STAINLESS STEEL PRICE FORECASTING

Case Outokumpu Oyj

Lappeenranta–Lahti University of Technology LUT

Master's Programme in Business Analytics, Master's Thesis

2024

Elisa Kallela

Examiners: Post-Doctoral Researcher, Shahid Ahmad Bhat

Associate Professor, Jan Stoklasa

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Business and Management

Business Administration

Elisa Kallela

Stainless steel price forecasting

Master’s thesis

2024

97 pages, 35 figures, 33 tables and 4 appendices

Examiner(s): Post-Doctoral Researcher Shahid Ahmad Bhat and Associate Professor Jan Stoklasa

Keywords: Time series forecasting, Exponential smoothing, Autoregressive integrated moving average, Vector autoregressive, Random forest regression, Stainless steel, Case study

Price forecasting enables businesses to make informed decisions, manage risks, optimise resource allocation, and stay competitive in dynamic markets. This thesis seeks to investigate various quantitative forecasting methods to predict case company’s pricing data. The thesis compares four different forecasting methods: two univariate models and two multivariate models which utilise external data.

The pricing dataset encompasses Net Reference Prices (NRP) and consists of 33 observations from January 2021 to September 2023. Six clusters are chosen from the pricing dataset for this study based on their volumes, grades, and customers, to ensure both the cluster's significance to the company and a diverse range representation of the clusters in the study. The exponential smoothing model, the family of autoregressive integrated moving average (ARIMA), vector autoregressive (VAR) and random forest regression models are tested for each of the six clusters the performance of the models will be assessed by their ability to predict the final three observations of the time series.

Results of the study suggest that for each of the six clusters both exponential smoothing and random forest regression outperformed the benchmark model which consisted of forecasts made by experts in the company. Based on the mean absolute percentage errors, random forest regression demonstrated highest performance compared to the other models in three out of six clusters, exponential smoothing model had the highest performance in two clusters and ARIMA model in one. Moreover, across all clusters, random forest regression outperformed at least one univariate model, while VAR was outperformed by at least one univariate model in all cases.

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Elisa Kallela

Ruostumattoman teräksen hinnanennustaminen

Kauppatieteiden pro gradu -tutkielma

2024

97 sivua, 35 kuvaa, 33 taulukkoa ja 4 liitettä

Tarkastaja(t): Tutkijatohtori Shadid Ahmed Bhat ja Apulaisprofessori Jan Stoklasa

Avainsanat: Aikasarja ennustaminen, eksponetiaalinen tasoitus, autoregressiivinen intergroitu liukuva keskiarvo, vektori autoregressivinen, satunnainen metsä regressio, ruostumaton teräs, tapaustutkimus

Hinnan ennustamisen avulla yritykset voivat tehdä tietoon perustuvia päätöksiä, hallita riskejä, optimoida resurssien allokointia sekä pysyä kilpailukykyisinä dynaamisilla markkinoilla. Tämä tutkimus pyrkii tutkimaan erilaisia kvantitatiivisia hinnan ennustamismalleja ennustakseen tapausyrityksen hinnoitteludataa. Tutkimuksessa käytettiin neljä eri kvantitatiivista mallia, joista kaksi ovat yhden muuttujan mallia ja kaksi monimuuttuja mallia, joissa hyödynnetään myös ulkoista dataa.

Hinnoitteludata kostui netto vertailuhinnoista (NRP) ja siinä oli 33 havaintoa tammikuusta 2021 syyskuuhun 2023. Tutkimuksessa käytettyyn aineistoon oli valittu kuusi erilaista muuttujaa hinnoitteludatasta, jotka vaalittiin niiden volyymien, laatujen ja asiakkuuksien perusteella. Tarkoituksena oli varmistaa sekä muuttujien merkitys yritykselle, että muuttujien monipuolinen edustus tutkimuksessa. Eksponentiaalinen tasoitusmalli, autoregressiivisen integroidun liukuvan keskiarvon (ARIMA) perhe, vektori autoregressiivisen (VAR) ja satunnainen metsä regressio mallit testatiin näille kuudelle klusterille ja mallit arvioitiin sen perusteella, kuinka hyvin ne pystyvät ennustamaan aikasarjan viimeiset kolme havaintoa.

Eksponentiaalinen tasoitus ja satunnainen metsä regressio suoriutuivat paremmin kuin bencmark-mallina toimivat eksperttien ennusteet jokaisessa kuudessa muuttujassa. Keskimääräisen absoluuttisen prosentuaalisen virheiden perusteella, satunnainen metsä regressio menestyi parhaiten kolmessa kuudesta muuttujasta verrattaessa muihin malleihin, eksponentiaalinen tasoitus menestyi parhaiten kahdessa muuttujassa ja ARIMA-malli yhdessä kulsterissa. Lisäksi satunnainen metsärergessio suoriutui paremmin kuin ainakin yksi yhden muuttujan malli kaikissa tapauksissa, kun taas vektori-autoregressiivinen mallia paremmin suoriutui ainakin yksi yhden muuttujan malli kaikissa tapauksissa.

ACKNOWLEDGEMENTS

With the submission of this thesis, I end my academic journey at LUT. I am eternally grateful for the experiences I had at LUT, which have turned my years a student into something truly amazing. As I leave behind my life as a student, I carry with me numerous cherished memories, valuable friendships, enhanced skills, and newfound knowledge.

I would like to express my gratitude to my supervisor Shahid Bhat for his guidance throughout this work. He provided me with great support throughout the entire writing process.

I want to extend my gratitude to Outokumpu Oyj. Special thanks to my supervisor, Matti Aksela, for guiding me and giving me support throughout the entire process. Additionally, I am grateful to Kari Hänninen for suggesting this thesis topic and assisting me with pricing-related matters throughout the project.

In addition, I want to convey my thanks to friends and family for their encouragement during this thesis writing process and the continuous support they have provided throughout my entire academic journey.

05.02.2024

Elisa Kallela

ABBREVIATIONS

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Aikake's Information Criteria
AR	Autoregressive
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with Exogenous Inputs
ESTER	Euro Short Term Rate
GDP	Gross Domestic Product
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MSE	Mean Squared Error
NN	Neural Networks
NRP	Net Reference Price
PACF	Partial Autocorrelation Function
PFEM	Point Forecast Error Measure
RSS	Residual Sum of Squares
SARIMA	Seasonal Autoregressive Integrated Moving Average
SMAPE	Symmetric Mean Absolute Percentage
SBIC	Schwarz's Bayesian Information Criteria
SES	Simple Exponential Smoothing
VAR	Vector Auto-Regressive
VEC	Vector Error Correction

Table of contents

Appendices

Figures

Tables

1. Introduction

Forecasting in business plays a crucial role in activities such as planning, strategizing and decision-making (Thomakos, Wood, Ioakimidis & Papagiannakis 2023). The purpose of forecasting is to predict what might happen in the future (Waller & Fawcett 2013) which allows more efficient decision-making, better cost control and quicker reactions to different scenarios in the future (Sanders 2017, 5-6).

The steel industry is considered as one of the key industries that serves many large and small industries. Steel industry growth reflects on economic growth as a result of them being causally related and impacted by one another (Mehmanpazir, Khalili-DAmghani & Hafezalkotob 2019). The steel industry is considered highly cyclical, which makes forecasting a vital tool for steel companies. However, this cyclicality also poses challenges for accurate forecasting.

The aim of this thesis is to study different quantitative forecasting methods for predicting Net Reference Prices (NRPs) for Outokumpu Oyj, a globally known stainless steel producer. A quantitative approach to price forecasting can serve as a vital tool and bring significant value to the company. Four different quantitative approaches, exponential smoothing models, the family of Autoregressive Integrated Moving Average (ARIMA), Vector Autoregressive (VAR) model and random forest regression, are compared in order to find what method among these demonstrates the highest performance. Using these methods in forecasting allows for a comprehensive analysis, utilising the company's historical price data in univariate time series models and incorporating relevant external data in the multivariate analyses.

1.1. Previous research

Over the last few decades, forecasting metal prices has become popular, with numerous researchers applying different techniques to improve the accuracy of forecasting metal prices (Ozdemir, Bulus & Zor 2022). Moreover, forecasting the price of steel has been broadly pursued subject among researchers. Adli (2020) forecasted 6-month steel prices in Turkey and revealed that the ARIMA model with explanatory variables (ARIMAX) could not perform superior results compared to ARIMA models. Mancke (1968) studied the U.S. steel industry employing multiple regression analysis to examine how traffic and imports influenced domestic steel prices. Wu & Zhu (2012) utilized Neural Network (NN) models to forecast the price of eight steel products from the steel market in China. Chou, CC & Yang (2012) explored the connection between crude oil prices and global steel prices with the VAR model, discovering a long-run positive relationship between the variables. Adli & Sener (2021) found that the non-stationary VAR model performs better than the vector error correction (VEC) model when forecasting the United States steel prices index using cointegrated variables.

In addition, forecasting the steel supply and demand has been investigated in previous literature. Rogers (1987) developed a supply and demand model for the American steel industry using regression analysis. Chen, Clements, Roberts & Weber (1991) used the VEC model to forecast crude steel consumption in China. Azadeh, Neshat, Mardan & Saberi (2013) forecasted steel demand using the Neural Networks, fuzzy regression and linear regression with US and Iran steel consumption. Mehmanpazir et al. (2019) employed multiple logarithmic regression analyses to fit supply and demand functions for Iran’s steel market. Igarashi, Kakiuchi, Daigo, Matsun & Adachi (2008) used logistic and conventional regression models were used to predict steel consumption in Asia, with factors such as population growth, economic expansion, and demand patterns for finished products like automobiles and electric appliances.

While there are many research studies published in literature related to the steel industry forecasting analysis, to best of our knowledge none of these research studies have addresses specifically the forecasting of stainless steel. Therefore, investigating the stainless steel price forecasting is crucial, providing valuable insights and tools for industries and stakeholders.

1.2. Research questions

The main objective of this thesis is to study how the four different quantitative forecasting methods can forecast the company’s NRPs. The prices are grouped into several clusters based on the existing contracts, geographical areas, and product grades. Through the integration of historical data, the thesis seeks to enhance the Company’s ability to make datadriven decisions and optimize strategies. The findings of this study collectively affirm the potential benefits of integrating statistical methods into Outokumpu's pricing strategies, promising more accurate and adaptable forecasts that can contribute to better decision-making and ultimately enhance the company's competitive edge in the market. Ultimately, the thesis aspires to investigate if these methods could be valuable tool that forecasts NRPs for the different pricing clusters.

The objectives of the study are divided into three research questions as follows:

1. *Which of the exponential smoothing, ARIMA models, VAR models or random forest regression can outperform the used judgemental forecasts?*

Outokumpu, like many other companies, is already doing regular judgemental forecasts. The first question explores can the forecasting process be improved by utilizing quantitative methods. The judgemental forecasts which are made by the experts in the company serve as a benchmark in this study. This implies that in order to the model to perform well, the model needs to surpass the judgemental forecasts.

2. *Is there a difference in performance of the exponential smoothing, ARIMA model, VAR models or random forest regression?*

NRP forecasting is a unique phenomenon and different methods have different strength. If quantitative methods are better (question 1), the subsequent question seeks to identify the best method among them. Therefore, the models will be also compared against each other. The performance of the models will be assessed by their ability to predict the final three observations in the time series.

3. *Is there a difference in performance between the multivariate models (VAR or random forest regression) and the univariate models (exponential smoothing and ARIMA)?*

In principle, multivariate models contain more information so they should provide better forecast than univariate models. The third and final question explores and confirm this assertion. In this study, two models are univariate models considering solely the historical pricing data, while the remaining two models incorporate additional external data as well. The selection of external data for this study is done thoughtfully, considering factors that are observed to have an impact on stainless steel pricing.

The forecasting accuracy is measured with mean absolute percentage error (MAPE) and through training and test splitting. The current forecast is done based on judgemental forecast, which will be used as a benchmark model that gives minimum performance requirements for the models, which means that the forecasting models need to be able to forecast the last three observations better than the benchmark model.

1.3. Research methods

The data used in this research consists of Outokumpu's historical NRPs clustered in different pricing groups, along with additional external data. The pricing data has several clusters, which are divided based on contracts, geographical area, and stainless-steel grades. For this study, six clusters are selected to test the forecasting models. The external dataset consists of three variables: average monthly nickel price, Gross Domestic Product (GDP) and Euro Short Term Rate (ESTER). The datasets contain 33 monthly observations from January 2021 to September 2023, falling into the category of a small dataset in the context of forecasting. As these datasets are characterized as time series data, time series models are the logical choices for model selection.

Previous studies have focused on comparing different models in time series forecasting with limited dataset and the results have concluded that using more complicated techniques doesn't always lead to superior outcomes (Yu & Swartz 2006; Makridakis and Hibon 2000). For example, Yu & Schwartz (2006) compared fuzzy time series and grey theory to moving average and exponential smoothing models to forecast annual tourism demand. The study found that the length of the time series used to fit the model influenced forecasting accuracy to some extent, but no clear pattern emerged. The results indicate that with short time series, complex models don't generate more accurate forecasts than simple traditional models and the choice of error measure doesn't significantly affect the ranking. (Yu & Schwartz 2006). Moreover, there is no single model in forecasting that consistently outperforms other models in terms of forecasting accuracy (Song & Li 2008).

The methods selected to use in this study are exponential smoothing models, the family of autoregressive integrated moving average (ARIMA) models, Vector Autoregressive (VAR) model and random forest regression. ARIMA and exponential smoothing models can be considered traditional time series methods as they are widely used in business forecasting analysis (Box 1991; Snyder, Koehler & Ord 2002). These methods are univariate time series models which focus on a single time series variable. The univariate models are compared to each other and multivariate models, VAR and random forest regression. VAR and random forest are multivariate models which deal with multiple time series variables simultaneously. Incorporating multivariate models that utilize external data in this study provides a more comprehensive context and captures additional relevant factors.

Each model built will be evaluated based on how well it is able to forecast the last three observations in the time series. This is evaluated by plotting the original data with the forecasted values, providing a clear indication of the model's forecasting accuracy for the last three observations. Moreover, judgemental forecasts, which are made by experts in the company, are used as a benchmark and compared by plotting them against both forecasted values and actual values. To further compare the models, the Mean Absolute Percentage Error (MAPE) is calculated for the models and judgemental forecasts. This metric provides a quantitative measure of the accuracy of the model by evaluating the percentage difference between the predicted and actual values.

1.4. Structure of the thesis

The study is divided into 5 main chapters which are divided into subchapters as necessary. In chapter 2, the literature review is be conducted. The next chapter will go more deeply into what is forecasting and forecasting in business. Then commonly used forecasting methods are introduced in detail in chapter 2 as well. After the literature review, the case company Outokumpu Oyj is introduced with a description of their current price forecasting method in chapter 3. The empirical part of this study begins in chapter 3 with a description of the data set and results of the used models in chapter 4. After the empirical part, the conclusions of the work are presented in chapter 5, where research questions of the work are answered, the content of the work is reviewed and analysed, and finally, purpose topics for further research are presented.

2. A brief overview of forecasting models

Forecasting is attempting to predict the future. Forecasting plays a pivotal role in all organizational operations. For instance, in the field of marketing, forecasting is crucial for estimating future demand and sales, as well as predicting market sizes, emerging competitors' trends, and shifts in consumer preferences. Meanwhile, in finance, forecasting is used to evaluate financial performance, determine capital investment needs, and predict stock prices and investment portfolio returns. In order to make successful forecast, the phenomenon needs to be somewhat repetitive or deterministic and it cannot be chaotic or completely stochastic. For example, forecasting individual stock prices has proven challenging, as evidenced by the long-term success of index funds shows. (Sanders 2017, 5-11).

Armstrong (2001, 1-3) describes stages of forecasting as shown in Figure 1. Forecasting consists of six stages: formulating the problem, obtaining information, selecting methods, implementing methods, evaluating methods and using forecasts. It begins with defining the forecasting problem and collecting relevant data. After that, suitable forecasting methods are chosen and applied to the data. Subsequently, the accuracy of the forecasts is evaluated and if the forecasts are not meeting the desired level of accuracy, this stage may lead back to considering alternative methods, as shown in Figure 1. Finally, the generated forecasts can be used for example to help decision-making.

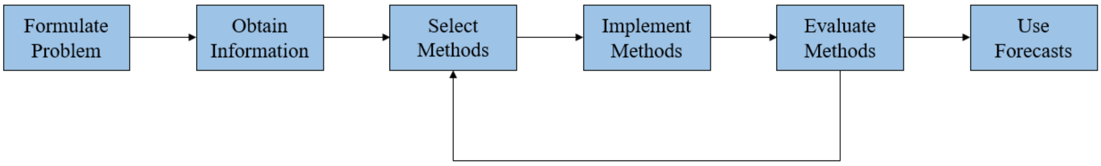


Figure 1: Stages of forecasting (Armstrong 2001, 1)

Forecasting methods can be categorized into two primary branches: judgmental or qualitative and statistical or quantitative methods (Sanders 2017, 50; Mentzer & Moon 2005). Judgemental methods typically rely on the opinions of experts to make predictions. Quantitative methods utilize collected data to forecast a future quantity or quantities of interest. (Sanders 2017, 51). The quantitative methods can be further divided into time series methods and explanatory methods. Time series methods use historical data to make predictions and explanatory methods seek to comprehend how factors affect predicted variables. (Mentzer & Moon 2005). The next subchapter will discuss how to select the right forecasting method.

2.1. Forecasting method selection

According to Armstrong (2001, 365-370) judgemental methods are preferred when there are major changes in the forecasting area, forecasts are done frequently and/or there are disagreements among decision makers. For selecting a quantitative method, the level of knowledge about relationships, the amount of change involved, the type of data, the need for policy analysis, and the extent of domain knowledge should be taken into consideration.

Moreover, Armstrong (2001, 365) described six ways to select quantitative forecasting methods: convenience, market popularity, structured judgement, statistical criteria, relative track records and principles from published research. Armstong (2001, 366) doesn't recommend using the convenience or market popularity since convenience has a high risk and market popularity may not be related as it overlooks some methods. Structured judgment involves forecasters developing explicit criteria first and then rating various methods against them. Statistical criteria, such as distribution of errors or statistical significance, are recommended to use together with structured judgement. Relative track records are the comparative performance of various methods and principles methods that have worked well in similar situations in the past. Relative track records can be expensive but useful, meanwhile, principles from published research can be a low-cost and effective approach but require a lot of work. (Armstrong 2001, 366-376).

Accuracy is often seen as the most important criterion in forecasting (Yokum & Armstrong 1995). However, Yokum & Armstrong (1995) suggest that criteria for evaluating forecasting methods may vary depending on the specific contact. The authors recommend researchers consider a variety of criteria beyond accuracy when comparing forecasting methods, such as ease of interpretation, use, flexibility, employing available data, and implementation (Yokum & Armstrong 1995).

2.2. Judgemental forecasting

Judgemental forecasting methods are the most common methods in business practice. Judgemental forecasting methods rely on the subjective assessments and judgements of individuals such as managers, sales staff, or customers. Consequently, these forecasts are inherently subjective and subjected to numerous human biases. The judgmental forecast offers notable advantages such as their ability to consider exceptional or unique events. However, they are also highly susceptible to human cognitive limitations and biases such as short-term memory and optimism. (Sanders 2017, 52).

Forecasters may use different types of judgement heuristics to make judgemental forecasts. The heuristic used depends on the nature of the information available to the forecaster (Harvey 2007). Heuristics related to forecasting are for example representativeness (Kahneman and Tversky 1973) and anchor-and-adjustment (Hogarth & Makridakis 1981; Lawrence & O'Connor 1992). In cases where forecasters possess information about correlated variables, the representativeness heuristic is employed. This involves selecting a variable that represents the one to be predicted. (Kahneman and Tversky 1973). When forecasting future values based on past data, people often use anchor-and-adjustment heuristics, adjusting their forecast from a reference point such as the last data point to account for trends or autocorrelation (Lawrence & O'Connor 1992).

2.3. Time series methods

Time series models analyse the historical data, time series, and predict the continuation of historical patterns (Makridakis & Wheelwright 1989, 23). Time series models are categorized into two main types: univariate and multivariate. Univariate time series models focus on predicting future values by analysing single time serie variable, meanwhile, multivariate time series models make forecasts based on multiple time series variables. Essentially, univariate models focus on understanding and predicting the behaviour of a single variable over time, while multivariate models consider the relationship between multiple variables making multiple predictions. (Brooks 2008, 206).

There are four types of main patterns in time series: trend, cyclical, seasonal and irregular pattern. The time series can include only one pattern or combination of the patterns. (Sanders 2017, 78-79). A trend is present when data shows a long-term increase or trend decrease. This type of pattern can be found in various economic indicators such as the gross national product as it exhibits a continuous trend in their movement over time. A cyclical pattern occurs when the data exhibit rises and cyclical falls that don't follow a consistent timeframe. In economic contexts, these are usually due to economic fluctuations such as those associated with the business cycle. For instance, product sales, such as steel often display this kind of pattern. (Makridakis & Wheelwright 1989, 24).

A seasonal pattern exists when a dataset is affected by recurring season factors, such as the quarter of the year, month, or day of the week. This pattern can be found for example in sales of products such as ice cream. An irregular pattern refers to movement in the dataset that is unrelated to a seasonal or cyclical pattern. These irregular patterns can manifest as isolated and unpredictable events, like natural disasters, that disrupt the otherwise expected patterns within the time series data. The primary difference between a seasonal and a cyclical pattern is that the seasonal pattern is of a constant length and recurs on a regular periodic basis. In contrast, the cyclical patterns have varied in length. Moreover, cyclical patterns tend to be longer than seasonal patterns. (Makridakis & Wheelwright 1989, 24-25).

Another important characteristic to investigate in a time series is stationary which is a statical property of a time series. A time series is said to be weakly stationary if the time series has constant mean, constant variance and the covariance is independent of time. A strictly stationary time series is one where the distribution of the values remains the same through time. Some time series forecasting models require the time series to be stationary, due to their constant statistical properties. For example, if in an autoregressive model the time series is stationary, the coefficients will exhibit the unfortunate property that previous values of the error term will have a non-declining effect on the current value of y_t as time progresses. (Brooks 2008, 216). The weak stationary is enough to ensure the statistical properties, therefore in this study when talking about stationarity, weak stationarity is meant.

Often time series are not stationary and require transformation. The most common way to transform non-stationary time series to stationary is using first or second-order differencing. Differencing involves computing the differences between consecutive observations in a time series. The primary goal in differencing is to remove the trend or seasonality in a time series, which often makes the series non-stationary. (Brooks 2008, 220).

2.3.1. Exponential smoothing

Exponential smoothing (Brown 1956) is a forecasting technique that calculates an average of past data points. It relies on the weighted average of past observations, where the weight decreases exponentially as one goes further back in time. The appropriate exponential smoothing method depends on the characteristics of the data. For instance, if there's no clear trend or seasonal pattern, the appropriate model is Single (or Simple) Exponential Smoothing (SES): $f_{t+1} = \alpha * y_t + (1 - \alpha)f_t$ (1)

where f_{t+1} is the desired forecasted value, f_t is the latest forecasted value, y_t is the actual value at time t and α is the exponential smoothing parameter which ranges from 0 to 1. (Petropoulos et al. 2022).

In order to use SES, the initial forecast and the exponential smoothing parameter need to be estimated. The standard way to estimate both of these has been by minimising the sum of squares of the one-step ahead forecast errors (Petropoulos et al. 2022).

The exponential smoothing model capable of handling local trends is known as double smoothing or Holt's method (Holt 2004, originally published in 1957). The double smoothing model is a two-component model, including level and trend components. Holt's method with additive trend can be characterised as follows:

$$L_t = \alpha * y_t + (1 - \alpha) * (L_{t-1} + T_{t-1}) \tag{2}$$

$$T_t = \beta * (L_t - L_{t-1}) + (1 - \beta) * T_{t-1} \tag{3}$$

$$f_{t+1} = L_t + T_t * k \tag{4}$$

Where L_t is the level component, L_{t-1} is the previous level, T_t is the trend component, $T_{t-1}$is the previous trend, α is the smoothing parameter for the level, β is the smoothing parameter for the trend and k is the number of forecasts into the future. (Holt 2004)

An extension of the double smoothing method is the triple smoothing method, also known as the Holt-Winters method (Winters 1960). The triple smoothing method adds a third component, seasonality, to the level and trend component from double smoothing. The Holt-Winters method with additive trend and seasonality can be expressed as:

$$L_t = \alpha * (y_t - S_{t-m}) + (1 - \alpha) * (L_{t-1} + T_{t-1}) \tag{5}$$

$$T_t = \beta * (L_t - L_{t-1}) + (1 - \beta) * T_{t-1} \tag{6}$$

$$S_t = \gamma * (y_t - L_t - T_{t-1}) + (1 - \gamma) * S_{t-m} \tag{7}$$

$$f_{t+1} = (L_t + T_t) * S_{t+h-m*(k+1)} \tag{8}$$

Where S_t is the seasonal component, m is the number of periods in season, S_{t-m} is the previous trend with previous seasonality component and γ is the smoothing parameter for seasonality.

The trend and seasonal components can take either an additive or multiplicative form, leading to Holt’s method having two iterations, while Holt-Winters has four iterations. The additive approach for time series data is considered as the sum of its components as seen in the formulas above, while multiplicative is the product of its components. (Koehler, Snyder & Ord 2001).

2.3.2. ARIMA models

The family of Autoregressive Integrated Moving Average (ARIMA) models (Box, George, Jenkins & Gwilym 1967) is often used in time series forecasting. It includes models such as Moving Average (MA), Autoregressive (AR) and a combination of these to ARIMA models. The Moving Average of order q, MA(q), process uses a linear combination of past white noise error terms. MA model smooths out the noise in the data by averaging values over consecutive time periods. The order term ‘q’ tells how many past error terms are considered in the prediction. MA(q) process can be expressed as follows:

$$y_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + .. + \theta_q u_{t-q} \tag{9}$$

where μ is the constant, u_t is a white noise error term at time t, u_{t-q} is the lagged white noise error term , θ_q are the parameters and q is the order of the model. (Brooks 2008, 2006-212).

On the other hand, the Autoregressive Process of order p, AR(p), signifies that the present value of y is determined solely by the previous values of y, with the addition of an error term. The AR(p) can be expressed as,

$$y_t = \mu + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + .. + \emptyset_p y_{t-p} + u_t \tag{10}$$

where μ is the constant, y_t is the observed value at time t, y_{t-p} is the lagged value, $\emptyset_p$ are the parameters, u_t is a white noise term and p is the order of the model (Brooks 2008, 215).

Autoregressive Moving Average (ARMA) model is a combination of AR and MA processes. ARMA model states that the current value of some series y depends linearly on its previous values plus a combination of current and previous values of a white noise error term. (Brooks 2008, 233). The model can be written as,

$$y_t = \mu + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + .. + \emptyset_p y_{t-p} + \theta_1 u_{t-1} + \theta_2 u_{t-2} + .. + \theta_q u_{t-q} + u_t \tag{11}$$

θ_q are the moving average parameters and $\emptyset_p$ are the autoregressive parameters. (Brooks 2008, 215).

To determine which of the ARIMA family models to select and the order of the model, autocorrelation function (ACF) and partial autocorrelation function (PACF) are typically plotted. The ACF measures the degree of correlation between the observations of time series that are separated by k time units (Box 2016). The PACF measures the correlation between an observation from k periods ago and the present observation while adjusting for the influences of observations at intermediate lags. (Brooks 2008, 222). Table 1 summarizes how the characterises of each model are shown in the functions. For example, if the AFC graph seems to be geometrically decaying and in the PACF graph the lags become abruptly zero after lag 3, the AR(3) model could be appropriate.

Table 1: Characteristics of AR, MA and ARMA process in ACF and PACF

	AR(p)	MA(q)	ARMA
ACF	Geometrically decaying	The cutoff to zero after lag p	Geometrically decaying
PACF	Cutoff to zero after lag p	Geometrically decaying	Geometrically decaying

Another technique to help determine the right model and order is information criteria. Information criteria consist of a combination of two components: one that relies on the residual sum of squares (RSS), and another that accounts for the reduction in degrees of freedom caused by adding extra parameters. The two most popular information criteria are Akaike’s information criteria (AIC) (1974), and Schwarz’s Bayesian information criteria (SBIC) (1978). SBIC is strongly consistent but inefficient and AIC is not consistent but is generally more efficient, meaning that as sample size increases, AIC it may not always select the true underlying model, yet it is generally more efficient, tending to choose models with better predictive performance in finite sample sizes. (Brooks 2008, 232-233). Therefore, in this study, AIC is used. AIC can be formulated as:

$$AIC = \ln(\hat{\sigma}^2) + \frac{2k}{T} \tag{12}$$

where $\hat{\sigma}^2$ is the residual variance (also equivalent to the residual sum of squares divided by the number of observations), k is the total number of parameters estimated and T is the sample size.

The ARIMA modelling, as opposed to ARMA modelling, includes the extra letter ‘I’ which stands for ‘integrated’. An integrated autoregressive process is characterized by having a root on the unit circle. Generally, researchers transform the variable by differencing it as required, and then they construct an ARMA model based on these differenced values. An ARMA(p, q) model applied to a variable differenced ‘d’ times is essentially equivalent to an ARIMA(p, d, q) model applied to the original data. (Brooks 2008, 233).

2.3.3. VAR model

A vector autoregressive (VAR) model (Sims 1980) is a multivariate time series model which deals with multiple time series variables. The VAR model consists of a collection of linear regression equations that explain how endogenous variables change over time. In each equation, every variable is expressed as a function of lagged values of all the variables in the system. Like other time series models, VAR models require the series to be stationary. (Petropoulos et al. 2022). VAR(1) model with two variables is formulated as follows:

$$y_{1\mathrm{t}} = B_1 + \alpha_{1,1} * y_{1,\mathrm{t}-1} + \alpha_{1,2} * y_{2,t-1} + u_{1t} \qquad (13)$$

$$y_{2\mathrm{t}} = B_2 + \alpha_{2,1} * y_{1,\mathrm{t}-1} + \alpha_{2,2} * y_{2,\mathrm{t}-1} + u_{2t} \qquad (14)$$

Where B_1 and B_1 are the constants, $\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}$and $\alpha_{2,2}$ are the coefficients, $y_{1,t-1}$ and $y_{2,\mathrm{t}-1}$ are the lagged values at the time point $t-1$, u_{1t} and $u_{2t}$are the white noise terms (Brooks 2008, 327). The formula shows that VAR requires rather many parameters to be estimated, which leads to a requirement of a large number of observations for larger orders of VAR models.

2.4. Explanatory methods

The explanatory models assume that the variable being forecast is related to other variables (Sanders 2017, 78). The purpose of the explanatory model is to discover the form of the relationship and apply it to predict future values of the target variable. In explanatory forecasting, it is assumed that any change in inputs will have a predictable impact on the system's output assuming the explanatory relationship will not change. (Makridakis & Wheelwright 1989, 8-10).

A type of typical explanatory model is linear regression. This scenario is used when there are only two variables. When more variables are added, multiple regression will be used. Multiple regression extends regression by looking at a relationship between the independent variable and multiple dependent variables. The general formula for multiple linear regression is as follows:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_k * x_k \qquad (15)$$

where y is the dependent variable, β_0 is the intercept term, $\beta_1 \dots \beta_\mathrm{k}$ are the regression coefficients and $x_1 \dots x_\mathrm{k}$ are the independent variables. (Sanders 2017, 85-86).

2.4.1. Tree-based regression

Another approach to regression is tree-based methods. Three-based methods are simple and useful for interpretation. In tree-based methods, several trees are created and then combined to generate a single consensus prediction. For regression tasks, the prediction is typically the mean or median of predictions made by each individual tree. (James, Witten, Hastie, Tibshirani & Taylor 2023). Two common tree-based regressions are bootstrap aggregating or bagging (Breiman 1996) and random forest regression (Kam 1995).

The bagging regression technique involves creating multiple decision trees on different subsets of the training data and then combining their predictions. The fundamental concept behind bagging is to train multiple trees on different data subsets and average their predictions. The trees are developed separately using random samples of the observations, resulting in the trees being quite similar to each other. One issue with bagging arises when there is one exceptionally strong predictor in the dataset, alongside several moderately strong predictors. In such cases, bagging may not significantly reduce variance compared to a single tree. Random forests address this issue by restricting each split to consider only a subset of the predictors. (James et al. 2023).

In the construction of decision trees within a random forest, each time a split in a tree is considered, a random sample of predictors is chosen as split candidates from the full set of predictors. Predictions from individual trees are averaged to obtain the final ensemble prediction. It is common to use a larger number of trees, in random forests, as in many cases by increasing the number of trees doesn’t lead to overfitting, but there’s a risk of underfitting if the number of trees is too small. Random forest is effective for capturing complex relationships in data, reducing overfitting, and providing robust predictions in various regression scenarios. (James et al. 2023)

2.5. Evaluation and validation of forecasting models

Evaluation and validation are crucial in forecasting as they collectively assess and ensure the performance, generalization, and reliability of models, guiding the selection to make informed decisions based on accurate and meaningful results. In this study, the train-test split is used, which means that the dataset is split into two subsets: a training set used to train the model and a separate test set used to evaluate its performance.

2.5.1. Accuracy measures

There are many common point forecast error measures (PFEM), such as the mean squared error (MSE), mean absolute error (MAE), mean absolute scaled error (MASE), and mean absolute percentage error (MAPE) to evaluate the forecast accuracy. (Petropoulos et al. 2022) In this study, to evaluate the model performance MAPE is used, which is a commonly used metric for evaluating the accuracy of a forecasting model (Makridakis & Wheelwright 1989, 43). MAPE is typically expressed as a percentage, and it provides a measure of the average absolute percentage difference between the predicted and actual values. The lower the MAPE, the better the forecasting accuracy. The MSE and MAPE formulas are defined as follows:

$$MSE\ = \frac{1}{n}\sum_{i=1}^{n}(y_t - f_t)^2, \tag{16}$$

$$MAPE\ = \frac{1}{n}\sum_{i=1}^{n} |\frac{y_t - f_t}{y_t}| \times 100 \tag{17}$$

where n is the number of observations, y_t is the observed value at time t and f_t is the forecasted value at time t. The MAPE is not without drawbacks. The most critical drawback is that it is biased towards low forecasts (Armstrong & Collopy 1992). Therefore, Symmetric mean absolute percentage error (SMAPE) was proposed to handle the drawbacks of the MAPE metric. The max value of SMPAE is 200%. (Chicco, Warrens & Jurman 2021). The SMAPE formula is defined as follows:

$$SMAPE\ = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - f_t|}{(|y_t| + |Y_t|)/2} \tag{18}$$

where n is the number of observations, y_t is the actual value at time t and f_t is the forecasted value at time t. (Chicco et al. 2021). In this study, both MAPE and SMAPE are chosen as the accuracy metrics due to their comparability and their presentation in percentage form, which conceals the absolute values, as the actual NRP values cannot be directly revealed.

2.5.2. Time series cross-validation

Cross-validation is a methodology for determining the optimal model and parameters by iteratively training and testing on different data subsets. It involves multiple rounds of the train-test split with varying data for each iteration. (James et al. 2023). In the context of time series data, time series cross-validation is employed to evaluate predictive models. This technique divides time-ordered data into consecutive segments, training the model on earlier segments and testing on later ones. The process is repeated for each segment, simulating how well the model generalizes to unseen future data, crucial in time series analysis where observation order matters. Common types of time series cross-validation include “rolling" or "expanding" window methods, where the training set gradually incorporates more data over time, and "fixed origin" methods, where a fixed training window is used with a sliding testing window. (Deng 2023).

Figure 2 illustrates an example of time series cross-validation divination, showcasing the dataset used in this study with 33 observations. In this scenario, the time series is divided into 4 folds, and the testing set always consists of the most recent 3 observations. The first fold contains the whole dataset. The second fold contains the same dataset except the last three observations. Consequently, the training se in the second fold comprises 27 observations and test set consist of the last three observations. This pattern repeats for the remaining folds, where the last three observations are consistently excluded, leading to a new test set. This is the same split that is used in this study.

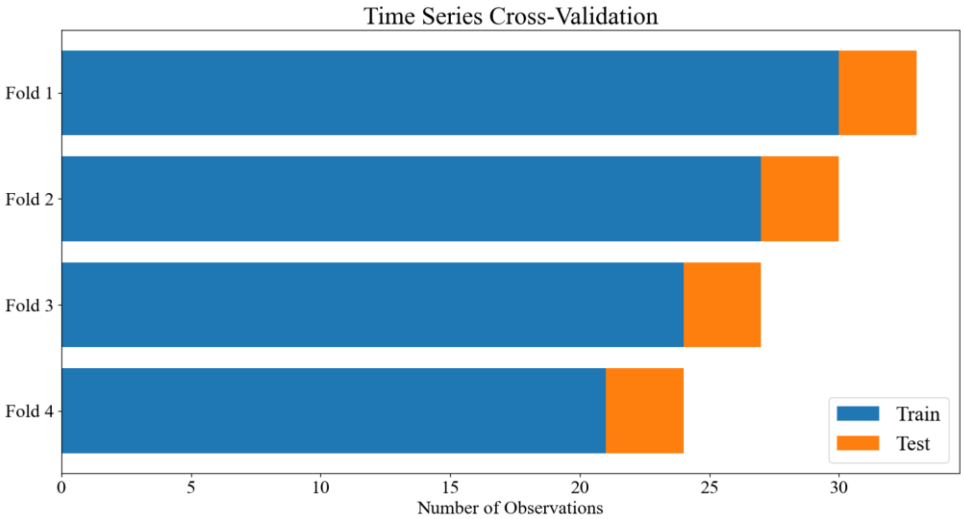


Figure 2: Time series cross-validation

2.5.3. Residuals

When assessing the model's performance, it's crucial to examine the residuals. Residuals are the differences between the observed and predicted values. If the residuals exhibit characteristics of white noise, implying no discernible pattern or correlation, the model can be considered a good fit. In contrast, if residuals show a pattern or correlation, it indicates that there are still aspects of the data that the model has not captured. In such cases, it might be necessary to consider an alternative model. The autocorrelation can be tested for example with the Ljun-Box test, which tests whether the residuals are autocorrelated. (Petropoulos et al. 2022).

3. Case Outokumpu Oyj

This chapter provides a thorough exploration of the case company, as well as, describing the company’s current approach to price forecasting and the extensive dataset utilized in this study. A brief introduction to the company is followed by an overview of their existing price forecasting methodology. Subsequently, a detailed examination is conducted on the comprehensive set of internal and external datasets used in this study.

3.1. Outokumpu Oyj

Outokumpu Oyj is a stainless steel producer headquartered in Helsinki, Finland. Outokumpu is the market leader in cold-rolled stainless steel in Europe and the second-largest stainless steel producer in the Americas. Outokumpu produces stainless steel in mills in Finland, Germany, Mexico, Sweden and the US. In addition, Outokumpu is the owner of the largest known chromite reserves in Europe, which is located in Finland. (Outokumpu 2022).

Outokumpu is known for the quality of its products and expertise in stainless steel. The customers of Outokumpu use their stainless steel in the construction of bridges and buildings, produce cars, trains and trucks as well as in the production of various household appliances and utensils. Outokumpu’s customer base is all over the world in the European, Middle-Eastern, Asian and African markets. Outokumpu’s vision is to be the customer’s first choice in sustainable stainless steel. (Outokumpu 2022).

Sustainability is one of the key drivers in Outokumpu. Outokumpu is the global leader in sustainable stainless steel and their stainless steel has the smallest carbon footprint on the market. Outokumpu has a strong track record in sustainable performance and ambitious climate targets. Outokumpu’s total carbon footprint is less than 30% of the global average. As well as Outokumpu’s regular production being the current sustainable leader in the industry, Outokumpu has an emission-minimized Circle Green stainless steel line. The Circle Green product line has up to 92% lower carbon footprint than the average. (Outokumpu 2022).

Europe is Outokumpu’s biggest business area and Europe brings 66% of the Company’s net sales. Moreover, Outokumpu is the cost leader in high-volume stainless steel products in Europe. (Outokumpu 2023). In the year 2022, the sales for the European business area were 6,266 million euro (Outokumpu 2022).

3.2. Price forecasting at Outokumpu

At the moment, Outokumpu Oyj is using a judgmental approach to forecast Net Reference Prices in the specific business lines in Europe. At Outokumpu, five experts are forecasting NRPs every month for the end of the year and next year at the moment. The prices are grouped into several clusters based on the existing contracts, geographical areas, and product grades. The experts are responsible for the clusters based on the market area. The experts make their forecasts based on already placed orders, old and new customers and history of the prices. Moreover, the experts ensure that the price forecasts are in line with the volume forecasts made by the supply department.

The number of clusters has been continuously changing. Before organizational changes in 2022, the number of price clusters was nine for the whole Europe business area. The increase in the number of clusters was made possible by a higher level of granularity included in deployed Enterprise Resource Planning software SAP which allows the exact NRP to be calculated. Now the forecasting is done for each business line separately in the business area of Europe, taking into account order intake and how future monthly invoicing will be derived from it. With the improved data quality, a statistical approach can be introduced and bring significant value to the company.

External organizations have made statistical price forecasting for Outokumpu; however, these forecasts have typically produced a very limited number of forecasts. External organizations have forecasted for example monthly total order intake and total price for specific products in Europe and Asia based on market data.

3.3. Case data

The next part describes the dataset comprehensively. The data description and empirical research of this study are done via Python. The data set consists of Outokumpu’s price data and external data. Outokumpu’s price data consists of the historical actual NRPs and judgemental forecasts. The external data consists of nickel price, Gross Domestic Product (GDP) and Euro Short Term Rate (ESTER).

3.3.1. NRPs

The NRP data contains 33 observations from January 2021 to September 2023. The pricing data has several clusters, which are divided based on contracts, geographical area, and stainless-steel grades. For this study, six clusters are selected to test the forecasting models. The six clusters are chosen based on their volumes, grades and customers, with the goal of ensuring both the significance of the cluster to the company and a diverse range representation of the clusters in the study. This data is sensitive and contains information that is not allowed to be published, therefore, the NRP values are not shown in this research.

Each cluster with its NRPs is plotted in the Figure 3. Based on Figure 3, none of the clusters seems to show any clear trends. The highest peaks occurred in 2021 in all clusters except for clusters 1 and 4. Clusters 2, 3 and 5 reached their highest points in December 2021, while cluster 6 reached its highest peak in November 2021. Cluster 1, on the other hand, reached its highest peak in April 2022 and cluster 4 in April 2023. The lowest drops happened in 2023 expect for clusters 1 and 4. More precisely, clusters 2 and 5 experienced their lowest values in NRPs in June 2023, while cluster 3 experienced its lowest point in February 2023 and cluster 6 in December 2023. Finally, cluster 1 had its lowest value in January 2021 and cluster 4 in May 2021.

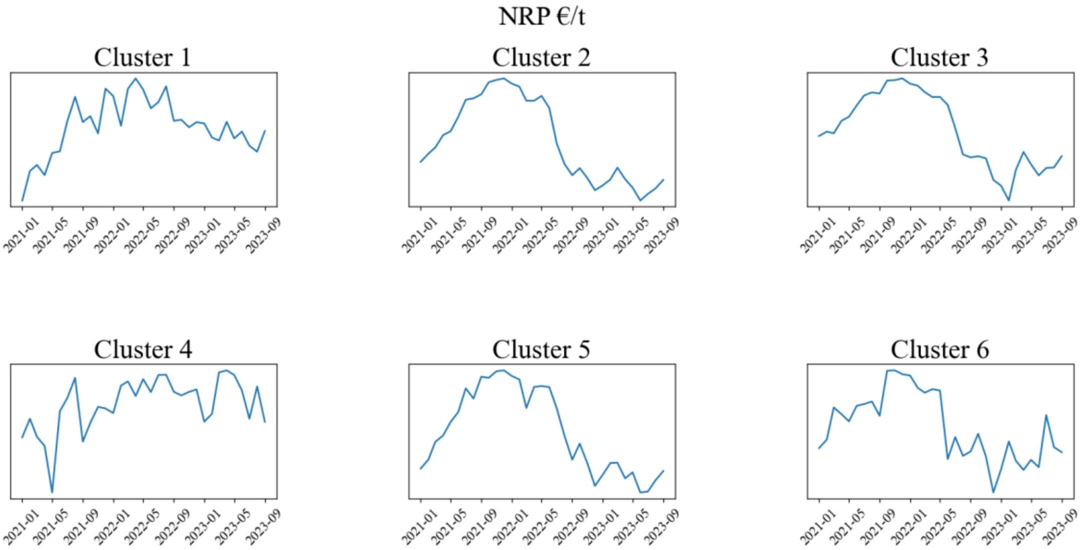


Figure 3: NRPs

For each cluster, stationarity is tested with the Augmented Dickey-Fuller (ADF) test. The results of the ADF tests are presented in Table 2. The results indicate that only two of the clusters, clusters 3 and 4, are stationary with p-values of 0.042 and 0.013. The p-values are less than 0.05, meaning that we can reject the null hypothesis with a 5% confidence level. Other clusters fail to reject the null hypothesis with high p-values which indicates that time series are non-stationarity.

Table 2: Augmented Dickey-Fuller (ADF) test results on original NRPs

Cluster	Test statistic	1%	5%	10%	p-value
Cluster 1	-2.758	-3.770	-3.005	-2.643	0.065
Cluster 2	-1.725	-3.689	-2.972	-2.625	0.418
Cluster 3	-2.933	-3.711	-2.981	-2.630	**0.042**
Cluster 4	-3.348	-3.654	-2.957	-2.618	**0.013**
Cluster 5	-0.996	-3.654	-2.957	-2.618	0.755
Cluster 6	-1.436	-3.738	-2.992	-2.636	0.565

Since four clusters are non-stationary, first-degree differencing is done for these datasets to ensure stationary. After the new series is created, stationery is again tested with the ADF tests. The results of the second ADF tests are presented in Table 3. The result shows that cluster 5 is now stationary. With a confidence level of 1% the null hypothesis can be rejected in cluster 5. The null hypothesis cannot be rejected in other cases. Therefore, second-degree differencing is done for clusters 1, 2 and 6.

Table 3: ADF test results on first-degree differenced NRPs

Cluster	Test statistic	1%	5%	10%	p-value
Cluster 1	-1.416	-3.700	-2.976	-2.628	0.574
Cluster 2	-1.529	-3.689	-2.972	-2.625	0.519
Cluster 5	-5.003	-3.661	-2.961	-2.619	**0.000**
Cluster 6	-1.840	-3.753	-2.998	-2.639	0.361

Stationarity is again tested with the second-degree differenced series for clusters 1, 2 and 6. The results of the third ADF test are presented in Table 4. The results show that clusters 1 and 2 are now stationary. With a confidence level of 1% the null hypothesis can be rejected in these cases, but not in the case of cluster 6. Therefore, for cluster 6, third-degree differencing is performed, and stationarity is again tested.

Table 4: ADF test results on second-degree differenced NRPs

Cluster	Test statistic	1%	5%	10%	p-value
Cluster 1	-8.128	-3.700	-2.976	-2.628	**0.000**
Cluster 2	-5.912	-3.689	-2.972	-2.625	**0.000**
Cluster 6	-1.956	-3.738	-2.992	-2.636	0.306

The final results of the ADF test on the third-degree differenced cluster 6 show that the cluster 6 time series is stationary. The results are shown in Table 5. The p-value is 0.000 and, therefore, the null hypothesis can be rejected. All the differenced clusters are now stationary and the differenced datasets a plotted in Appendix 1. From the Appendix 1, we can see now that the datasets appear relatively uniform and stable throughout the observed time period.

Table 5: Dickey-Fuller test results on third-degree differenced NRPs

Cluster	Test statistic	1%	5%	10%	p-value
Cluster 6	-7.232	-3.737	-2.992	-2.636	0.000

3.3.2. Nickel price

Nickel is the essential metal alloy in stainless-steel production (Outokumpu 2022), which means that NRPs can be exposed to price changes in nickel. The monthly nickel prices are extracted from of the London Metal Exchange (LME). The nickel prices are nickel cash settlements. The LME data is in USD/Tonne; therefore, it’s transformed to Euro/Tonne with the average monthly exchange rates from LME. The monthly average nickel prices from January 2021 to September 2023 are plotted in Figure 4.

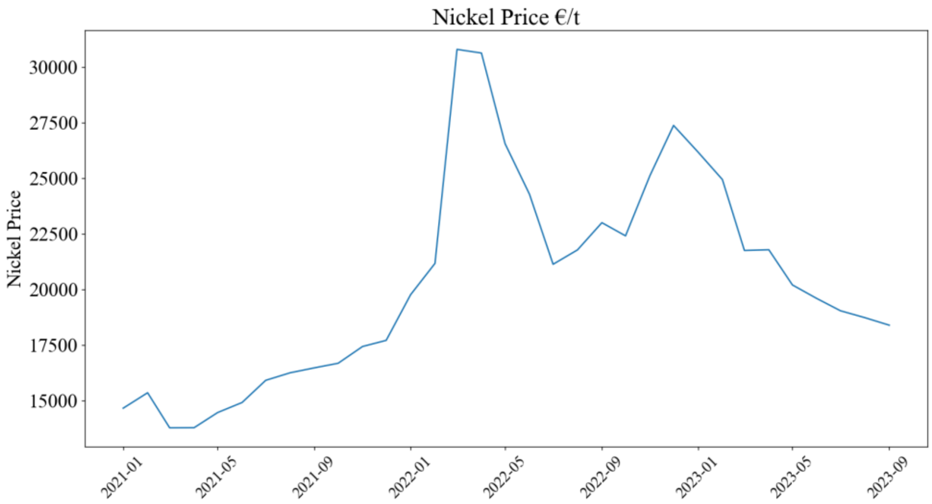


Figure 4: Monthly nickel price (LME)

From Figure 4, it is evident that the nickel was the most expensive in March of 2022. During the same month at least 3 price clusters experienced notable drops in their values. Moreover, the nickel price had quite a high peak in January 2023 as well, when many price clusters experienced drops in their NRPs again. On the other hand, the nickel price was the cheapest in March 2021. During the same month, three price clusters experienced their lowest values in NRPs. All of the clusters are visualized together in a single figure with each external data in Appendix 2. This visualization utilizes the same scaling technique known as max-min scaling, where variables are transformed by subtracting the minimum value and dividing the range of the variable.

The correlation matrix for the clusters and external data is presented in Appendix 3. From the matrix, we can see that nickel has a strong positive correlation with cluster 1 (0.604) and cluster 4 (0.540). The strong positive correlation indicates that as one variable increases, the other variable tends to increase as well. With clusters 2 (-0.129), 3 (-0.226) and 6 (-0.222) nickel price has a weak negative correlation. A weak negative correlation indicates that as one variable increases, the other variable tends to decrease slightly, but the relationship is not very strong. For cluster 5 the correlation is close to zero, indicating no correlation at all.

3.3.3. Eurozone GDP

Economic growth impacts steel industry growth (Mehmanpazir et al. 2019), therefore, Gross Domestic Product (GDP) growth in the Euro area is chosen as one of the external variables. The Eurozone GDP data was collected from Refinitiv Eikon. The GDP data is expressed as a percentage change year over year, meaning that it measures how much Eurozone GDP has changed compared to the same period in the previous ear.

The Eurozone GDP is expressed quarterly and plotted in Figure 5. From the Figure 5, we can see that the data had the highest peak in the second quarter of 2021. This means that the GDP changed 14.8 compared to the same quarter in 2020. This can be related to the COVID-19 pandemic. In the second quarter of 2020, the whole world experienced a sharp economic contraction due to the pandemic and associated lockdowns. By the second quarter of 2021, many Eurozone countries had started reopening their economies as they lifted COVID-19 restrictions.

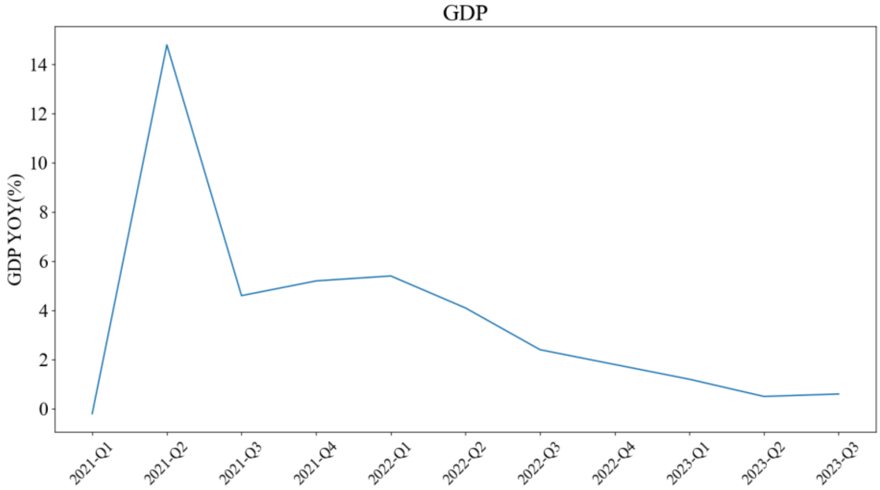


Figure 5: Quarterly Eurozone GDP (Refinitiv Eikon 2023)

The European GDP data is converted into monthly rates to align with the monthly nature of the price data aimed to predict. The data is converted into monthly values by dividing each quarterly rate by 3 to get the equivalent monthly rate. The correlation matrix in Appendix 3, shows that GDP shows the highest correlation with cluster 2 (0.512) and cluster 3 (0.517). On the other hand, the smallest correlation value is observed in cluster 1 (-0.030), indicating a lack of correlation between GDP and cluster 1. Cluster 4 exhibits a moderate negative correlation of -0.445, while clusters 5 and 6 exhibit a moderate positive correlation with GDP, having values of 0.485 and 0.454.

3.3.4. Interest rate

Interest rates can have an indirect effect on steel prices through their impact on the overall economy and the cost structure of the steel industry. The interest rate used in this research is the Euro Short Term Rate (ESTER). The ESTER or €STR rate is a measure of the overnight borrowing costs of banks in the Eurozone. It shows the interest rates at which banks lend to each other without requiring collateral. This rate is published every Euro system TARGET2 business day based on transactions that are conducted and settled on the previous TARGET2 business day with maturity. These transactions are considered to be conducted at arm's length and represent unbiased market rates. (European Central Bank 2023). The monthly short-term interest rates from January 2021 to September 2023 are plotted in Figure 6. The

Figure shows that the interest rates were negative until August of 2022, after which they began to increase steadily, following a nearly linear line.

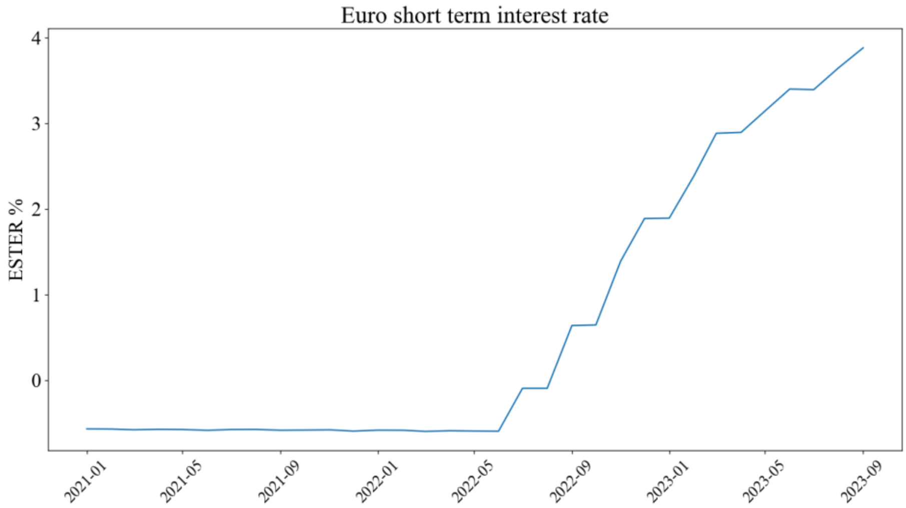


Figure 6: Eurozone short-term interest rate (Refinitiv Eikon 2023)

The correlation matrix in Appendix 3, shows that ESTER exhibits a very strong negative correlation with cluster 2 (-0.806), cluster 3 (0.803) and cluster 5 (-0.796). In addition, cluster 6 exhibits a strong negative correlation of -0.642 and cluster 1 has a weak negative correlation of -0.180. Cluster 4 is the only cluster that exhibits a positive correlation of 0.261 with ESTER.

3.3.5. Statistical analysis and stationarity of external data

Table 6 presents data statistics for the external variables, including nickel prices, GDP growth and interest rate. The data shows notable variability, with standard deviations reflecting dispersion. The range for each variable is substantial, with significant fluctuations in nickel prices. GDP growth and interest rates vary from negative to positive with close fluctuations.

Table 6: External data statistics

	Nickel (€/t)	GDP (%)	Interest rate (%)
mean	20 369.144	1.224	0.651
std	4 711.669	1.354	1.636
min	13 788.970	-0.067	-0.593
max	30 788.340	4.933	3.880
median	19 767.010	0.8	-0.566

For external data, stationarity is also tested with the Augmented Dickey-Fuller test. Table 7 displays the test outcomes, revealing non-stationarities for all variables with p-values of 0.304, 0.293 and 0.834. Therefore, each variable requires at least first-degree differencing. In addition, Table 7 summarizes the outcomes of both first-degree and second-degree differencing if needed. The results suggest that achieving stationarity nickel price and GDP require first-degree differencing, while interest rate requires second-degree differencing. All the differenced variables are now stationary and the differenced datasets a plotted in Appendix 1 as well. From the Appendix 1, we can see now that the datasets appear relatively uniform and stable throughout the observed time period.

Table 7: Stationarity tests for external data

	Test statistic	1%	5%	10%	p-value
Nickel price	-1.960	-3.661	-2.961	-2.619	0.304
GDP	-1.986	-3.752	-3.000	-2.639	0.293
Interest rate	-0.746	-3.680	-2.968	-2.623	0.834
		First-degree	differencing		
Nickel Price	-4.236	-3.661	-2.961	-2.619	**0.001**
GDP	-13.718	-3.679	-2.968	-2.623	**0.000**
Interest rate	-1.248	-3.679	-2.968	-2.623	0.653
		Second-degree	differencing		
Interest rate	-9.018	3.679	-2.968	-2.623	**0.000**

4. Analysis and implementation of forecasting models

This chapter describes the method implementation of the research. First, exponential smoothing models are tested with the original NRP datasets. The datasets are partitioned into training and testing sets, with the training set including all observations except the last three (circa 91% of the dataset), and the testing set consisting of the last three observations (circa 9% of the dataset). Subsequently, the model is trained using the training data and evaluated using the testing data. After exponential smoothing models, ARIMA and VAR models are tested. These models require datasets to be stationarity, and as such, they employ either the original data or if required the differenced data. After ARIMA and VAR models, random forest regression is tested as well.

The models are built and tested in Python. In this research, 5% is used as the confidence level which means that the null hypothesis is rejected at a 0.05 level. The results of the research are therefore valid with 95% confidence and at the same time, the probability of error is 5%.

4.1. Exponential smoothing

Simple exponential smoothing (SES), double smoothing and triple smoothing are all tested for each cluster and the model with the lowest MAPE is selected for further analyses. It’s important to note that due to the nature of the model, the SES model provides a constant prediction for the entire forecasting horizon. On the other hand, Holt’s method provides a linear outcome that incorporates both level and trend components. The Holt-Winters takes this a step further, providing a more complex outcome by including level, trend, and seasonality components.

The optimal smoothing pattern (alpha) level in the SES model is selected by a loop built in Python that calculates MAPEs for each model with alpha levels from 0.1 to 0.9 with a 0.1 gap and finds the alpha that results in the lowest MAPE. The alpha with the lowest MAPE is selected to use in the SES model.

4.1.1. Cluster 1

For cluster 1, the optimal alpha level was found as 0.6, therefore the SES model is built with this alpha value. Subsequently, the double and triple smoothing models are as well evaluated for cluster 1. The outcomes of these evaluations, specifically the Mean Absolute Percentage Errors (MAPEs), are detailed in Table 8, showcasing the accuracy of different exponential smoothing models applied within the cluster.

The judgemental forecasts, serving as the benchmark, exhibit a MAPE of 7.08%. Among all models, the double smoothing model with additive trend demonstrates the lowest MAPE at 4.08% and double smoothing models with multiplicative trend has a slightly higher MAPE of 4.30%. The SES model has slightly higher 5.08% MAPE than both of the double smoothing models. The triple smoothing models are not far away, exhibiting MAPEs ranging from 9.68 to 11.45%.

Table 8: Exponential smoothing models MAPEs for cluster 1

Model	MAPE
Judgemental forecast (Benchmark)	7.08%
Simple Exponential Smoothing	5.08%
Double Smoothing, Trend: Multiplicative (Mul.)	4.30%
Double Smoothing, Trend: Additive (Add.)	**4.08%**
Triple Smoothing, Trend: Mul., Seasonal: Mul.	10.67%
Triple Smoothing, Trend: Mul., Seasonal: Add	11.45%
Triple Smoothing, Trend: Add., Seasonal: Add.	10.46%
Triple Smoothing, Trend: Add., Seasonal: Mul.	9.68%

Since the double smoothing model with the additive trend model exhibits the lowest MAPE, this model is used for further analyses. The model exhibits an AIC of 266.52**.** The double smoothing model is applied to make the predictions, and the results are compared with both the original values and judgemental forecasts, as depicted in Figure 7. From the Figure 7, it seems that the model is following the trends of the actual values but with delay. Moreover, the drops and peaks are underestimated. When predicting the last three observations, the forecasted values decrease slightly. The forecasted values are higher than the actual values in the first two observations, but since the actual value increases in the last observations, the forecasted value decreases lower than the actual values. Furthermore, the model's predictions appear to be a better fit than the judgemental forecasts, as the judgemental forecasts are predicted to be lower than the actual values.

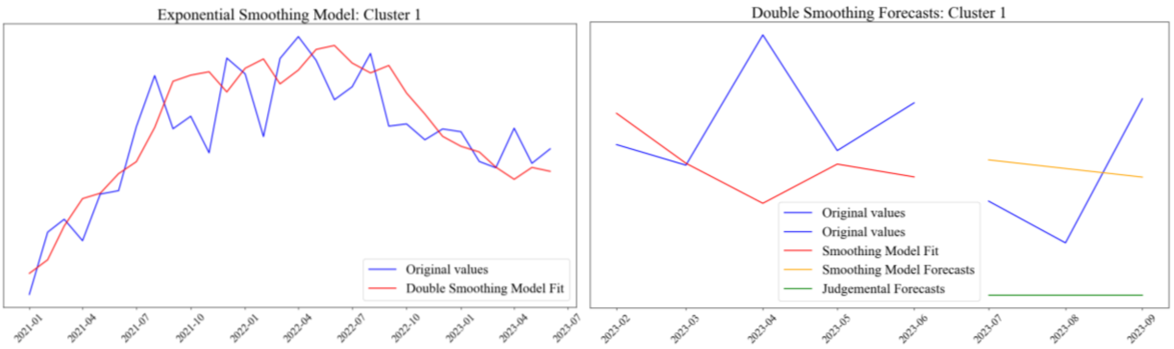


Figure 7: Double Smoothing model for cluster 1

4.1.2. Cluster 2

In cluster 2, the ideal alpha level was found as 0.4, leading to the construction of an SES model using this alpha value. The MAPEs for each exponential smoothing model are presented in Table 9. The judgemental forecast, serving as the benchmark, has a MAPE of 45.50%. Among the exponential smoothing models, the SES model demonstrates the lowest MAPE at 28.24%, which is lower than the benchmark model as well. Double smoothing models with multiplicative and additive trends have MAPEs of 89.70% and 155.82%, indicating that the multiplicative is a more suitable trend in double smoothing models. The triple smoothing models' MAPEs range from 85.34% to 198.34% with the lowest MAPE found in models containing multiplicative trend and multiplicative seasonality.

Table 9: Exponential smoothing models MAPEs for cluster 2

Model	MAPE
Judgemental forecast (Benchmark)	45.50%
Simple Exponential Smoothing	**28.24%**
Double Smoothing, Trend: Multiplicative (Mul.)	89.70%
Double Smoothing, Trend: Additive (Add.)	155.82%
Triple Smoothing, Trend: Mul., Seasonal: Mul.	85.34%
Triple Smoothing, Trend: Mul., Seasonal: Add	141.78%
Triple Smoothing, Trend: Add., Seasonal: Add.	198.34%
Triple Smoothing, Trend: Add., Seasonal: Mul.	102.71%

Since the SES model exhibits the lowest MAPE, this model is used for further analyses. The model's AIC is measured at 342.11. The SES model is utilized, and the model's fit and predictions are plotted in Figure 8. Overall, the model seems to effectively capture the trends of the actual values but with some delay. From the Figure 8, we can see that for the second observation, the model is a perfect fit when examining the predictions for the last three observations. However, as the actual values show a linear increasing trend, the forecasted values are higher in the first observation and lower in the last observation. Moreover, it is evident that the model's predictions appear to be a better fit than the judgemental forecasts, which are forecasted to be clearly lower than the actual values and the model's forecasts.

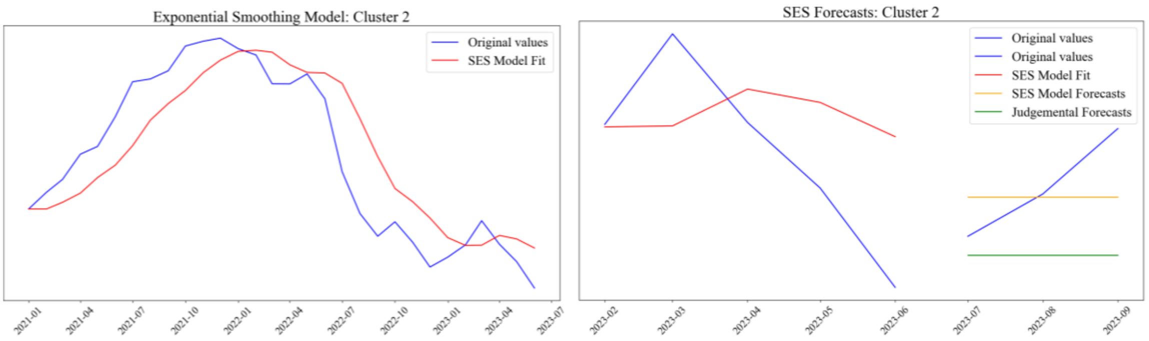


Figure 8: SES for cluster 2

4.1.3. Cluster 3

Due to the nature of the cluster 3 data, exponential smoothing models with multiplicative trends or seasonality cannot be used, therefore, only the SES model, double smoothing model with additive trend and triple smoothing model with additive trend and additive seasonality are tested. For cluster 3, the optimal alpha level was found as 0.3, therefore the SES model is built with this alpha value. The MAPEs for each exponential smoothing models are presented in Table 10. The judgemental forecast, serving as the benchmark, has a MAPE of 108.77%. In comparison, the SES model demonstrates lower MAPE at 80.00%, which is the lowest among all exponential smoothing models as well. While MAPE of 80.00% is deemed high, it remains lower than the benchmark model, making it acceptable in this study. The double smoothing model with an additive trend exhibits a MAPE of 309.21%, while the triple smoothing model with an additive trend and additive seasonality shows an even higher MAPE at 528.79%.

Table 10: Exponential smoothing models MAPEs for cluster 3

Model	MAPE
Judgemental forecast (Benchmark)	108.77%
Simple Exponential Smoothing	**80.00%**
Double Smoothing, Trend: Additive (Add.)	309.21%
Triple Smoothing, Trend: Add., Seasonal: Add.	528.79%

Since the SES model exhibits the lowest MAPE of the exponential smoothing models, this model is used for further analyses. The model exhibits an AIC of 577.51. The SES model is applied to forecast predictions for the last three observations and plotted in Figure 9. The model appears to follow the trend from the actual values but underestimating most of the values. When predicting the last three observations, the forecasted values are close to the first two actual values, however, as the last value increases, the model falls behind in accuracy. Moreover, the Figure shows that the forecasted values are not quite far from the judgemental forecasts but still being a better fit as they are closer to the actual values.

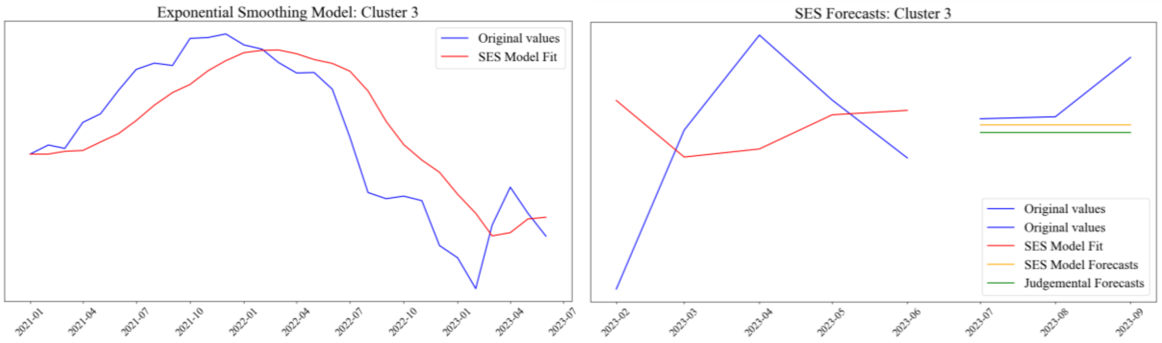


Figure 9: SES for cluster 3

4.1.4. Cluster 4

Due to the nature of the cluster 4 data, smoothing models with multiplicative trends or seasonality cannot be used, therefore, only the SES model, double smoothing model with additive trend and triple smoothing model with additive trend and additive seasonality are tested. For cluster 4, the optimal alpha level is found to be 0.1, which is utilized to construct the SES model. Table 11 provides the MAPEs for the exponential smoothing models applied to cluster 4. The judgemental forecast, serving as the benchmark, shows a MAPE of 65.19%. The SES model exhibits a lower MAPE at 26.62%, which is additionally the lowest MAPE of all models. The double smoothing model with an additive trend displays a MAPE of 45.57% and the triple smoothing model with an additive trend and seasonality MAPE of 55.15%. Moreover, these both are lower than the benchmark.

Table 11: Exponential smoothing models MAPEs for cluster 4

Model	MAPE
Judgemental forecast (Benchmark)	65.20%
Simple Exponential Smoothing	**26.62%**
Double Smoothing, Trend = Additive (Add.)	45.57%
Triple Smoothing, Trend: Add., Seasonal: Add.	55.15%

Since the SES model exhibits the lowest MAPE, this model is used for further analyses. The model exhibits an AIC of 329.55. The SES model and predictions are plotted in Figure 10.

The model fit exhibits signs of underfitting, as it fails to capture the inherent patterns and lacks alignment with the observed trends in the actual values. When predicting the last three observations, the forecasted values appear to be in proximity to the most recent peak observed in the actual values. The model's predictions continue to demonstrate a superior fit than the judgemental forecasts as judgemental forecasts are predicted to be higher than the actual values.

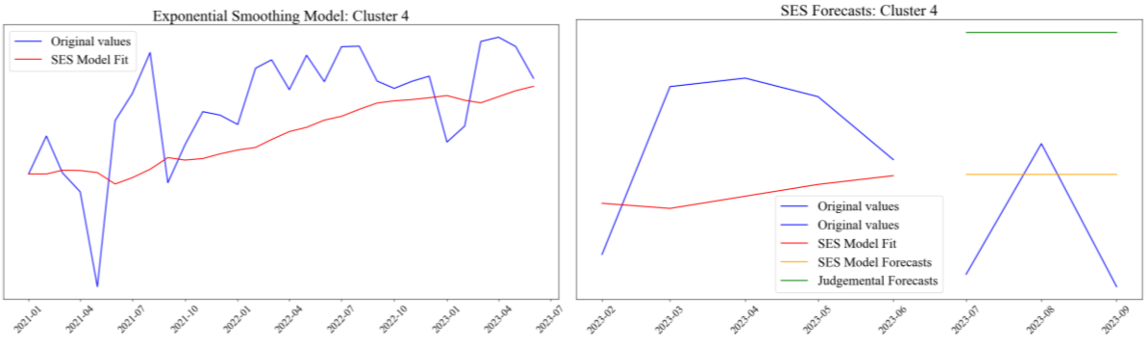


Figure 10: SES for cluster 4

4.1.5. Cluster 5

For cluster 5, the optimal alpha level was found as 0.4 and utilized to build the SES model. The MAPEs for each exponential smoothing models are presented in Table 12. The judgemental forecast, serving as the benchmark, has a MAPE of 14.77% and the SES model exhibits a lower MAPE of 12.12%. The double smoothing models exhibit MAPEs of 24.21% and 30.91%. Among the triple smoothing models, the lowest MAPE is exhibited with multiplicative trend and multiplicative seasonality at 27.89%. The other triple smoothing models exhibit MAPEs ranging from 30.71% to 53.08%.

Table 12: Exponential smoothing models MAPEs for cluster 5

Model	MAPE
Judgemental forecast (Benchmark)	14.77%
Simple Exponential Smoothing (SES)	**12.12%**
Double Smoothing, Trend: Multiplicative (Mul.)	24.21%
Double Smoothing, Trend: Additive (Add.)	30.91%
Triple Smoothing, Trend: Mul., Seasonal: Mul.	27.89%
Triple Smoothing, Trend: Mul., Seasonal: Add	30.71%
Triple Smoothing, Trend: Add., Seasonal: Add.	53.08%
Triple Smoothing, Trend: Add., Seasonal: Mul.	31.62%

Since the simple smoothing model exhibits the lowest MAPE, this model is selected for further analysis. The model exhibits an AIC of 275.70. The fit of the model and predictions for the last three observations are plotted in Figure 11. The model appears to exhibit a good fit, effectively capturing the trends of the actual values but with a delay. When examining the predictions for the last three observations, the model predicts the second value perfectly but overestimates the first value and underestimates the last value as the actual values show an increase. Furthermore, the model's predictions appear to still be a better fit than the judgemental forecasts as the judgemental forecast are close to the actual value for the first observation.

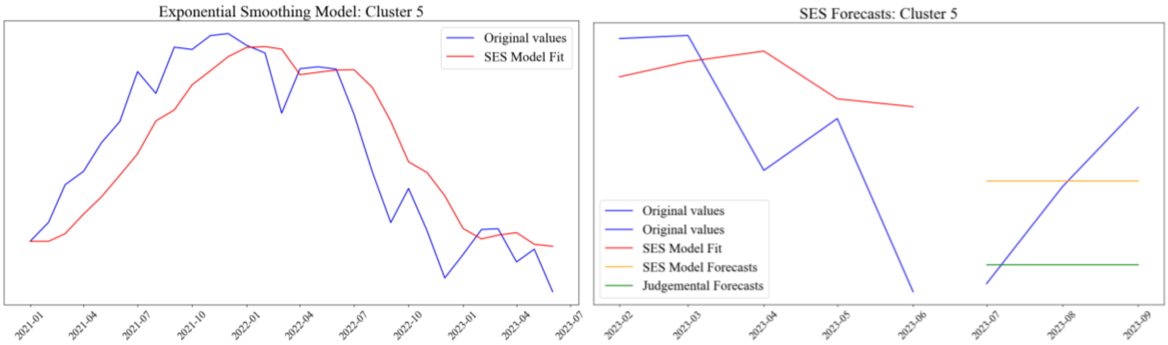


Figure 11: SES for cluster 5

4.1.6. Cluster 6

Due to the nature of the cluster 6 data, smoothing models with multiplicative trends or seasonality cannot be used, therefore, only the SES model, double smoothing model with additive trend and triple smoothing model with additive trend and additive seasonality are tested. In cluster 6, the optimal alpha was determined to be 0.1 which is utilized to construct the SES model. Table 13 provides the MAPEs for the exponential smoothing models applied to cluster 6. The judgemental forecast, serving as the benchmark, shows a significantly high MAPE of 458.72%. The SES model exhibits the lowest MAPE at 133.03%. MAPE of 133.03% is considered to be high, but it is lower than the benchmark model and therefore acceptable in this study. When we move from the SES model to double smoothing and triple smoothing models, there is a noticeable increase in the MAPEs. Specifically, the MAPE for the double smoothing model rises significantly to 567.43%, and for the triple smoothing mode, it further increases to 750.56%. This shift indicates a substantial decrease in the accuracy of the predictions as we move from simpler to more complex smoothing models.

Table 13: Exponential smoothing models MAPEs for cluster 6

Model	MAPE
Judgemental forecast (Benchmark)	458.72%
Simple Exponential Smoothing	**133.03%**
Double Smoothing, Trend: Additive (Add.)	567.43%
Triple Smoothing, Trend: Add., Seasonal: Add.	750.56%

As a result of the SES model exhibiting the lowest MAPE, this model is selected for further analyses. The model’s AIC is 384.16. The model fit and predictions from the SES model are plotted in Figure 12. The model fit portrayed in Figure 12 suggests underfitting, indicating an inability to grasp underlying patterns and a lack of alignment with the observed trends in the actual values. When examining the predictions for the last three observations, it’s evident that the model’s predictions appear to be a better fit than the judgemental forecasts as they are close to the last two actual values.

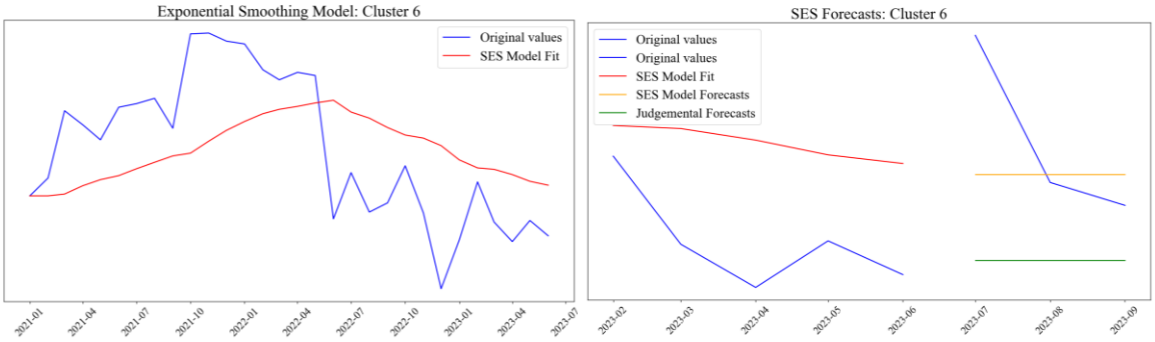


Figure 12: SES results for cluster 6

4.1.7. Validation for exponential smoothing models

After analysing the exponential smoothing models, it is evident that SES models were used in most cases, with the one exception occurring in cluster 1, where the lowest MAPE was found in the double smoothing model with an additive trend. As described in the data description, none of the clusters exhibited clear evidence of trends or seasonality. Consequently, it is logical that among the exponential smoothing models, the SES model performs best given that it doesn’t consider trend or seasonality components, unlike the other models. Overall, the exponential smoothing models performed well in over half of the models. Four clusters exhibit MAPEs under 29% (clusters 1, 2, 4 and 5) while the MAPEs of the others (clusters 4 and 6) range from 80.00% to 133.03%. Moreover, the exponential smoothing models outperformed judgemental forecasts across all clusters.

For each cluster, we are examining the impact on the model's MAPEs when altering the forecasting horizon to six months. In this scenario, the training dataset consists of 27 observations, and the testing dataset consists of 6 observations. Across all clusters, except for Cluster 4, the MAPEs exhibit an increase. Specifically, cluster 4 experienced a decrease in MAPE from 26.62% to 20.57. Cluster 1 rises from 4.74% to 9.99%, cluster 2 rises from 28.24% to 161.16%, cluster 3 increases from 80.00% to 282.34%, cluster 5 climbs from 12.12% to 29.23%, and cluster 6 ascends from 133.03% to 253.26%. The increased MAPEs collectively suggest that the accuracy of the forecasting model is decreasing as we extend the prediction period.

Moreover, the best-performing model for each cluster undergoes time series cross-validation, where datasets are partitioned into four folds. The folds are partitioned as described in chapter 4, where the fourth and last fold contains 26 observations, 23 in the training set and the next three in the test set. Table 14 presents the results for each fold and the average of MAPEs in the folds. From Table 14, it is evident that for cluster 1 and cluster 4, the MAPEs decrease or don't increase a lot compared to the first folds MAPE, indicating that the model performing well across the diverse subset of the dataset. For clusters 2, 3 and 6, two or more folds exhibit high MAPEs, which indicates potential issues or outliers in the data or the models. Cluster 5 has both low and high MAPEs across the folds, indicating variability of the model's performance across different subsets of the data. However, it is important to note that the high variability between the folds can be resulted due to the small dataset. Cross-validation with a small dataset can lead to instability in model evaluation, as the limited data may result in high variability between folds.

Table 14: MAPEs of cross-validation for Exponential Smoothing Models

Cluster	Model	Fold 1	Fold 2	Fold 3	Fold 4
Cluster 1	Double Smoothing	4.08%	8.75%	1.34%	6.74%
Cluster 2	Simple Smoothing	28.24%	229.61%	28.28%	135.42%
Cluster 3	Simple Smoothing	80.00%	391.79%	159.73%	390.59%
Cluster 4	Simple Smoothing	26.62%	17.85%	24.39%	9.17%
Cluster 5	Simple Smoothing	12.12%	29.43%	6.21%	50.06%
Cluster 6	Simple Smoothing	133.03%	173.85%	195.50%	262.24%

4.2. ARIMA

For each price cluster, ARIMA models are estimated either with the original data or differenced data. For each price cluster ACF and PACF are plotted and described. Auto_arima function from the pmdarima package in Python is used to help identify the most optimal parameters for an ARIMA model by finding the model that has the lowest AIC. When the optimal model is found, the last three observations are estimated and plotted against the original value and judgemental forecasted values.

4.2.1. Cluster 1

As previously mentioned, the dataset for cluster 1 exhibits non-stationary behaviour and therefore, the second-degree differenced dataset is used to estimate the parameters. The ACF and PACF are presented in Figure 13. For cluster 1, the ACF has a statistically significant lag at lag 1 and PACF shows three statistically significant lags at lag 1, 2 and 4. Moreover, the PACF seems to be geometrically decaying after lag 4. This indicates that appropriate models could be for example MA(1) or AR(4).

The auto_arima search found the lowest AIC value 329.74 1 in ARIMA(4,2,0) model. The results of this model are presented in Table 15. The results indicate that in ARIMA(4,2,0) all the orders are statistically significant with p-values smaller than 0.05. The Ljung-Box is performed as well with the models to test the autocorrelation of the residuals. The results are presented in the same Table 15. The results of the Ljung-Box test indicate that there is no autocorrelation detected in the model with a p-value of 0.96.

The ARIMA(4,2,0) model is chosen to build the forecasts for cluster 1. The model’s fit and the predictions for the last three observations are plotted in Figure 14. From the Figure 14, it seems that the model fits the training data well; however, overestimating some drops and peaks. When predicting the last three observations, it appears that the model starts to predict the values higher than the actual values. However, in the last observation the model correctly forecasts an increase and the last value approaches close to the actual value. Moreover, the forecasted values seem to be a better fit as the judgemental forecasts are predicted to be lower than the model’s predictions and actual values. The MAPEs corroborate the findings from the graphs. The MAPE for the predictions with the ARIMA model is 4.74% while for the judgemental forecasts, it is slightly higher at 7.08%. This suggests that the ARIMA model performs better in predictions.

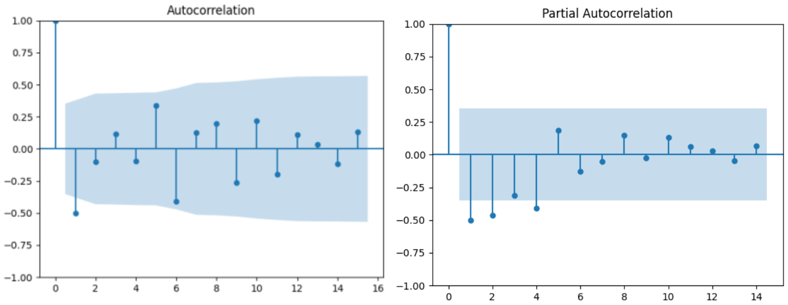


Figure 13: ACF and PACF of cluster 1

Table 15: ARIMA(4,2,0) results for cluster 1

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.279	0.159	-8.062	**0.000**	-1.590	-0.968
ar.L2	-1.375	0.193	-7.135	**0.000**	-1.753	-0.997
ar.L3	-1.016	0.190	-5.337	**0.000**	-1.389	-0.643
Ar.L4	-0.591	0.191	-3.089	**0.002**	-0.966	-0.216
Sigma2	44723.946	1675.759	2.819	**0.005**	1439.517	8008.374
Ljung-Box (L1) (Q)	0.00					
Prob(Q)	0.96					

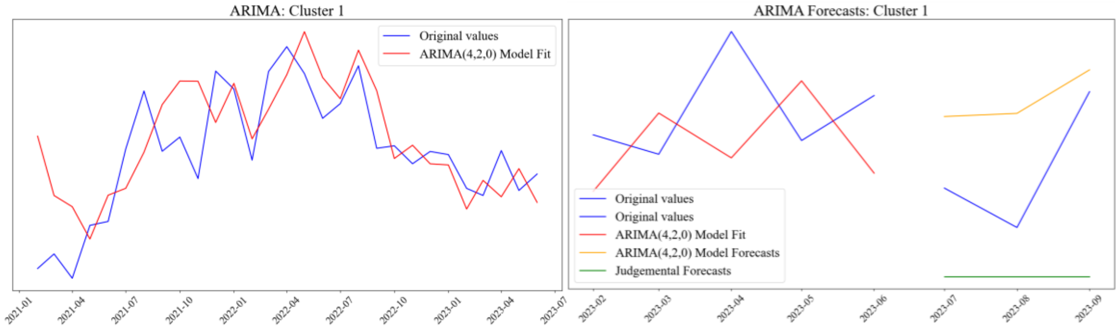


Figure 14: ARIMA(4,2,0) plots for cluster 1

4.2.2. Cluster 2

The cluster 2 data require second-degree differencing to ensure stationary, therefore, the differenced data is used to estimate the parameters. The ACF and PACF of cluster 2 are presented in Figure 15. The ACF doesn't have any significant lags, however, the PACF shows one statistically significant lag at lag 3. This indicates that the AR(3) model might be appropriate.

The auto_arima search found the lowest AIC value of 364.01 in the ARIMA(3,2,0) model. The results of the model are presented in Table 16. The results indicate that only the estimated variance of the residual (sigma2) is statistically significant. The results of the Ljung-Box test indicate that there is no autocorrelation detected in the model with a p-value of 0.75.

The ARIMA(3,2,0) model is chosen to build predictions for cluster 2. The model's fit on training data and the predictions for the last three observations are plotted in Figure 16. From the Figure 16, it seems that the model fits the training data's trends well, however, overestimating some drops and peaks. When predicting the last three observations, it appears that the model predicts the values to be lower than the actual values. Moreover, the model predicts a decrease while the original values increase. The judgemental forecasts seem to perform better compared to the model as the forecasts are closer to the actual values. The MAPEs confirm these findings; the MAPE for judgemental forecasts is 45.50% and for the ARIMA model 115.55%.

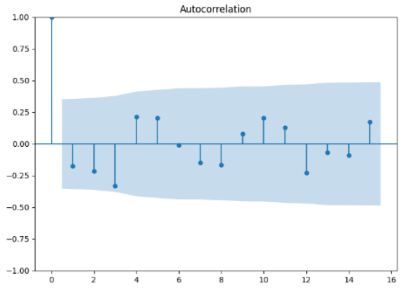


Figure 15: ACF and PACF for cluster 2

Table 16: ARIMA(3,2,0) results

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.3562	0.280	-1.270	0.204	-0.906	0.193
ar.L2	-0.3729	0.230	-1.620	0.105	-0.824	0.078
ar.L3	-0.4346	0.229	-1.902	0.057	-0.883	0.013
Sigma2	18920.0	67209.786	2.624	**0.009**	4787.178	33000.0
Ljung-Box (L1) (Q)	0.10					
Prob(Q)	0.75					

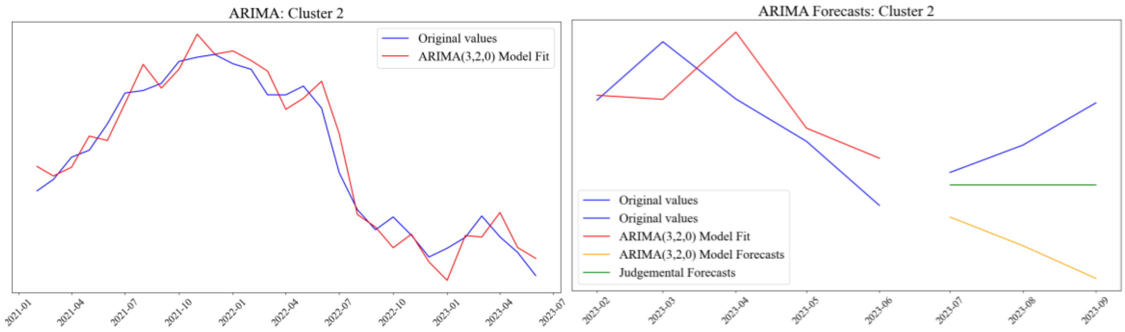


Figure 16: ARIMA(3,2,0) plots for cluster 2

4.2.3. Cluster 3

The dataset for cluster 3 is stationarity, therefore there is no need for differencing and the original dataset is used for estimating the parameters. The ACF and PACF of cluster 3 are presented below in Figure 17. The ACF has two statistically significant lags at lag 1 and lag 2. Moreover, the PACF shows three statistically significant lags at lag 1, 2 and 7. This indicates appropriate models to be, for example, AR models with q = 7 or MA models with p = 1 or 2.

The auto_arima found the lowest AIC value of 423.37 in ARIMA(1,0,1). The results of the ARIMA(1,0,1) are presented in Table 17. The results show autoregressive order of 1 (ar.L1) and sigma2 are statistically significant. The results of the Ljung-Box test indicate that there is no autocorrelation detected in either of the models with a p-value of 0.68.

The ARIMA(1,0,1) model is used to build predictions for cluster 3. The model's fit and the predictions for the last three observations are plotted in Figure 18. As depicted in the Figure 18, the model seems to fit the model quite well. When predicting the last three observations, the model initially forecasts values lower than the actual values, but it anticipates a rising trend. Subsequently, the real values also increase, causing the model to align more closely with the actual values. Moreover, the model surpasses the judgemental values during the second observation. From the Figure, it seems that the model's accuracy is close to the judgemental forecast's accuracy. The MAPE for the forecasted values with ARIMA is 114.51% while for the judgemental forecasts, it is slightly lower at 108.77%. This suggests that the model doesn't outperform the judgemental forecasts.

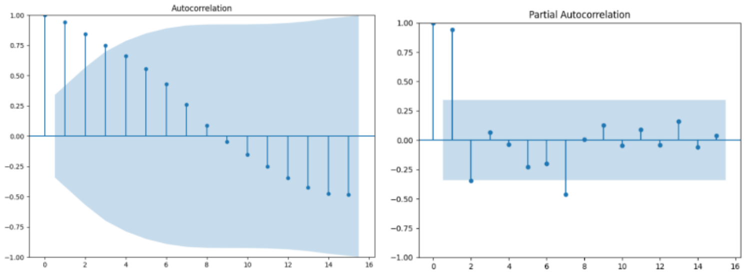


Figure 17: ACF and PACF of cluster 3

Table 17: ARIMA(1,0,1) results for cluster 3

	coef	std err	z	P>|z|	[0.025	0.975]
const	768.788	873.344	0.880	0.379	-942.936	2480.510
ar.L1	0.913	0.076	12.028	**0.000**	0.764	1.062
ma.L1	0.344	0.258	1.334	0.182	-0.161	0.848
Sigma2	55520.0	14200.0	3.906	**0.000**	27700.0	83400.0
Ljung-Box (L1) (Q)	0.17					
Prob(Q)	0.68					

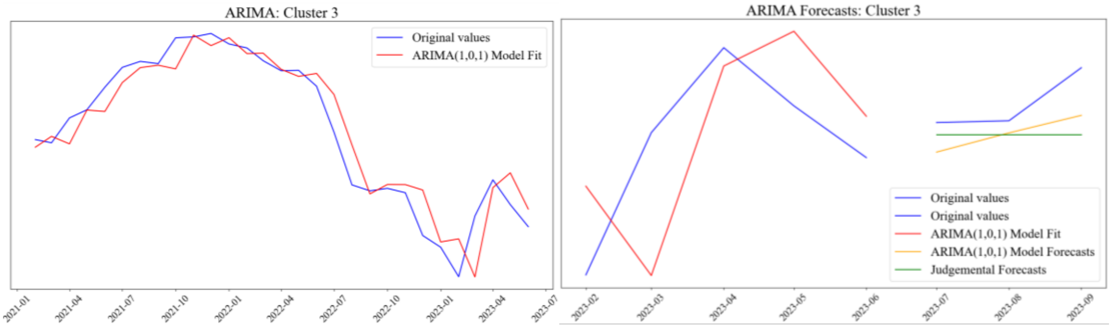


Figure 18: ARIMA(1,0,1) plots for cluster 3

4.2.4. Cluster 4

The dataset for cluster 4 is stationarity, therefore there is no need for differencing and the original dataset is used for estimating the parameters. The ACF and PACF are presented in Figure 19, and both figures seem to have one significant lag at lag 1. Moreover, both functions seem to be geometrically decaying. This indicates that the ARMA(1,1) model for an appropriate model.

Auto_arima found the lowest AIC value of 449.36 in ARIMA(1,0,0) ergo AR(1). The results of the AR(1) are presented in Table 18. The Table 18 shows the autoregressive order of 1 (ar.L1), constant term (const) and estimated variance of the residual (simga2) are all statistically significant. The results of the Ljung-Box test indicate that there is no autocorrelation detected in either of the models with a p-value of 0.89.

The model's fit and the predictions for the last three observations are plotted in Figure 20. The model appears to follow the actual values, however, consistently underestimates both drops and peaks in the data. When predicting the last three values, the model predicts the values to be higher than the actual values. Nevertheless, as the actual values increase during the second observation, the model gradually converges towards the actual values. Unfortunately, when the actual values experience a significant decline in the last observation, the model is unable to follow the drop. Based on the visual representation, it appears that the model's predictions are closer to the actual values compared to the judgemental forecasts which are predicted to be too high. The same conclusion is supported by the MAPEs. Specifically, the model exhibits a MAPE of 23.68% whereas the MAPE for the judgemental forecast is higher at 65.20%.

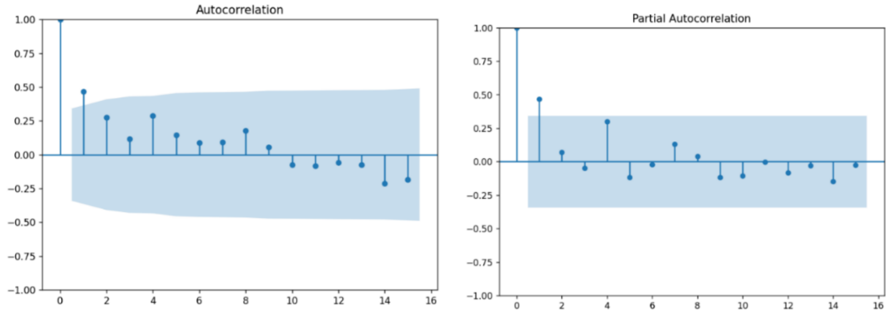


Figure 19: ACF and PACF of cluster 4

Table 18: ARIMA(1,0,0) results for cluster 4

	coef	std err	z	P>|z|	[0.025	0.975]
const	1507.829	184.700	8.164	**0.000**	1145.824	1869.834
Ar.L1	0.534	0.119	4.500	**0.000**	0.302	0.767
sigma2	152100.0	37200.0	4.083	**0.000**	79100.0	225000.0
Ljung-Box (L1) (Q)	0.02					
Prob(Q)	0.89					

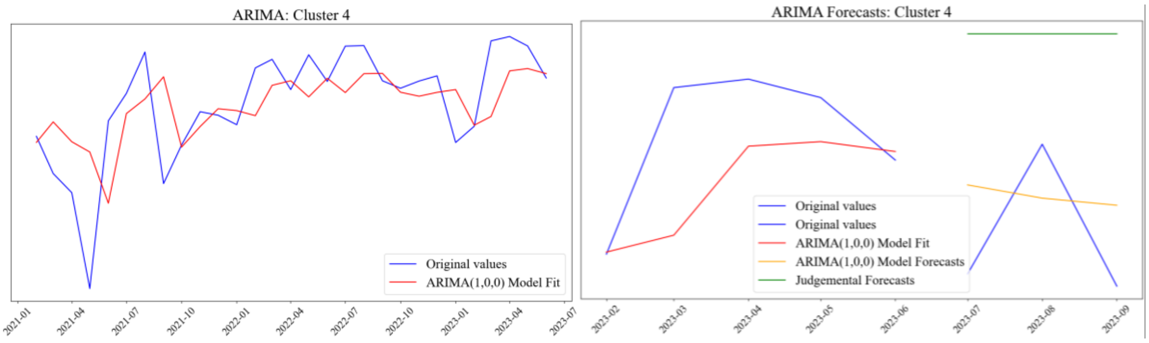


Figure 20: ARIMA(1,0,0) plots for cluster 4

4.2.5. Cluster 5

The dataset for cluster 5 requires first-degree differencing to achieve stationarity. Therefore, the differenced dataset is used for estimating the parameters. The ACF and PACF graphs don’t show any statistically significant lags in Figure 21. The lowest AIC value of 409.65 is found in ARIMA(0,1,0). The ARIMA(0,1,0) model is a first-order differencing model with no autoregressive or moving average components, also referred to as a random walk model. The results of the ARIMA(0,1,0) model are presented in Table 19. The Table shows that only sigma2 is statistically significant. The results of the Ljung-Box test indicate that there is no autocorrelation detected in either of the models with a p-value of 0.60.

The model fit and predictions are plotted in Figure 22. The model appears to exhibit a good fit, effectively capturing the trends of the actual values but with a delay. The model forecasts all of the observations to have the same value as it is the random walk model and doesn't have any autoregressive or moving average components. Therefore, the model is unable to forecast the increasing trend in the last three observations. Moreover, based on the Figure 22, the judgemental forecasts appear to be a better fit than the model's predictions. The judgemental forecasts seem to be closer to the actual values than the model's predictions. The MAPE confirms these findings. The MAPE for the model is 17.09% and for the judgemental forecast, it is slightly lower at 14.76%.

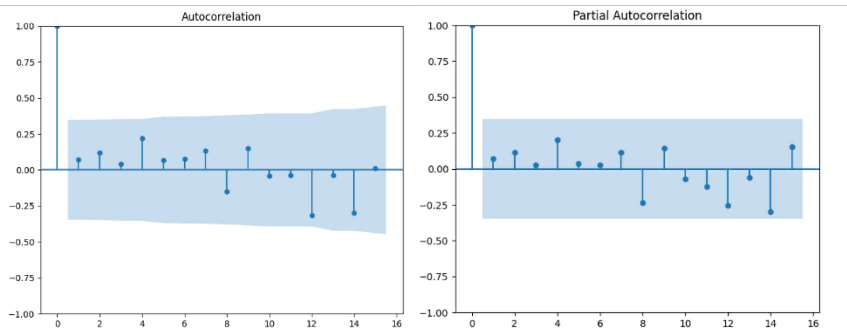


Figure 21: ACF and PACF of cluster 5

Table 19: ARIMA(0,1,0) results for cluster 5

	coef	std err	z	P>|z|	[0.025	0.975]
Sigma2	21360.0	8109.180	2.634	**0.008**	5464.478	37300.0
Ljung-Box (L1) (Q)	0.12					
Prob(Q)	0.73					

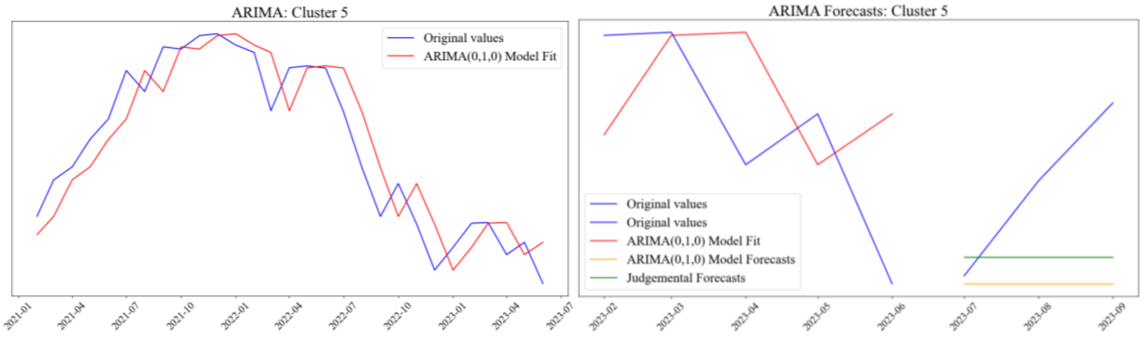


Figure 22: ARIMA(0,1,0) plots for cluster 5

4.2.6. Cluster 6

The data for cluster 6 is not stationarity and requires third-degree differencing to achieve stationarity. Therefore, third-degree differenced data is used to estimate the parameters. The ACF and PACF are presented in the Figure 23. The ACF shows one significant lag at lag 1 and PACF shows three significant lags at lag 1, 2 and 3. Moreover, the PACF seems to be geometrically decaying, indicating an AR model. The lowest AIC value was found in the ARIMA(6,3,0) model with AIC of 410.27. The model is presented in Table 20. Table 20 shows that all the orders and sigma2 in the model are statistically significant. Moreover, the Ljung-Box test shows that there is no autocorrelation detected in the model with a p-value of 0.84.

The ARIMA(6,3,0) model is chosen to build predictions for cluster 6 and the model fit and predictions for the last three observations are plotted in Figure 24. The model doesn't seem to be a good fit for the data based on the Figure 24. For example, in the other half of 2022, the model's values decrease when the actual values increase and vice versa. When predicting the last three observations, it seems that the model predicts a peak in the last three observations. However, during this period, the actual values are declining, and a peak in the values occurred earlier in the sequence. The judgemental values are notably lower than the actual values, indicating a more accurate predictions than the model. The MAPE for judgemental forecast is at 458.72% and for the model is even higher at 543.33%. This suggests that both models struggle to make accurate forecasts, yet judgemental forecasts still outperform the model.

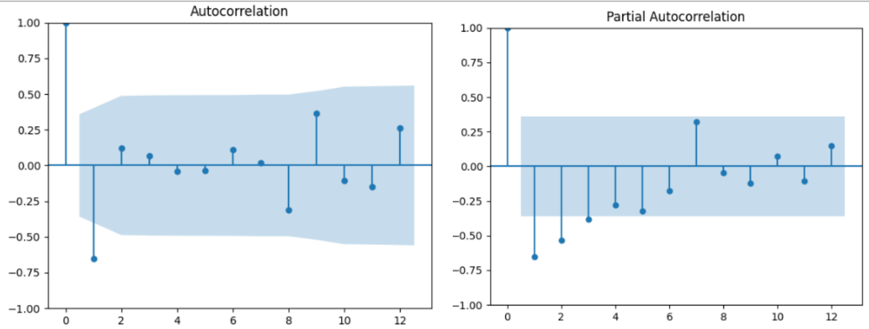


Figure 23: ACF and PACF of cluster 6

Table 20: ARIMA(6,3,0) results for cluster 6

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.876	0.149	-12.576	**0.000**	-2.168	-1.583
ar.L2	- 2.449	0.298	-8.226	**0.000**	-3.033	-1.866
ar.L3	-2.541	0.421	-6.031	**0.000**	-3.367	-1.715
ar.L4	-2.161	0.479	-4.508	**0.000**	-3.101	-1.222
ar.L5	-1.511	0.362	-4.175	**0.000**	-2.221	-0.802
ar.L6	-0.669	0.244	-2.737	**0.006**	-1.147	-0.190
sigma2	115100.0	44800.0	2.571	**0.010**	27400.0	203000.0
Ljung-Box (L1) (Q)	0.04					
Prob(Q)	0.84					

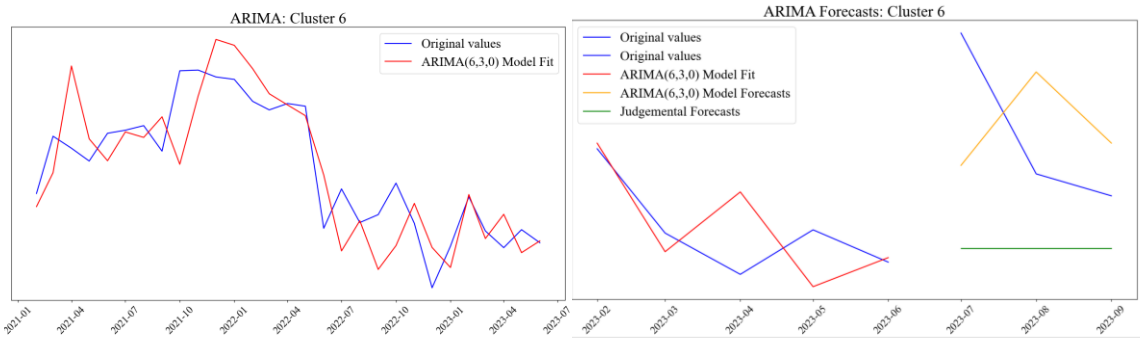


Figure 24: ARIMA(6,3,0) results for cluster 5

4.2.7. Validation for ARIMA models

After analysing the ARIMA models, AR and ARMA models were mostly used. It is also interesting to note that clusters 1 and 2 were second-degree differenced and cluster 6 was third-degree differenced and all found the best model in AR. Cluster 1 demonstrated strong performance when using the AR(4) model on the second-degree differenced dataset, achieving a MAPE of 4.74%. Conversely, clusters 2 and 6 showed poor performance with MAPEs exceeding 110% for AR(3) and AR(6) models on the second and third-degree differenced datasets. In the case of cluster 6 where especially high MAPE were detected, the models used high orders AR(6). This indicates that there might be too complex and there might be overfitting. Moreover, only the models for clusters 1 and 4 could outperform the judgemental forecasts.

For each cluster, the forecasting horizon is changed to six months to observe the changes in MAPEs. In this scenario, the training dataset consists of 27 observations, and the testing dataset consists of 6 observations. Across all clusters, except for Cluster 4, the MAPEs exhibit an increase. Specifically, Cluster 4 experienced a decrease in MAPE from 23.68% to 17.89%. Conversely, cluster 1 rises from 4.74% to 6.41%, cluster 2 from 115.55% to 359.37%, Cluster 3 from 114.51% to 380.18%, cluster 5 from 17.09% to 32.23%, and cluster 6 from 543.33% to 784.47%. The increased MAPEs collectively suggest that the accuracy of the forecasting model is decreasing as we extend the prediction period.

Table 21 presents the results of the cross-validation for each cluster. From the Table 21, it is evident that for cluster 1 and cluster 4, the MAPEs decrease or are similar to the first folds MAPE, indicating that the model performing well across the diverse subsets of the dataset. For clusters 2 and 3, two or more folds exhibit high MAPEs, which indicates potential issues or outliers in the data or the models. Cluster 2 and 3 exhibit both lower and higher MAPEs compared to the first fold, indicating variability of the model's performance across different subsets of the data. Cluster 5 exhibits slightly higher MAPEs across the folds, meanwhile, cluster 6 MAPEs decrease across the folds.

Table 21: MAPEs of Cross-validation for ARIMA models

Cluster	Model	Fold 1	Fold 2	Fold 3	Fold 4
Cluster 1	ARIMA(4,2,0)	4.74%	5.87%	2.65%	2.02%
Cluster 2	ARIMA(3,2,0)	115.55%	435.20%	106.31%	179.26%
Cluster 3	ARMA(1,1)	114.51%	407.72%	64.58%	128.57%
Cluster 4	AR(1)	23.68%	15.92%	25.13%	11.87%
Cluster 5	ARIMA(0,1,0)	17.09%	32.43%	24.70%	22.34%
Cluster 6	ARIMA(6,3,0)	543.33%	98.44%	276.09%	327.18%

4.3. VAR models

For each price cluster, VAR models are estimated either with the original train data or differenced train data. For each cluster and appropriate lag order (p) for the VAR(p) model is determined by fitting VAR models with progressively higher orders and selecting the order that results in the model with the lowest AIC. Given that the dataset consists of 33 observations, the maximum order of lag for the VAR model is set at 3. This is due to higher-order VAR models requiring a larger amount of data. When the optimal model is found, the last three observations are estimated and plotted against the original test data and judgemental forecasted values.

Before building the VAR models, relationships between NRPs and external data are examined with Granger's Causality test. This test assesses the null hypothesis that the explanatory variable does not Granger-cause the response variable, indicating that the explanatory variable does not have a significant impact on the responsible variable. Table 22 provides a summary of the p-values obtained from these tests for each variable with the maximum number of lags at 3. In Table 22, the rows represent the responses (Y), while the columns correspond to the predictor series (X).

The findings from the Granger causality test indicate that there is a statistically significant influence of cluster 4 and GDP on cluster 1. The p-values associated with these influences are 0.017 and 0.006, leading to the rejection of the null hypothesis and indicating an impact on the response variable. In the case of Cluster 2, there is evidence of influence from nickel price and cluster 3 and 5 with associated p-values of 0.024, 0.009 and 0.012. In the case of cluster 3, ESTER and clusters 2, 4 and 5 can be said to have an impact on cluster 3 with p-values of 0.000 0.043, 0.014 and 0.006, respectively. Cluster 4 appears to be influenced by only external variable GDP as suggested by the p-value of 0.000, meanwhile, cluster 4 and nickel appear to have an impact on cluster 5 with the p-values of 0.029 and 0.033. Lastly, for cluster 6 all the explanatory variables have high p-values, indicating no impact on cluster 6 from any variable. Therefore, constructing a VAR model for the cluster 6 is not feasible since it requires multiple variables.

Table 22: Granger causality

y/x	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Nickel	GDP	ESTER
Cluster 1	-	0.322	0.716	**0.017**	0.370	0.276	0.308	**0.006**	0.103
Cluster 2	0.356	-	**0.009**	0.287	**0.012**	0.180	**0.024**	0.443	0.368
Cluster 3	0.467	**0.043**	-	**0.014**	**0.006**	0.024	0.642	0.687	**0.000**
Cluster 4	0.509	0.874	0.738	-	0.299	0.252	0.525	**0.000**	0.673
Cluster 5	0.428	0.166	0.399	**0.029**	-	0.248	**0.033**	0.181	0.458
Cluster 6	0.470	0.543	0.076	0.236	0.225	-	0.335	0.566	0.303

4.3.1. Cluster 1

Cluster 1 is influenced by Cluster 4 and GDP based on the Granger causality test; therefore, these variables are used in the VAR model. The lowest AIC value was found in the VAR(3) model at 18.78. However, the results in Appendix 4, show that at lag three (L3) both of the explanatory variables GDP and cluster 4 exhibit high p-values, indicating that they are not statistically significant and don’t impact cluster 1. Therefore, lag order 3 is changed to 2 and VAR(2) is constructed and examined further. The results of the VAR(2) model are presented in Table 23 below. The results indicate that cluster 1 is statistically significant at lag 1 (L1.C1) and lag 2 (L2.C1), cluster 4 at lag 1 (L1.C4) and GDP at lag 2 (L2.GDP) with a confidence level of 5%. Moreover, L2.C4 can be said to be statistically significant with a confidence level of 10% as its p-value is under 0.10.

The VAR(2) is used to plot predictions and compare them against the original values and judgemental forecasts in Figure 25. The model seems to fit the model well; however, some drops and peaks are underestimated. When predicting the last 3 values, the Figure 25 shows that the model correctly predicts a drop in the values. However, the model’s projected decline is not as pronounced as the actual values. From the visual representation, the model seems still to be a better fit than the judgemental values as the judgemental forecasts are predicted to be essentially lower than the actual values. This is supported by MAPE, where the MAPE of the VAR model is 4.41% and for the judgemental forecasts it stands at 7.08%.

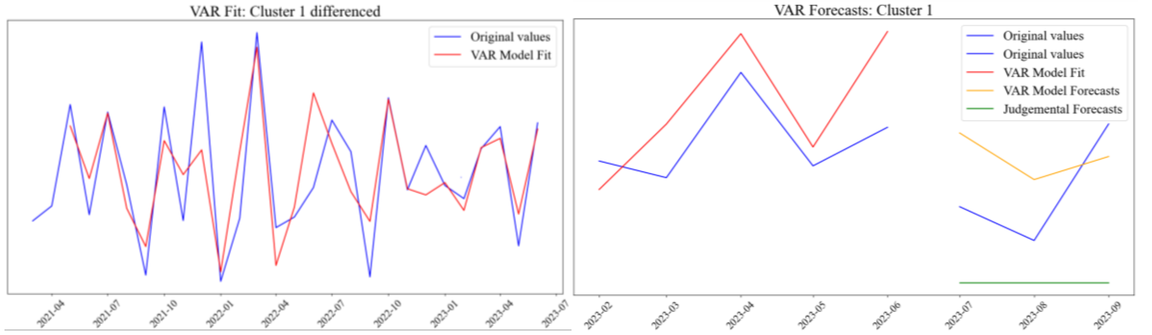


Figure 25: VAR(2) results for cluster 1

Table 23: VAR(2) results for cluster 1 equation with cluster 4 and GDP

	coefficient	Std. error	t-stat	prob
Const	-86.256	102.784	-0.839	0.401
L1.C1	-0.905	0.172	-5.252	**0.000**
L1.C4	0.145	0.068	2.126	**0.033**
L1.GDP	7.822	17.984	0.435	0.664
L2.C1	-0.517	0.168	-3.068	**0.002**
L2.C4	-0.093	0.048	-1.921	0.055
L2.GDP	47.182	23.359	2.020	**0.043**

4.3.2. Cluster 2

In the case of cluster 2, the Ganger causality test revealed that nickel price, cluster 3 and cluster 5 have an influence on cluster 2. The lowest AIC is found in the VAR(2) model of 46.36. The results of the VAR(2) model are presented in Appendix 4. The results show that none of the coefficients are statistically significant. Cluster 5 exhibits the highest p-values and, therefore, the model is tested again without this variable. The AIC of the VAR(2) with cluster 3 and nickel price is 37.11 which is the smaller AIC of the three VAR models. The results of the VAR(2) model with nickel price and cluster 3 are presented in Appendix 4 as well. From the result, we can see that both explanatory variables are statistically significant at lag 1 but none of the variables are statistically significant at lag 2. Therefore, VAR(1) is built for further analysis. The results of the VAR(1) model are presented in Table 24.

The model’s fitted values and predictions are still plotted in Figure 26. The Figure 26 shows that the model that the model seems to follow the actual values directions well, however, most of the drops and peaks are underestimated. When predicting the last three observations, the model seems to correctly predict a decrease. Moreover, the model predicts excessively lower values than the actual values. It is evident that the judgemental forecasts are a better fit based on the Figure 26. Furthermore, the MAPE for the model is 168.27% and 45.50% for the judgemental forecasts, indicating that the judgemental forecasts are better fit.

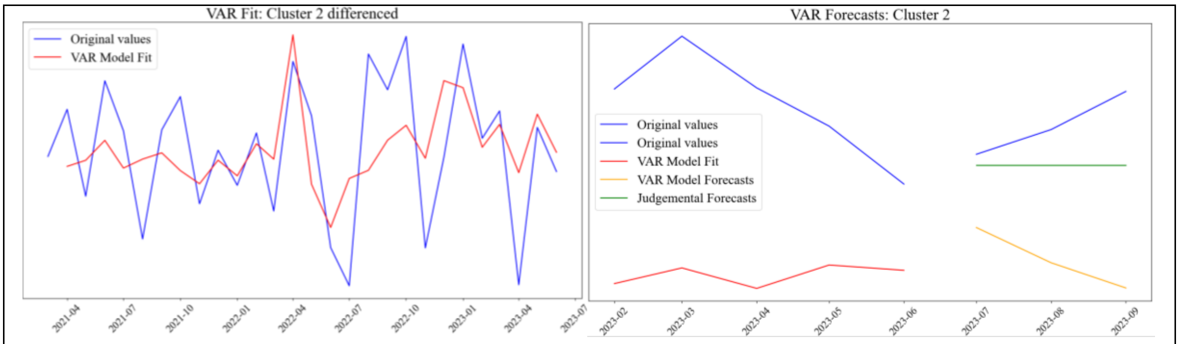


Figure 26: VAR(1) results for cluster 2

Table 24: VAR(1) results for cluster 2 equation with nickel price and cluster 3

	coefficient	Std. error	t-stat	prob
Const	37.839	42.896	0.882	0.378
L1.C2	-0.155	0.182	-0.853	0.394
L1.Nickel	0.031	0.012	2.545	**0.011**
L1.C3	-0.067	0.039	-1.715	0.086

4.3.3. Cluster 3

The Granger causality test found evidence of clusters 2, 4, 5 and ESTER having an influence on cluster 3. With these variables, the lowest AIC was again found in the VAR(3) model at 36.12. The results of the VAR(3) model, in Appendix 4, show that clusters 2, 4 and 5 are not statistically significant at any lag and, therefore, they are removed from the model. The lowest AIC with only ESTER as an explanatory variable was found again in the VAR(3) model at 8.11, which is considerably lower than the AIC of the VAR(3) model with 5 variables. The results of the model are presented in Table 25.

The model is fitted, and predictions are plotted in Figure 27. The model seems to fit the train values well to the actual data. When predicting the last 3 observations, it's clear that the model initially forecasts values higher than the actual ones, showing a slight decrease in the last 3 observations. The predictions closely match the actual values in the first two observations, but as the real values increase in the last observation and the model's predictions decrease slightly, the alignment is not good in the final observations. Still, the Figure 27 shows that the model seems to be a better fit than the judgemental forecast, and this is supported by MAPE as well. The MAPE for the VAR model is 79.66% while the judgemental forecast has a higher MAPE at 108.77%.

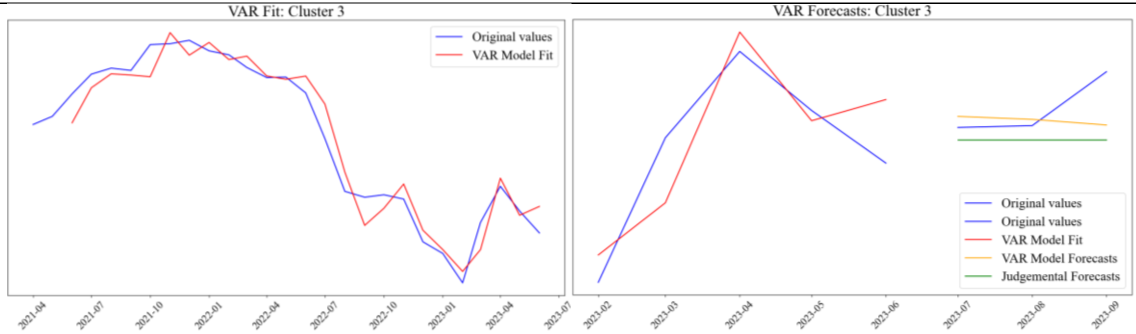


Figure 27: VAR(3) results for cluster 3

Table 25: VAR(3) for cluster 3 equation with ESTER

	coefficient	Std. error	t-stat	prob
Const	-0.962	65.900	-0.015	0.988
L1.C3	1.484	0.181	8.176	**0.000**
L1.ESTER	-322.804	239.280	-1.349	0.177
L2.C3	-0.843	0.304	-2.774	**0.006**
L2.ESTER	-805.553	341.801	-2.357	**0.018**
L3.C3	0.342	0.200	1.714	0.087
L3.ESTER	-878.345	247.330	-3.551	**0.000**

4.3.4. Cluster 4

The Granger causality test found only GDP having an influence on cluster 4. The lowest AIC of 9.62 is in the VAR(3) model. The results of the VAR(3) are in Appendix 4 and when examining the results of VAR(3), it is evident GDP has a high p-value at lag three (L3.GDP), and therefore VAR(2) model is used. The AIC of the VAR(2) model is 11.60. The model is presented in Table 26 below, where we can see that L1.C4, L1.GDP an L2.GDP are all statistically significant.

The model’s fitted values and the predictions are plotted in Figure 28. From the Figure 28, it seems that the model fits the values quite well as it follows the directions of the actual values. Figure 28 shows that the model predicts a small drop although the actual values exhibit a peak. The first and last predicted value seems to have a significant error term compared to the actual values. Nevertheless, the model still seems to be a more suitable fit than the judgemental forecasts, considering that the judgemental forecasts are consistently predicted too high. The MAPE for the VAR model is 27.83% and for judgemental forecasts, it is 65.20%.

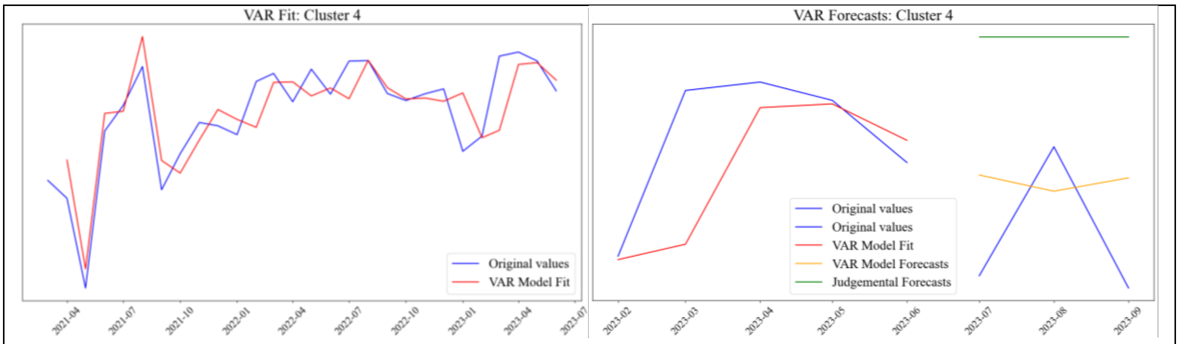


Figure 28: VAR(2) results for cluster 4

Table 26: VAR(2) for cluster 4 equation with GDP

	coefficient	Std. error	t-stat	prob
Const	484.420	253.937	1.908	0.056
L1.C4	0.775	0.167	4.647	**0.000**
L1.GDP	-171.588	46.339	-3.703	**0.000**
L2.C4	-0.074	0.127	-0.577	0.564
L2.GDP	221.667	57.940	3.826	**0.000**

4.3.5. Cluster 5

For cluster 5, The Ganger causality test found evidence of nickel price and cluster 4 having an influence. With these variables, the lowest AIC was found in the VAR(1) model at 37.36. The results of the VAR(1) model are presented in Appendix 4. The results show that none of the variables are statistically significant at a 5% threshold, however, cluster 4 is statistically significant at lag 1 (L1.C4) with a 10% threshold as the p-value is 0.083 and smaller than 0.10. Therefore, the model is tested again without the nickel price and the lowest

AIC of 21.94 is still found in VAR(1). The results of the VAR(1) model are presented below in Table 27 and now cluster 4 is statistically significant at lag 1.

The model fitted model to the differenced data is plotted in Figure 29. The model appears to follow the actual values, yet consistently underestimates both drops and peaks in the differenced data. The predicted values of the VAR(1) model are plotted against the original values as well in Figure 29. The Figure 29 shows that the model's predictions a clearly smaller than the actual values. Moreover, the actual values exhibit an increase while the model predicts a small decrease. From the Figure 29, it's evident that the judgemental forecasts seem to be a better fit than the model. The MAPE for the VAR model is 20.39% and for the judgemental forecast 14.77% indicating that the judgemental forecast outperforms the model.

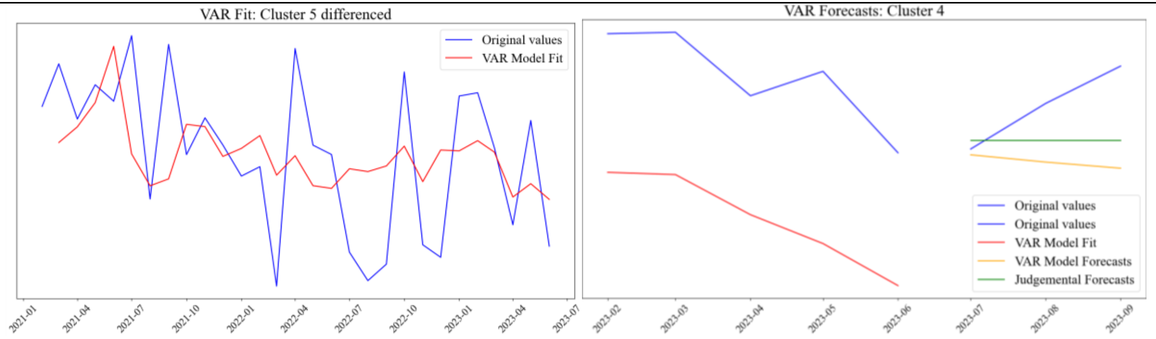


Figure 29: VAR(1) results for cluster 5

Table 27: VAR(1) for cluster 5 equation with cluster 4

	coefficient	Std. error	t-stat	prob
Const	588.705	1.885	1.885	0.059
L1.C5	0.571	0.638	0.894	0.371
L1.C4	0.625	0.199	3.139	**0.002**

4.3.1. Validation for VAR models

In most of the cases VAR(2) or VA(1) models were used, except for cluster 3 where VAR(3) was found as the most suitable model. VAR models incorporating GDP demonstrated the best performance, with the smallest MAPEs of 4.41% and 27.83% for clusters 1 and 4. Moreover, the models for clusters 1, 3 and 4 outperformed the judgemental forecasts. The poorest performance with MAPE of 168.27% was found in VAR model cluster 2 where the explanatory variable was cluster 3. In addition to cluster 2, cluster 5 couldn't outperform the judgemental forecast.

When changing the forecasting horizon to six months, the MAPEs increase in two clusters and decrease in three clusters. Cluster 1 decreases from 4.41% to 3.46%, Cluster 2 rises from 168.27% to 261.62%, Cluster 3 increases from 79.66% to 136.82%, Cluster 4 decreases from 27.83% to 16.62% and Cluster 5 drops from 20.39% to 13.60%. The decreased MAPES, indicates an improvement in the accuracy of the VAR models over a six-month horizon, therefore, suggesting that the models might perform better when making predictions over a longer horizon.

VAR models are complicated due to the temporal nature of time and therefore in the cross-validation, only new one-fold is tested. Implementing more folds in cross-validation with VAR models would demand a substantial amount of data and the data in this research is limited. The fold 2 that is tested contains 30 observations, where 27 are in training and 3 are in testing. The results of the cross-validation are presented in Table 28 below. From the Table 28, we can see that three clusters of MAPEs decreased substantially in fold 2. Cluster 1 decreased from 4.41% to 1.84%, Cluster 4 dropped from 27.83% to 4.51% and Cluster 5 decreased from 20.39% to 14.37%, indicating an improvement in the model's accuracy. On the other hand, the other two clusters, cluster 2 and 3, experienced a substantial increase in the MAPEs. Cluster 2 rises from 168.27% to 740.46% and cluster 3 increases from 76.66% to 305.80%.

Table 28: MAPEs of Cross-validation for VAR models

Cluster	Model	Fold 1	Fold 2
Cluster 1	VAR(2)	4.41%	1.84%
Cluster 2	VAR(1)	168.27%	740.46%
Cluster 3	VAR(3)	79.66%	305.80%
Cluster 4	VAR(2)	27.83%	4.51%
Cluster 5	VAR(1)	20.39%	14.37%

4.4. Random Forest Regression

For every price cluster, random forest regression models are constructed using external data. The monthly NRPs serve as the response variable, while the external data acts as explanatory variables. In addition, the NRPs from the previous month are included as an explanatory variable "NRP month before". As the first value January 2021 doesn't have a month before value, it is excluded from all datasets. Therefore, the datasets used in random forest regression comprise 32 observations, where 29 observations are in the training set and the last 3 in the test set. The optimal number of trees is examined by plotting the number of trees against the model's MAPEs for each cluster.

4.4.1. Cluster 1

For cluster 1, the number of trees plot clearly indicates that the lowest MAPE is found when using only one tree. Therefore, for cluster 1 random forest is built with one tree. The model's fit, as depicted in Figure 30, exhibits good performance at certain data points but struggles at others. For example, in the first few values, the model fits perfectly but in the tenth value when the actual values show an increase the model predicts a substantial decrease. Moreover, in the forecast plot, it is evident that the model struggles to predict the significant increase in the last value. When predicting the last three values, the model tends to overestimate the initial two values, predicting them to be higher than the actual values. Despite these discrepancies, the model's predictions appear more accurate compared to the judgemental forecast, as confirmed by the MAPE values as well. The model achieves a MAPE of 4.11% outperforming the judgemental forecasts with a MAPE of 7.08%.

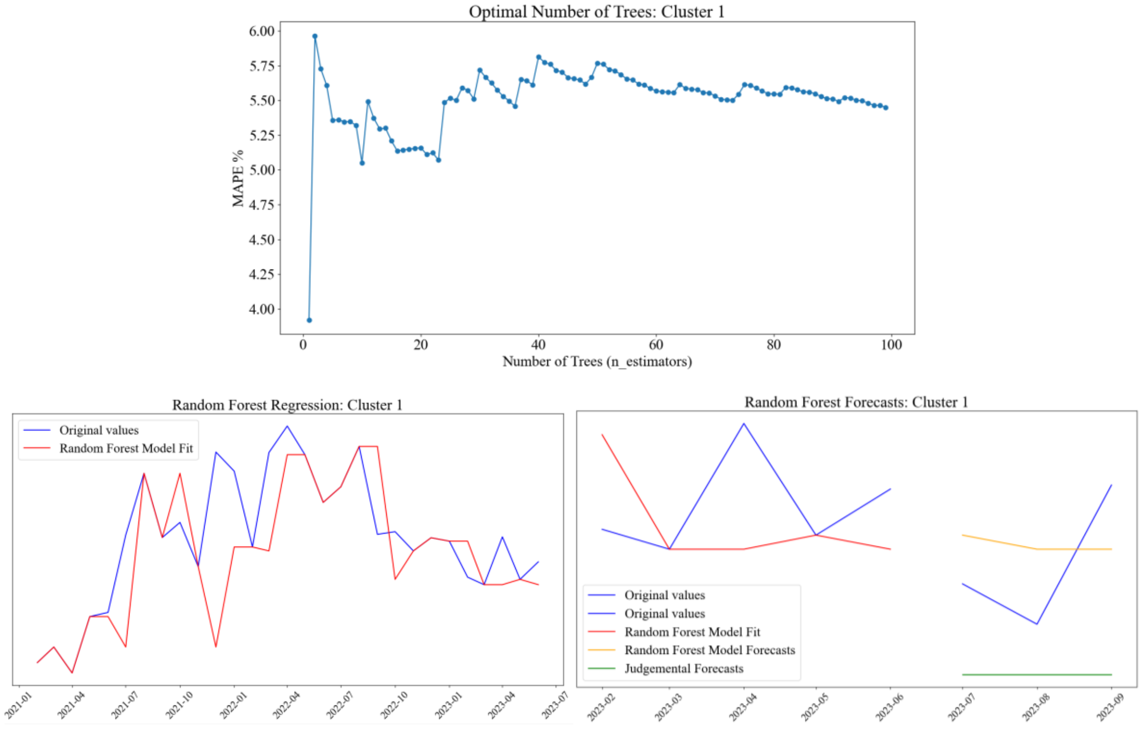


Figure 30: Random forest result for cluster 1

4.4.2. Cluster 2

For cluster 2, the number of trees plot clearly indicates that the lowest MAPE is found close to the end of 100 trees, more precisely at 98 trees. Therefore, a random forest is built with 98 trees. The model seems to fit the values well based on the Figure 31. However, when predicting the last 3 values the model struggles the predict the increase that happens in the actual values. Additionally, the model tends to underestimate the last two values, predicting them to be lower than the actual values. Still, the model’s predictions appear more accurate compared to the judgemental forecast, as they are even lower than the model’s forecast. This is supported by MAPEs, where the model exhibits a MAPE of 28.08% and the judgemental forecasts exhibit a MAPE of 45.50%.

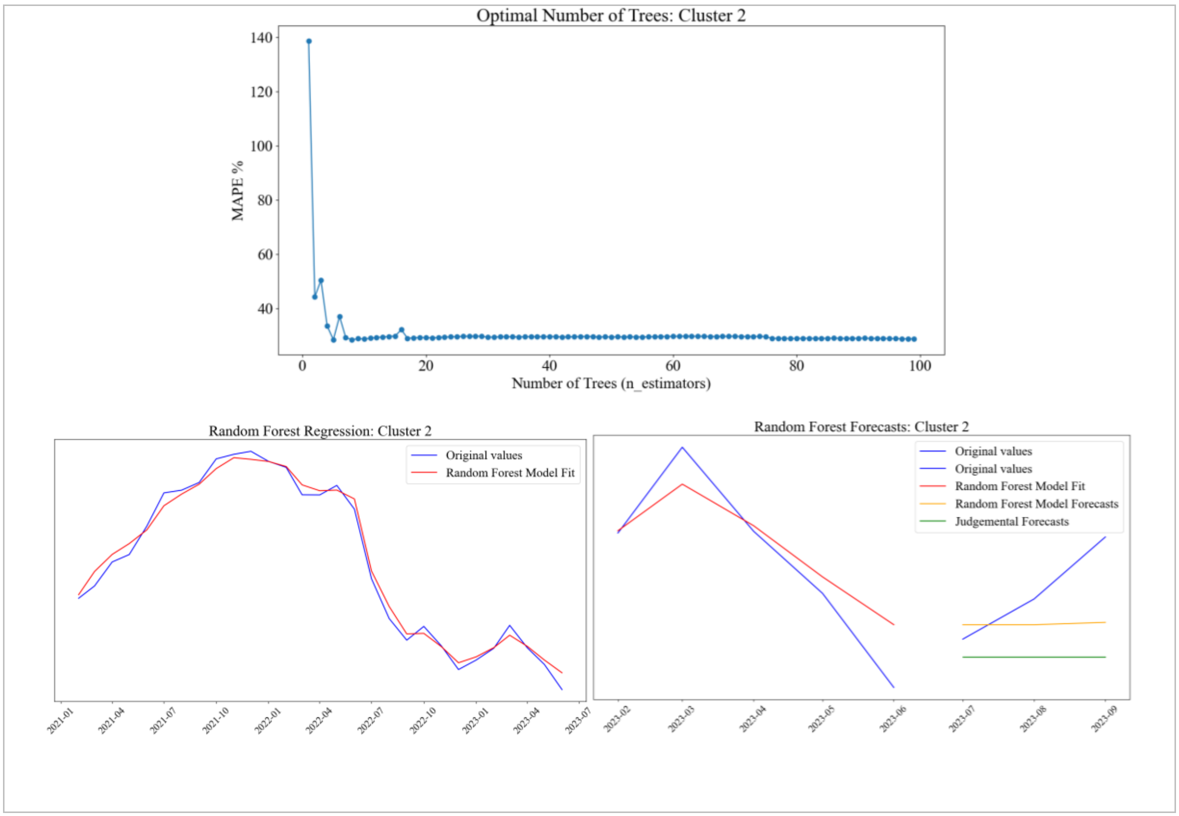


Figure 31: Random forest result for cluster 2

4.4.3. Cluster 3

For cluster 3, Figure 32 indicates that the lowest MAPE is found using two trees. Therefore, a random forest is built with two trees. The model seems to fit the model well; however, some drops and peaks are underestimated. When predicting the last 3 values the model seems to predict the first two values almost perfectly. However, the model struggles to predict the increase in the last value. The model's predictions appear to be a better fit than the judgemental forecast, as they are even lower than the model's forecast. Conclusion supported by MAPEs, where the model exhibits MAPE of 62.23% and the judgemental forecasts exhibit MAPE of 108.77%.

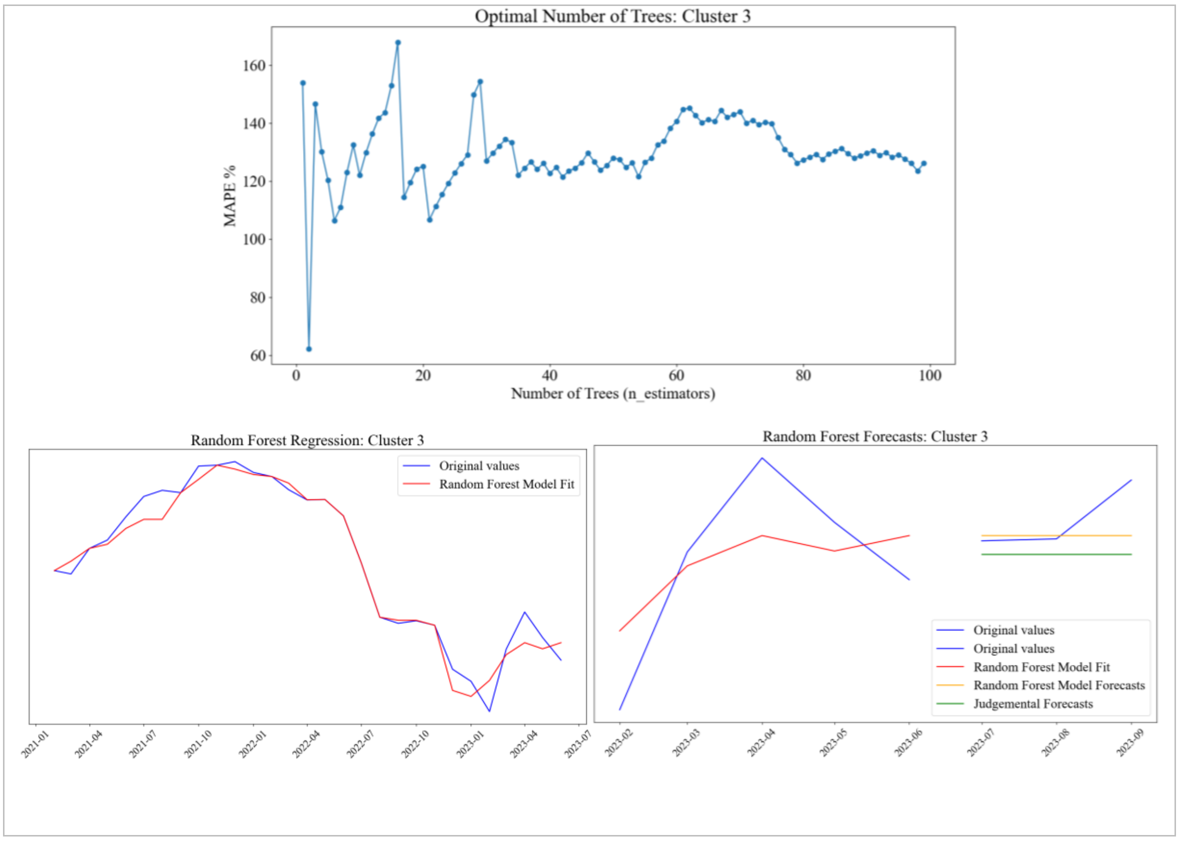


Figure 32: Random forest results for cluster 3

4.4.4. Cluster 4

For cluster 4, the number of trees plot indicates that the lowest MAPE is found using 8 trees based on the Figure 33, and therefore, a random forest is built with 8 trees. The model seems to struggle to fit accurately the substantial drops and peaks in the beginning well based on the Figure 33. However, its performance noticeably improves as we progress through the observations. In forecasting the last three values, the model struggles to anticipate the initial increase, and furthermore, it overestimates the first value compared to the actual observation. However, as the observed values start to decrease, the model also forecasts a decline, though not to the same extent as the actual values. The model’s predictions appear to be a better fit than the judgemental forecast, as they are even higher than the model’s forecast. Conclusion supported by MAPEs, where the model exhibits MAPE of 11.75% and the judgemental forecasts exhibit MAPE of 65.20%.

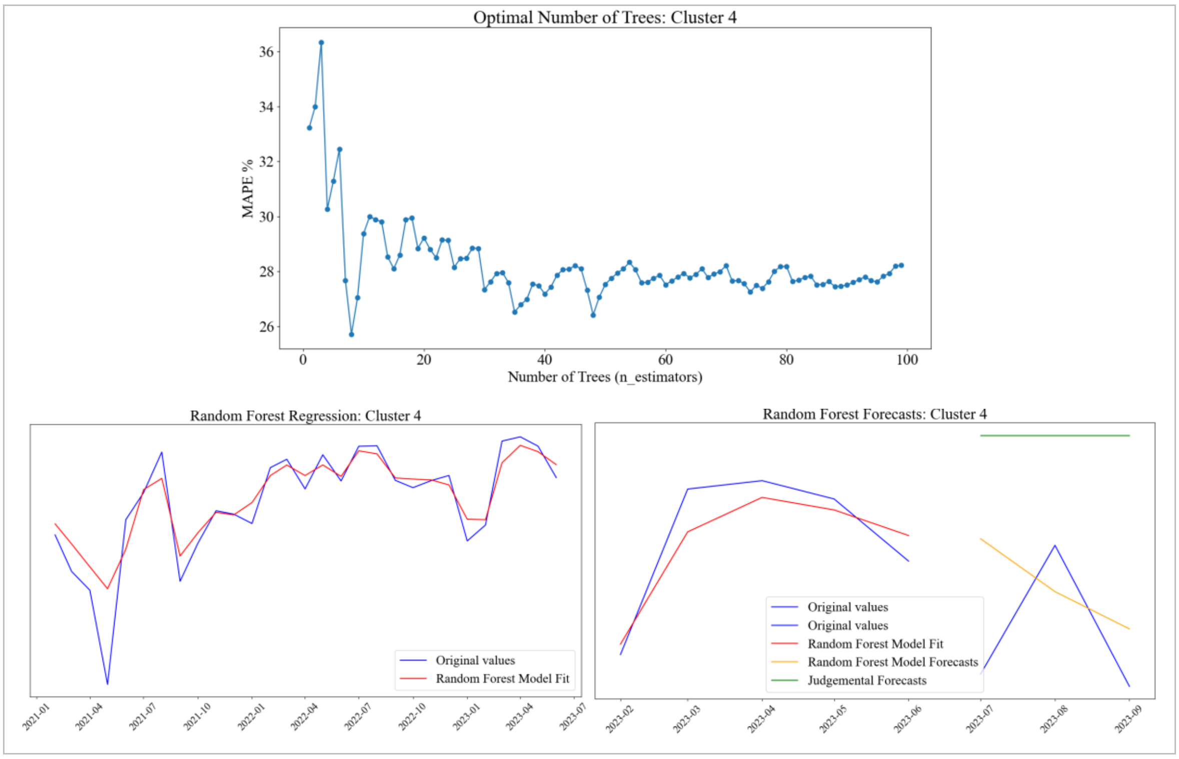


Figure 33: Random Forest results for cluster 4

4.4.5. Cluster 5

For cluster 5, the number of trees plot clearly indicates that the lowest MAPE is found using four trees based on Figure 34. Therefore, a random forest is built with four trees. The model seems to fit the model well; however, some drops and peaks are underestimated based on Figure 34. When predicting the last 3 values the model consistently predicts the same value throughout the entire forecasting periods, while the actual values show an increase. The model predicts the second value perfectly but overestimates the first value and underestimates the last value as the actual values show an increase. The model’s predictions appear to still be a better fit than the judgemental forecast. The model exhibits a MAPE of 11.75% and the judgemental forecasts exhibit a MAPE of 14.77%.

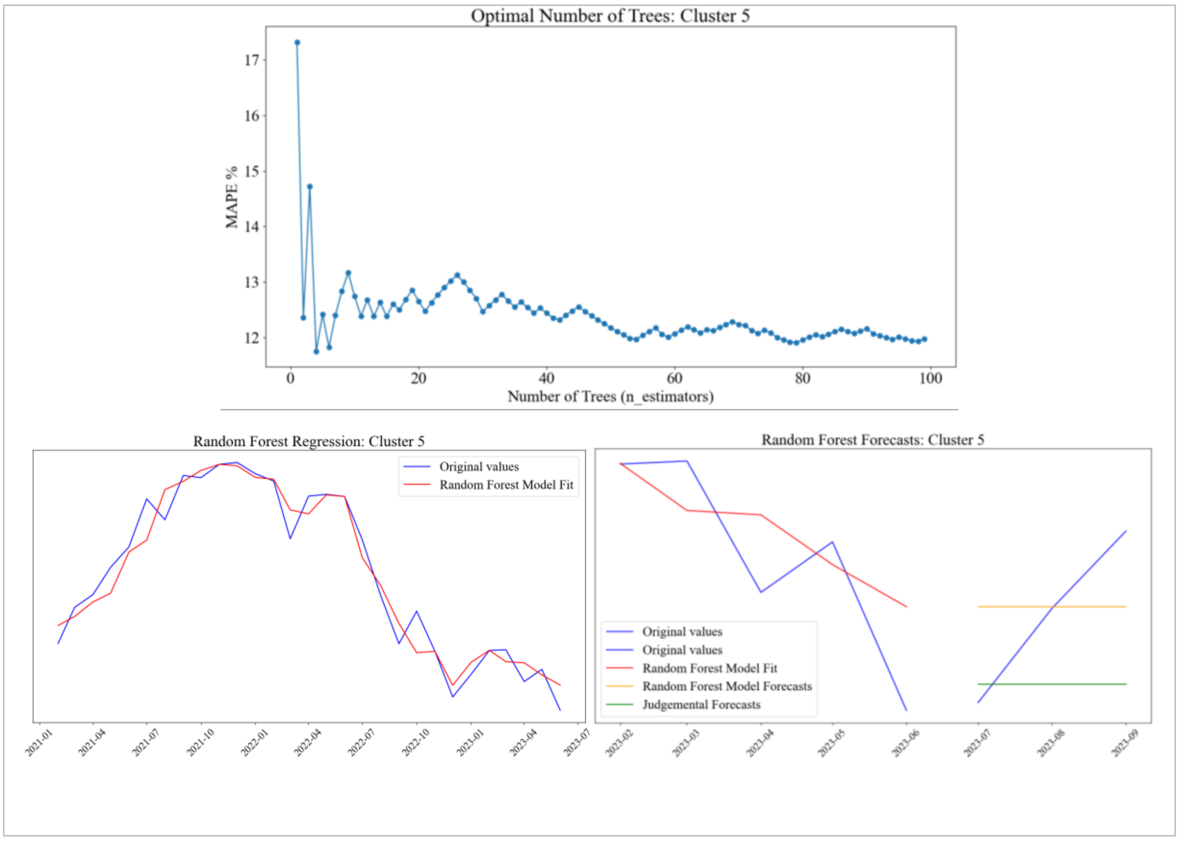


Figure 34: Random forest results for cluster 5

4.4.6. Cluster 6

For cluster 6, the number of trees plot clearly indicates that the lowest MAPE is found using 13 trees based on Figure 35. Therefore, a random forest is built with 13 trees. The model appears to follow the actual values, yet consistently underestimates both drops and peaks in the data based on Figure 35. In forecasting the last three values, the model foresees both an increase and a decrease. Specifically, for the first two values, the model anticipates an increase, contrary to the decrease in actual values. However, an alignment between the model's predictions and the actual values occurs at the second value. In the final value, the model accurately predicts a decrease, although the decrease is larger than what is observed in the actual values. The model exhibits a MAPE of 181.71% and the judgemental forecasts exhibit a MAPE of 458.72% indicating a better fit with the model.

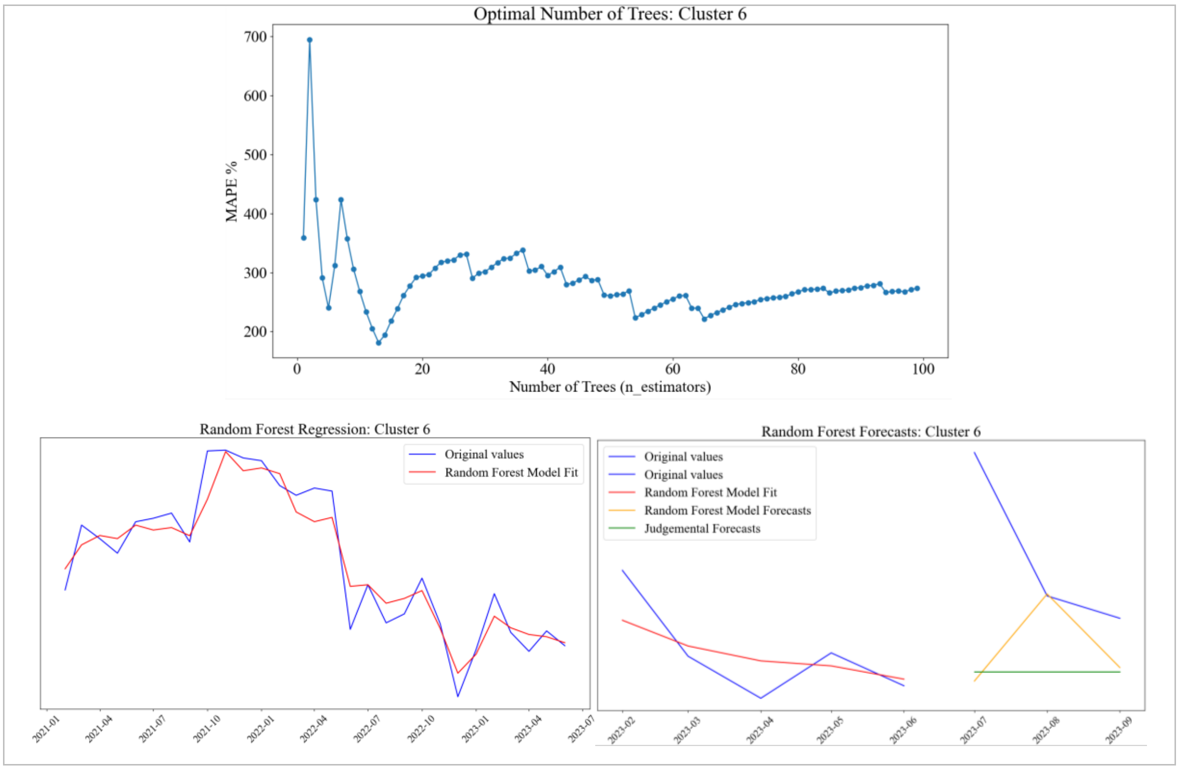


Figure 35: Random Forest results for cluster 6

Table 29 summarizes the importance of each variable in the random forest model built before. The importance represents how much adding a certain variable improves the accuracy of predictions. Table 29 reveals that, in cluster 1, the foremost influential variable is the NRP month before. In cluster 2, GDP emerges as the pivotal variable, while in cluster 3, ESTER is the most influential variable, with GDP closely following. Cluster 4 is characterized by the nickel price as the most critical variable. For cluster 5, the most important variable is the NRP month before, and for cluster 6, it is ESTER. Interestingly, the variable with the lowest significance is the nickel price in clusters 1, 2, 3, and 5. In cluster 4, the least influential variable is ESTER, and in cluster 6, it is GDP.

Table 29: Importance of each variable in the random forest regressions

Cluster	Nickel Price	GDP	ESTER	NRP month before
Cluster 1	0.068	0.138	0.079	0.714
Cluster 2	0.034	0.426	0.298	0.242
Cluster 3	0.043	0.398	0.493	0.066
Cluster 4	0.487	0.128	0.076	0.309
Cluster 5	0.026	0.255	0.056	0.663
Cluster 6	0.135	0.059	0.509	0.297

4.4.7. Validation for random forest regressions

After analysing the random forest regressions, it is interesting to notice that each model identified the optimal number of trees differently. Notably, all models utilized thirteen trees or fewer, except for Cluster 2, where the best-performing model was found to be composed of 98 trees. Overall, the random forest regressions performed well in over a half of the models. Four clusters exhibit MAPEs under 29% (clusters 1, 2, 4 and 5) while the MAPEs of the others (clusters 4 and 6) exhibited MAPEs of 62.23% and 181.71%. However, it is important to note that other models and judgemental forecasts struggled to predict these clusters as well. Moreover, the random forest models outperformed judgemental forecasts across all clusters.

For each cluster, we are examining the impact on the model's MAPEs when altering the forecasting horizon to six months. Across all clusters, the MAPEs exhibit an increase. Cluster 1 rises from 4.11% to 28.08%, Cluster 2 rises from 89.22% to 260.40%, Cluster 3 increases from 62.23% to 106.13%%, Cluster 4 sees an increase from 25.71 to 28.08%. Cluster 5 climbs from 11.75% to 34.36%, and Cluster 6 ascends from 181.71% to 251.11%. The increased MAPEs collectively suggest that the accuracy of the forecasting model is decreasing as you extend the prediction period.

Table 30 displays the outcomes for each cross-validation folds. Observing the Table 30, it is apparent that in the case of clusters 1 and 4, the MAPEs either decrease or exhibit marginal increases compared to the MAPE in the initial fold. This suggests that the model performs consistently well across diverse subsets of the dataset for these clusters. Conversely, for clusters 2 and 5, there is a substantial decrease in MAPE for one-fold, while the other two folds witness increases. This inconsistency suggests that the model might be sensitive to variations in certain subsets of the data, performing well in some cases but less effectively in others. In the case of cluster 3, the MAPEs increase compared to the first fold and for cluster 6 the MAPEs decrease across the folds, indicating improved performance or adaptability of the model to diverse data samples within Cluster 6. Again, it is important to note that the high variability between the folds can be resulted due to the small dataset.

Table 30: MAPEs of Cross-validation for random forest models

Cluster	Number of trees	Fold 1	Fold 2	Fold 3	Fold 4
Cluster 1	1	4.11%	1.63%	5.39%	8.47%
Cluster 2	98	28.08%	358.80%	7.32%%	62.30%
Cluster 3	2	62.23%	138.55%	139.72%	65.58%
Cluster 4	8	25.71%	20.19%	29.85%	11.71%
Cluster 5	4	11.75%	34.76%	7.14%	40.02%
Cluster 6	13	181.71%	162.55%	94.31%	99.18%

5. Result discussions and conclusions

In this final chapter, the results of the implemented forecasting models are analysed, and the insights derived from these models are discussed. Moreover, we address the research questions that guided our study and the main findings, explaining the details and nuances we discovered. Additionally, the limitations of this study are discussed, and possible future research directions are discussed.

Table 31 below presents the MAPE results from all used forecasting methods for a three-month time horizon. For cluster 1 the exponential smoothing model exhibited the best performance with a MAPE of 4.08%, closely followed by the random forest with a slightly higher MAPE of 4.11%. Moving on to cluster 2, the random forest showed the best performance with a MAPE of 28.08% followed by exponential smoothing with a close MAPE of 28.24%. For cluster 3, random forest outperformed other models' performance with the MAPE of 62.23%. For cluster 4, the ARIMA model stood out with the lowest MAPE of 23.68%. Cluster 5 shows MAPEs ranging from 11% to 34% for the models with the lowest 11.75% found in the random forest and followed by exponential smoothing model closely with a MAPE of 12.12%. Lastly, for cluster 6 the lowest MAPE is found in the SES model at 133.03%.

Table 31: MAPEs of models for three-month time horizon

	Judgemental Forecasts (Benchmark)	Exponential smoothing models	ARIMA	VAR	Random forest
Cluster 1	7.08%	**4.08%**	4.74%	4.41%	4.11%
Cluster 2	45.50%	28.24%	115.55%	168.27%	**28.08%**
Cluster 3	108.77%	80.00%	114.51%	79.66%	**62.23%**
Cluster 4	65.20%	26.62%	**23.68%**	27.83%	25.71%
Cluster 5	14.77%	12.12%	17.09%	33.80%	**11.75%**
Cluster 6	458.72%	**133.03%**	543.33%	-	181.71%

As discussed in the second chapter, reservations and uncertainties surround the use of MAPE. Therefore, the Symmetrical Mean of Absolute Percentage (SMAPE) is computed for each cluster and model over a three-month time horizon, with the results presented in Table 32. Overall, the SMAPE outcomes remain consistent with the MAPE; random forest attains the lowest SMAPEs in three clusters, exponential in two, and ARIMA in one. However, variations emerge when comparing the performance of each model for the cluster. Despite random forest achieving the lowest MAPE in cluster 2, upon considering SMAPE, exponential smoothing emerges as the optimal model. Additionally, changes are noted in cluster 6, where, despite exponential smoothing initially exhibiting the lowest MAPE, the lowest SMAPE is now observed in random forest.

Table 32: SMAPEs of the models for three-month time horizon

	Judgemental Forecasts (Benchmark)	Exponential smoothing models	ARIMA	VAR	Random forest
Cluster 1	7.41%	**4.07%**	4.60%	4.32%	4.08%
Cluster 2	62.36%	**27.88%**	179.64%	200.0%	33.96%
Cluster 3	98.86%	84.33%	109.94%	102.32%	**80.69%**
Cluster 4	47.34%	22.95%	**21.49%**	23.92%	22.44%
Cluster 5	16.36%	11.55%	19.55%	43.01%	**11.29%**
Cluster 6	177.89%	139.77%	176.11%	-	**116.49%**

In conclusion from both Tables 31 and 32, across all clusters, at least two quantitative models consequently outperform judgemental forecasts. For cluster 1, all models performed better than the judgemental forecasts. Cluster 2 shows only exponential smoothing model and random forest outperforming judgemental forecasts. For cluster 3, random forest, VAR and exponential smoothing models performed better than the judgemental forecasts based on MAPEs and based on SMAPEs only random forest and exponential smoothing outperformed judgemental forecasts. For cluster 4, all models outperformed judgemental forecasts. For both, clusters 5 and 6, exponential smoothing and random forest models could outperform the judgemental based on MAPEs, but based on SMAPEs, ARIMA could outperform the judgemental forecast as well with exponential smoothing and random forest in cluster 6.

For each model, we also changed the forecasting time horizon to six month and performed cross validation to see how the model's performance was affected. Changing the forecasting time horizon from three months to six months resulted in widespread rise in MAPEs across various clusters in all models, indicating an overall decline in forecasting model accuracy as we extend the prediction period. Moreover, the cross-validation outcomes for all models revealed that the MAPEs fluctuate a lot across the folds within nearly every cluster. These fluctuations in MAPE values imply potential variations in the models' effectiveness, underscoring the importance of the starting point for forecasting. Nevertheless, it is crucial to bear in mind that the dataset is small which poses challenges in the cross-validation process. The high variability between the folds can be resulted due to the dataset being small.

5.1. Answering the objectives

The first objective was to find out find out can quantitative methods outperform the benchmark model. The judgemental forecasts were used as a benchmark in this study. The exponential smoothing model and random forest regression performed better than the judgemental forecasts in all clusters, the VAR model in three cases out of five and the ARIMA models outperformed judgemental forecasts in two clusters out of six clusters. Therefore, we can say that quantitative methods can outperform the benchmark model. The SMAPE values in table 32 of the exponential smoothing and random forest are close to half of the benchmark (expect cluster 5). Therefore, it is evident that exponential smoothing and random forest have potential to outperform judgemental forecasts. Although the results are rather convincing, the dataset is limited, and the reliability of the models needs to be verified. Judgmental forecasts for an extended historical period are unavailable, limiting our ability to compare them. As a result, we can only assess them in relation to this specific dataset and specific test set. Not all methods outperform the benchmark in every case which indicates that the methods should be used to improve judgemental forecasts, not replace it.

The second objective of this study was to find if there is a difference in performance of the quantitative models used. For each cluster MAPEs and SMAPEs were calculated, and random forest models were able to outperform the other models in three clusters based on MAPEs and SMAPEs. In the remaining three clusters random forest was second best. Out of these methods random forest is the most performant. However, it does not outperform other clearly and consistently. The differences in MAPEs between exponential smoothing model and random forest are minimal for example in cluster 1 and 2. Therefore, more validation and larger data sets are required to establish the best method among these.

The third and final objective of this study was to find out if there is a difference in performance between the multivariate models used and the univariate models. Here the results are somewhat mixed. Based on the MAPEs, random forest was able to outperform both univariate models exponential smoothing and ARIMA models in three clusters (clusters 2, 3 and 5) out of six and in the other three clusters (clusters 1, 4 and 6), it was able to outperform one univariate model. VAR models were able to outperform two univariate models in one cluster (cluster 3) and one univariate model in another cluster (cluster 1). Exponential smoothing model was able outperform both multivariate models in two clusters (clusters 1 and 6) and in three clusters (clusters 2, 4 and 5) it was able to outperform one multivariate model VAR. ARIMA models were able to outperform both two multivariate models in one cluster (cluster 4) and one multivariate model in 2 clusters (clusters 2 and 5). Based on SMAPEs in table 32, random forest regression was able to outperform at least one univariate model in all clusters and exponential smoothing model was able outperform both multivariate models in five clusters. ARIMA models were able to outperform at least one multivariate model in three clusters. VAR outperformed one univariate model in two clusters, but it was also outperformed by at least one univariate model in all cases. Therefore, multivariable models do not always outperform models, but random forest seems to be mostly better than the univariable models.

Table 33: Answers to the research questions

Research question	Answers
Which of the exponential smoothing, ARIMA models, VAR models or rand forest regression can outperform the used judgemental forecasts?	Across all six clusters, two quantitative models exponential smoothing and random forest regression outperformed judgemental forecasts.
Is there a difference in performance of the exponential smoothing, ARIMA model, VAR model or random forest regression?	The random forest regression demonstrated highest performance compared to other models in three out of six clusters and performed second bets in the remaining three clusters. Exponential smoothing model exhibited highest performance compared to other models in two clusters and ARIMA model in one cluster.
Is there a difference in performance of the exponential smoothing, ARIMA model, VAR model or random forest regression?	Across all clusters, random forest regression outperformed at least one univariate model, while VAR was outperformed by at least one univariate model in all six cluster.

As the random forest regression with external data emerged as the model with most accurate results for the company's dataset. It is important to note that incorporating additional variables and external data presents challenges due to the need to obtain values monthly. However, the three external variables used in this random forest regression are easily available as public information with public information, and forecasts for these variables are also easily obtainable from public sources. Therefore, it is advantageous to use additional variables in this case.

5.2. Research limitations

The study has several limitations which highlight the need for additional research. First, the research is constrained by the limited dataset, and this limitation is notably evident when considering the accuracy as well. The dataset consists of 33 observations and is divided into training and test sets, with the training test containing the first 30 observations and the test's last 3 observations. With a larger and more diverse dataset, the models can better capture the underlying patterns in the data, thereby improving their ability to make accurate predictions.

Second, the models used in this research are not without limitations. In the case of the exponential smoothing model, it relies on weighted averages of past observations which can make it struggle to capture and predict complex patterns or sudden shifts in the data. The ARIMA models require careful selection of model parameters with the risk of inaccurate forecasts if inappropriate values are chosen for the parameters. While VAR models can handle multiple time series variables, the complexity increases with the number of variables, and estimating parameters becomes more challenging. Random forest regression, in the other hand, has potential to overfitting and lack of understandability in complex models.

Third, when selecting the best model order for each method only one test set was used. As the dataset was small, cross-validation was not performed in the model selection. This means that we cannot say that the models used for each method in this study are always the best. Cross-validation with a small dataset can lead to instability in model evaluation, as the limited data may result in high variability between folds, and it can lead to challenges to obtain reliable performance estimates. However, cross-validation was performed for each model chosen. The cross-validation relieved that the performances varied between folds, which confirms the fact that time series cross-validation for small data set is difficult.

Fourth, the data set is from an era that was characterised by Covid-19 pandemic which has a major impact on all business transactions. It is likely that the normal cyclic or seasonal variations are hidden by changes due to Covid-19 pandemic. Finally, it is important to acknowledge that this research has employed a limited number of methods, suggesting that potentially valuable approaches may have been overlooked. While this study’s findings may not encompass the full spectrum of available methodologies, future research can broaden the scope for a more comprehensive understanding.

5.3. Future research

When considering potential future research perspectives, we have already highlighted that expanding the research to include new model families which were not tested in this research, could provide interesting insights. In this research, more complex models were left out due to the limited number of observations in the datasets. For example, more complex methods such as neural network methods and fuzzy set theory could be included in future research.

Moreover, the development of hybrid models could be a valuable direction for future exploration. Hybrid models combine the strengths of traditional time series methods with the flexibility and learning capabilities of more advanced approaches.

An alternative angle for future research could be changing the external variables or adding variables from internal data to contribute to a more comprehensive understanding of the factors influencing the predictions. Additionally, adding different measurement methods in terms of accuracy could bring more insights. Unlike this research which evaluated the measurement accuracy from the prediction error direction, future research may benefit from alternative accuracy measurements.

References

Adli, K. (2020). Forecasting steel prices using ARIMAX model: A case study of Turkey. *The International Journal of Business Management and Technology.*

Adli, K.A., & Sener, U. (2021). Forecasting of the US Steel Prices with LVAR and VEC Models. *Business and Economics Research Journal*, *12*(3), 509-522.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Boston (MA): Kluwer Academic.

Armstrong, J. S., Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecast*ing, 08, 69–80.

Azadeh, Neshat, N., Mardan, E., & Saberi, M. (2013). Optimization of steel demand forecasting with complex and uncertain economic inputs by an integrated neural network–fuzzy mathematical programming approach. *International Journal of Advanced Manufacturing Technology*, 65(5-8), 833–841.

Box, G. (1991). Understanding Exponential Smoothing: A Simple Way to Forecast Sale and Inventory. *Quality Engineering*, 4(1), 143-51

Box, G. E. P. (2016). *Time series analysis: forecasting and contro*l. Fifth edition. Hoboken, New Jersey: Wiley.

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123-140.

Brooks, C. (2014). Introductory econometrics for finance. 3rd ed. Cambridge: Cambridge University Press.

Brown, R.G. (1956). Exponential Smoothing for Predicting Demand. Cambridge, Massachusetts: Arthur D. Little InC.

Chen, D., Clements, K. W., Roberts, E. J., & Weber, E. J. (1991). Forecasting steel demand in China. *Resource Policy*, 173,196–210

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Computer Science*, 7, 623–e623.

Chou, M.T, CC, S., & Yang, Y.L. (2012). Review of Economics & Finance A Study of the Dynamic Relationship between Crude Oil Price. *Better Advances Press, Canada in Its Journal Review of Economics & Finance*, *2*(May), 30–42.

Deng, A. (2023). Time series cross validation: A theoretical result and finite sample performance. *Economics Letters*, 233, 111369-.

European Central Bank. (2023). *Overview of the euro short-term rate (€str), European Central Bank.* [Accessed: 30 October 2023]. Available at: https://www.ecb.europa.eu/stats/financial_markets_and_interest_rates/euro_short-term_rate/html/eurostr_overview.en.html#calc

Havery, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & Reasoning*, 13(1), 5-24.

Ho, T. K. (1995). Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Monteral, QC, 278-282

Hogarth, R. M, & Makridakis, S. (1981). Forecasting and planning: An evaluation. *Management Science*, 27(2), 115-138.

Holt, C. C. (2004). Forecasting seasonals and trend by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1), 5-10.

Igarashi, Y., Kakiuchi, E., Daigo, I., Matsuno, Y., & Adachi, Y. (2008). Estimation of steel consumption and obsolete scrap generation in Japan and Asian countries in the future. *ISIJ international*, 48(5), 696-704.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. 1st ed. Cham: Springer International Publishing

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103(3), 582-91; discussion 592-6.

Koehler A., Snyder, R., & Ord. K. (2001). Forecasting Models and Prediction Intervals for the Multiplicative Holt–Winters Method. *International journal of forecasting*, 17, 269–286.

Lawrence, M., & O'Connor, M. (1992). Exploring judgemental forecasting. *International journal of forecasting*, 8(1), 15-26.

Lin, C. B., Su, S. F., & Hsu, Y. T. (2001). High-Precision Forecast Using Grey Models. International Journal of Systems Science, 32 (5), 609-19.

Makridakis, S. & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International journal of forecasting*, 16 (4), 451–476.

Makridakis, S. G. & Wheelwright, S. C. (1989). *Forecasting methods for management*. 5th ed. New York (NY): Wiley & Sons.

Mehmanpazir, F., Khalili-Damghani, K. and Hafezalkotob, A. (2019). Modeling steel supply and demand functions using logarithmic multiple regression analysis (case study: Steel industry in Iran). *Resources Policy*, 63, 101409.

Mancke, R. (1968). The Determinants of Steel Prices in the US .: 1947-65. *The Journal of Industrial Economics, 16*(2), 147–160.

Mentzer, J. T. & Moon, M. A. (2005). *Sales Forecasting Management – A Demand Management Approach*. Thousand Oaks, CA, Sage.

Outokumpu. (2022). Annual report 2022. Outokumpu Oyj. [Accessed 2 October 2023]. Available at: Raportit ja esitykset 2022 | Outokumpu.

Outokumpu. 2023. Business area Europe. [Accessed 2 October 2023]. Available at: Business Area Europe | Outokumpu.

Ozdemir, Buluş, K., & Zor, K. (2022). Medium- to long-term nickel price forecasting using LSTM and GRU networks. *Resources Policy, 78*, 102906–.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. (2022). Forecasting: Theory and Practice. *International Journal of Forecasting*, 38(3), 705–871.

Rogers, R. (1987). Unobservable transactions price and the measurement of a supply and demand model for the American steel industry. *J Bus & Econ Stat*, 5(3):407–416.

Sanders, N.R. (2017). *Forecasting fundamentals*. First edition. New York, New York: Business Expert Press.

Song, H. & Li, G. (2008). Tourism demand modelling and forecasting—A review of recent research. *Tourism management*, 29 (2), 203–220.

Snyder, R.D., Koehler, A.B. and Ord, J.K. (2002). Forecasting for Inventory Control with Exponential Smoothing. *International Journal of Forecasting*, 18(1), 5-18.

Thomakos, Wood, G., Ioakimidis, M., & Papagiannakis, G. (2023). ShoTS Forecasting: Short Time Series Forecasting for Management Research. *Biritish Journal of Management*, 34(2), 539-554.

Waller, M.A. & Fawcett, S.E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. *Journal of business logistics*, 34 (2), 77-84.

Winters, P.R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3). 324-342.

Wu, B., & Zhu, Q. (2012). Week-ahead price forecasting for steel market based on RBF NN and ASW. *ICSESS 2012 - Proceedings of 2012 IEEE 3rd International Conference on Software Engineering and Service Science, 3*, 729–732

Yokuma, J. T. & Armstrong, J. S. (1995). Beyond accuracy: Comparison of criteria used to select forecasting methods. *International journal of forecasting.* 11(4), 591-597

Yu, G. & Schwartz, Z. (2006). Forecasting Short Time-Series Tourism Demand with Artificial Intelligence Models. *Journal of travel research*. 45 (2), 194–203.

Appendix 1. Figures of differenced variables

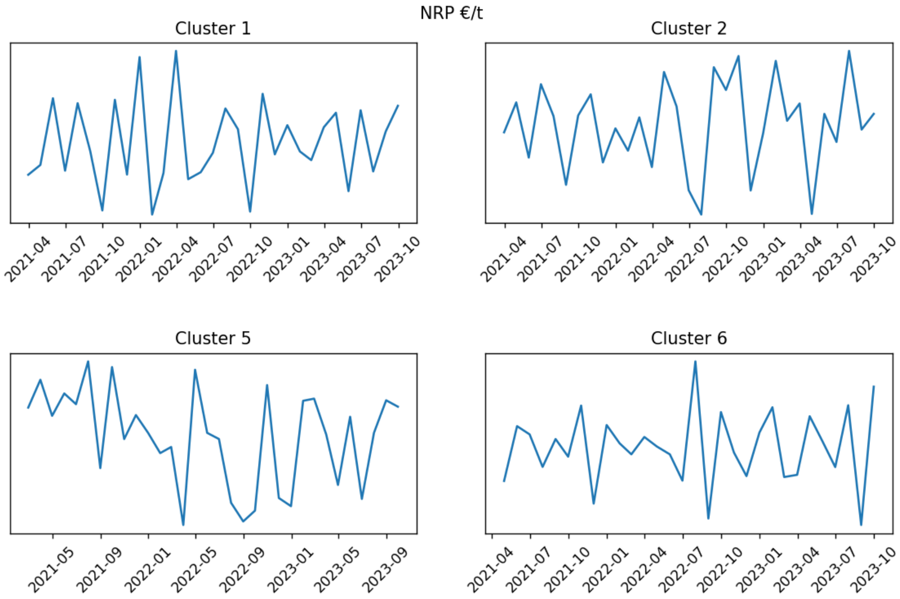


Figure 1: Differenced NRPs

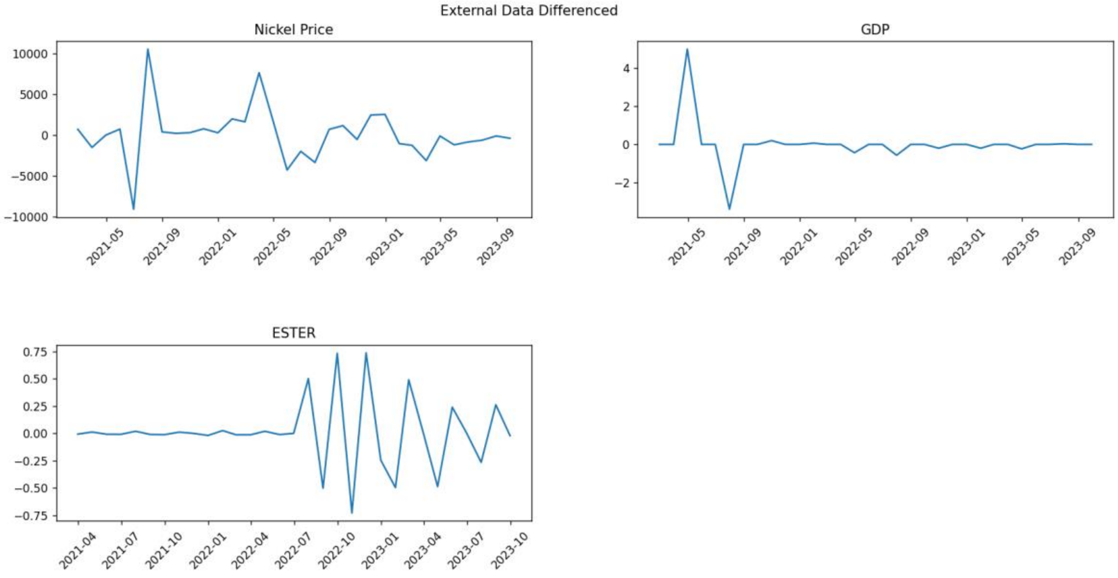


Figure 2: External data differenc

Appendix 2. Figures of NRPs and External data

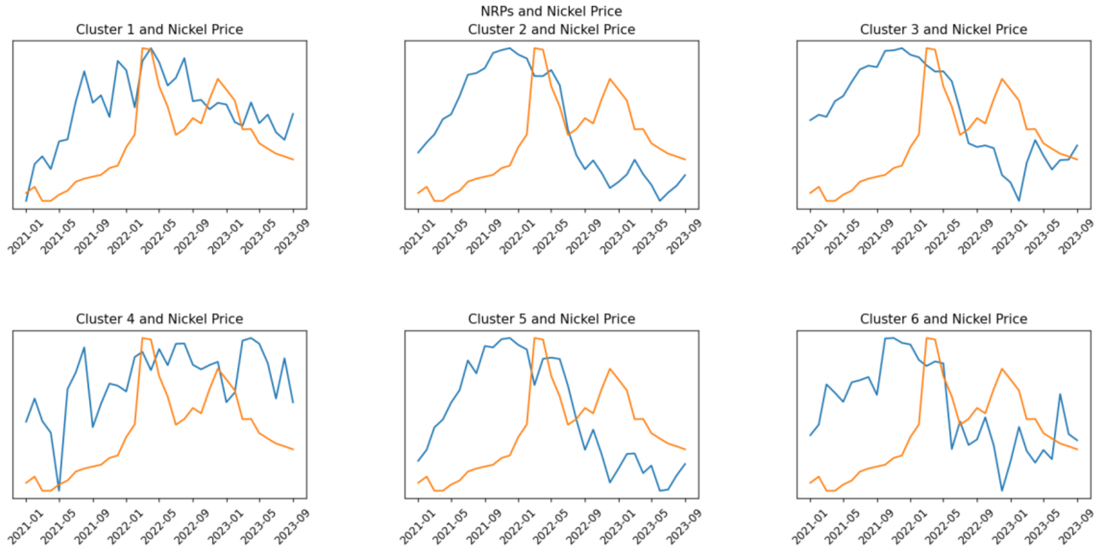


Figure 1: Clusters and Nickel Price

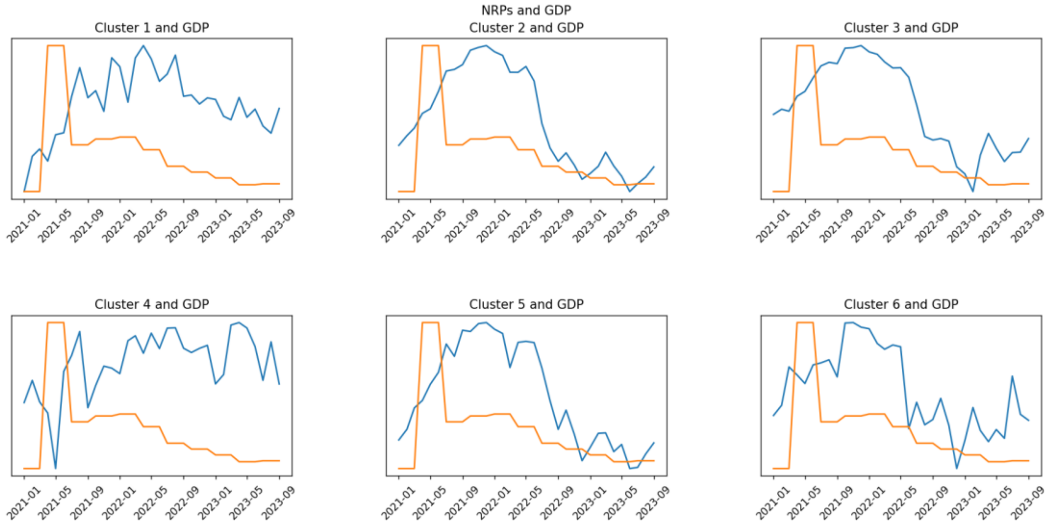
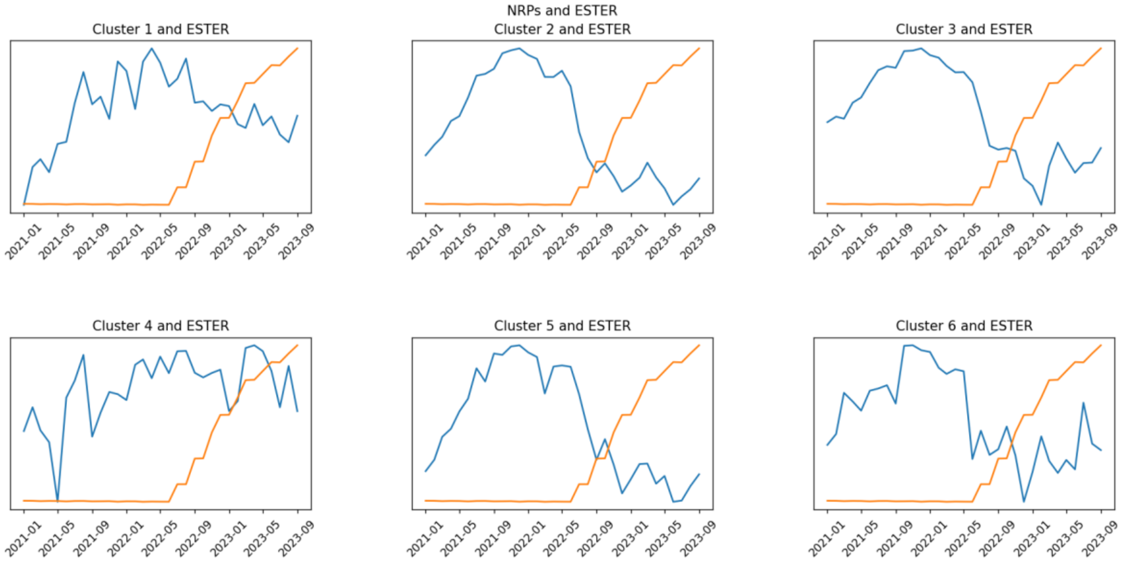


Figure 2: Clusters and Eurozone GDP



NRPs and ESTER

Cluster 1 and ESTER

Cluster 2 and ESTER

Cluster 3 and ESTER

Cluster 4 and ESTER

Cluster 5 and ESTER

Cluster 6 and ESTER

Figure 3: Clusters and ESTER

Appendix 3. Correlations

Table 1: Correlation matrix

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Nickle Price	ESTER	GDP
Cluster1	1.000	0.406	0.333	0.525	0.498	0.265	0.604	-0.180	-0.030
Cluster2	0.406	1.000	0.967	-0.112	0.974	0.857	-0.129	-0.806	0.512
Cluster3	0.333	0.967	1.000	-0.131	0.926	0.843	-0.226	-0.803	0.517
Cluster4	0.525	-0.112	-0.131	1.000	-0.058	-0.212	0.540	0.261	-0.445
Cluster5	0.498	0.974	0.926	-0.058	1.000	0.809	-0.073	-0.796	0.485
Cluster6	0.265	0.857	0.843	-0.212	0.809	1.000	-0222	-0.642	0.454
Nickel	0.604	-0.129	-0.226	0.540	-0.073	-0.222	1.000	0.186	-0.291
ESTER	-0.179	-0.806	-0.803	0.261	-0.796	-0.642	0.186	1.000	-0.516
GDP	-0.039	-0.512	0.512	-0.445	0.485	0.454	-0.291	-0.516	1.00

Appendix 4. VAR results

Cluster 1

Table 1: VAR(3) for cluster 1 equation with cluster 4 and GDP

	coefficient	Std. error	t-stat	prob
Const	-78.543	109.910	-0.715	0.475
L1.C1	-1.224	0.205	-5.987	**0.000**
L1.C4	0.177	0.082	2.157	**0.031**
L1.GDP	-4598	37.584	-0.122	0.903
L2.C1	-0.909	0.254	-3.573	**0.000**
L2.C4	-0.065	0.099	-0.659	0.510
L2.GDP	47.546	22.192	2.143	**0.032**
L3.C1	-0.470	0.188	-2.506	**0.012**
L3.C4	-0.071	0.065	-1.106	0.269
L3.GDP	-6.568	30.996	-0.212	0.832

Cluster 2

Table 3: VAR(2) for cluster 2 equation with nickel price, cluster 3 and 5

	coefficient	Std. error	t-stat	prob
Const	9.806	45.285	0.217	0.829
L1.C2	-0.237	0.217	-1.093	0.274
L1.Nickel	0.022	0.014	1-584	0.113
L1.C3	-0.225	0.141	-1.598	0.110
L1.C5	-0.142	0.268	-0.531	0.596
L2.C2	-0.171	0.198	-0.865	0.387
L2.Nickel	0.016	0.015	1.084	0.278
L2.C3	0.173	0.137	1.260	0.208
L2.C5	-0.046	0.281	-0.162	0.871

VAR(2) results for cluster 2 equation with nickel price and cluster 3

	coefficient	Std. error	t-stat	prob
Const	17.064	40.497	0.421	0.673
L1.C2	-0.263	0.191	-1.377	0.168
L1.Nickel	0.026	0.012	2.256	**0.024**
L1.C3	-0.257	0.111	-2.322	**0.020**
L2.C1	-0.194	0.180	-1.074	0.283
L2.Nickel	0.014	0.013	1.138	0.255
L2.C3	0.197	0.118	1.670	0.095

Cluster 3

Table 5: VAR(3) for cluster 3 equation with cluster 2, 4, 5 and ESTER

	coefficient	Std. error	t-stat	prob
Const	448.985	328.047	1.369	0.171
L1.C3	1.287	0.3145	4.092	**0.000**
L1.C2	0.579	0.788	0.735	0.462
L1.C4	-0.188	0.161	-1.172	0.241
L1.C5	-0.349	0.637	-0.547	0.584
L1.ESTER	-138.577	349.763	-0.396	0.692
L2.C3	-0.492	0.536	-0.918	0.359
L2.C2	0.184	0.473	0.388	0.698
L2.C4	0.035	0.177	0.199	0.843
L2.C5	0.105	0.578	0.182	0.855
L2.ESTER	-636.693	519.378	-1.226	0.220
L3.C3	0.131	0.339	0.387	0.699
L3.C2	-0.376	0.480	-0.783	0.434
L3.C4	-0.105	0.170	-0.618	0.537
L3.C5	0.471	0.699	0.674	0.500
L3.ESTER	-840.689	378.740	-2.220	**0.025**

Cluster 4

Table 8: VAR(3) for cluster 4 with GDP

	coefficient	Std. error	t-stat	prob
Const	500.497	263.145	1.902	0.057
L1.C4	0.720	0.204	3.522	**0.000**
L1.GDP	-2323.084	46.984	-4.748	**0.000**
L2.C4	-0.299	0.220	-1.359	0.174
L2.GDP	203.572	56.029	3.633	**0.000**
L3.C4	0.287	0.130	2.198	**0.028**
L3.GDP	-34.033	69.175	-0.492	0.623