Multivariate Time Series Forecasting of Earnings Before Interests and Taxes

Implementing a VARMA model for the corrugated production business at Stora Enso

Master's thesis in Complex Adaptive Systems

FRIDA MOBERG

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Multivariate Time Series Forecasting of Earnings Before Interests and Taxes

Implementing a VARMA model for the corrugated production business at Stora Enso

FRIDA MOBERG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
Gothenburg, Sweden 2022

Multivariate Time Series Forecasting of Earnings Before Interests and Taxes
Implementing a VARMA model for the corrugated production business at Stora Enso
FRIDA MOBERG

Supervisor: Elin Eriksson, Stora Enso
Supervisor: Olof Zetterqvist, Applied Mathematics and Statistics
Examiner: Umberto Picchini, Applied Mathematics and Statistics

Master's Thesis 2022
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Multivariate Time Series Forecasting of Earnings Before Interests and Taxes
Implementing a VARMA model for the corrugated production business at Stora Enso
FRIDA MOBERG
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Financial forecasting is an important tool for companies when planning their operations and structuring their organization. At the Packaging Solutions division at Stora Enso they work with three different forecasts for their corrugated business, where one of them is a 15-months rolling forecast. This forecast is updated each month, which is a time consuming process. It is desired to decrease that time to make more time available for analysing the result from the forecast and plan for actions. One of the most important outcomes is the forecast for Earnings Before Interest and Taxes (EBIT). Therefore, this project aims to create an automatic model that can make the forecasting process for EBIT more efficient and more accurate. The input data consists of 16 different features, such as raw material costs, end product price and maintenance costs, that are recorded monthly between 2014-2021, for seven different countries and three regions. Since the amount of data is relatively small and multivariate, the vector autoregressive moving average (VARMA) model was selected. During the model training, five different combinations of features were tested for all countries and regions. The result showed that the accuracy increased compared to the company model for four countries and one region. It could be seen that the countries that had more stable EBIT data worked well with the VARMA model while the ones that included sudden increases or decreases were more difficult to model, as expected. To conclude, the VARMA model is a good option to make the forecasting process more efficient but the model would benefit from some fine adjustments before it can be implemented in the daily work at Stora Enso.

Keywords: Time series, Financial forecasting, Multivariate forecasting, VARMA, Earnings Before Interests and Taxes

Acknowledgements

Fist of all I would like to thank my supervisors Elin Eriksson and Olof Zetterqvist. Olof who has shown great support, answered all of my questions and helped me solve any obstacle along the way. Elin who has discussed ideas, provided me with all information I could possible need from Stora Enso and connected me with experts at Stora Enso that taught me everything about forecasting at your company. To all of you, I would also like to express my gratitude. Then, I want to share a special thank you to my examiner Umberto Picchini and to Stora Enso for giving me the opportunity and means to perform this project.

Lastly, I want to thank my dear family and friends. Your encouragement and endless support throughout my five years studies at Chalmers means everything.
Thank you!

Frida Moberg, Gothenburg, May 2022

Contents

List of Figures

List of Tables

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ACF	Autocorrelation Function
ADF	Agumented Dickey-Fuller test
AIC	Akaike Information Criterion
AR	Autoregression
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
EBIT	Earnings Before Interests and Taxes
GARCH	Generalised Autoregressive Conditional Heteroscedasticit
MA	Moving Average
MAE	Mean Absolute Error
MASE	Mean Absolute Scaled Error
PACF	Partial Autocorrelation Function
QQ-plot	Quantile-Quantile plot
RMSE	Root Mean Square Error
SARMA	Seasonal Autoregressive Moving Average
VARMA	Vector Autoregressive Moving Average

1
Introduction

This master's thesis will, in collaboration with Stora Enso, investigate how the company can digitalize their forecasting process at the Packaging solution division. In this chapter an introduction to financial forecasting is given together with the aim of this master's thesis and a summary of previous work within the field. This is followed by a presentation of how the Packaging Solutions division is working with forecasting today, and a section describing the data used in the project. All together leading into the problem formulation that outlines the rest of the thesis.

1.1 Background

Trying to predict the future is a challenge people have taken on for centuries, it is known that 600 BC they looked at the distribution of maggots in sheep liver to foretell the future [1]. Since then the history of forecasts tells us about a lot of different methods and applications. As our possibilities for collecting and storing data has increased in later years, data analysis and statistical approaches to forecasting have proven to be the most accurate. However, the demands on a forecast varies for each application. One scenario could be deciding whether to build new apartments in a city, that requires a forecast over how many that are estimated to move into the city during the upcoming years. In another scenario, a hospital is planning their staffing demand, they will need to forecast the patient flow during the upcoming months, which can depend on flu season or holiday celebrations. In a third scenario, the distribution of telecommunication routing is based on a forecast for calls during the next few minutes. Taken together, forecasts can have a varying time horizon and depend on several factors, such as seasonal trends. Therefore, there is not only one forecast model that fits in all cases [1].

An accurate forecast brings great advantages to the users, but inaccurate forecasts can be dangerous and lead to big losses. If, for example a company predicts a large increase in profit, and decides to scale up, but then misses out on the success, it could even mean the end for the company [2]. Forecasts, should in other words, be handled with care and while remembering the risks, forecasts are an excellent asset in many cases. One of the major user fields is within the finance sector, where forecasts are considered before making investments, starting new projects and during commodity price setting, to mention a few [3].

The forest industry group Stora Enso provides renewable solutions in packaging, biomaterials, wooden construction and paper. It is a company that operates in multiple countries and they use forecasts in their everyday work. The company is divided into several divisions and one of them is named Packaging Solutions. They are responsible for the companies corrugated business for which they work with three different types of forecasts, a 15-months rolling forecast, an annual budget and a 10-year financial frame [4]. The different forecasts have different purposes, but in general they use them to plan projects, investments, new hires and pricing of their products. The company also use the forecasts to provide guidance to the stock market. The people that create the forecasts at the Packaging Solutions division manually look at different factors, such as pricing of raw materials, and proceed from that to create a forecast for their profit. Hence, the manual method requires a lot of knowledge about the different variables and the current conditions. Having a lot of insight in the process makes the users well aware of what is causing changes in profit, which is important due to the risks of forecasts. However, this manual workflow is also resource and time consuming. Because of that, it was suggested by the Packaging Solutions division, that a digital tool could make the forecasting process more efficient and more accurate. If the time consumed by creating the forecast decreased, that would release time for people with a lot of expertise to instead analyze the result, consider its consequences and plan for relevant actions. Today, little has been done regarding digitalization of the forecasting processes at Stora Enso, and this master's thesis can be seen as the initialization of that process [4].

Therefore, this project will include a research that localizes the fields where an automated forecasting model could be beneficial and possible to implement, in addition to a data analysis in order to find an appropriate model to use. It will also include an attempt to create, test and evaluate a model. During the evaluation the automatic model will be compared to the model used at Stora Enso today, referred to as the company model. Followed by feedback on how Stora Enso in general and the Packaging Solutions division in particular, can continue to work with similar solutions.

1.2 Aim

The aim of this project is to create an automated forecasting model that can be implemented for the corrugated production business at the Packaging Solutions division at Stora Enso. This model should make the forecasting more efficient and more accurate than the manual methods used today. Furthermore, the aim is to provide feedback on how Stora Enso can continue to work with similar solutions.

1.3 Literature review

Financial time series are characterized by asymmetry, non-correlated serial dependence and unstable clustering which makes it a challenge to produce accurate forecasts [5]. Furthermore, there is a wide range of different desired outcomes from financial forecasting. The majority of research is done with data collected from the stock market, which is usually collected daily and the forecast has a prediction window of a week or a month [6], [7]. But there is also research done with data collected on more rare occasion, for example monthly, that have a longer prediction window [8], [9]. In addition, the prediction quality can vary, in some cases it does not need to be as precise as possible, but rather show if the trend is increasing or decreasing, which is enough information for someone trading on the stock market [7]. While in other cases it needs to be more exact, for example if a company is planning on investments [8]. Therefore, there have been different attempts to find good model approaches, resulting in several suggestions of different models for financial time series [10].

One suggestion is to use the classical approaches, which are the models Moving average (MA) and Autoregression (AR). They are based on smoothing techniques and linear regression [11]. MA and AR can be used on their own or combined into an ARMA model. These models have been frequently used for several years and have developed into different variations, such as seasonal ARMA (SARMA) or ARIMA where the I stand for integrated and indicates a differentiation step [12]. They can also be used for multivariate time series, then Vector is added to the name giving VAR, VMA or combined as VARMA [10]. Another classical model is the Generalized autoregressive conditional heteroscedasticity (GARCH), that make use of the clustering effects that appears in financial time series [6]. There are also researches where ARMA and GARCH have been combined in different ways [6]. The benefits of using classical models are that they are robust, they can work with a small amount of data and the result is easy to interpret. In addition, they have been standard for a long time, which has resulted in that there are several programs to facilitate implementation. On the other hand, they make several assumptions, are prone to overfitting and are linear, meaning that they cannot model relations that are not linear. In addition, classical models can be very computationally heavy, especially for high-dimensional inputs [11].

Another suggestion is to use more advanced machine learning algorithms. Among those are support vector regression [13], random forest [14], ridge regression [15] and Lasso-models [16]. These are more effective than the classical models when using high dimensional data and the random forest model is non-linear and flexible to model non-linearity [13]. Furthermore, these algorithms do not put any distributional assumptions to the input data [15]. The challenges with more advanced machine learning algorithms are in general that they require arbitrary parameter selection and optimization technique selection, which require a lot of knowledge about the methods or heavy computations through trial and error. To solve this, automatic selection methods have been developed, called Auto machine learning (AutoML), which are showing promising results [17], [18]. The drawback of using a more automated solution is that interpretation of the results gets harder and the models are perceived as more complex than the classical models [13].

The most recent studies investigate different deep learning algorithms. Within deep learning, natural language processing is a main field, and within that field transformers are outperforming other methods. Since there are similarities between natural language data and time series data, applying transformers for financial time series forecasting has proven successful [19]. In addition to transformers, other deep learning methods have also been applied, for example convolutional neural networks, recurrent neural networks and long short-term memory in different combinations and variations [11]. The deep learning algorithms show reliability and good performance for financial time series forecasts but require a lot of data, which leads to that preprocessing data and the model training can be computationally heavy and time consuming. In addition, similar to the machine learning algorithms, the deep networks are perceived as complex which will make the result interpretation suffer [11].

As it comes to other machine learning and deep learning methods, the proportion of research done with stock market data increases even further, while research done with other types of financial data is more rare. Wasserbacher and Spindler discuss in the article *Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls* that this could be due to the economical earnings that are possible from only a small improvement within stock market forecasting [8] . While other fields usually demand a larger improvement to find value in adapting a new method. However, Wasserbacher and Spindler further discuss that there is likely a change in the near future, as larger companies have started to collect and store large amount of data in a way that will allow for improved machine learning algorithms in the future. In addition, the machine learning algorithms are improving and are becoming more publicly available [8].

1.4 Forecasting at Stora Enso

As mentioned previously, the Packaging Solutions division at Stora Enso is currently working with three different forecasts for their corrugated business, a 15-months rolling forecast, an annual budget and a 10-year financial frame. The rolling forecast is updated each month and is therefore the one that is most time consuming to create [4]. At the same time, it usually has a good accuracy for the upcoming four months, which is considered to be an important time frame. The rolling forecast is primarily used for planning in the near future, looking into the next month, the current quarter and the end of the current year, in that order. It is a frequently used tool for pricing and planning within the division [20],[21]. The rolling forecast for corrugated business covers *End product sales, Variable costs* and *Fixed cost*. These three subjects are divided into 15 different contributing features, and results in a final 16th feature named *Earnings before interests and taxes* (EBIT). A list of all features is presented in Table 1.1. The EBIT forecast is then what is most considered during planning within the division. The result from the EBIT forecast is also what the Packaging Solutions division use when they provide guidance to the stock market [4].

As the corrugated part of the division operates in seven different countries, there are separate forecasts created for each country. In addition, the separate forecasts are combined into three larger regions. Due to confidentiality the countries and regions will not be specified but instead the countries are marked from A-G and the regions are marked from I-III, where region I covers the countries B+C+D, region II adds E to the previous, making it cover B+C+D+E, and region III covers F+G [4]. Creating the 15-months forecasts is a similar process within each region but some details can vary between the regions, see section 1.4.1. However, according to the ones creating the forecasts today, the most contributing features to EBIT are in general raw material costs and end product sales [20],[21]. Raw material costs are independent of internal factors and the main driver for end product price setting, where the price setting lags behind raw material cost changes with about three to four months. The price multiplied by delivered end product volume results in the sales. By that, the raw material cost is not only the main expense but also linked to the final sales result [22]. Fixed costs are also a large expense, though these are usually more stable and considered in the annual budget with occasional updates through the year. Included in the fixed costs is maintenance cost, which is regulated by a fixed plan, but if it deviates from the plan it can affect the EBIT forecast a lot [20],[21]. In addition to the numbers of different features, there are trends and seasonality that are considered when creating the EBIT forecast. Most features have an annual seasonality, for example the delivered volume, that typically increases during the second half of the year. The raw material cost is an exception from this as it has a longer seasonal period of approximately 1.5-2 years [4].

This manual method of creating the rolling forecast is appreciated for being flexible and easy to overview, which leads to a large understanding of the forecast results [20], [23]. The most difficult parts are considered to be estimation of delivered volume and unforeseen expenses. Delivered volume is unstable and since it is a main contributor to the result it can lead to large deviations in the resulting EBIT forecast. Unforeseen expenses can also lead to a forecast that deviates from the true values [21].

There has not been any previous attempt to automate EBIT forecasting at Stora Enso, but in 2018 they did a project where they investigated a dynamic pricing tool, in which they included a forecasting model for price. That project presented a promising model but struggled with data collection and found problems with very local factors that affected the result and was not continued [24], [25].

1.4.1 Country and region specific properties

Above are the general principles that applies to all countries. However, the countries and regions are different in some respects, which are explained bellow.

Country A
Country A is not a part of any region. It is one out of the two largest countries in the aspect of delivered volume and consist of three corrugated units. Country A has the highest EBIT. It operates in their local currency but the forecast is presented in Euros according to the current exchange rate [22].

Region I: B,C,D
Region I is the smallest region and there is only one corrugated unit, located in country B. Countries C and D instead buy their corrugated material from country B and there are therefore internal costs considered when the forecast is put together. As a result EBIT in C and D are about half of that in B [22].

Region II: B,C,D,E
This region adds country E to region I. Country E is the other out of the two largest countries in the aspect of delivered volume. Country E operates very similar to country A as it also consists of three corrugated units and has a high EBIT. In addition, country E is operating in their local currency but the forecast is presented in Euros according to the current exchange rate [22].

Region III: F,G
The countries F and G operates similar to each other. Regarding size of delivered volume and EBIT, they are smaller than countries A and E but larger than B, C and D. Countries F and G are differ from the other countries in a way that they have higher personnel costs, which also reflects on their fixed costs as some wages are included under that post. Like A and E, country G is operating in a local currency but presents their forecast according to the current exchange rate. Country F operates in Euros [4].

1.4.2 Data

The data presented in this report is received from Stora Enso. It is historical data that can also be found in their annual reports. However, to remove unnecessary focus of the company economic situation, the units presented are standardized on an index scale starting at 100 units. Figure 1.1 shows the sum of all countries joined EBIT plotted using the index scale.

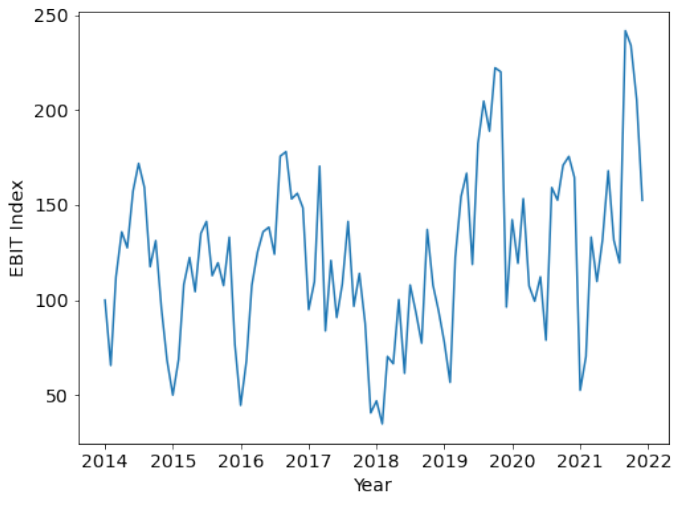


Figure 1.1: Example of data used in the project. The plot shows EBIT for all seven countries, each month from 2014-01 to 2021-12 on a scaled unit.

All data was collected from the Packaging Solutions division and consisted of monthly balance for each feature and each country. The data reached from 2014-01, earlier material was not accessible due to a reorganization, to 2021-12. That gives a total of 96 data points per feature. The raw data contained multiple posts beyond to the 16 relevant features, for example summations of features and old features that are not considered anymore. They were removed and the remaining features included in the analysis are presented in Table 1.1.

Table 1.1: All features considered in the analysis presented together with their measured unit. The table also comment the features that are not self-explanatory.

Feature	Unit	Comment
Production	m^2	Produced end product
Deliveries	m^2	Delivered end product
Inventory	m^2	End product in stock
Sales	EUR	Earnings from deliveries
Price	EUR/m^2	Price on deliveries
Other operating income	EUR	
Change in finished goods inventory	EUR	
Transportation freight costs total	EUR	
Energy costs total	EUR	
Other operating variable costs total	EUR	
End product purchases total	EUR	Raw material costs
Personnel costs	EUR	
Maintenance materials services costs total	EUR	
Other fixed costs	EUR	
Depreciations and impairment charges	EUR	
Operational EBIT	EUR	EBIT

1.5 Problem definition and limitations

In order to create an automated forecasting model that can improve the forecasting at Stora Enso, there are four main issues to consider. First, an initial research of how the forecast is created today, to be able to decide what the purpose of the automatic forecast should be. The content from this research is found in section 1.4. Based on the information received it was decided to focus this master's thesis on creating a model for the corrugated business at the Packaging Solutions division and to focus on the EBIT forecast, since it is one of the most important forecasts considered in the daily work at Stora Enso. Furthermore, it was decided to recreate the 15-months rolling forecast because it is frequently used in the daily work and because it is time consuming to update it each month. It was clear that the most considered features in the company model are *Price*, *Deliveries* and *End product purchases total*, so it should be considered if they are necessary for the new model as well. From the initial research a good model was defined as an accurate model that can predict peaks and sudden changes. The initial research also concluded that a good model should keep a close connection to the input, to make it easy to understand, if for example it is a certain feature that affect, why the forecast comes out as it does.

The second issue is to investigate the data. In this case the data, as presented in section 1.4.2, consist of many different features and several countries. But since each feature only has 96 data points, it refrains from using deep learning models and some machine learning models, as they require larger amount of data for training to receive a good result. This limitation in combination with the condition of easy interpretation, lead to the decision of using a classical model. Since the data is multivariate, the appropriate model is the VARMA model.

When a model has been selected, the third issue is to, create, test and evaluate the model. Due to the available time span, it was decided to test five different feature setups, see Table 3.1, on all countries and regions. The feature setups were constructed based on the initial research and on correlation analysis. For the model evaluation it is necessary to recall the definition of a good model stated above. One of the main objectives is the model accuracy, therefore the evaluation for this project includes calculation of root mean square error (RMSE) to measure the accuracy. RMSE can also be obtained from the company model, making a comparison possible. To compare the models between each country it is necessary to use an additional metric called mean absolute scaled error (MASE). RMSE is affected by the order of the data and since the EBIT varies a lot between the countries, RMSE cannot be compared between countries. MASE on the other hand, uses a scaling factor that makes it possible to compare the result between the countries. The mathematical theory of RMSE and MASE is found in section 2.2.4. The evaluation of how well the model captures specific peaks is more difficult, as the available metrics are more complex. Instead the results were plotted and compared in a more approximate manner. It is also difficult to measure how easy the results are to interpret, but it was discussed and considered in the results. The quality of the model was evaluated through residual analysis, both investigating correlation between the residuals and the distribution of the residuals.

As the last fourth issue comes suggestions for future work. Based on the evaluation, this part should provide recommendations on how the Packaging Solutions division can develop the concept further or conclude that it is better to stay with the company model.

2
Theory

The theory chapter will start with an introduction to time series and their properties of stationarity, trends, seasonality and correlation. After follows a description of the different models that are used to create the forecast, an explanation of the model selection criterion applied and the evaluation metrics used.

2.1 Time series

A time series is defined as a set of observations y_t, collected at time t. If the observations are made for fixed time intervals, e.g. every second, every hour or every year, it gives a discrete time series. Whereas if the observations are made on a continuous interval the time series is a continuous time series [10]. Both discrete and continuous time series can typically be decomposed into three parts

$$y_t = f(T_t, S_t, \varepsilon_t) \tag{2.1}$$

where T_t is the trend component, S_t is the seasonal component and ε_t is the error, all at time t [26]. The trend component is a smooth function that changes slowly and captures only the general changes, neglecting temporary fluctuations. The seasonal component is a periodic function with period d, and describes any repeating pattern in the series. A simple example would be sun lotion sale in Sweden, the sale will increase during the summer, therefore there is seasonality in the data meaning that, if the sale each day is recorded, $d = 365$, or if the sale each month is recorded $d = 12$. At the same time the sun lotion sale increases every year due to an increased knowledge of sun damage and therefore the trend component will be some increasing function. What is not described by the trend component or the seasonal component will be classified as residuals [10].

If the time series shows a pattern where the seasonal trend magnitude increases, the function is best expressed as multiplicative, see function 2.2. If there is no magnitude increase, the function is instead expressed as additive, see function 2.3 [26].

$$y_t = T_t \times S_t \times \varepsilon_t \tag{2.2}$$

$$y_t = T_t + S_t + \varepsilon_t \tag{2.3}$$

ε_t is stationary, meaning that it is independent of time. It is common to assume stationary data when using most analyzing tools and modeling approaches, therefore the purpose is often to first eliminate T_t and S_t, then use ε_t for modelling and forecasting [10].

2.1.1 Trend and seasonal adjustment

There are a few different ways to find and remove the trend and seasonal components, the processes are named trend adjustment and seasonal adjustment. One of the standard procedures is to use differencing and it can be used for both trend and seasonal adjustment. The differencing formula is presented in equation 2.4

$$\nabla_d y_t = y_t - y_{t-d} \tag{2.4}$$

where ∇_d is the differencing operator and d is the lag-order. For trend adjustment d is set equal to 1 and for seasonal adjustment d is equal to the seasonal period [1].

If both trend and seasonality are present in the data, it is recommended to perform seasonal adjustment first since that might result in an immediate stationary time series where trend adjustment is needless. Whereas performing trend adjustment first will not remove any seasonality. Furthermore, it is recommended to always select d so it can be explained by a seasonal cycle. It is possible to select d as any number but that might result in loss of interpretability for the result [1].

Differencing is a quick method that does not entail any parameter estimations or other heavy calculations and it is therefore easy to use in data preprocessing. There is one drawback that appears due to the fact that equation 2.4 does not apply for $t < d$. This means that the first d data points cannot be transformed [10].

2.1.2 Agumented Dickey-Fuller test

To analyze whether the remaining data after trend or seasonal adjustment is stationary, an Augmented Dickey-Fuller (ADF) test can be conducted. It is a statistical hypothesis test with the null and alternative hypothesis presented in 2.5 [10].

$$\begin{cases} H_0: & \text{Time series is non-stationary} \\ H_1: & \text{Time series is stationary} \end{cases} \tag{2.5}$$

The ADF-test is, as the name implies, an augmented version of the Dickey-Fuller test. The Dickey-Fuller test is a unit root test, meaning that the null hypothesis is that the root equals unity. If a time series has a unit root, it is not stationary. When the Dickey-Fuller test is extended into an ADF-test it can work for more complex models and the ADF-test is therefore standard. The ADF-model is presented in equation 2.6 [10].

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + ... + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \tag{2.6}$$

where α is a constant, β is the time trend coefficient, γ is the coefficient of the first lag on y_t, p is the number of lags that is included in the model, δ is a constant, $\Delta y_t = y_t - y_{t-1}$ and ε_t is an error term. The value of p is typically set to maxlag, determined by using t-statistics testing for significance of each lag or set to the lag that minimizes Akike information criterion (AIC) or Bayesian information criterion (BIC). AIC and BIC are explained bellow in section 2.2.3. The coefficient of interest in the ADF-test is γ, and since equation 2.6, is different from the original Dickey-Fuller test, the hypothesis that corresponds to equation 2.5, is as follows in equation 2.7 [10].

$$\begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma < 0 \end{cases} \tag{2.7}$$

The estimated $\hat{\gamma}$ is then divided by the standard error to obtain the test statistic. If the test statistic is less (more negative) than the critical value, the null hypothesis can be rejected in favor of the alternative hypothesis, meaning that the time series is stationary [10].

2.1.3 Correlation

To make a good prediction of the feature in focus, it is desired to find features that correlates with the feature in focus, while not correlating with each other. Since a forecasting model depends on previous time steps, called lags, it is necessary to have correlation for lags greater than zero in order to be able to create a prediction. The correlation between two variables (X, Y) can be calculated using equation 2.8

$$r_k = \frac{\sum_{t=k+1}^{T}(x_t - \bar{x})(y_{t-k} - \bar{y})}{\sqrt{\sum_t^T (x_t - \bar{x})^2}\sqrt{\sum_t^T (y_t - \bar{y})^2}} \ , k < T \tag{2.8}$$

where t is the current time step, T is the last time step in the X time series, k is the lag, $\bar{x}$ is the mean of X and $\bar{y}$ is the mean of Y [27]. k can vary to measure correlations at different lags for each feature. In addition, Y can be equal to X but at a different lag. In that case the resulting correlation is called auto correlation function (ACF) [27]. In Figure 2.1 an example of ACF for *EBIT* from region III is presented.

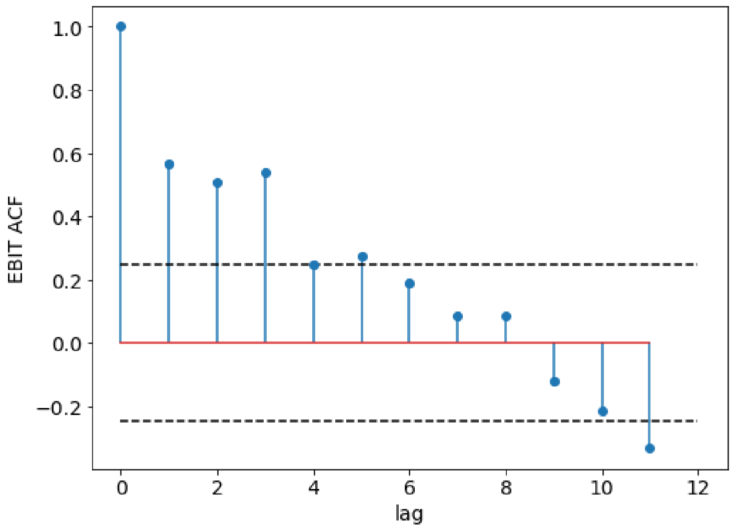


Figure 2.1: ACF for EBIT data from region III at different lags. The horizontal lines represent the critical values at ± 0.247, corresponding to a 5% significance level.

The dashed horizontal lines in Figure 2.1 represents the critical value, also known as the confidence interval, if the correlation reaches outside these lines, the correlation is statistically significant at the selected significance level. The confidence interval is obtained using equation 2.9

$$ci = \left[-\frac{z_{\alpha/2}}{\sqrt{N}}, \frac{z_{\alpha/2}}{\sqrt{N}}\right] \tag{2.9}$$

where N is the number of observations and $z_{\alpha/2}$ is the z-score taken from the normal distribution at the selected significance level α. The significance level is typically 5%, which corresponds to $z = 1.96$ [10]. In Figure 2.1 it can be seen that correlation at lag 2,3,4 and 11 reaches outside the critical values, meaning that these lags potentially can be used for prediction. An immediate consequence from equation 2.8 is though that as k increase, the amount of data to compute the correlation on will decrease. Therefore, correlation for large lags does not contribute with enough information to be considered. In the example, Figure 2.1, it can therefore be questioned if the correlation at lag 11 is important to include for prediction purpose [28].

One thing to remember is that correlation is only measured at linear scale, if there are non-linear relations, which is possible, it will not show in this type of analysis [28]. Another way of measuring correlation is through partial correlation. Partial correlation also measures the correlation on a linear scale, but measure the correlation between residuals from linear regression of the time series. The equation for partial correlation is equal to equation 2.8, but replace the observed values from (X, Y) with the residuals. Like ACF this can be done on the same time series at different lags, giving the partial auto correlation function (PACF) [10].

2.2 Time series models

As mentioned in the Literature review section 1.3, there are multiple ways of modelling time series. In this project the VARMA model was selected. To understand the VARMA model, this section will start by introducing the theory for modelling univariate cases before scaling up to the multivariate case that applies the VARMA model.

2.2.1 Univariate time series models

One way of modelling univariate time series is by creating a linear regression model of the time series according to equation 2.10

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{2.10}$$

where c is the intercept, $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)$ are the parameters that need to be estimated in the model, ε_t is independent and identically distributed noise (iid noise) and p is the auto-regressive trend order. The optimal p is given by the PACF plot and corresponds to the number of significant lags. This is called an Autoregressive model and is typically denoted AR(p) [10].

Another way of modelling univariate time series is by creating a model for the error terms. The error terms, ε_t, are given by taking the difference between the observed values and the predicted values. Then, the error terms are modelled as a linear function according to equation 2.11

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \tag{2.11}$$

where c is the intercept, $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)$ are the parameters that need to be estimated in the model, ε_t is iid noise and q is the moving-average trend order. The optimal q is given by the ACF plot and corresponds to the number of significant lags. This is called a Moving average model and is typically denoted MA(q) [10].

An Autoregressive moving average (ARMA) model combines the AR(p) model with the MA(q) model and therefore has both p and q as required input parameters [10]. The equation is presented bellow in 2.12.

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2.12}$$

where c is the intercept, ε_t is iid noise, p is the AR order and q is the MA order. The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ need to be estimated in the model and that is done by maximizing the loglikelihood function [1].

The likelihood function is derived from the probability of receiving a certain estimation given a set of parameters, $\boldsymbol{\theta}$. If the error $\varepsilon_i = y_i - \hat{y}_i(\boldsymbol{\theta})$, between the observation y_i and the estimation $\hat{y}_i$ can be assumed as normally distributed, the likelihood function is given by 2.13,

$$\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta}) \tag{2.13}$$

where σ_i is the standard deviation of the normal distribution, according to the assumption $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. By maximizing the likelihood function, the probability that the observations and estimations belong to the same distribution is maximized. In practice, taking the logarithm of equation 2.13 simplifies the calculations, giving the loglikelihood [29].

2.2.2 Vector Autoregressive Moving Average (VARMA)

If the dataset contains multiple variables, they can be modeled using the vector version of the three mentioned models above. The combined model is therefore called VARMA and it models each variable as a linear combination of the past values of itself and the past values of the other variables [10]. The VARMA(p, q) process on a general form with VAR order p, MA order q and K features can be written as in equation 2.14.

$$y_t = c + \sum_{i=1}^{p} A_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} M_i \varepsilon_{t-i} \tag{2.14}$$

where $y_t = (y_{1t}, ..., y_{Kt})'$ is the $K \times 1$ vector with the observations for each feature at time t, A_i and M_i are fixed $K \times K$ coefficient matrices, $c = (c_1, ..., c_K)'$ is a fixed $K \times 1$ vector of intercept terms, $\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{Kt})'$ is a K-dimensional iid noise.

Including more variables in the model will increase the number of parameters in the model, as it is given by K(1+pK+qK) where K is the number of variables. The more parameters included the heavier the computations get and the larger the errors tend to get. Therefore, it is desired to only include the variables that contribute with information to the model [10]. The VARMA parameters are also estimated using the loglikelihood function in 2.13. The estimation of VARMA has a fundamental problem, the parameter estimation is not unique, meaning that there might be several parameters sets that give the same prediction. This results in that if the actual parameter values are of interest, the nonuniqueness needs to be considered, see [29] on more details on how to do that. However, in this project, focus is on the forecast which is not affected by the nonuniqueness problem [29].

2.2.3 Akaike and Bayesian information criterion

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are tools used for model selection when the set of models is finite. They are both based on information theory and aim to identify the model that can explain the data using the smallest possible number of parameters. The balance to find an appropriate number of parameters is hard, where too few will give an underfitted model that is not able to find the information in the data. To many parameters on the other hand would lead to overfitting of the training data and the model not being able to generalize onto the test data. AIC tries to find the balance by minimizing the negative likelihood while at the same time penalize as the number of parameters increases, see equation 2.15

$$AIC = -2 * \log(\mathcal{L}) + 2p \tag{2.15}$$

where $\mathcal{L}$ is maximum value of the likelihood function, equation 2.13, for the model and p is the number of parameters in the model. The models with smaller AIC are preferable chosen. BIC will instead find the model that is most probable given the data, according to equation

$$BIC = -2 * \log(\mathcal{L}) + p * \log(n) \tag{2.16}$$

where n is the sample size. Similar to AIC, the model selected using BIC is the one that gives the smallest score [30]. It has been seen in studies that AIC has the tendency to select larger models that result in overfitting, while BIC selects smaller models. Therefore, it is recommended to use an approach that includes both criteria [31]. Important to note is that AIC and BIC do not measure absolute quality of the model, but only quality in relation to the other models tested [30].

2.2.4 Evaluation metrics

There are several different ways to evaluate a forecast. Among those, root mean square error (RMSE) is one of the most common metrics. It calculates the difference between the predicted value and the observed value at each time step according to equation 2.17

$$RMSE = \sqrt{\sum_{t=t_0}^{t_{n-1}} \frac{(\hat{y}_t - y_t)^2}{n}} \tag{2.17}$$

where t_0 is the starting point of the forecast, $\hat{y}_t$ is the predicted value at time t, y_t is the actual value at time t and n is the number of observed values. From the equation it can be seen that RMSE increases if the predicted value is far from the observed value. That means, RMSE should be as small as possible. In addition to evaluation, RMSE is common to use during model training. The aim is to select the parameters that minimizes RMSE, though it needs to be done carefully due to the risk of overfitting.

RMSE does not include any normalizing step, making it dependent on the scale of the samples. If the order is small that will result in a small RMSE and if it is large, it will result in a larger RMSE. Therefore, it is not possible to compare RMSE between data sets that are not on the same scale. To be able to compare models on different scales, another metric can be used, called mean absolute scaled error (MASE). MASE first computes the mean absolute error (MAE) according to equation 2.18

$$MAE = \frac{1}{n} \sum_{t=t_0}^{t_{n-1}} |\hat{y}_t - y_t| \tag{2.18}$$

where t_0 is the starting point of the forecast, $\hat{y}_t$ is the predicted value at time t, y_t is the actual value at time t and n is the number of observed values. Then the MAE of the naive model is computed. The naive model simply sets the prediction equal to the previous value, which results in equation 2.19.

$$MAE_{\text{naive}} = \frac{1}{n-1} \sum_{t=t_1}^{t_{n-1}} |y_t - y_{t-1}| \tag{2.19}$$

If the data is seasonal it is also possible to have a naive model that sets the prediction equal to the observed value of the previous season, meaning replacing y_{t-1} with y_{t-d} where d is the seasonal period. When both MAE and $\text{MAE}_{\text{naive}}$ are determined MAE is normalized using $\text{MAE}_{\text{naive}}$, giving MASE according to equation 2.20.

$$MASE = \frac{MAE}{MAE_{\text{naive}}} \tag{2.20}$$

Hence, MASE < 1 indicates that the model is better than the naive model and MASE > 1 indicated that the naive model is better than the model. That means models used on data at different scales can be compared on how accurate they are in relation to the naive model.

In addition to accuracy measures, a model can be evaluated through residual analysis. Since normally distributed residuals is a common assumption of a model, that can be checked by looking at quantile-quantile plot (QQ-plot). The QQ-plot has the sample quantiles on the y-axis and the examined distribution quantiles on the x-axis. If the aim is to find out if the residuals is normally distributed, the examined distribution quantiles are taken from the normal distribution. This results in a plot where the points should follow a line if the sample is normally distributed and the $x = y$ line if the sample is standard normally distributed.

The residual correlations can also be investigated in a process similar to what is described in 2.1.3. A good model should capture the residual correlation, giving no significant residual correlation for lags larger than zero. If there is still a lot of correlation between the residuals after modelling, it is possible to repeat the modelling procedure until all correlation for lags larger than zero has been captured by the model.

3
Methods

Bellow follows the methods for creating the model in chronological order. The sections include data collection and preprocessing, code modules used, hyperparameter search, model training and finally the evaluation process. Motivations for different approaches and model selection are based on the reasoning and limitations presented in the Problem definition section 1.5.

3.1 Data collection and preprocessing

The initial part of this project was to collect information about the current forecasting method at Stora Enso. This was done through interviews and emails. Thereafter, the data was collected. The details of how the raw files were converted are specified in Appendix A. When all data files were completed, they were imported to the program where the model was created.

The program code was written in Python using Statmodels for creating the VARMA model and scikit-learn for preprocessing data and evaluation of the model. In addition, common modules such as numpy, math and scipy were used.

The data was preprocessed through standardization and seasonal adjusted through differentiation using period 12. After that, the data was tested for stationary using the ADF-test as described in section 2.1.2. Initially only the data for region III was tested with the ADF-test due to time limitations. As seen in Table 4.1, almost all data was stationary for region III after the seasonal adjustment. Therefore, it was decided to not do any further preprocessing. When the data for the other countries and regions were tested with the ADF-test, they did not show as much stationarity as the data from region III, but since *EBIT* was stationary for a majority of the countries and regions, see Table 4.1, there was no additional preprocessing done on these data sets either.

3.2 Feature selection

During the project, five different feature setups was tested, see Table 1.1. The features for setup 1 and 2 were based on information received from the initial research of forecasting at Stora Enso and included different combinations of the features that are the most considered in the company model. The features for setup 3, 4 and 5 were instead based on the correlation plots. The correlation between the features was calculated pairwise according to equation 2.8 and plotted. Since EBIT forecasting is the goal, the correlation between each feature and EBIT was plotted separately in a second plot, an example of this is presented in Figure 2.1, where the correlation between the different features and EBIT is plotted for country D. It can be seen

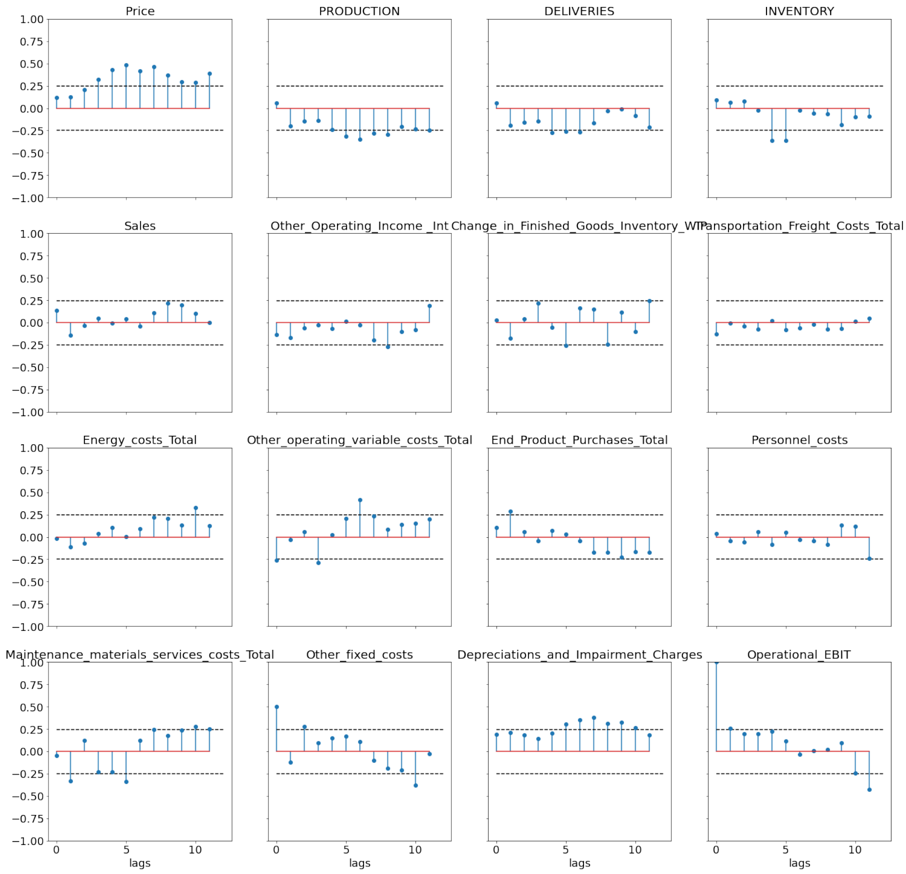


Figure 3.1: Correlation between the different features and EBIT for country D. The horizontal lines represent the critical values at ± 0.247, corresponding to a 5% significance level.

in Figure 3.1 that there are significant correlation for lag larger than zero between *EBIT* and the features *Price, Production, Inventory, Other operating variable costs, End product purchases, Maintenance materials services costs total, Other fixed costs* and *Depreciations and impairment charges*. This procedure was repeated for every country and region, and the features that were significant in most countries and regions were selected for the different setups. Setup 5 includes the features that showed significant correlation with EBIT in 7, 8 or 9 countries and regions. In setup 3 the features that have high correlation with the other included features were removed. Finally setup 4 combined setup 1 and 3 by replacing *Deliveries* with *Price*. The same setups were applied to all countries and regions. After the feature selection, the data for each country or region was divided into a training, a validation and a test set. Due to the low amount of data the validation set included 6 data points and the test set 15 data points, as 15 is the desired length of the resulting forecast. The remaining 75 points constituted the training set.

Table 3.1: This table specifies the variables included in each feature setup.

Setup 1	Price End product purchases total Operational EBIT
Setup 2	Deliveries Sales End product purchases total Operational EBIT
Setup 3	Deliveries Energy costs total End product purchases total Maintenance materials services costs total Other fixed costs Operational EBIT
Setup 4	Price Energy costs total End product purchases total Maintenance materials services costs total Other fixed costs Operational EBIT
Setup 5	Production Deliveries Energy costs total End product purchases total Maintenance materials services costs total Other fixed costs Operational EBIT

3.3 Hyperparameter search

The hyperparameters, p and q needed for VARMA were determined through a hyperparameter search. The search was carried out separately for each country or region and for each feature setup, and tested all combinations of p and q in the range 0-6. The limit was set to 6 since the low amount of data lead to that correlation for larger lags is based on very short time series, as explained in section 2.1.3, and for efficiency reasons. For each combination, the model trained on the training set and evaluated on the validation set. Evaluated AIC, BIC and RMSE were stored. The hyperparameter search procedure is described in algorithm 1. Three different combinations of hyperparameters were then selected. The first combination was the p and q that minimized the sum of AIC and BIC. The second combination was the p and q that minimized the sum of RMSE for all features included in the VARMA model. Finally, the last combination was the p and q that minimized RMSE for EBIT.

Algorithm 1 Pseudocode for hyperparameter search

```
  Input: Train data set, validation data set, (p,q)-grid, feature setup
Normalize the train data set
Seasonally adjust the train data set
for p,q in (p,q)-grid do
    Fit VARMA model with order = p,q
    Get model AIC and BIC
    Use model to make prediction of length = len(validation)
    Remove standardization and invert differentiation of result
    for feature in feature setup do
        Get RMSE using the validation set
    end for
    Append AIC, BIC and RMSE to results
end for
  Output: Results dataframe
```

3.4 Model training and evaluation

Three models per feature setup and country or region were then trained, one for each selected p, q combination. Followed by a prediction step, using prediction window equal to 15 according to the aim of creating a 15-months forecast. The model was evaluated by analyzing the significance of each coefficient in the EBIT prediction. Furthermore, the residuals from the model was plotted in a QQ-plot, to test if they could be considered normally distributed. The residuals were also checked for correlation pairwise against each other. Finally, the prediction was compared to the test set to calculate RMSE and MASE according to equations 2.17 and 2.20 respectively. Five setups and three models per setup resulted in 15 different RMSE and MASE scores per country or region. The best model for each country was defined as the model that had the lowest RMSE for their EBIT prediction.

RMSE and MASE was also calculated for the company model by comparing the 15 months forecast from the company model with the test set. The result was compared to the RMSE and MASE for the best VARMA model for each country or region.

4

Results

The following sections presents the results from the feature selection, hyperparameter search and the VARMA model evaluation on each country and region. Since there were many countries and regions, only a few figures are highlighted in this chapter to maintain clarity. They are selected to represent the range of results received. The complete collection of figures can be found in Appendix B.

4.1 Preprocessing

After the normalization and seasonal adjustment, the ADF-test was performed at a 5% significance level. The result is presented in Table 4.1.

Table 4.1: Results from the ADF-test on a 5% significance level. The table presents the results for every feature included in the best feature setup for each country and region. If the feature time series is stationary it is marked as "S", if it is non-stationary it is marked "NS" and if it is not represented in the feature setup it is marked with a "-".

Feature	A	B	C	D	E	F	G	I	II	III
Production	-	-	-	-	-	-	-	NS	-	-
Deliveries	-	NS	S	-	NS	-	S	NS	-	NS
Sales	-	S	NS	-	-	-	NS	-	-	-
Price	S	-	-	NS	-	NS	-	-	NS	-
Energy costs total	-	-	-	-	NS	-	-	S	NS	S
End product purchases total	NS	NS	NS	NS	NS	NS	NS	NS	NS	S
Maintenance materials services costs total	-	-	-	-	S	-	-	S	S	S
Other fixed costs	-	-	-	-	S	-	-	S	S	S
EBIT	NS	S	S	S	NS	S	S	NS	NS	S

In Table 4.1 it can be seen that for region III all features are stationary except *Deliveries*. *EBIT* is stationary for a majority of the countries and regions while *End product purchases total* is only stationary for region III.

4.2 Feature selection and hyperparameter search

Table 4.2 bellow presents the feature setup (as specified in Table 1.1), selection criterion and hyperparameters that resulted in the lowest RMSE for each country or region. From Table 4.2 it can be seen that feature setups 1 and 2 are the most occurring ones for the countries, while the regions and all countries together received better results for feature setups 3, 4 and 5.

Table 4.2: Feature setup, selection criterion and hyperparameters that resulted in the lowest RMSE for each country or region.

Country/Region	Best feature setup	Selection criterion	p,q
A	1	AIC and BIC	2,2
B	2	AIC and BIC	1,3
C	2	AIC and BIC	3,2
D	1	All features RMSE	0,1
E	3	EBIT RMSE	0,2
F	1	EBIT RMSE	4,2
G	2	AIC and BIC	6,5
I	5	All features RMSE	0,4
II	4	EBIT RMSE	1,6
III	3	All features RMSE	2,4
All countries	3	All features RMSE	6,6

A similar pattern can be seen in the selection criterion where the countries received the best results using AIC and BIC while the regions performed better using minimization of RMSE. Regarding the hyperparameters p and q there are no particular pattern.

4.3 VARMA model evaluation and comparison with company model

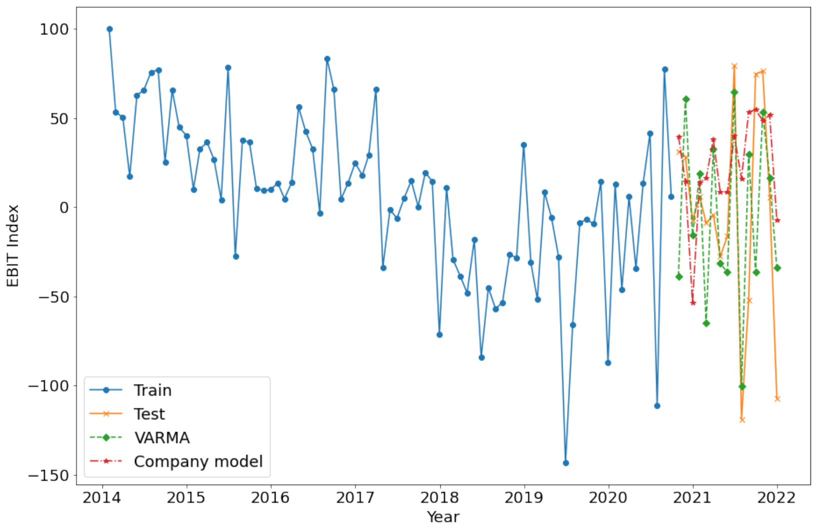
When using the p, q combinations and feature setups specified in Table 4.2, the predictions received the RMSE and MASE scores presented in Table 4.3. Table 4.3 also includes the RMSE and MASE scores for the company model. Furthermore, Table 4.3 includes a column with the percentage difference between the VARMA model RMSE and the company model RMSE, as a measure of how much the VARMA model increased or decreased the accuracy. The percentage difference was computed by dividing RMSE from the company model with the RMSE from the VARMA model. The table is sorted with the country showing the largest percentage improvement at the top. It can be seen in Table 4.3 that country F results in the largest improvement of RMSE compared to the company model. Also G, the other country in region III, shows a good improvement and so does region III as a whole.

Table 4.3: MASE and RMSE for the VARMA model with the lowest RMSE for each country and region, compared to the MASE and RMSE for the company model for each country and region.

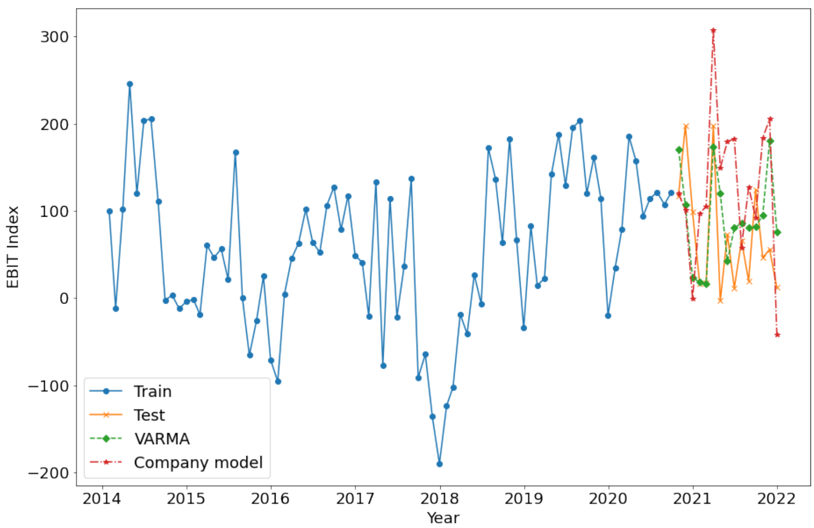
Country/ Region	VARMA MASE	Company MASE	VARMA RMSE $\times 10^5$	Company RMSE $\times 10^5$	Difference %
F	0.70	1.17	3.15	4.99	+58
A	1.05	1.35	5.49	7.19	+31
III	0.69	0.94	7.29	8.75	+20
G	0.69	0.81	5.22	6.15	+18
B	0.75	0.76	2.29	2.38	+4
C	1.21	1.01	0.84	0.79	-6
All countries	1.04	0.98	14.73	13.16	-11
II	1.30	1.14	9.10	7.70	-15
I	1.25	1.05	4.38	3.50	-20
D	2.39	1.89	2.27	1.78	-22
E	1.60	1.22	7.98	5.05	- 37

In addition, these three has the lowest MASE, indicating that they are the best models compared to the other VARMA models. Country A also makes a large improvement in accuracy, but does not result in as low MASE score as country B, that has a low MASE but does not improve accuracy as much as F, A, III and G. The remaining countries and regions did not improve the accuracy compared to the company model, and they show a relatively high MASE. In particular country E resulted in the largest increase of RMSE and country D resulted in the largest MASE.

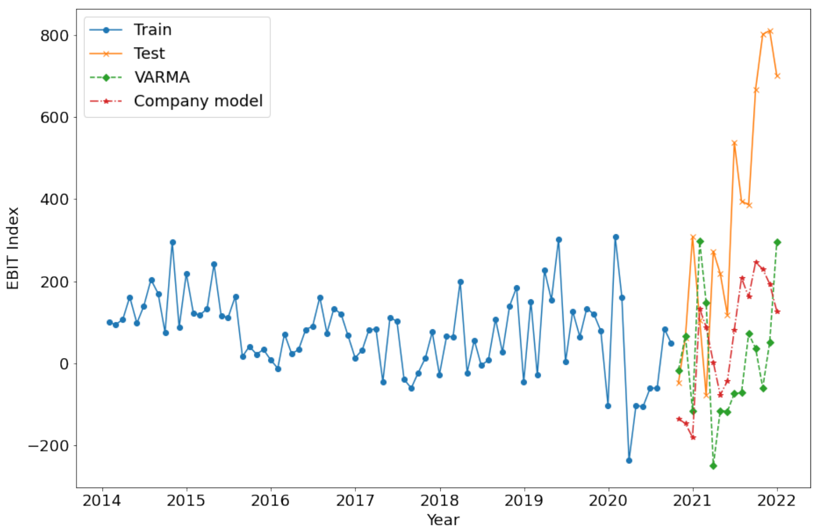
A selection of the predictions are illustrated in Figure 4.1, figures representing all predictions are found in Appendix B.1. In each plot the blue dot line represents the training data, the yellow cross line represents the test data, meaning the values that the predictions should be as close to as possible. The dashed green diamond line represents forecast of the VARMA model and the semi dashed red star line represents the forecast of the company model. Figures 4.1a and 4.1b shows the result from country G and F respectively, which are two countries that improved RMSE according to Table 4.3. That can also be seen from the figure, as the VARMA model prediction follows the test data better than the company model prediction. The other two plots, Figure 4.1c and 4.1d, instead show the results from country D that gave the highest MASE score and country E that resulted in largest increase of RMSE compared to the company model. In both these cases it is clear from the figure that the VARMA predictions do not follow the test data very well. Figure 4.1 further shows that the VARMA model predictions has a large variation and attempts to capture peeks and sudden changes. In figures 4.1a and 4.1b the predictions captures most shifts correctly. While in Figure 4.1c the VARMA model does not follow the test data increase at the end of the prediction and in Figure 4.1d, the VARMA predicts the peaks incorrectly, as it increases EBIT when the observed value drops.



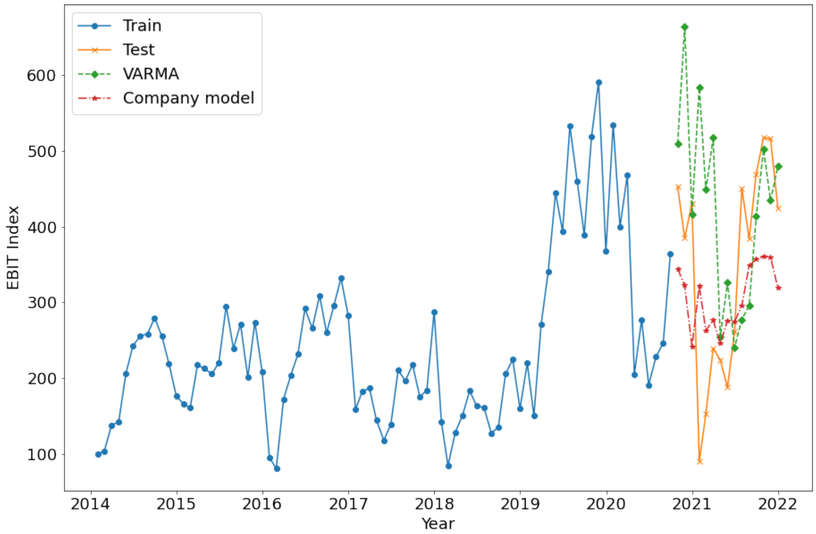
(a) EBIT forecast for country G. This forecast gave the lowest MASE score.



(b) EBIT forecast for country F. This forecast showed the best improvement in RMSE from the company model.



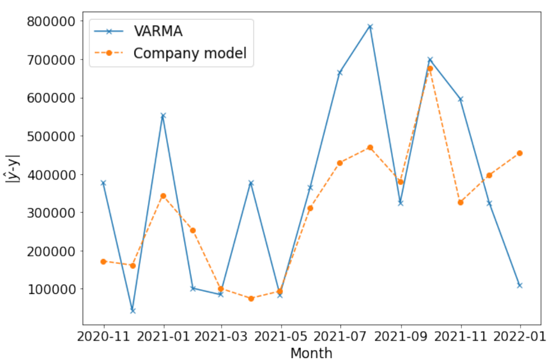
(c) EBIT forecast for country D. This forecast gave the highest MASE score.



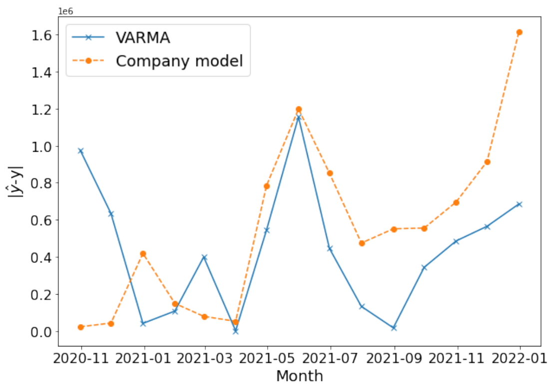
(d) EBIT forecast for country E. This forecast showed the largest deterioration in RMSE compared to the company model.

Figure 4.1: EBIT forecasts for four different countries. The blue dot line represents the training data, the yellow cross line represents the test data, meaning the values that the predictions should be as close to as possible. The dashed green diamond line represents forecast of the VARMA model and the semi dashed red star line represents the forecast of the company model.

To compare the deviations from the observed values with the VARMA model and the company model, they were plotted in separate plots. The plot for region I and the plot for country A is presented in Figure 4.2. The plots for the remaining countries and regions are found in Appendix B.2.



(a) Deviation from the observed value for region I.



(b) Deviation from the observed value for country A.

Figure 4.2: Deviation from the observed value each month. The blue cross line shows the deviation for the VARMA model and the yellow dot line represents the deviation from the company model.

Since a low RMSE indicates better accuracy, the line that is below the other line in Figures 4.2a and 4.2b corresponds to the model that has the best accuracy at that point. It can be seen that in both region I and country A, the company model performs better in the beginning of the forecast, while the VARMA model performs better in the end of the forecast. It can also be seen that the two models in general follow similar trends, indicating that it is the same parts in the observed data that are more difficult or easier to predict for both models. This pattern shows for a majority of the countries and regions, with a few exceptions, see Appendix B.2.

4.4 Model quality evaluation

The model quality was evaluated through analysing residual correlation and distribution. A typical result of the correlation analysis is presented in Figure 4.3 The full result of the correlation is included in Appendix B.3. From Figure 4.3 it can be seen that there are only a few significant correlations for lags that are larger than zero. That is indicating that the model has managed to describe the correlations well. There is significant correlation for when lag is equal to zero which is expected since a forecast uses the previous data points for prediction, and therefore the lag equal to zero correlations are not included in model.

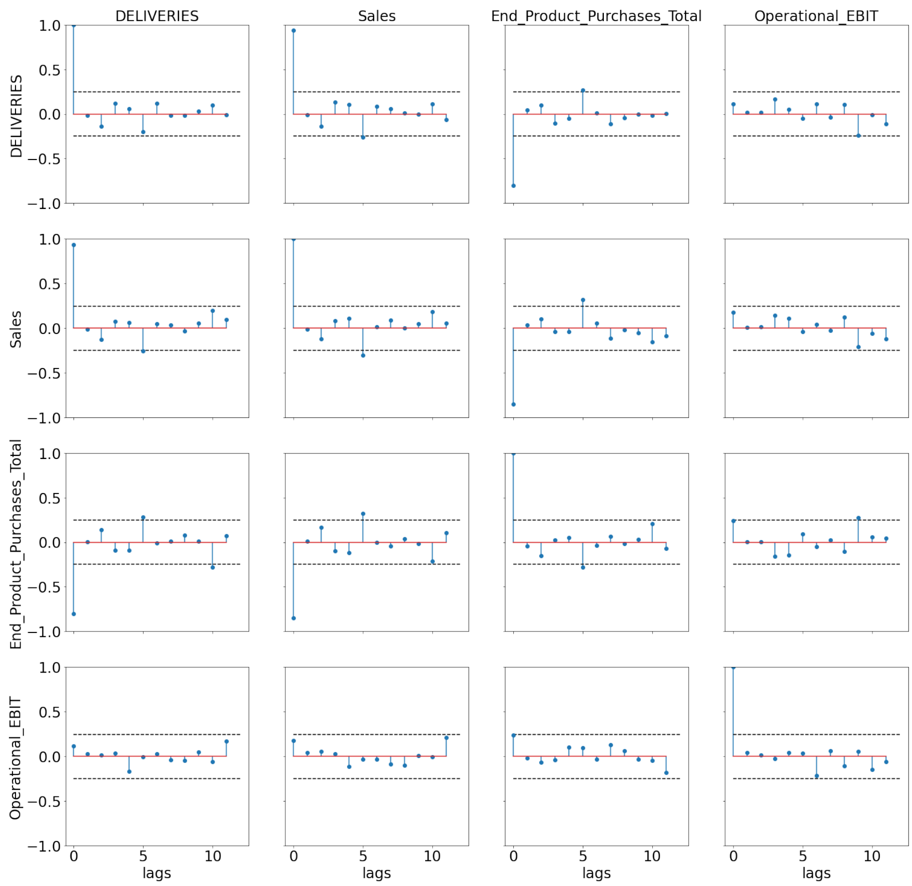


Figure 4.3: Residual correlation for all features included in the model applied to country C.

A typical result of the distribution is presented in Figure 4.4. The full result of the distribution is included in Appendix B.4. In Figure 4.4 it can be seen that for the majority of features, the residuals do follow a straight line, indicating that they follow a normal distribution. Some of them follow the red line, indicating that their distributed with unity standard deviation. The others are showing a more horizontal pattern. A horizontal tendency indicates that the distribution contains more extreme values than expected from the standard normal distribution, meaning that the standard deviation is larger than unity. This is also known as heavy tails in the normal distribution.

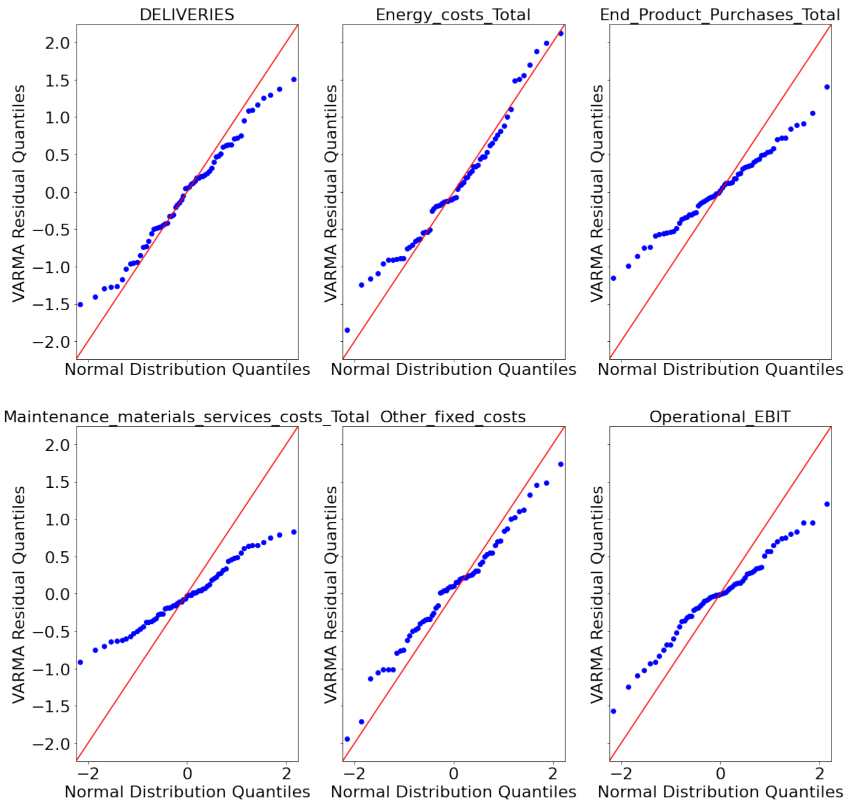


Figure 4.4: Normal distribution QQ-plots with the residuals for all features included in the model applied to region III.

5
Discussion

This chapter discuss the results, starting with investigating the different VARMA model designs that gave the lowest RMSE for each country. Then presenting suggestions on reasons behind why the VARMA model works well on certain countries, and in front of all why it does not work on the others. Taken together, the discussion will end in suggestions on how Stora Enso can continue to work with automatic financial forecasting in the future.

5.1 VARMA model design

As seen in Table 4.2 and 4.3, the feature setups including fewer features give better results for all countries, except for country E. Country E and the regions get better results when using feature setups that include more features. Since country E has one of the highest EBIT of all countries and since the regions also have high EBIT, that means that the models fitting data with higher EBIT, prefers more features. Another possible explanation can be found by looking at the results presented in both Table 4.2 and 4.3. By comparing the results with the different feature setups, it can be seen that when the VARMA model improves, the best feature setups are 1 or 2 and when the VARMA model does not improve, the best feature setups are 3,4 or 5. This could indicate that when the model struggles to perform, it favors to include more features and by that also more parameters. However, this is not true in all cases, country C and D do not improve but still has best feature setup 2 respectively 1 and region III improves while feature setup 3 gives the best results. It is clear that different countries and regions prefer different feature setups. Therefore, it could also be argued that looking at the correlations for each feature and each country separately would explain the difference. In this project, the same feature setups were applied to all countries, but an alternative approach would be to select features differently for every country and region. For example, as stated in section 1.4, the correlation between *EBIT* and *Other fixed costs* are particularly high in region III, therefore it is likely that a setup including *Other fixed costs*, performs better. From this analysis it can be concluded that the features that were stated as important from the people creating the company model (feature setup 1 and 2), also are important for the VARMA model. When the data series is more difficult to forecast, it seems like a good idea to increase the number of features considered. Though there was no feature that stood out as vital for a good prediction in all cases, but it varied between the different countries and regions.

The preferred selection criteria show a similar pattern to the feature setups. When the VARMA models improve the accuracy, they apply AIC and BIC. While the VARMA models that do not improve, apply minimizing RMSE. This fact could possibly be explained by how BIC is defined. As presented in section 2.2.3, AIC and BIC works well to find the model that is the best one in comparison with the other models tested. Though, if none of the models tested are good, AIC and BIC might favor opposite models. Therefore, when adding them together they will not be able to find the best model but end up choosing a model with an average performance. Therefore, in those cases minimizing RMSE works better.

Table 4.2 also presents the hyperparameters, as mentioned, there are no particular pattern showing up in that selection. Sometimes $p > q$, sometimes $p < q$ and in a couple of cases they are equal. It can be noted that $p = 0$ for country E, D and region I, which means that the AR par of the model is not considered, only the MA part of the model. But since country A, G and region I do not have anything in common, it is hard to tell if there is a specific reason behind that. One could argue that the hyperparameter search should be larger, for example between (0,12), since 12 is the seasonal period. Increasing the hyperparameter search would result in more inefficient training, since the time needed for each model fitting increases with the size of p and q and of course it would increase the number of model fittings needed. It is possible that larger hyperparameters could improve the results, but since the correlation gets more unreliable as the lag increases, see section 2.1.3, it would likely only increase the training time without increasing the performance.

In this analysis, the estimated model parameters have not been investigated very closely, they did not show any unexpected or extraordinary. Therefore, it is not a big problem that the parameter estimation is nonunique. Perhaps the nonuniqeness is a reason to that there is nothing to notice about the parameter settings. The nonuniqeness makes the interpretation of the results a bit more difficult, from a singular solution it would be straight forward to look at the parameter set and investigate which feature had the largest impact on the EBIT prediction. The parameter set can still give that information but it should be noted that there might be several parameters sets that can give similar results when using this method. However, the VARMA model outputs predictions for all input features, which can help to draw conclusions on which feature that affects certain peaks sudden changes. The prediction for the other features are in general not as accurate as the EBIT prediction, since the model is optimized for EBIT, but they can still be a good tool for interpretation.

5.2 Why does the VARMA model perform well on some countries and regions but not on all?

The result in Table 4.3 shows that region III and its countries F and G improve the accuracy very well and have the lowest MASE. Since F and G are very similar in how they operate, it is expected that if the model works well on one of them, it should also work well on the other. It is not necessary that it works well for the region as there might be other correlations appearing when adding countries together, but in this case it does work well.

The countries A and E also operate very similar to each other and should with the same reasoning as for F and G, either both improve the accuracy and have low MASE or both do not improve the accuracy and have high MASE. Though this is not the case, country A is found as number 2 in Table 4.3 and country E is at the bottom. That brings out the question of why the VARMA model does not perform well on country E. One possible explanation can be given by looking at Figure 4.1d, where it shows that the training data for country E expresses a unique behavior as it has a steady trend from 2014 to 2019, but then in 2019 EBIT suddenly increases to about the double amount and by that the trend completely changes. EBIT stays at a new high level for a year before it drops in 2020 and becomes very unstable throughout the forecast window. The sudden increase in 2019 is likely hard for the VARMA model to capture and model, leading to that the prediction does not work well. It can also be seen in Figure 4.1d that the prediction from the VARMA model is very similar to the peak in 2019, indicating that the seasonal component had a high influence on the prediction. An additional explanation to the discrepant behaviour between country A and E could be that they operate in different currencies and that the exchange rate has influenced the result. On another note, the company model is not able to capture the peaks of the observed values either. However, the company model is taking a safer approach and stays at an average value. Therefore, it receives a good RMSE score, but the MASE for the company model is relatively high compared to the other countries. By that, country E was particularly difficult to model due to the unstable input data. In addition, when being aware of these difficulties it does not come as a surprise that the region including country E, namely region II, is difficult for the VARMA model to perform on.

The VARMA model in general shows a good capability to capture peaks and sudden drops. Sometimes it pays off, as in Figure 4.1a where the VARMA model very well captures the large drop appearing in middle of 2019. While as for country E, the VARMA model makes peaks in the opposite direction compared to the observed value, as seen in Figure 4.1d, which immediate results in a high RMSE for the whole prediction, as discussed above. It is possible that the VARMA model tendency to have a high variation comes from using differencing as seasonal adjustment method. As mentioned in section 2.1.1 there are other methods for seasonal adjusment, but these would likely lead to a prediction that is more stable around a mean value.

The countries of region I show both improvements of RMSE with low MASE and increased RMSE with high MASE. In region I, this behavior is not unexpected since there are differences between the countries. Country B is the only country in region I that has a corrugated site while countries C and D do not. It is therefore not strange that the model works better for country B than for countries C and D. Looking closer at the result in Figure 4.1c, it can be seen that EBIT for country D in the forecast window drastically changes from the previous trend in the training data, which further explains why the VARMA model fails in its prediction. Looking at the MASE for the company model, country D is also the most difficult one to predict with the company model. Therefore, the improvement for country D is better than the one for country E, even though MASE for the VARMA model applied to country D is outstanding high.

The result from residual analysis in general shows that the VARMA model is able to model the correlations between the features. This conclusion can be drawn since in Figure 4.3, there are not many lags where there are significant correlation. The same can also be seen from the remaining residual correlation plots in Appendix B. In Figure 4.4 it is found that the assumption of normally distributed residuals is met for most variables, though there are variables where the condition is not met. That could indicate that other models or repeated modelling of the residuals would give better results.

5.3 Suggestions for future work

As mentioned in the introduction chapter, this master's thesis is an initial attempt to digitalize the forecasting process at Stora Enso, meaning that their work improving the method is not finished here. Therefore, suggestions on what to bring into their future work is presented below.

5.3.1 Data collection

The data was one of the major limitations in this project. Increasing the amount of data is a key to improved results, regardless of choice of method. If it is possible to recover data from before 2014, that would be one possible solution. Otherwise, carefully continue to store the monthly data is important. If a similar event occurs as in 2014, where old data was not comparable, it should be considered how to transfer data into the new format to not be forced to waste the data.

It was discussed to collect data on a weekly or even daily basis, to be able to collect more data faster. However, that is not likely to improve the results for this type of forecast, as the forecast will predict on the scale that is used by the input meaning that for a forecast for 15 months will need monthly data. Also, there are several variables in the input data that it does not make sense to collect more frequently since they are monthly costs.

5.3.2 Preprocessing and feature selection

Regarding the preprocessing, there is one possible improvement that should be tried. In this project only the seasonal difference was adjusted for without further differencing to save more data for training. However, as explained in the Method section 3.1, not all data was stationary after the preprocessing, see Table 4.1. Since VARMA assumes the data to be stationary, it would likely perform better if all data were stationary. Applying differencing with lag equal to unity once or twice usually gives stationary data and that could improve the results.

Another way to improve the results could be to approach each country separately. In this study, all countries and regions were tested on the same features, but as the results show, there are differences between the countries and regions that affect the result. Treating each country would likely lead to different feature setups for some countries, which could improve the forecasts. It should also show improvement in the residual correlation graphs, where there would be less correlation left as the model could capture more if the exact right features are included. However, this will be more time consuming.

It is also possible to treat the data for each input variable separately. During the preprocessing in this project, they were all seasonally adjusted by using the same period, but it is for example known that the raw material costs has a longer period. The issue that is raised if different seasonal periods are considered, is that the resulting data after adjustment would be of different length. As mentioned in 2.1.1, for seasonal period d, the first d data points in the series will not be able to use in the training since they cannot be transformed. A longer period would mean that more data is needed for the adjustment, and less data can be used for training. Since the data volume is already low, it might then be a good idea to consider other seasonal adjustment methods, though the other methods available are usually more complex and time inefficient. Considering the cons mentioned, it is possible that the accuracy would not improve by treating each input variable separately, but it is an approach that could be tested.

5.3.3 Model selection and evaluation

The result provided by the VARMA model do fulfill the aim of being more accurate and more efficient than the manual way for some of the countries and for region III. However, not in all cases. The changes mentioned above could improve the result further so that should be tested firsthand. It could also be interesting to try fit the VARMA model on all the countries simultaneously. That of course is more computationally heavy, but on the other hand it will only need one run to receive results for all countries. If there are correlation between countries, running all countries simultaneously would likely bring good results.

If none of the suggested measures work to improve the results further, another model approach could be tested. In that case, a Random Forest model would be interesting to investigate. It should work with the amount of data provided and it is not too complicated to still give an easy interpretable result.

5.3.4 Application and usage

A challenging part during this master's thesis project was to account for the high number of different countries and regions during each process. It could therefore be easier to select one country to focus on and create a model that works very well for that country, then move on to the next one and so on. That would not be as efficient but would most certainly give better results, as this study shows that there are differences between the countries that affect the results. Here it could also be considered if it can be expected to get improved results from an automatic model on all countries and regions included. The countries that do not improve their accuracy using an automatic model show an unstable trend which make the forecasting more difficult, both for the VARMA model and for the company model. With the results received here, it seems like a good idea to create an automatic model for the more stable countries in region III and for country A. Doing that could release time to for example put into forecasting the unstable countries and regions, resulting in improved forecasting overall.

Finally, it is interesting to investigate how the model could be used. This study focused on creating a model that would replace the current way of doing the 15-months rolling forecasting, but as seen in Figure 4.2, the automatic model outperforms the company model at the end of the forecast, but is not as good during the months that are near in time. Therefore, it could be interesting to try the automatic approach for the annual budget or the 10-year financial frame, since they focus more on the long-term perspective. Alternatively, the VARMA model could be adjusted to put more weight on minimizing the deviation in the nearer time frame. The method here puts equal weights to minimizing the error through the whole forecast window, but that could be adjusted. The VARMA model allows for weighting any lag which makes it possible to optimize the prediction for a selected time frame.

6
Conclusion

Taken together, this master's thesis shows that there are good conditions for creating an automatic 15-months rolling forecast for the Packaging Solutions division at Stora Enso. In countries A,B F and G and in region III the VARMA model works well with the preprocessing and feature setups used and could be implemented for a more efficient and accurate forecasting. In the improved countries, the VARMA model is able to predict peaks and sudden changes, and is also easy to interpret thanks to prediction of all input features. For the remaining countries and regions, changes such as increased preprocessing, more data or another model selection, needs to be made before an automatic model could be useful. However, it is suggested to begin looking at implementing the automatic forecasting in the more stable countries, before proceeding to the countries and regions that show larger variance in their EBIT.

Bibliography

1. Hyndman, R. J. & Athanasopoulos, G. *Forecasting: principles and practice* (OTexts, 2018).
2. Hall, M. *Digital Equipment Corporation* Encyclopedia Britannica. Accessed 14 March 2022. [Online].
3. Sezer, O. B., Gudelek, M. U. & Ozbayoglu, A. M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing* **90,** 106181 (2020).
4. Eriksson, E. Personal communication. Jan. 2022.
5. Taylor, S. J. *Modelling financial time series* (world scientific, 2008).
6. Radha, S. & Thenmozhi, M. *Forecasting short term interest rates using ARMA, ARMA-GARCH and ARMA-EGARCH models* in *Indian Institute of Capital Markets 9th Capital Markets Conference Paper* (2006).
7. Huang, W., Nakamori, Y. & Wang, S.-Y. Forecasting stock market movement direction with support vector machine. *Computers & operations research* **32,** 2513–2522 (2005).
8. Wasserbacher, H. & Spindler, M. Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. *Digital Finance,* 1–26 (2021).
9. Chen, Y., He, K. & Tso, G. K. Forecasting crude oil prices: a deep learning based model. *Procedia computer science* **122,** 300–307 (2017).
10. Brockwell, P. J. & Davis, R. A. *Introduction to time series and forecasting* (Springer, 2016).
11. Wan, R., Mei, S., Wang, J., Liu, M. & Yang, F. Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting. *Electronics* **8,** 876 (2019).
12. Ediger, V. Ş. & Akar, S. ARIMA forecasting of primary energy demand by fuel in Turkey. *Energy policy* **35,** 1701–1708 (2007).
13. Sapankevych, N. I. & Sankar, R. Time series prediction using support vector machines: a survey. *IEEE computational intelligence magazine* **4,** 24–38 (2009).
14. Hamidi, O., Tapak, L., Abbasi, H. & Maryanaji, Z. Application of random forest time series, support vector regression and multivariate adaptive regression splines models in prediction of snowfall (a case study of Alvand in the middle Zagros, Iran). *Theoretical and Applied Climatology* **134,** 769–776 (2018).
15. Lima, C. H. & Lall, U. Climate informed monthly streamflow forecasts for the Brazilian hydropower network using a periodic ridge regression model. *Journal of hydrology* **380,** 438–449 (2010).

16. Li, J. & Chen, W. Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting* **30,** 996–1015 (2014).
17. Xu, Z., Tu, W.-W. & Guyon, I. *AutoML Meets Time Series Regression Design and Analysis of the AutoSeries Challenge* in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2021), 36–51.
18. Javeri, I. Y., Toutiaee, M., Arpinar, I. B., Miller, J. A. & Miller, T. W. *Improving Neural Networks for Time-Series Forecasting using Data Augmentation and AutoML* in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)* (2021), 1–8.
19. Othan, D., Kilimci, Z. H. & Uysal, M. *Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models* in *International Conference on Innovative and Intelligent Technologies (ICIIT-19), Istanbul, Turkey* (2019).
20. Kruglov, A. Personal communication. Jan. 2022.
21. Abolina, B. Personal communication. Jan. 2022.
22. Claramunt, C. Personal communication. Jan. 2022.
23. Ekberg, D. Personal communication. Jan. 2022.
24. Von Hofsten, T. Personal communication. Jan. 2022.
25. Benlaieb, M. Personal communication. Feb. 2022.
26. Damrongkulkamjorn, P. & Churueang, P. *Monthly energy forecasting using decomposition method with application of seasonal ARIMA* in *2005 International Power Engineering Conference* (2005), 1–229.
27. Koprinska, I., Rana, M. & Agelidis, V. G. Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems* **82,** 29–40 (2015).
28. Yu, L. & Liu, H. *Feature selection for high-dimensional data: A fast correlation-based filter solution* in *Proceedings of the 20th international conference on machine learning (ICML-03)* (2003), 856–863.
29. Lütkepohl, H. *New introduction to multiple time series analysis* (Springer Science & Business Media, 2005).
30. Acquah, H. D.-G. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics* **2,** 001–006 (2010).
31. Kuha, J. AIC and BIC: Comparisons of assumptions and performance. *Sociological methods & research* **33,** 188–229 (2004).

A

Appendix 1: Data file compilation

In this appendix the details of reconstructing the input data files are presented for the purpose of being able to recreate the results from the raw data.

All data was collected from Packaging solutions rolling forecasting files. It was received in excel files, one per year starting from 2015 to 2021. The file from 2015 also contained data from 2014, giving in total, data reaching from 2014-01 to 2021-12. Each file contained a lot more information than the data that was necessary for implementation, for example summations of features and old features that are not considered anymore. There were also a few features that was not specified for every country and together with the irrelevant features, these were not transferred to the csv-files. An exception was the feature *Price*, which for example was not specified for some countries, but could be calculated by dividing the feature *Sales* with *Deliveries*. New files were created with only the monthly data for the features presented in Table 1.1. This time one file was created per country or region and stored in csv-format to facilitate importation to the program.

In addition to the files of observed data, csv-files with data for the existing forecast were created. Similar to the files with observed data, one file per country or region was created, containing the forecast created in 2020-09. The forecast from 2020-10 reaches over 15 months, meaning that the final month in the forecast is 2021-12, which corresponds to the final month of observed data. Country C however, did not have a 15 month forecast in 2020-09, but only one reaching to 2020-12. Therefore, the forecast file with the existing model for county C was created by combining the forecast from 2020-19 with the one created in 2020-12.

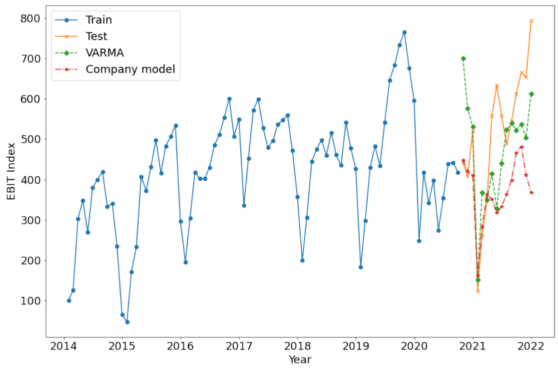
B

Appendix 2: Complete results collection

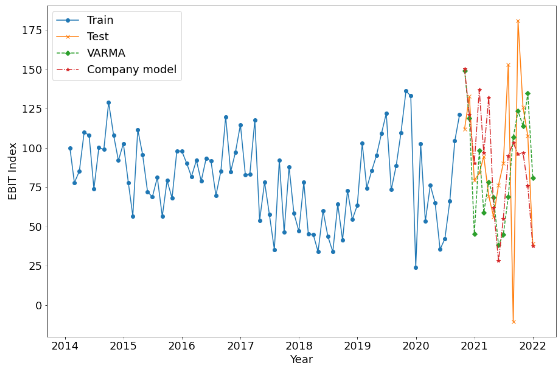
Since the Results chapter only highlights a few countries and regions, this appendix provides the complete collection of results for all countries and regions. First all EBIT forecasts are presented, followed by the comparison figures of the deviation between the true value and the VARMA model prediction compared to the deviation between the true value and the company model prediction. Then the residual correlation plots are presented, before finishing with the residual QQ-plots.

B.1 EBIT forecasts for all countries and regions

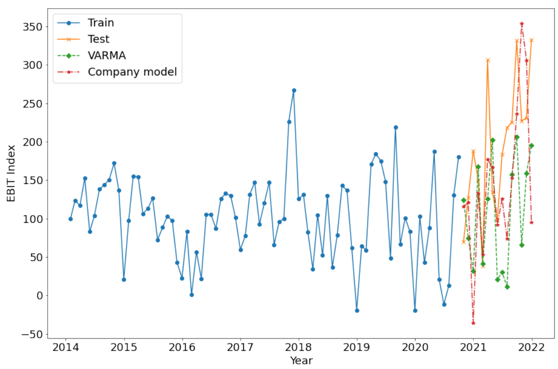
Figure B.1 presents the EBIT forecasts from the VARMA model and the company model.



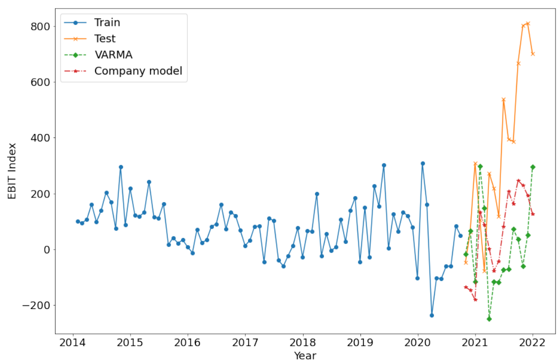
(a) EBIT forecast for country A.



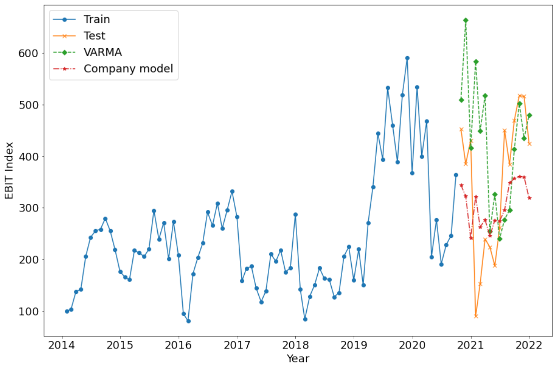
(b) EBIT forecast for country B.



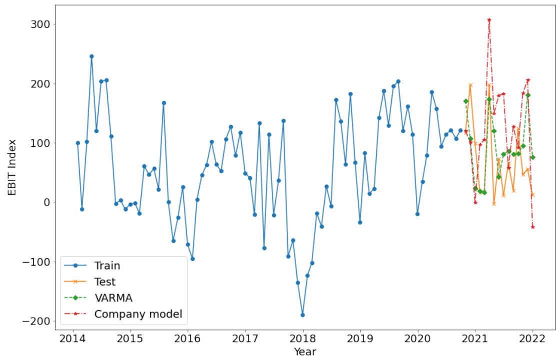
(c) EBIT forecast for country C.



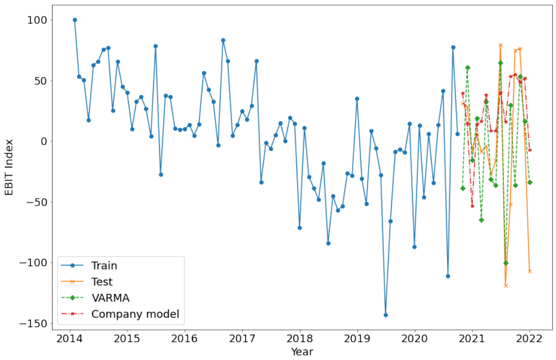
(d) EBIT forecast for country D.



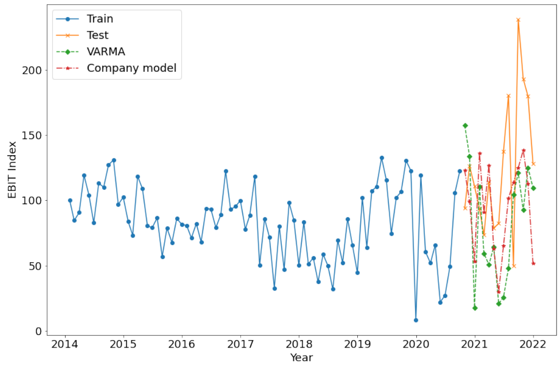
(e) EBIT forecast for country E.



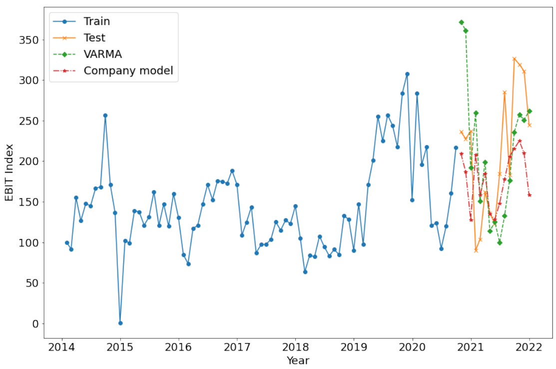
(f) EBIT forecast for country F.



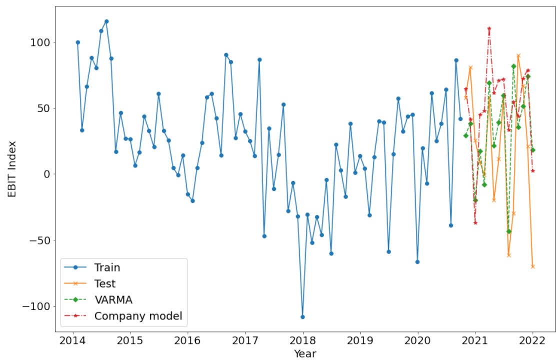
(g) EBIT forecast for country G.



(h) EBIT forecast for region I.



(i) EBIT forecast for region II.

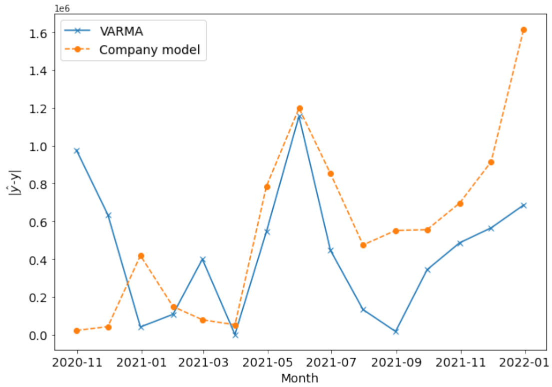


(j) EBIT forecast for region III.

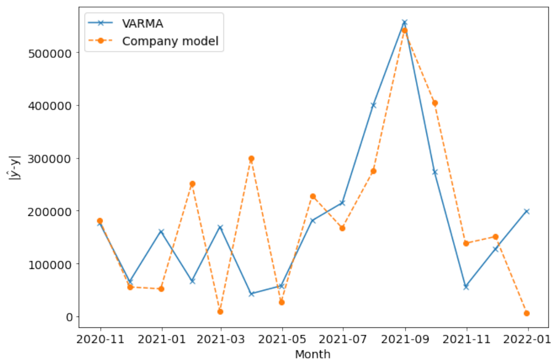
Figure B.1: EBIT forecasts for all countries and regions.

B.2 Deviation comparison between the VARMA model and the company model

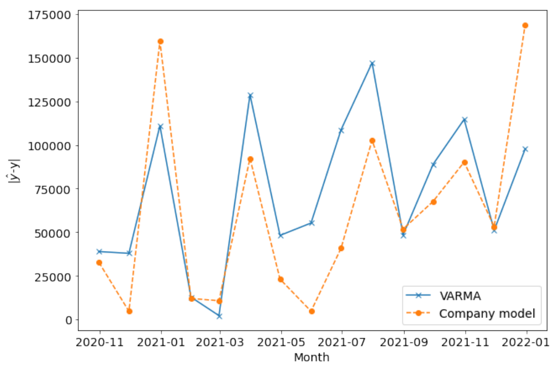
Figure B.2 presents the deviation between the VARMA model and the observed value compared to the deviation between the company model and the observed value.



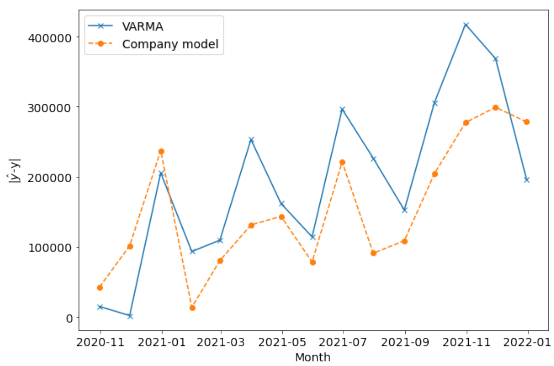
(a) $|\hat{y} - y|$ for country A.



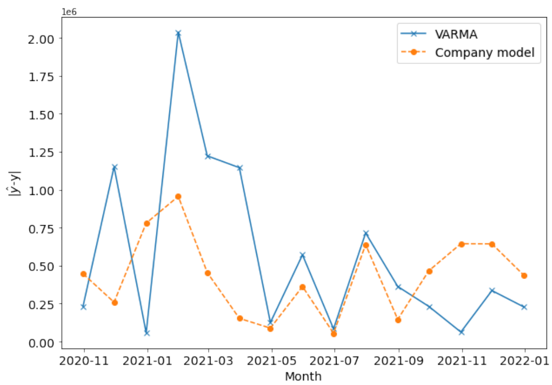
(b) $|\hat{y} - y|$ for country B.



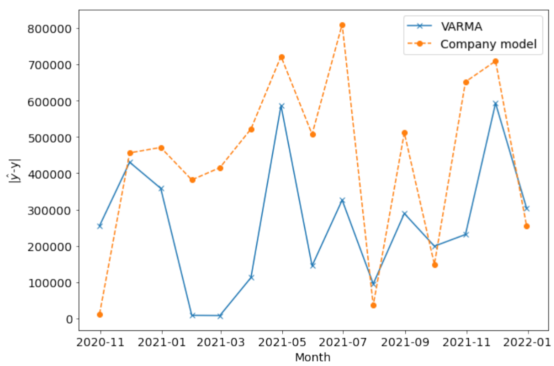
(c) $|\hat{y} - y|$ for country C.



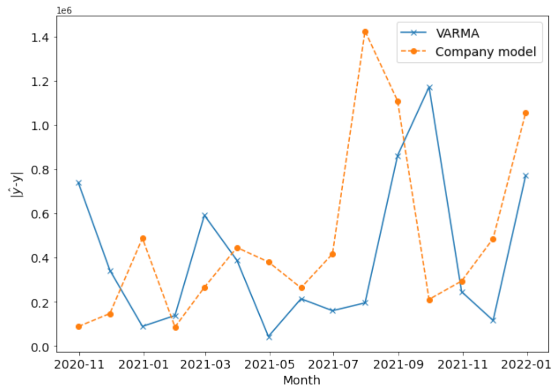
(d) $|\hat{y} - y|$ for country D.



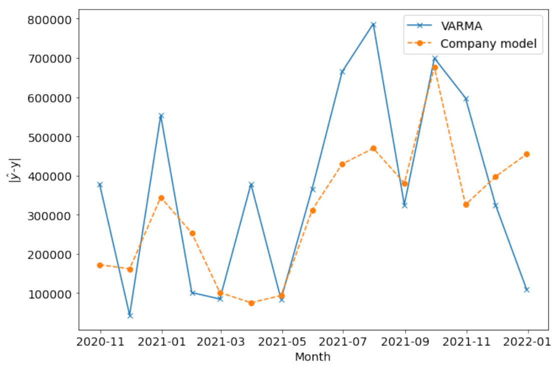
(e) $|\hat{y} - y|$ for country E.



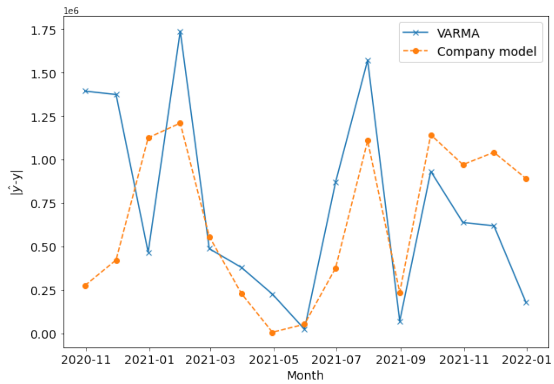
(f) $|\hat{y} - y|$ for country F.



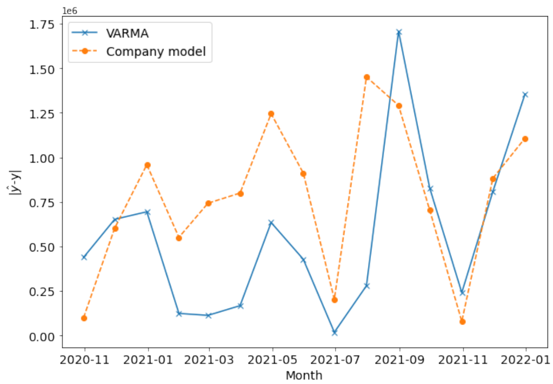
(g) $|\hat{y} - y|$ for country G.



(h) $|\hat{y} - y|$ for region I.



(i) $|\hat{y} - y|$ for region II.

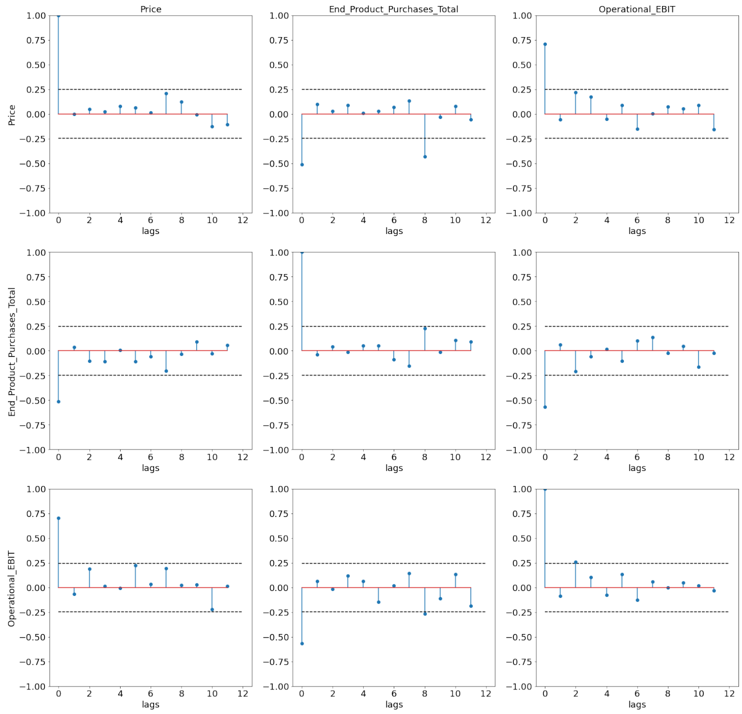


(j) $|\hat{y} - y|$ for region III.

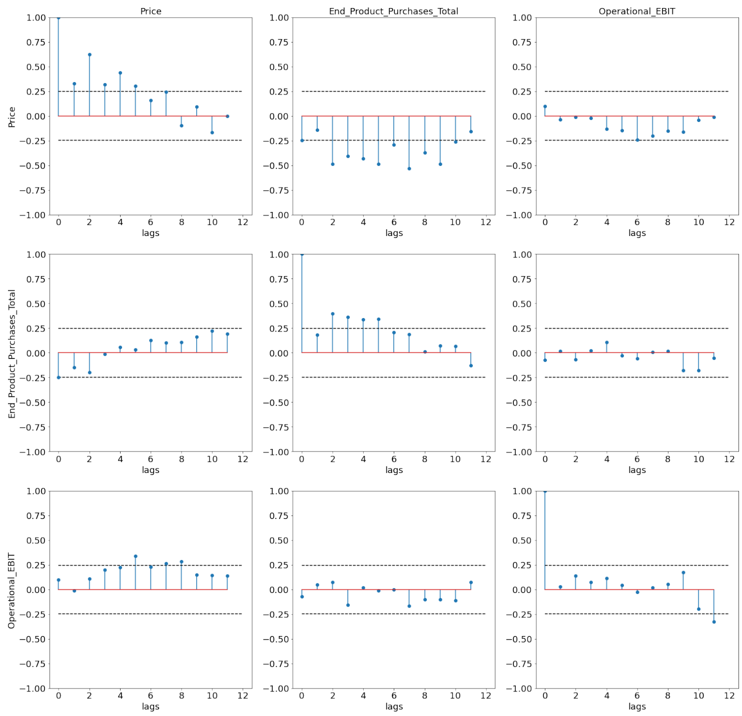
Figure B.2: Difference between the true value and the predictions from the VARMA model and the company model for all countries and regions.

B.3 Residual correlation analysis

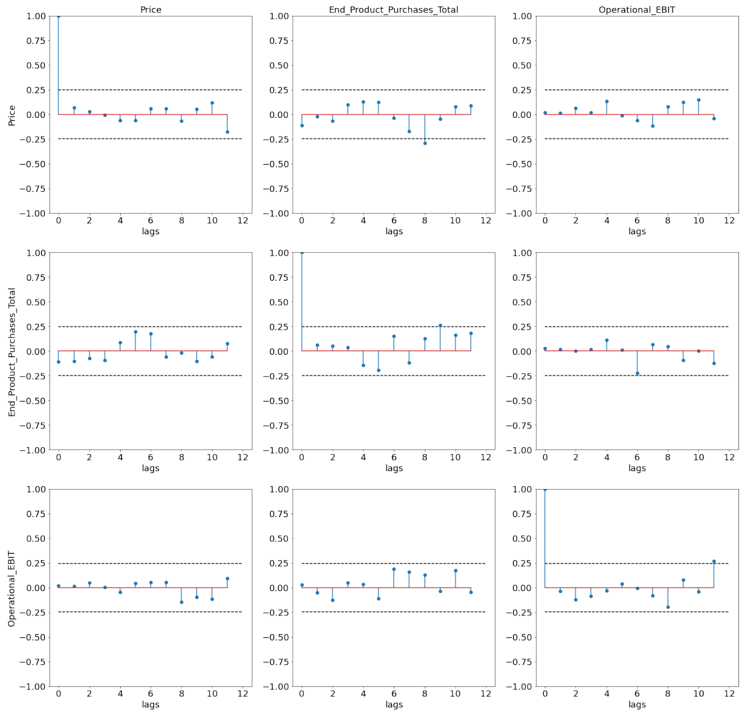
In figures B.3, B.4, B.5 and B.6 the correlations between the residuals from the VARMA model are found.



(a) Residual correlation for country A.

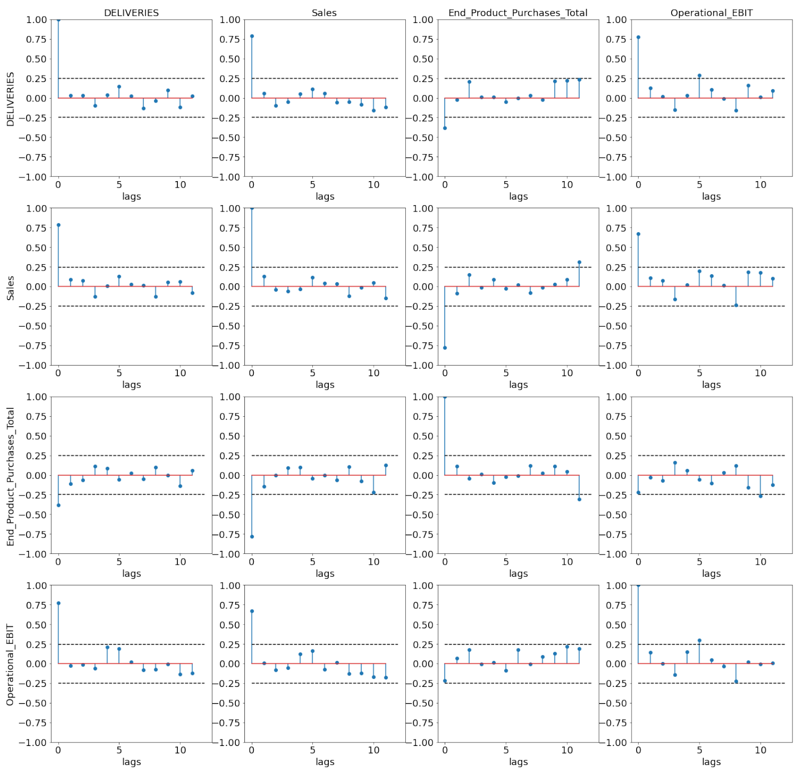


(b) Residual correlation for country D.



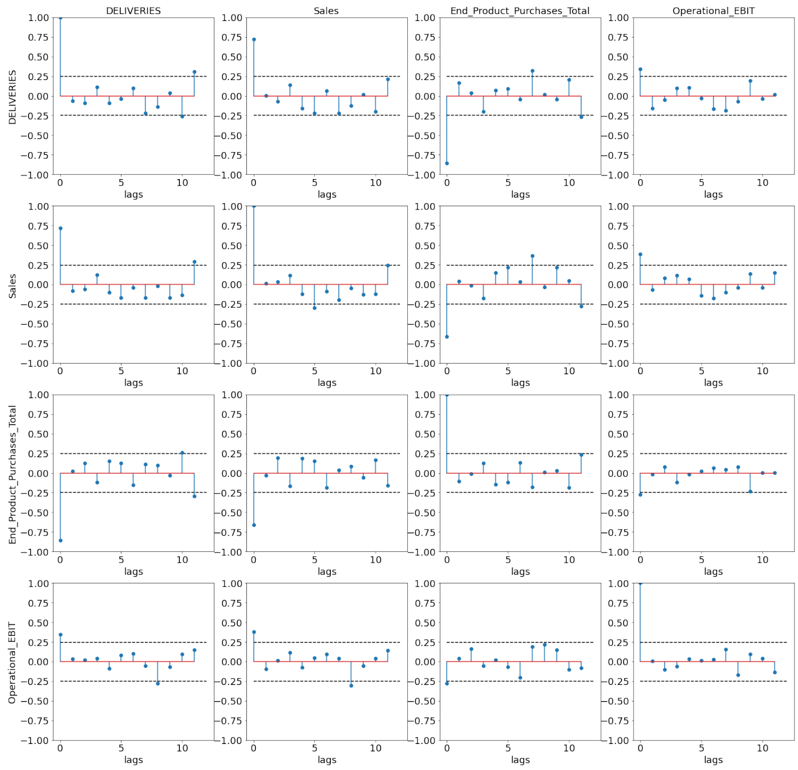
(c) Residual correlation for country F.

Figure B.3: Residual correlation for the countries applying feature setup 1.



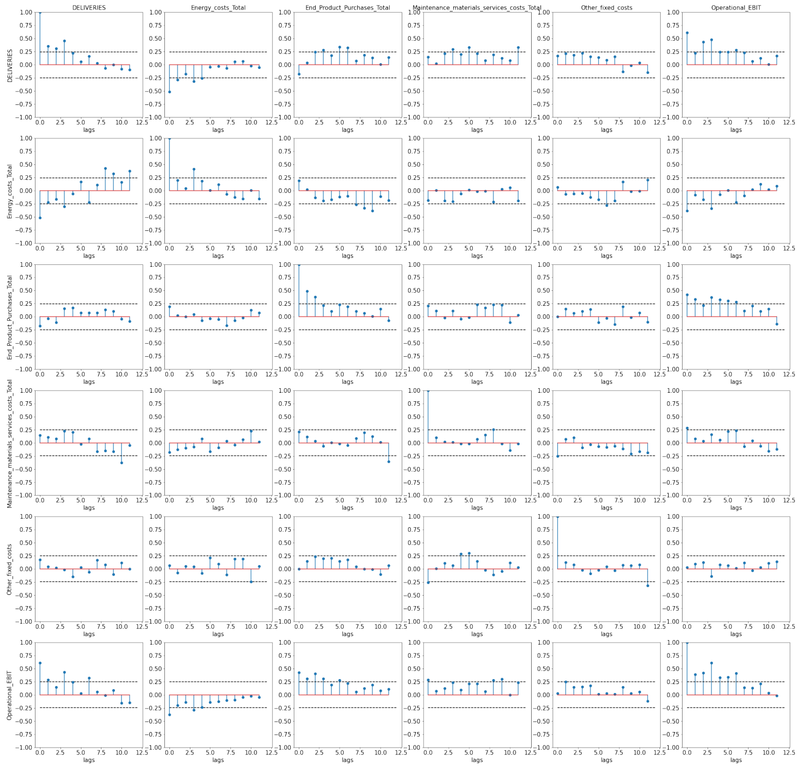
(a) Residual correlation for country B.

(b) Residual correlation for country C.

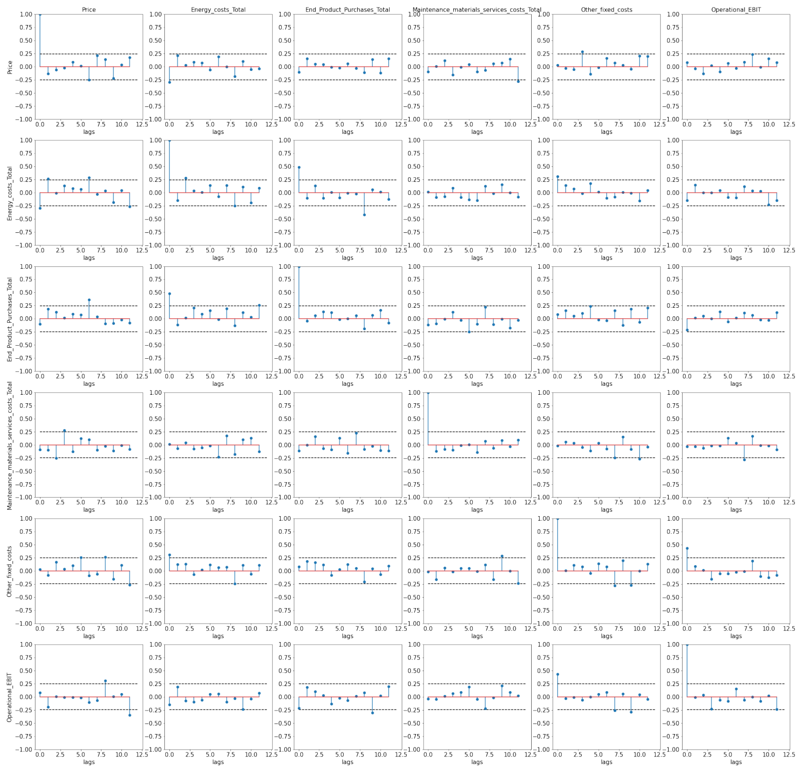


(c) Residual correlation for country G.

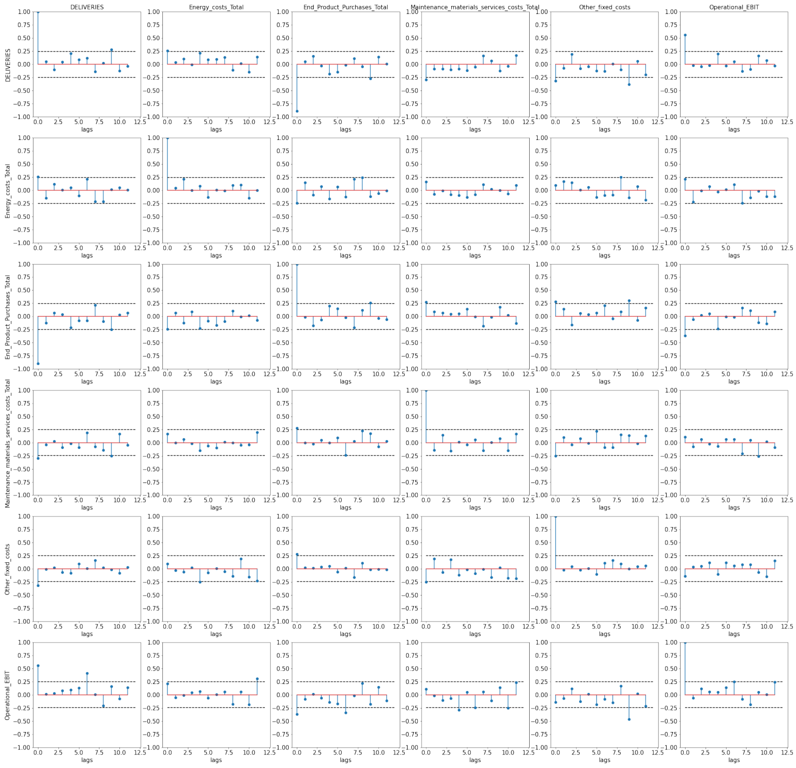
Figure B.4: Residual correlation for the countries applying feature setup 2.



(a) Residual correlation for country E.



(b) Residual correlation for region II.



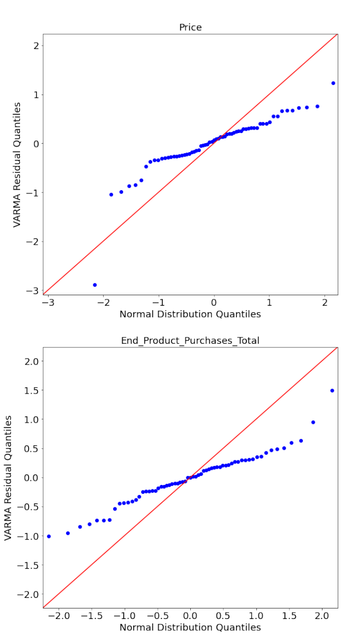
(c) Residual correlation for region III.

Figure B.5: Residual correlation for the countries applying feature setup 3 and 5.

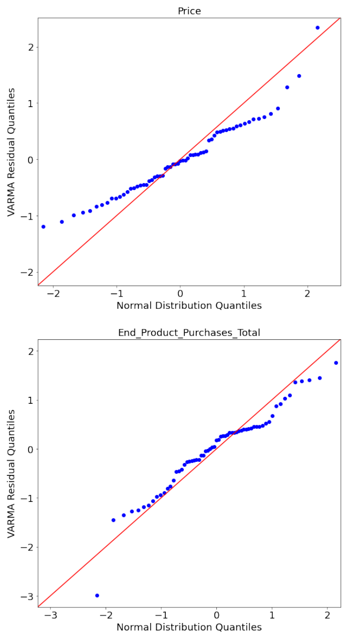
Figure B.6: Residual correlation for region II, applying feature setup 4.

B.4 QQ-plots

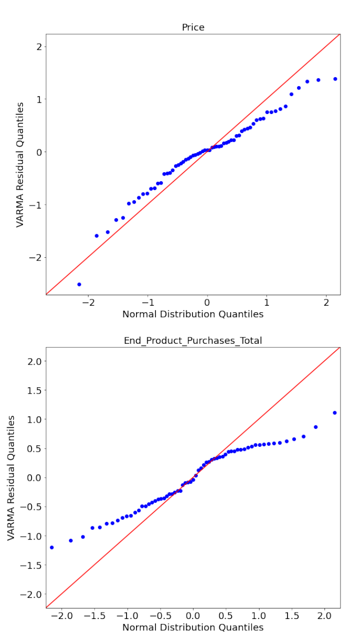
In figures B.7, B.8 and B.9 the QQ-plots for the residuals can be seen.



(a) QQ-plots for country A.

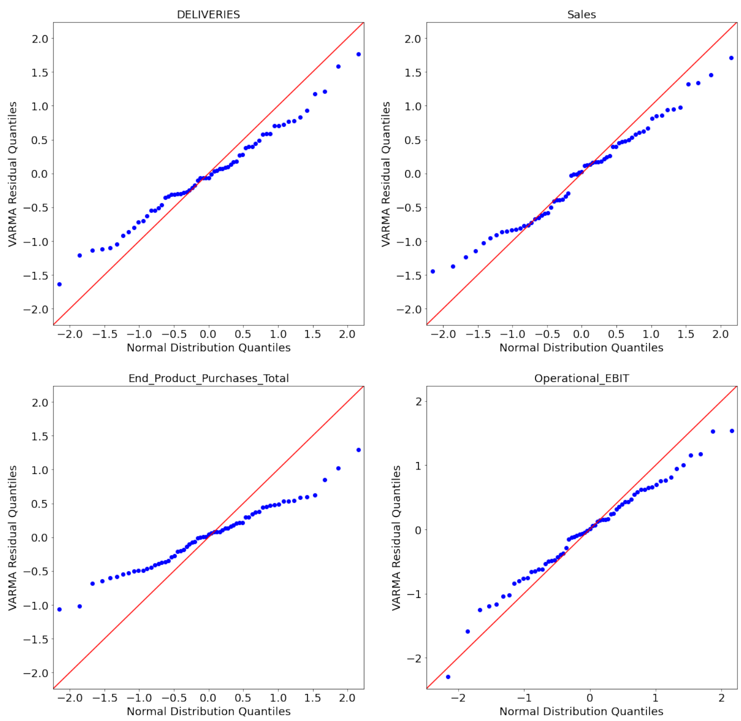


(b) QQ-plots for country D.

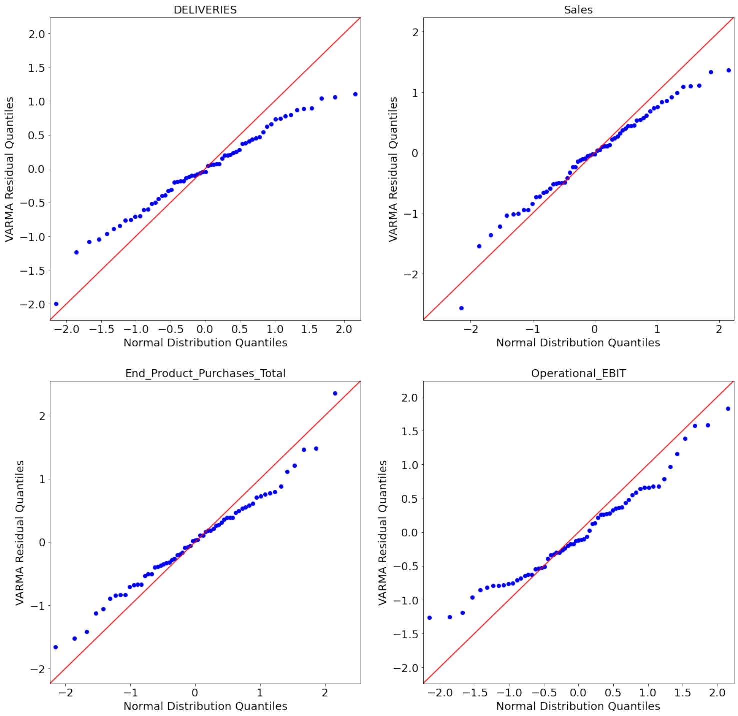


(c) QQ-plots for country F.

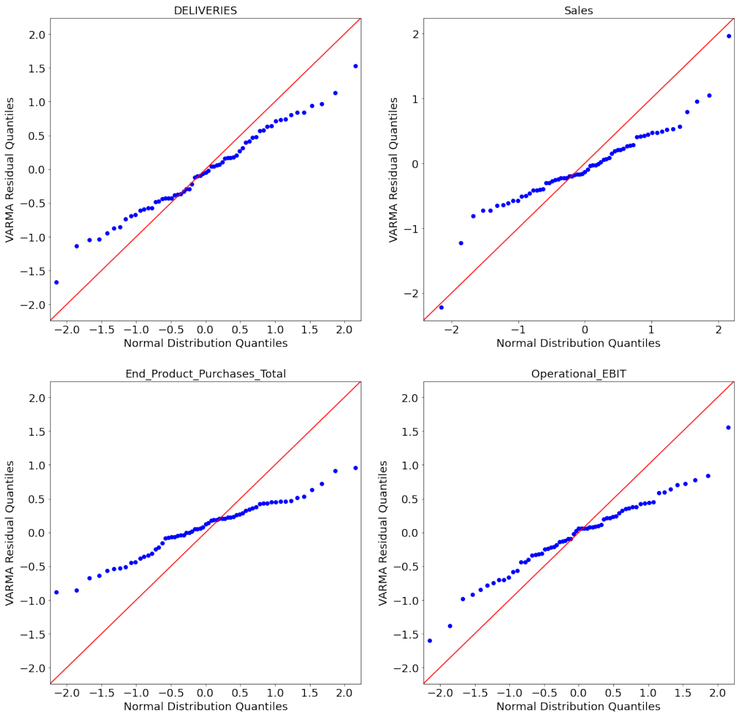
Figure B.7: QQ-plots for countries applying feature setup 1



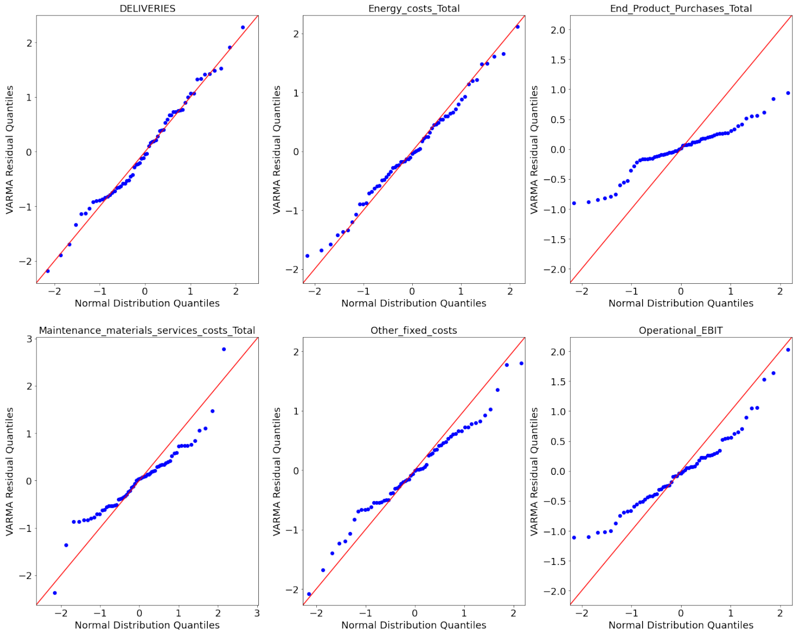
(a) QQ-plots for country B.



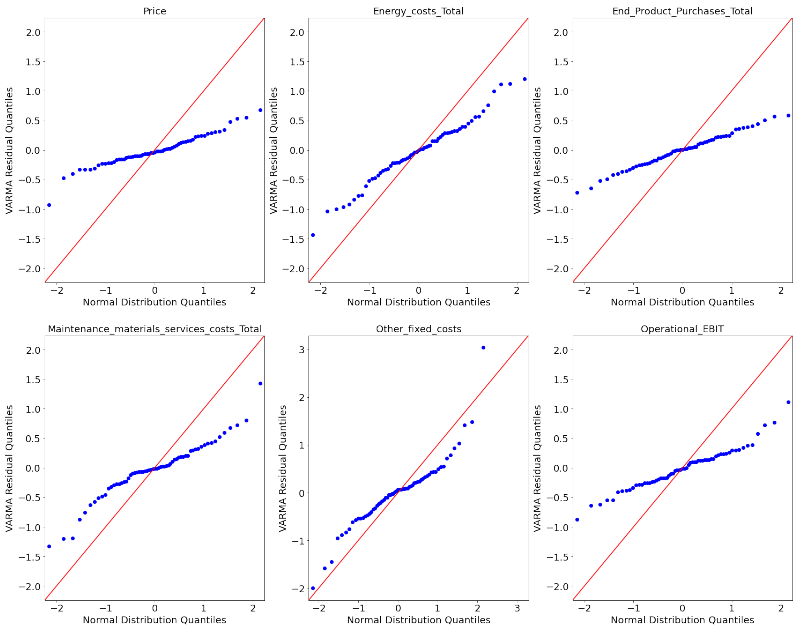
(b) QQ-plots for country C.



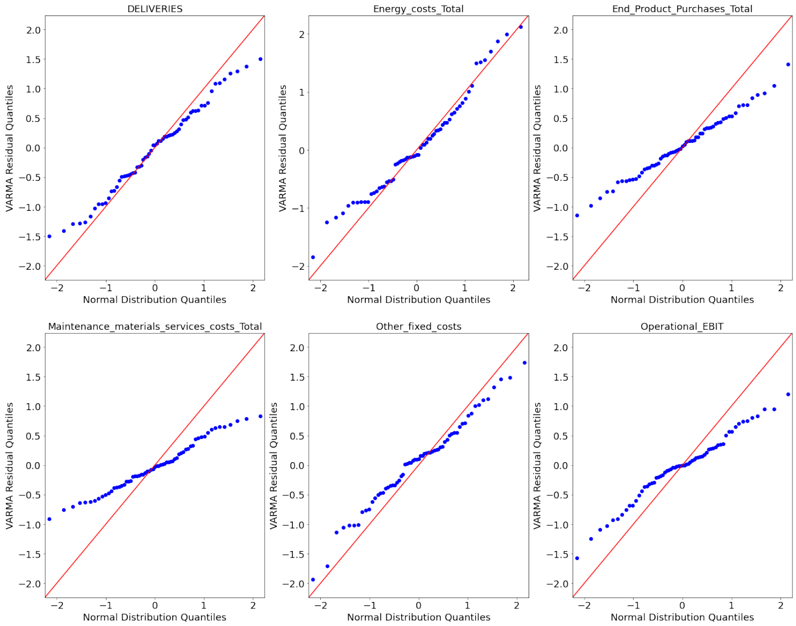
(c) QQ-plots for country G.



(d) QQ-plots for country E.



(e) QQ-plots for region II.



(f) QQ-plots for region III.

Figure B.8: QQ-plots for countries and regions applying feature setup 2,3 and 5.

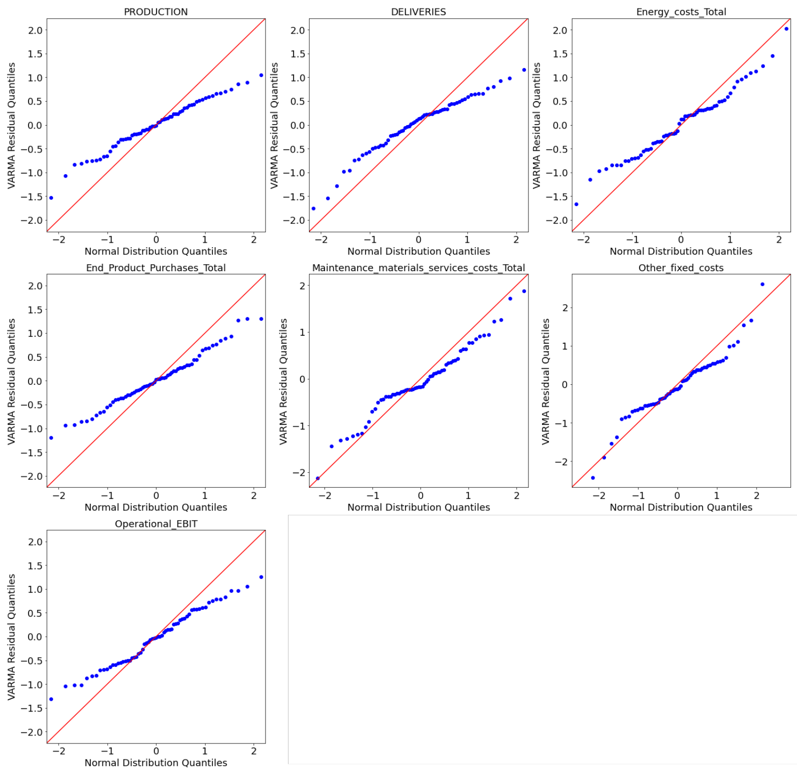


Figure B.9: QQ-plots for region III, applying feature setup 4

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

AVANCEZ 1829

CHALMERS
UNIVERSITY OF TECHNOLOGY