Pricing of Liquefied Petroleum Gas in North-West Europe

Daniel Engström

Master's thesis
2014:E44

Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematical Statistics

CENTRUM SCIENTIARUM MATHEMATICARUM

TYP AV DOKUMENT ☒ Examensarbete ☐ Delrapport	☐ Kompendium ☐ Rapport	**DOKUMENTBETECKNING** **LUTFMS—3259—2014**

INSTITUTION

Matematikcentrum. Matematisk statistik, Lunds universitet, Box 118, 221 00 LUND

FÖRFATTARE

Daniel Engström

DOKUMENTTITEL OCH UNDERTITEL

Pricing of Liquefied Petroleum Gas in North-West Europe

SAMMANFATTNING

Liquefied Petroleum Gas (LPG) is a flammable mixture of hydrocarbon gases, mainly propane and butane, used for various heating purposes and as vehicle fuel. This thesis focuses on examining the LPG market, evaluating a couple of driving factor hypotheses for the propane price and developing a model for forecasting of future propane prices in North-West Europe, both on daily and monthly horizons. It is shown that especially two factors, crude oil (brent) and naphtha, which is a light crude oil distillate produced when rening crude oil, have strong relation to propane and may aff
ect the propane price. Unfortunately, no good model to forecast the daily propane price was developed. For the monthly average price, the proposed models perform badly when forecasting the actual price, but one of the models, an AR(12) model, which can forecast the direction of propane price movements one and two months forward in time is presented. The AR(12) model is able to forecast the correct price movement direction with an accuracy of over 70%. This result is good, and shows that the AR(12) model is a useful tool in the LPG trading.

NYCKELORD

DOKUMENTTITEL OCH UNDERTITEL - SVENSK ÖVERSÄTTNING AV UTLÄNDSK ORIGINALTITEL

UTGIVNINGSDATUM år 2014 | mån 8	**ANTAL SID**	**SPRÅK** ☐ svenska ☒ engelska ☐ annat

ÖVRIGA BIBLIOGRAFISKA UPPGIFTER	**ISSN**
	ISBN
	2014:E44

Signature ______________________ Date 3/9 - 2014

Abstract

Liquefied Petroleum Gas (LPG) is a flammable mixture of hydrocarbon gases, mainly propane and butane, used for various heating purposes and as vehicle fuel. This thesis focuses on examining the LPG market, evaluating a couple of driving factor hypotheses for the propane price and developing a model for forecasting of future propane prices in North-West Europe, both on daily and monthly horizons. It is shown that especially two factors, crude oil (brent) and naphtha, which is a light crude oil distillate produced when refining crude oil, have strong relation to propane and may affect the propane price. Unfortunately, no good model to forecast the daily propane price was developed. For the monthly average price, the proposed models perform badly when forecasting the actual price, but one of the models, an AR(12) model, which can forecast the direction of propane price movements one and two months forward in time is presented. The AR(12) model is able to forecast the correct price movement direction with an accuracy of over 70%. This result is good, and shows that the AR(12) model is a useful tool in the LPG trading.

Acknowledgements

I would like to thank my supervisor Daniel Hjortström and Lotta Grinneland, for their help and guidance throughout the work, and for giving me the opportunity to write the thesis. My examiner at LTH, Johan Lindström also deserves a special thanks for his input and comments on the work. Finally, I would like to thank my family and friends for all the support.

Daniel Engström
Malmö, July 2014

Abbreviations and Symbols

AIC	-	Akaike's Information Criteria
AR	-	AutoRegressive
ARA	-	Amsterdam, Rotterdam and Antwerpen region
ARMA	-	AutoRegressive Moving Average
ARMAX	-	AutoRegressive Moving Average with Exogenous Inputs
BIC	-	Bayesian Information Criteria
CIF	-	Cost Insurance Freight
EUR/MWh	-	Euros per Megawatt Hour
FOB	-	Free On Board
GARCH	-	Generalized Autoregressive Conditional Heteroskedasticity
LNG	-	Liquefied Natural Gas
LPG	-	Liquefied Petroleum Gas
MSE	-	Mean Squared Error
NWE	-	North-West Europe
OTC	-	Over the Counter
RMSE	-	Root Mean Squared Error
USD	-	U.S. Dollars
USD/bbl	-	U.S. Dollars per barrel
USD/t	-	U.S. Dollars per tonne
WSS	-	Wide (or Weak) Sense Stationary
$\mathbb{E}\left(X\right)$	-	Expected value of X
$\mathbb{V}\left(X\right)$	-	Variance of X
$\mathbb{C}\left(X,Y\right)$	-	Covariance between X and Y
$\mathcal{N}\left(\mu,\sigma^2\right)$	-	Normal distribution with expected value μ and variance σ^2

Contents

List of Figures

List of Tables

1 Introduction

Liquefied Petroleum Gas (LPG) is a flammable mixture of hydrocarbon gases, mainly propane and butane, used for various heating purposes and as vehicle fuel. LPG is a bi-product from oil refining and natural gas processing. Even though LPG is a relatively small energy source, over 240 million tons are consumed worldwide each year (The World LP Gas Association, 2010). Over 73 million tons tons are traded each year with a potential trading value of over 50 billion USD. The size of individual LPG trades vary from a few tons to over 45000 tons. Companies that trade LPG are naturally interested in making the best possible trades. Some companies may also want to secure their positions to limit their risk exposure or to buy/sell LPG at a pre-determined price. This can be done using financial instruments, mainly financial SWAP contracts, which let the two parts of the contract swap their cash flows with each other at some point in the future. To be able to make good trades and to secure favourable swap and forward prices, knowledge of factors that affect the LPG price is of great interest.

One problem with the LPG trading is the absence of a ”true” marketplace where trades are done. Instead all trading agreements, both the physical and the financial contracts, are made over the counter (OTC), most often using a broker. This makes the trading of LPG different from crude oil and natural gas trading, where an organized market exists. In the absence of an exchange market, several pricing institutes present market prices of LPG in different regions based on some formula and interviews with the companies and brokers that trade LPG. These listed prices indicate the current price level, but does not provide a guaranteed price for trading. To complicate matters further, the listed price often vary among different institutes. The bid-ask spread is usually big in the listed prices and on trading days where no physical trading takes place strange price jumps with no obvious logical explanation are common. In addition the LPG price is very volatile, with no typical patterns, making it hard to predict future prices.

The price indicators given by the pricing institutes are different among the institutes and uncertain. The high values that are involved in the LPG trading generates good trading opportunities for players that can predict future price movements. Because of this, it is important for a company to know what affects the price and how to take advantage of this knowledge in its trading strategies.

1.1 Aim and Limitations

This thesis focuses on examining the LPG market and analysing a number of hypotheses regarding driving factors of the LPG price. These hypotheses will be evaluated using statistical analysis. Hopefully this will lead to insights regarding what really affects the price which can be used to develop a model for the LPG price. The aim is to find a suitable model that can be used to forecast future prices or at least predict the most truly direction and size of price changes for propane in the North-West European market.

This thesis will focus on the propane trading in the North-West European region, a small part of the global LPG market. All data presented in this study comes from this region, if not explicitly specified. There are probably large similarities between this market and the other LPG markets, but each regional market probably has its own unique characteristics. The data used is assumed to be the real market prices even though the actual price in almost all trades deviates from the listed prices since trades are made OTC with no limitations.

1.2 Disposition

In Chapter 2, some facts regarding LPG, the product chain and the LPG market are presented. In Chapter 3 the different hypotheses regarding driving factors of the LPG price together with arguments/reasons why these factors may affect the price are given. In Chapter 4 the statistical tools and models that will be used throughout the thesis are presented together with some important definitions. Chapter 5 contains a short presentation of the data on which the analysis is based. The work of evaluating hypotheses and finding appropriate models for the propane price is described in Chapter 6. Results of how the different models performs together with comparison of the different models are presented in Chapter 7. Finally the results are summarized and discussed in Chapter 8.

2 Liquefied Petroleum Gas

In this chapter a brief explanation of the LPG product and the LPG market is given. The LPG product chain is described and some important definitions are given.

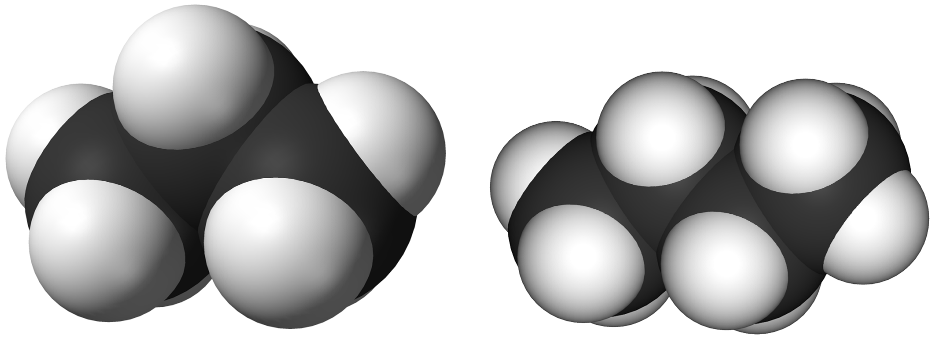
2.1 What is LPG?

LPG is the product name for a flammable mixture of hydrocarbon gases which mostly consist of propane and butane. The gas mixtures may have different contents, but they are still called LPG even though they have varying characteristics (see Table 2.1 and Figure 2.1).

LPG originates from crude oil or natural gas and is thus a fossil fuel. Compared to other fossil fuels such as oil and coal, LPG has lower particle emissions, lower NO_x emissions and lower sulphur content reducing the emission of air pollutions due to combustion. In addition LPG emits 33% less CO_2 than coal and 15% less than heating oil (European LPG Association, 2009). Another advantage with LPG is the distribution and availability. Even though it is a gas, LPG is mostly stored and transported in liquid form, heavily reducing its volume. Cooling or pressurising the gas transforms it into liquid phase, providing a great advantage since in liquid form much more energy can be distributed in an efficient way. The possibility to easily transform the gas into liquid is one of the major reasons why LPG is a popular fuel in rural and remote places with no gas network.

	Propane	Butane
Liquid density (t/m^3)	$0.50 - 0.51$	$0.57 - 0.58$
Gas density (kg/m^3)	$1.40 - 1.55$	$1.90 - 210$
Ratio gas/liquid	274	233
Boiling point (C^o)	-45	-2
Latent heat vaporization (KJ/kg)	358	372
Flammability limit	$2.2 - 10.0\%$	$1.8 - 9.0\%$
Calorific value (MJ/kg)	50.4	49.5
Minimum ignition temperature (C^o)	460	410

Table 2.1: Typical properties for propane and butane (The World LP Gas Association, 2010).



(a) Propane C_3H_8. (b) Butane C_4H_{10}.

Figure 2.1: Propane and Butane molecules.

2.1.1 Production

LPG is produced from two sources, crude oil and raw natural gas. After extracting crude oil from oil fields the crude oil is transported to a refinery where different products are separated. About $2-4\%$ of the crude oil is separated as LPG which is one of the lightest distillates (Energigas Sverige, 2008). LPG is just a small fraction of the refined crude oil and it is a bi-product with a ”small” value compared to gasoline ($\sim 45\%$ of crude oil), diesel and heating oil ($\sim 29\%$ of crude oil) and jet fuel kerosene ($\sim 9\%$ of crude oil) (U.S. Energy Information Administration, 2013). A little less than 50% of the LPG consumed in the world is produced from crude oil (The World LP Gas Association, 2010).

The rest of the LPG is produced from natural gas. When raw natural gas is extracted from gas fields it will go through a gas processing plant where it is cleaned and non-methane hydrocarbons are separated. About 3% of the raw natural gas consist of propane and butane and is separated as LPG (Energigas Sverige, 2008). Even when LPG is produced from natural gas it is a bi-product with ”small” value compared to the main product, methane.

2.1.2 Distribution

When the LPG has been separated it is transformed into liquid phase. In liquid phase the energy content per volume unit is about 250 times higher than in gas phase, making it more suitable for transport. From the refineries and gas processing plants LPG is typically transported by sea using specially built gas tankers. The possible shipping size varies from a couple of hundred tons up to 45000 tons (Shelley, 2003). As an alternative to shipping by sea, LPG can be loaded in rail road tanker cars or in tanker trucks for transport to customers. It is also common to bottle LPG in small bottles ($4-200\,\mathrm{kg}$) which are then transported by trucks or trains.

Storage

Larger LPG distributors and trading companies often have some kind of LPG terminal where they can store LPG. There are two different kind of storages, cooled and pressurised. The cooled storages are often old underground caverns equipped with a cooling system that keeps the temperature under the LPG boiling point. This temperature is typically below -35^0C. The pressurised storages are typically smaller than the cooled and instead of cooling, high pressure is used to keep the LPG in liquid phase.

Storage is an important part of the LPG business since the production stream of LPG is rather constant throughout the year but demand for heating purposes varies. For a trading company the storage may create value by using a buy and hold strategy when LPG prices are expected to increase. The LPG storage capacity is limited and building new storages is often very expensive.

2.1.3 Use of LPG

LPG is a very flexible energy source that can be used for almost any heating purpose. Most people relates LPG to bottled gas for barbecue, camping and cooking, but this is just a small part of the total use. The main use of LPG varies among countries and continents. Below the six different consumption sectors are listed together with the LPG world consumption share (The World LP Gas Association, 2010).

- Domestic ($\sim$ 47%)
- Industry ($\sim$ 11%)
- Agriculture ($\sim$ 2%)
- Transport ($\sim$ 9%)
- Refinery ($\sim$ 5%)
- Petrochemical industry ($\sim$ 26%)

Domestic

The domestic sector is the largest consuming sector of LPG in the world. In the domestic sector most of the LPG is used for space and water heating. Cooking is another large application, especially in Asia. LPG in the domestic sector is often a replacement for natural gas in rural areas where no natural gas network exists.

Industry

In the industrial sector LPG can be used for many different things. LPG is often used when a hot flame with high energy and clean burning characteristics is needed. Applications include steel mill melting furnaces, coffee roasting, drying of laundry, and ceramic production.

Agriculture

The agricultural sector uses LPG for drying of fruit and grain and heating of pig and chicken rearing, where a clean energy source is needed. Other applications are weed, and pest control and grass cutting fuel where LPG replaces chemicals and gasoline.

Transport

In the transport sector LPG is often referred to as autogas. Autogas can be used as vehicle fuel in cars, trucks and fork-lifts. The autogas sector grows constantly, but the number of vehicles and filling stations varies among countries. Hot air balloons are another consumer in the transport sector.

Refinery

As mentioned before, LPG is a bi-product of crude oil refining. About $2-4\,\%$ of crude oil is separated as LPG in refineries. The refinery process starts with heating of crude oil to high temperatures so that the different components evaporate before being separated by condensation. How the crude oil is heated varies between refineries, with some refineries using the produced LPG for heating. At refineries, LPG is also used to heat oil in storage tanks.

Petrochemical Industry

The petrochemical industry is the worlds second largest consumer of LPG. Here LPG is used as a feedstock for cracking to produce ethylene, propylene, butadiene, and other petrochemical products.

2.2 The LPG Market

LPG is used as an energy source all around the globe. This makes LPG an international cargo that is traded on an international market. Even if the LPG market is international, it is divided into regional markets since the transportation costs between different regions are high. However, inter-regional trades are common since there are regions where production does not cover consumption. Since this thesis focuses on the north-west European market a short description of this and the global market is provided below.

2.2.1 Global Market

Some regions of the world heavily affect the global LPG market. It is especially the large producing and consuming countries that are of interest. In Table 2.2 and Table 2.3 the Top-10 producers and consumers in 2009 are presented. The major regions that affect the global LPG market are the Middle East, Japan, Far-East (South-East Asia, including China, South-Korea, and India), Africa (mainly Algeria and Nigeria), and the U.S.

The Middle East is a large LPG producing region with a large surplus. It is the largest

	Production (million tons)	Share of Global %
USA	47.78	19.8
Saudi Arabia	24.63	10.2
China	19.11	7.9
Russian Federation	11.38	4.7
India	10.35	4.3
Canada	8.99	3.7
Algeria	8.07	3.4
United Arab Emirates	7.96	3.3
Mexico	6.70	2.8
Norway	6.52	2.7

Table 2.2: Top 10 LPG producing countries in the world in 2009 (The World LP Gas Association, 2010).

	Consumption (million tons)	Share of Global %
USA	54.07	22.5
China	23.17	9.6
Japan	16.44	6.8
India	12.53	5.2
Saudi Arabia	12.05	5.0
Russian Federation	9.53	4.0
South Korea	9.24	3.8
Mexico	9.08	3.8
Brazil	6.94	2.9
Canada	6.11	2.5

Table 2.3: Top 10 LPG consuming countries in the world in 2009 (The World LP Gas Association, 2010).

exporting region on the LPG market. Most of the excess LPG is traded and shipped to Japan and the Far-East. Smaller quantities are traded and shipped to Europe, U.S., and Central and South America. Since large parts of the LPG in Japan and Far-East region originates from the Middle East, the price in Asia is affected by the price in the Middle East.

Africa, and especially Algeria and Nigeria is the second largest exporting region on the LPG market. The surplus in this region is primarily traded and shipped to Europe and America.

Japan and South Korea are large LPG consuming countries with limited own production. Thus, they have to import large quantities to meet demand. The same is true for China and India even though they produce significant amounts of LPG domestically. The shortage of LPG in Japan and the Far-East results in higher prices in this region

	Consumption (million tons)	Share of total consumption %
Domestic	1.39	12.4
Industry	1.82	16.2
Agriculture	0.13	1.1
Transport	0.94	8.4
Refinery	0.77	6.9
Petrochemical	6.17	55.0
Total	11.21	100

Table 2.4: Usage of LPG in different sectors in NWE 2009 (European LPG Association, 2010).

compared to the rest of the world. The price in Japan and the Far-East is important for the rest of the world since it creates arbitrage opportunities by shipping LPG to that region instead of the local region if the price difference is high enough.

U.S. is the largest producer and consumer of LPG in the world. At the moment U.S. needs to import LPG to meet demand, but production is expected to grow. Most of the U.S. imports come from Africa. U.S. has the lowest LPG price in the world, but the shortage within the country limits the export to other regions.

2.2.2 North-West European Market

The European market is divided into three parts, North-West Europe (NWE), Mediterranean, and Eastern Europe. Most of the LPG in the NWE and Mediterranean market comes from oil and gas fields in the North Sea. The rest is imported from the Middle East, Africa, and Eastern Europe. The NWE market has its largest trading hubs in the Amsterdam-Rotterdam-Antwerpen (ARA) region and is sometimes called the ARA region.

The NWE region includes Belgium, Denmark, Germany, Ireland, Netherlands, Norway, Sweden, the UK, and the northern parts of France. The region consumes about 11 million tons LPG per year, which is a small part of the global market. The use of LPG in different sectors for NWE are presented in Table 2.4; the petrochemical industry dominates the use of LPG in NWE. Compared to the rest of the world, domestic use is particularly low.

Within NWE LPG is transported mainly by rail road tanker cars or trucks. For larger transports, small pressurised ships or barges are used if the destination is reachable by water.

2.3 Trading LPG

There exists no international marketplace for LPG. All trading commitments, both physically and financially, are made OTC where two parts agrees to trade LPG. Often a broker helps companies to find a suitable trading partner and to set up the deal. It is common that propane and butane are considered as separate products and not as LPG in trades because of their different characteristics.

Since there is no marketplace where LPG is exchanged there are no quoted prices. Instead several pricing institutes (Argus, Platts, Icis, etc.) quote their own prices based on trades that are made and interviews with trading companies and brokers. These prices are considered as good indicators of the current market price of LPG.

2.3.1 Physical Trading

Physical trading is the trading where LPG is exchanged between the two parties of the contract. The trading agreements can be constructed in many different ways and is unique for each trade. The contract contains the amount of LPG, which gas (propane, butane or a specified mix) that should be traded and the price. The price is often, but not always, linked to one of the pricing institutes monthly average price plus some discount or fee. The trading contract also states how the LPG should be delivered. There are two important terms that often occurs, CIF (Cost Insurance Freight) and FOB (Free on Board). If the trade is CIF the transportation cost to the buyer is included in the price. If the trade is FOB the buyer has to arrange and pay for the transport. Because of this, the listed FOB prices are, in general, lower than the listed CIF prices.

2.3.2 Financial Trading

Financial trading is the trading where no exchange of LPG necessarily takes place, this include option and swap contracts. These contracts can be used to secure a future price and hedge against price movements which can limit the risk of large unpredictable losses. On the LPG market it is mostly SWAP contracts that are traded.

SWAP Contract

A SWAP contract is a financial agreement where two parts accepts to swap their cash flows at a future point in time. The buying part of a SWAP contract with strike price K and maturity at time T gets the pay-off

$$P_T - K$$

at time T, where P_T is the price of LPG at time T. At the same time the selling side gets the pay-off

$$K - P_T.$$

If $P_T < K$ the buying side has to pay the difference to the selling side. If $P_T > K$ the opposite takes place. The price P_T is often connected to the monthly average price from some pricing institute. Often the SWAP contracts are referred to as Forward contracts even though no physical transaction of LPG takes place as in a normal Forward contract.

3 Driving Factor Hypotheses

There exists a number of hypotheses regarding possible driving factors of the LPG price. The hypotheses are based on assumptions that directly or indirectly may affect the price. In this chapter the hypotheses that will be examined in this thesis are presented.

3.1 Crude Oil

The crude oil hypothesis, that the LPG price is related to the crude oil price, exists for two reasons. First, LPG is produced by refining crude oil. Second, oil can be seen as an alternative heating fuel in many applications, with the same flexibility as LPG. So if the price differs too much consumers will switch to the cheaper fuel. Almost all prices of energy sources has some connection to the oil price, making this a very natural hypothesis.

The crude oil hypothesis will be evaluated by examining the correlation between log-returns of propane and crude oil.

3.2 Natural Gas

In the same way as one can assume that the LPG price is connected to the crude oil price one can assume that it is connected to the natural gas price; a little more than half of LPG is produced from natural gas. However, the alternative fuel hypotheses is not as strong in this case. Natural gas does not have the same flexibility as oil and LPG since it is harder to distribute. Natural gas often requires a gas network even though Liquefied Natural Gas (LNG) is becoming increasingly popular. Because of this, natural gas is not a competitive alternative fuel in markets without a gas network.

Historically the natural gas price is closely connected to the oil price. This connection is slowly disappearing, since natural gas today is traded separately from crude oil. The historical connection between crude oil and natural gas prices makes the natural gas and crude oil hypotheses related to each other. The natural gas hypothesis is evaluated in the same way as the crude oil hypothesis.

Feedstock source	Ethylene weight %	Propylene weight %	Butadiene weight %	Aromatics weight %	Other weight %
Ethane	84.0	1.4	1.4	0.4	12.8
Propane	45.0	14.0	2.0	3.5	35.5
Butane	44.0	17.3	3.0	3.4	32.3
Naphtha	34.4	14.4	4.9	14.0	32.3
Gas oil	25.5	13.5	4.9	12.8	43.3

Table 3.1: Weight of petrochemical products from steam cracking for different feedstocks (Staff, 2001).

3.3 Naphtha

The naphtha hypothesis is particularly strong in regions with much petrochemical industry, like NWE. Naphtha, just like LPG, is one of the light distillates in the crude oil refining process. The main use of naphtha is as a high octane component in gasoline and as feedstock for cracking in petrochemical industry. Many crackers can alter their use of feedstock and chooses the most cost efficient according to the demand of petrochemical products. However, the choice of feedstock for cracking affects the outcome of products, so demand for different products may also influence the choice of feedstock. In Table 3.1 the weight of petrochemical products for some different feedstock for steam cracking is presented. Since the petrochemical industry is a large consumer of LPG the demand from this sector may affect the LPG price. It is common that two parameters, naphtha-spread and naphtha-ratio, are discussed in the LPG pricing. The naphtha-spread is the difference in price between LPG and naphtha

$$P_{LPG} - P_{Naphtha}$$

and the naphtha-ratio is the quotient

$$\frac{P_{LPG}}{P_{Naphtha}}.$$

These two parameters, together with correlation between log-returns of propane and naphtha will be used to evaluate the hypothesis.

3.4 Seasonal Variation

The seasonal hypothesis exists since the LPG price often is lower during spring and summer than during winter. When LPG is used for space heating, typically in the domestic sector, the consumption depends on outdoor temperature. Therefore, one expects a connection between the LPG price and temperature. Another reason for a seasonal variation is that the production of LPG is relatively constant throughout the year while demand varies. The storage shortage makes it hard to store LPG for later use, which sometimes forces producers to dump the LPG price when their storages are filled up during periods of low demand. Demand is typically low during summer months because of lower heating needs and lower industrial use due to vacations.

Another possible reason for a seasonal variation is industrial and economical cycles. In good market conditions, demand from industry is high and companies are more willing to pay a higher price for feedstocks, such as LPG. In a bad economic environment, demand is typically low and companies are not willing to take the risk of purchasing LPG in large quantities. The market condition can be quantified using stock and industry indexes.

The seasonal hypothesis can be evaluated by analysing how changes in temperature and demand affects the LPG price. The Euro Stoxx 50[1] index will be used as market measure to evaluate how the industrial and economic situation affects LPG price.

3.5 East-West Spread

The East-West spread is a frequently discussed phenomenon in the LPG trading. It is defined as the difference in LPG prices between a western region, like NWE or Mediterranean, and an eastern region, like Japan or China

$$P_{west} - P_{east}.$$

Sometimes, when the actual difference in prices are unimportant the East-West ratio is used instead

$$\frac{P_{west}}{P_{east}}.$$

The strongest argument for its importance is that arbitrage opportunities will arise if the spread becomes too large. The arbitrage opportunity is to buy LPG in region with low price (western region) and sell in region with high price (eastern region). When the spread is larger than the shipping cost between the regions, an arbitrage opportunity exists. To utilise an arbitrage opportunity, low transport costs are necessary, which requires large trading volumes. In NWE a very limited number of companies, mainly because of limited storage capacity, are able to handle large trades when arbitrage opportunities occurs. Since the arbitrage opportunities are almost impossible to utilise, it is sometimes claimed that the East-West spread is a purely psychological thing, created by traders whom have large financial positions in LPG, in order to drive the prices in their favour and maximize profit.

The east-west spread hypothesis will be evaluated by analysing how changes in the spread affect the propane price in NWE. There may exist maximum/minimum levels of the spread which can be utilised to create upper and lower bounds for the price.

3.6 Other Factors

There are more hypotheses about driving factors of the LPG price. These other factors are not examined further in this thesis, but some of them are discussed shortly below.

[1] The Euro Stoxx 50 index is the leading blue chip index in Europe and consist of 50 sector leading stocks from the eurozone (STOXX, 2014).

Changes in Supply and Demand

As for all other products, it is reasonable to assume that the LPG price depends on supply and demand. According to basic microeconomic theory (see for instance, Axelsson et al., 1998; Varian, 2009), limited supply and high demand leads to higher prices, while large supply and limited demand leads to lower prices. Thus, if supply or demand changes, the LPG price should change.

The reason for not examining the supply and demand hypothesis further is the limited information available regarding supply and demand in NWE. A correct analysis requires knowledge of consumer preferences and how supply and demand changes. Collecting that information is an extensive work which goes beyond the scope of this thesis. If information regarding supply and demand in NWE were available, it would have been interesting to see how the changes affect the LPG price. Both supply and demand could have been used as external signals in ARMAX models (see Chapter 4), which is common in electricity price models for instance.

Psychology

All financial markets are full of psychology and believes about the future. These believes may affect price, especially in markets with low liquidity, like the LPG market. If many traders believe that the price will rise or fall, the price often follow these believes, even if the market conditions are unchanged. Worries about recession, wars, and changed regulations are other examples of psychological factors which can affect the market situation, and thus the price. The market psychology is almost impossible to analyse from a statistical point of view. Most psychological factors are hard to predict, making it hard to include them in statistical models. Therefore, the psychology of the LPG market will not be examined further in this thesis.

Shipping Availability

Large LPG trades are typically transported by sea in specially built gas tankers. When two parts agree to trade LPG it is decided if the price is CIF or FOB. For CIF contracts, the buyer knows what the price will be, delivered and ready to use. For FOB contracts, the buyer has to arrange and pay the shipping by itself and thus the "final" price is uncertain.

Some companies own or leases the gas tankers on a permanent basis, while others hire a gas tankers for each trade. To obtain a low price, including the shipping costs, it is important to have a smooth logistic chain without delays or waiting time for loading and unloading.

The effect of shipping will not be examined further because the shipping only affects the transportation part of the LPG price (if the price is CIF) and not the price of the LPG itself.

4 Theory

In this chapter the statistical tools, methods and models used throughout the work are presented.

4.1 Time Series Analysis

A time series

$$\{y_t; t = 0, 1, 2, \ldots\}$$

is a sequence of observations which are ordered in time. An example of a time series is a realization of the stochastic process

$$\{Y_t; t = 0, 1, 2, \ldots\}$$

where Y_t is a family of random variables and t is an index set. Time series analysis is the statistical discipline concerned with analysing, characterising, and forecasting time series.

When performing time series analysis one typically assume an underlying model, which the time series follows and accepts or rejects this model through a process of estimation and validation. In time series analysis it is preferable to work with stationary series since there are much more powerful tools to use for stationary series.

Definition 1. *A stochastic process X_t is Wide Sense Stationary or Weak Sense Stationary (WSS) if*

1. $\mathbb{E}(X_t) = \mu < \infty, \forall t$
2. $\mathbb{V}(X_t) = \sigma^2 < \infty, \forall t$
3. $\mathbb{C}(X_t, X_{t+\tau}) = \gamma(\tau), \forall t.$

In words a stochastic process is WSS if it has constant finite mean, constant finite variance and autocorrelation which only depends on the lag τ, over time.

Unfortunately most time series are not stationary. However, it is often possible to find a suitable transformation of the data, making the series stationary. When dealing with prices the most popular transformation is to use the return instead of true prices. To stabilize the variance the log-return is the standard choice for financial time series.

Definition 2. *The log-return of a series P_t is defined as*

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln\left(P_t\right) - \ln\left(P_{t-1}\right).$$

The different types of models considered in this work are presented below but first two important definitions are given.

Definition 3. *The autocorrelation function (ACF) is the cross correlation of a signal X_t with itself. It is defined as*

$$\gamma(s,t) = \frac{\mathbb{E}\left[(X_t - \mu_t)(X_s - \mu_s)\right]}{\sigma_t \sigma_s}$$

where μ_t is the mean and σ_t is the standard deviation of the stochastic process X at time t. If X_t is WSS the autocorrelation function is symmetric and simplifies to

$$\gamma(\tau) = \frac{\mathbb{E}\left[(X_t - \mu)(X_{t+\tau} - \mu)\right]}{\sigma^2} = \gamma(-\tau).$$

Definition 4. *The partial autocorrelation function (PACF) of a stochastic process X_t is the autocorrelation between X_t and $X_{t+\tau}$ with the linear dependence of X_{t+1} to $X_{t+\tau-1}$ removed. The PACF is thus the regression coefficients Φ_l which minimizes*

$$\{\Phi_l\} = \min \mathbb{E}\left[\left(X_t - \Phi_0 - \sum_{l=1}^{\tau} \Phi_l X_{t+l}\right)^2\right].$$

4.2 Time Series Models

To simplify the notation when presenting the models the shift operator z is used.

Definition 5. *The shift operator z is defined as*

$$x_t z^{-k} = x_{t-k}.$$

In the following e_t will represent a white noise process.

Definition 6. *A white noise process e_t is a stochastic process with the following properties*

1. $\mathbb{E}\left(e_t\right) = 0$, $\forall t$
2. $\gamma(t) = \sigma^2$, $t = 0$
3. $\gamma(t) = 0$, $t \neq 0$

where $\gamma(t)$ is the autocorrelation function. e_t is often assumed to follow a specific distribution. In this thesis, the normal distribution is assumed, yielding a Gaussian white noise.

4.2.1 ARMA Models

Autoregressive Moving Average (ARMA) models are mixtures of two of the most basic models, the autoregressive (AR) and the moving average (MA).

Definition 7. *An autoregressive model of order p, AR(p), is described by*

$$A(z)y_t = e_t$$

where

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}.$$

The AR models uses a weighted sum of past observations to describe the current state. AR models are very common and used in many applications. To identify an AR model it is possible to use the ACF and PACF; the ACF will typically have a slow exponential decay and the PACF will be zero for lags greater than the order of the process p, $\Phi_i = 0, i > p$.

Definition 8. *A Moving Average model of order q, MA(q), is described by*

$$y_t = C(z)e_t$$

where

$$C(z) = 1 + \sum_{i=1}^{q} c_i z^{-i}.$$

The MA model is very similar to the AR model with the difference that MA models observations using a weighted sum of past error terms instead of past observations. As for AR models the ACF and PACF can be used to find a suitable MA model. For a MA model of order q the ACF will be zero for lags greater than the order of the process q, $r(\tau) = 0, \tau > q$ while the PACF will have an exponential decay. If the AR and MA models are combined an ARMA model is obtained.

Definition 9. *The ARMA model of order* (p, q) *is described by*

$$A(z)y_t = C(z)e_t$$

where

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$$

$$C(z) = 1 + \sum_{i=1}^{q} c_i z^{-i}.$$

In ARMA models both past observations and error terms are used to describe the current state. ARMA models are a bit harder to identify than AR and MA models. Typically one has to use a trial and error method to find the correct orders, but both the ACF and PACF can be helpful in this work.

4.2.2 ARMAX Models

To further extend the ARMA models it is possible to include external signals to the model which is then called an ARMAX model.

Definition 10. *An ARMAX model is described by*

$$A(z)y_t = B(z)x_t + C(z)e_t$$

where x_t is the external signal and

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-i}$$

$$B(z) = z^{-d} \sum_{i=0}^{k-1} b_i z^{-i}$$

$$C(z) = 1 + \sum_{i=1}^{q} c_i z^{-i}.$$

An ARMAX model has one additional component, the $B(z)$ polynomial, compared to the ARMA model. The $B(z)$ polynomial, of order k, describes how the external signal affects the observations. The effect from the external signal can be delayed, which is described by d in the equations. It is possible to include more than one external signal in the model, in which case each signal gets its own $B(z)$ polynomial with individual orders and delays.

4.2.3 GARCH Models

Another type of model that is very common in economics and financial time series are the Autoregressive Conditional Heteroskedasticity (ARCH) models introduced in (Engle, 1982). In ARCH models the variance σ_t^2 of an error term ϵ_t is modelled as an MA process which changes over time. An extension of the ARCH model is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model, introduced in (Bollerslev, 1986), where the variance is modelled as an ARMA process instead. That the volatility of prices and stock returns exhibits time varying volatility clustering is a well known fact which makes GARCH models suitable for these problems.

Definition 11. *A GARCH(p,q) model of the process y_t is defined as*

$$y_t = \mu + \epsilon_t$$

where

$$\epsilon_t = \sigma_t \eta_t$$

where η_t is a white noise with zero mean and constant variance $\mathbb{V}(\eta_t) = 1$ and

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2.$$

To ensure positive variance and stability the following restrictions on the parameters must be fulfilled

$$\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1$$
$$\omega > 0 \quad \alpha_i > 0 \quad \beta_i > 0.$$

Often a GARCH(1,1) model is enough to get a satisfactory volatility model and higher order GARCH models are rarely used. The GARCH model can be extended and combined with other types of models to create a better estimate. In this thesis three combinations of the GARCH will be used, these are presented below.

EGARCH

The volatility of stock returns and prices is often asymmetric and increases more after large losses than after large increases. The traditional GARCH model is sometimes criticized because it does not catch this effect. To get rid of this problem the Exponential GARCH (EGARCH) where introduced (Nelson, 1991). In EGARCH model the logarithm of the variance is modelled instead of the variance directly.

Definition 12. *An EGARCH(p,q) model of the process y_t is defined as*

$$y_t = \mu + \epsilon_t$$

where

$$\epsilon_t = \sigma_t \eta_t$$

where η_t is a white noise with zero mean and unit variance and

$$\ln \sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \left(\theta \eta_{t-i} + \lambda \left(|\eta_{t-i}| - \mathbb{E}\left[|\eta_{t-i}| \right] \right) \right) + \sum_{i=1}^{p} \beta_i \ln \sigma_{t-i}^2.$$

There are two new parameters compared to the traditional GARCH. The parameter θ determines the sign effect of shocks which allow for asymmetry in the model. The parameter λ determines the size effect and how much a large negative or positive shock should affect the volatility. Compared to traditional GARCH the EGARCH has the advantage that there are no restrictions on the parameters since $\ln \sigma_t^2$ can be both positive and negative while σ_t^2 always has to be positive.

AR-GARCH

The AR-GARCH is a combination of AR and GARCH where the process y_t is assumed to follow an AR process instead of a random walk.

Definition 13. *The AR(k)-GARCH(p,q) model of the process y_t is the same as the GARCH model with the following modification in the equation of y_t*

$$A(z) y_t = \mu + \epsilon_t$$

where

$$A(z) = 1 + \sum_{i=1}^{k} \Phi_i z^{-i}.$$

GARCH-M

The GARCH in Mean (GARCH-M) model, inspired by (Engle et al., 1987), is an extension of the GARCH model where the error variance enters in mean equation of y_t.

Definition 14. *The GARCH-M model of y_t is the same as the GARCH model with the following difference in the mean equation*

$$y_t = \mu + \lambda g(\sigma_t^2) + \epsilon_t.$$

The function $g(x)$ can take any form but often $g(x) = x$ or $g(x) = \sqrt{x}$, which corresponds to the variance or volatility, is used.

4.3 Estimation

When a model has been selected the model parameters have to be estimated. Estimation of parameters can be done using different methods, for example least squares, maximum likelihood and method of moments (Jakobsson, 2012). The different methods are based on different theories and all have their different advantages and disadvantages. For estimation the model is assumed to be correct and parameter values that provide the best fit to data are determined. In estimation it is important to compute the variance of the estimated parameters to asses if they are significant. In this work two estimation techniques are used, maximum likelihood estimation and the prediction error method.

4.3.1 Maximum Likelihood

In maximum likelihood the parameters are estimated by finding the parameter values that maximizes the likelihood function which is the joint conditional probability density given the parameter values. Let $\boldsymbol{\theta}$ denote the vector containing all the parameters that should be estimated. Given the data $x_1, x_2, \ldots, x_N$ the likelihood function is

$$L(\boldsymbol{\theta}) = f_{X_1,X_2,\ldots,X_N}\left(x_1, x_2, \ldots, x_N|\boldsymbol{\theta}\right)$$

where $f_{X_1,X_2,\ldots,X_N}\left(x_1, x_2, \ldots, x_N|\boldsymbol{\theta}\right)$ is the joint conditional probability density function. If the observed data x_i is assumed to be independent, which often is the case, the likelihood function simplifies to

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta}).$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ is now found by solving

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

which is typically done using a numerical optimisation method. Many times it is easier to solve this problem using the log-likelihood function instead

$$l\left(\boldsymbol{\theta}\right) = \ln\left(L(\boldsymbol{\theta})\right) = \sum_{i=1}^{N} \ln\left(f\left(x_i|\boldsymbol{\theta}\right)\right).$$

The logarithm is a monotonically increasing function and the parameters that maximize the log-likelihood function are the same as those that maximize the likelihood function. For dynamic models, such as ARMA and GARCH, where the current observation depends on past observations, the likelihood function has to be modified. Given the past observations $x_{t-1},\, \ldots,\, x_1$, the current observation x_t, only depends on the white noise component e_t, which is independent of the earlier noise terms. Then the joint conditional probability density function can be decoupled into two parts yielding the likelihood function

$$\begin{aligned} L(\boldsymbol{\theta}) &= f_{X_1,\ldots,X_N}\left(x_1,\ldots,x_N|\boldsymbol{\theta}\right) \\ &= f_{X_{p+1},\ldots,X_N}(x_{p+1},\ldots,x_N|x_p,\, \ldots,\, x_1,\, \boldsymbol{\theta}) f_{X_1,\ldots,X_p}(x_1,\ldots x_p|\boldsymbol{\theta}) \\ &= \left(\prod_{i=p+1}^{N} f_{X_{p+1},\ldots,X_i}\left(x_{p+1},\ldots,x_i|x_{i-1},\, \ldots,\, x_1,\, \boldsymbol{\theta}\right)\right) f_{X_1,\ldots,X_p}\left(x_1,\ldots x_p|\boldsymbol{\theta}\right), \end{aligned}$$

where p is the AR order. Asymptotically, it can be shown that the prior distribution, $f_{X_1,\ldots,X_p}\left(x_1,\ldots x_p|\boldsymbol{\theta}\right)$, can be ignored and the maximum likelihood estimation is obtained by maximizing

$$L(\boldsymbol{\theta}) = \left(\prod_{i=p+1}^{N} f_{X_{p+1},\ldots,X_i}\left(x_{p+1},\ldots,x_i|x_{i-1},\, \ldots,\, x_1,\, \boldsymbol{\theta}\right)\right).$$

Under some regularity conditions the maximum likelihood estimate is asymptotically efficient and normally distributed (Jakobsson, 2012)

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\mathcal{I}}^{-1}\right), \quad \text{as } N \to \infty$$

or equivalently

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{1}{N}\boldsymbol{\mathcal{I}}^{-1}\right), \tag{4.1}$$

where $\boldsymbol{\mathcal{I}}^{-1}$ is the inverse of the Fischer information matrix:

$$\{\boldsymbol{\mathcal{I}}\}_{i,j} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta})\right]. \tag{4.2}$$

Using equation 4.1 and 4.2, it is possible to create appropriate confidence intervals for the parameter values. More information about maximum likelihood estimation can be found in Blom et al. (2005); Jakobsson (2012); Madsen et al. (2004).

4.3.2 Prediction Error Method

The prediction error method is very similar to the least square method where the parameters are estimated by minimizing the sum of squared residuals from a model. The difference is that in prediction error method the residual is defined as the difference between the measurement and the one-step ahead prediction

$$\epsilon_{t+1|t}(\boldsymbol{\theta}) = y_{t+1} - \hat{y}_{t+1|t}(\boldsymbol{\theta})$$

where $\hat{y}_{t+1|t}(\boldsymbol{\theta})$ denotes the one-step prediction given the parameters $\boldsymbol{\theta}$. The parameters are then found by minimizing the squared prediction errors

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_t \epsilon^2_{t+1|t}(\boldsymbol{\theta}),$$

typically using numerical methods. The reason for using the prediction error method instead of the easier least square method is that the variance of the parameter estimates are typically lower when using the prediction error method. Let $S_{\boldsymbol{\theta}}$ denote the sum of the squared residuals, the variance of the residuals is estimated as

$$\hat{\sigma}^2_\epsilon = \frac{S_{\boldsymbol{\theta}}}{N-1}.$$

If it is assumed that the residuals are white, like in the models above, the prediction error estimates are consistent and asymptotically normal distributed with variance

$$\mathbb{V}\left(\hat{\boldsymbol{\theta}}\right) = 2\sigma^2_\epsilon \boldsymbol{H}^{-1}$$

where $\boldsymbol{H}$ is the Hessian matrix of $S_{\boldsymbol{\theta}}$ defined as

$$\{h_{i,j}\} = \frac{\partial^2 S_{\boldsymbol{\theta}}}{\partial\theta_i \partial\theta_j}.$$

From this it is possible to create confidence intervals for the estimated parameters. The prediction error method is described in greater detail by Jakobsson (2012).

4.4 Model Order Selection

As mentioned in Section 4.2.1, the ACF and PACF can be used to find appropriate model orders for MA and AR processes. For other models there are no similar tools to use. One possibility is to use information theoretic models to estimate suitable model orders (Jakobsson, 2012). In these methods, a penalized likelihood function is minimized to obtain estimates of model order. Adding more parameters to a model, i.e. increasing the order, will always result in a better likelihood fit. The problem is the risk of over-fitting the data by using too high orders. In the information theoretic models, the likelihood function is penalized according to the number of parameters to estimate, in order to overcome the over-fitting problem. The two different information criteria that will be used in this thesis are described below.

4.4.1 Bayesian Information Criteria

The Bayesian information criteria, discussed in Schwarz (1978); Kashyap (1982), is defined as

$$BIC(p) = -2\ln(\mathbf{L}) + p\ln(N)$$

where p is the number of estimated parameters, $\mathbf{L}$ is the maximized likelihood function, and N is the number of samples. By minimizing $BIC(p)$ over all possible p, an estimate of the model order is obtained.

4.4.2 Akaike's Information Criteria

Akaike's information criteria (AIC) (Akaike, 1974), is defined as

$$AIC(p) = -\ln(\mathbf{L}) + 2p$$

where p is the number of estimated parameters and $\mathbf{L}$ is the maximized likelihood function. By minimizing $AIC(p)$, model order estimates are obtained.

If different model order estimates are obtained from AIC and BIC, one should remember that BIC penalizes the number of parameters more than AIC and that AIC tends to over-estimate model order. It is also important to know that neither AIC or BIC gives information of how good the models are, only what number of parameters that seem to be optimal.

4.5 Validation

When a model has been selected and the parameters are estimated the performance of the model has to be validated. Typically this validation is done using another dataset than the estimation to check that the model works properly out-of-sample.

4.5.1 Whiteness Tests

In the validation procedure the residuals of the model are analysed. All models assume that there are error terms e_t containing the part the model is unable to explain. If the model is accurate the errors should be white noise, which means that there is no structure left in the error terms. To test if the error terms are white, there are whiteness tests that can be used. The whiteness tests that are used in this work are presented below. All the tests are based on the fact that ACF and PACF of a white noise process will be asymptotically normally distributed with

$$\mathcal{N}\left(0, \frac{1}{N}\right)$$

for all lags $\tau \neq 0$. It should be mentioned that these test do not say anything about the distribution of the noise which typically is assumed to be normal, student's t or generalized error distributed. If the noise term e_t in the models is rewritten as

$$e_t = \sigma_t \eta_t$$

where η_t is a white noise with $\mathbb{E}(\eta_t) = 0$ and $\mathbb{V}(\eta_t) = 1$ for all t, we allow for time varying volatility. In ARMA models it is assumed that the volatility is constant, $\sigma_t = \sigma$, while it changes in GARCH models. To get correct results, the normalized residuals

$$\frac{e_t}{\sigma_t}$$

should be used in the tests. The tests presented below are described in greater detail by Brooks (2008); Jakobsson (2012).

Box-Pierce Test

The Box-Pierce Q-statistics test (Box and Pierce, D. A., 1970) is a test which test the hypothesis that the initial ACF coefficients of the residuals are not significantly different from zero, using the Q-statistics

$$Q = N \sum_{i=1}^{K} \gamma^2(i)$$

where N is the sample size, γ is the autocorrelation function and K is the number of considered correlations, typically $15 \leq K \leq 25$. If the residual is a white noise process it can be shown that Q is asymptotically χ^2 distributed with K degrees of freedom, $Q \in \chi^2(K)$. The hypothesis that the initial ACF coefficients of the residuals are not significantly different from zero (and thus not a white noise) is rejected with significance α if

$$Q > \chi^2_{1-\alpha}(K)$$

where $\chi^2_{1-\alpha}(K)$ is the α-quantile of the χ^2 distribution with K degrees of freedom. Since the Q-statistics is only asymptotically χ^2 distributed the test may be poor for small sample sizes.

Ljung-Box-Pierce Test

The Ljung-Box-Pierce test (Ljung and Box, G. E. P., 1978) is very similar to the Box-Pierce test and tests the same hypothesis. The small modifications is in the test statistics Q which is formed as

$$Q = N(N+2) \sum_{i=1}^{K} \frac{\gamma^2(i)}{N-i} \tag{4.3}$$

which is a better approximation of the χ^2 distribution. The Ljung-Box-Pierce test yields better performance in general but may still be poor for small sample sizes.

McLeod-Li Test

In the McLeod-Li test (McLeod and Li, W. K., 1983) equation 4.3 is used but γ^2 is replaced by the ACF coefficients of the squared residuals instead of the residuals themselves. The McLeod-Li test examines higher order dependences than the previous tests and it is sensitive to the assumption that the residuals are normal distributed, which makes it unreliable for non-normal data.

Monti Test

The Monti test (Monti, 1994) is also based on equation 4.3 but in this test γ^2 is replaced by the PACF coefficients of the residuals instead of the ACF. The hypothesis that the initial PACF coefficients are not significantly different from zero is rejected in the same way as in the Box-Pierce test.

Sign Change Test

Another way to test whether the noise is white or not is to look at the sign change of the residuals. Since the noise is independent and identically distributed, with a symmetric distribution with zero mean, the probability of a sign change of a residual from one time point to the next one should be about 50%. Thus the number of sign changes, P, should follow the binomial distribution

$$P \in \text{Bin}\left(N-1, \frac{1}{2}\right).$$

For large sample sizes this can be approximated as a normal distribution

$$\mathcal{N}\left(\frac{N-1}{2}, \frac{N-1}{4}\right)$$

which can be used to form a confidence interval of P and test the hypothesis that the sign change frequency is about 50%.

4.5.2 Prediction Error

Another way to validate the model is to perform predictions, one or several steps ahead, and compare these with the true data. If the model is good the predictions should be fairly close to the real data. For some time series, like financial time series, prediction can be really hard since the series has a large noise component. Then one has to keep in mind that the predictions might be poor even when the model is quiet good. To analyse and compare how well the models perform in prediction one can use the mean squared error (MSE)

$$MSE = \frac{1}{n-k} \sum_{i=1}^{n-k} \left(y_{t+k} - \hat{y}_{t+k|t}\right)^2$$

where $\hat{y}_{t+k|t}$ is the k-step prediction from time t. Sometimes the root mean square error (RMSE) which is just the square root of the MSE is used instead to get a normalized value.

4.5.3 Hit Rate

Financial time series are known to have large noise terms which makes them hard to predict accurately. Bad prediction results may lead to rejection of models that could be useful even if the prediction is bad. If we only care about whether the price is most likely to increase or decrease, and not exactly how large the increase or decrease is, then we can use the hit rate for validation. The hit rate, HR, for a model of the series y_t is defined as

$$HR = \frac{1}{N} \sum_{t=1}^{N} \mathbb{1}_{\{\text{sign}(y_t)=\text{sign}(\hat{y}_t)\}} \tag{4.4}$$

where $\mathbb{1}$ is the indicator function, sign(y_t) is the sign of y_t and $\hat{y}_t$ is the model prediction of y_t. The sum is taken over all available predictions. Using the hit rate an estimate of the probability of predicting the return in the right direction is obtained. In equation 4.4 the hit rate is defined using equal signs of prediction and true data. This could be modified and extended. As an example a "hit" can be when the difference between prediction and true data is smaller than a certain threshold.

To evaluate how good a hit rate is, the hit rate can be compared with the expected hit rate from guessing. For financial data, a random walk model with equal probability of positive and negative return is often assumed (Björk, 2009; Brooks, 2008; Madsen et al., 2004). Thus, the number of hits when guessing is expected to follow the binomial distribution $X \in \text{Bin}\left(n, \frac{1}{2}\right)$, giving the expected hit rate $\mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{2} = 50\%$. For large samples, the binomial distribution can be approximated using the normal distribution

$$X \in \text{Bin}\left(n, p\right) \approx \mathcal{N}\left(np, np(p-1)\right)$$

which can be used to create confidence intervals for the hit rate (Blom et al., 2005). If the obtained hit rate from a model is significantly higher than the hit rate from guessing, the model can be considered as good.

4.6 Prediction

When a model has been estimated and validated it is possible to use it for prediction and forecasting of future values of a time series. This is often the most interesting part of time series analysis and can be very useful if the predictions are accurate. Some models can be used to predict the actual outcome of a time series while others, like GARCH models, are mostly used to predict the variance of the noise term and not the outcome. In the following let Ω_t denote all available information at time point t. The k-step prediction, at time point t, of a time series y_t is then formed as the conditional expectation

$$\hat{y}_{t+k|\Omega_t} = \mathbb{E}\left[y_{t+k}|\Omega_t\right].$$

How the expectation is calculated depends on the underlying model used to describe y_t. Below, predictions for ARMAX and GARCH models are described. More about prediction of time series can be found in Brooks (2008); Jakobsson (2012); Lindgren et al. (2009); Madsen et al. (2004)

4.6.1 Linear Models

The linear ARMA model of the time series y_t is described by

$$A(z)y_t = C(z)e_t$$

which may be rewritten as

$$y_t = \frac{C(z)}{A(z)}e_t = \sum_{i=0}^{\infty} \psi_i e_{t-i}$$

after expanding the polynomial division. In a k-step prediction we only know the values of y_t up to time point t. Then the polynomial division can be divided into two parts, one that handle the unknown future values and one that handle the past observed values.

$$\begin{aligned}
y_{t+k} &= \frac{C(z)}{A(z)}e_{t+k} \\
&= \left(F(z) + z^{-k}\frac{G(z)}{A(z)}\right)e_{t+k} \\
&= F(z)e_{t+k} + \frac{G(z)}{A(z)}e_t \\
&= F(z)e_{t+k} + \sum_{i=k}^{\infty}\psi_i e_{t+k-i}
\end{aligned}$$

where $F(z)$ is a polynomial of order $k-1$ and $G(z)$ and $F(z)$ satisfies the Diophantine equation

$$C(z) = A(z)F(z) + z^{-k}G(z).$$

By rewriting

$$e_t = \frac{A(z)}{C(z)}y_t$$

and inserting this into the equation we get

$$y_{t+k} = F(z)e_{t+k} + \frac{G(z)A(z)}{A(z)C(z)}y_t = F(z)e_{t+k} + \frac{G(z)}{C(z)}y_t$$

and the k-step prediction is then given by

$$\hat{y}_{t+k|t} = \mathbb{E}\left[y_{t+k}|\Omega_t\right] = \mathbb{E}\left[F(z)e_{t+k} + \frac{G(z)}{C(z)}y_t \middle| \Omega_t\right] = \frac{G(z)}{C(z)}y_t.$$

For an ARMAX model with the external signal x_t

$$A(z)y_t = B(z)x_t + C(z)e_t$$

the prediction is very similar. Start by forming the equation for y_{t+k}, using the rewritten form of $C(z)$ above, as

$$\begin{aligned}
y_{t+k} &= \frac{C(z)}{C(z)}y_{t+k} \\
&= \frac{1}{C(z)}\left(A(z)F(z) + z^{-k}G(z)\right)y_{t+k} \\
&= \frac{1}{C(z)}\left(F(z)A(z)y_{t+k} + G(z)y_t\right) \\
&= \frac{1}{C(z)}\left(F(z)C(z)e_{t+k} + F(z)B(z)x_{t+k} + G(z)y_t\right) \\
&= F(z)e_{t+k} + \frac{F(z)B(z)}{C(z)}x_{t+k} + \frac{G(z)}{C(z)}y_t \\
&= F(z)e_{t+k} + H(z)x_{t+k} + \frac{I(z)}{C(z)}x_t + \frac{G(z)}{C(z)}y_t
\end{aligned}$$

where the polynomial $H(z)$ is of order $k-1$ and $H(z)$ and $I(z)$ satisfies the Diophantine equation

$$F(z)B(z) = C(z)H(z) + z^{-k}I(z).$$

Then the k-step prediction is formed as

$$\begin{aligned}
\hat{y}_{t+k|t} &= \mathbb{E}\left[y_{t+k}|\Omega_t\right] = \mathbb{E}\left[F(z)e_{t+k} + H(z)x_{t+k} + \frac{I(z)}{C(z)}x_t\, \frac{G(z)}{C(z)}y_t \middle| \Omega_t\right] \\
&= H(z)\mathbb{E}\left[x_{t+k}|\Omega_t\right] + \frac{I(z)}{C(z)}x_t + \frac{G(z)}{C(z)}y_t.
\end{aligned}$$

The ARMAX prediction is more complicated than the ARMA prediction since the expected values of future input signals are needed.

4.6.2 GARCH Models

For the original GARCH model of y_t where

$$y_t = \mu + \epsilon_t$$

$$\epsilon_t = \sigma_t \eta_t$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$$

with $\mathbb{E}\left[\eta_t\right] = 0$ and $\mathbb{V}\left[\eta_t\right] = 1$ the prediction of y_t is very easy since $\mathbb{E}\left[\epsilon_t\right] = 0$. This gives the k-step prediction as

$$\hat{y}_{t+k|t} = \mathbb{E}\left[y_{t+k}|\Omega_t\right] = \mathbb{E}\left[\mu + \epsilon_{t+k}|\Omega_t\right] = \mu.$$

This prediction is rather uninteresting and often the variance $\mathbb{V}\left[y_t\right] = \sigma_t^2$ is more interesting to predict. First one should note that $\mathbb{E}\left[\epsilon_t^2\right] = \mathbb{E}\left[\sigma_t^2\right]$ since $\mathbb{E}\left[\epsilon_t\right] = 0$. Let σ^2 denote the unconditional expectation of the variance at any time. This gives the unconditional expectation of the variance as

$$\begin{aligned}
\sigma^2 = \mathbb{E}\left[\sigma_t^2\right] &= \mathbb{E}\left[\omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2\right] \\
&= \omega + \sum_{i=1}^{q} \alpha_i \mathbb{E}\left[\epsilon_{t-i}^2\right] + \sum_{i=1}^{p} \beta_i \mathbb{E}\left[\sigma_{t-i}^2\right] \\
&= \omega + \sum_{i=1}^{q} \alpha_i \mathbb{E}\left[\sigma_{t-i}^2\right] + \sum_{i=1}^{p} \beta_i \mathbb{E}\left[\sigma_{t-i}^2\right] \\
&= \omega + \sigma^2 \left(\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i\right) \\
&= \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{i=1}^{p} \beta_i}.
\end{aligned}$$

A k-step prediction of the variance will converge towards σ^2 as k increases which makes GARCH models bad for long term predictions of the variance. For shorter prediction horizons they can be very useful though. If we start with the 1-step prediction it is given by $\mathbb{E}\left[\sigma_{t+1}|\Omega_t\right]$ where Ω_t as before denotes all available information at time point t,giving the following 1-step prediction

$$\begin{aligned}\hat{\sigma}^2_{t+1|t} &= \mathbb{E}\left[\sigma^2_{t+1}|\Omega_t\right] = \mathbb{E}\left[\omega + \sum_{i=1}^{q}\alpha_i\epsilon^2_{t+1-i} + \sum_{i=1}^{p}\beta_i\sigma^2_{t+1-i}\middle|\Omega_t\right]\\ &= \omega + \sum_{i=1}^{q}\alpha_i\epsilon^2_{t+1-i} + \sum_{i=1}^{p}\beta_i\sigma^2_{t+1-i}\end{aligned}$$

since all the values are known at time t. The 2-step prediction is a bit more complicated since it includes unknown factors. If we use that $\mathbb{E}\left[\epsilon^2_{t+k}|\Omega_t\right] = \mathbb{E}\left[\sigma^2_{t+k}|\Omega_t\right]$ we get

$$\begin{aligned}\hat{\sigma}^2_{t+2|t} &= \mathbb{E}\left[\sigma^2_{t+2}|\Omega_t\right] = \mathbb{E}\left[\omega + \sum_{i=1}^{q}\alpha_i\epsilon^2_{t+2-i} + \sum_{i=1}^{p}\beta_i\sigma^2_{t+2-i}\middle|\Omega_t\right]\\ &= \omega + \mathbb{E}\left[\alpha_1\epsilon^2_{t+1} + \beta_1\sigma^2_{t+1}|\Omega_t\right] + \sum_{i=2}^{q}\alpha_i\epsilon^2_{t+2-i} + \sum_{i=2}^{p}\beta_i\sigma^2_{t+2-i}\\ &= \omega + (\alpha_i + \beta_i)\,\mathbb{E}\left[\sigma^2_{t+1}|\Omega_t\right] + \sum_{i=2}^{q}\alpha_i\epsilon^2_{t+2-i} + \sum_{i=2}^{p}\beta_i\sigma^2_{t+2-i}\\ &= \omega + (\alpha_i + \beta_i)\,\hat{\sigma}^2_{t+1|t} + \sum_{i=2}^{q}\alpha_i\epsilon^2_{t+2-i} + \sum_{i=2}^{p}\beta_i\sigma^2_{t+2-i}\end{aligned}$$

so we use the 1-step prediction in the 2-step prediction as well. Giving the following recursion for a general k-step prediction, where we assume that $\alpha_i = 0$ if $i > q$ and $\beta_i = 0$ if $i > p$,

$$\begin{aligned}\hat{\sigma}^2_{t+k|t} &= \mathbb{E}\left[\sigma^2_{t+k}|\Omega_t\right] = \mathbb{E}\left[\omega + \sum_{i=1}^{\infty}\alpha_i\epsilon^2_{t+k-i} + \sum_{i=1}^{\infty}\beta_i\sigma^2_{t+k-i}\middle|\Omega_t\right]\\ &= \omega + \mathbb{E}\left[\sum_{i=1}^{k-1}\alpha_i\epsilon^2_{t+k-i} + \sum_{i=1}^{k-1}\beta_i\sigma^2_{t+k-i}\middle|\Omega_t\right] + \sum_{i=k}^{\infty}\alpha_i\epsilon^2_{t+k-i} + \sum_{i=k}^{\infty}\beta_i\sigma^2_{t+k-i}\\ &= \omega + \sum_{i=1}^{k-1}(\alpha_i + \beta_i)\,\mathbb{E}\left[\sigma^2_{t+k-i}|\Omega_t\right] + \sum_{i=k}^{\infty}\alpha_i\epsilon^2_{t+k-i} + \sum_{i=k}^{\infty}\beta_i\sigma^2_{t+k-i}\\ &= \omega + \sum_{i=1}^{k-1}(\alpha_i + \beta_i)\,\hat{\sigma}^2_{t+k-i|t} + \sum_{i=k}^{\infty}\alpha_i\epsilon^2_{t+k-i} + \sum_{i=k}^{\infty}\beta_i\sigma^2_{t+k-i}\end{aligned}$$

which will converge to the unconditional variance σ^2 as k grows.

5 Presentation of the Data

To evaluate the driving factor hypotheses and to find a suitable model for the propane price in NWE, different types of data has been collected. The data contains price series of propane, butane, naphtha, natural gas, brent[1] crude oil, and Euro Stoxx 50 index. In addition temperature data for NWE and swap/forward prices of propane, naphtha and Brent have been collected. The data consist of daily (all available trading days) observations in the period starting on the 4th of January 1999 until the 28th of February 2014. Some series have up to 30 missing data points in this period which are then replaced by the average of previous and following observations. If more than three series have missing data for a specific date this date is removed from the series. This gives a total of 3803 daily observations or 182 months averages available for the analysis.

The price series of propane and naphtha as well as the swap prices for propane, naphtha and Brent were collected. All these prices are the CIF price in the ARA region given in USD/t. There is also one series for the propane price in Japan. There are six different propane swaps with maturity $1-6$ months forward in time. For naphtha there are three different swaps with maturity $1-3$ months forward in time. For Brent there are three different forward contracts with maturity $1-3$ months forward in time. It is important to note that the observed swap prices correspond to the strike price in the swap contract which is based on the monthly average price during that month.

The crude oil spot price series, in this thesis corresponding to the Brent price, is available from U.S. Energy Information Administration (2014). The Brent price series was chosen since it is one of the most traded crude oils and originates from the North Sea, close to NWE. The Brent price is the FOB spot price given in USD/bbl.

The natural gas data are available from Gaspoint Nordic (2014). The natural gas data starts on the 5th of March 2008, making the price series much shorter than the other. The marketplace "Gaspoint Nordic" were established in late 2007 and before this date there were no trades. There are possibly other sources of data that could have been used. The natural gas price is given in EUR/MWh and not in USD as the other price series.

The Euro Stoxx 50 index data is available from STOXX (2014). The used series is the index price in USD. This was chosen to minimize the currency effect when analysing the

[1] Brent Crude is the name of the sweet light crude oil sourced from the North Sea. Brent, together with West Texas Intermediate (WTI) crude, are the most well know crude oils and are very important for the global oil prices.

data since the other prices are given in USD.

The temperature data that are used are collected from European Climate Assessment and Dataset (2014) and Wunderground (2014). Since I was unable to find daily average temperatures for the whole NWE region, it was created using average temperatures from a couple of places in the region. How the temperature data was computed is described in detail in Appendix A.

6 Modeling the LPG Price

In this chapter the work of finding an appropriate model for the propane price in NWE is described. This includes transforming and de-trending the data, evaluation of the driving factor hypotheses, and estimation and validation of the different models.

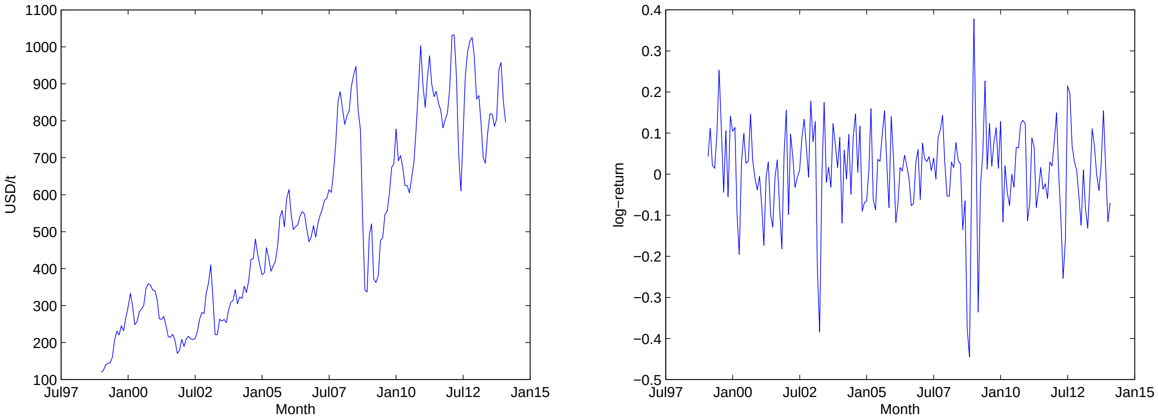
6.1 Preparation of the Data

As mentioned before the data consists of daily prices and monthly average prices from January 1999 to February 2014. The monthly average data is shown in Figure 6.1. The true prices exhibit a growing trend and is clearly non-stationary. To overcome this problem the data is transformed to log-returns, making the series more stationary even though it looks like the variance still varies over time. This property is typical for financial data and one of the reasons why GARCH models may be useful. Since the log-returns are more stationary it will be used instead of true prices for modeling. If it is possible to model the log-returns this can be used for the true prices since they are related by

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \quad \Leftrightarrow \quad P_t = P_{t-1}e^{r_t}.$$

The other price series, for Naphtha, Brent and Euro Stoxx 50 index as well as the East-West ratio, are also transformed using log-returns for the same reason as above. The temperature series is also transformed since it seems logical to expect that it is the change in temperature from one time point to another that affects the propane price and not the exact temperature at the current time point. For temperature, the log-return is not a good choice for transformation. Instead the one step difference $\Delta T_t = T_t - T_{t-1}$ is used.

To estimate and validate different models the data is divided into two parts. One which is used for estimation of the parameters and one part used for validation of the estimated models. It is of course possible to use the estimation data for validation as well but this is cheating because then the model is already adapted to the validation data. The models should work for future (unknown) data if it should be of any use, why it is better to validate it over a dataset that is not included in the estimation. In this thesis, about two thirds of the data will be used for estimation, and the rest for validation. The estimation dataset starts in January 1999 and ends in December 2009. The validation dataset starts in January 2010 and ends in February 2014. For the monthly average data there are only 181 log-return observations available which may affect the result.



(a) Propane price (USD/t) in NWE.

(b) Propane log-returns in NWE.

Figure 6.1: Monthly average propane price (USD/t) in NWE and the corresponding log-returns from January 1999 to February 2014.

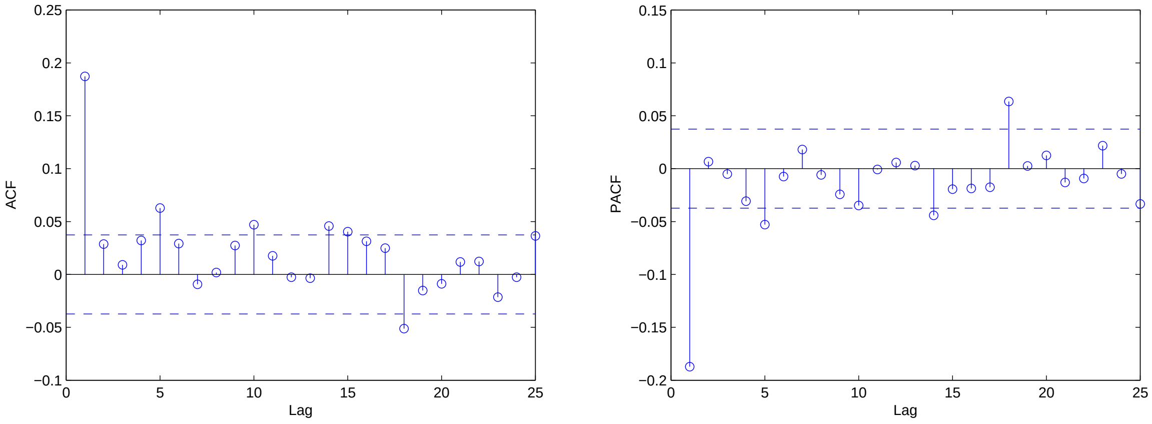
6.1.1 Analysis of Propane Log-Returns

Even though the variance varies over time it will be assumed that the log-returns are WSS. Thus, the ACF and PACF will be used to analyse the dependency in the return series and suggest suitable models. The ACF and PACF for lag $1-25$ of the estimation part of daily log-returns are shown in Figure 6.2 and for monthly average log-returns in Figure 6.3.

For the daily log-returns the ACF is significant for lag 1. There are also significant values for some higher lags but these are rather small. This suggests that a MA(1) model may be appropriate to describe the daily log-returns. The PACF shows a similar result, where the lag 1 coefficient is large and negative, which suggest that an AR(1) model may be useful.

For the monthly average log-returns the result is very similar. The ACF has one large coefficient at lag 1 indicating that a MA(1) model could be a good choice. The PACF has two significant values at lag 1 and 2 suggesting an AR(2) model. The PACF has also significant values at lag 12 and 18, corresponding to one and one and a half year back in time, which may indicate some seasonal dependencies.

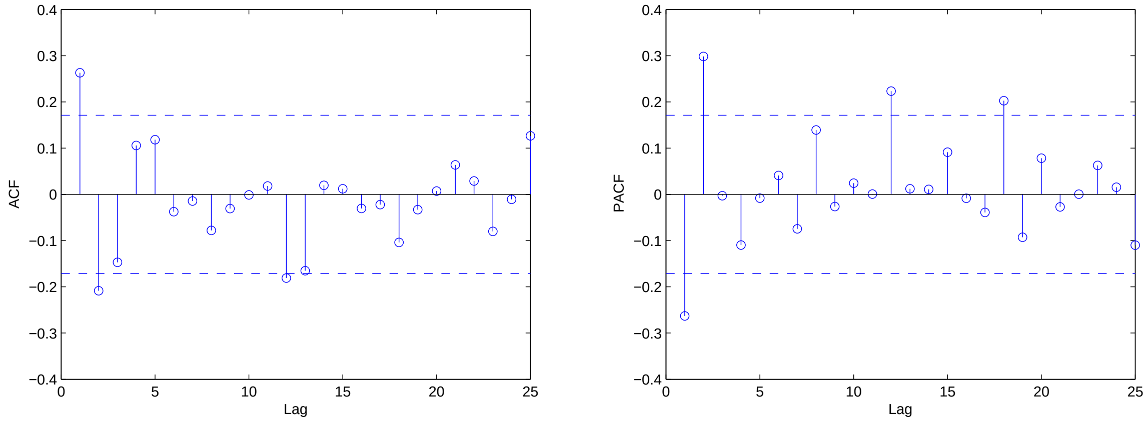
The short analysis above, shows that there are some dependency in the propane log-returns which can be included in a model. Before continuing the modeling work, the hypotheses will be evaluated, which hopefully gives more information about which factors that affect the log-returns.



(a) ACF

(b) PACF

Figure 6.2: ACF and PACF for estimation part of daily propane log-returns in NWE together with 95% confidence bounds.



(a) ACF

(b) PACF

Figure 6.3: ACF and PACF for estimation estimation part of monthly average propane log-returns in NWE together with 95% confidence bounds.

Lag	Brent	Natural gas	Naphtha	Temperature	Euro Stoxx 50 index	East-West ratio
0	**0.4374**	0.0202	**0.5104**	**−0.0004**	**0.2767**	**0.6521**
1	**0.0952**	0.0019	**0.0933**	−0.0003	0.0435	**0.1023**
2	**0.0490**	**0.0272**	**0.0374**	0.0000	0.0000	−0.0031
3	0.0185	**0.0377**	**0.0473**	−0.0003	−0.0077	−0.0257
4	0.0247	−0.0044	0.0265	−0.0002	**0.0533**	−0.0034

Table 6.1: Regression parameter β for daily simple linear regression of propane log-returns using different explanatory variables and lags. Values in bold face are significant at the 5% level.

Lag	Brent	Natural gas	Naphtha	Temperature	Euro Stoxx 50 index	East-West spread
0	**0.7729**	**0.3549**	**0.8387**	−0.0042	**0.6549**	**0.7286**
1	**0.2112**	−0.0689	**0.1890**	−0.0021	**0.7956**	0.2640
2	−0.0908	−0.2641	**−0.2299**	0.0025	0.2954	−0.1706
3	−0.0092	−0.0065	−0.1048	0.0022	−0.2451	0.0721
4	0.0413	−0.0416	0.1033	0.0037	−0.1112	0.0726

Table 6.2: Regression parameter β for monthly average simple linear regression of propane log-returns using different explanatory variables and lags. Values in bold face are significant at the 5% level.

6.2 Evaluation of the Hypotheses

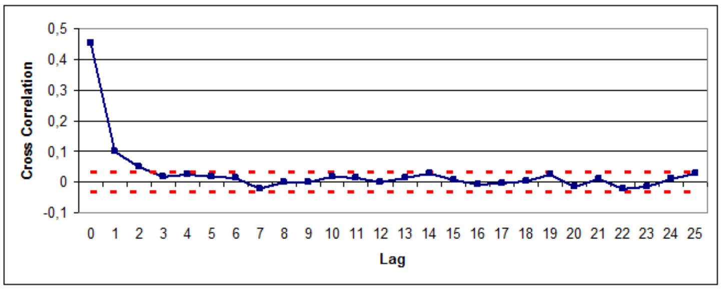
To evaluate the hypotheses all available data will be used, both estimation and validation data. The hypotheses are evaluated by analysing the cross correlation and the regression parameter β, in the simple linear regression

$$r_t = \alpha + \beta x_{t-l} + e_t$$

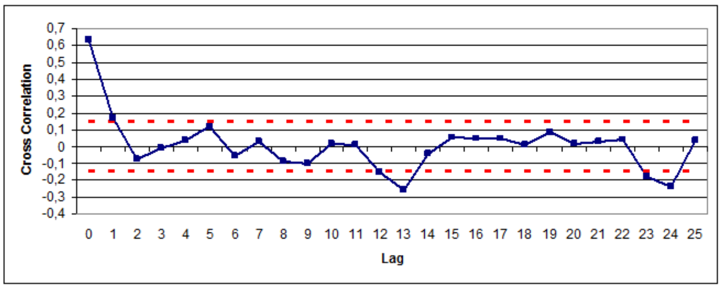
where r_t is the propane log-return and x_{t-l} is the explanatory variable lagged l steps. The estimated regression parameter for daily returns are shown in Table 6.1 and for monthly average returns in Table 6.2. The cross correlation between propane log-returns and other returns for different lags are shown in Figure 6.4-6.9.

6.2.1 Crude Oil

As can be seen in Figure 6.4 and Table 6.1 and 6.2, there is strong correlation between Brent and propane returns at lag zero, both for daily and monthly average data. For daily returns the regression parameters for the first two lags are significantly different from zero, which indicates that they can be used to explain future returns. For monthly average data it is only the first lag regression parameter that is significantly different from zero. There are also some negative correlation for lag 13 and 24, which will be examined further when constructing the models. To conclude, there is a connection between Brent and propane returns and the crude oil hypothesis is accepted.



(a) Daily.



(b) Monthly average.

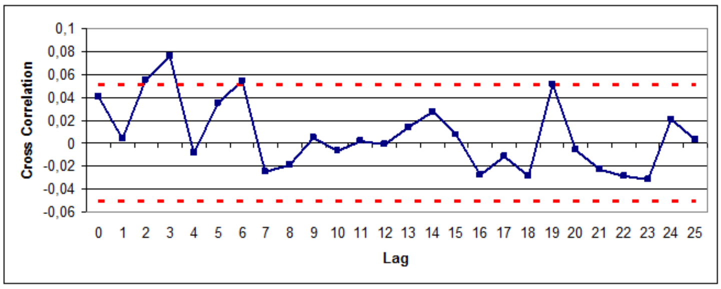
Figure 6.4: Cross correlation between propane and Brent log-returns for lag 0 to 25 together with 95% confidence bounds.

6.2.2 Natural Gas

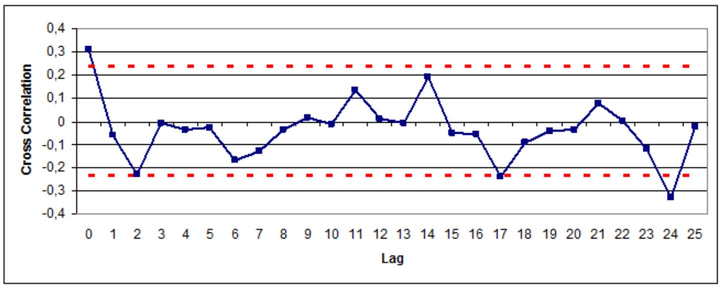
The cross correlation between natural gas and propane log-returns are shown in Figure 6.5. Due to the small amount of available data for natural gas the confidence bounds are much wider than for the other variables. For the daily data there are some correlation for lag 2 and 3, but it is small and will probably not explain the propane return. For monthly average data there are correlation at lag zero and lag 24. Even in this case the correlation is just a little bit over the confidence bounds and are probably not useful. The regression parameters in Table 6.1 and 6.2 verifies that there are no significant relation between natural gas and propane more than a very small connection for lag 2 and 3 for daily data and for lag 0 for monthly average data. Therefore the natural gas hypothesis is rejected for daily data and accepted for monthly average data, even though it seems much weaker than the crude oil hypothesis.

6.2.3 Naphtha

According to Table 6.1 and 6.2 above, the naphtha returns for lag $0-3$ for daily data and for lag $0-2$ for monthly average data can be used to explain the propane return. Naphtha is the explanatory variable with most significant regression parameters. As can be seen in Figure 6.6 the cross correlation is also higher than for the other factors, both for daily and monthly average data. Based on this information, the hypothesis that propane and naphtha prices are related is accepted.

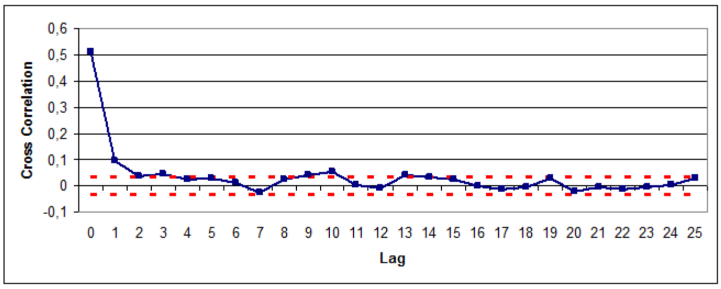


(a) Daily.

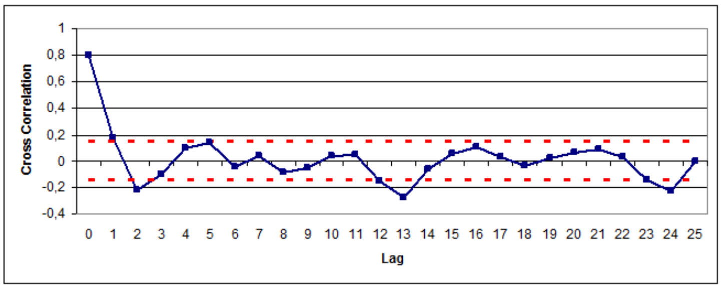


(b) monthly average.

Figure 6.5: Cross correlation between propane and natural gas log-returns for lag 0 to 25 together with 95% confidence bounds.

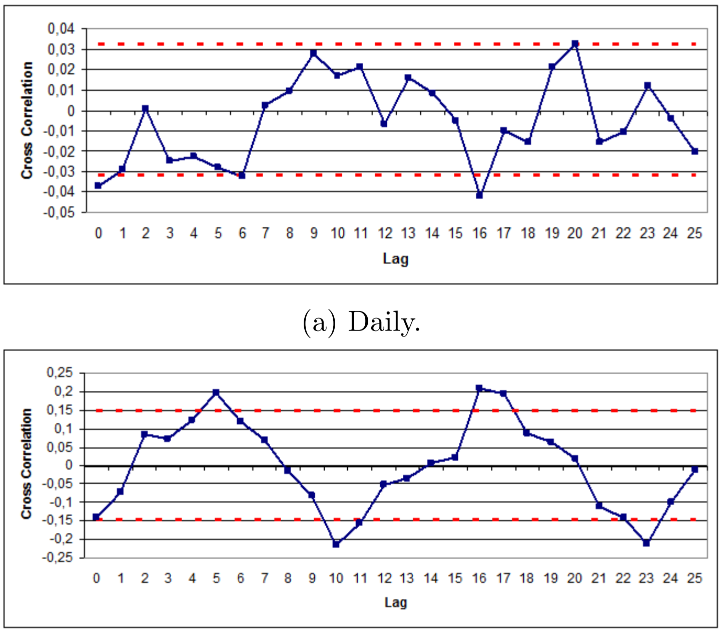


(a) Daily.



(b) monthly average.

Figure 6.6: Cross correlation between propane and naphtha log-returns for lag 0 to 25 together with 95% confidence bounds.



(a) Daily.

(b) monthly average.

Figure 6.7: Cross correlation between propane log-returns and one step difference temperature for lag 0 to 25 together with 95% confidence bounds.

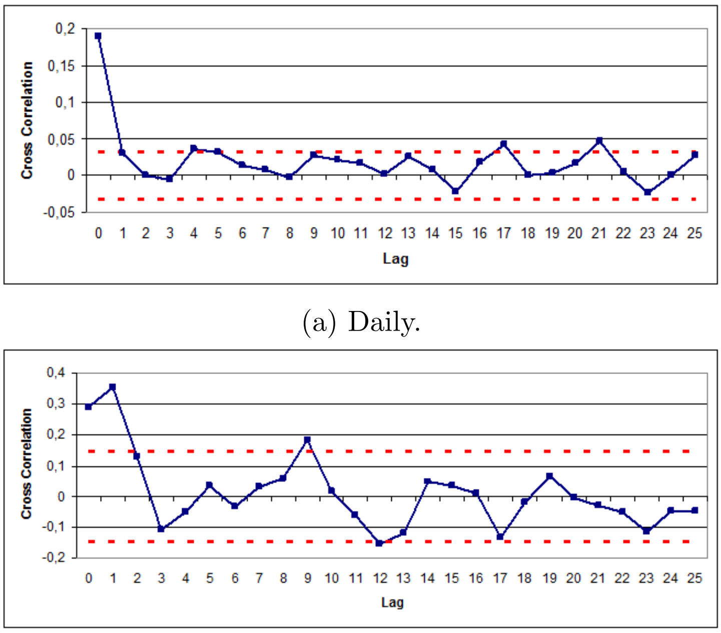
6.2.4 Seasonal Variation

Temperature

Table 6.1 and 6.2 as well as Figure 6.7 do not show much relation between propane returns and change in temperature, neither on daily nor monthly scale. The only significant regression parameter is for lag zero for daily data, where the parameter is negative. The negative value confirms the hypothesis that a decrease in temperature increases the price, but this relation seems to be very weak. The cross correlation has some significant values but these are very close to the confidence bounds. It is also hard to interpret why temperature changes 10 and 23 months from the current state would affect the return. The temperature hypothesis is rejected for these reasons.

Euro Stoxx 50 Index

For the Euro Stoxx 50 index, one interesting thing for monthly average data is observed in the cross correlation in Figure 6.8. The correlation between propane and the index return is significant for both lag zero and lag 1, but it is larger for lag 1 than for lag zero. This is interesting since it shows that previous values of the index return, which is observable, contains more information about the propane return one step forward in time than the index return in that time step. For daily data it is only the lag zero correlation that is significantly different from zero. The hypothesis about a connection between propane and Euro Stoxx 50 index returns is accepted for monthly average data. Since the Euro Stoxx 50 index is the factor used to quantify market and industrial conditions, accepting the hypothesis implies a belief that market and industrial conditions affect the propane price.



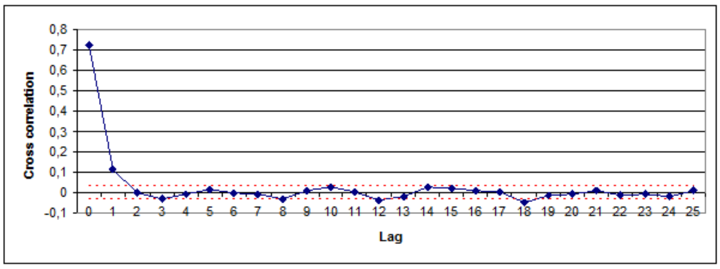
(a) Daily.

(b) monthly average.

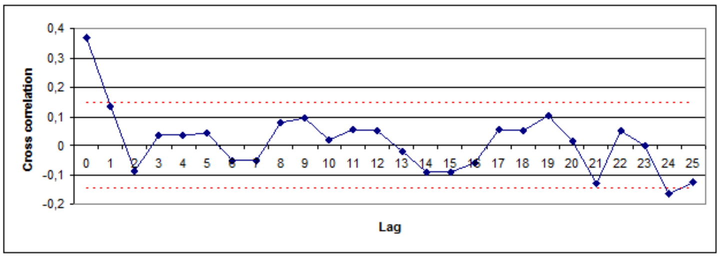
Figure 6.8: Cross correlation between propane and Euro Stoxx 50 index log-returns for lag 0 to 25 together with 95% confidence bounds.

6.2.5 East-West Spread

To evaluate the East-West spread hypothesis the East-West ratio will be used. The ratio is normalized compared to the real spread, making it more suitable since the real spread varies a lot over time while the ratio is more stable. The regression parameters and the cross correlations in Figure 6.9 shows that a relation between the NWE propane price and the East-West ratio exists. For lag zero this is obvious, since the ratio is defined using the price. Therefore, correlation for lags greater than zero are the only ones of interest. There are no such correlation for the monthly average data, but there is one significant correlation at lag 1 for the daily data. The East-West spread hypothesis is rejected for monthly average data and accepted for daily data.



(a) Daily.



(b) monthly average.

Figure 6.9: Cross correlation between propane and East-West ratio log-returns for lag 0 to 25 together with 95% confidence bounds.

6.3 Incorporating Hypotheses in Models

The evaluation of the driving factor hypotheses shows that the daily propane returns are related to the Brent, naphtha and East-West ratio returns while the monthly average returns are related to the Brent, naphtha, natural gas, and Euro Stoxx 50 index returns. Since the returns are related to each other, it is desirable to incorporate them when modeling the propane returns.

In this thesis the driving factors will be incorporated as external signals in ARMAX models. Since the aim is to predict propane returns at future time points, correlation with factors for lags greater than zero is of particular interest. If there is correlation for lag zero only, future values for the external signal have to be predicted before it is possible to predict the propane returns. This requires an accurate model for the external signal, a complex task that will not be investigated further in this thesis. Another approach is to use the SWAP and forward prices for months forward in time as the expected future value, since these values can be seen as the market's expectation of future prices. This is of course only possible for the factors where future contracts are available, namely Brent and naphtha. The SWAP and forward contract are only available for the average month price and cannot be used for the daily returns. For the other factors, ARMAX models with delayed external signal are used to overcome the problem. Since the natural gas return only shows correlation for lag zero in the monthly average case, and no SWAP/forward prices are available, gas prices will not be used as external signal in the ARMAX models.

6.4 Model Selection

The model selection work for daily and monthly average propane log-return is theoretically similar, but since the series have different properties and driving factor hypotheses, the two horizons will be considered separately. To simplify the work of calculating likelihood functions, it will be assumed that the driving noise in the models is normally distributed, even though most of the residuals seems to originate from a distribution with heavier tails.

From the analysis in section 6.1.1, it is expected that a time varying volatility model with AR, MA or ARMA structure, possibly with external signals consisting of Brent, naphtha, Euro Stoxx 50, and East-West ratio returns, will be a suitable model for the propane log-return. In ARMAX models, it is assumed that the noise term has constant variance over time. It is clear from Figure 6.1 that the propane log-returns do not have constant volatility. But if it is possible to model the volatility, the normalized log-returns

$$\frac{r_t}{\sigma_t}$$

can be assumed to have constant variance. Therefore a two step procedure where the volatility is modelled, followed by ARMAX modelling of the normalized returns is done.

	GARCH	EGARCH
Number of Parameters	3	4
Log-likelihood	2603.8	2602.6
Ljung-Box-Pierce test	Fail	Fail
McLeod-Li test	Pass	Pass
Monti test	Pass	Pass
Sign change test	Fail	Fail

Table 6.3: Validation results for volatility models of daily propane log-returns.

6.4.1 Volatility Modeling

The volatility will be modelled using the GARCH(1,1) model introduced in Definition 11 and the EGARCH(1,1) model given in Definition 12. The models are estimated for the estimation dataset and validated using the validation dataset. The daily and monthly average returns will be considered separately since they have different characteristics.

Daily Returns

In Table 6.3, validation results for the different volatility models are shown. Both models performs very similar, and passes two out of four whiteness tests. The GARCH model has a slightly higher log-likelihood value, indicating that, under the normal assumption, the GARCH model fits the data better than the EGARCH model. The number of parameters is smaller in the GARCH model, which also motivates that GARCH is a better choice than EGARCH. In the EGARCH model, the asymmetry parameter θ is not significant, showing that there is no asymmetry in the volatility data. Based on the above discussion, the GARCH(1,1) model is used to model the volatility for the daily propane log-returns.

Monthly Average Returns

Validation results for the different volatility models are shown in In Table 6.4. As for daily data, the results for GARCH and EGARCH are very similar. The log-likelihood shows that the GARCH model fits better under normal assumption. The EGARCH model passes three of four whiteness tests, while the GARCH model passes two. The asymmetry parameter θ in the EGARCH model is not significantly different from zero. Based on this, the GARCH model is chosen to model the volatility also for the monthly average propane log-returns.

6.4.2 ARMAX Modeling

Using a GARCH(1,1) model for volatility, the normalized returns can be computed. The normalized returns are then modelled using different ARMAX models. For both daily and monthly average data, ARMA models with no exogenous signals will be tested as a first step. For daily data, the ARMA models will be followed by ARMAX models, with external signals of brent, naphtha, and east-west ratio returns with at least one step delay. All external signals will be considered both separately and in combinations, by

	GARCH	EGARCH
Number of Parameters	3	4
Log-likelihood	47.35	46.65
Ljung-Box-Pierce test	Fail	Fail
McLeod-Li test	Pass	Pass
Monti test	Fail	Pass
Sign change test	Pass	Pass

Table 6.4: Validation results for volatility models of month propane log-returns.

	AR(1)	AR(2)	MA(1)	MA(2)	ARMA(1,1)	ARMA(2,2)
AIC	2941.4	2945.8	**2940.7**	2945.7	2945.9	2948.6
BIC	2946.4	2955.7	**2945.7**	2955.6	2955.8	2968.5
Ljung-Box-Pierce test	Pass	Pass	Pass	Pass	Pass	Pass
McLeod-Li test	Pass	Pass	Pass	Pass	Pass	Pass
Monti test	Pass	Pass	Pass	Pass	Pass	Pass
Sign change test	Fail	Pass	Fail	Pass	Pass	Pass

Table 6.5: Validation results for AR, MA and ARMA models for normalized daily propane log-returns. The lowest AIC and BIC values are in bold face.

allowing for more than one external signal in the ARMAX models. The external signals are delayed because there are no swap or forward prices available for the daily data. The monthly average returns will be modelled using ARMAX models with and without delays, and using brent, naphtha, and Euro Stoxx 50 index returns as external signals.

Daily Returns

The autocorrelation and partial autocorrelation functions in Figure 6.2 indicates that AR or MA models of order 1 to 2 should be sufficient. In Table 6.5, validation results for different AR, MA and ARMA models are shown together with AIC and BIC values. All models passes the whiteness tests, except the AR(1) and MA(1) model which both fail the sign change test. Both AIC and BIC have minimum values for the MA(1) model. Increasing the model order or extending to an ARMA model does not improve the result and seems to over fit the data. Both the AR(1) and MA(1) models perform similarly on the validation data. To choose one of the models, they are validated over the estimation data as well. Over the estimation data, the AR(1) model fits the data better than the MA(1) model. Therefore, the normalized returns are modelled using an AR(1) model.

Hopefully, it will be possible to improve the model further by extending the AR(1) model to an ARX model by adding exogenous signals. To choose suitable model orders and delays for the exogenous signals AIC and BIC will be used. In Table 6.6 and 6.7, AIC and BIC values for different exogenous signals of different orders and delays are presented. As can be seen in the tables, both AIC and BIC yields the same model order and delay

	Brent			Naphtha			East-West ratio		
Order	1	2	3	1	2	3	1	2	3
1	2944.3	2946.0	2947.7	2944.0	2946.4	2946.8	2944.1	2945.8	2947.7
Delay 2	**2943.1**	2945.5	2946.2	2944.0	2945.1	2946.2	**2943.4**	2946.6	2946.8
3	2943.8	2945.3	2946.3	**2942.5**	2944.5	2945.4	2944.6	2947.5	2948.7

Table 6.6: AIC values for ARX models of different orders and delays for normalized daily propane log-returns using different exogenous signals. The lowest value for each signal is in bold face.

	Brent			Naphtha			East-West ratio		
Order	1	2	3	1	2	3	1	2	3
1	2954.2	2960.9	2967.5	2953.9	2961.3	2966.7	2954.0	2960.7	2967.6
Delay 2	**2953.0**	2960.4	2966.0	2953.9	2960.0	2966.1	**2953.3**	2961.5	2968.4
3	2953.7	2960.1	2966.1	**2952.4**	2959.4	2965.2	2954.5	2962.3	2968.5

Table 6.7: BIC values for ARX models of different orders and delays for normalized daily propane log-returns using different exogenous signals. The lowest value for each signal is in bold face.

estimates. For brent, a first order model with two steps delay is suggested. For naphtha, a first order model with three steps delay is suggested. For the East-West ratio, a first order model with two steps delay is suggested. The naphtha model, is the model with lowest AIC and BIC values, which indicates that it fits the data better. The AIC and BIC values are higher for all ARX models than for the simpler AR(1) model, which indicates that adding the exogenous signals may not improve the result as wanted. In the next chapter all three exogenous models and the simple AR(1) model will be used for forecasting.

It is also possible to allow for more than one exogenous signal in the model. Here, a model with brent, naphtha and East-West ratio returns as inputs will be developed. To choose appropriate model orders and delays, AIC and BIC will be used. Since the number of possible models is very large, the AIC and BIC values will not be presented. When allowing for more than one exogenous signal, the lowest AIC and BIC values are again given by the simple AR(1) model. Therefore, no model with more than one exogenous signal will be investigated further.

Monthly Average Returns

The same procedure as for daily returns is now repeated for monthly average returns. Figure 6.3 suggest that an MA(2) or AR(2) model might be appropriate. In Table 6.8, validation results for different AR, MA and ARMA models are shown together with AIC and BIC values. As shown in the table, all the models passes most of the whiteness tests and can be considered to have white residuals. The lowest AIC and BIC values are generated by the MA(1) model. The AIC value for the AR(2) model is almost as low as for the MA(1) model. Combining AR and MA models in an ARMA model does not improve the result compared to increasing the order of the AR or MA model. The

	AR(1)	AR(2)	MA(1)	MA(2)	ARMA(1,1)	ARMA(2,2)
AIC	130.15	128.60	**128.34**	129.35	130.33	134.34
BIC	132.06	132.42	**130.25**	133.18	134.15	141.99
Ljung-Box-Pierce test	Fail	Pass	Pass	Fail	Pass	Pass
McLeod-Li test	Pass	Pass	Pass	Pass	Pass	Pass
Monti test	Pass	Pass	Pass	Pass	Pass	Pass
Sign change test	Pass	Pass	Pass	Pass	Pass	Pass

Table 6.8: Validation results for AR, MA and ARMA models for normalized monthly average propane log-returns. The lowest AIC and BIC values are in bold face.

		Brent			Naphtha			Euro Stoxx 50		
	Order	1	2	3	1	2	3	1	2	3
	0	**123.37**	124.48	126.71	**118.99**	119.11	121.69	–	–	–
Delay	1	131.08	133.00	135.33	**130.23**	133.68	137.04	130.35	132.78	133.99
	2	**130.53**	133.24	135.09	131.62	134.87	137.71	**130.10**	131.68	133.97
	3	131.26	133.35	135.17	130.91	134.33	136.44	130.46	132.96	135.39

Table 6.9: AIC values for ARX models of different orders and delays for normalized monthly average propane log-returns using different exogenous signals. The lowest values, both with and without delay, for each signal are in bold face.

validation dataset is limited to only 50 samples, so to choose between MA(1) and AR(2) validation is done over the estimation data too. Over the estimation data, the AR(2) model generates a better fit. Therefore the AR(2) model is chosen as a first model for the normalized propane monthly average log-returns.

As mentioned earlier, the PACF in Figure 6.3 indicates that there are some dependency for lag 12. Therefore the AR(2) model is extended by adding a parameter for lag 12. In fact that is an AR(12) model where parameter 3 to 11 are forced to zero, giving three parameters to estimate. The AR(12) model, passes two of four whiteness tests and has AIC value 125.23 and BIC value 130.97. Both the AIC and BIC values are lower than for the AR(2) model. Both the AR(2) and AR(12) model will be used for forecasting in the next chapter.

As for daily data, the AR models will be extended to ARX models with exogenous signals of brent, naphtha and Euro Stoxx 50 index returns, which hopefully can improve the performance. For brent and naphtha, swap and forward prices are available. By using them as predictions of future prices when forecasting, exogenous signals with no delay can be used. Table 6.9 and 6.10 shows AIC and BIC values for the AR(2) model extended with exogenous signals of different orders and delays. The tables shows that the minimum AIC and BIC values are for the ARX model with a first order input of naphtha with zero delay. All the models with zero delay have lower AIC and/or BIC values than the original AR(2) model. The models with one or more steps delay all have higher AIC and BIC values than the original AR(2) model. The higher AIC and BIC values, indicates

		Brent			Naphtha			Euro Stoxx 50		
	Order	1	2	3	1	2	3	1	2	3
	0	**129.11**	132.13	136.27	**124.72**	126.76	131.25	–	–	–
Delay	1	136.82	140.65	144.89	**135.97**	141.33	146.60	136.08	140.42	143.55
	2	**136.27**	140.89	144.65	137.35	142.52	147.27	**135.83**	139.33	143.53
	3	137.00	141.00	144.73	136.65	141.98	146.00	136.20	140.60	144.95

Table 6.10: BIC values for ARX models of different orders and delays for normalized monthly average propane log-returns using different exogenous signals. The lowest values, both with and without delay, for each signal are in bold face.

that to improve the AR(2) model ARX models with zero delay should be used. If it is possible to improve the result using zero delay models will be evaluated further in the next chapter where forecasting is done. The ARX models that will be used for forecasting is, first order brent with zero delay, first order naphtha with zero delay, first order brent with two steps delay, first order naphtha with one step delay and first order Euro Stoxx 50 with two steps delay. Extending the AR(12) model to an ARX model with the same inputs as above yields higher AIC and BIC values for all combinations and these models will not be considered further.

When allowing for more than one exogenous signal, the minimum AIC (120.45) and BIC (130.00) values is given by the ARX model with second order naphtha with zero delay and first order Euro Stoxx 50 with four steps delay. These AIC and BIC values are higher than for the zero delayed naphtha ARX model. Therefore, this model will not be considered further.

7 Models and Forecasting Performance

In the previous chapter a couple of models were developed for the propane returns. Now they will be used for forecasting the propane returns and evaluated according to how well they perform. The forecasting is done over the validation dataset. The models will be evaluated using two different measures, the root mean squared error (RMSE) and the hit rate. The root mean squared error shows how close the forecasts is to the actual outcome while the hit rate shows if the model is able to forecast the direction of future movements. The RMSE will be calculated for the estimates of both the log-return and true price.

All models that will be used for forecasting are of ARX type

$$A(z)r_t = B(z)x_t + e_t. \tag{7.1}$$

The ARX models are used to forecast the normalized propane log-return. A forecast of the true log-return is generated by multiplying the normalized log-return forecast with the forecast of the volatility. As mentioned before, the volatility is modelled using a GARCH(1,1) model

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2. \tag{7.2}$$

Below the model parameters and the forecasting results are presented. First, the models for daily data will be considered. Then, the models for monthly average data will be considered.

7.1 Daily

The daily volatility is modelled with the GARCH(1,1) model in equation 7.2 with estimated parameter values

$$\omega = 7.32 \cdot 10^{-6} \quad \alpha = 0.0742 \quad \beta = 0.9152.$$

In the previous chapter, four different models for the normalized propane log-return were suggested. The estimated polynomials in equation 7.1 for the different models are presented below.

1. The simple AR(1) model is given by

$$A(z) = 1 - 0.1447z^{-1}.$$

	AR(1)	ARX (brent)	ARX (naphtha)	ARX (EW-ratio)
Hit Rate (%)	51.52	52.48	50.86	51.24
RMSE (return)	0.0203	0.0203	0.0203	0.0203
RMSE (price)	16.94	16.95	16.93	16.94

Table 7.1: Forecasting results for the different models for daily propane data. Hit rates in bold face are significantly better than guessing at the 5% level.

2. The ARX model with brent as input, denoted ARX (brent), is given by

$$A(z) = 1 - 0.1412z^{-1} \qquad B(z) = 1.3380z^{-2}.$$

3. The ARX model with naphtha as input signal, denoted ARX (naphtha), is given by

$$A(z) = 1 - 0.1412z^{-1} \qquad B(z) = 1.3680z^{-3}.$$

4. The ARX model with EW-ratio as input signal, denoted ARX (EW-ratio), is given by

$$A(z) = 1 - 0.1450z^{-1} \qquad B(z) = -0.7303z^{-2}.$$

The 1-step forecasting results for these models are presented in Table 7.1. As can be seen, all the models performs similarly with a hit rate just over 50% and a root mean squared error of 16.9 USD/t, compared to the daily average price change of ±7.8 USD/t. Since the forecast is only one day ahead, the results are very poor. It also shows that adding external signals to the AR(1) model does not improve forecasting performance. The 1-step predictions are bad and predictions over longer horizons are even worse and will not be discussed further. To conclude, all the models for daily propane prices perform poorly. If one would use a model for the daily propane price, the simple AR(1) model is probably the best choice.

7.2 Monthly Average

The monthly average volatility is modelled using a GARCH(1,1) model from equation 7.2 with estimated parameter values

$$\omega = 0.0030 \quad \alpha = 0.4092 \quad \beta = 0.3971.$$

In the previous chapter, two AR models, three ARX models with delayed exogenous signals and two ARX models with no delay of the exogenous signals were suggested. The estimated polynomials in equation 7.1 for the different models are presented in Table 7.2 and the forecasting results are presented in Table 7.3.

	AR		ARX delayed			ARX forward	
	AR(2)	AR(12)	brent	naphtha	Euro Stoxx 50	brent	naphtha
$A(z)$	$1 - 0.2553z^{-1}$ $+0.09825z^{-2}$	$1 - 0.2678z^{-1}$ $+0.1272z^{-2}$ $+0.2178z^{-12}$	$1 - 0.255z^{-1}$ $+0.1158z^{-2}$	$1 - 0.2898z^{-1}$ $+0.08913z^{-2}$	$1 - 0.2425z^{-1}$ $+0.1104z^{-2}$	$1 - 0.1352z^{-1}$ $+0.09011z^{-2}$	$1 + 0.03183z^{-1}$ $+0.01061z^{-2}$
$B(z)$			$0.2604z^{-2}$	$-0.414z^{-1}$	$1.085z^{-2}$	6.348	6.489

Table 7.2: Estimated polynomials for the different models for monthly average normalized propane log-returns.

		AR		ARX delayed			ARX forward	
		AR(2)	AR(12)	brent	naphtha	Euro Stoxx 50	brent	naphtha
1-step	Hit Rate (%)	**68.75**	**71.05**	**68.75**	**70.83**	**66.67**	**64.58**	52.08
	RMSE (return)	0.0837	0.0887	0.0836	0.0831	0.0835	0.0876	0.0972
	RMSE (price)	70.06	74.42	70.01	69.79	69.80	73.00	80.70
2-step	Hit Rate (%)	61.70	**67.57**	63.83	61.70	57.45	53.19	46.81
	RMSE (return)	0.0949	0.1003	0.0949	0.0954	0.940	0.0948	0.0966
	RMSE (price)	115.77	128.30	116.48	114.09	115.21	126.64	138.80
3-step	Hit Rate (%)	56.52	55.56	56.52	58.70	56.52	47.83	47.83
	RMSE (return)	0.0961	0.1020	0.0958	0.0962	0.0959	0.0967	0.0975
	RMSE (price)	149.61	171.57	151.33	146.18	149.45	162.43	169.19

Table 7.3: Forecasting results for the different models for monthly average propane data. Hit rates in bold face are significantly better than guessing at the 5% level.

The forecasting results in Table 7.3 show that all models have difficulties to predict future propane prices. The model with lowest root mean squared error for the true price is the ARX model with delayed naphtha as input. Using that model, the mean prediction error is around 70 USD/t for one step, 114 USD/t for two steps and 146 USD/t for three steps, which is high since the propane price is in the range $600 - 1050$ USD/t over the period. Comparing the ARX model with the simple AR(2) model, the RMSE for true prices are just a little bit better for the ARX model than for the AR(2) model. To predict future true propane prices the best choice is to use either the ARX with delayed naphtha as input or the simple AR(2) model.

Since all models performs bad when forecasting true prices, the hit rate may be a more interesting measure to analyse. The AR(12) model has the highest hit rate for one and two steps forecasting, with around 71% and 68% hits respectively, which is significantly better than the hit rate from guessing. Over the validation period, 48% of the true log-returns were negative and 52% were positive, which shows that the random walk assumption holds. The AR(12) model performs better than guessing for one and two steps forecasting, which shows that it is able to predict the direction of future price movements well. To predict future price movement direction the best choice is to use the AR(12) model.

The ARX models without delay, where swap and forward prices are used as future values, performs worse than all the other models. They have lower hit rate and larger RMSE. This is a bit surprisingly, but it shows that the forward and swap prices are no good predictors of the future prices, even though they can be considered as the market expectation prices.

8 Conclusions and Discussion

The aim of this thesis was to get a deeper insight into factors that affect the propane price in NWE and to develop a model for prediction of future propane prices, both on daily and monthly horizons. Throughout the thesis, driving factor hypothesis have been evaluated and models have been developed, thus fulfilling the aim. The driving factor analysis shows interesting results while the forecasting performance results are mixed, despite this all the results presented in the thesis can be useful in propane trading.

Starting with the driving factor hypotheses, it is shown that there are especially two factors, brent and naphtha, that are related to the propane price. A bit surprisingly, the naphtha seems to affect the propane price more than brent. The most probable reason for the strong naphtha connection, is the petrochemical industry, which is a large consumer of both naphtha and propane in NWE. The strong naphtha connection might be a local fact and it would have been interesting to know if the naphtha price is more related to the propane price all around the globe.

The natural gas hypothesis, which was partly based on the same reasoning as the brent hypothesis, were rejected. In the narrow future, the amount of propane produced from natural gas will increase and natural gas will replace oil in some applications. Therefore, propane might be more related to natural gas than naphtha in the future.

Another interesting result, is that no relation between the outdoor temperature and the propane price was found. Still, there are some seasonal variation in the propane price, but it is not because of the temperature which was the original hypothesis. Instead it is shown that the variation may be related to business cycles and the overall economic situation, in this work quantified by the Euro Stoxx 50 index.

The East-West spread, was shown to have limited relation to the propane price. There are of course arbitrage opportunities arising from a large spread, but it is almost impossible for companies in NWE to utilise these opportunities. Partly because of the high transportation costs, but also because of high propane prices within NWE compared to the U.S..

Based on an analysis of the propane prices itself and the analysis of driving factor hypothesis, a couple of models for the propane price were developed, both for daily and monthly average prices. The models for the daily propane price were unfortunately not very successful. The best model, an AR(1) model, performed bad in forecasting and had the same hit rate as a guessing. These results were disappointing, but shows how hard it is to find good models for financial time series. It is also worth mentioning that the NWE propane market, especially the daily trading, is very illiquid and there are many days where no trades take place but the listed price changes heavily anyway. The unexplained changes are hard for a model to capture. All the trades are made OTC, and some deals are un-official, which is impossible to incorporate in the models.

For the monthly average propane price, which is the one of largest interest, the model performance results are better, but still not good. It is very hard to forecast the propane price forward in time, but some models performs better than others. The simple AR(2) model and the ARX with delayed naphtha as exogenous signal perform best in the RMSE sense. The most interesting result is that the AR(12) model can forecast the direction of the propane price movements one and two months forward with an accuracy of over 70% and 65% respectively. This is a good result that could be very useful, but it says nothing about the size of future price movements. One thing, that could be used as an indicator of the movement size, is the volatility forecast.

To develop this work further, at least two things can be done. First, supply and demand of propane are ignored in this thesis, but it is reasonable to assume that supply and demand affect the propane price. If it is possible to quantify the difference between supply and demand, incorporating the difference in an ARX model may improve the forecasting. Secondly, in this work the models are assumed to be constant over time. When market conditions change, the estimated parameters in the models for normalized returns and volatility may change. To capture these changes, one approach is to re-estimate the models every time step, or at least after a certain number of time points. Such re-estimation has not been tested in the thesis, but it would have been interesting to see if it could improve the results.

References

Akaike, H. A New Look at Statistical Model Identification. *IEEE Transaction on Automatic Control*, 19:716–723, 1974.

Axelsson, R., Holmlund, B., Jacobsson, R., Löfgren, K-G., and Puu, T. *Mikroekonomi*. Studentlitteratur, 2nd edition, 1998.

Bass, R. F. *Stochastic Processes*. Cambridge University Press, 2011.

Björk, T. *Arbitrage Theory in Continuous Time*. Oxford University Press, 3rd edition, 2009.

Blom, G., Enger, J., Englund, G., Grandell, J., and Holst, L. *Sannolikhetsteori och Statistikteori med Tillämpningar*. Studentlitteratur, 5th edition, 2005.

Bollerslev, T. Generalized Autoregressive Conditional Heteroscedasticity. *Econometrica*, 31:307–327, 1986.

Box, G. E. P. and Pierce, D. A. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 65:1509–1526, 1970.

Brooks, C. *Introductory Econometrics for Finance*. Cambridge University Press, 2008.

Energigas Sverige, . Gasol, March 2008. URL `http://energigas.se/Energigaser/~/media/Files/www_energigas_se/Publikationer/Infomaterial/Gasolbroschyr.ashx`. [Acessed 2014-02-24].

Engle, R. F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, 50:987–1007, 1982.

Engle, R. F., Lilien, D. M., and Robins, R. P. Estimating Time Varying Risk Premia in the Term Stucture: The ARCH-M Model. *Econometrica*, 55:391–407, 1987.

European Climate Assessment and Dataset, . Daily Mean Temperature TG, 2014. URL `http://www.ecad.eu/utils/downloadfile.php?file=download/ECA_blend_tg.zip`. [Accessed 2014-03-27].

European LPG Association, . The LPG Industry Roadmap, 2009. URL `http://aegpl.eu/media/16783/the%20lpg%20industry%20roadmap%20ed.%202009.pdf`. [Accessed 2014-04-08].

European LPG Association, . European LPG Sector Overview 2010, September 2010.

Franke, J., Härdle, W., and Hafner, C. M. *Statistics of Financial Markets: An Introduction.* Springer, 2004.

Gaspoint Nordic, . European Brent Spot Price FOB, 2014. URL `http://www.gaspointnordic.com/market-data?t=1397042719`. [Accessed 2014-04-14].

Jakobsson, A. *Time Series Analysis and Signal Modeling.* Book draft version: 121015, 2012.

Kashyap, R. L. Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:99–104, 1982.

Kingman, J. F. C. *Poisson Processes.* Oxford University Press, 1993.

Lindgren, G., Rootzén, H., and Sandsten, M. *Stationary Stochastic Processes.* Lund University: Centre for Mathematical Sciences Mathematical Statistics, 2009.

Ljung, G. M. and Box, G. E. P. On a Measure of a Lack of Fit in Time Series Models. *Biometrika*, 65:297–303, 1978.

Madsen, H., Nygaard Nielsen, J., Lindström, E., Baadsgaard, M., and Holst, J. *Statistics in Finance.* Lund Institute of Technology: Centre for Mathematical Science, 2004.

McLeod, A. J. and Li, W. K. Diagnostic Checking ARMA Time Series Models using Squared-Residual Correlations. *Journal of Time Series Analysis*, 4:269–273, 1983.

Monti, A. C. A Proposal for a Residual Autocorrelation Test in Linear Models. *Biometrika*, 81:776–780, 1994.

Nelson, D. B. Conditional Heteroscedasticity in Asset Returns: A new Approach. *Econometrica*, 59:347–370, 1991.

Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.*, 6:461–464, 1978.

Shelley, C. *The Story of LPG.* Poten & Partners, 2nd edition, May 2003.

Staff, . Petrochemical Processes 2001. *Hydrocarbon Processing*, pages 71–246, 2001.

STOXX, . STOXX homepage, 2014. URL `www.stoxx.com`. [Accessed 2014-06-23].

The World LP Gas Association, . Statistical Review of Global LP Gas, September 2010.

U.S. Energy Information Administration, . U.S. Refinery Yield, September 2013. URL `http://www.eia.gov/dnav/pet/pet_pnp_pct_dc_nus_pct_a.htm`. [Accessed 2014-04-10].

U.S. Energy Information Administration, . European Brent Spot Price FOB, 2014. URL `http://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RBRTE&f=D`. [Accessed 2014-03-15].

Varian, H. R. *Intermediate Microeconomics: A Modern Approach.* Norton & Company, 8th edition, 2009.

Wang, P. *Financial Econometrics.* Routledge, 2003.

Wunderground, . Väder Historik för London, United Kingdom, 2014. URL `http://www.wunderground.com/history/airport/EGLL/2004/12/1/MonthlyHistory.html`. [Accessed 2014-03-27].

Appendix A - Temperature Data

Unfortunately there are no average temperature data available for the NWE region. Since it is desirable to investigate the seasonal variation and how the temperature affects the LPG price, temperature data for the whole region has to be created. Of course one could have taken the temperature from just one place in NWE and used it as an indicator for the temperature in the whole region since it is probably either cold or hot in the whole region. Here an weighted average temperature which is described below will be used instead.

The usage area of LPG that is most sensitive to outdoor temperature is space heating, which is included in the domestic sector of LPG consumers. To get an average temperature that reflects the demand of LPG for space heating a weighted average of temperatures from one point in every country in NWE (Gent-Belgium, Copenhagen-Denmark, Berlin-Germany, Dublin-Ireland, Amsterdam-Netherlands, Oslo-Norway, Stockholm-Sweden and London-United Kingdom) are used as the regional average temperature. The weights for each point is calculated as the LPG consumption used in the domestic sector within the country divided by the total consumption of LPG in the domestic sector in NWE. The consumption of LPG in the domestic sector (European LPG Association, 2010) and the weights used in the weighted average is shown in Table A.1.

As is clear from the table the weighted average temperature will heavily depend on the temperature in Berlin and London. All the temperature data is available from European Climate Assessment and Dataset (2014) except for London which are available from Wunderground (2014).

From the weighted average temperature it is possible to create "normal" temperature in NWE for the different months. Since the data consists of only 15 years of observations one should keep in mind that the "normal" temperature may be incorrect. The "normal" temperatures are more often based on averages over longer periods, typically 30 year. The "normal" temperatures were simply calculated as the mean of the weighted average temperature over the different months and is shown in Table A.2.

	Domestic sector consumption (kt)	Weight
Belgium	95	0.0685
Denmark	14	0.0101
Germany	838	0.6042
Great Britain	311	0.2242
Ireland	55	0.0397
Netherlands	55	0.0397
Norway	14	0.0101
Sweden	5	0.0036
Total	**1387**	**1**

Table A.1: Usage of LPG in the domestic sector for every country in NWE together with the weight for the weighted average temperature in NWE.

	Normal temperaure	Standard deviation
January	2.9	1.8
February	3.2	1.9
March	5.6	1.7
April	10.3	1.3
May	14.2	1.3
June	17.2	1.0
July	19.2	1.7
August	18.8	1.1
September	15.2	1.4
October	10.8	1.6
November	6.6	1.0
December	3.2	2.2

Table A.2: Monthly ”normal” weighted average temperature in NWE and the corresponding standard deviation.

Master’s Theses in Mathematical Sciences 2014:E44
ISSN 1404-6342

LUTFMS-3259-2014

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/