Charles University

Faculty of Social Sciences
Institute of Economic Studies

BACHELORS'S THESIS

Forecasting oil prices volatility with Google searches

Author: **Ekaterina Tolstoguzova**

Supervisor: **doc. PhDr. Ladislav Krištoufek, Ph.D.**

Academic Year: **2018/2019**

Declaration of Authorship

I hereby proclaim that I wrote my bachelor thesis on my own under the leadership of my supervisor and that the references include all resources and literature I have used.

I grant a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 31, 2019

Signature

Acknowledgments

I am grateful especially to my thesis supervisor doc. PhDr. Ladislav Krištoufek Ph.D. for his valuable comments and suggestions. I would also like to thank my family and friends, who supported me through my entire studies.

Abstract

Oil market pricing is highly susceptible to geopolitical and economic events. With the rapid development of information technology, energy market can quickly get external information shocks through the Internet. This thesis examines the relationship between prices of three oil benchmarks, CBOE Crude Oil Volatility Index, and Google search queries. We built VAR model to study Granger causality and to provide impulse response analysis. Results indicate both one side and two-side causal relationship between oil-related series and most of the search queries. Out-of sample forecasting with measures of predictive accuracy and Diebold-Mariano test demonstrated that Google trends can improve short-run prediction potential only for models with WTI price and volatility index.

JEL Classification	E37, G17, Q43, Q47
Keywords	Google Trends, oil, VAR, search query, price volatility, nowcasting

Author's e-mail	katerina.tev@gmail.com
Supervisor's e-mail	ladislav.kristoufek@fsv.cuni.cz

Abstrakt

Kombinácia zrýchlujucej sa dynamiky obchodovania na trhu s ropou a rapídneho rozvoja technológií umožňuje jednoduchý prenos externých informačných šokov cez internet. V tejto práci skúmame vzájomné vzťahy medzi troma referenčnými cenami ropy, CBOE Cruide oil indexom volatility a Google vyhľadávaniami. Za účelom testovania Grangerovej kauzality a uskutučnenia impulse-response analýzy sme vytvorili VAR model. Výsledky ukazujú jednostranné aj obojstranný pričinný vzťah medzi ropnými cenami, OVX a Google vyhľadávaniami. Out-of sample predpovede a Diebold-Marianov test nám ukázali je možné využiť Google trends na zlepšenie krátkodobej predikciu v prípade modelu s WTI cenami a indexom volatility.

Klasifikace JEL	E37, G17, Q43, Q47
Klíčové slova	Google Trends, ropa, VAR, vyhledávací dotaz, volatilita, nowcasting

E-mail autora	katerina.tev@gmail.com
E-mail vedoucího práce	ladislav.kristoufek@fsv.cuni.cz

Contents

List of Tables

List of Figures

Acronyms

ADF Augmented Dickey-Fuller test

AIC Akaike Information Criterion

BIC Schwartz Bayesian Information Criterion

CBOE Chicago Board Options Exchange

GSVI Google Search Volume Index

GT Google Trends

HAC Heteroskedasticity and Autocorrelation Consistent

KPSS Kwiatkowski-Phillips-Schmidt-Shin test

MAE Mean Absolute Error

OECD Organisation for Economic Co-operation and Development

OLS Ordinary Least Squares

OVX Oil Volatility Index

RMSE Root Mean Squared Error

U.S. United States

VAR Vector Autoregression

WTI West Texas Intermediate

Thesis Proposal

Author	Ekaterina Tolstoguzova
Supervisor	doc. PhDr. Ladislav Krištoufek, Ph.D.
Proposed topic	Forecasting oil prices volatility with Google searches

Research question and motivation

Oil and its trade underlie international relations, so they are often a reflection of what is happening in the world. Oil markets are very dynamic. Often, changes in their prices depend on many geopolitical factors.

In recent decades, new methods of estimating and forecasting macroeconomic indicators have been developed. In 2009, H. Choi and H.Varian put forward the hypothesis that the query statistics in Google should correlate with the current level of business activity, and it may also be useful for a short-term forecast. The idea is that a set of characteristic keywords is revealed, and then a graph is constructed based on the quantity of search queries. We immediately see how the public's interest in this sector is changing, and we can make assumptions about how the demand for corresponding shares will change in this connection.

Hypotheses

1. Google searches data proves to be useful in short term forecasting of consumer behaviour.

2. The search query categories can be successfully utilized to nowcast oil prices volatility.

Methodology

I will study the partial effect of searching for individual words using standard OLS time series method and vector autoregression. Also I will test how these effects will be changed over time trough moving estimation window. For estimating prices volatility I will be using Garman-Klass estimator and CBOE Crude Oil Volatility Index.

Contribution

In recent years, oil has acquired the status of a "world currency", because the stability of the economy largely depends on it. The price of oil has become an important indicator of the state of the world economy. The main purpose of this work is to help to understand the market, whether it is possible to predict, and pre-warn changes in the oil market using Google searches. Relying on it, governments could competently adapt their international policy.

Outline

1. Introduction
2. Literature Review & Theoretical Background
3. Data
4. Methodology
5. Results
6. Conclusion

Core bibliography

1 BOSLER, Fabian T. MODELS FOR OIL PRICE PREDICTION AND FORECASTING. 2010. Master of Science in Applied Mathematics. SAN DIEGO STATE UNIVERSITY.

2 CHOI, HYUNYOUNG a HAL VARIAN. Predicting the Present with Google Trends. THE ECONOMIC RECORD. 2012, 88(SPECIAL ISSUE), 2-9. DOI: 10.1111/j.1475-4932.2012.00809.x.

3 PREIS, Tobias, Helen SUSANNAH MOAT a H. Eugene STANLEY. Quantifying Trading Behavior in Financial Markets Using Google Trends. SCIENTIFIC REPORTS. 2013, (3), 1-6. DOI: 10.1038/srep01684.

4 PAVLICEK, Jaroslav a Ladislav KRISTOUFEK. Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries. PLOS ONE. 10(5). DOI: 10.1371/journal.pone.0127084.

5 CURME, Chester, Tobias PREIS, H. Eugene STANLEY a Helen Susannah MOAT. Quantifying the semantics of search behavior before stock market moves. PNAS. 2014, 111(32), 11600–11605 —. DOI: 10.1073/pnas.1324054111.

6 LATOEIRO, Pedro, Sofia B. RAMOS a Helena VEIGA. Predictability of stock market activity using Google search queries. 2013. Universidad Carlos III de Madrid.

Author

Supervisor

Chapter 1

Introduction

Nowadays oil is the most important mineral for humans. Life of the modern society is inconceivable without this valuable raw material. Oil plays one of the main role in maintenance of the global balance between fuel and energy. This raw material is used not only in production of fuel for vehicles, but it is also widely applied in energy and chemical industries. For example, IEA asserts that road transport is dependent on oil by 92 %. Oil accounts for 36% of the energy that is absorbed in the world and produces 9% of the electricity of the entire planet, (A.N. Brodunov 2015).

Oil has become the main type of energy raw material, as a result its economic and political importance in the world has increased. Presence of our own oil resources, possibility of organizing export of oil and oil products allow various countries to achieve significant success in economic and social development. In recent years, oil has acquired the status of "world currency", since the stability of the economy largely depends on it. The price of oil has become an important indicator of the state of the global economy. For a long time, factors, which affect past, present and future oil price levels and their fluctuations, are the main subjects of analysis for many scientists, politicians, energy experts and economists. Oil price's forecasts can now be found not only in the scientific literature, but also in the government reports, various political discussions and analytical publications of banks.

The oil market is very multifaceted and contains a complex operating structure. The pricing mechanism covers the entire planet and is of great interest to scientific researchers. One of the most significant is the study by Kilian (2009). He studied demand shocks, supply shocks and extraneous factors that influence oil demand applying the SVAR model. The results demonstrated that all three

variables have a significant impact on prices in the energy market and economic activity. Ratti & Vespignani (2013) extend the previous study by including an additional factor in the model, such as global real money stocks. A close relationship was found between global real money reserves and oil prices, as well as the fact that the greatest impact was during the period of price increases from 2009 to 2011. Kilian & Murphy (2014) supplemented the primary study. They showed that future supply shock has a significant effect on the price of oil. Also, scientists from Japan in their study Takuji Fueki & Tamanyu (2018) proved that the shocks of future demand and future supply can explain about 30-35 percent of the volatility of oil prices. Contrariwise, Marco Lorusso (2018) in their research paper studied how changes in the oil market can affect the macroeconomic indicators of Great Britain. They concluded that oil shocks play an important role in changing the UK's main macroeconomic indicators such as GDP growth, unemployment, inflation and nominal interest rates.

The whole interest of researchers is related to the fact that the prices of oil and petroleum products concern both producers and consumers. Changes in the dynamics of oil prices affect the level of costs in all sectors of the manufacturing industry without exception. Furthermore, fluctuations in world oil prices and the oil market situation lead to serious changes in the economic policies of both oil-producing countries and countries, in which industry is based on imported oil. Higher energy costs increase the cost of producing and transporting everything. With the growth of cost, which slows down production, the incomes of enterprises fall, and the stock market is experiencing a decline. For consumers, the rise in gasoline prices scares those who, seeing the loss of their purchasing power, reduce their spending on non-essential goods, which adversely affects the sales of companies. It also has a negative effect on economic growth and share prices. The fall in oil prices helps consumers reduce the cost of living and save money that can be spent on more expensive purchases. In most cases, this implies a reduction in transportation costs, which leads to a lower cost of living and lower inflation. When it comes to the impact of falling oil prices on the economy, this usually means good news for oil importers, such as Europe, China, India, Japan. For oil exporters, such as OPEC or Russia, a fall in oil prices has the opposite effect - it reduces the cost of their exports and leads to a decrease in the trade surplus. To summarize we can conclude that changes in oil prices directly or indirectly affect everyone. Therefore, more and more people are interested in oil prices, their ups and downs, and consequently, they are looking for additional information about crude oil with a constantly

increasing interest.

Information is one of the most important components of the market. Market pricing is no longer bound only by the law of supply and demand, but is also strictly determined by information, its content, volume and transfer. Nowadays, it has embraced a completely new meaning thanks to the Internet. It accelerates the receipt of external information about important events all over the world. Thanks to the Internet and popular search engines, information easily and instantly reaches players in the energy and oil market and thus influences their decisions. The popularization of the Internet has led to the formation of a new direction of analytics called Google Econometrics. Google is no longer perceived only as a search engine. Google tools, such as Google Trends or Google Correlate, provides publicly available search query data that you can use to display people's intentions and interests in real time. This type of analysis is often referred to as "nowcasting", Giannone Domenico (2008), because it tries to explain the current situation rather than predict future activities. Choi & Varian (2009a) was the first study, in which search queries were used to improve the prediction of economic indicators. They have assumed that when any political or important economic events occur, people try to understand what this may mean for the market and begin to look for information that could suggest the right decision. After that, many researchers began to use Google data to forecast and nowcast social and economic indicators.

Due to the fact that crude oil is one of the most important commodities in the global markets, it can be assumed that building models that show the relationship between oil prices and Web data can provide valuable information for predicting price changes in other markets. This study finds out whether public opinion, reflected in Internet search queries, influences decisions made by participants in the energy market. The relationship between Google Trends and crude oil is examined by applying vector autoregressive model. And the predictable ability of search queries is studying through rolling fixed window method.

The rest of the paper is structured as follows: Chapter 2 provides a brief review of the literature; Chapter 3 describes the data and their selection process; Chapter 4 defines the used methodological framework. The empirical results and discussion are presented in Chapter 5, and finally, Chapter 6 concludes this thesis.

Chapter 2

Literature Review

2.1 Internet searches in economics and finance

Recently, the number of papers, in which it is investigated whether data from Internet search engines, such as Google, can improve nowcasts or short-term forecasts of economic and financial variables, is growing rapidly. A rather large part of "Google econometrics" literature is devoted to establishing causal relationships and predicting labour market in different countries.

One of the earliest papers on this topic was written by Nikolaos Askitas (2009), it studies the correlation between Google Trends data and the unemployment rate in Germany. For the research, they selected keywords in German related to unemployment and job search and divided them into 4 groups. After estimating several error-correction models (ECM) and comparing the results using Bayes-Schwartz information criterion (BIC), they concluded that using data from third and fourth weeks of the previous month helps to predict the unemployment rate for the current month. Analyzing data from Israel Suhoy (2009) concluded that query indices could help to estimate the current economic downturn in the country. Also, it was found that the predictive ability of keywords is not constant over time. Influenced by these studies Choi & Varian (2009b) examined relationships between unemployment, welfare-related searches and U.S. Initial Claims for unemployment benefits. Based on AR model and on the obtained mean absolute percentage error (MAPE), they demonstrated that Google Trends data can improve predictions of initial jobless benefit claims.

Francesco D'Amuri (2010) also focused on analyzing the unemployment rate in the US, but considering only one search query "jobs". Comparing about 500 ARMA models with various data conversions, as well as several non-

linear models (SETAR, LSTAR, AAR), they concluded that Google Trends data could improve forecast accuracy by two months in advance. Similar results were obtained by Meltem Gulenay Chadwick (2012). Studying Turkish data and applying RMSE and Modified Diebold-Mariano test, they found that the models with better nowcasting capabilities always included Google data. Y. Fondeur (2013) also confirms these findings. This academic work was based on data on French unemployment for 15- to 24-year olds, and more advanced methods, such as modified version of the Kalman filter, were used in it.

In an updated version of their first two papers Choi & Varian (2012) examined not only unemployment, but also whether search queries can be used to improve the prediction of many other economic indicators, such as automobile sales, tourist arrivals and consumer confidence. They used basic methods with Google data for nowcasting. After estimating the simple seasonal AR models and constructing a rolling window forecast, they concluded that models with Google Trends data improved the prediction ability for all selected economic variables. Gary Koop (2013) were also not limited to investigating one indicator. For the analysis of nine macroeconomic variables, they chose an unusual method for this area of research - the method of dynamic model selection (DMS), which simplifies working with time-varying models. Results proved that search queries could be successfully used for the nowcasting of macroeconomic aggregates. It was found that Google data had a more significant effect on prediction, when models included probabilities of search queries, rather than using them as normal regression variables.

The financial industry also attracts researchers studying Internet data. For example, Zhi Da & Gao (2011) suggested that Search Volume Index (SVI) could be a good indicator of investor behavior. They studied the causal relationships between Google data and Russell 3000 stock tickers. The results confirmed the hypothesis and showed a high correlation in the short-term. They also provided evidence that SVI's effect is stronger for retail investors. A similar study conducted by Nikolaos Vlastakis (2012). Analyzing data of week closing stock prices and related values of the S&P 500 index and VIX index, they concluded that SVI data reinforce the significant impact of information demand on individual stock in terms of historical volatility and trading volume. It was also found that the demand for information increases if the level of risk aversion increases.

Also in 2013, researchers at Warwick Business School published an experiment, Tobias Preis (2013), in which the Google search engine was used as a

tool to predict trends in the stock market. They were able to find a correlation between the increase in the number of search queries related to various political and economic topics, and the subsequent collapse of stock markets. A specially created investment game simulator was used to identify the relationship between such requests. When the number of search queries decreases, the computer practically "buys" stocks, and with an increase in the number of requests for "crises" and similar events, it closes long positions. The most reliable for the United States was the word "debt". By tracking markets only on it, scientists increased their hypothetical securities portfolio by 326% in just seven years. When modeling a standard trading strategy that did not take into account the frequency of search queries, they managed to achieve an increase of only 16%.

Further, the cryptocurrency market was analyzed by Google econometrics researchers. Martina Matta (2015) studied the predictive ability of social and web search media on bitcoin price fluctuation. They achieved that tweets and Google Trends data has a strong positive correlation with Bitcoin's price. Also, Kristoufek (2015) analyzed the investor's interest in Bitcoin. For this aim were utilized data obtained from Google and Wikipedia for word "Bitcoin". It was found that there is a positive correlation between searches on both engines and Bitcoin's price fluctuation in the long run. Rishanki Jain (2018) examined the relationship between people's behavior on Internet and changes in Bitcoin's price. ARIMA and ARIMAX models were built for price prediction. They obtained that there exist a strong correlation between Twitter volume and cryptocurrency price. Additionally, it was concluded that Google Trends data can improve the forecasting of Bitcoin price.

Seung-Pyo Jun (2018) collected the most cited 657 scientific papers that utilized Google Trends data in last ten years. They study the impact of using Big Data from web searches on researches. They obtained that pharmaceutical field was the first one that web data to predict different epidemic. However, it turned out that in recent years, Google econometrics researchers are most actively studying economics and business. Results also demonstrated the trend of using extra media sources to get more accurate Internet data analysis and to overcome the limitation of data obtained only from search engines.

2.2 Google data and oil market

As we see, there is a lot of literature confirming that the study of human behavior in the Internet can help in predicting various economic and financial phenomena. In the last few years researchers have published some studies on the application of data from search engines for analyzing and forecasting prices in the oil market.

Jian-Feng Guo (2013) were the first who applied Google search queries to analyze the energy market. They divided the selected keywords into four groups: oil price, oil demand, the financial crisis of 2008 and the Libyan war of 2011. Brent prices were presented as an economic variable, as it is one of the most common oil benchmark in the world. To analyze this data, they used co-integration methods and a modified EGARCH model. The results show a long-term correlation between Brent prices and some search queries. They also found that in the short term there is an asymmetrical effect between the influence of positive and negative public sentiment and the volatility of oil prices.

Dean Fantazzini (2014) in their study provided a set of multivariate models for predicting the real oil price. For this they used Google data and various economic and energy aggregates. After applying various robustness tests, it was found that in the short term, models containing both macroeconomic indicators and the Google index statistically perform better than other forecasts. However, multivariate models that only include Google data are best suited for medium and long-term predictions for up to 24 steps ahead.

In Xin Li (2015), Google search volume index (GSVI) was used to study the relationship between the different trader positions that were obtained from COT reports and crude oil prices. The recursive out-of-sample forecast method was used to determine, whether the model with search queries outranks other models. The results of the study demonstrate that GSVI has an impact on non-commercial and non-reporting traders and has a positive correlation with the volatility of oil prices.

Unlike other studies, I. Campos (2017) focused on studying the financial side of the energy market, rather than on the physical oil benchmarks. Four HAR models were built to model the CBOE Crude Oil Volatility Index using a specially developed abnormal search volume index and traditional macro-financial indicators. The standard out-of-sample methodology based on the constructed models was applied for prediction. Both in modeling and forecast-

ing, the results showed that ASVI has a positive correlation with oil volatility. It has also been proven that Google data can bring additional information to a model that helps in predicting.

In the article by Mohammed Elshendy (2018) were used 4 media platforms (Twitter, Wikipedia, Google Trends, and GDELT) to examine whether Internet data allows you to predict changes in WTI crude oil prices. For the analysis, the researchers built the ARIMAX model, because it allows adding different external variables, and also indicates how much they contributed to the forecast. A strong interaction was found between the search queries and keywords from all the above mentioned online resources and energy market indicators. In particular, Google trends have a positive correlation with oil prices and also have the highest prediction ability at a three-day lag.

One of the latest studies exploring the relationship between Google data and the oil market was presented by researchers from China at the IEEE International Conference on Big Data 2018, Tao *et al.* (2018). The study is based on collecting a large set of keywords. Using a special filtering algorithm, search queries were divided into 7 sections so-called factors search volume indices (FSVI). For prediction, model ARMAX was built based on Google data and WTI crude oil prices. After estimating the model they found that the model with FSVI variables improves the forecasting of energy prices by a quarter.

This study differs from the research papers discussed above in several ways. Firstly, we collected a unique set of keywords, and the time series of each search query was used as a separate variable in the models. This is useful because it allows to see the significance of a particular search term in the analysis. In addition, this thesis is based on the three main oil prices, as well as their average value, and helps to imitate a more accurate market situation. Besides, CBOE OVX is added to reflect oil volatility on a global level.

Chapter 3

Data

Three main time series are used to study the effect of Internet data on forecasting the volatility of oil prices. The first one is the set of keywords time series that was obtained from Google Trends. The next one is CBOE Crude Oil Volatility Index from Yahoo!. And the last one crude oil spot prices that was collected from The World Bank. The constructed data set contains 140 observations, all monthly data are available from May 2007 to December 2018, where 01.05.2007 is the earliest date for which monthly Volatility Index is available.

3.1 Google Trends

Google Trends is a publicly accessible web application that has been available since 2012.It is based on Google statistics and shows the frequency of searching for a particular term in relation to the total volume of search queries. The data do not reflect the absolute volume of the search, but only the relative popularity of a particular query at a specific time and in a specific geographic location. The analysis does not contain queries that have a too small number of searches. Also, it does not take into account queries that were entered several times within a short period of time by the same user. As a result, Google search volume index (GSVI) is built and it varies from 0 to 100, the maximum value is assigned to 100. A value of zero is also assigned when the absolute number of searches for a specific query is below the minimum limit. Data are normalized and can vary by a few percent from day to day as GSVI is calculated applying the sampling method.

The service provides various settings that can help to obtain more detailed information about queries. For example, the time filter can be set for any

period for which data are available, that is, for any time interval since January 2004. Weekly data can be obtained for most time periods. One hour is the smallest interval in open access on the official site [1], in this case, time series have a minute frequency. However, the service automatically sets the monthly frequency for intervals exceeding 5 years.

It is possible to clarify the meaning of the word by denoting the category the search query relates to. This feature is especially useful for search queries that have several completely different meanings. For example, the word "apple" have several meanings. Category "Technology" can help to focus the analysis on the company Apple Inc.

The next useful feature of Google Trends is the ability to get an idea of the popularity of a query for a specific region. You can find out the popularity of the query in Google by country, region and even cities. In addition, you can choose the source of the analyzed data: Web search, images, news, as well as search on YouTube and Google Shopping. Furthermore, the service allows you to compare up to five different queries or compare the popularity of a search query in up to five different geographical locations.

There are various ways to download the Google Trends time series. One of the possibilities is to download the CSV file from the official site. The next opportunity is to apply special packages and functions in various programming languages and statistical software. For this study, the programming language R and the specially developed googletrendsR package[2] were used. It allows to get data directly in RStudio and, moreover, it retains the possibility of using all the configurations available on the official Google website.

3.2 Selecting keywords

Correct keywords selection is a crucial factor in Google econometrics. This is one of the most complex and controversial problems in web search based studies. There are various methods and hypotheses used to choose optimal search queries. Most scientific works use so-called economic intuition. This method is that researchers select words based on the field of study. The basic logical criterion is preliminary economic knowledge that a specific keyword can fully correspond to the phenomenon under study. For example, R. Kulkarni (2009) analyzed housing prices.They were based on the hypothesis that due to falling

[1] `http://trends.google.com/trends`
[2] `http://github.com/PMassicotte/gtrendsR`

home prices in the US, homeowners decide to refinance their property. Therefore, as the main selected keywords were "house for sale", "home refinance", and "home value". Choi & Varian (2012), Tuhkuri (2015) in their researches of the unemployment rate based on the logic that people who have lost their jobs will not only look for words related directly to job search but will also search information about the unemployment benefit system.

The next possible option for selecting keywords is Google Correlate service[3]. Google launched it as Google Trends auxiliary service to improve search queries analysis. It is a fully automatic method to analyze not a particular query, but time series data. The service compares the data to get a list of requests with similar time series patterns. As in the Google trend, it is not based on absolute search volumes, but the relative values. It should be noted that Google Correlate can not be regarded as a final source of data for research. Since service calculate only correlation between the data, which should not be equal to causality. It is also not possible to work with the logarithmic or difference form data and it may have a negative effect on more detailed studies.

The optimal amount of keywords for research remains the following controversial issue in Google econometrics. Some researchers focus only on the time series of a single query. Francesco D'Amuri (2010) selected only one keyword to study US unemployment. Also, Pavlicek & Kristoufek (2015) used only the word "job" in four languages corresponding to the Visegrad Group countries for which the study was conducted. Similarly, Choi & Varian (2012) restricted their travel analysis by single subcategory "Hong Kong". At the same time, another studies use a large set of keywords and examine each search query individually as well as combining them into specific groups. For example, Tobias Preis (2013) selected 98 related to finance keywords to analyze stock markets. Nikolaos Askitas (2009) used four groups that contained from one to eight search queries to analyze the impact of Google trends on unemployment forecasting in Germany.

Stéphanie Combes (2016) studied how Google trend data improves forecasting. They applied different methods and analyzed whether the number and variety of searches give much more useful and accurate information. They got the result that the introduction of additional search queries time series data model does not always have a positive effect on forecasting. They conclude that it is almost impossible to obtain the best forecasts for data with a wide range of series using only automatic methods and without any human intervention.

[3] `https://www.google.com/trends/correlate`

Search term	Minimum	1Q	Median	Mean	3Q	Maximum	Kurtosis	Skewness
oil stock	13.0	18.0	26.0	29.8	38.0	100.0	7.0	1.7
oil price	11.0	17.0	28.0	33.6	44.0	100.0	3.9	1.2
oil reserves	9.0	16.0	20.0	23.4	26.0	100.0	14.6	2.9
oil production	27.0	36.0	39.0	42.2	46.0	96.0	7.6	1.8
news oil	14.0	18.0	22.0	28.9	37.0	100.0	8.5	1.9
world oil	25.0	30.0	33.0	35.84	38.0	100.0	16.1	3.0
crude oil	15.0	23.7	32.5	37.3	44.0	100.0	4.7	1.5
brent oil	5.0	9.0	14.5	22.6	31.0	100.0	6.5	1.8
baker hughes	21.0	32.0	40.0	41.31	51.0	100.0	5.4	0.9
petroleum	34.0	40.0	47.5	47.9	54.0	75.0	2.7	0.4
opec	9.0	17.0	21.5	24.9	29.3	65.0	4.6	1.4
wti	8.0	13.0	18.0	26.0	34.3	100.0	5.9	1.7
iea	31.0	43.0	51.0	56.7	69.3	99.0	2.3	0.6

Table 3.1: First group of selected keywords

Therefore, it is very important to pay special attention to the keywords selection, analyze the various groups of words and select only those that present the most significant impact on the study.

For this study, two groups of search queries were used. The first group of words was selected based on economic knowledge of the oil market and assumptions about the people's behaviour who are interested in information about the energy market. The preliminary list contained five search terms such as "crude oil", "oil price", "wti", "brent", and "oil stock". It was based on economic intuition and according to Google econometrics researches that are related to the oil market(Mohamad Afkhami (2017), Jian-Feng Guo (2013), Dean Fantazzini (2014)). The next step was to find out a similar time series using Google Correlate. However, service does not give an opportunity to analyze worldwide data. This assumption does not consider the objectives of this study, because oil market is usually considered as global. Therefore, we analyzed the keywords from the initial list for seven countries[4] separately. This method helps to approach the worldwide model since these countries account for almost 2/3 of the global population and the global GDP[5]. Five of the most correlated queries were obtained for each initial search term, that is, 25 time series for every national level. But only 19 of them co-exist for data from all countries. We selected time series that do not consist of zero value in Google Trends and get the final list that contains 13 keywords. Selected search terms

[4]USA, Brazil, Russia, China, Germany, Japan, UK

[5]According to OECD statistics

and their descriptive statistics are presented in table 3.1. All words that were obtained are in English. Most of them are closely related to the topic of oil and do not require additional explanations and specific knowledge. Also, there are two words that mean the names of organizations associated with the crude oil market. One of them is International Energy Agency (IEA) that control the supply and demand of energy resources around the world. And Baker Hughes, a GE company (BHGE), it is one of the largest companies that provides services in the field of crude oil. The Google Trends restriction options were not applied, so each time series was collected based on worldwide data from each available category from 2004 to the present.

The second group of keywords contains the names of states that have a significant impact on the global oil market. This list includes all countries of the Organization of Petroleum Exporting Countries (OPEC) and four countries that play an important role in the energy market: Russia, USA, China, and Canada. The choice was based on the hypothesis that major political and social changes in these countries could affect the oil price. Therefore, to obtain a more accurate analysis, data from Google Trend were restricted to the category "News".

3.3 Oil price related data

There are three major oil benchmarks that are Brent Crude, WTI (West Texas Intermediate), and Dubai Crude. They are crucially important in the formation of oil prices. International agencies publish price quotes for these benchmarks, which are subsequently used by traders. Each marker variety is a benchmark for a specific part of the world:

- Brent is a reference for pricing for the European and Asian markets. According to ICE Futures, it is the leading global price benchmark, which is associated with the barrel cost of about 70 percent of all exported oil grades;
- WTI is the marker variety for the prices of the countries of the western hemisphere (mainly the USA and Canada). Also, it is a price benchmark for these countries. It is extracted from the USA in the state of Texas. There is a high demand for this type of oil in the USA and China.
- Dubai Crude is crude oil that is extracted from United Arab Emirates.

It is usually sold in countries of the Asia-Pacific region. The variety is the basis for pricing for export crude oil in the Persian Gulf.

Spot oil price data were obtained from The World Bank. Monthly data are available since 1960 for Dubai Crude, 1979 for Brent, and 1982 for WTI, and are expressed in dollars for barrel. Figure 3.1 demonstrates a historical change in oil price from January 1982 to December 2018 for all three oil benchmarks that were discussed above. As we see with the collapse of the Soviet Union, the price of oil has slightly "settled down", and it's fluctuations in the 90s of the last century were around $ 20 per barrel. A sharp jump in spot oil prices occurred during the 2008 crisis. Then the crude oil price jumped to a historical maximum to 130 dollars per barrel. And during the crisis of 2011-2013 price quotes briefly dropped below $ 100 but the average annual price was 108.56 dollars. All examined time series have absolutely similar trend and moreover, they take almost the same values during all time except period between 2011 and 2014, when WTI prices were significantly lower than the other two. This was due to increased shale oil production in the USA.

For a more global interpretation of the energy market, this thesis considers not only each described oil standard separately, but also their average value. Volatility is a statistical indicator that characterizes price change over time.

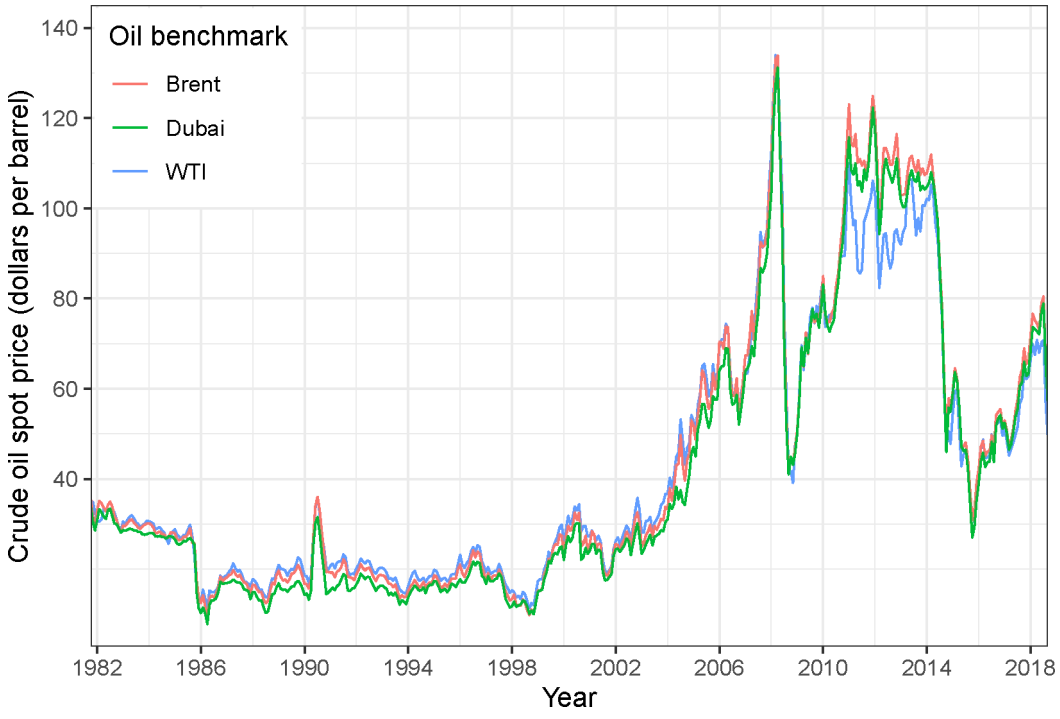

Figure 3.1: Oil price history

Volatility indices are indicators of fear among market participants, reflecting their actions and expectations for the near future. Crude Oil Volatility Index (OVX) has been provided by Chicago Board Option Exchange since May 2007, calculated using VIX methodology. Unlike other indexes that represent the past, OVX index expresses the current market estimate of the expected 30-day volatility in crude oil prices. The VIX is constructed using the Black-Scholes option pricing model to calculate implied volatility for a number of stock index options. These data are combined to give a full assessment of market expectations regarding volatility in the short term. Mathematically, the volatility index is expressed as a percentage with reference to a specific period.

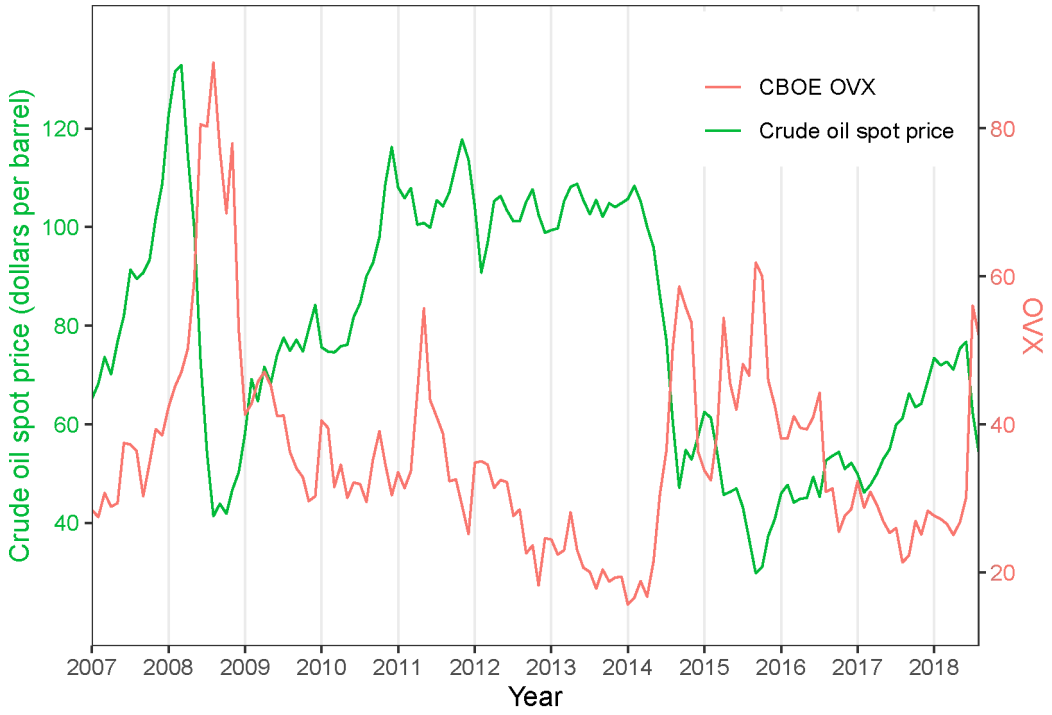

Figure 3.2: Relationship between prices and volatility index

Figure 3.2 demonstrates the relationship between the volatility index and average value of oil prices from May 2007 to June 2018. The graph clearly shows a negative correlation between the two time series. That is, the maxima of the OVX index coincide or precede the minima of oil price. Conversely, the minimum OVX values coincide with peaks in oil contracts. For example, when prices dropped tragically in 2008, the volatility index reached a historic high.

Chapter 4

Methodology

This chapter presents theoretical remarks and the most methods that are used in this thesis. Firstly, unit root and stationarity tests are introduced to analyze the dynamic of data. The next subsection demonstrates the OLS methods and related tests. Then VAR and Granger Causality method are discussed. And at the end, a prediction method and measurements of quality of forecasts are described.

4.1 Unit root test – Stationarity

A lot of technics require assumption that time series have to be stationary. According to Wooldridge (2012) stochastic process $\{z_t : t = 1, 2 \dots\}$ is said to be stationary if its main properties stay unchanged across time.

A weak stationary (or covariance stationary) process has three conditions:

(i) $\mathrm{E}(z_t) = \mathrm{E}(z_{t-1}) = \mu$, for $\forall t$

(ii) $\mathrm{Var}(z_t) = \sigma^2 < \infty$, for $\forall t$

(iii) $\mathrm{Cov}(z_t, z_{t-k}) = \gamma_k$, for $\forall t, k \geq 1$

In other words, it means the mean and variance of the stochastic process are constant over time, and there are no seasonality or autocorrelation.

There are several formal statistical tests for stationarity. In this paper are provided two of them: Augmented Dickey-Fuller test and Kwiatkowski-Phillips-Schmidt-Shin test.

4.1.1 Augmented Dickey-Fuller test (ADF)

Dickey-Fuller test (as well as its augmented version) is considered one of the basic ones for detecting a unit root. It was first introduced in Dickey & Fuller (1979).

The test assumes the construction of a simple autoregression model:

$$\Delta z_t = \delta + \beta y_{t-1} + \gamma t + \sum_{i=1}^{k} \alpha_i \Delta y_{t-i} + \epsilon_t$$

where α is an intercept constant called a drift,γ is the coefficient on a time trend

$$H_0 : \beta = 0, \text{ presence of a unit root}$$
$$H_1 : \beta < 0 \text{ , time series is integrated of order 0}$$

Ordinary t test is calculated to check the hypotheses:

$$DF = \frac{\hat{\beta}}{se(\hat{\beta})}$$

ADF statistics by meaning and formula is Student's statistics, but it has a slightly different distribution, therefore other critical values are used. If the statistical value lies to the left of the critical value (critical values are negative) at a given significance level, then the null hypothesis of a unit root is rejected, and the process is recognized as stationary. Otherwise, the hypothesis is not rejected, and the process may contain unit roots, that is, non-stationary (integrated) time series.

4.1.2 Kwiatkowski, Phillips, Schmidt and Shin Test (KPSS)

Unlike the Dickey-Fuller test, KPSS criterion (Denis Kwiatkowski & Shin (1992)) considers the hypothesis that the time series are stationary, as the null hypothesis.

$$z_t = \delta + \beta t + \theta \sum_{i=0}^{t} \xi_i + \epsilon_t$$

The essence of the test is that if a random walk occurs in this process, this will lead to systematic deviations from the trend in some parts of the series. Two competing hypotheses are put forward:

$$H_0 : \beta = 0$$
$$H_1 : \beta < 0$$

This test is checked by LM-statistics (Lagrange multipliers test) calculated by the formula:

$$KPSS = \frac{1}{n^2 s^2} \sum_{t=1}^{n} S_t^2$$

The process of testing the hypothesis in KPSS test is identical to the generally accepted one: if the calculated value is less than the table value, the researcher has no reason to reject the null hypothesis of stationarity.

4.2 OLS regression analysis

Ordinary least squares is a method for finding optimal linear regression parameters, such that the sum of squared errors is minimal. In the case of linear regression, the minimization problem is mathematically described in the following way:

$$\sum_{i=1}^{n} (y_i + \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \to min,$$

where y_i is an actual value, $\hat{y}_i$ is an estimated value, and e_i^2 is the squared errors of regression.

According to Wooldridge (2012), there are five assumptions that have to be satisfied for correct interpretation in the analysis of the usual OLS standard errors, t and F statistics. Firstly, linearity and weak dependence have to present. Besides, there should be no perfect collinearity between independent variables. Also, there are assumptions of zero condition mean and homoscedasticity, and the last one is the abscence of autocorellation of errors. To verify last two assumptions White test and Breusch–Godfrey test are implemented in this thesis.

4.2.1 White test

White test is a universal procedure for checking the heteroscedasticity of random errors in a linear regression model that does not impose special restrictions on the structure of heteroscedasticity. It was introduced by White (1980).

The test uses the regression residuals estimated using the ordinary least squares method. For the test, auxiliary regression of the squares of these residues is estimated for all regressors, their squares, and pairwise products:

$$e_t^2 = \alpha_0 + \alpha^T x_t + x_t^T A x_t + u_t,$$

where e_t is residuals and x_t is independent variables from original regression, α_0 is a constant, α is a linear coefficient vector and A is a coefficient matrix for squares and pairwise products of variables.

The test verifies the null hypothesis of the absence of heteroscedasticity. That is, model errors are assumed to be homoscedastic — with constant variance. LM-statistics is used to test this hypothesis, $LM = nR^2$, where R^2 is the auxiliary regression determination coefficient and n is the number of observations.

4.2.2 Breusch–Godfrey test

Breusch–Godfrey test is the procedure used in econometrics to test autocorrelation in random errors of regression models. The test is performed using an auxiliary model, in which the dependent variable is the residuals of the estimated model. This auxiliary model has the form:

$$e_t = \alpha_0 + \sum_{i=1}^{k} a_i X_{it} + \sum_{s=1}^{m} \rho_s e_{t-s} + \epsilon_t,$$

where ρ is a residual autocorrelation coefficient.

The null hypothesis is verified that all coefficients for residuals are simultaneously equal to zero. The verification is carried out using the corresponding LM-statistics, equal to $R^2 n$, where R^2 is the coefficient of determination of the auxiliary model and n is the sample size.

4.3 Vector Autoregression (VAR)

Vector autoregression model was proposed by Sims (1980), who demonstrated its advantages in analyzing economic time series. This type of model is usually used for systems for predicting interrelated time series and for analyzing the dynamic effects of random disturbances on a system of variables. VAR is a system of equations in which each endogenous variable is represented by a linear combination of all variables at previous periods.

Mathematical representation of the vector autoregression model of order p:

$$y_t = \alpha_0 + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + \epsilon_t = \alpha_0 + \sum_{m=1}^{p} \Phi_m y_{t-m} + \epsilon_t \ ,$$

y_t is a k-dimensional vector of observed time series variables,
α is a $(k \times 1)$ vector of constants,
$\Phi_1 \dots \Phi_p$ are $(k \times k)$ matrices of coefficients,
$\{\epsilon_t\}$ is a sequence of serially uncorrelated errors.

4.3.1 Lag Lenght Selection

One of the drawbacks of the VAR model is considered to be uncertainty in the choice of a suitable lag length. Because an excessive amount of lags increases forecast errors, while omitting the necessary lags can cause estimation bias. However, applying of various information criteria can be one way to solve this problem. The most common criteria are Akaike information criterion (AIC), Bayesian information criterion (BIC) and Hannan–Quinn information criterion (HQ).

$$
\begin{aligned}
AIC &= 2k - 2ln(\hat{L}) \\
BIC &= ln(n)k - 2ln(\hat{L}) \\
HQ &= -2L_{max} - 2kln(ln(n))
\end{aligned}
$$

Using different criteria can lead to the choice of different models. Though there are also recommendations that describe the advantages and disadvantages of using a specific information criteria depending on the sample size, data frequency, and other conditions, Lutkepohl (2005).

4.3.2 Stability of VAR process

Stability is one of the basic conditions of a VAR model and checks whether a model is a good indicator of how time series has changed over time.

By mathematical lemma VAR model is stable if

$$det(I_{Kp} - \Phi z) \neq 0 \text{ for } |z| \leq 1, \text{ where}$$

$$
\Phi = \begin{pmatrix}
\Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\
I_K & 0 & \cdots & 0 & 0 \\
0 & I_K & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \\
0 & 0 & \cdots & I_K & 0
\end{pmatrix}
$$

According to Canova (2007), it means that VAR process can be defined as stable if all eigenvalues of Φ have modulus less or equal than one.

4.3.3 Granger Causality

Granger causality is a widely used statistical concept that formalizes a causal relationship for time series. The definition and testing procedure was proposed in Granger (1969). The basic idea is that the degree of influence of one system on another is estimated by changing the accuracy of predicting the behavior of the first system when introducing information on changes in the second system into the predictive model.

Formally, assume $X = x_1, x_2, \dots x_T$ and $Y = y_1, y_2, \dots y_T$ are time series. There is Granger causality between series $x_t \rightarrow y_t$ if the variance of optimal linear predictor of $\hat{y}_{t+1}$ based on $y_1, \dots y_t, x_1, \dots x_t$ is smaller than the optimal linear predictor of $\hat{y}_{t+1}$

$$E\Big((\hat{y}_{t+1} - yt + 1)^2 | y_1, \dots y_t, x_1, \dots x_t\Big) \leq E\Big((\hat{y}_{t+1} - y_{t+1})^2 | y_1, \dots y_t\Big)$$

In the context of the VAR model, the variable x_t "Granger-causes" y_t, if the lag coefficients x_t in the first equation are statistically significant and the coefficients at lags y_t in the second equation are statistically insignificant and vice versa. There may also be cases of bivariate causality or mutual independence.

4.4 Forecasting

One of the main tasks of forecasting time series is to choose between several potential models. Also, sometimes the amount of data available may not be sufficient to provide a statistically significant estimation of market parameters. One of the most effective ways to solve these problems is a method of rolling window estimation. Its main idea is to create pseudo-new observations using sequential samples. In this thesis, fixed window method is used. It is based on the fact that the window, that is, the number of selected observations (ΔT), remains unchanged and moves one observation at a time. Then the total number of returns is $T - \Delta T + 1$. This method allows you to calculate RMSE and MAE to verify the quality of the predictive ability of the model.

4.4.1 AR(1) model

To implement forecasting procedure AR(1) model is chosen. Two models are stated for future comparison. Reference model has form:

$$OIL_t = \beta_0 + \beta_1 OIL_{t-1} + \epsilon_t \tag{4.1}$$

where OIL is one of the oil price related variable.

And the comparison model presents as follow:

$$OIL_t = \beta_0 + \beta_1 OIL_{t-1} + \sum_{i=1}^{I} \gamma_i GT_{it} + \epsilon_t \tag{4.2}$$

where GT is addition explanatory variable that represents set of search queries.

4.4.2 Mean Absolute Error

MAE calculates the average value of forecast errors without taking into account their direction. This is the average for the sample of the absolute value of the difference between the forecast value and the actual observation. In other words, this measure indicates how large the error can be expected from the forecast on average.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} |x_j - \hat{x}_j|$$

where x_j is a true value and $\hat{x}_j$ is a forecast.

4.4.3 Root Mean Squared Error

RMSE is another alternative to test how reliably the model which is chosen as a forecast generator describes the retrospective of the phenomenon under study.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

where y_j is a true value and $\hat{y}_j$ is a prediction.

4.4.4 Diebold-Mariano test

Diebold-Mariano test is statistical test that allows you to compare the quality of the forecasts of the time series of two predictive models. It was introduced by Diebold & Mariano (2002). This test is resistant to various deviations from standard assumptions about the properties of the prediction error. Thus, there is no need to satisfy all classical assumptions.

Two competing hypotheses are put forward:

$$H_0 : E\{\epsilon_t^1\}_{t0}^T = E\{\epsilon_t^2\}_{t0}^T$$
$$H_1 : E\{\epsilon_t^1\}_{t0}^T \neq E\{\epsilon_t^2\}_{t0}^T$$

so, the null hypothesis states that the quality of forecasts is the same against the alternative that there are differences.

Diebold-Mariano statistics is calculated as:

$$S = \frac{\bar{d}}{\sqrt{L\hat{R}V_d/T}}$$

where $\bar{d}$ is the mean loss dierential and $L\hat{R}V_d$ corresponds to a consistent estimate of the asymptotic variance of $\sqrt{T}\bar{d}$. According to Diebold & Mariano (2002), this statistic follows normal distribution.

Chapter 5

Empirical results

This section presents the most statistically significant empirical results using the methodology from the previous chapter. The first four subsections demonstrate the constructed models and test results. The last subsection briefly discusses the interpretation of the results.

5.1 Stationarity

Firstly, following Kristoufek (2013) we transformed our time series into three basic forms, that are logarithmic, difference and logarithmic difference. It is known that all criteria that check the belonging of a time series to the class of stationary or non-stationary processes have some drawbacks or limitations. Therefore, we performed ADF and KPSS tests simultaneously, since they have opposite zero and alternative hypotheses and in combination give more reliable results for analyzing the series of their belonging to a particular class.

Test results provide that both original and logarithmic forms are non-stationary and have a unit root. So, since all time series in both difference forms conform to the assumption of stationarity and do not contain a unit root, we choose the logarithmic difference for Google search queries as well as for oil price data. It allows us to get a more accurate interpretation of our subsequent analysis.

The detailed results of ADF and KPSS tests for logarithmic and logarithmic difference forms of all time series are summarized in Table 5.1.

	ADF		KPSS	
Google Trends	log	diff log	log	diff log
oil stock	-0.0059	9.1774***	0.0592	0.0289
oil price	0.2538	-7.2842***	0.152	0.0608
oil reserves	-0.5798	-8.9114***	0.447**	0.0194
oil production	-0.3671	-9.2147***	0.256	0.0162
news oil	-0.0561	-8.8198***	0.22	0.0255
world oil	-0.356	-11.3354***	0.429*	0.0158
crude oil	-0.0337	-8.1999***	0.104	0.0448
brent oil	-0.081	-8.4639***	0.518**	0.0395
baker hughes	-0.4554	-11.8394***	0.518**	0.178
petroleum	-0.8473	-13.2534***	0.739**	0.0182
opec	-0.3965	-9.0276***	0.14	0.0328
wti	0.0122	-8.7639***	0.2	0.0431
iea	-0.9089	-14.5509**	1.39***	0.0142
kuwait	-0.3163	-12.1737***	1.93***	0.0324
libya	-1.5125	-9.1212***	0.167	0.0475
uae	0.0913	-11.3162***	1.22***	0.0235
angola	-0.7409	-12.2528***	0.161	0.0299
venezuela	-0.6791	-13.4402***	1.32***	0.0176
saudi arabia	-0.0176	-11.9851***	2.15***	0.0116
iran	-0.7132	-12.6286***	0.927***	0.0168
iraq	-2.0532**	-10.5898***	0.451*	0.032
ngeiria	0.5184	-11.267***	0.301	0.0999
russia	-0.0787	-12.0631***	0.726***	0.0346
china	-0.334	-11.6906***	0.772 ***	0.0307
usa	-0.6966	-12.2199***	0.658**	0.0138
canada	-0.7293	-11.2868***	1.37***	0.0138
Oil series	log	diff log	log	diff log
WTI	-0.4477	-5.5731***	0.176	0.0541
Brent	-0.3988	-5.9276***	0.156	0.0586
Dubai	-0.3446	-5.7801***	0.148	0.0588
Average	-0.4009	-5.6534***	0.158	0.057
OVX	0.103	-7.913***	0.114	0.0571

Note: *, ** , *** denotes rejection of null hypothesis at the 10%, 5% and 1% level of significance, respectively.

Table 5.1: ADF and KPSS tests

5.2 Basic relationship

To ensure that our variables have the potential for further analysis, we run a simple OLS regression and examine the relationship between the logarithmic difference of the oil price data and the logarithmic difference of Google search query data.

$$\Delta log(OIL)_t = \beta_0 + \sum_{i=1}^{I} \gamma_i \Delta log(GT)_{it} + \epsilon_t \tag{5.1}$$

The endogenous variable OIL is corresponds to one of five time series that are Brent, WTI, Dubai oil prices, their average and volatility index. And the exogenous variable GT presents set of time series of all selected keywords from the first group(Table 3.1) and the second group(Table A.1).

White test was used to control the heteroskedasticity of errors. We do not reject the null hypothesis and assume homoskedasticity for all models except model with WTI oil price as an endogenous variable since its p-value is 0,08912. For more detailed results see Table A.2. For control the autocorrelation of errors was used LM-test. Results are summarized in Table A.3. Only for two models, we do not reject the null hypothesis of the absence of autocorrelation, which are models with Brent oil price and OVX. Other models failed the test up to 12 lags and demonstrated strong autocorrelation. Thus, for these models we selected Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors.

The best results are represented by the model with a crude oil volatility index. Its R^2 is 0.309, it means that approximately 30% of OVX variance is predictable by Google search queries. Five variables are statistically significant at 10% significant level. Four of them have positive sing of coefficient and it means that these variables move together with our dependent variable. And variable "libya" has coefficient that is equal to -1.89 and thus this search term suggests a negative effect on volatility index series. The rest of the models have adjusted R^2 in the interval between 0.05 to 0.1 and have only two statistically significant variables. It corresponds to the assumption that the model does not fit good and describes the data insufficiently.

More detailed information about regressions outputs is provided in Appendix B.

5.3 Vector Autoregression

To identify causality relationships, we implemented five VAR models based on our obtained logarithmic difference data. As in the case of linear regression, we added all Google Trends series and one of five oil series to each model. The number of lags was chosen using the information criteria described in the previous chapter. Based on them, vector autoregression model of order one was selected in each case, thus VAR(1). Asymptotic Portmanteau test was provided to verify the stability of our models. Its null hypothesis states that there is no autocorrelation in the residuals of a model and, therefore, the model can be considered as stable. Also, we created the plot of inverse roots (Figure A.1). According to Lutkepohl (2005), the estimated VAR is considered as stable if all roots have modulus less than one and lie inside the unit circle. Based on the test results and inverse roots of characteristic polynomial analysis, we concluded that all models are stable and can be used for the following tests.

Wald test was performed to verify Granger causality between variables of obtained models. Results demonstraed both one-sided and two-sided causality between price series and some Google Trends. For relations in which two-side causality was detected, we conducted an impulse response analysis. More detailed descriptions of the provided procedure and collected results of each model are presented below.

5.3.1 Granger causality and impulse response analysis

According to Wald test results crude oil volatility index has one-side causality with three variables("brent oil", "wti", "angola") that are following the assumption that these search queries Granger-cause OVX. There is a single two-side Granger causality in this case, that is the relation between OVX and Google Trend "libya". Based on impulse response function (Figure 5.1) we see that these two variables have a greater effect in the first two lags. There are 0.05, responses from the log-differenced of this search query variable to schock in OVX. Response falls to zero until 3 steps ahead.

Oil price data [1] have a rather similar causal relation with Google Trends. For all four cases there exist two-side Granger causality with two search queries ("oil reserves","iea"). Google Trends "oil reserves" positively responses to impulse in oil prices and go up to 0.023 in first lag. However, after second lag

[1] Average price, WTI, Brent, Dubai

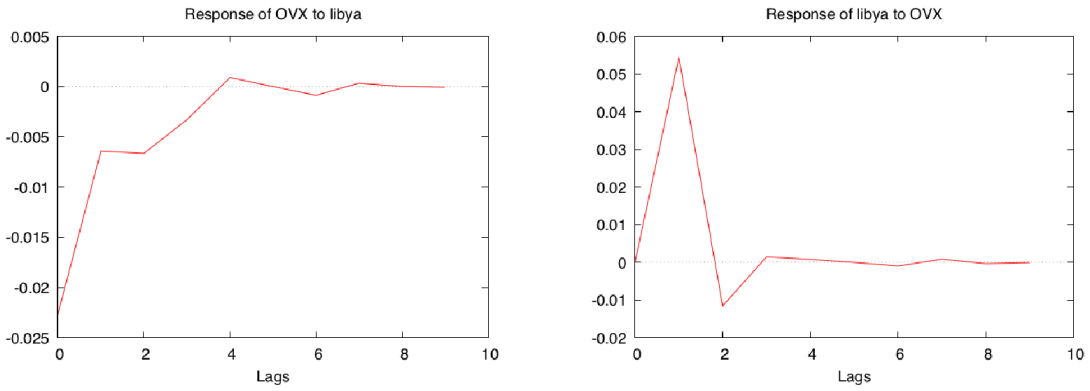

Figure 5.1: IRF OVX - "libya"

the effects decrease and start to tend to zero. Figure 5.3 demonstrates that the shock of oil prices adversely affects the search query "iea" in the first lag. All responses get to the zero value until 7 steps ahead.

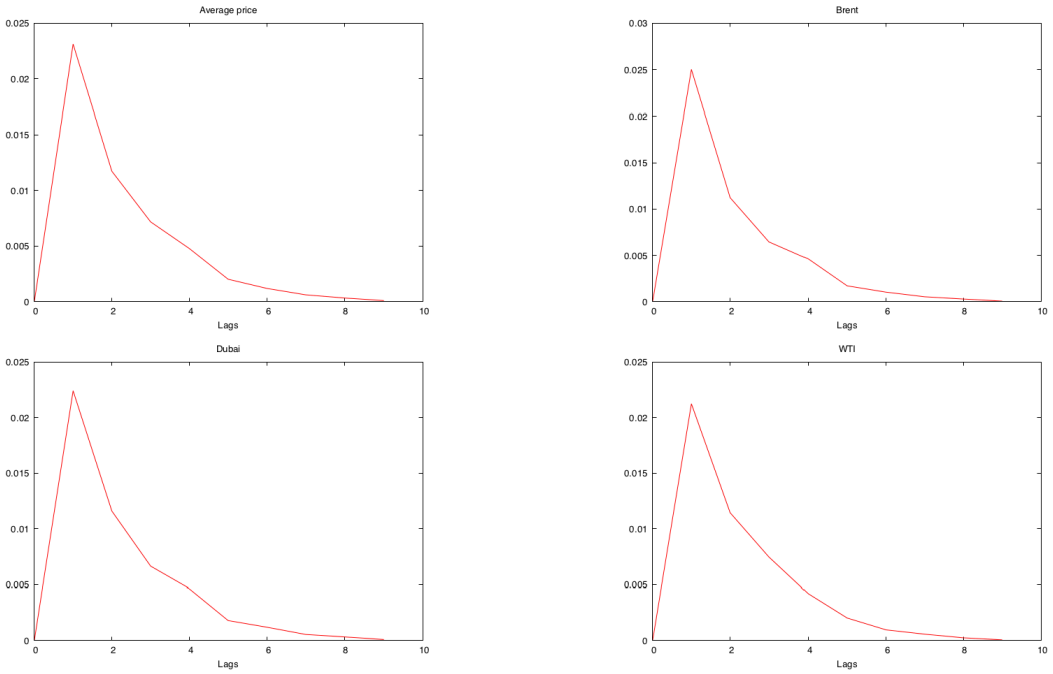

Figure 5.2: Response of "oil reserves" to Oil price

According to Wald test all oil price variables are Granger causes logarithmic difference of "world oil" series. Most of the one-side causal relationships found are associated with keywords indicating the names of countries, that are "uae", "venezuela", "russia", "china", and "canada". WTI has a two-side causality with the search query "canada" and they positively response on the shocks

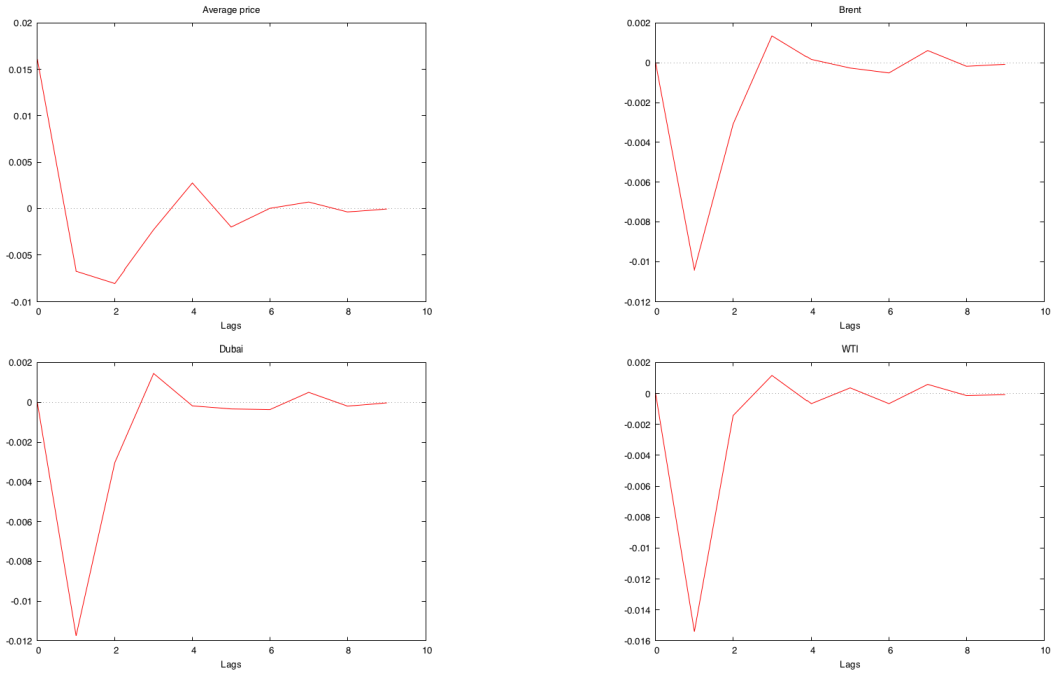

Figure 5.3: Response of "iea" to Oil price

to each other in first lag (Figure 5.4). More detailed results of Wald test are summarized in Table A.4 - Table A.8.

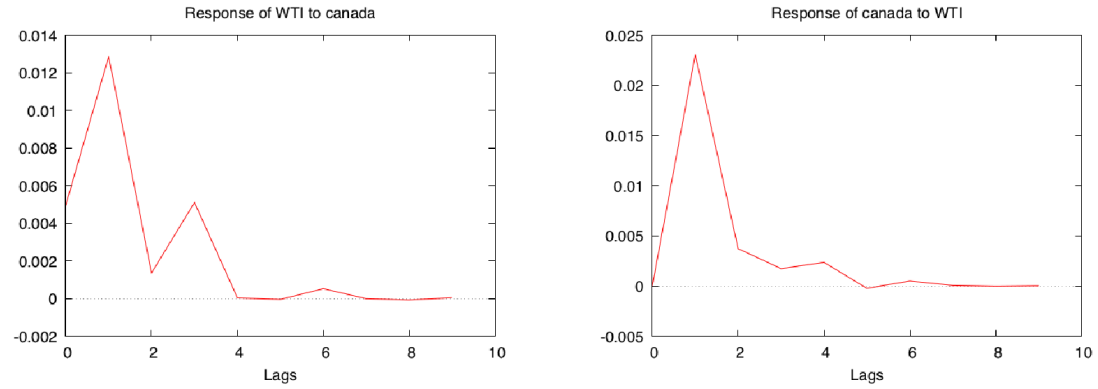

Figure 5.4: IRF WTI - "canada"

5.4 Forecasting

Rolling fixed window procedure was implicated by using AR(1) models that were stated in equations 4.1 and 4.2. The window size was stated by 50 observations that corresponds to our in-sample part of the data set. Thus, as a result, we obtained 90 returns as one observation of the data set was lost

because of the differencing of non-stationary time series. As in the previous cases, logarithmic difference data were used to build AR(1) models both reference and comparative one with Google Trends. Further, for simplicity, the name of the model corresponds to the name of its dependent variable.

Most of obtained results suggest that Google Trends improve the models and therefore improve the quality of forecasting changes in oil prices. In particular, in OVX model after adding Google Trends RMSE decreased from 0.15 to 0.11, that is, almost 26% and MAE decreased from 0.11 to 0.09 by 18%. The null hypothesis of DM-test can be reject at 5% significant level, so we came to the conclusion that OVX model with Google Trend has better predicting ability.

Dubai and Brent models have almost similar results. Presence of Google Trends in models reduces RMSE by almost 16% and MAE by almost 11%. However, we can not reject the null hypothesis of DM-test even at 10% significant level. Thus, it demonstrates that Google Trends did not significantly improve these models.

Adding of Google Trends in the WTI and Average prices models causes the decrease in RMSE by almost 17% and in MAE by almost 10%. With DM statistics of 2.116 and 1.189 we rejected the null hypothesis at 10% significant level for both models. It allows us to conclude that in these case, models with Google Trends variables has better quality of prediction.

5.5 Discussion

The results show that the majority of causal relationships in our analysis are between oil prices and keywords from the second group that is related to news of the selected countries. As we know, according to Kilian (2009) and the subsequent literature around it, the oil market is strongly associated with various geopolitical and economic phenomena. This supports the idea that not only keywords that are directly related to a particular economic sector can improve forecasting, as it was demonstrated in many studies. But also, in the case of the oil market, search queries related to events occurring in countries with significant influence in the export and import of oil have causality with oil price volatility. It is also supported in Jian-Feng Guo (2013), where Libya war related keywords were successfully applied for oil price nowcasting. Also, Tao *et al.* (2018) used keywords set of geopolitical events and performed that it is helpful in the prediction of oil price volatility.

Measures of prediction accuracy RMSE and MAE demonstrated adding of

Google Trends into models causes improvement of their forecasting ability. But according to results of Diebold-Mariano test, these changes are not statistically significant for models with Brent and Dubai price. From these results it can be concluded that Google trends cannot help in predicting of these oil benchmarks. However, more sophisticated research by Jian-Feng Guo (2013) refute this assertion by performed that Google Trends significantly affect Brent price both short-run and long-run. This prompted us to conclude that our forecasting models are not ideal for the oil sector of the energy market and need to be improved in future studies.

Chapter 6

Conclusion

This thesis explored the relationship between specific Google search queries and crude oil price volatility. The main propose was to establish presence causality relations between these data and examined the forecasting potential of Google Trends series in case of the oil market.

Provided that the ADF and KPSS tests showed that the original series are non-stationary, data were used in the logarithmic difference form.

OLS regression was specified to examined the contemporaneous relationship between search queries and each of oil-related series individually. The model with OVX performed the best results with five statistically significant independent variables and adjusted R^2 is almost 0,3. We captured a positive effect of Google Trends on volatility index as the most significant coefficients are positive.

After applying Wilde test on VAR model, both unilateral and bilateral causal relationships were discovered. Most often Granger causality related to keywords that mean country names. This supports the idea that not only search queries that are directly related to the oil market can influence the volatility of oil prices but also search terms about the geopolitical events occurring in countries that play an important role in the energy sector of the economy. However, the results of the impulse response analysis for these cases showed that the response to shock exists in two sides but it is quite small. And it does not allow us to state with certainty about a strong causality between oil price volatility and Google trends.

Forecasting ability of Internet data was studied by providing out-of-sample procedure with two comparative AR(1) models. Using two measures of prediction accuracy RMSE and MAE, we found that all models are improved by

adding Google search queries. However, DM-test performed that these improvements are statistically significant only in three models. Thus, results demonstrated that Google Trends significantly helpful for the prediction of CBOE Crude Oil Volatility Index, WTI price, and the average price of three benchmarks.

To summarize, the research supports the hypothesis that Google trends are related to the volatility of oil prices and improve their forecasting models in some cases. However, some results and models need to be adjusted and supplemented. Further studies can reduce these research gaps and take full advantage of econometric analysis using search queries. We can offer two ways to improve the results obtained in this thesis. The first one is to use a deeper analysis and more sophisticated econometric models using the available data. The second possibility is a more accurate selection of keywords from Google search queries or the addition of data from other social networks. It will allow you to get a more complete picture of the market participants behaviour on the Internet.

Bibliography

A.N. Brodunov, K.G. Bunevich, V. L. (2015): "Analysis of factors affecting the stability of the ruble in the conditions of macroeconomic uncertainty." *Journal of Moscow Witte University* **16(1)**: pp. 24–29.

Canova, F. (2007): *Methods for Applied Macroeconomic Research.* Princeton University Press.

Choi, H. & H. Varian (2009a): "Predicting the Present with Google Trends." *Technical report, Google Inc.* .

Choi, H. & H. Varian (2009b): "Predicting Initial Claims for Unemployment Benefits." *Technical report, Google Inc.* .

Choi, H. & H. Varian (2012): "Predicting the Present with Google Trends." *THE ECONOMIC RECORD* **88**: pp. 2–9.

Dean Fantazzini, N. F. (2014): "Forecasting the real price of oil using online search data." *Int. J. Computational Economics and Econometrics,* **4(1/2)**: pp. 4–31.

Denis Kwiatkowski, Peter C.B. Phillips, P. S. & Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root?" *Journal of Econometrics* **54(1992)**: pp. 159–178.

Dickey, D. A. & W. A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American statistical association* **74(366)**: pp. 427–431.

Diebold, F. X. & R. S. Mariano (2002): "Journal of Business Economic Statistics." *Econometrica* **20(1)**: pp. 134–144.

Francesco D'Amuri, J. M. (2010): "Google it!" Forecasting the US Unemployment Rate with a Google Job Search index." *FEEM Working Paper* **31**.

Gary Koop, L. O. (2013): "Macroeconomic Nowcasting Using Google Probabilities." *Working Paper, University of Strathclyde and ECB.* .

Giannone Domenico, Reichlin Lucrezia, S. D. (2008): "Nowcasting: The real-time informational content of macroeconomic data." *Journal of Monetary Economics, Elsevier* **55(4)**: pp. 665–676.

Granger, C. W. J. (1969): "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* **37(3)**: pp. 424–438.

I. Campos, G. Cortazar, T. R. (2017): "Modeling and predicting oil VIX: Internet search volume versustraditional mariables." *Energy Economics* **66**: p. 194–204.

Jian-Feng Guo, Q. J. (2013): "How does market concern derived from the Internet affect oil prices?" *Elsevier Ltd.* **112**: p. 1536–1543.

Kilian, L. (2009): "Not all oil price shocks are alike: disentangling demand and supply shocks in the crude oil market." *American Economic Review* **99(3)**: p. 1053−1069.

Kilian, L. & D. P. Murphy (2014): "The role of inventories and speculative trading in the global market for crude oil." *Journal of Applied Econometrics* **29**: p. 454−478.

Kristoufek, L. (2013): "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era." *Scientific Reports* **3(3415)**.

Kristoufek, L. (2015): "What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis." *PLoS ONE* **10(4)**: pp. 17–27.

Lutkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis.* Springer.

Marco Lorusso, L. P. (2018): "Causes and consequences of oil price shocks on the UK economy." *Economic Modelling* **72**: pp. 223–236.

Martina Matta, Ilaria Lunesu, M. M. (2015): "Bitcoin Spread Prediction Using Social And Web Search Media." *UMAP Workshops* .

Meltem Gulenay Chadwick, G. S. (2012): "Nowcasting Unemployment Rate in Turkey: Let's Ask Google." *Central Bank of the Republic of Turkey* **12/18**.

Mohamad Afkhami, LindseyCormack, H. G. (2017): "Google search keywords that best predict energy price volatility." *Energy Economics* **67**: pp. 17–27.

Mohammed Elshendy, Andrea Fronzetti Colladon, E. B. P. A. G. (2018): "Using four different online media sources to forecast the crude oil price." *Journal of Information Science* **44(3)**: p. 408–421.

Nikolaos Askitas, K. F. Z. (2009): "Google Econometrics and Unemployment Forecasting." *Applied Economics Quarterly* **55(2)**: pp. 107–120.

Nikolaos Vlastakis, R. N. M. (2012): "Information demand and stock market volatility." *Journal of Banking Finance* **36(6)**: p. 1808–1821.

Pavlicek, J. & L. Kristoufek (2015): "Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries." *PLoS ONE* **10(5)**.

R. Kulkarni, K. E. Haynes, R. R. S. J. H. P. P. (2009): "FORECASTING HOUSING PRICES WITH GOOGLE ECONOMETRICS." *SCHOOL OF PUBLIC POLICY* **(2009-10)**.

Ratti, R. A. & J. L. Vespignani (2013): "Why are crude oil prices high when global activity is weak?" *Economics Letters* **121(1)**: p. 133−136.

Rishanki Jain, Rosie Nguyen, L. T. T. M. (2018): "Bitcoin Price Forecasting using Web Search and Social Media Data." *Oklahoma State University* **3601**.

Seung-Pyo Jun, Hyoung SunYoo, S. C. (2018): "Ten years of research change using Google Trends: From the perspective of big data utilizations and applications." *Technological Forecasting and Social Change* **130**: pp. 69–87.

Sims, C. A. (1980): "Macroeconomics and Reality." *Econometrica* **48(1)**: pp. 1–48.

Stéphanie Combes, C. B. (2016): "Nowcasting with Google Trends, the more is not always the better." *Conference: CARMA 2016 - 1st International Conference on Advanced Research Methods and Analytics* .

Suhoy, T. (2009): "Query Indices and a 2008 Downturn: Israeli Data ." *Discussion paper series. Research Department, Bank of Israel* .

Takuji Fueki, Hiroka Higashi, N. H. J. N. S. O. & Y. Tamanyu (2018): "Identifying oil price shocks and their consequences: the role of expectations in the crude oil market." *BIS Working Papers* **725**.

Tao, R., X. Zhang, & L. Zhao (2018): "Forecasting crude oil prices based on an internet search driven model." *2018 IEEE International Conference on Big Data (Big Data)* pp. 4156–4161.

Tobias Preis, H. S. M. . H. E. S. (2013): "Quantifying Trading Behavior in Financial Markets Using Google Trends." *SCIENTIFIC REPORTS* **3**: pp. 1–6.

Tuhkuri, J. (2015): *Big Data: Do Google Searches Predict Unemployment?* Master's thesis, University of Helsinki.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* **48(4)**: p. 817–838.

Wooldridge, J. M. (2012): *Introductory Econometrics: A Modern Approach,Fifth Edition.* South-Western, Cengage Learning.

Xin Li, ShouyangWang, X. Z. (2015): "How does Google search affect trader positions and crude oil prices?" *Elsevier B.V.* **49**: p. 162–171.

Y. Fondeur, F. K. (2013): "Can Google data help predict French youth unemployment?" *Economic Modelling, Elsevier* **30**: p. 117–125.

Zhi Da, J. E. & P. Gao (2011): "In Search of Attention." *THE JOURNAL OF FINANCE* **66(5)**: p. 1461–1499.

Appendix A

Data description and test results

Search term	Minimum	1Q	Median	Mean	3Q	Maximum	Kurtosis	Skewness
kuwait	35.0	47.0	57.0	57.5	66.0	96.0	0.3	-0.6
libya	2.0	3.0	3.0	5.9	4.0	100.0	6.0	44.5
uae	32.0	38.0	42.0	43.9	48.3	100.0	2.5.6	13.9
angola	13.0	19.0	22.0	25.9	27.0	100.0	2.9	10.5
venezuela	15.0	21.0	26.0	28.4	30.3	100.0	2.8	10.5
saudi arabia	12.0	22.0	26.0	28.8	32.0	100.0	2.8	11.4
iran	9.0	12.0	16.0	16.9	19.0	100.0	6.0	52.4
iraq	2.0	4.0	6.0	8.3	9.0	34.0	1.8	3.1
nigeria	12.0	19.0	39.5	36.5	48.3	100.0	0.2	-0.1
russia	16.0	21.0	27.5	35.7	41.0	100.0	1.5	1.5
china	41.0	50.0	56.0	58.3	64.0	93.0	0.9	0.5
usa	5.0	6.0	7.0	8.3	8.0	100.0	10.3	112.0
canada	12.0	14.0	17.0	18.9	22.0	56.0	2.4	8.6

Table A.1: Descriptive statistics of second group of selected keywords

Model	LM-statistic	p-value
Average price	62.869	0.143705
Brent	58.8776	0.238383
WTI	66.5329	0.0846415
Dubai	59.9393	0.209913
OVX	54.2037	0.390386

Note: The name of the model corresponds to its dependent variable

Table A.2: White Test results

Model	LMF-statistic	p-value
Average price	1.73459	0.0703868
Brent	1.46457	0.150347
WTI	1.74387	0.0685058
Dubai	1.70039	0.0777437
OVX	1.37266	0.191591

Note: The name of the model corresponds to its dependent variable

Table A.3: Breusch–Godfrey Test results

H0	OVX does not cause GT		GT does not cause OVX	
GT	Test statistic	p-value	Test statistic	p-value
oil stock	0.14	0.71	1.8	0.18
oil price	0.016	0.9	0.73	0.39
oil production	0.3	0.58	1.3	0.26
oil reserves	0.24	0.63	0.14	0.71
news oil	0.17	0.68	1.0	0.31
world oil	0.39	0.53	1.2	0.28
brent oil	0.33	0.57	**3.4**	**0.066**
crude oil	0.0019	0.71	1.9	0.17
baker hughes	0.24	0.63	0.057	0.81
petroleum	0.15	0.7	1.2	0.28
opec	1.0	0.31	0.00095	0.98
wti	0.011	0.92	**3.3**	**0.071**
iea	0.077	0.78	0.2	0.89
kuwait	0.00051	0.98	1.3	0.25
libya	**4.8**	**0.028**	**2.7**	**0.098**
uae	0.81	0.37	0.031	0.86
angola	0.5	0.48	**4.6**	**0.032**
venezuela	0.025	0.88	1.8	0.18
saudi arabia	0.48	0.49	0.27	0.61
iran	1.2	0.28	1.1	0.28
iraq	0.096	0.76	1.2	0.28
nigeria	0.4	0.53	0.45	0.5
russia	**2.7**	**0.098**	0.16	0.68
china	**2.8**	**0.089**	0.51	0.47
usa	0.017	0.9	1.5	0.22
canada	0.59	0.44	0.85	0.36

Note: Relationships in which the Granger causality is present are in bold

Table A.4: Granger causality relationships between OVX and GT

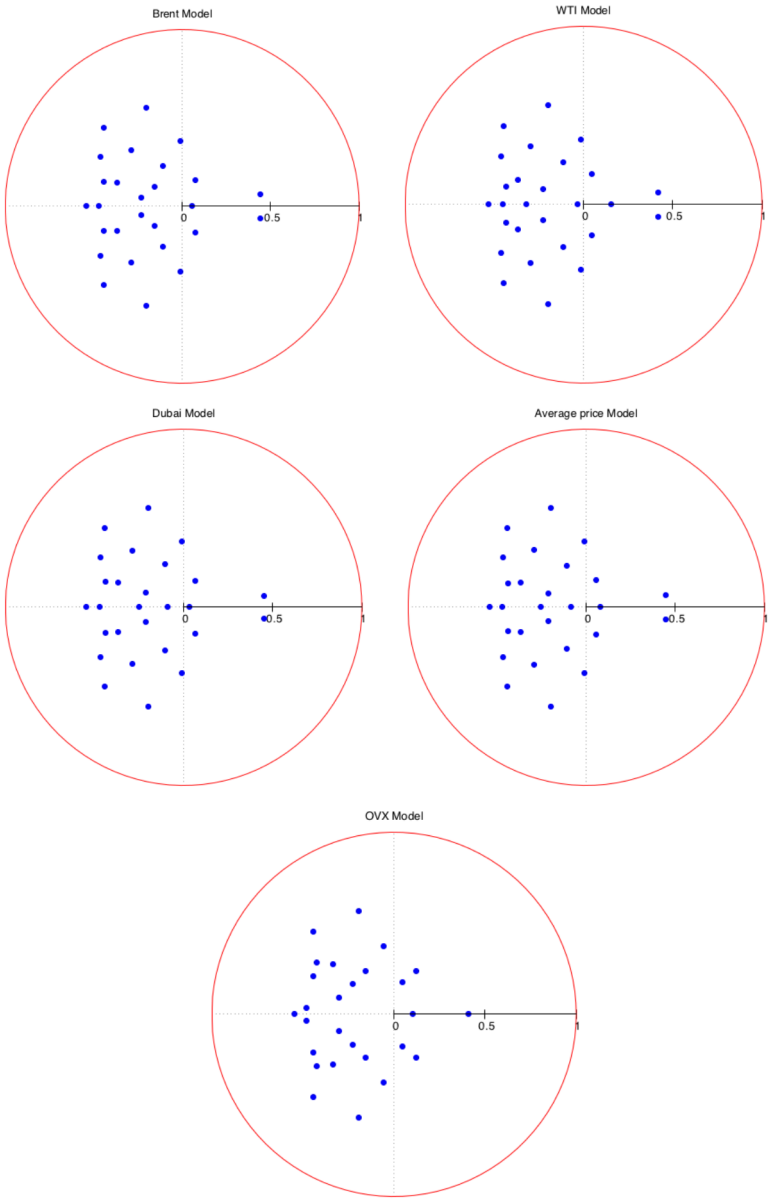

Figure A.1: Inverse roots of a characteristic polynomial

H0	Av.price does not cause GT		GT does not cause Av.price	
GT	Test statistic	p-value	Test statistic	p-value
oil stock	0.00014	0.99	**4.9**	**0.027**
oil price	0.36	0.55	**3.8**	**0.053**
oil production	0.046	0.83	0.68	0.41
oil reserves	**4.0**	**0.044**	**10.2**	**0.0014**
news oil	1.7	0.19	0.0038	0.95
world oil	**3.6**	**0.056**	0.56	0.45
brent oil	0.29	0.59	0.46	0.5
crude oil	0.00064	0.98	0.43	0.51
baker hughes	0.87	0.35	1.6	0.21
petroleum	0.029	0.87	0.92	0.34
opec	1.9	0.17	0.26	0.61
wti	0.041	0.84	2.3	0.13
iea	**4.0**	**0.04**	**3.3**	**0.071**
kuwait	0.57	0.45	1.1	0.29
libya	0.094	0.76	0.68	0.41
uae	**8.3**	**0.0041**	0.064	0.8
angola	0.034	0.85	2.1	0.15
venezuela	0.41	0.52	**10.9**	**0.00095**
saudi arabia	0.14	0.71	0.47	0.49
iran	0.69	0.41	**2.7**	**0.098**
iraq	0.14	0.71	0.12	0.73
nigeria	1.5	0.22	0.12	0.73
russia	**2.9**	**0.088**	**3.5**	**0.06**
china	0.13	0.72	**4.2**	**0.041**
usa	0.00064	0.98	0.0058	0.94
canada	1.7	0.2	**4.6**	**0.031**

Note: Relationships in which the Granger causality is present are in bold

Table A.5: Granger causality relationships between Average price and GT

H0	Brent does not cause GT		GT does not cause Brent	
GT	Test statistic	p-value	Test statistic	p-value
oil stock	0.0032	0.95	**5.8**	**0.016**
oil price	0.36	0.55	2.3	0.13
oil production	0.07	0.79	0.82	0.37
oil reserves	**4.6**	**0.032**	**10.4**	**0.0013**
news oil	1.7	0.19	0.035	0.85
world oil	**4.0**	**0.046**	0.27	0.6
brent oil	0.23	0.63	0.75	0.39
crude oil	0.00057	0.99	0.49	0.48
baker hughes	0.71	0.4	1.4	0.24
petroleum	0.077	0.78	0.95	0.33
opec	1.8	0.18	0.43	0.51
wti	0.0089	0.92	1.8	0.18
iea	**2.7**	**0.099**	**2.7**	**0.097**
kuwait	0.7	0.4	0.89	0.35
libya	0.024	0.88	0.74	0.39
uae	**8.4**	**0.0037**	0.041	0.84
angola	0.034	0.85	1.5	0.22
venezuela	0.17	0.68	**11.2**	**0.00082**
saudi arabia	0.18	0.67	0.35	0.55
iran	0.58	0.44	2.0	0.16
iraq	0.25	0.62	0.41	0.52
nigeria	1.3	0.25	0.11	0.74
russia	2.1	0.15	**3.8**	**0.052**
china	0.12	0.73	**3.8**	**0.051**
usa	0.011	0.92	0.015	0.9
canada	1.0	0.32	**4.7**	**0.031**

Note: Relationships in which the Granger causality is present are in bold

Table A.6: Granger causality relationships between Brent and GT

H0	Dubai does not cause GT		GT does not cause Dubai	
GT	Test statistic	p-value	Test statistic	p-value
oil stock	0.041	0.84	**4.4**	**0.036**
oil price	0.13	0.72	2.7	0.1
oil production	0.045	0.83	1.6	0.21
oil reserves	**3.9**	**0.048**	**9.3**	**0.0023**
news oil	1.8	0.18	0.14	0.71
world oil	**4.2**	**0.04**	0.47	0.49
brent oil	0.55	0.46	0.61	0.44
crude oil	0.029	0.86	0.36	0.55
baker hughes	0.74	0.39	1.2	0.28
petroleum	0.062	0.8	0.49	0.48
opec	1.5	0.22	0.93	0.34
wti	0.14	0.71	1.8	0.18
iea	**3.7**	**0.055**	**2.8**	**0.9**
kuwait	0.7	0.4	1.1	0.3
libya	0.067	0.8	0.34	0.56
uae	**8.5**	**0.0035**	0.00093	0.98
angola	0.074	0.79	1.3	0.25
venezuela	0.34	0.56	**11.6**	**0.00065**
saudi arabia	0.15	0.7	0.27	0.6
iran	0.6	0.44	2.5	0.11
iraq	0.29	0.59	0.16	0.69
nigeria	0.86	0.35	0.18	0.67
russia	2.4	0.12	**5.3**	**0.021**
china	0.11	0.74	**3.9**	**0.047**
usa	0.0048	0.94	0.1	0.75
canada	1.0	0.31	**5.1**	**0.024**

Note: Relationships in which the Granger causality is present are in bold

Table A.7: Granger causality relationships between Dubai and GT

H0	WTI does not cause GT		GT does not cause WTI	
GT	Test statistic	p-value	Test statistic	p-value
oil stock	0.01	0.92	**4.0**	**0.047**
oil price	0.63	0.43	**5.8**	**0.016**
oil production	0.021	0.88	0.07	0.79
oil reserves	**3.1**	**0.076**	**9.4**	**0.0021**
news oil	1.4	0.23	0.11	0.74
world oil	2.4	0.12	1.1	0.3
brent oil	0.13	0.72	0.092	0.76
crude oil	0.009	0.92	0.32	0.57
baker hughes	1.1	0.29	1.9	1.7
petroleum	0.0015	0.97	1.3	0.25
opec	2.2	0.13	0.0049	0.94
wti	0.016	0.9	**3.1**	**0.079**
iea	**5.8**	**0.016**	**3.9**	**0.048**
kuwait	0.37	0.7	1.3	0.25
libya	0.26	0.61	0.93	0.33
uae	**6.7**	**0.093**	0.31	0.58
angola	0.011	0.92	**3.2**	**0.075**
venezuela	0.83	0.36	**8.6**	**0.00035**
saudi arabia	0.071	0.79	0.86	0.35
iran	0.8	0.37	**2.7**	**0.1**
iraq	0.052	0.94	0.00012	0.99
nigeria	2.4	0.12	0.048	0.83
russia	**4.3**	**0.039**	1.5	0.22
china	0.16	0.69	**4.2**	**0.041**
usa	0.0015	0.97	0.09	0.76
canada	**3.4**	**0.066**	**3.5**	**0.061**

Note: Relationships in which the Granger causality is present are in bold

Table A.8: Granger causality relationships between WTI and GT

Appendix B

Outputs of regressions

	Coefficient	Std. Error	t-ratio	p-value
const	0.00166331	0.00830271	0.2003	0.8416
ld_iea	0.0713563	0.0568625	1.255	0.2121
ld_petroleum	0.226122	0.165658	1.365	0.1750
ld_bakerhughes	0.0178765	0.0451415	0.3960	0.6929
ld_brentoil	−0.000559578	0.0810326	−0.006906	0.9945
ld_wti	−0.0413508	0.103401	−0.3999	0.6900
ld_crudeoil	−0.145158	0.123422	−1.176	0.2420
ld_worldoil	0.0757586	0.0588865	1.287	0.2009
ld_opec	−0.0284253	0.0288588	−0.9850	0.3268
ld_newsoil	−0.0326247	0.0413568	−0.7889	0.4319
ld_oilproduction	−0.0677927	0.0954471	−0.7103	0.4790
ld_oilreserves	0.0588654	0.0542227	1.086	0.2800
ld_oilprice	0.148512	0.103041	1.441	0.1523
ld_oilstock	−0.132575	0.0807580	−1.642	0.1035
ld_kuwait	0.0150712	0.0307460	0.4902	0.6250
ld_libya	0.0205968	0.0210923	0.9765	0.3309
ld_uae	0.0164858	0.0517138	0.3188	0.7505
ld_angola	0.0263880	0.0208137	1.268	0.2075
ld_venezuela	−0.00169936	0.0195559	−0.08690	0.9309
ld_saudiarabia	−0.0366296	0.0239002	−1.533	0.1282
ld_iraq	−0.0129703	0.0149743	−0.8662	0.3882
ld_iran	0.0485647	0.0183079	2.653	0.0091
ld_nigeria	−0.0117634	0.0565260	−0.2081	0.8355
ld_russia	0.00666961	0.0369465	0.1805	0.8571
ld_china	0.0292629	0.0545007	0.5369	0.5924
ld_usa	−0.0480250	0.0177084	−2.712	0.0077
ld_canada	0.0261396	0.0236131	1.107	0.2707

Table B.1: OLS regression with dependent variable Average price

	Coefficient	Std. Error	t-ratio	p-value
const	0.00168560	0.00765658	0.2202	0.8262
ld_iea	0.0603550	0.0738298	0.8175	0.4154
ld_petroleum	0.218438	0.184920	1.181	0.2400
ld_bakerhughes	0.0104164	0.0627718	0.1659	0.8685
ld_brentoil	0.0157838	0.0668199	0.2362	0.8137
ld_wti	−0.0318231	0.0910877	−0.3494	0.7275
ld_crudeoil	−0.117296	0.163480	−0.7175	0.4746
ld_worldoil	0.0768983	0.0862772	0.8913	0.3747
ld_opec	−0.0266123	0.0310887	−0.8560	0.3938
ld_newsoil	−0.0294670	0.0626301	−0.4705	0.6389
ld_oilproduction	−0.0615876	0.112970	−0.5452	0.5867
ld_oilreserves	0.0599427	0.0550005	1.090	0.2781
ld_oilprice	0.0891766	0.123885	0.7198	0.4731
ld_oilstock	−0.127549	0.0811034	−1.573	0.1186
ld_kuwait	0.0107532	0.0497633	0.2161	0.8293
ld_libya	0.0243394	0.0234105	1.040	0.3007
ld_uae	0.0387594	0.0745866	0.5197	0.6043
ld_angola	0.0252850	0.0215213	1.175	0.2425
ld_venezuela	−0.00282874	0.0220327	−0.1284	0.8981
ld_saudiarabia	−0.0390014	0.0249695	−1.562	0.1211
ld_iraq	−0.0119383	0.0282304	−0.4229	0.6732
ld_iran	0.0454142	0.0249873	1.817	0.0718
ld_nigeria	−0.00657394	0.0587581	−0.1119	0.9111
ld_russia	0.00476278	0.0352079	0.1353	0.8926
ld_china	0.0223149	0.0727691	0.3067	0.7597
ld_usa	−0.0458621	0.0240005	−1.911	0.0586
ld_canada	0.0281611	0.0413060	0.6818	0.4968

Mean dependent var	−0.001282	S.D. dependent var	0.090488
Sum squared resid	0.887387	S.E. of regression	0.089012
R^2	0.214662	Adjusted R^2	0.032352
$F(26, 112)$	1.177455	P-value(F)	0.273850
Log-likelihood	154.0169	Akaike criterion	−254.0338
Schwarz criterion	−174.8030	Hannan–Quinn	−221.8365
$\hat{\rho}$	0.284381	Durbin–Watson	1.417204

Table B.2: OLS regression with dependent variable Brent

	Coefficient	Std. Error	t-ratio	p-value
const	0.00118119	0.00851460	0.1387	0.8899
ld_iea	0.0950382	0.0632117	1.503	0.1355
ld_petroleum	0.218427	0.164462	1.328	0.1868
ld_bakerhughes	0.0299149	0.0460877	0.6491	0.5176
ld_brentoil	−0.0330546	0.0834796	−0.3960	0.6929
ld_wti	−0.0598307	0.112133	−0.5336	0.5947
ld_crudeoil	−0.215232	0.124833	−1.724	0.0874
ld_worldoil	0.0652311	0.0572928	1.139	0.2573
ld_opec	−0.0273409	0.0300157	−0.9109	0.3643
ld_newsoil	−0.0519299	0.0401507	−1.293	0.1985
ld_oilproduction	−0.0560358	0.0973186	−0.5758	0.5659
ld_oilreserves	0.0557725	0.0554616	1.006	0.3168
ld_oilprice	0.253381	0.107768	2.351	0.0205
ld_oilstock	−0.100015	0.0875040	−1.143	0.2555
ld_kuwait	0.0180066	0.0340582	0.5287	0.5981
ld_libya	0.00760025	0.0201323	0.3775	0.7065
ld_uae	−0.0233702	0.0534826	−0.4370	0.6630
ld_angola	0.0316637	0.0214925	1.473	0.1435
ld_venezuela	0.00143238	0.0197325	0.07259	0.9423
ld_saudiarabia	−0.0283831	0.0223108	−1.272	0.2059
ld_iraq	−0.00925154	0.0160452	−0.5766	0.5654
ld_iran	0.0457225	0.0191112	2.392	0.0184
ld_nigeria	−0.0171760	0.0646887	−0.2655	0.7911
ld_russia	0.00620481	0.0368924	0.1682	0.8667
ld_china	0.0377513	0.0560196	0.6739	0.5018
ld_usa	−0.0456012	0.0180354	−2.528	0.0129
ld_canada	0.0284242	0.0236770	1.200	0.2325

Mean dependent var	−0.001867	S.D. dependent var	0.093531
Sum squared resid	0.899154	S.E. of regression	0.089600
R^2	0.255190	Adjusted R^2	0.082288
$F(26, 112)$	2.968598	P-value(F)	0.000039
Log-likelihood	153.1014	Akaike criterion	−252.2029
Schwarz criterion	−172.9721	Hannan–Quinn	−220.0056
$\hat{\rho}$	0.276625	Durbin–Watson	1.425637

Table B.3: OLS regression with dependent variable WTI

	Coefficient	Std. Error	t-ratio	p-value
const	0.00211171	0.00835044	0.2529	0.8008
ld_iea	0.0587510	0.0556960	1.055	0.2938
ld_petroleum	0.243446	0.172568	1.411	0.1611
ld_bakerhughes	0.0138582	0.0477663	0.2901	0.7723
ld_brentoil	0.0129086	0.0799607	0.1614	0.8720
ld_wti	−0.0326957	0.0984125	−0.3322	0.7403
ld_crudeoil	−0.104275	0.125306	−0.8322	0.4071
ld_worldoil	0.0846631	0.0636380	1.330	0.1861
ld_opec	−0.0304943	0.0297260	−1.026	0.3072
ld_newsoil	−0.0156515	0.0445400	−0.3514	0.7259
ld_oilproduction	−0.0877761	0.0958481	−0.9158	0.3617
ld_oilreserves	0.0619014	0.0538476	1.150	0.2528
ld_oilprice	0.107290	0.103375	1.038	0.3016
ld_oilstock	−0.172226	0.0805391	−2.138	0.0347
ld_kuwait	0.0171221	0.0303433	0.5643	0.5737
ld_libya	0.0291227	0.0220661	1.320	0.1896
ld_uae	0.0323025	0.0518452	0.6231	0.5345
ld_angola	0.0222517	0.0202480	1.099	0.2741
ld_venezuela	−0.00360837	0.0202653	−0.1781	0.8590
ld_saudiarabia	−0.0424716	0.0255301	−1.664	0.0990
ld_iraq	−0.0175359	0.0147848	−1.186	0.2381
ld_iran	0.0541245	0.0185361	2.920	0.0042
ld_nigeria	−0.0129209	0.0538485	−0.2399	0.8108
ld_russia	0.00845776	0.0370777	0.2281	0.8200
ld_china	0.0277423	0.0544470	0.5095	0.6114
ld_usa	−0.0535334	0.0180392	−2.968	0.0037
ld_canada	0.0225760	0.0262027	0.8616	0.3908

Mean dependent var	−0.000962	S.D. dependent var	0.092035
Sum squared resid	0.878183	S.E. of regression	0.088549
R^2	0.248730	Adjusted R^2	0.074328
$F(26, 112)$	2.704461	P-value(F)	0.000164
Log-likelihood	154.7416	Akaike criterion	−255.4832
Schwarz criterion	−176.2524	Hannan–Quinn	−223.2859
$\hat{\rho}$	0.297729	Durbin–Watson	1.395773

Table B.4: OLS regression with dependent variable Dubai

	Coefficient	Std. Error	t-ratio	p-value
const	−0.000227003	0.0111597	−0.02034	0.9838
ld_iea	−0.0222811	0.107609	−0.2071	0.8363
ld_petroleum	−0.309550	0.269526	−1.148	0.2532
ld_bakerhughes	0.156700	0.0914918	1.713	0.0895
ld_brentoil	0.0359481	0.0973921	0.3691	0.7127
ld_wti	0.0761373	0.132763	0.5735	0.5675
ld_crudeoil	0.300225	0.238277	1.260	0.2103
ld_worldoil	−0.0312529	0.125752	−0.2485	0.8042
ld_opec	0.0261028	0.0453127	0.5761	0.5657
ld_newsoil	0.0371208	0.0912853	0.4066	0.6850
ld_oilproduction	0.122578	0.164657	0.7444	0.4582
ld_oilreserves	−0.0872669	0.0801650	−1.089	0.2787
ld_oilprice	−0.0220924	0.180566	−0.1224	0.9028
ld_oilstock	0.0974225	0.118211	0.8241	0.4116
ld_kuwait	0.143807	0.0725315	1.983	0.0499
ld_libya	−0.0644821	0.0341215	−1.890	0.0614
ld_uae	0.127639	0.108712	1.174	0.2428
ld_angola	0.0141497	0.0313680	0.4511	0.6528
ld_venezuela	−0.0203282	0.0321133	−0.6330	0.5280
ld_saudiarabia	−0.0169275	0.0363939	−0.4651	0.6427
ld_iraq	8.95153e–005	0.0411466	0.002176	0.9983
ld_iran	−0.0339246	0.0364198	−0.9315	0.3536
ld_nigeria	−0.0393838	0.0856418	−0.4599	0.6465
ld_russia	0.132850	0.0513166	2.589	0.0109
ld_china	−0.157284	0.106063	−1.483	0.1409
ld_usa	0.0782601	0.0349815	2.237	0.0273
ld_canada	−0.0684540	0.0602048	−1.137	0.2580

Mean dependent var	0.004323	S.D. dependent var	0.156025
Sum squared resid	1.885162	S.E. of regression	0.129737
R^2	0.438845	Adjusted R^2	0.308577
$F(26, 112)$	3.368782	P-value(F)	4.46e–06
Log-likelihood	101.6495	Akaike criterion	−149.2990
Schwarz criterion	−70.06824	Hannan–Quinn	−117.1018
$\hat{\rho}$	−0.161348	Durbin–Watson	2.316438

Table B.5: OLS regression with dependent variable OVX