARMA(p,q) is a combination of an AR(p) and MA(q) model. An autoregressive moving average (ARMA) model therefore also only uses the past values of the response variable to generate an estimate. The ARMA model is considered to be a workhorse amongst forecasters and the popularity of ARMA models can be explained by a few qualities according to Elliott and Timmermann (2016): (i) one of the advantages is the minimal demand placed on the required information set to generate the model. The model only needs historical data for the variable and can be estimated even if the forecaster has no idea of what fundamentals are driving change in the variable of interest assuming the fundamentals do not change often/strongly. It is, (ii), commonly used as a benchmark against more complicated models to show if there is any added value from using more advanced models. (iii) The utilization of ARMA models is theoretically motivated by the Wold representation theorem showing that all covariance stationary processes can be modeled as some moving average process. (iv) ARMA models have proved to be hard to beat in empirical work given that the lag order of the model is determined sensibly according to Elliott and Timmermann (2016). They proceed to argue that ARMA models are well suited for capturing persistence in economic variables. An ARMA(p,q) model is specified as:

$$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} \tag{12}$$

where pth is the highest autoregressive order and qth is the highest MA order that the model takes of the stationary variable y_t (Elliott and Timmermann, 2016).

Vector Auto-regressive,VAR(p), models are, as previous models mentioned, considered a workhorse within the field (Elliott and Timmermann, 2016). The inputs needed from the modeler are what variables to include, y_t, and the lag order, p of said variables. It is also of importance that the modeler selects sensible variables and imposes restrictions on the covariance matrix if needed. As the VAR(p) model is multivariate, y_t becomes a vector of size $(n \times 1)$ and the information set is extended as well which gives $Z_t = (y_{it}, y_{it-1}, ...)_{i=1}^{n}$. VAR models can amongst other things be used to estimate a covariance stationary multivariate process (Elliott and Timmermann, 2016). A vector auto-regressive model of order p, VAR(p), is denoted as:

$$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + u_t, \tag{13}$$

by Elliott and Timmermann (2016) where, A_i is a matrix of size $(n \times n)$ containing the auto-regressive coefficients for each i and $E[u_t u_t'] = \sum$. c is a vector $(n \times 1)$ containing the intercepts.

2.7 Evaluating Forecast Methods and Models

Economists are often faced with the problem of determining the relative merit of several forecasting methods according to Giacomini and White (2006). One solution is to develop out-of-sample tests that compare the different forecasts given that a general loss function is constructed. We consider four different tests to evaluate the performance of the forecasting models used in this thesis. The tests that we choose to use are; Diebold and Mariano (1995), DM, the Clark and West (2007), CW, mean squared forecast error (MSFE) adjusted t-statistic, Giacomini and White (2006), GW, and the Campbell and Thompson (2008) out-of-sample R^2.

Campbell and Thompson (2008) propose an out-of-sample R^2 statistic for evaluation of forecast performance. They compute the out-of-sample R^2 statistic on asset returns, r, as

$$R^2_{OS} = 1 - \frac{\sum_{t=1}^{T}(r_t - \hat{r}_t)^2}{\sum_{t=1}^{T}(r_t - \bar{r}_t)^2}, \tag{14}$$

where $\hat{r}_t$ is the predicted return at time, t and $\bar{r}_t$ is the average return. If R^2_{OS} is greater than zero it implies that the prediction has generated a lower mean squared prediction error than the benchmark average return. A small positive R^2_{OS} can generate large economic returns in-spite of having a relatively modest explanatory power. It then implies that a large positive R^2_{OS} most likely is too good to believe. Quoting Campbell and Thompson (2008) *"The saying "If you're so smart, why aren't you rich?" applies with great force here, and should lead investors to suspect that highly successful predictive regressions are spurious."* Campbell and Thompson (2008) find that their out-of-sample R^2 increases with time horizon given that the predictor variable is persistent. As horizons increase the denominator is subject to change, as the average return changes with horizon. For a short horizon i.e. one day the average returns is expected to be small and subsequently average return for a one month horizon is expected to be larger.

Diebold and Mariano (1995) tests the hypothesis that there is no difference in accuracy between two different forecast models. The forecast errors associated with the competing models are denoted as $[o_{it}]_{t=1}^{T}$ and $[o_{jt}]_{t=1}^{T}$ where T is the sample size and t is time. The Diebold and Mariano (1995) test statistic evaluates the null hypothesis:

$$E[d_t] = 0 \tag{15}$$

where, $d_t \equiv [g(o_{it}) - g(o_{jt})]$ and $g(\cdot)$ denotes the loss function. The test use an estimate d_t which represents the difference between the loss associated with the error for each of the two forecast models. Which results in the null hypothesis that the population mean of d is equal to zero. Under the assumption that d has a short memory and is covariance stationary the asymptotic distribution of $\bar{d}$ can be derived as follows according to Diebold and Mariano (1995):

$$\sqrt{T}(\bar{d} - \mu) \xrightarrow{d} N(0, 2\pi f_d(0)), \tag{16}$$

where the sample mean $\bar{d}$ is calculated as,

$$\bar{d} = \frac{1}{T}\sum_{=1}^{T}[g(o_{it}) - g(o_{jt})] \tag{17}$$

and the density of d at frequency zero is given by,

$$f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} E[(d_t - \mu)(d_{t-\tau} - \mu)] \tag{18}$$

where the population mean loss differential is equivalent to μ. In large samples the $\bar{d}$ is assumed to be approximately μ and distributed as $2\Phi f_d(0)/T$. Thus the Diebold and Mariano (1995) test statistic is computed as:

$$S_1 = \frac{\bar{d}}{\sqrt{\frac{2\Phi \hat{f}_d(0)}{T}}} \tag{19}$$

where $\hat{f}_d(0)$ is a consistent estimate of $f_d(0)$.

$$\hat{f}_d(0) = sum_{\tau=-(T-1)}^{(T-1)} 1(\frac{\tau}{S(T)})((1/T)\sum_{t=|\tau|+1}^{T}(d_t - \bar{d})(d_{t-|\tau|} - \bar{d})) \tag{20}$$

where $S(T)$ denotes the truncation lag and the lag window is computed as $1(\tau/S(T))$ by Diebold and Mariano (1995) in the asymptotic test. Costantini and Kunst (2011) argue that the Diebold and Mariano (1995) test statistic is invalid when evaluating nested models because the asymptotic properties would no longer hold. They argue that the test tends to be biased towards the simpler model. Diebold (2015) clarifies that the test originally was not intended for model comparison. However, the model is simple and it can easily be extended so that the loss differential can be explained by external variables such as business cycle, inflation rate. Such a change would result in the test moving conditional to the external variable that we want to condition on.

According to Diebold (2015) the Diebold and Mariano (1995) is great at comparing forecasts but when comparing models with several parameters the situation calls for something more nuanced. The Diebold and Mariano (1995) is still highly relevant in the situation but the test might not be optimal for model comparison. Diebold (2015) argues that pseudo out-of-sample comparisons, a category of tests including Diebold and Mariano (1995), might not be the optimal route for model comparison. Furthermore, they state that full sample comparisons seem preferable assuming that a true model exist.

Clark and West (2006) out-of-sample mean squared prediction error, MSPE, tests whether or not there is a difference between two MSPEs. It does so by evaluating whether the series follows a martingale difference or not. The test is developed by Clark and West (2006) as an extension of existing tests to be able to account for nested models. The null hypothesis of the Clark and West (2006) test is that there is no difference between the populations of the two mean squared prediction errors. In other terms there is zero difference in MSPEs.

$$H_0(\text{model 1}) : y_t = e_t \tag{21}$$

$$H_A(\text{model 2}) : y_t = X_t'\beta + e_t \tag{22}$$

i.e., $\beta = 0$ under the null hypothesis and $\beta \neq 0$ under the alternative hypothesis. Expectations based on current and past X's and e's are expressed as E_{t-1} and

X_t is a vector of variables. This thesis implements a one-tailed Clark and West (2006) test with a null hypothesis of the mean squared forecast error of model one being smaller or equal to the mean squared forecast error of model two. The alternative hypothesis is thus that mean squared forecast error of model one is larger than that of model two. If the null hypothesis is rejected model two performs better than model one. e_t is assumed to be a zero mean martingale difference under the null and alternative hypothesis:

$$E_{t-1}e_t \equiv E(e_t|X_t, e_{t-1}, X_{t-1}, e_{t-2}...) = 0 \tag{23}$$

Sample size is $T+1$. P observations are used for the predictions and R is the last in-sample observation. The first prediction is denoted as $(R+1)$, $(R+2)$,..., with the final prediction being $T+1$. This gives us $T+1 = R+P$. The prediction errors of the models are given by:

$$\hat{\sigma}_1^2 \equiv P^{-1} \sum_{t=T-P+1}^{T} y^2_{t_{+1}} = \text{MSPE model 1} \tag{24}$$

$$\hat{\sigma}_2^2 \equiv P^{-1} \sum_{t=T-P+1}^{T} (y^2_{t_{+1}} - X'_{t+1}\hat{\beta}_t)^2 = \text{ MSPE model 2} \tag{25}$$

Clark and West (2006) use $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ to test the null hypothesis that $\hat{\sigma}_1^2 - \hat{\sigma}_2^2 = 0$. They also suggest that one ought to use an adjustment to the MSPE when testing the martingale difference hypothesis to be able to use a standard normal distribution. The adjustment is given by $E(X'_{t+1}\hat{\beta}_t)^2)$. Including the adjustment then generates the following formula to compute the Clark and West (2006) test statistic:

$$\hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - [p^{-1} \sum_{t=T-P+1}^{T} (X'_{t+1}\hat{\beta}_t)^2]) \equiv \hat{\sigma}_1^2 - (\hat{\sigma}_2^2 - adj.) = \bar{f}(MSPE - adjusted) \tag{26}$$

One of the assumptions that Clark and West (2006) make is that stationarity holds. If stationarity does not hold the calculations in equation 26 need to be adjusted. The test favors models with more parameters. Meaning that even if the test indicates that the model containing more parameters performs better than a model with fewer parameters it does not imply that in any finite sample the larger model is better at producing out-of-sample forecasts (Elliott and Timmermann, 2016).

Giacomini and White (2006) propose a framework designed for use in situations when the forecasting model is likely to be misspecified. The framework they develop is to be used to evaluate out-of-sample predictive ability and selection of forecast designs (Elliott and Timmermann, 2016). Expanding on Diebold and Mariano (1995), Giacomini and White (2006) introduce two new innovations: (1) The finite sample properties are preserved asymptotically in the environment in which they derive their test statistic. (2) Introducing accommodations to evaluate different objectives, such as which model is more accurate at a future date or which model has on average been more accurate.

The null hypothesis of the Giacomini and White (2006) test is that the difference in expected value of the loss function of model one and the expected loss of model two is equal to zero (Elliott and Timmermann, 2016).

$$H_0 : E[L_{t+k}(Y_{t+k}, f_t(\hat{\beta}^*_{mt})) - L_{t+k}(Y_{t+k}, g_t(\hat{\beta}^*_{mt}))|G_t] = 0 \tag{27}$$

where m denotes the model, $f_t(\beta^*_m)$, and $g_t(\beta^*_m)$ are the forecasts for k steps ahead Y_{t+k} and $L_{t+k}(\cdot)$ denotes the loss function. The parameter estimates are denoted as $\hat{\beta}^*_{1t}$ and $\hat{\beta}^*_{2t}$ and G_t is some information set available at time t. In other words the null can be explained as: given the information available at time t it cannot be predicted which model will generate the most accurate estimate at time $t+k$. The Giacomini and White (2006) test statistic follows a Chi-squared distribution and the null is evaluated accordingly $T^h_{m,n,k} > \chi^2_{q,1-\alpha}$. Where α equals the significance level and $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2_q distribution with k degrees of freedom. n is defined as $n \equiv T - k - m + 1$ where T is the entire sample size and m is the size of the estimation window. If the test statistic is positive it implies that model two outperforms model one and if the test statistic is negative the opposite is implied. The Giacomini and White (2006) test statistic, $T^h_{m,n,k}$, is computed as:

$$T^h_{m,n,k} = n\bar{Z}'_{m,n}\tilde{\Omega}^{-1}_n\bar{Z}_{m,n} \tag{28}$$

where,

$$\bar{Z}_{m,n} \equiv n^{-1}\sum_{t=m}^{T-k} h_{tm,t+k}, \tag{29}$$

$$\begin{aligned}\tilde{\Omega}_n \equiv n^{-1}\sum_{t=m}^{T-k} Z_{m,t+k}Z'_{m,t+k} + n^{-1}\sum_{t=j}^{k-1} w_{n,j}*\\ \sum_{t=m+j}^{T-k}[Z_{m,t+k}Z'_{m,t+k-j} + Z_{m,t+k-j}Z'_{m,t+k}]\end{aligned} \tag{30}$$

where $w_{n,j}$ is a weighting matrix. As $n \to \infty$ $w_{n,j} \to 1$ for each $j = 1, ..., k-1$ (Newey and West, 1986; Giacomini and White, 2006).

The strength of Giacomini and White (2006) test allows for quite general estimation methods to produce forecasts. In addition to that the test is able to deal with both nested and non-nested models as well as capturing the effect of estimation uncertainty on relative performance of different forecast models. The primary drawback of Giacomini and White (2006) test statistic according to Clark and McCracken (2010) is that the approach cannot be utilized with a recursive scheme. If no other restrictions/assumptions are imposed on β as the sample size increases the estimated $\hat{\beta}_t$, assuming that $\hat{\beta}_t$ is a pseudo-true parameter, are going to be consistent with population β and thus there will be no estimation error in the unconstrained case. In the mis-specified model case the asymptotic characteristics only apply when the moving window being implemented is small compared to the number of out-of-sample observations (Clark and McCracken, 2010).

3 Data and Implementation

In this section we will go over our data selection, sample construction, model implementation, bandwidth selection, and testing procedures. To begin with, we utilize a symmetric loss function as our goal is to estimate future prices and we assume that over and under estimation in our case is equally costly. As the goal of our thesis is to evaluate what forecast model produce the most accurate estimates of future values, a symmetrical loss function is implemented. One of the assumptions for several of our models, including the ARMA, and some of our test, including the Clark and West (2006) test, is stationarity. In order for the explanatory and response variables to be stationary we have modeled the returns. The returns are tested for stationarity using the Augmented Dickey Fuller test.

3.1 Data

We used monthly data for the LME copper spot prices (LOCADY) denoted in United States Dollars (USD) retrieved from Bloomberg. We selected the LOCADY as a proxy for industrial metal commodity markets. The LOCADY was selected as the it is used by practitioners. Furthermore, Buncic and Moretto (2015) also utilised monthly copper prices from LME, which makes the results comparable.

In an attempt to reduce risk of autocorrelation in forecast errors for our time series we have opted to use the last value every month to estimate our models. Alternatively, we could have used daily data and estimate one month ahead which would have increased our sample size significantly. However, the autocorrelation between forecast errors for the one month estimate made on March first and March second will be greater than that of two estimates made on February first and March first.

In order to be able to test the models out-of-sample forecasting performance the initial sample, containing 315 months of observations between November 1995 and January 2022, are divided into a training set, upon which the models are estimated, and a testing set which makes up our out-of-sample observations. We retain out-of-sample observations for evaluation in order to circumvent selecting a over fitted model in the process of model evaluation. The training set contains 265 observations between November 1995 and November 2017. The test set contains 50 observations between December 2017 and January 2022. The selection of sample split is made arbitrarily as the trade-of between allowing for a large training sample to optimise the models is weighted against the value of a larger amount of out-of-sample observations available for out-of-sample testing. Given the emphasis we have selected to attribute to bandwidth optimization we split the sample into training and test respectively containing 85% and 15% of the entire sample. Resulting out-of-sample observations is 47 with the last observation being October 2021.

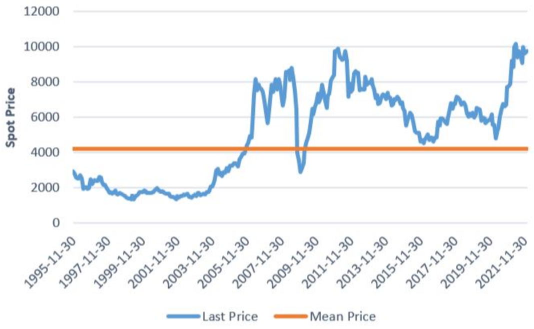

Figure 1: Monthly Spot Price of LOCADY on LME

Due to the assumptions of stationarity associated with several of our models, for instance the ARMA processes, and our tests, for instance the Clark and West (2006) test, we test the LOCADY training sample for stationarity using an ADF test. The test indicated that the the LOCADY training sample contain unit roots suggesting that it is not stationary, and therefore not appropriate for our suggested methodology.

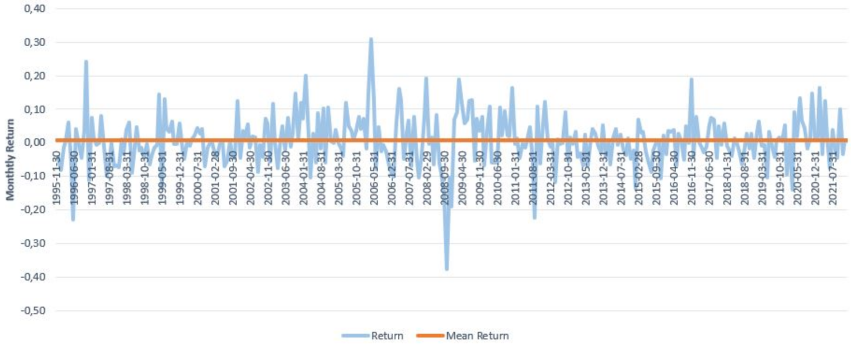

Figure 2: Monthly Returns of LOCADY

In order to overcome the lack of stationary properties for the LOCADY we chose to construct percentage returns and test the newly created return training sample for stationarity. We find that the returns training sample is indicated to be stationary. It is visually apparent in Figure 2 that the returns display mean reverting qualities as it oscillates around the mean represented by the orange line. Furthermore, it is apparent that the mean returns are not equal to zero. The mean return is 0.0061 which translates to 0.61% average monthly return. 0.61% average monthly return cumulative over a year translates to approximately 7.57% yearly returns. Notably, we only test for stationarity in the training sample as the information provided beyond this point is to be considered unavailable to us if we want to maintain the integrity of the out-of-sample observations, test data.

3.2 Implementation

A significant part of our implementation is about model selection. The model selection is heavily affected by the transferability, and applicability of the models to other commodities and time series. Therefore, the vast majority of the models, all except the VAR, utilize exclusively the historical spot price data for the commodity it forecasts, in this case the LOCADY, as its information set. The VAR(p) model utilizes additional explanatory variables apart from the historical LOCADY data. In total we create 10 models that are then evaluated against one another.

Model 1 is a simple auto-regressive model, AR(p). The order of the AR model is decided using the Akaike Information Criterion, (AIC), and Bayesian Information Criterion, BIC. AIC and BIC resulted in conflicting lag length suggestions. AIC suggest a lag order of $p = 11$ and BIC suggests $p = 1$. We decide to go with lag order $p = 1$ as indicated by the Bayesian Information Criterion. We choose BIC over AIC, because BIC allows for variability in the parameter. Below is our constructed model where t is time, p equals lag length, a is a constant, ϕ is the coefficient and y is copper return.

$$y_{t+1} = a + \phi_1 y_t + \phi_{i-p} y_{t-p} + \varepsilon_t \tag{31}$$

Model 2 and Model 3 are moving average, MA(q), models. We use two MA models. The first MA model is based on three months moving average and the second on the six month moving average. We chose the orders three and six since we have monthly data and wanted to evaluate the effect of one and two quarters. We decided to not include three and four quarters to reduce model size and computation time. Below is our constructed model where t is time, q equals lag length, a is a constant, θ is the coefficient and y is copper return.

$$y_{t+1} = a + \theta_1 \varepsilon_t + \theta_i \varepsilon_{t-q} + \varepsilon_t \tag{32}$$

Model 4, Model 5 and Model 6 are auto-regressive moving average models, ARMA(p,q). We construct an ARMA(1,1), ARMA(1,3) and ARMA(1,6) model. As they consider data from the past month, the past quarter by month and the past half a year. The *pth* order is selected using AIC as in Model 1 and the order of q is selected as in Model 2 and 3. The forecast model utilized is:

$$f_{t1|}t = \phi_1 f_{t+h-1|t} + \phi_2 f_{t+h2|t} + ... + \phi_{p+h|t} + \theta_h \varepsilon_t + ... + \theta_q \varepsilon_{t-q+h} \tag{33}$$

Model 7 is a vector auto-regressive model which includes six variables. One distinguishing feature separating the construction of the VAR model from the other models is the sample size upon which it is trained and tested. The variables included in the model are selected based on previous literature. The model utilizes the United States Dollar (USD) to Chilean Pesos (CHP) exchange rate which is relevant since Chile is the largest exporter of copper in the world (Otgochuluu et al., 2021). Additionally, the export of copper accounts for half of Chile's total exports (Díaz et al., 2021). Therefore, one of the main drivers of CHP should be external demand for copper. The trade of copper is conducted in USD which suggests that the relationship between the two currencies is correlated with the price of copper.

US Industrial Production Index is included, as Roache and Rossi (2010) find that it to produces relevant coefficients for other metallic commodities. The Michigan Surveys of Consumers sentiment index, (SENTI), is included as a macro-variable as Nguyen and Walther (2020) find that SENTI has an effect on commodities. Sethi et al. (2014) suggests that it is correlated with the price of silver, which they classify as a very useful industrial commodity despite it being defined as a precious metal. Kilian and Zhou (2018) global real economic activity index is added as a variable due to its suggested ability to forecast commodity prices and the argument made by Gargano and Timmermann (2014) that commodity markets are global. The corresponding three months forward contract of copper is included as it contains information about the markets expectation of the future copper prices.

Lastly, the inventory levels of copper as reported for LME is included attributed to the fundamental theories of supply and demand where change in the supply of a given commodity, in this case copper, could have an affect on the price of the commodity. Due to lower availability of historical data for some of the six included variables, the VAR is created using data between January of 2012 and December 2021. We expect the variables to have an inter-temporal effect on each other as there might be delayed effects from inventory changes, forward contracts as well as from consumer sentiment and US industrial production. The lag order of the VAR is selected in a similar manner to how the bandwidth is determined, described in more detail in section (3.3). Instead of running window size on the x-axis we have estimated the model for different orders of p. MSFE in-sample is minimized when p=1 we therefore choose to use a VAR(1) model.

Model 8 is a simple ensemble of models comprising of a combination of the other forecast models. We generate the portfolio by equally weighting the MA(3), MA(6), AR(1), ARMA(1,1), ARMA(1,3) and ARMA(1,6) models. We refer to the combination as "EQW".

Model 9 is an ensemble of three models. However, in this combination we only utilize the top three models for each forecast horizon. They are ranked by their in-sample R^2 values after each model has been optimized. The ensemble then contains the top three models that have been optimized individually. We refer to the combination as "TOP3". In the short term ensemble the "TOP3" contains ARMA(1,6), MA(6) and MA(3). For the mid term forecast horizon "TOP3" is comprised of the AR(1), ARMA(1,1) and ARMA(1,6) model. Lastly, for the three step ahead it consists of the AR(1), ARMA(1,6) and MA(3).

Model 10 is our benchmark. The benchmark is created from the notion that current price is the best prediction for future price. If a model fails to outperform the benchmark it indicates that the *status quo* is an equally or more accurate estimate of future copper prices.

3.3 Bandwidth optimization

We implement a moving window to get time-varying parameters as a way to deal with model instability (Johannes et al., 2014). Furthermore, a moving window mitigates the effects of persistent predictors associated with an expanding window. The bandwidth of the moving window is optimised by computing the mean squared forecast error, MSFE, in-sample. In order to reduce computation time of the code the window size increases in jumps of 6 months which drastically decreases the number of windows which are to be evaluated. We chose to identify the minimum MSFE for each forecast horizon. We evaluate the models on one, two and three month ahead forecasts. However, there is a possibility of the MSFE being less accurate for the larger window sizes due to less data-points being available to evaluate the bandwidth on. The reduction in available observations in sample is a direct function of the window size being implemented.

An example is seen in Figure 3.3 from which bandwidth selection is made for the different forecast horizons of the AR(1) model. The number of data points used to evaluate the in-sample MSFE is reduced as window size increases. Figure 3.3 indicates a drastic decrease in MSFE as bandwidth increases above 140. We decide to evaluate amongst the window sizes below 140 to avoid having too few data points to properly evaluate the MSFE in-sample. Subsequently, we select a bandwidth of 138 observations for all three horizons of Model 1 as denoted in Table 1. The bandwidth optimisation is conducted in the same manner for all models subject to optimisation. Table 1 contains the bandwidths selected for Model 1 to Model 7 for each forecast horizon.

Model 8, Model 9 and Model 10 are not subject to bandwidth optimization. Model 8, and Model 9 are weighted combinations of the other models and therefore do not use a specified bandwidth but rater consider the forecasts made by the models they contains. Model 10 is a benchmark suggesting that

Figure 3: MSFE for bandwidth from 0 to 180 for AR(1), MA(3), and MA(6)

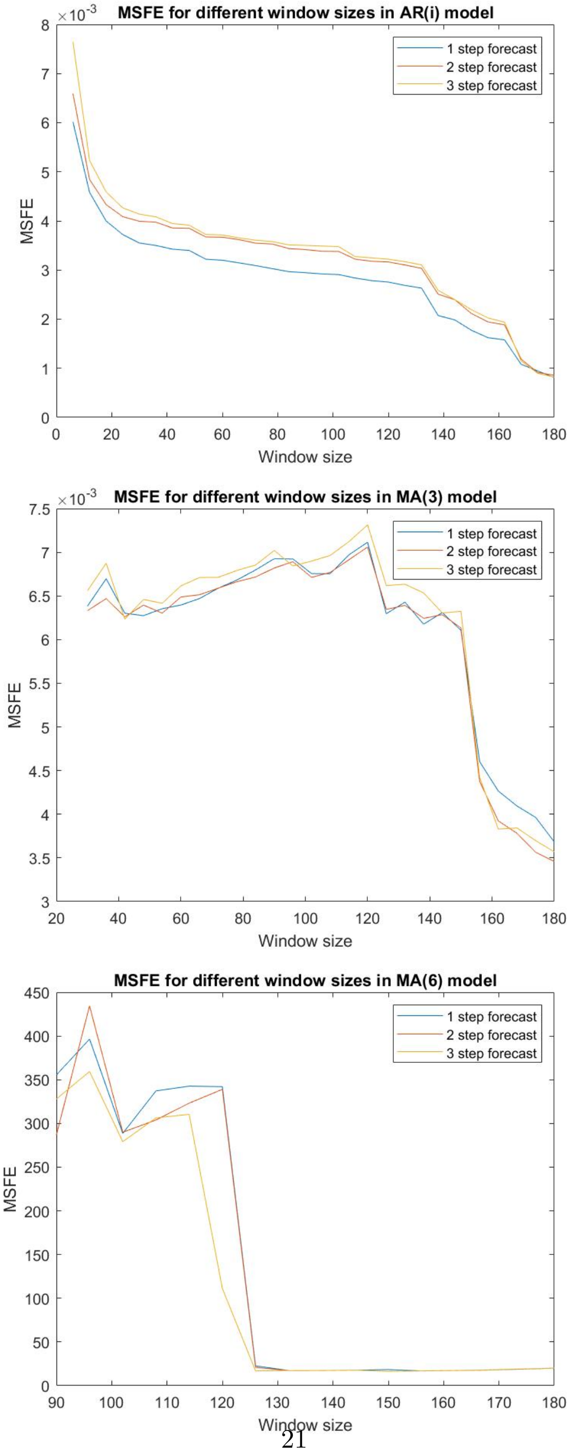

Table 1: Bandwidth selected for model and horizon

Model number	Type of model	Short-term	Mid-term	Long-term
Model 1	AR(1)	138	138	138
Model 2	MA(3)	42	42	42
Model 3	MA(6)	102	102	102
Model 4	ARMA(1,1)	30	42	42
Model 5	ARMA(1,3)	60	60	60
Model 6	ARMA(1,6)	136	136	136
Model 7	VAR(1)	36	30	30

the monthly returns are equal to zero for all periods and is as such not subject to time-varying estimation and does not need a bandwidth.

3.4 Test implementation

In this section, we shortly explain how we have applied the tests introduced in section 2.6. In the tables that the test are presented the column represents model one and the rows represent model two implemented in the test. The Campbell and Thompson (2008) out-of-sample R^2 we compute ourselves in line with the math presented in section 2.6. For the Giacomini and White (2006) test statistic we utilize the original code file posted by Giacomini and White (2006) in association with the paper. For the implementation and computation of the Clark and West (2006) test statistic we utilize the function created by Bang (2022). The Diebold and Mariano (1995) test is implemented via the function created by Ibisevic (2022).

4 Results

The results of the four tests are evaluated below, starting with the Campbell and Thompson (2008) out-of-sample R^2 test results. The out-of-sample R^2 statistic is followed by presentations and evaluations of the short-, mid-, and long-term forecast performance. The test statistics for the Diebold and Mariano (1995) test, Clark and West (2006) test, and Giacomini and White (2006) test are presented for each forecast horizon. Lastly, there is a general discussion of the models performance across the battery of tests and the discrepancies between the tests.

4.1 Campbell and Thompson (2008) out-of-sample R^2 test

We found that the most notable results in Table 2 is the positive out-of-sample R^2 value for the one-step-ahead forecast. The out-of-sample R^2 is a test evaluating the forecast performance for a given model relative to the average returns. Evaluating the test results seen in Table 2 the columns contain the models whereas the rows denote the forecast horizon. For example, in column six the model is ARMA(1,6) and for row one which is the one-step-ahead forecast, we observe a test statistics of $R^2 = 0.0326$. The positive test statistic is surprising as the surface level implication is that the rather simple models tested perform well enough to profitably trade on. However, keeping in mind the warning from Campbell and Thompson (2008) that a surprisingly good test statistic may indicate issues with the data, the application of the test, false positives or the models implying that results may in fact be too good to be true.

One possible explanation is that the monthly return on the spot price of copper excludes some other costs associated with the trading of copper, such as storage and transportation. In this case, it may not be possible to make profits with any of our models due to the additional costs of shipping and storage, for a minimum period of a month, exceeding the potential profits indicated by our Campbell and Thompson (2008) test statistics. Correspondingly, the financial actors in this market would not be relying on the predictions from a forecasting model for trading unless the given model outperforms the historical average with a margin large enough to cover the additional costs associated with trading. Thus, a potential explanation for why we obtain positive test statistics with relatively simple models, is that the market is unable to trade on their low magnitudes of explanatory power due to additional costs of trading this commodity.

Table 2: Out-Of-Sample R^2

	ma(3)	ma(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random walk
1 step	-0,0246	-0,0940	-0,0291	-0,1154	-0,0622	0,0326	-0,2889	0,0006	0,0127	-0,0224
2 step	-0,1256	-0,1639	0,0138	0,0327	-0,1116	0,0485	0,0168	-0,0079	0,0470	-0,0224
3 step	-0,1423	-0,2330	-0,0035	-0,0472	-0,1818	-0,0313	-0,0380	-0,0767	-0,0354	-0,0224

In an attempt to clarify if the positive results and support the aforementioned notion that the test statistics are caused by unobserved trading costs, the same models and the Campbell and Thompson (2008) out-of-sample R^2 statistic is computed for the London Metal Exchange three-month-forward contracts. The statistic is presented in Table A7, which can be found in the appendix. The test on tree-month-forward contracts produces mostly negative out-of-sample R^2 values. For the two-step-ahead MA(6), the two-steps-ahead "EQW", and the two-steps-ahead "TOP3" we find positive out-of-sample R^2 values. A notable observation in the test statistic is that the AR(1), VAR(6), ARMA(1,1) and ARMA(1,6) models which generated positive R^2 statistics for our original data produce exclusively negative test statistics for the three-month-forward case. Notably, as the positive values are produced by different models in this case providing a slight indication that the potential flaw is not related to any one model.

Alternatively, Campbell and Thompson (2008) find that out-of-sample R^2 increases with time horizons when the predictor variables are persistent. Applying their findings can lend an explanation for the positive out-of-sample R^2 statistics. However, as the forecast horizons being used are not particularly short, ranging between one and three months, the length of the forecast horizon is therefore not likely to be the source of the surprising results.

Another aspect which may contribute to, or cause the surprising values is if the assumption of stationarity found within the training sample does not hold for the test sample. Considering the time period in question, which is between September 2017 and October 2021, when we experienced global turmoil across most markets during the COVID-19 pandemic. The turmoil could have a large enough impact on the price of copper to break the stationarity of returns. To control for stationarity throughout the training and testing samples, we ran an ADF test on the full sample containing all 315 observations. The test indicated that despite the COVID-19 related noise in the market the percentage returns on LOCADY still maintained stationarity.

Demonstrating the surprising results associated with the positive Campbell and Thompson (2008) out-of-sample R^2 values we applied a simple buy and short scheme using our ARMA(1,6) short-run model. The scheme buys copper when our ARMA(1,6) model indicates that the price will increase over the coming month and shorts copper prices when the model indicates that the prices will decrease the coming month. The scheme is simple and does not account for any trading cost. In the test sample the investment scheme based on our ARMA(1,6) predictions produces 14.84% average annual return. Given the simplicity of the ARMA(1,6) model and the expected market engagement in the LME copper market, 14.84% is highly surprising and should not be possible. Furthermore, assuming that financial actors are only interested in the real returns on their investment, the real profitability is still to be considered high. According to Ha et al. (2021) and the World Bank database, the average yearly inflation globally during the estimated period from 2017 through 2021 was 4.15%. Which means that that the yearly real returns would exceed 10%. The indicated high profitability in our simple model supports the suggestion that there exists additional costs associated with the trade of copper on the LOCADY.

The unlikely occurrence of positive out-of-sample R^2 values may be also caused by our small number of out-of sample observations $n = 47$. It is possible that our out-of-sample R^2 test results either are over or under estimated due to our small group of out-of-sample observations deviating from the true population. However, given the notion that the practical trade of the spot prices may carry additional costs, we do not find it evident that the Campbell and Thompson (2008) test statistics are overestimated due to a small test sample. Reconsidering the control test we ran on a smaller data set for three month futures, and subsequently as a result on a smaller group of out-of-sample observations with $n_{out-of-sample} = 36$. In this case for the three month future contracts, which is assumed to be a market with extensive amount of interaction from financial actors, most of the out-of-sample R^2 are negative. The few positive values are likely overestimates due to sample size and the sample in question deviating from the population.

The Campbell and Thompson (2008) out-of-sample R^2 results, despite the surprising positive statistics, provides a ranking of the competing models for each horizon as seen in Table A8 in the Appendix. Evaluating the different rankings for each horizon we see that the short-term forecast favours ARMA(1,6), the "TOP3" and the "EQW". The benchmark performs above the remaining six models. For the mid-term forecasts the ARMA(1,6) and the "TOP3" maintain their rankings as the top two models. However, for the long-term forecast, the AR(1) performs the best followed by the benchmark, the ARMA(1,6) and then "TOP3" in that order. The result indicates that the ARMA(1,6) and the "TOP3" consistently are amongst the best performers. Furthermore, the ARMA(1,6) outperforms both of the ensemble's for every h step ahead forecast. It is also evident that no model outperforms the benchmark over all three periods.

Comparing the different horizons, we note that six out of the ten competing models increase their out-of-sample R^2 statistic between the short-term and mid-term forecast horizons. For example, the third column which is the AR(1) model, has a greater reported out-of-sample R^2 in the second row, representing the mid-term forecast, than it does in the first row which is the short-term forecast. The increase is surprising as our models are expected to perform worse the further ahead we predict. Especially, since the models forecast monthly returns which is assumed to be stationary and is tested for stationarity by applying an Augmented Dickey Fuller test. Subsequently, the stationarity of the monthly returns imply that the historical mean should hold and perform better compared to the model forecasts as the forecasting horizon increases. The long-term forecast do behave as expected and decrease relative to the mid-term forecasts. However, three of the ten models produce lower out-of-sample R^2 values for the long-term than they do for the short-term horizon.

Despite the potential issues with the accuracy of the out-of-sample R^2 values, we do observe that Buncic and Moretto (2015) similarly found that a relatively simple OLS model produced significantly larger explanatory powers than we found in our models. Their result of close to 10% out-of-sample R^2 for a simple OLS model suggest that our simpler models producing just under 5% are in line with their findings.

4.2 Short-term forecasts

The Clark and West (2006) test, Diebold and Mariano (1995) test and Giacomini and White (2006) tests are relative tests comparing the performance of one forecasting model with the performance of another forecasting model.

Table 3: One step CW Test Statistic

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		2,3588***	1,233	1,6089*	1,5392*	0,628	3,0122***	0,4903	0,2248	1,122
MA(6)	0,2388		0,41	0,8112	0,268	-0,4769	2,611***	-1,3499	-1,0178	0,2328
AR(1)	1,1674	1,2036		1,8085**	1,1209	-0,1622	2,6523***	0,4067	0,4392	0,3496
ARMA(1,1)	0,5702	0,4915	0,4136		0,4666	-0,3115	2,7477***	-0,5221	-0,8738	0,006
ARMA(1,3)	0,3449	1,3053*	0,8752	1,0936		0,0832	2,6487***	-0,1754	-0,1675	0,7861
ARMA(1,6)	1,5332*	1,8975**	1,3941*	2,3581***	1,3104*		3,0869***	1,3254*	1,3718*	1,5087*
VAR(1)+	1,851**	1,2513	0,373	0,6736	1,2148	0,1789		0,6446	0,9929	0,3335
EQW	1,0833	1,9462**	0,925	1,6497**	0,9903	-0,0147	2,6366***		-0,3845	0,8633
TOP3	0,9946	1,7651**	1,1719	1,9935**	1,0034	0,3511	2,4422***	0,7417		1,1319
Random Walk	1,1383	1,2024	0,5993	1,777**	1,0944	0,1182	2,8234***	0,385	0,3647	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%
+ estimated on a smaller sample

The Clark and West (2006) is a one-tailed-test which tests the null hypothesis, that the MSPE of model one is equal to or smaller than that of model two. The test results for the short-term forecasts are presented in Table 3 where model one is denoted in the column and model two in the row. For example, in the second column where model one is the MA(6) and the first row where model two is the MA(3) we obtain the test statistic $t = 2.3588$ which is significant at 99% as indicated by the three asterisks. Significant test statistics indicate that model two perform significantly better than model one at some significance level which is represented in the table by asterisks. In this case it implies that the MA(3) outperforms the MA(6).

Table 4: One step GW Test Statistic

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	"TOP3"	Random Walk
MA(3)		0,427	6,465**	-0,416	0,899	0,018	-1,718	0,128	-0,102	5,832*
MA(6)	-0,427		4,756*	-1,549	3,712	-0,293	-3,385	-1,076	-1,885	3,939
AR(1)	-6,465**	-4,756*		-4,252	-12,296***	-4,844*	-4,671*	-9,744***	-9,078**	1,895
ARMA(1,1)	0,416	1,549	4,252		1,148	1,402	-2,899	1,973	0,985	5,359*
ARMA(1,3)	-0,899	-3,712	12,296***	-1,148		-7,694**	-2,969	-10,693***	-5,786*	9,682***
ARMA(1,6)	-0,018	0,293	4,844*	-1,402	7,694**		-2,176	0,581	-0,379	2,388
VAR(1)+	1,718	3,385	4,671*	2,899	2,969	2,176		3,296	2,687	5,171*
EQW	-0,128	1,076	9,744***	-1,973	10,693***	-0,581	-3,296		-4,304	7,757**
TOP3	0,102	1,885	9,078**	-0,985	5,786*	0,379	-2,687	4,304		8,652**
Random Walk	-5,832*	-3,939	-1,895	-5,359*	-9,682***	-2,388	-5,171*	-7,757**	-8,652**	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%
+ estimated on a smaller sample

Both the Diebold and Mariano (1995) test and the Giacomini and White (2006) test are two-tailed tests. The two models are presented using the same structure. The test results for the short-term forecasts are presented in Table 5 for the Diebold and Mariano (1995) test statistics and in Table 4 for Giacomini and White (2006) test statistic where model one is denoted in the column and model two is denoted in the row. A positive result indicates that model two outperforms model one. Negative test statistics indicate the opposite. For example, in Table 4 the third column being the AR(1) and the fifth row being the ARMA(1,3) we obtain a test statistic $t = 12.296$. The test statistic indicates that the ARMA(1,3) outperforms the AR(1) as we reject the null hypothesis with significance level of 99%. Due to the two-tailed nature of the tests switching the models around yields the same test statistics but with the opposite sign. For example, using our previous example but considering the fifth column being the ARMA(1,3) and the third row being the AR(1) we obtain a test statistic $t = -12.296$. The test statistic indicates that the ARMA(1,3), outperforms the AR(1) as the null hypothesis is rejected with 99% confidence.

As the three tests compare two models at a time it indicates the relationship between the two models. The relationships indicated by the tests are not transferable which can create inconsistencies. For example, in Table 4 ARMA(1,6) outperforms the ARMA(1,3) at significance 95% and the ARMA(1,3) outperforms the benchmark at 95% significance. However, the ARMA(1,6) does not outperform the benchmark with significance greater than 90%. This lack of transferability of the relationships between models result in performance of each model being evaluated with greater emphasis on the amount of other models they outperform or are outperformed by, and less on individual relationships.

Evaluating the short-term forecast test statistics from each of the three tests we observe that they deviate slightly from one another. The GW test statistics indicate that the ”TOP3” and ”EQW” perform the best. Both models outperform three other models including the benchmark model at 95% significance. Neither model is outperformed at significance greater than 90% by any of the other models. The ARMA(1,6) outperforms two other models at significance 90%. The ARMA(1,6) does not outperform the benchmark. Similarly, the CW test statistic indicates that ARMA(1,6), the ”EQW” and ”TOP3” perform well. However, unlike the GW test the CW test indicates that the ARMA(1,6) outperforms every other model at significance greater than 90%. Furthermore, the CW test statistics indicates that the MA(3) outperforms more models than the ”EQW” whereas the GW test indicates that the ”EQW” performs slightly better than the MA(3). The GW statistic deviates from the CW as the ARMA(1,3) performs really well in the test but average in the CW test. Furthermore, the CW test statistics deviate form the GW as it indicate both the VAR(1) and the ARMA(1,1) are outperformed by the benchmark at significance greater than 95%.

Table 5: One step DM Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		0,3272	-0,4765	-0,2549	1,3977	-0,7034	0,4871	-0,5885	-0,6694	-0,4344
MA(6)	-0,3272		-0,7458	-0,4889	0,6755	-1,1703	0,5498	-1,3303	-1,2519	-0,7124
AR(1)	0,4765	0,7458		0,8322	0,7329	-0,2149	1,3639	0,3221	0,1078	0,5434
ARMA(1,1)	0,2549	0,4889	-0,8322		0,5707	-0,8637	1,3088	-0,1385	-0,4651	-0,7136
ARMA(1,3)	-1,3977	-0,6755	-0,7329	-0,5707		-0,9553	0,0563	-0,9542	-0,9635	-0,7102
ARMA(1,6)	0,7034	1,1703	0,2149	0,8637	0,9553		1,4498	0,9215	0,777	0,4546
VAR+	-0,4871	-0,5498	-1,3639	-1,3088	-0,0563	-1,4498		-1,2758	-1,0457	-1,3776
EQW	0,5885	1,3303	-0,3221	0,1385	0,9542	-0,9215	1,2758		-1,0026	-0,2198
TOP3	0,6694	1,2519	-0,1078	0,4651	0,9635	-0,777	1,0457	1,0026		0,0454
Random Walk	0,4344	0,7124	-0,5434	0,7136	0,7102	-0,4546	1,3776	0,2198	-0,0454	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

The DM test statistics presented in Table 5 show no values with significance greater than 90%. For reference, the CW test returns 29 significant results and the GW test returns 31 significant results. The discrepancies imply that the DM is more conservative than the other two tests.

4.3 Mid-term forecasts

The mid-term forecast test statistics are slightly inconsistent between the GW test and the CW test. Considering the test results presented in Table A5 and Table A3, both found in the appendix, we see that the GW test, favours the ARMA(1,1) and the "EQW". The ARMA(1,6) and the "TOP3" perform rather poorly. For the CW test the ARMA(1,6) and the "TOP3" performing the best with the ARMA(1,1) doing well and the equally weighted portfolio performance being average. We observe that the two test struggle to agree on what forecasting model performs the best.

Evaluating the models mid-term forecast performances against the benchmark we find that the CW test results and the GW test results deviate slightly but that they are mostly consistent. CW indicates that three models, the ARMA(1,6), the "TOP3", and the AR(1) outperform the benchmark for the mid-term forecast. The GW test statistics indicate that six of the models, including ARMA(1,6), the "TOP3", and the AR(1), outperform the benchmark model. Noticeably, the CW indicates that two models are out performed by the benchmark whereas the GW test does not indicate that any model is out performed by the benchmark. Generally, we find that the two tests agree on the what model is stronger, however, they disagree on whether a relationship is significant.

Taking into consideration the DM test, we find that the test produces fewer significant results for the mid-term forecasts than the other tests. The DM test produces more conservative test statistics which is consistent with what we found for the short-run forecast. Notably, the mid-term forecast test statistics, seen in Table A1, do contain three relationships with significance. The test indicates that the AR(1) and the "TOP3" outperform the benchmark model. Furthermore the "EQW" outperforms the MA(6). The increase in significant results, especially against the benchmark, supports the pattern seen in the Campbell and Thompson (2008) out-of-sample R^2 that the models perform better in the mid-term forecast than they do for the short-term forecast.

4.4 Long-term forecasts

For the long-term forecast the CW test seen in Table A6 in the appendix and the GW test seen in Table A4 in the appendix both indicate that the ARMA(1,1) and the "TOP3" are amongst the best performing models. However, the GW indicates that the ARMA(1,1) is the top performer whereas the CW test indicates that it's performance is average. Similarly, the AR(1) is indicated to out-perform most other models in the GW test, whereas it is indicated to perform average in the CW test. The two tests are consistent in favouring the "TOP3" as a top performer they do not however agree on which other models perform well.

Notably, the CW test indicates that none of the models outperforms the benchmark model. The benchmark is, however, indicated to outperform the MA(6), the ARMA(1,3), and the VAR(1). The GW test alternatively indicates that four models outperform the benchmark. The "TOP3", the ARMA(1,3), the ARMA(1,6), and the AR(1) are all indicated to outperform the benchmark, whereas the benchmark is not indicated to outperform any of the models. The tests do not contradict each other as there is no overlap between the group indicated to be worse than the benchmark in the CW test and the group indicated to be better than the benchmark in the GW test. The test results do however disagree on the significance in the individual relationships between models.

Evaluating Table A2, found in the appendix, it supports the notion that the DM test is more conservative than the other tests. Similar to the short-term forecast test statistics the long-term forecast test statistics do not indicate that any model is better or worse than any other model.

4.5 General discussion

Considering the results of the Campbell and Thompson (2008) out-of-sample R^2 test, Clark and West (2006) test, Giacomini and White (2006) test, and the Diebold and Mariano (1995) test performed on the ten models for short-term forecast performance, mid-term forecast performance, and long-term forecast performance, certain trends in model performance and discrepancies between tests become evident.

The discrepancies between the tests can be a result of the test's varying ability to handle the different model designs. For instance, the GW test struggles with accuracy if the moving window is not small in relation to the number of out-of-sample observations as the assumed asymptotic characteristics of the test may not hold any longer. With bandwidth for some of our models exceeding

100, and the out-of-sample observation is $n = 47$, the test statistic might not comply with the underlying necessary assumptions. Similarly for the DM test, Costantini and Kunst (2011) suggest that the test is poorly suited for nested models as it is biased towards the simpler models. The creators of the test argue that the model struggles with comparisons of models containing multiple parameters Diebold (2015).

Our small number of out-of-sample observations can cause either Type 1 or Type 2 errors depending on how the different test are defined. Type 1 errors are false positives which is the rejection of a null hypothesis when it should not have been rejected. One example of a Type 1 error is if a DM test indicates that an AR(p) model performs better than an ARMA(p,q) containing the AR(p) model. Costantini and Kunst (2011) argues that the DM test is biased preferring simpler models and struggles with nested models. If the test result indicates that the AR(p) model outperforms the ARMA(p,q) model, when the the AR(p) model actually perform worse or as well as the ARMA(p,q), this would be a Type 1 error. Type 2 errors are the opposite, that is the failure to reject a null hypothesis that should have been rejected. For instance, if a GW test fails to indicate that there is a difference in performance between an AR(p) and a MA(q) model, when there actually is a significant difference in performance between the two models. In this case, the models may have used a bandwidth significantly greater than the testing sample which can result in the assumptions for the test being violated. The broken assumptions in turn may render the test unable to accurately compare the model performance, increasing the likelihood of both Type 1 and Type 2 errors. The likelihood of either of the two types of errors are negatively correlated with size of the test sample. Considering the potential occurrence of Type 1 or Type 2 errors to our test statistics, with the out-of-sample observations being limited to $n = 47$, it may be that the sample does in fact deviate significantly enough from the true population. The cause of deviation between the three test's respective results may be due to Type 1 and Type 2 errors.

Both the CW test and the GW test produce overall rather consistent results. The DM test on the other hand produces drastically more conservative test statistics only indicating three significant relationships across all 135 tested relationships. The deviation could either be due to Type 1 errors occurring within both the CW test and the GW test. Alternatively, the DM test is subject to Type 2 errors. Considering the findings in Buncic and Moretto (2015) who used the CW test and found significant results compared to their benchmark, which is consistent with our results of our CW test. It might be that our tests have low power which increases the likelihood of Type 2 errors. The power of a test indicates its ability to reject the null hypothesis when the alternative hypothesis is true. The power of a test is positively correlated with sample size. If the tests have weak power towards the alternative hypothesis that might help explain the inconsistency of our test statistics. Assuming that sample size, effect and alpha remain unchanged a one-tailed test is more powerful than a two-tailed test (Borenstein, 1998). Intuitively, we would expect the DM test to be less powerful than the CW test as they contain the same data and the DM is a two-tailed test whereas the CW is a one-tailed test. However, we do not know what power the tests have against the alternative hypothesis as we do not conduct a power analysis. The power of the tests will need to be verified in a future study in order to determine whether power is the cause of the inconsistent test results.

Despite the discrepancies some trends are consistent throughout the different tests. The forecasts seem to perform better against the benchmark in the mid-term than the short term across all four tests. The ARMA(1,6), the "TOP3", and the "EQW" do well in three of the four tests with the Diebold and Mariano (1995) indicating that no model is significantly better or worse than any other model in the short run. In the short-term these models also outperform the benchmark consistently across three of the tests. The mid-term forecast test results indicate that the "TOP3" and the ARMA(1,1) perform the best consistently across three out of the four test statistics. The top models yet again outperform the benchmark across the three tests. However, the long-term test statistics are less consistent across the tests, only consistently indicating that the models are unable to outperform the benchmark model. Only one model, the AR(1), outperforms the benchmark in more than one test.

Across the three horizons we see further discrepancies as the rankings and relationships alter for all tests. For instance, the test results in the short-term generally favors the ARMA(1,6) model, whereas the ARMA(1,6) performs poorly in the long-term. Similarly, the AR(1) model performs well in the mid,- and long-term but struggles in the short-term relative to the other models. Overall the forecasts seem to perform decently in the short-term, they do better in the mid-term, however, in the long-term they perform poorly struggling to outperform the benchmark.

5 Conclusion

Primarily, our results are consistent with Buncic and Moretto (2015) in that several of our rather simple models produce positive out-of-sample R^2 values. The results indicate that the AR(1) and ARMA(1,6) are strong estimators for the monthly returns of copper. However, the two models are not indicated to consistently outperform the model combinations options, "TOP3" and "EQW", across the tests. It is evident that the models struggle to predict further than two months ahead. The four different test statistics are inconsistent as the Diebold and Mariano (1995) test indicates far less significant relationships than the remaining three tests. The other three tests are more consistent with one another and, additionally, with the findings of Buncic and Moretto (2015). Lastly we find, that the trading of LOCADY on the LME market may contain additional costs potentially rendering the models unprofitable if implemented in a trading scenario.

6 Further Research

Given the surprising tendencies demonstrated by the positive out-of-sample R^2 values obtained by our rather standard models we suggest that further research is appropriate to investigate why the LME copper market is susceptible to be explained by basic models. Subsequently, considering models including additional cost of trading on the LOCADY may be appropriate to examine whether it is actually possible to profit from trading. Furthermore, we suggest that the power of our tests would need to be be verified. If the power of the tests are weak, adapting our methodology on a larger data set or higher data frequency could be interesting as it should increase the power of the tests and thus the quality of the results.

References

Bang, R. (2022). Clark-west test for forecasting performance.

Borenstein, M. (1998). The shift from significance testing to effect size estimation.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika 52*(3), 345–370.

Buncic, D. and C. Moretto (2015). Forecasting copper prices with dynamic averaging and selection models. *The North American Journal of Economics and Finance 33*, 1–38.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of financial studies 21*(4), 1509–1531.

Cheung, Y.-W. and K. S. Lai (1995). Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics 13*(3), 277–280.

Clark, T. E. and M. W. McCracken (2010). Testing for unconditional predictive ability. *Federal Reserve Bank of St. Louis Working Paper No.*

Clark, T. E. and F. Ravazzolo (2012). The macroeconomic forecasting performance of autoregressive models with alternative specifications of time-varying volatility.

Clark, T. E. and K. D. West (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of econometrics 135*(1-2), 155–186.

Clark, T. E. and K. D. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics 138*(1), 291–311.

Cochrane, J. H. (2009). *Asset pricing: Revised edition.* Princeton university press.

Costantini, M. and R. M. Kunst (2011). On the usefulness of the diebold-mariano test in the selection of prediction models: some monte carlo evidence.

Díaz, J. D., E. Hansen, and G. Cabrera (2021). Economic drivers of commodity volatility: The case of copper. *Resources Policy 73*, 102224.

Dickey, D. A. and W. A. Fuller (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica 49*(4), 1057–1072.

Diebold, F. and R. Mariano (1995). Comparing predictive accuracy. *Journal of business economic statistics 13*(3), 253–263.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of business economic statistics 33*(1), 1–1.

Elliott, G. and A. Timmermann (2005). Optimal forecast combination under regime switching. *International economic review (Philadelphia) 46*(4), 1081–1102.

Elliott, G. and A. Timmermann (2016). *Economic Forecasting*. Princeton University Press.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy 81*(3), 607–636.

Fildes, R. and S. Makridakis (1988). Forecasting and loss functions. *International Journal of Forecasting 4*(4), 545–550.

Gargano, A. and A. Timmermann (2014). Forecasting commodity price indexes using macroeconomic and financial predictors. *International Journal of Forecasting 30*(3), 825–843.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Granger, C. W. (1999). Outline of forecast theory using generalized cost functions. *Spanish Economic Review 1*(2), 161–173.

Ha, J., M. A. Kose, and F. Ohnsorge (2021). One-stop source: A global database of inflation.

Hansen, P. R. and A. Timmermann (2012). Choice of sample split in out-of-sample forecast evaluation .

Hansen, P. R. and A. Timmermann (2015). Comment. *Journal of Business & Economic Statistics 33*(1), 17–21.

Heinze, G., C. Wallisch, and D. Dunkler (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical journal 60*(3), 431–449.

Ibisevic, S. (2022). Diebold-mariano test statistic.

Johannes, M., A. Korteweg, and N. Polson (2014). Sequential learning, predictability, and optimal portfolio returns. *The Journal of Finance 69*(2), 611–644.

Kilian, L. and X. Zhou (2018). Modeling fluctuations in the global demand for commodities. *Journal of International Money and Finance 88*, 54–78.

Koehrsen, W. (2018). Overfitting vs. underfitting: A complete example. *Towards Data Science*.

Kriechbaumer, T., A. Angus, D. Parsons, and M. R. Casado (2014). An improved wavelet–arima approach for forecasting metal prices. *Resources Policy 39*, 32–41.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics 54*(1), 159–178.

Lee, T.-H. (2007, 04). Loss functions in time series forecasting.

Lee, T.-H. (2008). Loss functions in time series forecasting. *International encyclopedia of the social sciences*, 495–502.

Liu, C., Z. Hu, Y. Li, and S. Liu (2017). Forecasting copper prices by decision tree learning. *Resources Policy 52*, 427–434.

Newey, W. K. and K. D. West (1986, April). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. Working Paper 55, National Bureau of Economic Research.

Nguyen, D. K. and T. Walther (2020). Modeling and forecasting commodity market volatility with long-term economic and financial variables. *Journal of Forecasting 39*(2), 126–142.

Otgochuluu, C., L. Altangerel, G. Battur, C. Khashchuluun, and G. Dorjsundui (2021). A game theory application in the copper market. *Resources Policy 70*, 101931.

Palachy, S. (2019). Stationarity in time series analysis. *Recuperado desde https://towardsdatascience. com/stationarity-in-time-series-analysis-90c94f27322*.

Pesaran, M. and A. Timmermann (2002). Market timing and return prediction under model instability. *Journal of Empirical Finance 9*(5), 495–510.

Roache, S. K. and M. Rossi (2010). The effects of economic news on commodity prices. *The Quarterly Review of Economics and Finance 50*(3), 377–385.

Sethi, N., S. Gupta, K. Nagar, B. H. Rajan, and T. MahaveerUniversity (2014). Impact of consumer sentiments on precious metals. *International Journal 1*(1).

Stambaugh, R. F. (1999). Predictive regressions. *Journal of financial economics 54*(3), 375–421.

Timmermann, A. (2018). Forecasting methods in finance. *Annual review of financial economics 10*(1), 449–479.

Wagenmakers, E.-J. and S. Farrell (2004). Aic model selection using akaike weights. *Psychonomic bulletin & review 11*(1), 192–196.

A Appendix: Additional plots and tables

Table A1: Two step DM Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		-0,6302	-0,8448	-0,9249	-0,756	-0,959	-0,7571	-0,9452	-0,931	-0,7647
MA(6)	0,6302		-1,0525	-1,2493	0,5207	-1,452	-0,8332	-1,8463*	-1,286	-0,8861
AR(1)	0,8448	1,0525		-0,8836	0,8544	0,1281	0,2376	0,5959	-0,6509	2,0879**
ARMA(1,1)	0,9249	1,2493	0,8836		0,9633	0,6423	0,6015	0,8728	0,5608	1,5714
ARMA(1,3)	0,756	-0,5207	-0,8544	-0,9633		-1,0141	-0,7457	-1,0168	-0,971	-0,7486
ARMA(1,6)	0,959	1,452	-0,1281	-0,6423	1,0141		0,0247	0,9883	-0,6378	0,5654
VAR(1)+	0,7571	0,8332	-0,2376	-0,6015	0,7457	-0,0247		0,7865	0,1287	0,5095
EQW	0,9452	1,8463*	-0,5959	-0,8728	1,0168	-0,9883	-0,7865		-0,8853	-0,2925
TOP3	0,931	1,286	0,6509	-0,5608	0,971	0,6378	-0,1287	0,8853		1,7962*
Random Walk	0,7647	0,8861	-2,0879**	-1,5714	0,7486	-0,5654	-0,5095	0,2925	-1,7962*	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

Table A2: Three step DM Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		-0,477	-0,8954	-0,7291	-0,639	-0,9171	-0,9316	-0,8985	-0,9785	-0,8458
MA(6)	0,477		-1,1247	-0,8534	-0,3687	-1,2147	-1,1698	-1,3915	-1,477	-1,0709
AR(1)	0,8954	1,1247		0,4082	0,9926	0,7304	0,1507	0,86	0,7108	1,3184
ARMA(1,1)	0,7291	0,8534	-0,4082		0,7558	0,1104	-0,2118	0,4188	0,2554	-0,0647
ARMA(1,3)	0,639	0,3687	-0,9926	-0,7558		-1,0334	-1,0417	-1,1068	-1,2048	-0,9302
ARMA(1,6)	0,9171	1,2147	-0,7304	-0,1104	1,0334		-0,7208	0,7192	0,5234	-0,3233
VAR(1)+	0,9316	1,1698	-0,1507	0,2118	1,0417	0,7208		0,8438	0,7194	0,3607
EQW	0,8985	1,3915	-0,86	-0,4188	1,1068	-0,7192	-0,8438		-0,8443	-0,6947
TOP3	0,9785	1,477	-0,7108	-0,2554	1,2048	-0,5234	-0,7194	0,8443		-0,4848
Random Walk	0,8458	1,0709	-1,3184	0,0647	0,9302	0,3233	-0,3607	0,6947	0,4848	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

Table A3: Two step GW Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		4,924*	5,339*	-3,168	0,521	-0,458	-4,363	-0,75	-2,005	7,159**
MA(6)	-4,924*		-2,496	-2,947	-10,266***	-2,463	-4,451	-3,904	-2,377	2,646
AR(1)	-5,339*	2,496		-8,607**	-6,85**	-0,86	-2,185	-7,139**	-3,545	22,037***
ARMA(1,1)	3,168	2,947	8,607**		3,022	5,939*	3,698	3,435	9,296***	9,319***
ARMA(1,3)	-0,521	10,266***	6,85**	-3,022		-4,054	-7,11**	-5,253*	-4,229	7,217**
ARMA(1,6)	0,458	2,463	0,86	-5,939*	4,054		-4,322	1,074	1,45	3,976
VAR(1)+	4,363	4,451	2,185	-3,698	7,11**	4,322		0,876	0,95	2,482
EQW	0,75	3,904	7,139**	-3,435	5,253*	-1,074	-0,876		-2,412	7,971**
TOP3	2,005	2,377	3,545	-9,296***	4,229	-1,45	-0,95	2,412		9,803***
Random Walk	-7,159**	-2,646	-22,037***	-9,319***	-7,217**	-3,976	-2,482	-7,971**	-9,803***	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

Table A4: Three step GW Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		4,842*	2,626	-0,142	4,528	-1,564	-5,016*	3,729	-1,445	4,424
MA(6)	-4,842*		-2,81	-8,527**	-9,756***	-3,907	-4,982*	-5,306*	-5,416*	-0,491
AR(1)	-2,626	2,81		-3,385	3,417	-6,166**	-2,037	0,399	-7,054**	32,597***
ARMA(1,1)	0,142	8,527**	3,385		8,576**	-4,841*	-3,339	9,402***	8,713**	11,21***
ARMA(1,3)	-4,528	9,756***	-3,417	-8,576**		-7,423**	-2,625	-9,658***	-9,275***	2,665
ARMA(1,6)	1,564	3,907	6,166**	4,841*	7,423**		-0,707	2,009	5,34*	9,572***
VAR(1)+	5,016*	4,982*	2,037	3,339	2,625	0,707		2,965	1,697	4,298
EQW	-3,729	5,306*	-0,399	-9,402***	9,658***	-2,009	-2,965		-5,314*	2,856
TOP3	1,445	5,416*	7,054**	-8,713**	9,275***	-5,34*	-1,697	5,314*		11,564***
Random Walk	-4,424	0,491	-32,597***	-11,21***	-2,665	-9,572***	-4,298	-2,856	-11,564***	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

Table A5: Two step CW Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		1,1701	0,1826	-0,0085	0,6516	-0,4102	1,2661	-0,4949	-0,2411	0,4084
MA(6)	0,199		-0,6781	-1,0583	-0,0584	-1,3373	1,1226	-2,4774	-1,3279	-0,3393
AR(1)	1,173	1,6967**		0,2258	1,2467	0,1963	2,3172**	0,6464	-0,6209	2,0072**
ARMA(1,1)	1,1969	2,0563**	0,8408		1,223	0,5722	2,0294**	0,8449	-0,0193	1,4204*
ARMA(1,3)	0,8787	1,5245*	0,2761	-0,0758		-0,3344	1,3335*	-0,5077	-0,2147	0,5636
ARMA(1,6)	1,4253*	2,5455***	1,1994	1,084	1,4988*		2,3815***	1,5134*	0,5476	1,7218**
VAR(1)+	1,4843*	1,933**	0,829	0,929	1,5995*	0,8941		1,0782	1,0795	1,2
EQW	1,147	2,98***	0,2318	-0,0954	1,2381	-0,5466	1,672**		-0,5136	0,6918
TOP3	1,2612	2,1334**	1,2173	0,8552	1,3253*	0,4227	2,0711**	0,9918		1,9248**
Random Walk	1,1019	1,5278*	-1,6355	-0,1736	1,1788	-0,1332	1,9243**	0,397	-1,0604	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

Table A6: Three step CW Test Statistics

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	VAR(1)+	EQW	TOP3	Random Walk
MA(3)		1,5455*	-0,2993	-0,1672	1,1478	-0,1717	0,7647	-0,4042	-0,7182	-0,2462
MA(6)	-0,3509		-1,2505	-0,9868	-0,0464	-0,9595	0,3973	-2,0801	-1,89	-1,2675
AR(1)	1,2489	2,0944**		1,0883	1,7836**	1,0475	1,5092*	1,2985*	0,737	1,2029
ARMA(1,1)	0,8177	1,6969**	-0,1688		1,215	0,4205	1,1819	0,5539	0,1846	0,132
ARMA(1,3)	0,0055	1,9544**	-0,5941	-0,5623		-0,4044	0,5068	-1,1012	-1,0754	-0,5459
ARMA(1,6)	1,3356*	2,4182***	-0,0095	0,7629	1,9629**		1,3581*	1,3601*	0,5053	0,4301
VAR(1)+	1,7315**	2,4161***	1,025	1,1752	2,1559**	1,4522*		1,0495	1,1798	1,2641
EQW	0,905	2,5981***	-0,6347	-0,121	1,79**	-0,2034	1,3751*		-1,3791	-0,4807
TOP3	1,2985*	2,5977***	-0,2434	0,4809	2,1251**	0,2926	1,3545*	1,7823**		0,0653
Random Walk	1,192	2,0645**	-0,8621	0,9074	1,7402**	0,7418	1,4179*	1,1784	0,5116	

* denotes that it is significant at 90%, ** denotes that it is significant at 95%, *** denotes that it is significant at 99%

+ estimated on a smaller sample

Table A7: Forward contract out-of-sample R^2

	MA(3)	MA(6)	AR(1)	ARMA(1,1)	ARMA(1,3)	ARMA(1,6)	EQW	TOP3
1 step	-0,1303	-0,0687	-1,0506	-0,1650	-0,0356	-0,1346	-0,0664	-0,0373
2 step	-0,0832	0,0936	-0,7729	-0,1218	-0,0306	-0,0891	0,0388	0,0168
3 step	-0,2277	-0,0846	-1,5177	-0,0957	-0,1499	-0,2464	-0,1694	-0,0857

Table A8: Out-of-sample R^2 ranking from largest to smallest for each horizon

1step		2step		3step	
Model	R^2	**Model**	R^2	**Model**	R^2
ARMA(1,6)	0,0326	ARMA(1,6)	0,0485	AR(1)	-0,0035
TOP3	0,0127	TOP3	0,0470	Random walk	-0,0224
EQW	0,0006	ARMA(1,1)	0,0327	ARMA(1,6)	-0,0313
Random walk	-0,0224	VAR(1)+	0,0168	TOP3	-0,0354
MA(3)	-0,0246	AR(1)	0,0138	VAR(1)+	-0,0380
AR(1)	-0,0291	EQW	-0,0079	ARMA(1,1)	-0,0472
ARMA(1,3)	-0,0622	Random walk	-0,0224	EQW	-0,0767
MA(6)	-0,0940	ARMA(1,3)	-0,1116	MA(3)	-0,1423
ARMA(1,1)	-0,1154	MA(3)	-0,1256	ARMA(1,3)	-0,1818
VAR(1)+	-0,2889	MA(6)	-0,1639	MA(6)	-0,2330

Table A9: Out-of-sample R^2 rankings for horizon each and model

	1step	**2step**	**3step**	**Average**
ARMA(1,6)	1	1	3	1,67
TOP3	2	2	4	2,67
AR(1)	6	5	1	4,00
Random walk	4	7	2	4,33
EQW	3	6	7	5,33
ARMA(1,1)	9	3	6	6,00
VAR(1)+	10	4	5	6,33
ARMA(1,3)	7	8	6	7,00
MA(3)	5	9	8	7,33
MA(6)	8	10	10	9,33

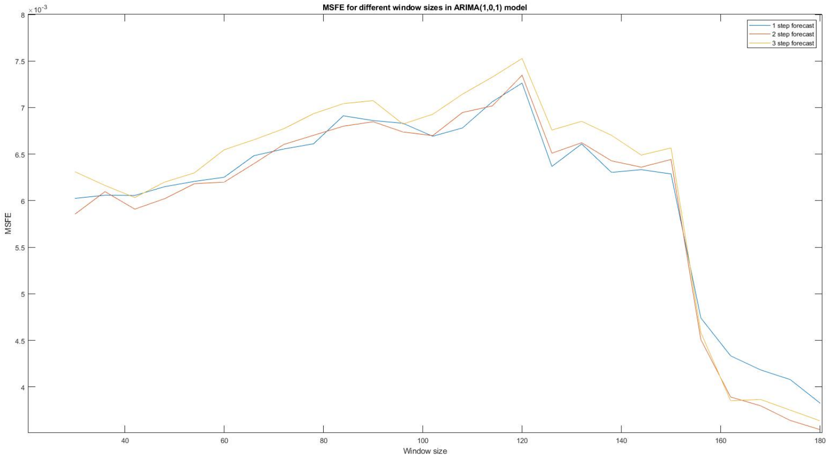

Figure A1: MSFE ARMA(1,1)

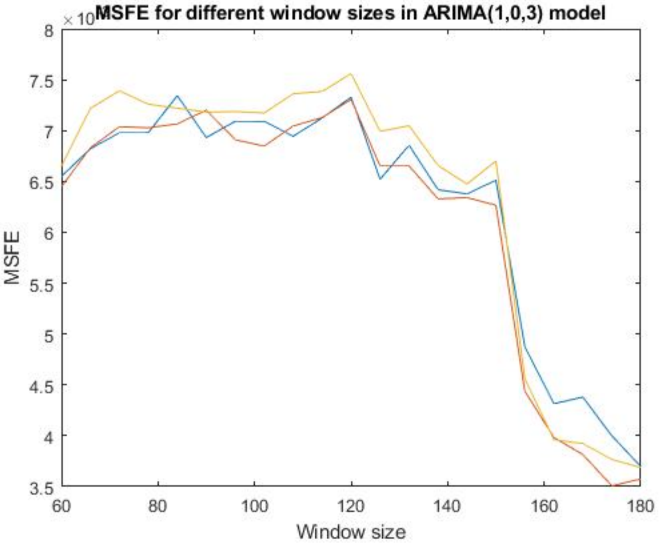

Figure A2: MSFE ARMA(1,3)

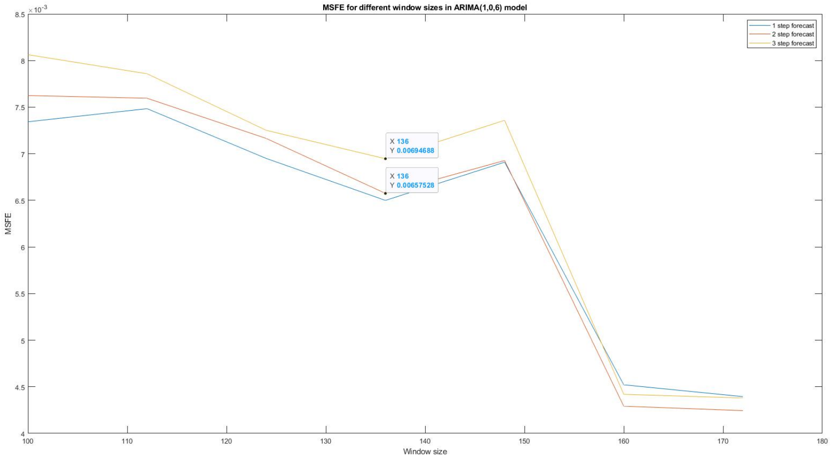

Figure A3: MSFE ARMA(1,6)