Predicting Crude Oil Market Value Based on Crude Oil Properties

Emil Kauppi

School of Science

Bachelor's thesis
Espoo 27.8.2020

Supervisor

Asst. Prof. Pauliina Ilmonen

Advisor

DSc (Tech.) Jarno Kohonen

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Emil Kauppi

Title Predicting Crude Oil Market Value Based on Crude Oil Properties

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Sciences **Code of major** SCI3029

Teacher in charge Asst. Prof. Pauliina Ilmonen

Advisor DSc (Tech.) Jarno Kohonen

Date 27.8.2020 **Number of pages** 22+1 **Language** English

Abstract

Crude oil market price prediction holds an extensive research field. Methods for predicting crude oil prices range from different econometric methods to deep learning frameworks. The research is mostly focusing on the most common (measured in market volume) crude oil types and benchmark crude oils. Therefore, it is of interest to study the pricing relationships of other crude oils and how they depend on crude oils' properties.

Crude oil is obtained from deep within the earth and extracted by drilling. Crude oil itself does not have any demand on the market, but the products able to be refined from it have. The extraction of market products from crude oil is done at dedicated oil refineries. Different product streams with different market values arise from a barrel of crude oil. Hence, more valuable product streams yield a higher price for the crude oil.

We motivate the use of a set of linear models with simple crude oil market assumptions. Efficient markets with a marginal refinery are assumed. The marginal refinery is defined as the simplest configuration of a refinery, a straight run refinery, including only one distillation unit. The marginal refinery is operating in a break-even environment, hence setting the price of crude oil.

This thesis aims to validate the predictive power of evolving product prices and crude oil properties in predicting crude oil market differences. The difference is the difference between a benchmark crude oil and crude oil itself. Two different linear regression models and ARIMAX modeling are used in making predictions. Prediction error metrics for the models are gathered. Results are then compared to the historical average value of the price difference, which is treated as a benchmark model.

We conclude that product prices indeed hold a predictive power in predicting the prices for most of the researched crude oils compared to the benchmark model used. However, the methodologies include flaws and simplifications, for instance, regarding market price outliers and assumptions of the market structure. Future research areas could include developing a unified crude pricing model and using a different marginal configuration on the market.

Keywords Crude Oil, Price Prediction, Linear Regression, ARIMAX, Feature Engineering, Marginal Configuration

Aalto University
School of Science

Aalto-universitetet, PB 11000, 00076 AALTO
www.aalto.fi
Sammandrag av kandidatarbetet

Författare Emil Kauppi

Titel Att förutse oljepriser med hjälp av egenskaper hos råoljor

Utbildningsprogram Teknisk fysik och matematik

Huvudämne Matematik och systemanalys **Huvudämnets kod** SCI3029

Ansvarslärare Bitr.prof. Pauliina Ilmonen

Handledare TkD Jarno Kohonen

Datum 27.8.2020 **Sidantal** 22+1 **Språk** Engelska

Sammandrag

Det existerar en bred och omfattande forskningslitteratur för förutsägande av råoljepriser. Metoderna omfattar allt ifrån traditionella ekonometriska metoder till maskininlärningstekniker. Denna litteratur är koncentrerad runt välkända, och till marknadsvolymen stora, råoljor. Det uppstår därmed ett behov för kartläggning av hur andra, till marknadsvolymen mindre, råoljor prissätts i förhållande till de viktigaste råoljorna, och i förhållande till varandra.

Råolja är en vätskeformad råvara vars ursprung är jordens berggrund. Råoljan i sig själv har ingen efterfrågan men däremot har de raffinerade produkterna det. Raffineringen görs på särskilda oljeraffinaderier där olika produktströmmar tas fram genom t.ex. kemisk destillering. Mängden värdefulla produktströmmar som kan produceras bestämmer råoljans värde på marknaden och härrör sig från råoljans kemiska egenskaper.

Målet med detta arbete är att modellera priser på råoljor som en funktion av priser på slutprodukter. Närmare sagt är priserna uttryckta som differensen till en referensråolja. Vi antar att det på marknaden existerar ett marginaloljeraffinaderi som opererar med nollresultat. Dessutom antar vi att detta raffinaderi innehar den enklast möjliga konfigurationen, och därmed innehåller enbart en destilleringsenhet. Dessa antaganden motiverar modellvalen i arbetet. Hypotesen för prissättningen på marknaden blir därmed att priserna i hög grad styrs av produktpriser på de raffinerade produkterna och därmed anknyter till råoljeegenskaperna.

För att bekräfta detta konstruerar vi två linjära regressionsmodeller samt en ARIMAX-modell, med produktpriser som förklarande variabler. Dessa modeller strävar sedan till att förutse prisdifferensen till referensoljan. Resultaten jämförs med historiska, viktade medeltal för att fastställa att produktpriserna och råoljans egenskaper har någon förmåga att förutsäga priser.

Resultaten visar att de metoder och antaganden som används i viss mån förklarar prisutvecklingen hos de råoljor som undersöktes. Det enkla antagandet som gjordes gällande marknadsstrukturen är en brist som kunde forskas vidare. Andra förslag på vidare forskning är t.ex. utveckling av en enhetlig prismodell och hur man beaktar yttre, negativa marknadshändelser i modellerna.

Nyckelord Råolja, Råoljepriser, Oljeraffinering, Marginalaktör, Förutsägning av marknadspriser, Regressionanalys, ARIMAX, Tidsserieanalys

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Emil Kauppi

Työn nimi Raakaöljyhinnan ennustaminen raakaöljyominaisuuksia käyttäen

Koulutusohjelma Teknillinen Fysiikka ja Matematiikka

Pääaine Matematiikka ja Systeemianalyysi **Pääaineen koodi** SCI3029

Vastuuopettaja Apul.prof. Pauliina Ilmonen

Työn ohjaaja TkT Jarno Kohonen

Päivämäärä 27.8.2020 **Sivumäärä** 22+1 **Kieli** Englanti

Tiivistelmä

Raakaöljyjen hintamaailmojen ennustamiseen on olemassa mittava tutkimuskirjallisuus. Ennustusmenetelmät vaihtelevat neuroverkoista perinteisiin ekonometrisiin menetelmiin. Tutkimuskirjallisuuden puutteet ovat menetelmien keskittyminen myyntimääriltään suurimpiin raakaöljyihin. Tämän myötä muodostuu tarve pienempien raakaöljyjen hintamaailman hinnoittelun tutkimiseen suhteessa toisiinsa ja niin sanottuihin referenssilaatuihin nähden.

Raakaöljy on syvältä maan sisältä lähtöisin oleva nestemäinen raaka-aine, jolle ei itsessään ole kysyntää. Näin ollen, raakaöljy jalostetaan, saaden erilaisia tuotejakeita. Yksinkertaisimmillaan tämä saavutetaan pelkän tislauksen avulla. Tuotejakeet riippuvat raakaöljyjen ominaisuuksista ja mitä arvokkaampia tuotejakeita raakaöljystä on mahdollista saada, sitä arvokkaampi raakaöljy on kyseessä.

Tämän kandidaatintyön tavoite on tutkia raakaöljyjen hintojen muodostumista jalostettujen tuotehintojen funktiona. Olettamalla marginaalitoimija, joka toimii nollatuloksella, ja edelleen olettamalla tämä suoratislaus jalostamoksi voidaan olettaa nämä riippuvuudet linaarisiksi. Tämän perusteella voidaan ongelmaa lähestyä perinteisten linaariregressiomallien ja ARIMAX mallien kautta.

Työssä käytetään kahta eri linaarista regressiomallia ja yhtä ARIMAX mallia hintaerotusten mallintamiseen, erotuksen ollessa referenssilaatuun. Ennustustuloksia vertaillaan niin sanottuun benchmark malliin, joka tässä työssä määritellään olemaan painotettu historiallinen keskiarvo hintaerotuksista referenssilaatuun.

Työn tulokset osoittavat että suurimmassa osassa tutkituista raakaöljyistä lopputuotteiden hinnoilla ja raakaöljyn ominaisuuksilla on ennustuskykyä raakaöljyn hinnankehitykselle. Ennustusvirheet ovat pienemmät kaikille työssä esitellyille raakaöljyille benchmark malliin nähden. Mallinuksilla on silti puutteita esimerkiksi markkinashokkien huomioimisessa, eli mallin aikasarjapisteiden painotuksessa. Jatkotutkimusehdotuksiin kuuluvat esimerkiksi yhtenäisen raakaöljymallin kehittäminen ja toisen marginaalijalostajatyypin käyttö.

Avainsanat Raakaöljy, Raakaöljyn hinta, Ennustaminen, Lineaarinen regressioanalyysi, Aikasarja-analyysi, ARIMAX-malli, Öljyn jalostus, Marginaalijalostaja

Contents

1 Introduction

Predicting oil market prices is an important task, both for the different players in the industry and societies in general. It is because oil is the primary fuel source and still an essential source of energy in the modern economy. In 2004, more than 161 different crude oils were trading on the market (see The International Crude Oil Market Handbook, 2006). Fluctuations in the oil price may have significant impacts on a global scale, as seen in the oil crisis of the 1970s. Therefore, it is desirable to predict the development of crude oil prices, not at least for the oil refining companies producing the products from crude oils.

The refining companies' goal is to optimize the product yields from the crude oils to be as profitable as possible concerning the purchases of feedstocks, consisting mostly of crude oils. In order to make successful decisions on purchases of crude oil and the planning of refinery operations, it is essential to understand the price development of the crude oils in the future.

Benchmark crude oils have attracted much attention in academia. There are many methods for predicting future price development, with new findings and methods published at a constant rate. Benchmark crude oils are used as reference prices in the oil market. These benchmark crude oils include Brent Crude, which is a basket of crude oils extracted from the North Sea, WTI (West Texas Intermediate) (Inkpen and Moffett, 2011) extracted mostly from Midwest and Gulf Coast regions of the USA and Urals crude extracted in Russia (McKinsey, 2020d). However, for most crude oils, the market valuations with respect to each other are unclear.

Strong correlations between price movements are observable because of strong market integration. Variations in the price differences, with respect to a benchmark crude oil, could be caused by numerous factors. In the short run, the variation could be caused by factors such as shortages in supply or changes in market conditions. However, in the long run, the value on the market should reflect the crude oil properties, compared to a benchmark crude. Crude oil quality and properties determine the product yields of products obtained from refined crude oil after being processed at an oil refinery.

The prices of product fractions are often described on the market with a product spread, which describes the price of an oil product as a difference to some benchmark crude oil. The prices referred to in this thesis are price quotes, since the products are not trading according to some standard price list. The prices are determined on the market.

This thesis's objective is to address if the demand for different oil products has any predictive power in predicting crude oil price as a difference to a benchmark crude. If successful, with predictions of the desired benchmark crude oil and the expected product spreads set as predictors, an estimation for the price of crude oil can be made.

The thesis is organized as following. Background for the objective and a literature review is provided in chapter 2. The data used for the task and methods for developing and evaluating different pricing models are described in chapter 3. Results and evaluation of out-of-sample forecasts with the described models are provided in

chapter 4. Finally, conclusions from the analysis are made in chapter 5.

2 Background

Crude oil is a liquid compound, found deep within the earth, and extracted to the surface by drilling. The chemical composition of crude oil consists of a multitude of different hydrocarbon compounds. The geographic location of the source determines the crude oil's physical and chemical properties, originating from the geological processing in the location. Different structures in the hydrocarbons yield in varying properties such as density and predict other existing chemical compounds (see National Academies of Sciences et al., 2016). Properties determine the product yields possible to obtain from a specific crude oil when refined at oil refineries (see Fahim et al., 2010).

The products refined from crude oil have a particular demand on the market, which should reflect the crude oil market value. This type of reasoning refers to the gross product worth (GPW) of one crude oil barrel, which describes the total value of the product streams obtained from the processed crude oil at a refinery. A strong correlation of GPW and price of crude oil is observable, as seen in Secretariat (2011).

Besides the properties, crude oil value depends on the drilling capacity set to meet the demand. Therefore, prices are also affected in the reverse direction as well since higher crude oil prices (caused by, for instance, supply shortages) drive higher refining costs affecting product prices. That is to say, the properties of a specific crude oil should be reflected in its market valuation. Crude oils, which give higher fractions of in-demand products, are more valuable. This yields in pricing dynamics possible to be utilized when estimating crude oil prices.

2.1 Oil Refineries

Oil refineries have different configurations, and the simplest type of a refinery is a straight run, or topping, refinery, as described by McKinsey (2020b). A straight run refinery refers to a refinery that includes one distillation unit, only utilizing the distillation streams from crude oil, thus, producing only the first derivable products from crude oil. More advanced refinery types include, for instance, refineries with Cracking configuration (see McKinsey, 2020c), which further process the product streams in order to obtain more valuable products. An illustrative example of a straight run refinery is presented in figure 1.

Refining process operations are planned by optimizing the performance of the refinery. That is, the profit is maximized, and the operating costs are minimized. The refinery consists of different process units that perform different chemical operations on the feedstock, mostly crude oil, to transform crude oil into final products. The optimization problem is a linear programming problem, a mathematical model of the refinery. Constraints in the model being physical constraints related to the refinery and variables describing the possible feedstocks and sales of end products (see Parkash, 2003). Therefore, it is beneficial for the market players to understand

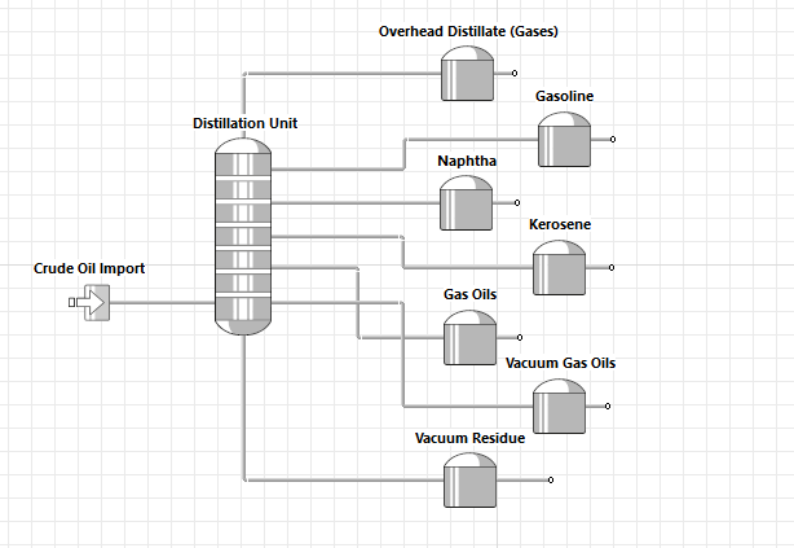


Figure 1: A schematic picture of a straight run refinery. Screen capture from Spiral Suite by Aveva (see AVEVA Group plc).

the price development of the oil market to plan these operations.

The market conditions for oil refineries are in this thesis assumed to follow effective markets. The effectiveness of the market is enhanced by assuming that all market players are using some optimization tool to optimize the operations and minimize costs of refining. Market pricing is also assumed to be dictated by a marginal refinery, or a marginal configuration (see McKinsey, 2020a). Therefore, we assume it is mostly oil refineries buying crude oil. The marginal refinery is a theoretical refinery operating in a break-even environment, therefore, setting the prices for feedstocks. That is, the refinery's purchase prices equal the value possible to be refined from the products. This marginal configuration is in this thesis defined as a straight run refinery. However, oil refineries do not operate in a break-even environment and are charging some premium on the products instead.

2.2 Crude Oil Price

The literature for predicting the real price of crude oil is extensive. The methods include a range of methods from more traditional time-series modeling to artificial neural networks. Lee and Huh (2017) compares the predictions of neural networks and ordinary least squares methods to a Bayesian model, with factors indicating supply and demand of crude oil as determinants. Ye et al. (2005), on the other hand, uses inventory data, which is easily available, to make short term predictions of the WTI spot market price. However, in our case, the interest is in the relationship of crude oil prices and the prices of refined products, using these to model the price of crude oil.

Verleger (1982) states that crude oils are valued on the market based on the value of refined products. The relationship is hypothesized as linear. The value of crude oil is a weighted average of the principal products refined from crude oil, deducted

with transportation costs and processing costs. This follows from the netback pricing formula in which transportation costs and refining margin is deducted from gross product worth. Couple of different regression models are used to prove that the official price for a crude oil set by OPEC countries (Organization of the Petroleum Exporting Countries) is determined by prevailing product prices. This hypothesis is later referred to as the Verleger Hypothesis (see Baumeister et al., 2018).

Asche et al. (2003) examines the relationship between crude oil and the prices of refined products. It is concluded that the price of Brent Crude is weakly exogenous in the price relationships that exist. Existing price relationships were found for the prices of gas oil, kerosene and naphtha, with respect to Brent crude. The exogenity indicates that the price of crude is driving the price of refined products in the long-run, not the other way around. However, the study also found that the price of refined products could influence the price of crude oil in the short run. Interestingly, heavy fuel oil does not influence the price of other refined products according to the study.

Liu and Ma (2014) proves that there exists a strong correlation between the price of crude oil and the refined products, concluding that past dynamics of the asset prices are useful in predicting future prices. The linearity of the relationships are questioned and should be noticed when modeling prices with linear relationships.

Baumeister et al. (2018) uses several prediction models, built on the prices of refined products and their market futures to explain the evolution in the market price of the benchmark crude oil WTI. Significant results are achieved for forecast horizons up to 24 months. Varying market conditions are taken into account in some of the models with time-varying parameters. It is noteworthy that the target was to predict the real price of oil with market futures and not by using some given predictions.

Bacon and Tordo (2004) constructs a regression model to determine the weight of different quality based factors in the market pricing as a difference to Brent crude. This is achieved by using pricing data from multiple crude oils and carrying out a panel data analysis of quality-based properties. The price discount with respect to Brent is determined based on quality differences, such as sulfur content or API gravity. Compared to the analysis done in this thesis, the achieved model is a static model, assuming constant price discounts for any point in time. However, the study indicates evidence for a static relationship between refined product prices and crude oil as well.

3 Methodology

Programming language R is used for statistical analysis and modeling (see R Core Team, 2020). Base functions or ggplot2 (see Wickham, 2016) are used for figures. Data preprocessing is done in Microsoft Excel (see Microsoft Corporation). Pricing data of different crude oils and product prices, as well as crude oil properties, are gathered and aggregated into sufficient data frames. In this section, the data used in this thesis is described, and the different methods, aiming to construct sufficient predictive functions, are presented.

3.1 Data

The data used for modeling include historical pricing data (all prices and quotes are differences to a benchmark crude oil) of a handful of crude oils and the product spreads. The time scope of the historical data is approximately five years of monthly data. Thus, the historical data includes around 60 data points (the time scope for different crude oils vary). Furthermore, the distillation yields and properties of the crude oils are used. The product streams of these yields are presented in table 1. In this thesis, results and analysis for three different crude oils are presented; Crude 1, Crude 2 and Crude 3.

The product streams of the different crude oils in the data are presented in table 1. These are then assigned to specific market quotes as the driving valuation factor.

Propane
Butane
Fuel Gas
Naphtha
Kerosene
Gasoil
Atmospheric Residue
Light Vacuum Gas Oil
Heavy Vacuum Gas Oil
Vacuum Residue

Table 1: All product streams, refined from crude oil, used.

Since product spreads are describing the price difference between the output stream and a benchmark crude, straight run value for the crude oil is then obtained by multiplying the yield of a product with the product spread. These variables are here on referred to as a randomly assigned number.

The data points from the historical data are divided into a training set and a test set. This provides the possibility to make an out-of-sample prediction and computing a prediction error on the prediction. The splitting ratio will be approximately 80/20 for the training set and test set, respectively.

3.2 Feature Engineering

The data provides several possible explanatory variables, but the length of the data is restricted. Sufficient methods for determining variable importance and their possible predictive power is one approach to restrict the variables used.

Variable subsetting is an intuitive, out-of-the-box method when aiming to pick features for regression models. The relatively small size of the data makes this approach sufficient. Akaike Information Criteria is used to determine the relative quality of the model, with respect to the number of variables. AIC is trading between

a simple model and a better performing model to find the best performing model. AIC-value is given by

$$AIC = -2l + 2p \tag{1}$$

where l is the log-likelihood of the model, measuring the goodness of the fit, and p number of estimated parameters in the model, as described by Hastie et al. (2009). A clear benefit of this measure is that it emphasizes a small number of variables with the relatively small data set.

Possible predictors to regression models can be chosen by subsetting and stepwise computing AIC values and choosing the predictors which seem to give the best results in terms of AIC-values. This can be carried out by several functions in R, but those available in the MASS package ((Venables and Ripley, 2002)) are used. The function uses a combined forward and backward selection, the downside being the computational expensiveness. However, since the size of the data is limited, it is sufficient enough. The benefits of carrying out feature selection are the possibility of examining what features seem to explain the price differences in different crude oils the best. This is beneficial information if developing multi-crude models.

3.3 Models & Evaluation

Models used to examine the objective are introduced here and referred to later in section 4. Three different models, plus one benchmark model, are used to address if product prices include predictive power for the crude oil price. The metrics used to evaluate the results are presented, as well.

3.3.1 A Non-Change Forecast

When modeling and predicting crude oil price development, it is useful to understand how much better the used models are performing compared to a non-changing forecast. The non-change forecast is predicting the next value of the target variable as a non-changing value. This approach is presented, for instance, in Baumeister et al. (2018). More precisely, the non-change value here is computed as a weighted average of past values

$$P^o_{t+k} = \bar{P}^o_t = \frac{1}{n}\sum_{t=1}^{n} \alpha_t P^o_t \tag{2}$$

where $\bar{P}^o_t$ is the weighted average of the price difference between a crude oil o and a benchmark crude oil at time t, and k is the point in time to be predicted. The weights α_t are constructed to be linear weights, emphasizing more recent history, and n is the size of data. The weighted average is computed for the training data, and the obtained value is used to predict prices in the test data.

This type of model is used as a benchmark in the predictions and is used to conclude how well a model is performing. By comparing the prediction error given by a non-changing model to the prediction errors of the models studied, we find out if the studied models presented below have any predictive power.

3.3.2 Regression Model

In the hypothesis, laid out by Verleger (1982), the value of a crude oil o, for a refiner is expressed as

$$V_t^o = \sum_{i=1}^{x} w_i P_{i,t} \tag{3}$$

where w is a technological constraint, i.e., how much of a product k with the price P_k a refiner is able to refine from the specific crude oil. The integer x refers to the number of products taken into consideration. Since the market contains multiple price takers, the constraint w is simplified to be the theoretical yield from the product k, which refers to the straight run marginal refinery presented in section 2. As mentioned before, this reasoning is also derived from the netback pricing formula of crude oil but does not include any freight costs, which is assumed to be a constant when training models. All this motivate the regression model

$$\hat{P}_t^o = \hat{\beta}_0 + (\sum_{i=1}^{x} \hat{\beta}_i w_i P_{i,t}) + \epsilon_t \tag{4}$$

where parameters $\hat{\beta}$ are estimated with ordinary least squares method (OLS) to $\hat{\beta}$. When using linear regression models, it is important to validate assumptions for using OLS. The assumptions, which include identically and independently distributed errors, and a minimum multicollinearity in the variables, are validated with different diagnostic plots, describing residuals' behavior. Autocorrelation plots are also used. Also, variance inflation factor (VIF) is used to determine if there exists multicollinearity in the variables. Estimation of parameters under maximum likelihood is evaluated by validating normality of residuals. This is especially required if confidence intervals assume this.

The regression model can be written to weigh recent data points more heavily. In that case, the following sum of squares is minimized to estimate parameters $\hat{\beta}$

$$\min. \sum_{t=1}^{n} \alpha_t [y_t - f(t, \beta)]^2 \tag{5}$$

where α_t is a weighting factor, and function f is the linear function trained in OLS. The weights are constructed to be linear, emphasizing more recent history. This is used because changing market conditions are most likely responsible for changing the pricing functions of crude oils. This effect could possibly be dampened by assigning weights to the data points in this way. Both models with weights and without weights are used.

3.3.3 Lagged Regression Model

To capture possible price adjustment phases in market pricing, a regression model that uses lagged variables will be constructed. A lagged model would include past values

of product prices, or crude price itself, to explain the current price. A distributed lag model is formulated as

$$\hat{P}_t^o = \hat{\beta}_0 + \sum_{i=1}^{x} \hat{\beta_{i1}} w_i P_{i,t-1} + \sum_{i=1}^{x} \hat{\beta_{i2}} w_i P_{i,t-2} + ... + \epsilon_t \tag{6}$$

where different lags of the product prices are trained with different parameters $\hat{\beta_{ih}}$, where h is the lag (see Baltagi, 2008). By investigating the lags, a perception of how quickly price adjustments are seen in the crude oil prices could be formed. However, the dynamics of the system could adjust the prices in the other direction, i.e. changes in crude oil prices could affect the prices of end products. Weights in equation 5 are included in the lagged model as well in order to emphasize recent history.

The difficulty here is to find the best combination of lagged values, since including all variables in equation 6 will include too many variables. This will be done by constructing a table with lagged variables and construct the model from there, minimizing AIC, as described in equation 1.

3.3.4 ARIMAX Model

It is reasonable to assume that there could be time-dependencies in the price differences themselves. More precisely, previous values of the time series could affect the current value. In order to take into account these dependencies, different types ARIMA models are applied. ARIMA models treat the time series of the variable y_t as a linear model by regressing previous values of the time series and the errors to explain the current value at time t. An ARIMA(p,d,q) is formulated as

$$\nabla^d y_t = \sum_{i=1}^{p} \phi_j y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} \tag{7}$$

where parameters ϕ_i and θ_i, are estimated for all i. The differencing operator ∇^d is used to stationarize the time series (if needed). The model can be further expanded to an ARIMAX model, which incorporates exogenous terms to the model, in our case, the straight run values of products. Hence, an ARIMAX model functions as a combined linear regression model and an ARIMA (equation 7) model. ARIMAX is formulated as

$$\nabla^d y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \sum_{i=1}^{x} \beta_i w_i P_{i,t} \tag{8}$$

where parameters ϕ_i, θ_i and β_i are estimated with autoregressive forecasting for all i. If confidence intervals are used, the normality of residuals need to be validated. These models have well-established computational tools in R, both as base functions and third-party libraries such as Hyndman et al. (2020), and are utilized. The models may be further extended with seasonal components, if needed, and are studied extensively in Box et al. (2015).

3.3.5 Evaluation Metrics

Two metrics for prediction error are used to evaluate the results of the computed forecasts. The graphical representation of the forecast is important since the price behavior of the crude oils is different. Some crude oil prices are more volatile than others, which yield larger absolute errors than those for more stable crude oils. Root mean square prediction error (RMSE) is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(P_t - \hat{P}_t)^2} \tag{9}$$

where P_i is the true price and $\hat{P}_i$ is the estimated price at time i (see Barnston, 1992). This metric is not directly comparable when analyzing different crude oils, which is why it is beneficial to evaluate the prediction error based on deviation in percentages. The mean absolute percentage error (see De Myttenaere et al., 2016) is defined as

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{P_t - \hat{P}_t}{P_t}\right| \tag{10}$$

These two error metrics are then used when evaluating the predictions made by the different models.

4 Results

Multiple different crude oils were analyzed and the prediction results for three of them are presented in this chapter. The results are also visualized together in order to show the price evaluation of the crude oils with respect to each other. The models presented in chapter 3.3 were used and compared.

4.1 Diagnostics

The models were trained and validated in R. The process of the analysis was first to build a linear regression model with all possible variables and then carry out feature engineering on the model to subset the most significant variables. Feature engineering was done as described in chapter 3.2. These variables were then assumed to be most significant when building other models as well.

In order to use linear models, the assumptions for linear models should be tested and validated. A correlation plot is used to visualize the correlation among the different features and crude oil and should exploit possible linear relationships. Correlation plot for Crude 1 is visualized in figure 2.

The correlation plot indicates multicollinearity among the variables. This should be addressed when building linear models. When constructing a linear model of the product spreads, it is assumed that it would be sufficient only to use one of the correlated variables to capture both variables. The aim is not to have any heavily correlated variables, in line with the assumptions of a linear model. Multicollinearity can also be evaluated using VIF factors. VIF factors yielded similar interpretations.

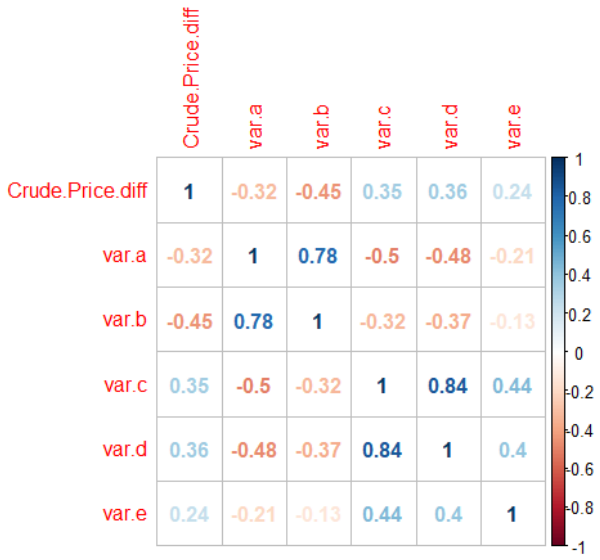


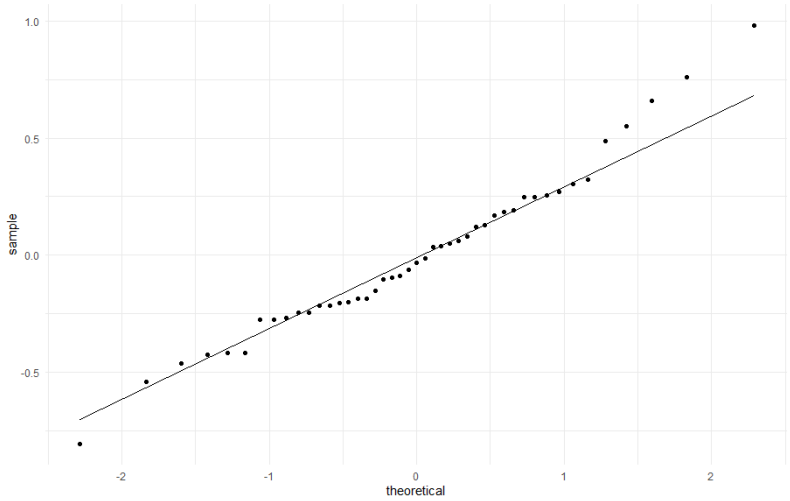
Figure 2: Correlation plot of the correlations between the price differences of Crude 1 and respective straight run products. Pearson correlation coefficient is used.

Naturally, the same phenomena were found for all other crude oils because the same product prices apply for all crude oils in this study, even though the crude properties bring a different scaling factor for each crude oil product.

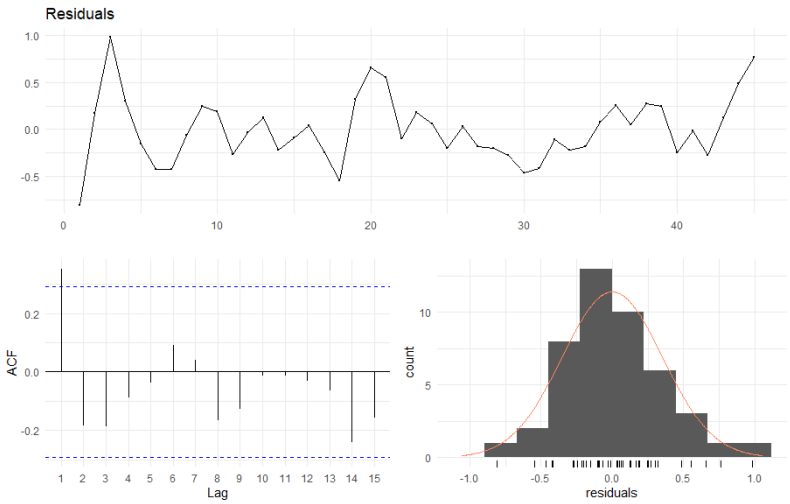
When fitting regular linear regression models, it is assumed that the residuals of the model are uncorrelated and normally distributed around zero. Figure 3 shows a subset of diagnostics plots that aim to address the assumptions. QQ-plot is plotting the quantiles from a normal distribution and the quantiles from residuals. Hence, a straight line would be a clear sign of normal distribution. In this case, the residuals fluctuate around the straight line, and the outermost residuals deviate from the straight line. Considering market outliers and other, possible exogenous factors, the distribution of residuals indicate a right-skewed distribution. This means caution if possible confidence intervals are used as well as indications that parameter estimations are not under maximum likelihood. Autocorrelation plot and time series of the residuals point to uncorrelated residuals.

The diagnostic plots for other crude oils yielded similar outcomes. Even though the details of the distributions of residuals varied slightly between crude oils, no significant correlations among residuals were found. A slightly skewed distribution of residuals would indicate that some more robust models could be tried out. Overall, the assumptions for a linear model seemed to be fulfilled to motivate the usage of linear models.

Out of the time series point of view, small trends were visible in some of the crudes



(a) QQ-plot for Crude 1 comparing the distribution of residuals to normal distribution.



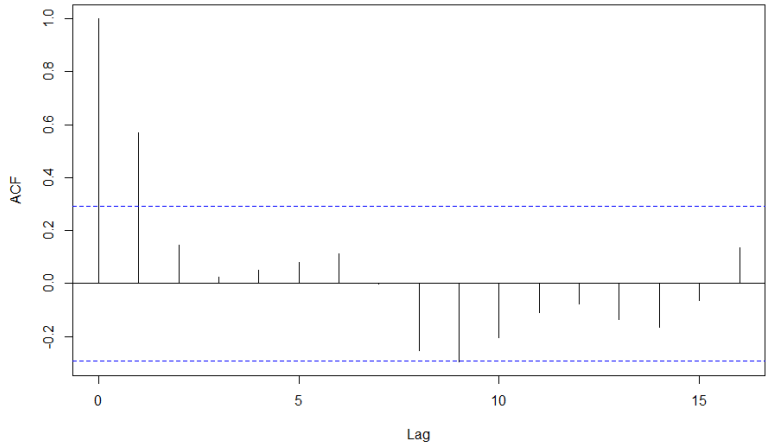
(b) Magnitudes, ACF and distribution of residuals.

Figure 3: A subset of diagnostic plots for Crude 1. The underlying model is a basic, feature engineered, regression model described in equation 4.

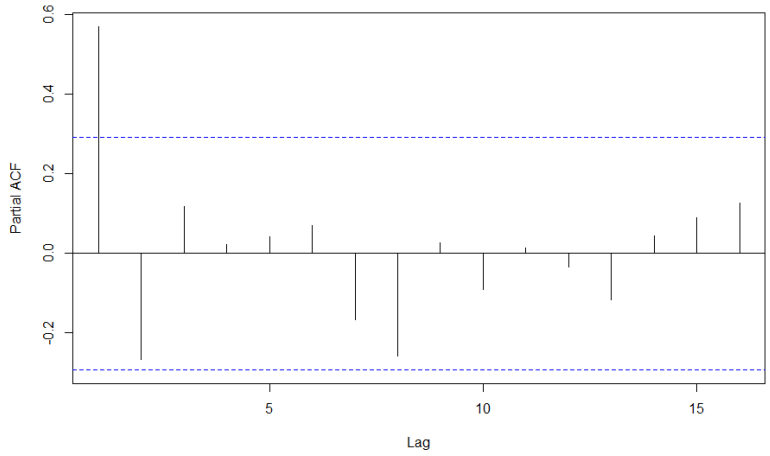
and were tried to be addressed by differencing the time series in ARIMAX modeling. When building ARIMAX models, autocorrelation plots and partial autocorrelation plots are examined to determine if there is a correlation between lagged time series values of the crude oil difference. In most cases, no significant correlations were found. Autocorrelation and partial autocorrelation plots for Crude 2 are found in figure 4.

4.2 Performance of Models

The prediction errors for all presented crude oils are found in tables A1, A2 and A3, in the appendix. The prediction errors are described by the metrics presented in



(a) ACF of Crude 2.



(b) PACF of Crude 2.

Figure 4: ACF and PACF plots for Crude 2 to examine possbile correlation among lagged price values. The blue dotted line describe the significance level of 95%.

chapter 3.3.5.

Noteworthy is that models using only crude oil straight run products as explanatory variables perform better than the benchmark models for all crude oils. This even though the fact that most crude oils do not have any clear trend for the span of the whole data set, making the short term trends partly explainable with the product prices. However, no conclusions about causalities are drawn.

A visualized plot of the prediction results is found in figure 5. The plot is divided by a vertical dashed line indicating the split between a training set and a test set. A dashed line always indicates the model result. The plot is, therefore, showing one part of the model fit as well as the out-of-sample prediction done for approximately one year ahead. Notably, there seem to be some lagged dependencies, but any clear

patterns are not detectable. The same conclusions can be drawn for most of the crude oils researched.

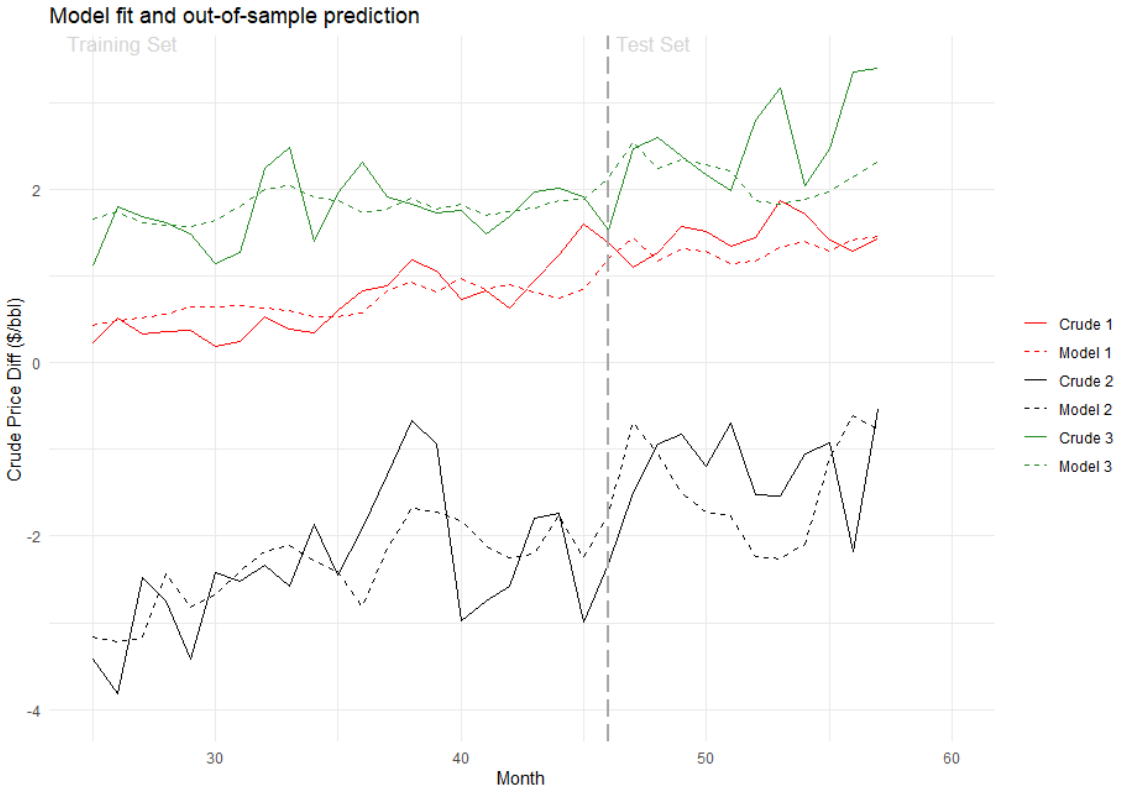


Figure 5: Crude oil price differences, plotted with the best performing model for each crude. The dashed line indicates the split between train and test set.

Bad performance of lagged regression models, presented in equation 6 is also recognizable. A common observation was an indication of overfitting to the training set, possibly caused by a large number of variables. Variable selection was most of the times yielding in models, which included more variables than the other regression models. This was interesting since a possible subset of variables also included the ones in the best performing model of regression models, demonstrating flaws in the method of variable selection. Overfitting could also be addressed with appropriate robustness to models, such as ridge regression.

There were only small differences in the performance of simple ARIMAX models and the basic regression models without weighting. This is probably because of the small weight of lagged values of the time series itself in the model.

5 Conclusions

As figure 5 shows, the product prices seem to, at least partly, drive the difference of the crude oil to the benchmark crude oil considered, indicating confirmation of the hypothesis given. Alternative approaches to predicting the development of the

price difference would be to assume the value of a specific crude oil to be constant, with respect to the benchmark crude oil. Reasonably, this constant could be some average of historical values, like the one used as benchmark model in this analysis. Thus, we have demonstrated that predicting the evolution of these price differences with the demand for refined products could yield more accurate estimations than this type of prediction.

A longer time frame of data would make it possible to consider a larger amount of variables, but then again, the market conditions change a lot in an extended time period. These market dynamics and the market outliers remain a challenge and would be beneficial to explore further. Market outliers indicate some exogenous shock, which then should be weighted down when training a model. These chocks most likely reflect something crude specific since an extensive chock to the oil market itself also affect the benchmark crude oil. However, demand-side chocks could affect the profitability of a specific crude oil resulting in changing valuation against the benchmark crude. Also, using interchangeable crude oils as predictors could be examined but would most likely be a very correlated predictor.

The underlying assumption that a marginal, straight run refinery determines the price of crude oil is quite a rough estimation and is knowingly not the case in the refinery business environment. One possible solution would be to simulate prices, using a different kind of marginal refinery as a basis of the market structure, such as a more complex refinery, including a hydrocracker. The assumption of the marginal refinery is in itself a rough estimate and assumes effective markets, but is generally used as an assumption for the pricing mechanisms. The yields from one barrel of crude oil in a more complex refinery are mathematically more complex and could require some nonlinear methods in the models.

This study was focusing on dealing with the researched crude oils separately. This means that the pricing dependencies among crude oils are not taken fully into account. Differences among crude oil properties are also only taken into account indirectly in the weights since the models are trained separately. One promising direction of development would be to combine the datasets, develop and evaluate a more unifying crude pricing model. A more unifying model would make it possible to better understand how expected changes in the product prices affect pricing of different types of crude oils. This kind of model would also make it possible to do pricing estimations for arbitrary chosen crude oils without the need for historical pricing data of the crude oil. However, evolving market conditions change the demand for different products and, therefore, the premiums expected from the refineries on the products refined. Maintaining such a model would be hard, and a suitable way to emphasize recent history more heavily would most likely be required.

To conclude, a sufficient background was given in order to understand how crude oils are valued on the market, and a hypothesis for predicting crude oil prices based on the product prices and the properties of crude oil was motivated. Models were described and technically compared in order to demonstrate that product prices indeed have predicting power in crude oil prices and perform better than the benchmark models used. Further development needs and flaws of the methodologies used were presented.

References

Frank Asche, Ole Gjølberg, and Teresa Völker. Price relationships in the petroleum market: an analysis of crude oil and refined product prices. *Energy Economics*, 25 (3):289–301, 2003.

AVEVA Group plc. Spiral Suite. URL `https://sw.aveva.com/plan-and-schedule`.

Robert Bacon and Silvana Tordo. Crude oil prices: Predicting price differentials based on quality. The World Bank Group. *Public policy for the private sector*, (275), 2004.

Badi H. Baltagi. Distributed lags and dynamic models. *Econometrics*, 2008.

Anthony G Barnston. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, 7(4):699–709, 1992.

Christiane Baumeister, Lutz Kilian, and Xiaoqing Zhou. Are product spreads useful for forecasting oil prices? An empirical evaluation of the Verleger hypothesis. *Macroeconomic Dynamics*, 22(3):562–580, 2018.

George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

Arnaud De Myttenaere, Boris Golden, Bénédicte Le Grand, and Fabrice Rossi. Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48, 2016.

Mohamed A. Fahim, Taher A. Alsahhaf, and Amal Elkilani. *Fundamentals of Petroleum Refining*. Elsevier, 2010. ISBN 978-0-444-52785-1.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846.

Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2020. URL `http://pkg.robjhyndman.com/forecast`. R package version 8.12.

Andrew C Inkpen and Michael H Moffett. *The Global Oil & Gas Industry: Management, Strategy & Finance*. PennWell Books, 2011.

Chul-Yong Lee and Sung-Yoon Huh. Forecasting long-term crude oil prices using a bayesian model with informative priors. *Sustainability*, 9(2):190, 2017.

Li Liu and Guofeng Ma. Cross-correlation between crude oil and refined product prices. *Physica A: Statistical Mechanics and its Applications*, 413:284–293, 2014.

Energy Insights By McKinsey. *Marginal Configuration*, 2020a. URL https://www.mckinseyenergyinsights.com/resources/refinery-reference-desk/marginal-configuration/.

Energy Insights By McKinsey. *Straight Run*, 2020b. URL https://www.mckinseyenergyinsights.com/resources/refinery-reference-desk/straight-run/.

Energy Insights By McKinsey. *Cracking Configuration*, 2020c. URL https://www.mckinseyenergyinsights.com/resources/refinery-reference-desk/cracking-configuration/.

Energy Insights By McKinsey. *Urals Crude*, 2020d. URL https://www.mckinseyenergyinsights.com/resources/refinery-reference-desk/urals-crude/.

Microsoft Corporation. Microsoft Excel. URL https://office.microsoft.com/excel.

Engineering National Academies of Sciences, Medicine, et al. *Spills of diluted bitumen from pipelines: A comparative study of environmental fate, effects, and response*. National Academies Press, 2016.

Surinder Parkash. *Refining Processes Handbook*. Elsevier, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Energy Charter Secretariat. Putting a Price on Energy. Oil Pricing Update, 2011.

Bill Venables and Brian Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Philip K Verleger. The determinants of official OPEC crude prices. *The Review of Economics and Statistics*, 1982.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Michael Ye, John Zyren, and Joanne Shore. A monthly crude oil spot price forecasting model using relative inventories. *International Journal of Forecasting*, 21(3):491–501, 2005.

A Prediction Errors

Following tables present the performance of the models in the metrics described in chapter 3.3.5.

Crude 1	RMSE ($/bbl)	MAPE (%)
Benchmark	0.782	51.55
Regression Model	0.411	15.44
Lag Model	0.912	59.93
ARIMAX Model	0.260	18.02

Table A1: Prediction errors for Crude 1.

Crude 2	RMSE ($/bbl)	MAPE (%)
Benchmark	1.367	138.38
Regression Model	0.792	58.15
Lag Model	1.146	98.27
ARIMAX Model	0.946	89.24

Table A2: Prediction errors for Crude 2.

Crude 3	RMSE ($/bbl)	MAPE (%)
Benchmark	0.956	29.78
Regression Model	0.714	20.46
Lag Model	1.146	98.27
ARIMAX Model	0.946	89.24

Table A3: Prediction errors for Crude 3.