Norwegian School of Economics
Bergen, Autumn, 2021

NHH

Forecasting the Price of Aluminium Using Machine Learning

Empirical comparison of machine learning and statistical methods.

Stina Johanne Mysen and Elisabeth Marie Thornton

Supervisor: Øivind Anti Nilsen

Master thesis, MSc in Economics and Business Administration,
Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Acknowledgement

This master thesis is a part of our master's degree in economy and administration, majoring in Business Analytics at the Norwegian School of Economics (NHH).

First and foremost, we would like to thank our supervisor Øivind Anti Nilsen for guiding us throughout this paper and contributing valuable input and feedback. Additionally, we want to thank Jonas Andersson and Johan Lyhagen for taking their time to provide us with useful inputs and their expert knowledge about forecasting in the process. Finally, we thank Hydro for providing us with significant data material to conduct the analysis and valuable insights into the aluminium market.

Norwegian School of Economics

Bergen, December 2021

Stina Johanne Mysen

Elisabeth Marie Thornton

Abstract

This thesis challenges statistical methods for forecasting aluminium's 3-month futures contract price with a machine learning technique. The goal is to find the superior model with a one-month horizon.

The first model we apply is the traditional statistical method, random walk, as a benchmark. After that, we examine the relationship between the futures and spot price by conducting the Johansen cointegration test. This test concludes that the two prices are cointegrated. Exploiting this relationship, we conduct a forecast using the more advanced statistical method, the vector error correction model (VECM). Lastly, we compare the results to the popular machine learning technique XGBoost.

To validate the models' predictions, we use the time series cross-validation technique called rolling cross-origin. After that, we evaluate the performance of the three models by comparing the root mean square error (RMSE) and mean absolute error (MAE).

Our research finds that the XGBoost method stays relatively stable through multiple robustness tests in contrast to the VECM, varying significantly for the majority of the tests. In conclusion, the most reliable and accurate model to use when forecasting the 3-month futures contract price is XGBoost.

Keywords – Aluminium, forecasting, time series analysis, machine learning, statistical methods, XGBoost, vector error correction model, random walk, cross-validation

Contents

List of Figures

List of Tables

1. Introduction

Throughout history there has been a fascination and interest in the ability to produce accurate forecasts. The most traditional methods are statistical forecasting approaches, such as the ARIMA model and vector autoregressive method. However, a new competitor has arrived in recent years, namely machine learning. Machine learning is a subfield of artificial intelligence, with the advantage of not being explicitly programmed but rather learning from historical data. The algorithm lets the system identify patterns that the human brain might overlook (Marr, 2016). Hence, there is potential in improving the accuracy of statistical forecasting methods with popular machine learning techniques.

This thesis focuses on forecasting the 3-months futures contract price of aluminium, with a one-month horizon. Aluminium is one of the most essential metals on earth due to its ease in manufacturing various products. Compared to other metals, aluminium has a high strength to weight ratio, is easy to form, suitable for mass production and easy to recycle. Its uses are many and vary from construction, automobile and aircraft manufacturing to food packages.

In terms of trading volumes, aluminium is the largest commodity exchange for metals in the world. A commodity exchange is a legal entity where future delivery contracts (futures) are traded. The world centre for trading metal is the London Metal Exchange (LME), trading about 3,7 billion tonnes of metals annually (UC RUSAL, 2015). It is rarely actual physical delivery of the metal but instead trading with professional market agents/traders using aluminium as a financial instrument (UC RUSAL, 2015). The 3-month futures contract is, without doubt, the most liquid and thus most traded contract compared to other contract lengths (Narin, 2019).

Forecasting aluminium's supply, demand, and price underpins future planning and investment decisions in production and processing industries. Therefore, a price forecast model is of interest to traders and other agents in the aluminium business as it contributes to better decision making.

The global economic stability is affected by the volatility in the aluminium market, especially concerning aluminium exporting countries (Kriechbaumer et al., 2013). Given its

versatility, numerous products used in everyday life are manufactured from aluminium. This makes it the second-ranked metal (after steel) in terms of volume consumption (UC RUSAL, 2015). Therefore, we are all more reliant on aluminium production than we might know, and an accurate forecasting model benefits agents without a direct connection to the aluminium business.

In the rankings of the world's largest aluminium producers, Western Europe is represented by the Norwegian company Hydro. The history of Hydro reaches back to 1905 when it started making valuable products out of natural resources and has since grown into the ninth biggest aluminium producer in the world (Bryhn & Gram, 2021). Due to their extensive market knowledge and important position in the industry, we use their guidance and data in this thesis.

Hydro is divided into different departments, including the Metal Sourcing & Trading department. Trading aluminium is one of its daily activities. Regarding forecasting aluminium *futures* prices, CRU[1] consulting firm provides Hydro with long-term annual forecasts. As CRU only conducts long-term forecasts, Hydro demands a short-term forecast to support the workers in trading strategies. However, some researchers, such as Dooley and Lenihan (2005) and Chen et al. (2010), examine short-term forecasts on the aluminium *spot* price. Hence, there is a gap in short-term forecast on the aluminium 3-month *futures* price.

To our knowledge, there is a lack of machine learning techniques when forecasting aluminium prices. This makes us curious whether we could achieve higher accuracy by exploring machine learning.

This thesis challenges statistical methods with machine learning when forecasting aluminium's 3-month futures contract price. To determine the superior model, we compare the accuracy of the statistical method VECM to the comprehensive machine learning technique XGBoost, using the simple statistical approach random walk as a benchmark.

[1] CRU market outlooks are not public, but Hydro has access. The outlooks were sent from Hydro. Their home page can be found at: https://www.crugroup.com/

2. Background

2.1 The aluminium market

In terms of supply, the biggest aluminium producing countries are China, followed by India and Russia. Despite Norway being a small country, it is a remarkably significant aluminium producer, ranked as the eighth largest (U.S. Geological Survey, 2021). The process, however, is not straightforward. The first step of the production is mining bauxite, a clay-like soil found in a belt around the equator (Hydro, 2021). Then, the bauxite is transported to plants worldwide, where alumina, or aluminium oxide, is extracted by a hot solution of caustic soda (Hydro, 2021). After that, the alumina is transported to the metal plant, further transformed into aluminium. Additional necessary raw materials in this refining process are electricity and carbon. The biggest alumina production countries in the world are China, Australia, and Brazil (U.S. Geological Survey, 2021).

Aluminium is a prevalent metal resulting in an increasing demand every year. More specifically, the average aluminium demand grows 5 – 7 % annually (UC RUSAL, 2015). This trend is expected to continue at an unstoppable rate (Doshi & Prasad, 2019). The aluminium market is commonly divided into two parts, China and the world excluding China. This is due to China's remarkable growth over the last decade (UC RUSAL, 2015). In terms of aluminium trading, China has a comparable metal exchange to LME, named Shanghai Futures Exchange (SHFE). Trade flows connect the Chinese and London metal markets as the difference in price can incentivise imports and exports of metals. This results in a rise in arbitrage activity between the two markets (LME, 2017).

2.2 Machine learning history

Machine learning has been a topic of conversation for years and has attracted increased interest as it continuously gets more relevant and revolutionary. To understand what machine learning is, the concept of artificial intelligence is essential to comprehend. Artificial intelligence is computer science used to create machines that can replicate how the human brain works, from the thinking processes to human behaviour. Such machines are called *intelligent machines* (McCarthy, 2007). A subfield of artificial intelligence is machine learning. Instead of having to be explicitly programmed, the machine learning algorithms let the computer learn from the past and improve independently.

Machine learning goes far back in time. In 1952, Arthur Samuel invents the first learning program on a computer. The thought process of the human brain is simulated in 1957 when Frank Rosenblatt designs the first neural network. In the 1990s, scientists shift the focus from knowledge-driven to data-driven, a huge step in machine learning (Marr, 2016). Consequently, this shift in focus results in the designing of programmes that manage vast amounts of data and can conclude thereafter.

In the years following, machine learning gets developed even further, being used by large platforms such as Google, Facebook, Amazon and Microsoft. Today, machine learning enables computers to beat professional chess players, communicate with humans and even drive cars autonomously. A computer's ability to analyse data, understand and interact with the world is growing at a remarkable rate (Marr, 2016). As Marr (2016) firmly believes, "machine learning will severely impact most industries and jobs within them, which is why every manager should have at least some grasp of what machine learning is and how it is evolving". Machine learning attracts increased attention in the forecasting community and is looked upon as a serious competitor to statistical approaches (Bontempi et al., 2012).

3. Literature review and established theories

In terms of commodities, several studies examine the accuracy of different forecast models. To the best of our knowledge, there is a lack of literature on forecasting aluminium with a high impact factor[2]. Therefore, we look at research concerning commodity prices beyond aluminium. After a deep dive into previous articles, it turns out that traditional statistical methods are widely used when forecasting commodity prices. When forecasting time series, machine learning techniques have got substantial interest over the last years, posing as a strong competitor to statistical methods.

3.1 Forecasting methods for commodity prices

As the pricing of commodity futures contracts have attracted substantial interest, two theories have been developed, namely, the Cost-of-Carry and the Risk Premium hypothesis (Chow et al.,2002). The theories differ when it comes to whether the commodity is storable. If the commodity is storable, the Cost-of-Carry can be applied, while the Risk Premium hypothesis is used when the commodity is not storable.

Kaldor (1939) and Working (1948;1949) are the first to formalise the Cost-of-Carry hypothesis. The basis of this hypothesis is that the futures prices are the underlying commodity spot price plus carrying cost. The return of buying the commodity at time *t* and selling it with a delivery date equal to time *T* is:

$$F(t,T) - S(t) = S(t)R(t,T) + W(t,T) - C(t,T) \tag{3.1}$$

where $F(t,T) - S(t)$ represents the return, $S(t)R(t,T)$ is the capital interest lost and $W(t,T)$ is the cost of storing the commodity. Having a commodity can be argued to hold a value, because of the flexibility one is provided due to, for example, unexpected demand.

[2] Impact factor is a measure of the frequency with which the average paper in a journal is cited during a year.

This advantage is called the convenience yield and is represented in the equation as $C(t,T)$ (Fama & French, 1987).

The risk premium hypothesis states that the return defined above can be divided into an expected premium and an expected change in the spot price (Fama & French, 1987). To express this, one can formulate the following equation:

$$F(t,T) - S(t) = E_t[P(t,T)] + E_t[S(T) - S(t)] \tag{3.2}$$

where $E_t[P(t,T)]$ represents the expected risk premium. The premium is defined as the bias of the futures price as a forecast of the future spot price. $E_t[S(T) - S(t)]$ is defined as the change in expected spot price (Fama & French, 1987).

Fama and French (1987) argue that the two equations 3.1 and 3.2 are alternatives to each other. The variation in the interest rate and the storage cost in the Cost-of-Carry hypothesis can translate into the expected change in the spot price in the latter equation.

Coppola (2008) considers the relationship between the spot price and futures price when forecasting the oil price. As many financial time series, including commodity spot and futures prices, are often non-stationary (Watkins & McAleer, 2006), there is a possibility of obtaining spurious results. This means that the model indicates a relationship between variables where they do not exist. Therefore, Coppola (2008) performs a cointegration test to explore the relationship between the price of futures contracts and spot prices before conducting the forecasting method. Conclusively, Coppola (2008) states that there exists a cointegration relationship, resulting in the use of the vector error correction model (VECM) when forecasting. Results from the study indicate that the VECM outperforms the random walk model (RWM) when forecasting the 1-month futures contract price.

Pradhan et al. (2021) also examine the relationship between the spot prices and futures prices of different commodities. As a result, they find that aluminium has a long-run equilibrium relationship between the spot and futures price. On the London Metal Exchange, the principle of futures-spot convergence holds precisely (Jia & Kang, 2021). This principle states that the futures contract price converges towards the spot price of the underlying

commodity, resulting in the price of the futures contract being roughly equal to the spot price at the delivery date (Chow et al.,2002). Therefore, it is reasonable to expect a long-run relationship between the two prices of a given commodity.

The auto-regressive integrated moving average (ARIMA) is another frequently used statistical method for forecasting commodities. Dooley and Lenihan's (2005) study applies the ARIMA method and compares it with lagged forward price models to forecast monthly lead and zinc spot prices. The conclusion of this research is not definitive. Regarding zinc, they do not find any conclusive findings suggesting that one method is superior to the other. On the other hand, the ARIMA method outperforms the lagged forward price models nine out of sixteen times for the lead price.

As previously mentioned, it is during the 1990s that the focus on machine learning changes from being knowledge-driven to being data-driven (Marr, 2016). In 1996, a study by Kohzadi et al. (1996) challenges the traditional ARIMA model with neural network models for forecasting commodity prices. Neural network is a branch of machine learning and has been significantly developed since then. Even in the early stages, the result of Kohzadi et al.'s study indicates that the neural network provides higher accuracy than the regular ARIMA with a three-year forecast horizon. Additionally, Kohzadi et al. state that neural network methods are suitable for other forecasting problems, like stocks, other financial prices, and commodity prices (Kohzadi et al., 1996). Later, Panella et al. (2012) research the use of a more advanced neural network approach for forecasting energy prices. They also find that neural networks compute accurate predictions on a long forecasting horizon.

In 2014 a new machine learning algorithm is introduced, namely Extreme Gradient Boosting (XGBoost). The algorithm is looked upon as a highly effective method used extensively in *Kaggle*[3]. XGBoost is used in 17 out of 29 challenge winning solutions in 2015, demonstrating its wide use among experts (Chen & Guestrin, 2016). A time series analysis by Gumus & Kiran uses the XGBoost for forecasting crude oil prices. Jabeur et al. (2021) also conduct XGBoost to forecast commodity prices, but they examine the gold price instead of oil. Jabeur et al. conclude that the method is superior to other machine learning approaches.

[3] *Kaggle* is a platform for competitions in Data Science

To our knowledge, there is a lack of literature where the XGBoost approach is used to forecast aluminium prices. This motivates us to examine if one can achieve higher accuracy than statistical methods by using the assumingly comprehensive machine learning method XGBoost.

3.2 Variables with predictive power over the aluminium price

Various factors affect the price of aluminium and knowing the interaction of these variables is difficult to ascertain. The metal market is complex, combined with stochastic economic processes influencing pricing, making price forecasting difficult (Kriechbaumer et al., 2014). In that manner, we investigate literature that includes explanatory variables when forecasting commodity prices.

Some countries are heavily dependent on the export of a given commodity, resulting in their currency following the world price of the primary commodity product (Chen et al., 2010). These currencies are called commodity currencies. In addition, Chen et al. state that these currencies have a robust predictive power over commodity prices, like aluminium. The theoretical foundation behind this statement is the following: firstly, commodity prices determine the country's nominal exchange rate value. This is because the commodity prices represent the terms of trade for the given country. Secondly, the nominal exchange rates incorporate future fundamental value expectations, such as commodity prices, and should assist in their prediction (Chen et al., 2010). However, the article by Chen et al. does not find robust evidence of the reverse relationship, namely the commodity price having predictive power over the exchange rate.

Gargano and Timmermann (2014) state the same matter as Chen et al., but instead of examining the forecast of commodity prices, they analyse commodity price indexes. The analysis determines that the two commodity currencies used, namely the Australian dollar - US dollar and the Indian rupee - US dollar exchange rates, have some predictive power when forecasting short term horizons. Not only do Gargano and Timmerman investigate the effect of commodity currencies when forecasting commodity prices indexes, but they also look at

what predictive power other macroeconomic and financial variables have. At a monthly horizon, the study implies that the long-term return, the lagged returns, and the US–Australian dollar exchange rate have some predictive power, especially over metals.

Another study that examines the determinants of metal prices is Labys et al.'s (1999), investigating the macroeconomic influence on the common factor. The common factor can be explained as the co-movement between metal prices. The study concludes that industrial activity is the variable with the highest predictive power over metal prices.

4. Methodology

Multiple statistical models and machine learning algorithms are relevant when forecasting the 3-month futures contract price of aluminium. Following our extensive research, we restrict this study to only look at two statistical forecasting methods, namely the random walk model (RWM) and the vector error correction model (VECM), and the machine learning technique XGBoost.

4.1 Forecast horizon

As there already exists long term forecasts from CRU, a short-term forecast is demanded by Hydro and other stakeholders in the aluminium business. A short-term forecast may help the stakeholders in production planning and control, identifying short-term cash requirements and necessary business adjustments. For aluminium traders especially, a short-term forecast may enhance decision making regarding trading strategies for the upcoming months. Hence, an important goal of this paper is to propose an applicable and well-performing forecasting model of the futures price for stakeholders in the aluminium business.

It is a fact that short-term forecasts are more accurate than long-term forecasts, as a long-term forecast increases the chance of unforeseen changes impacting future prices significantly (Nissi et al., 2021). The difference between long-term and short-term forecasts is simply the forecast horizon. The time span for a short-term forecast is less than one year, often ranging from one to three months (Wisdom IT Services India Pvt. Ltd, 2020), while long-term forecasts often have a horizon longer than one year.

A disadvantage of a longer forecasting horizon than one step ahead is that the future explanatory variables must be predicted to include them in the model. Apart from being more computationally expensive, the forecasting model will contain even more uncertainty as the predicted explanatory variables include additional errors and unknowns.

With the reasons mentioned above in mind, we find it most applicable to forecast one step ahead. In our case, this corresponds to using a one-month forecast horizon.

4.2 Cross-Validation

Forecasting accuracy tells us how well the model performs on new data. The only way to determine this is by splitting the data into two parts: a training set and a test set. The training set consists of the data we use to fit the model, while the test set is used to evaluate the performance of the fitted model (Hyndman & Athanasopoulos, 2021). In other words, we use the training set to fit a model that predicts observations in the test set.

The most standard approach for validating forecasting methods is *k-Fold* cross-validation. This method shuffles the data and splits them into *k*-folds, training the data on all folds except from one. The left-out fold represents the test set (James et al., 2013). When predicting forecasts for time series, however, we cannot change the order of the observations. Additionally, actual future data fall into the training set. This causes data leakage, and we achieve unrealistically good results as we predict values that are part of the data the model is training on (De Prado, 2018). We use the time series cross-validation approach called *Rolling Origin Cross-Validation* to avoid this.

4.2.1 Time Series Cross-Validation (Rolling Origin Cross-Validation)

Rolling Origin cross-validation (ROCV) is used to validate time series forecast performance. This approach starts with a forecasting origin, including *k* observations in the training set, forecasting *H* steps ahead. For every round, one observation is added to the training set while the test set moves one step forward (Hydman & Anthanasopoulos, 2021).

Figure 4.1 is an illustration of the incremental construction of the training set. The green blocks represent the training observations, while the blue block is the test set one performs the validation on. This thesis uses rolling origin with a constant holdout sample (test set) as we are only interested in predicting one step ahead. This process is repeated for all observations in the data set.

An advantage of using the ROCV is to detect if the model is overfitted. Overfitting occurs when the pattern observed by the model, based on the training set, does not fit the test set. In other words, the performance of the predictions in-sample is excellent, while forecasts

performed out-of-sample are less accurate (Hyndman & Athanasopoulos, 2021). A weakness of the ROCV is that one only uses a small part of the data set when computing the first predictions causing the errors to be calculated on inadequate training data (De Prado, 2018).

Training and test set

It is necessary to split our data into a training set and a test set to evaluate the models' predictions. When deciding the size of each set, the initial training set must be big enough to make predictions on a sufficient amount of the original data. Typically, the test set accounts for 20 % of the total data set (Hyndman & Athanasopoulos, 2021). Therefore, our initial training data spans from observation t = 1 to t = 90, corresponding to approximately 80 % of the data set. The initial test set will be t = 91, corresponding to August 2019.

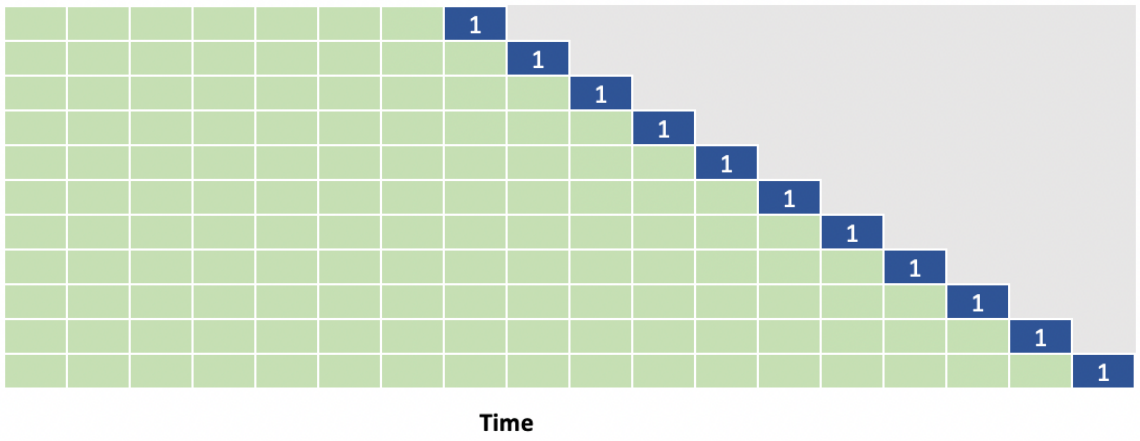


Figure 4.1: Example of time series cross-validation with validation at H = t + 1

4.3 Method Evaluation

After performing cross-validation on the models, we evaluate the forecasting performance by comparing evaluation metrics. Multiple metrics are available, where two commonly used are the root mean squared error (RMSE) and mean absolute error (MAE). The purpose of these metrics is to measure the error of the forecast model.

The main difference between the two metrics is that RMSE gives a relatively high weight to large forecast errors compared to MAE. As large forecast errors are undesirable in this thesis, we mainly focus on RMSE.

4.3.1 Root Mean Square Error

The root mean squared error (RMSE) is an evaluation metric that uses the standard deviation of the prediction errors. In order to make certain positive and negative deviations do not offset each other, the formula calculates the absolute error by taking the square of the deviation.

RMSE weights the large errors relatively high because the squaring of the error is computed before they are averaged. Therefore, the formula takes the squared root of a bigger number. The formula for the RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \widehat{Y}_t)^2} \tag{4.1}$$

where Y_t represents the actual value at observation t, $\widehat{Y}_t$ is the predicted value at time t and n is the number of predicted observations.

4.3.2 Mean Absolute Error

In the same way as RMSE does, MAE avoids the negative and positive deviation to offset each other. Instead of squaring the difference, MAE takes the absolute value of the deviation. In contrast to RMSE, MAE weight the errors equally as it simply takes the average of the absolute deviations. The formula for MAE is presented in equation 4.2.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{4.2}$$

where Y_t is the actual value at observation t, $\widehat{Y_t}$ represents the predicted value at time t and n is the number of predicted observations.

4.4 Bias-variance trade-off

There are two prediction errors that are necessary to comprehend when forecasting, namely bias and variance. The bias represents the difference between the average of the models' predictions and the actual value. Variance, however, tells us how the predictions are spread (James et al., 2013). In simpler terms, the variance tells us how the predictions change when the training data changes (Mottaghinejad, 2021).

When a models complexity increases, it gets more flexible, generating a much more comprehensive range of possible shapes to estimate functions (James et al.,2013). Restrictive models are less complex, for example linear regression, only generating linear functions. Figure 4.2 illustrates the bias-variance trade-off. As the complexity of the model increases, the variance increases but the bias decreases.

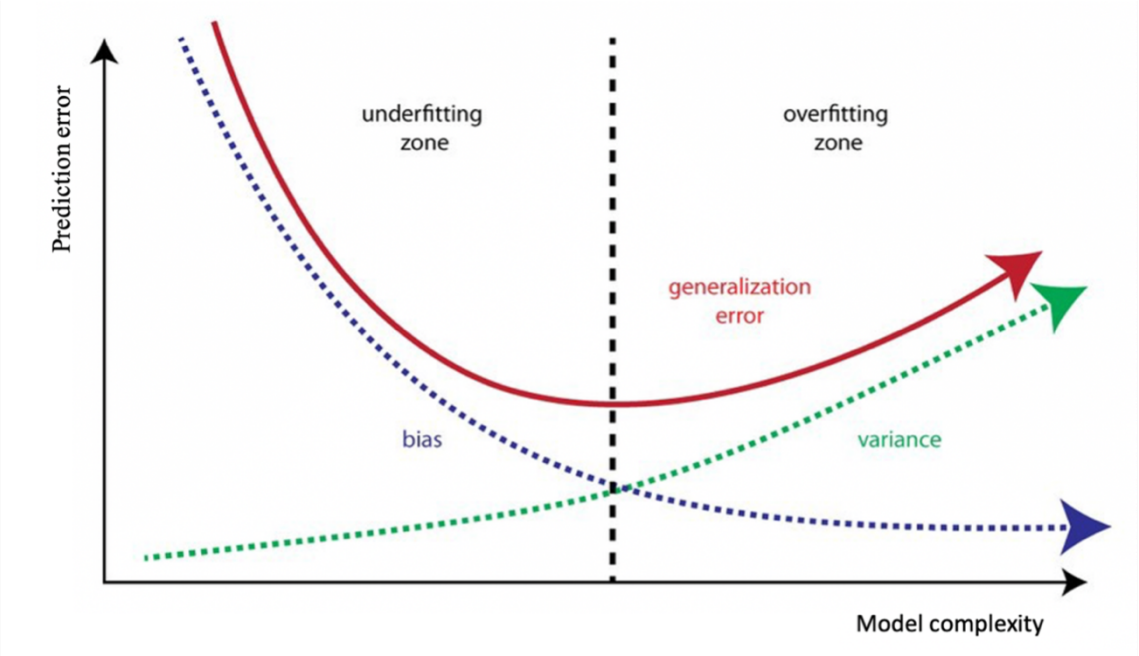


Figure 4.2: Illustration of the bias-variance trade-off. Source: (Mottaghinejad, 2021)

Models on the left-hand side achieve high bias but low variance. They risk underfitting the model, paying little attention to the training data. This results in an oversimplified model. Such a model will not predict accurate forecasts on the training or test set. On the contrary, one may achieve an overfitted model if the model pays much attention to the training data. Overfitted models have high variances but low bias. They achieve high accuracy on the training data but performs poorly on the test data, as they pay too much attention to the specific training set. One aims to find the perfect balance between overfitting and underfitting, illustrated as a black dotted line in figure 4.2.

4.5 Benchmark model – Random walk

When searching for a superior forecast, one must have a benchmark to compare the performance. A benchmark model provides a lower bound of the forecast accuracy. If the challenging model is below this, one can conclude that the model performs poorly.

A random walk model is an oversimplified approach, paying little attention to the training data. The restricted model assumes that the time series follows a random walk, where the

future value is always equal to the last observed value (Hyndman & Athanasopoulos, 2021), as expressed in equation 4.3:

$$\hat{y}_{+H} = y_t \tag{4.3}$$

where $\hat{y}_{+H}$ represents the forecast H step ahead, while y_t is the t last observations.

The random walk model, often known as the naive method, is a commonly used benchmark model in forecasting. This method has been shown to work surprisingly well on non-stationary financial data.

4.6 The VAR framework and VECM

Vector regressive models (VAR) are more complex than the random walk method and are part of the advanced statistical methods used for forecasting multivariate time series. A multivariate time series is when more than one time-dependent variable is present. In VAR, all variables are treated as endogenous, meaning they influence each other simultaneously. Hence, the model involves more than one equation, more specifically, as many equations as endogenous variables (Hyndman & Athanasopoulos, 2021).

The general VAR formula excluding the deterministic terms of a constant and a trend is expressed in equation 4.4.

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + B_1 x_t + B_2 x_{t-1} + \cdots + B_l x_{t-l} + \varepsilon_t \tag{4.4}$$

$$\text{where } y_t = \begin{bmatrix} y_{1,t} \\ \vdots \\ y_{k,t} \end{bmatrix} \text{ and } x_t = \begin{bmatrix} x_{1,t} \\ \vdots \\ x_{d,t} \end{bmatrix}$$

k is the number of endogenous variables affected by d number of exogenous variables. In other words, y_t is a k-dimensional stochastic time series for $t = 1,2, \dots, T$ and x_t represents a

d-dimensional matrix representing the exogenous variables. *B* denotes the corresponding coefficients for the exogenous variables. The lag structure for the endogen variables is represented by *p*, while the lag structure for the exogen variables is denoted as *l*. ε_t is the error term where the mean of each error is zero.

When the data is stationary, we forecast time series by simply fitting a VAR model. If the time series are non-stationary, one can take the differences[4] to make them stationary before fitting the VAR model. However, when the endogenous time series are cointegrated, it is necessary to include an error correction mechanism. This is done by fitting a vector error correction model (VECM) derived from Johansen's cointegration test (1988;1991) instead of a VAR (James et al., 2013). The VECM is a more flexible model than the VAR as it includes the error correction term. The Johansen cointegration test is described in subsection 4.6.1.

Assuming that there exists a cointegration relationship between the endogenous variables, one can transform equation 4.4 to the following formula, representing the VECM model:

$$\Delta y_t = \alpha\beta' y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + \sum_{i=0}^{l-1} B_{i+1} \Delta x_{t-i} + \mu_t + \varepsilon_t \tag{4.5}$$

$$\text{where } \beta' = \begin{bmatrix} \beta_{1,1} & \dots & \beta_{1,r} \\ \vdots & \dots & \vdots \\ \beta_{k,1} & \dots & \beta_{k,r} \end{bmatrix} \text{ and } \alpha = \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,r} \\ \vdots & \dots & \vdots \\ \alpha_{k,1} & \dots & \alpha_{k,r} \end{bmatrix}$$

β' and α, are two *k x r* matrices, where *k* is the number of endogen variables and *r* represents the number of cointegrated relations decided through the Johansen test. Hence, each column of β is the cointegrating vector. α denotes the speed of adjustment parameters. Larger values of α result in a faster convergence towards long-run equilibrium relationships between the variables when short-term deviations are present. $\alpha\beta' y_{t-1}$ is therefore the error correction term. Δ notes the first differences. The vector autoregressive (VAR) component in first

[4] Differencing includes subtracting the time series' consecutive observation. The equations for first- and second order differencing are: $y'_t = y_t - y_{t-1}$ *or* $y''_t = y'_t - y'_{t-1}$, respectively. According to Hyndman and Athanasopoulos (2021) it is almost never necessary to go beyond second-order differences.

differences is denoted as $\sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i}$, while μ_t represents a vector of length *k x 1* consisting of the constants for *k* equations.

4.6.1 Johansen cointegration test

Cointegration between time series is true if the time series have a long-run equilibrium relationship. This does not rule out the possibility of the time series deviating from the linear cointegration relationship in the short run. The prerequisite is that the time series eventually migrate to long-run equilibrium. Economic power is a cause of equilibrium, causing prices to move in relation to each other (Kočenda & Černý, 2015).

The Johansen cointegration (1988;1991) test is a method for determining cointegration relationships. It has the advantage of testing for numerous relationships simultaneously, as opposed to more straightforward tests like the widely used Engle-Granger approach by Engle and Granger (1987). Johansen modifies a vector autoregressive (VAR) framework as described in the former subsection.

One can use two different test statistics in the Johansen test: the trace ($\lambda trace$)- and maximum eigenvalue statistics (λmax)[5]. In order to identify the number of cointegrated relations *(r)*, one compares the test statistics to a critical value. The critical value we use is provided by Osterwald-Lenum (1992).

The hypothesis of the Johansen is divided into *k* stages, depending on how many endogenous variables are included in the VECM. The hypotheses of the VECM with *k* endogenous variables with the belonging *k* equations are explained in table 4.1.

[5] The trace- and maximum eigenvalue functions are found in Appendix A1

Table 4.1: The Johansen cointegration test's hypothesis

	$\lambda trace$ and λmax
Stage one	$H0: r = 0$
	$H1: r > 0$
Stage two	$H0: r \leq 1$
	$H1: r \leq 2$
	.
	.
	.
Stage k	$H0: r \leq k$-1
	$H1: r = k$

The first step has a null hypothesis stating no cointegrated relationship between the variables. If the null hypothesis is rejected (test statistics > critical value), one can claim H1, stating that *at least* one cointegrated relationship exists. Further, one moves to the next step, following the same procedure until the null hypothesis cannot be rejected. If the null hypothesis in the *k*th step is rejected, it implies that the time series are stationary.

4.6.2 Granger causality test

Correlation and causality are measures of how two variables move in relation to each other. In contrast to correlation, one needs to observe that at least one of the variables is causing the other to claim a causal relationship. However, correlated variables do not need to cause each other, but they move in coordination with each other. A variable that is correlated with the dependent variable may be expedient to include in the forecast model without a causal relationship (Hyndman & Athanasopoulos, 2021).

A widely used causality test is the Grangers causality test (1988)[6], which identify short-run causal relationships between time series. To be confident that a cointegration relationship exists, Granger (1988) states that there should be at least one-directional relationship present. The causality test is based on the vector autoregressive framework (VAR). In cases where the data is stationary, the F-test statistics [7] is appropriate to apply. Nevertheless, a study by Lütkepohl and Reimers (1992) find that these tests are applicable for non-stationary data if the model is based on a bivariate VAR process.

4.7 Tree Based Methods - XGBoost

A simple machine learning method is the tree-based approach. By portioning the explanatory variables into a series of leaves (also called sections), the method predicts each leaf. In figure 4.4, each of the vertical lines represents a leaf. These methods can be used for both regression and classification problems and have the advantage of being highly interpretable. This is because a tree structure illustrates the decision trees, making them intuitive to understand (James et al., 2013). A decision tree regarding the prediction of a baseball players salary is illustrated in figure 4.4.

[6] The Granger causality test is described in Appendix A2

[7] F-test is presented in Appendix A3

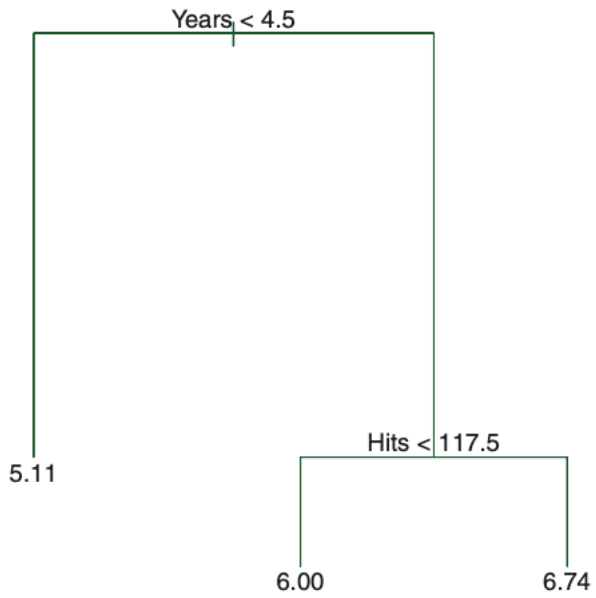


Figure 4.*3*: Illustration of a regression tree. Source: (James et al., 2013)[8].

Although the classic decision tree has its advantages in interpretation and simplicity, it often falls short compared to other methods when comparing accuracy measures. In addition, it is not very robust, resulting in significant changes in the estimated trees if the data changes slightly. In other words, there exists a chance of getting an overfitted model. However, approaches like boosted trees improve the predictive performance (James et al., 2013). This thesis will implement the highly flexible method boosted trees, further explained in the following subsections.

4.7.1 Boosted Trees and XGboost

Boosted trees aim to create multiple decision trees sequentially, meaning that the trees use information from the preceding trees. Each tree is grown on a modified version of the original data set using the residuals of the previous tree to fit the next model. This is done to improve the prediction where the previous model performed poorly (James et al., 2013).

[8] If the baseball player has played in the major leagues for less than 4.5 years, the predicted salary is equal to 5.11. However, if the player has played for more than 4.5 years, we are redirected to the right-hand leaf and encounter a second split. If the player has a hit rate below 117.5, the predicted salary is 6, while if the hit rate is above 117.5, the predicted salary is equal to 6.74.

The basis of boosting is weak learners. Weak learners are not very complex models and produce only slightly better predictions than models using random change. Boosting starts off by training one weak learner, then using the weak points to produce the next – specifically tuned to target the areas the previous weak learners missed. The set of weak learners will result in one single strong learner, which is a model made to give the best performance possible (Toprak, 2020). To explain how boosting works, a simplified algorithm described by James et al. (2013) is provided in Appendix A4.

Gradient boosting is a popular technique for implementing boosted trees. This approach forms the boosting into a numerical optimisation problem. The main idea is to first optimise a loss function, for example the squared errors. The algorithm then uses decision trees as the weak learners. The weak learners are added one by one, not changing the existing trees to minimise the loss function (Brownlee, 2021). A flexible implementation of Gradient boosting is the Extreme Gradient Boosting (XGBoost). The algorithm is highly effective and has got increasingly popular among experts.

A master thesis conducted by Nielsen, a student at NTNU in 2016, states that the reasoning for XGBoosts high performance is due to its adaptability. Even though it is highly flexible, it has the advantage of adjusting its flexibility for various sections of explanatory variables, proving that XGBoost considers the bias-variance trade-off. Hence, when the relationship between the variables is *simple*, XGBoost fits *simple* representations, while when handling more complex interactions, it fits more intricate functions (Nielsen, 2016).

XGBoost follows the principle for gradient boosting but adds multiple parameters. Both methods include the shrinkage parameter η, number of iterations M and a depth parameter. The shrinkage parameter affects the variance as it decides the influence of each decision tree. Setting the shrinkage parameter too high will increase the variance because data during the early iterations will highly influence the model (Chen & Guestrin, 2016). The number of iterations represents the number of decision trees to be produced, while the depth parameter decides the size of the decision trees. A large dept parameter will compute a more complex model and be more prone to overfitting (Chen & Guestrin, 2016; Nielsen, 2016).

The most critical parameters XGBoost includes are the *lambda* parameter, the *gamma* parameter and column subsampling. *Lambda* is a regularisation term on weights. Increasing this parameter means penalising the complexity of the trees to a greater extent. Furthermore, *gamma* specifies the minimum loss reduction that is required to make a further splitting of a leaf in the regression tree. In other words, if the loss reduction (squared error) is below this value, the model will stop splitting the leaf, meaning that the leaf becomes the terminal node of the tree (James et al. 2013). The purpose of gamma is to avoid overcomplicating the model and thereby combating overfitting. The column subsampling fraction allows the model to randomly select predictors for every iteration (Chen & Guestrin, 2016; Nielsen, 2016). An algorithm that uses column subsampling simplifies the implementation, especially when dealing with a large data set. XGBoost has another feature of being sparsity aware, meaning that the method can look past missing values, increasing the simplicity even further (Chen & Guestrin, 2016).

5. Data

The data in this thesis is a combination of data supplied by Hydro and from the Federal Reserve Economic Data (FRED).

5.1 Variable selection

As a proxy for global macroeconomic conditions, we choose to include variables from the US, China, Germany and Australia in our data set. Hydro has gathered data concerning the aluminium prices, both spot- and three-month futures contracts, as they have access to LME's historical prices. Additionally, the company has provided us with data on other variables that may affect the spot price, namely the German year-ahead power price, the index for coal contracts in Asia (NEWC Index) and the Shanghai Containerized Freight Index (SCFI Index). As we want to include more explanatory variables, we collect macroeconomic variables from FRED, namely industrial production in the US, the US dollar (USD)/Australian dollar (AUD) exchange rate, and the US interest rate Market Yield 3M[9].

Spot price

Earlier studies of other commodities presented in the literature review section indicate that spot and futures prices often follow each other. It is not unlikely that this is true for aluminium, suggesting that the spot price should be added to the data set.

Shanghai Containerized Freight Index

As the mining of bauxite, the extracting from bauxite to alumina and the refining of the alumina to aluminium take place in different parts of the world, it is logical to expect that transportation cost can be a determinant of the aluminium price.

[9] Industrial production, USD-AUD exchange rate and interest rate Market Yield 3M can be accessed from: https://fred.stlouisfed.org/

The Shanghai Containerized Freight Index includes a variety of routes from Shanghai to the rest of the world. The index is the most widely used for sea freight rates for imports from China worldwide (DSV Global Transport and Logistics, 2021). Hence, the index can be a good indicator of the global freight market.

German year-ahead power price

Commonwealth Scientific and Industrial Research Organisation calculates that the energy used to make the aluminium is 211 GJ per tonne, compared to 22.7 GJ per tonne for steel (Brooks, 2012). In other words, aluminium production is highly energy-intensive, and consequently, it is intuitively to investigate whether power prices can explain the variation in the aluminium spot price. Hence, we include Germany year ahead power prices in our data set as it is commonly used as a proxy for European power prices.

NEWC Index

Over 80% of aluminium production in China uses coal-fired power (Wood Mackenzie, 2021), implying that the coal price might be necessary when determining China's aluminium price. As the London and Chinese metal markets are somewhat connected, the Chinese coal price may impact the LME price. The NEWC Index is used as a benchmark for seaborne thermal coal prices in the Asia-Pacific region (globalCOAL, 2021), and hence can be a suitable indicator of the Chinese coal price.

In addition, coal is the most comprehensive primary energy source in Australia. The fact that Australia is the sixth biggest aluminium-producing country globally (U.S. Geological Survey, 2021) supports our argument of including The NEWC Index in our data set.

Industrial production and Market yield 3M

Inspired by Labys (1999) findings presented in the literature review section, industrial production is also part of our variables as it is a very informative indicator of development in industrial activity (ECB Economic Bulletin, 2016). In addition, we add an interest rate to our data set, which Labys (1999) also examine in his paper.

AUD/USD Exchange rate/ Commodity currency

Motivated by what Chen et al. (2010) and Gargano and Timmerman (2014) discover, we choose to include the commodity currency Australian dollar/USD exchange rate.

5.2 Data set up

The frequency of the raw data provided by Hydro is daily prices, while the US macroeconomic variables from FRED are monthly. An overview of the different time series and the pertaining periods and frequency are displayed in table 5.1.

Table 5.1*:* Overview of the time series with the time period and its original frequency

Time series/variables	**Time period**	**Frequency**
Aluminium 3-month futures price	1980.01.02 – 2021.09.30	Daily
Aluminium spot price	1957.02.01– 2021.09.30	Daily
Shanghai Containerized Freight Index	2011.09.01 – 2021.09.30	Daily
German year-ahead power price	2008.10.01 – 2021.09.30	Daily
NEWC Index	2009.06.10 – 2021.09.30	Daily
Industrial production	1919.01.01 – 2021.09.01	Monthly
AUD/USD exchange rate	1971.01.01 – 2021.09.01	Monthly
Market Yield 3M	1981.09.01 – 2021.09.01	Monthly

Although it is possible to conduct forecasts with different frequencies[10], we chose a simpler path in this thesis, only using a monthly frequency. We convert the daily data into monthly data by calculating the average daily prices for each month. In order to conduct forecasts including the time series mentioned above, the time series must contain the same period. Therefore, we only include data from September 2011 to September 2021 unless stated otherwise.

[10] An example of a forecast method which takes different frequencies into account is the mixed data sampling (MIDAS) model developed by Ghysels et al. (2007).

5.3 Initial analysis

When performing various forecasting methods, it is essential to understand the characteristics of the data set. By conducting initial analysis, we will obtain valuable insights into our data.

5.3.1 Dependent variable

The dependent variable of this paper is the aluminium 3-month futures contract price. A visualisation of the average monthly aluminium futures price from January 2011 to September 2021 is illustrated in figure 5.1. As we see from the summary statistics in table 5.2, the monthly futures price ranges from a minimum value of 1481.95 US Dollars (USD) to a maximum of 2898 USD. In addition, we see that the mean and median are not far from each other, which may indicate that the data is skewed.

Table 5.2: Summary statistics of the 3-month futures contract price

Min	Max	Median	Mean	1st Qu.	3rd Qu.
1481.95	2898.00	1901.22	1958.97	1779.87	2093.68

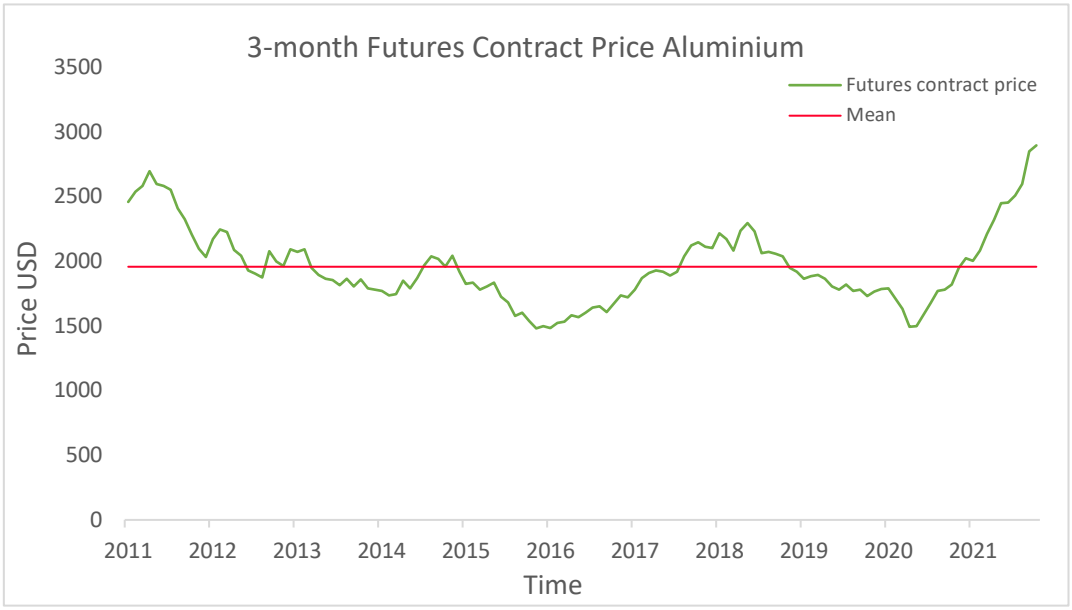


Figure 5.1: Plot of the 3-month futures contract price and its mean

As illustrated in figure 5.1, the futures price of aluminium has a rapid decrease from January 2020 till May 2020, followed by a continuous increase without any significant downwards adjustments towards September 2021. Two substantial events during this period are the outbreak of COVID-19 and a change in Chinese politics. As COVID-19 spreads globally at the start of 2020, many nations go into lockdown. This results in a dramatic reduction in aluminium manufacture, a disruption to supply chains and a significant reduction in demand. Not surprisingly, a decrease in demand results in a decrease in price, as demonstrated in figure 5.1. In August 2021, the LME hits the decades highest price of aluminium. As nations start opening again, the demand increases significantly as the world economy restarts. Additionally, the Chinese cut back on aluminium production due to climate goals in 2021 (Sanderson, 2021). With China being the main supplier of aluminium globally, the cutback leads to shortages putting increased pressure on the price.

Seasonality

Before conducting tests and forecasting methods, it is important to check if the dependent variable is strongly affected by seasonality. If this is true, it is necessary to seasonal adjust the data (Hyndman & Athanasopoulos, 2021). The seasonal plot of the aluminium 3-month futures contract price is displayed in figure 5.2. The data is plotted against the respective seasons, months in our case. It also allows us to spot if the pattern changes for a specific year. As one can see, there are no apparent patterns for each month. Hence, we do not take seasonality into account further in this analysis.

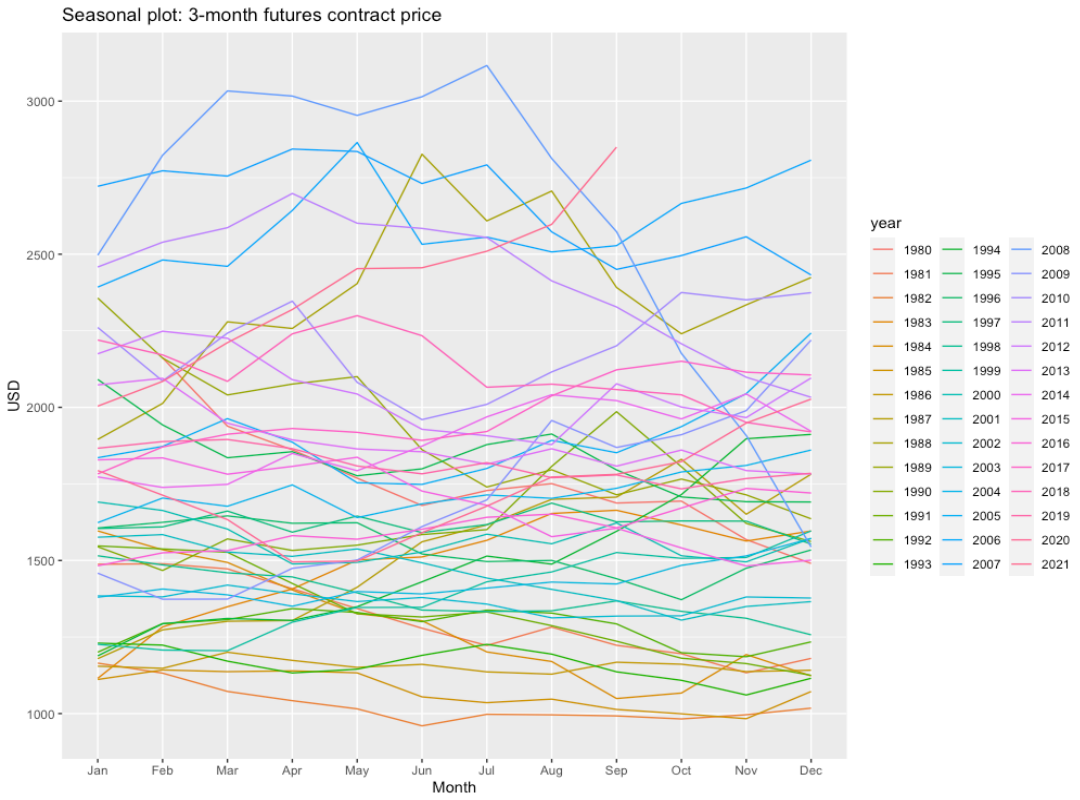


Figure 5.2*:* Seasonal plot of the 3-month futures contract price

5.3.2 Winsorizing

Throughout history, major events have affected macroeconomic factors, affecting both the demand and supply of aluminium. In the period we are examining, the most notable events are the outbreak of the coronavirus in 2020 and the cut back on aluminium production in China in 2021 (Sanderson, 2021). Therefore, the explanatory variables contain some extreme values. Such events will surely have impacts on a long-term horizon. Then again, we are forecasting a short-term horizon, implying that the outliers will not significantly affect the results. However, we cannot rule out this possibility and decide to winsorize the explanatory variables.

Winsorizing involves minimising the influence that the outliers may have on the results. We set an upper and lower border to the observations' 95% and 5% quantile, respectively. The observations that lie outside these borders will be set to the value of the closest border.

5.4 Stationarity and Unit Root Tests

To determine which models are applicable for our time series, we need to identify the stationarity through unit root testing. This is important as we intend to conduct either a VAR model or a VECM. As mentioned, the VAR model requires stationary data, while the VECM is necessary to use when the data are non-stationary.

According to Kočenda & Černý (2015), financial time series are often non-stationary due to economic growth. When the unit root is the reason for the changes in the statistical properties over time, such as the variance, the covariance or the mean, the times series is non-stationary.

5.4.1 ADF and KPSS test

A commonly used unit root test is the augmented Dickey-Fuller (ADF) test developed by Dickey and Fuller (1981). One can perform the ADF test including 1) an intercept, 2) both an intercept and a trend or 3) no deterministic terms. The null hypothesis states that the time series is non-stationary.

According to Wang and Tomek (2007) a constant is usually included in the equation when testing economic time series. Therefore, we only conduct tests where a constant is present. Additionally, we perform the test including a trend. The equation of the ADF test is presented below:

$$\Delta y_t = \alpha + \beta t + \delta y_{t-1} + \sum_{i=1}^{p} \theta_i \Delta y_{t-1} + \varepsilon_t \quad (5.1)$$

α is the intercept, β is the trend component and the δ refers to whether a unit root is present. If δ = 0, the null hypothesis is true (a unit root is present). The θ_i represents a matrix

including the lagged values of Δy[11]. Finally, the ε_t refers to the residuals of the model, which is the error term. If the test statistics (t-value) from the ADF test is smaller than the critical values taken from Dickey and Fuller (1981), the null hypothesis is rejected.

The results from the ADF test[12] reveal that the spot price and futures price are stationary in levels when including both a constant and a trend. However, when we include a constant but exclude the trend, the futures time series is the only one that appears to be stationary. The remaining time series are non-stationary in levels regardless of whether a trend is present or not. In the first order, the results indicate that all the time series are stationary except the Shanghai Containerized Freight Index (SHXFI) and the market yield 3M. However, we obtain stationary time series for these variables by taking the second difference.

The results from the ADF tests for the aluminium spot price and futures price are unclear as we only observe stationarity in some of the scenarios. A study by Nelson and Plosser (1982) presents that the ADF's performance is poor for near-unit root time series. Therefore, we want to investigate the stationarity for the two series with another unit root test, namely the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

In contrast to the ADF test, the null hypothesis of the KPSS is that there is no unit root, and in other words, the time series is stationary. Hence, testing the null hypothesis in opposite directions will increase the possibility of getting correct results (Nelson & Plosser, 1982). A more detailed explanation of the KPSS test is found in Appendix A7.

The results from the KPSS test[13] report that the futures price and spot price are non-stationary at a 1% level when including an intercept as well as when including both an intercept and a trend.

[11] The selection of lags for the ADF-test is presented in Appendix A5

[12] The ADF test results are reported in Appendix A6

[13] In Appendix A8 the KPSS test results are reported

Conclusively, the Shanghai Containerized Freight Index (SHXFI) and the market yield 3M are stationary in second-order I(2). The remaining time series are stationary in first order I(1).

5.4.2 Handling stationarity in different orders

While the VECM model transforms the dependent variable to stationary by taking the first difference, the dependent variable in XGBoost remains in levels. Consequently, one must handle the explanatory variables differently in the two methods.

The exogenous variables must be stationary in levels or be stationary in the same order as the dependent variable. Therefore, in the VECM, when the dependent variable is represented in its stationary form, the explanatory variables must also be transformed into stationary time series. In XGBoost, the explanatory variables may be included in levels if they are stationary. However, they must be transformed to stationary in the same order as the dependent variable if they are non-stationary.

5.5 Handling empty observations

The original data set spans from September 2011 to September 2021. However, differencing and lagging the variables leads to empty observations. As the VECM cannot forecast on empty observations, we omit the months without observations. In practice, this means that we do not train the forecasting models on the earliest observations. Hence, the data set we will apply in the forecast models will span from February 2012 to September 2021.

6. Model specifications

6.1 VECM

6.1.1 Deterministic terms

In the VECM, there is an option to include a constant or a constant and a trend. We decide this by exploring when the accuracy is the highest. As a result, we find that one should exclude both deterministic terms.

6.1.2 Lag selection of endogenous variables

An important aspect when fitting a VECM is to choose the optimal lag length of the endogenous variables. This can be challenging because of the trade-off between the model's goodness of fit and complexity. Suppose we do not include a sufficient number of lags. In that case, we might obtain a biased model due to autocorrelation[14] in the residuals, whiles when including too many lags, the test's predictive power decreases. We use information criteria to find the lag length of the endogenous variables that minimises the estimated information loss (Akaike, 1974). Our thesis will be using Akaike's Information Criterion (AIC) to determine the appropriate lag length, see Appendix A9 for further detail. The optimal lag for our data is 5.

6.1.3 Lag selection of exogenous variables

In order to select the proper lags for the exogenous variables, one should ideally look at the autocorrelation of the VECM's residuals. Such an approach is computationally expensive as it demands evaluating the autocorrelation for all possible combinations of lag lengths for the variables. Therefore, we follow a more straightforward approach where we specify a fixed lag length for all the exogenous variables and select the lags that are significant in the

[14] Autocorrelation in the residuals measures the correlation between a current residual value and its past values.

VECM. The fixed length is set by looking at the autocorrelation in the so-called autocorrelation function plot (ACF-plot). If the lags are within the significance lines, there is no autocorrelation.

In our case, the fixed lag length is 5 (see Appendix A10). The significant lagged values in the VECM are lag one, two and three of the SCFI Index (see Appendix A11).

6.2 XGBoost

6.2.1 Parameter tuning

As XGBoost has a long list of parameters, cross-validation can take a long time. Therefore, we follow a simplified and commonly used approach called the holdout strategy (Schneider, 1997). This strategy divides the data into a train- and test set and then tune the parameters that influence the model most.

The shrinkage parameter η and the number of iterations M should be decided simultaneously as they affect each other. A cost of decreasing the shrinkage parameter η is an increase in computational demand (M), again affecting the complexity decided through the depth parameter. Therefore, we start by tuning these main variables before we test the column-wise parameter and the lambda and gamma regularisation parameters.

The optimal hyperparameters are reported in table 6.1. The ideal tree depth is relatively low, making our model more shallow and less likely to overfit. Neither lambda nor gamma is present, indicating that the model does not need the regularisation terms.

Table 6.1: Hyperparameters of XGboost

Hyperparameter	**Value**
Boosting iterations	200
Learning Rate	0.19
Tree Depth	2
Column Sampling	0.8
Lambda	0
Gamma	0

6.2.2 Selection of explanatory variables and their lagged values

To compare the models, they must take the same variables and the corresponding significant lagged values into account. Therefore, we include the same variables and their significant lagged values selected from the VECM in the XGBoost approach.

7. Results

It appears that the models we use to forecast the aluminium 3-month futures contract price have a high degree of accuracy. This implies that the variables included in the data set are applicable.

In table 7.1, the accuracy measures of RMSE and MAE are reported for each of the methods conducted in this thesis. We see that the VECM outperforms the random walk model, which is in line with the conclusions drawn by Coppola (2008). However, the XGBoost is the superior model with the lowest RMSE and MAE. This conclusion underpins the findings from both Kaggle competitions and Jabeur et al. (2021), stating that XGBoost is a comprehensive model.

The following subsections will present each model's performance and attempt to explain why they behave the way they do.

Table 7.1: The results of all models, with ROCV starting at t = 90

		Test error	
	Method	RMSE	MAE
Benchmark	Random Walk	90.66	71.63
Advanced statistical	VECM	72.15	54.11
Machine learning	XGBoost	63.57	34.55

7.1 Random walk results

As reported from table 7.1, the random walk model performs poorly with an RMSE of 89.242 compared to the two challenging models. A visualisation of the forecast against the actual value is displayed in figure 7.1. We observe that the model predicts the precise value of the futures price with a one-day lag, which verifies equation 4.3.

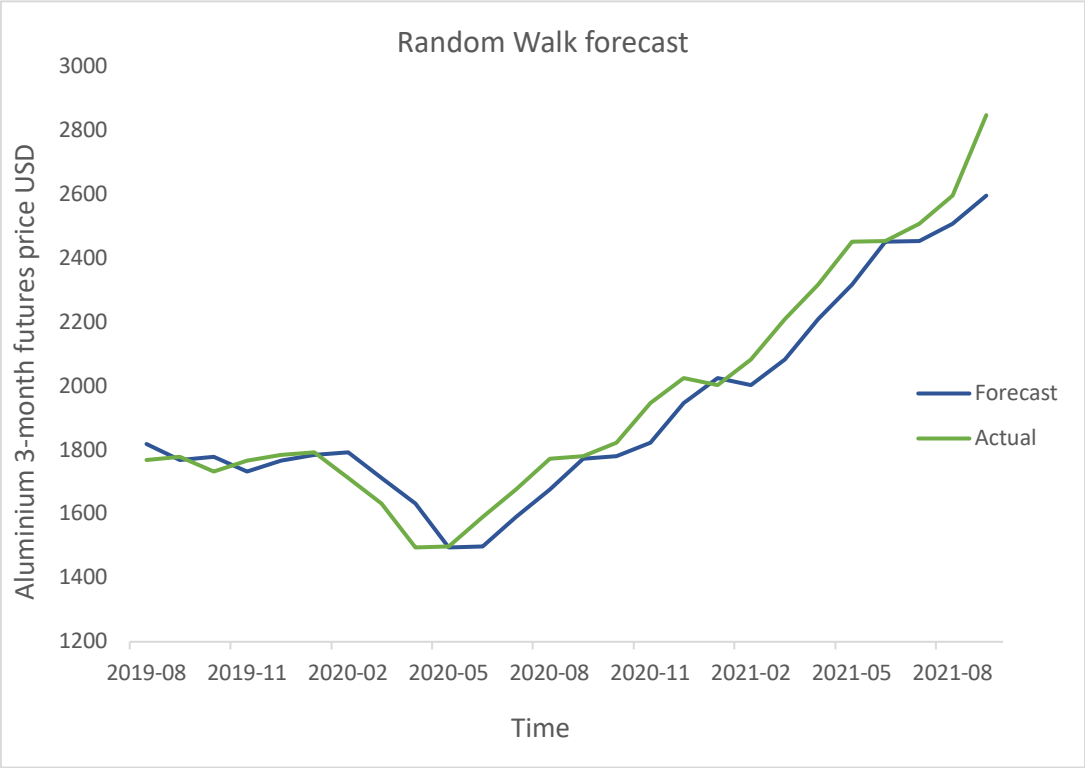


Figure 7.1: Visualisation of the random walk forecasts plotted against actual values

7.2 VECM

7.2.1 Cointegration and causality

Johansen cointegration test results

A prerequisite of applying the VECM is that it exists a cointegrated relationship. The result from the Johansen cointegration test is displayed in table 7.2. In this thesis, we base our analysis on a 5 % significance level, which is a commonly selected base value. We can reject the null hypothesis stating no cointegrated relationship in stage one. However, in stage two, we cannot reject the null hypothesis as $9.05 < 9.24$. Conclusively, we must keep the null hypothesis, stating that there exists one cointegrated relationship between the time series.

Table 7.2: Cointegration test results with spot and 3-month futures

	$\lambda trace$	10%	5%	1%
r = 0	41.83***	17.85	19.96	24.60
r <= 1	9.05*	7.52	9.24	12.97
	λmax			
r = 0	32.77***	13.75	15.67	20.20
r <= 1	9.05*	7.52	9.24	12.97

*Note: λtrace and λmax statistics marked with *, ** or *** are significant at a 10 %, 5 % or 1 % level, respectively.*

Granger causality test results

One can tell from the F-test in table 7.3 that both spot Granger cause the 3-month futures contract price and vice versa. Due to these results, one can conclude that a bidirectional Granger causality relationship exists, and Granger's requirement of at least one-directional relationship is fulfilled. This also supports the cointegration relationship reported in table 7.2. Hence, we can with confidence include spot and futures as endogenous variables in the VECM model.

Table 7.3: Granger causality test results

Dependent variable	**Independent variable**	**F-test**
3-month futures	Spot	7.55 ***
Spot	3-month futures	52.20***

*Note: F-test statistics marked with *, ** or *** are significant at a 10 %, 5 % or 1 % level, respectively.*

7.2.2 VECM forecast

As mentioned, the VECM outperforms the random walk model, with both a lower MAE and RMSE. From figure 7.2, we observe that the method delivers relatively accurate predictions with some deviations periodically. From early 2021 until today, the model underestimates

most of the time compared to the preceding years. It is considered that the underestimations throughout 2021 are partly due to the cutback in aluminium production in China as well as the global COVID-19 pandemic.

Although the VECM generally performs better than the random walk model, the deviations in the VECM are sometimes bigger than in the benchmark model. This may be due to unforeseen events not captured in the training data. As we know, the VECM includes macroeconomic variables to forecast, and it is not unlikely that unforeseen happenings will affect these variables. Therefore, it is not surprising that we observe that the VECM has bigger deviations than the random walk model, as the random walk model only considers the previous observation y_t when forecasting y_{t+1}.

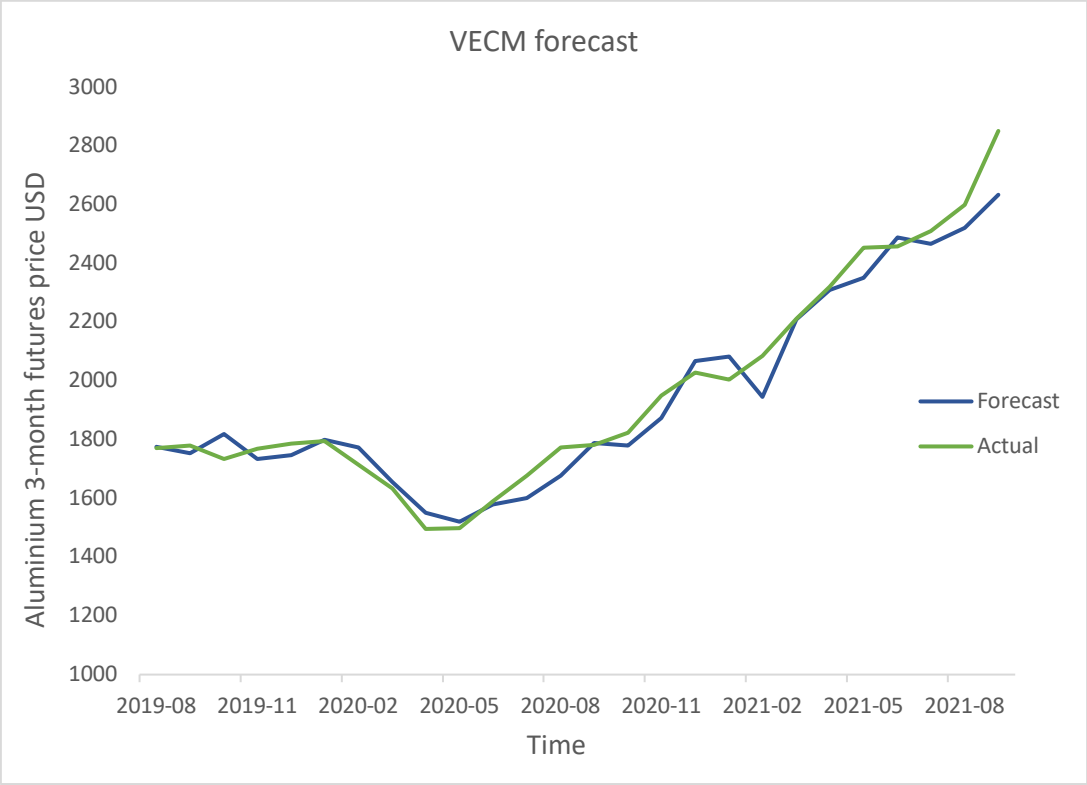


Figure 7.2: Visualisation of the VECM forecasts plotted against actual values

7.3 XGBoost forecast

Comparing XGBoost with the statistical methods, we find that XGBoost is the superior model measured with RMSE and MAE, presented in table 7.1.

Figure 7.3 implies very satisfactory accuracy in XGBoost compared to the VECM. In similarity to the VECM, XGBoost underestimates the futures price from 2021. However, comparing the plots for each model, XGBoost underestimates the prices to a larger extent.

Conversely, we observe XGBoost being the superior model before 2021, while for the following months, the VECM performs better. This indicates that the VECM produces better forecasts during the period with the macroeconomic shock of Chinese policy changes. However, one cannot conclude that the VECM is the preferred method to use for other unforeseen shocks on the economy, as XGBoost delivers a higher accuracy during for example the outbreak of COVID-19 in 2020. Furthermore, as we base our conclusions on the overall best accuracy for July 2019 to September 2021, we conclude that the XGBoost is the best performing model in forecasting the 3-month futures price of aluminium.

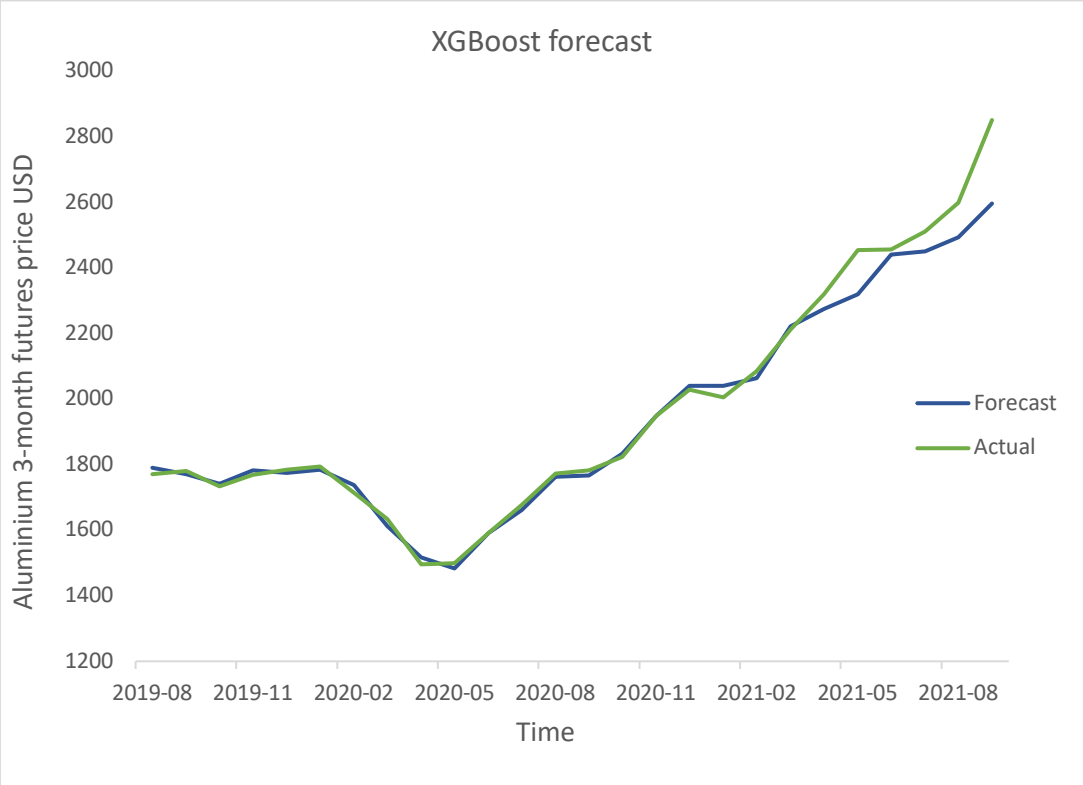


Figure 7.3: Visualisation of XGboost forecasts plotted against actual values

7.4 Robustness test of our findings

Our conclusion from the results is that XGBoost is the most accurate forecasting method to use when forecasting aluminium's 3-month futures contract price, looking at a one-month horizon. A robustness test of our results serves the goal of checking how the conclusion changes when the assumptions change. When conducting the analysis, one should incorporate the most important uncertainty into the model. Then again, model specifications will always be a difficult task, and through a robustness check, one can demonstrate that the main analysis is correct.

7.4.1 Including Exchange Rate as endogen

According to Chen et al. (2010), the Australian dollar is considered a commodity currency. As we state in the literature review, Chen et. al. conclude that commodity currencies do have predictive power over commodity prices, like aluminium prices. However, the article does not conclude with the commodity price having strong predictive power over the exchange rate. This contradicts the definition of commodity currencies, stating that they are affected by the commodity product. To explore this relationship, we conduct a robustness test of our findings where the exchange rate of USD-AUD is considered as an endogenous variable.

When including another endogenous variable, we perform the Johansen cointegration test again to detect the number of cointegrated relationships. Similarly to the cointegration test between the futures and the spot price in subsection 7.2.1, we consider a 5 % significance level. An output of the test is reported in Appendix A12.

The conclusion from the Johansen test is ambiguous. The trace test concludes that three cointegrated relations exist between the three time series. In other words, the time series are stationary. However, the maximum eigenvalue test states that there is one cointegrated relationship. As the main goal of our thesis is to attain the highest accuracy, we conduct forecasts for all possibilities. In other words, we perform the VECM with one cointegrated relationship and a vector autoregressive method (VAR) to observe when we achieve the best accuracy. Additionally, we perform XGBoost considering the exchange rate as endogenous. Random walk model is a univariate method, meaning that it only takes the 3-month futures

contract into account when forecasting. Therefore, we exclude random walk model as it is not expedient for this robustness test.

Table 7.4: Comparison of accuracy measures for the VECM (r = 1), VAR and XGBoost.

		Test error	
	Method	RMSE	MAE
Advanced statistical	VECM (r =1)	84.82	66.41
	VAR (r = 3)	86.98	67.54
Machine learning	XGBoost	65.86	36.42

As this thesis aims to find the highest accuracy, we compare the XGBoost to the statistical method that performs the best, namely the VECM, including one cointegrated relationship. XGBoost achieves an RMSE of 65.86 while the VECM, including one cointegrated relationship, obtains an RMSE of 84.82 (see table 7.4). Compared to the previous scenario where the exchange rate was considered as exogenous, the VECM increases its RMSE by approximately 17.5 %. The machine learning technique only increases by 3.6 %, implying that the VECM is more volatile when changing the assumption of endogenous variables. Again, the machine learning technique XGBoost is the superior model, indicating that our initial conclusion is robust.

Comparing the results of this robustness test to our main analysis, both the VECM and XGBoost perform worse than when we consider the exchange rate as exogenous. This contradicts the definition of commodity currencies following the underlying commodity. However, it underpins Chen et al.'s findings of exchange rates having predictive power over the commodity price but not vice versa. For achieving the highest accuracy, the conclusion from the robustness test is that our initial assumption is correct.

7.4.2 Excluding exogenous variables

Not surprisingly, the exogenous variables we include in the forecasting methods will affect how the models perform. To conclude that XGBoost outperforms the VECM due to the model itself and not because of the variables we include, we perform a test where we exclude the exogenous variables. We use the same period and number of lags for the endogen variables, namely the aluminium spot price and futures price.

Table 7.5: Accuracy measures of the VECM and XGBoost including futures and spot

		Test error	
	Method	RMSE	MAE
Advanced statistical	VECM	82.29	67.75
Machine learning	XGBoost	64.62	35.23

We see in table 7.5 that XGBoost and the VECM performs poorly compared to the accuracy achieved when including the exogenous variables. In the same manner as the previous test (including exchange rate as endogen), the increase in RMSE is significantly higher for the VECM than XGBoost, 14 % and 1.6 %, respectively. This indicates again that the VECM is more volatile to changes in the assumptions, while XGBoost is relatively robust.

Conclusively, XGBoost still outperforms the VECM, further substantiating our main conclusion.

7.4.3 Starting the ROCV in July 2012 (t=90)

According to Warner (1998), most time-series experts suggest that time-series analysis should base their model on at least 50 observations. The choice of starting the ROCV at time 90 was due to the Hyndman and Athanasopoulos (2021) suggestion that the train set should contain approximately 80 % of the test set. Additionally, it was a practical decision to minimise the computational time. With Warner's minimum requirement in mind, we want to test the robustness of our findings with an increased test set, starting at t = 50.

In contrast to the previous robustness test, the results from table 7.6 report an increase in accuracy for both models. For the VECM, this only comes to light when comparing the MAE as the RMSE is unchanged. However, the model performs just slightly better. XGBoost, on the other hand, has decreased its RMSE by 30 %, resulting in a very satisfactory accuracy.

XGBoost reaches significantly more accurate forecasts than the VECM, measured in RMSE and MAE. The robustness test again underpins our initial conclusion.

Table 7.6: The results of all models, with ROCV starting at t = 50

		Test error	
	Method	RMSE	MAE
Advanced statistical	VECM	72.15	53.24
Machine learning	XGBoost	44.41	24.74

7.4.4 Summary of the robustness tests

When changing the assumptions regarding endogenous variables, exogenous variables, and changes in the starting point of the cross-validation, we conclude in the same way as our main analysis: the machine learning technique XGBoost will in all scenarios outperform the statistical method VECM when forecasting the 3-month futures contract price of aluminium at one month horizon.

7.5 Critique

Based on the results obtained from the experiments, we manage to find a model that forecasts the aluminium 3-month futures contract price with a high degree of accuracy. Although ROCV helps detect overfitting, it does not eliminate the chance of getting an overfitted model. One reason for this is that all the training sets share the same first

observations. Hence, despite this thesis obtaining highly accurate predictions, we cannot rule out the possibility that the model is overfitted.

In general, a challenge of real-life forecasting is that it is not possible to predict every event that may occur. The corona pandemic is an excellent example of this. Although the rest of the world knew how the virus was evolving in China, no one predicted the precise outcome. As the macroeconomy is highly affected by politics, which is dependent on human interactions, forecasting variables like the aluminium price is challenging. According to Gilb's law of unreliability, "…any system which depends on human reliability is unreliable". Although this is a stringent law, one can argue that there will never be a data set that captures all information required to predict an outcome affected by human interactions. With that being said, we can still achieve satisfying predictions reducing the risk that will always be present when participating in a market.

7.5.1 Limited variable selection

With inputs from Hydro, we select variables we see as intuitive determinants of the futures price. Therefore, the variables are not a supplementary list, and we may have missed some essential variables. As reported, the XGBoost outperforms the VECM in all scenarios. We obtain the best accuracy when including all variables, suggesting that the variables do have predictive power. However, even though we obtain satisfying results with the selected variables, we cannot rule out the possibility that they could improve by including additional control variables.

7.5.2 Limitations of parameter tuning

The best way to prevent overfitting is to conduct a high degree of parameter tweaking. As mentioned, we tune the parameters in XGBoost. Because of limited capacity on our desktop computers, we do not check for all the combinations of the parameters. Hence, we will most likely not find the optimised parameters accurately, and they should be considered estimates. Therefore, it can be argued that a more comprehensive parameter optimisation could result in a lower probability of overfitting and an even higher accuracy.

7.5.3 Same lag length of endogen variables

In the conductance of the VECM, we select the lags of the exogenous variables by simply looking at which are significant, see Appendix A7. The general formulation of the VECM (equation 4.5) uses the same lag length for the endogenous variables. Furthermore, there is no option to include different lag lengths for the endogenous variables in the built-in function of the VECM in R. Therefore, enabling us to apply this built-in function, we do not optimise the lag length of the endogenous variables. It can be argued that not separating the lag lengths is reprehensible, as we may end up with more satisfying results by doing so.

7.6 Further research

Forecasting the aluminium 3-month futures contract price through this thesis has shown to be a complex task. Therefore, further research is recommended, and the most relevant is to do more research into the determinants of the aluminium price.

In addition, it would be expedient to assess the performance of other forecasting methods, such as the popular neural network approach. In addition to XGBoost, neural network forecasting methods have performed well when forecasting time series. As a matter-of-fact neural network-based methods were among the best entries in the well-known M4 forecasting competition (Benidis et al.,2020; Makridakis et al., 2018).

8. Conclusion

The goal of this thesis is to examine whether machine learning can outperform statistical methods when forecasting the aluminium's 3-month futures contract price. We have assessed three forecasting methods to answer this, more specifically the simple statistical method random walk, the more advanced statistical method VECM and the machine learning technique XGBoost. To verify our findings, we perform robustness tests.

This thesis manages to produce highly accurate forecasts, where XGBoost outperforms the statistical methods. As the demand of aluminium increases at a high pace, Hydro, among others, will benefit from applying this model as it enhances decisions regarding trading strategies that may provide higher returns.

The robustness tests indicate that our initial model, including all the exogenous variables and treating the AUD/USD exchange rate as exogenous, delivers the best accuracy. When changing the assumptions, XGBoost exclusively delivers the best accuracy and is the most stable model. These findings imply that XGBoost is the superior model. The VECM, on the other hand, has a more volatile accuracy. When including the AUD/USD exchange rate as endogenous and excluding the exogenous variables, the VECMs accuracy decreases significantly more than XGBoost. Additionally, when we start the ROCV at t = 50, XGBoost improves its RMSE, while the RMSE of the VECM remains unchanged, underpinning the conclusion that XGBoost is the preferred model.

The main goal of this thesis is to find the most applicable method with the highest accuracy. We propose that XGBoost, including exogenous variables, is the most suited model. In addition, the XGBoost delivers the most robust performance when changing the assumptions, suggesting it to be the most reliable model to apply.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control, 19*(6), 716-723.

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, B., Maddix, D., Turkmen, C., . . . Januschowski, T. (2020). Neural forecasting: Introduction and literature overview.

Bontempi, G., Taieb, S. B., & Borgne, Y. L. (2013). *Machine Learning Strategies for Time Series Forecasting* (Vol. 138). Berlin: Springer.

Brooks, G. (2012). *The trouble with aluminium*. Retrieved November 2021, from The Conversation: https://theconversation.com/the-trouble-with-aluminium-7245

Brownlee, J. (2021). Gradient Boosted Trees with XGBoost and scikit-learn. In *XGBoost in Python* (Vol. 1).

Bryhn, R., & Gram, T. (2021). *Norsk Hydro*. Retrieved September 2021, from Store norske leksikon: https://snl.no/Norsk_Hydro

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. 22). San Francisco, California: Association for Computing Machinery.

Chen, Y.-C., Rogoff, K. S., & Rossi, B. (2010). Can Exchange Rates Forecast Commodity Prices. *The Quarterly Journal of Economics, 125*(3), 1145-1194.

Chow, Y.-F., McAleer, M., & Sequeira, J. (2002). Pricing of Forward and Futures Contracts. *Economic Surveys, 14*(2), 215-253.

Coppola, A. (2007). Forecasting oil price movements: Exploiting the information in the futures market. *The Journal of Futures Markets, 28*(1), 34-56.

De Prado, M. L. (2018). *Advances in financial machine learning.* New Jersey: John Wiley & Sons.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica, 49*(1), 1057-1072.

Dooley, G., & Lenihan, H. (2005). An assessment of time series methods in metal price forecasting. *Resources Policy, 30*(3), pp. 208-217.

Doshi, Y., & Prasad, E. (2019). *Aluminium Market Report.* Allied Market Research.

DSV Global Transport and Logistics. (2021). *Shanghai Containerized Freight Index*. Retrieved November 2021, from DSV: https://www.dsv.com/en-nl/our-solutions/modes-of-transport/sea-freight/shanghai-containerized-freight-index

ECB Economic Bulletin. (2016). *Box 2 A closer look at differences between industrial gross value added and industrial production.* Retrieved November 2021, from European Central Bank: https://www.ecb.europa.eu/pub/pdf/other/eb201601_focus02.en.pdf

Engle, R. F., & Granger, C. W. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. pp. 251-276.

Fama, E. F., & French, K. R. (1987). Commodity Futures Prices: Some Evidence on Forecast Power, Premiums, and the Theory of Storage. *The Journal of Business, 60*(1), 55-73.

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

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. *Econometric Reviews, 26*, pp. 55-90.

globalCOAL. (2021). *The NEWC Index*. Retrieved November 2021, from globalCoal: https://www.globalcoal.com/coalprices/newcindex.cfm

Granger, C. W. (1988). Some recent development in a concept of causality. *Journal of econometrics, 39*(1-2), 199-211.

Gumus, M., & Kiran, M. S. (2017). Crude oil price forecasting using XGBoost. *International Conference on Computer Science and Engineering (UBMK).*

Hydro. (2021). *How is aluminium made?* Retrieved November 2021, from Hydro: https://www.hydro.com/en/aluminium/about-aluminium/how-its-made/

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (Vol. 2). otexts.

Jabeur, S. B., Mefteh-Wali, S., & Viviani, J. (2021). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Los Angeles, California, USA: Springer.

Jia, J., & Kang, S. B. (2021). Do the basis and other predictors of futures return also predict spot return with the same signs and magnitudes? Evidence from the LME,. *Journal of Commodity Markets, 21*.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *12*(2-3), pp. 231-254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive model. *Econometrica: Journal of the Econometric Society, 59*(6), 1551-1580.

Kaldor, N. (1939). Welfare Propositions of Economics and Interpersonal Comparisons of Utility. *The Economic Journal, 49*(195), 549-552.

Kočenda, E., & Černý, A. (2015). *Elements of time series econometrics: An applied approach.* Charles University in Prague, Karolinum Press.

Kohzadi , N., Boyd, M. S., Kermanshahi, B., & Kaastra, I. (1996). A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing, 10*(2), pp. 169-181.

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

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1-3), 159-178.

Labys, W. C., & Achouch, A. (1999). Metal prices and the business cycle. *Resources Policy, 25*(4), 229-238.

LME. (2017). *The Asian connection: how do London and Shanghai markets interact?* Retrieved November 2021, from LME: https://www.lme.com/Education/Online-resources/LME-insight/The-Asian-connection

Lütkepohl, H., & Reimers, H. E. (1992). Granger-causality in cointegrated VAR processes The case of the term structure. *Economics Letters, 40*(3), 263-268.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting, 34*(4), 802-808.

Marr, B. (2016). *A Short History of Machine Learning -- Every Manager Should Read*. Retrieved December 2021, from Forbes: https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/?sh=6d0d2f4a15e7

McCarthy, J. (2007). *What is artificial intelligence.* Stanford University, Computer Science Department, Stanford.

Mottaghinejad, S. (2021). *The intuition behind bias and variance*. Retrieved December 2021, from Towards Data Science: https://towardsdatascience.com/bias-and-variance-but-what-are-they-really-ac539817e171

Narin, A. (2019). *Understanding LME timespreads*. Retrieved November 2021, from linkedin: https://www.linkedin.com/pulse/understanding-lme-timespreads-narin-atwal/

Nelson, C. R., & Plosser, C. R. (1982). Trends and random walks in macroeconomic time series: some evidence and implications. *Journal of Monetary Economics, 10*(2), 139-162.

Ng, S., & Perron, P. (1995). Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag. *Journal of the American Statistical Association, 90*(429).

Nielsen, D. (2016). *Tree Boosting With XGBoost: Why Does XGBoost Win "Every" Machine Learning Competition?* Norwegian University of Science and Technology, Department of Mathematical Sciences, Trondheim.

Nissi, J., Småros, J., Ylinen, T., & Ala-Risku, T. (2021). *Measuring Forecast Accuracy: The Complete Guide*. Retrieved November 2021, from Relex: https://www.relexsolutions.com/resources/measuring-forecast-accuracy/

Osterwalder-Lenum, M. (1992). A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics. *Oxford Bulletin of Economics and statistics,, 54*(3), 461-472.

Panella, M., Barcellona, F., & D'Ecclesia, R. L. (2012). Forecasting Energy Commodity Prices Using Neural Networks. *Advances in Decision Sciences*.

Pradhan, R. P., Hall, J. H., & du Toit, E. (2021). The lead–lag relationship between spot and futures prices: Empirical evidence from the Indian commodity market. *Resources Policy, 70*.

Sanderson, H. (2021). *Aluminium prices hit decade high as Beijing warns against speculation*. Retrieved December 2021, from Financial Times: https://www.ft.com/content/42a5e710-3cfc-40e6-90df-f524d0319be3

Schneider, J. (1997). *Cross Validation*. Retrieved November 2021, from Carnegie Mellon University School of Computer Science: https://www.cs.cmu.edu/~schneide/tut5/node42.html

Schwartz, G. (1978). Tests for Unit Roots: A Monte Carlo Investigation. *The Annals of Statistics, 6*(2), 461–464.

Toprak, M. (2020). *Gradient Boosting and Weak Learners*. Retrieved November 2021, from Medium: https://medium.com/@toprak.mhmt/gradient-boosting-and-weak-learners-1f93726b6fbd

U.S. Geological Survey. (2021). *Mineral Commodity Summaries 2021.* Reston.

UC RUSAL. (2015). *How aluminium market works*. Retrieved November 2021, from All about aluminum: https://www.aluminiumleader.com/economics/how_aluminium_market_works/

Wang, D., & Tomek, W. G. (2007). Commodity Prices and Unit Root Tests. *American Journal of Agricultural Economics, 89*(4), 873-889.

Warner, R. M. (1998). *Spectral analysis of time-series data.* Guilford Press.

Watkins, C., & McAleer, M. (2006). Pricing of non-ferrous metals futures on the London Metal Exchange. *Applied Financial Economics, 16*(12), 858–880.

Wisdom IT Services India Pvt. Ltd. (2020). *SHORT, MEDIUM AND LONG-TERM FORECASTING - MARKETING MANAGEMENT*. Retrieved November 2021, from wisdomjobs: https://www.wisdomjobs.com/e-university/marketing-management-tutorial-294/short-medium-and-long-term-forecasting-9586.html

Wood Mackenzie. (2021). *Carbon neutrality goal forces Chinese aluminium smelters away from captive coal power*. Retrieved November 2021, from Wood Mackenzie: https://www.woodmac.com/press-releases/carbon-neutrality-goal-forces-chinese-aluminium-smelters-away-from-captive-coal-power/

Working, H. (1948). Theory of the Inverse Carrying Charge in Futures Markets. *Journal of Farm Economics, 30*(1), 1-28.

Working, H. (1949). The Theory of Price of Storage. *The American Economic Review, 39*(6), 1254-1262.

Appendix

A1 Trace and maximum likelihood functions

$$\lambda_{trace(r_0,n)} = -T \sum_{i=r_0+1}^{m} \ln (1 - \hat{\lambda}_i)$$

$$\lambda_{\max (r_0,r_0+1)} = -Tln(1 - \hat{\lambda}_{r_0+1})$$

T represents the number of observations and m is the number of linear independent relationships. $\hat{\lambda}_i$ denotes the eigenvalues computed from the VECM.

The trace test determines whether the rank of (r) is equal to 0 (r_0). Then it tests the answer against the alternative hypothesis, which is that the rank is bigger than 0 (r_0) and less than the number of endogenous variables (n). The null hypothesis of the maximum eigenvalue test is that the largest eigenvalue (r_0) is zero, while the alternative hypothesis states that it is r_0+1.

A2 The Granger causility test

When decomposing a bivariate VAR model into *k* equations for each of the *k* endogenous variables, one distinguishes between the dependent and independent endogenous variables. The two equations of a bivariate VAR model are presented below:

$$X_t = a_{1,1}x_{t-1} + a_{1,2}x_{t-2} + \cdots + a_{1,i}x_{t-p} + b_{1,1}y_{t-1} + b_{1,2}y_{t-2} + \cdots + b_{1,i}y_{t-p} + \varepsilon_t$$

$$Y_t = a_{2,1}y_{t-1} + a_{2,2}y_{t-2} + \cdots + a_{2,i}y_{t-p} + b_{2,1}x_{t-1} + b_{2,2}x_{t-2} + \cdots + b_{2,i}x_{t-p} + \varepsilon_t$$

The Granger causality is based on the coefficients to the lagged independent variables, denoted as $b_{1,t}$ and $b_{2,t}$ in the equations for t = 1,2…,T, for X and Y respectively. $a_{1,t}$ and $a_{2,t}$ represents the lagged dependent variables for t = 1,2…,T, for X and Y respectively. The lag structure is represented as p.

A Granger causality test that considers a bivariate VAR process may lead to four outcomes:

1. x is Granger causing y
2. y is Granger causing x
3. Both are Granger causing each other, in other words there exists a bidirectional causal relationship
4. None of the variables cause each other

A3 F-test

In order to see if the independent variables add valuable information to a model and test if they are significant, one can perform the F-test, expressed in the equation below:

$$F - test = \frac{(SRR_R - SRR_{UR})/p}{SRR_{UR}/(N - p - 1)}$$

where SRR represents the sum of squared residuals. R and UR denotes the restricted and unrestricted regression models, respectively. In addition, N denotes the number of observations and p is the number of parameters.

A4 Algorithm for boosting trees

Boosting for Regression Trees

1. Set $\hat{f}(x) = 0$ and residuals $r_i = y_i$ for all *i* in the training set.
2. For b = 1,2,…B, repeat:
    a. Fit model $\hat{f}(x)$ with *d* splits (*d* +1 terminal nodes) to the training data (X,*r*).
    b. Update $\hat{f}(x)$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

    c. Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$

3. Output the boosted model

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

A5 Lag selection ADF test

This thesis follows the Ng and Perron (1995) lag selection approach. The first step is to determine the maximum lag length. Schwert's (1989) rule of thumb for determining this parameter is presented in the equation below, where *pmax* represents the maximum number of lags:

$$p_{max} = 12 \left(\frac{T}{100}\right)^{1/4}$$

where T is denoted as number of observations in the time series.

After that, one gradually reduces the maximum number of lags until we have the length that provides us with a critical value of the last lagged differenced coefficient greater than 1.6.

A6 ADF test results

Time series	Levels		First difference		Second difference		Time period
	$\alpha \neq 0, \beta = 0$	$\alpha \neq 0, \beta \neq 0$	$\alpha \neq 0, \beta = 0$	$\alpha \neq 0, \beta \neq 0$	$\alpha \neq 0, \beta = 0$	$\alpha \neq 0, \beta \neq 0$	
Aluminium 3-month futures contract price	-2.93(20)**	-3.70(20)**	-5.57(19)***	-5.12(20)***	-	-	Feb 1980 – Sep 2021
Aluminium spot price	-1.56(20)	-3.90(20)**	-6.63(20)***	-6.62(20)***	-	-	Feb 1957– Sep 2021
Shanghai Containerized Freight Index	-2.31(6)	-2.07(7)	-1.45(12)	-3.90(8)**	-6.55 (10)***	-6.57(11)***	Sep 2011 – Sep 2021
German year ahead power price	-1.05(8)	-1.43(8)	-4.42(2)***	-4.81(2)***	-	-	Oct 2008 – Sep 2021
NEWC Index	-0.89(13)	-0.54(3)	-6.46(1)***	-6.70(1)***	-	-	Jun 2009 – Sep 2021
Industrial production	2.79(13)*	-2.81(13)	-7.48(2)***	-6.56(3)***	-	-	Jan 1919 – Sep 2021
AUD/USD exchange rate	1.77(9)	- 1.40(4)	-7.19(2)***	-5.56(3)***	-	-	Jan 1971 – Sep 2021
Market Yield 3M	-1.82(6)	-2.10(7)	-2.97(4)	-2.48(6)	-7.30(5)***	-5.93(7)***	Sep 1981 – Sep 2021

*Note: *, ** and denote significance on a 10 %, 5 % and a 1 % level, respectively*

A7 KPSS test

The KPSS test uses linear regression and ordinary least square (OLS) when estimating the model. The KPSS test is presented in the following equation:

$$x_t = r_t + \beta_t + \varepsilon_t$$

where β_t represents a deterministic trend, r_t is the random walk and ε_t is denoted as the stationary error. To reject the null hypothesis, the error must be significantly different from zero.

Opposed to the ADF fest, the lag selection in the KPSS test is more straightforward, as it selects lags using the following equation:

$$p = \frac{1}{4}\frac{T}{100}$$

where *T* denotes the time series length (Kwiatkowski et al., 1992).

A8 KPSS test results

	Levels		**Time period**
	μ	τ	
Aluminium 3-month futures contract price	2.72(5)***	0.22(5)***	Feb 1980 – Sep 2021
Aluminium spot price	8.10(6)***	0.35(6)***	Feb 1957– Sep 2021

*Note: *, ** and denote significance on a 10 %, 5 % and a 1 % level, respectively*

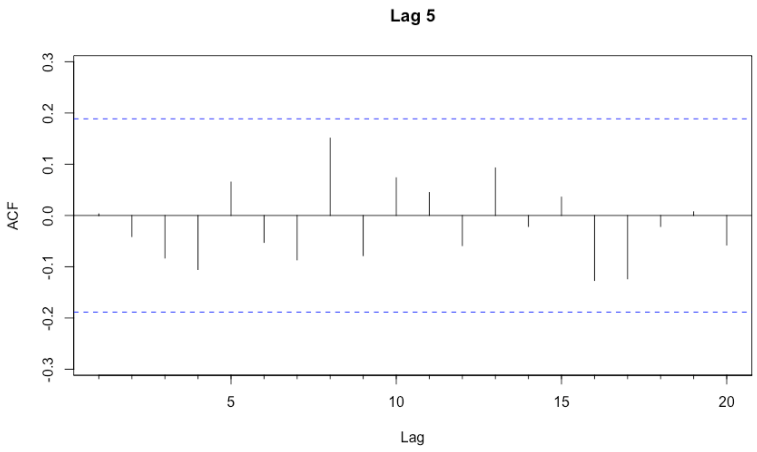
A9 Akaike's Information Criterion (AIC)

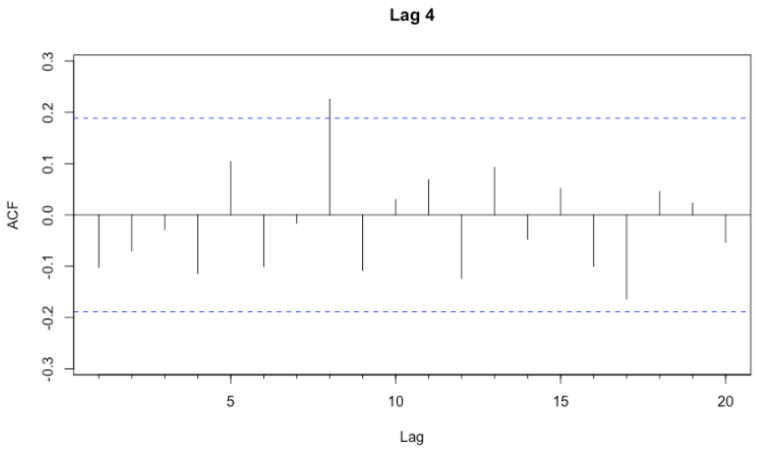
The general formula of AIC is the following:

$$AIC = -2(\log - \text{likelihood}) + 2k,$$

where log-likelihood is a measure of the model fit (the higher the number, the higher the fit) and K is the number parameters in the model.

A10 ACF-plots





A11 Selecting lagged values (output from the VECM)

AUD/USD lag 1	*Industrial production lag 1*	*Market yiel 3M lag 1*	*NEWX index lag 1*	*German power lag 1*	*SCFI index lag 1*
443.54(573.93)	1.81(9.0241)	70.51(126.22)	-0.64(1.43)	3.36(4.76)	**-0.19(0.08)****

AUD/USD lag 2	*Industrial production lag 2*	*Market yiel 3M lag 2*	*NEWX index lag 2*	*German power lag 2*	*SCFI index lag 2*
304.93(573.69)	7.43(8.8554)	83.26(143.65)	1.95(1.41)	1.49(4.79)	**-0.26(0.09)*****

AUD/USD lag 3	*Industrial production lag 3*	*Market yiel 3M lag 1*	*NEWX index lag 3*	*German power lag 3*	*SCFI index lag 3*
739.03(543.47)	-7.01(9.55)	46.93(154.05)	-0.61(1.48)	-2.14(4.82)	**-0.19(0.10)***

AUD/USD lag 4	*Industrial production lag 4*	*Market yiel 3M lag 4*	*NEWX index lag 4*	*German power lag 4*	*SCFI index lag 4*
546.54(535.19)	9.31(8.86)	57.63(148.28)	-0.32(1.63)	-0.76(4.89)	-0.10(0.08)

AUD/USD lag 5	*Industrial production lag 5*	*Market yiel 3M lag 5*	*NEWX index lag 5*	*German power lag 5*	*SCFI index lag 5*
328.17(505.42)	-1.65(7.97)	-53.50(109.30)	1.02(1.50)	-0.28(4.85)	-0.03(0.08)

*Note: The coefficients to the lagged values of the exogenous variables, with their corresponding standard error in ( ). *, ** and *** denote significance on a 10 %, 5 % and a 1 % level, respectively.*

A12 Cointegration test with exhange rate as endogeenous

	λtrace	10%	5%	1%
r = 0	62.32***	32.00	34.91	41.07
r <= 1	23.87**	17.85	19.96	24.6
r <= 2	9.96**	7.52	9.24	12.97
	λmax			
r = 0	38.45***	19.77	22.00	26.81
r <= 1	13.92*	13.75	15.67	20.20
r <= 2	9.96**	7.52	9.24	12.97

*Note: λtrace and λmax statistics marked with *, ** or *** are significant at a 10 %, 5 % or 1 % level, respectively.*