NLP for Analysis and Forecasting of Crude Oil Prices

Author:

Luigi Gifuni

Submitted in fulfilment of the requirements

for the Degree of Doctor of Philosophy in Economics

Adam Smith Business School, College of Social Science

University of Glasgow

October 2023

This page is intentionally left blank.

Abstract

The past fifteen years have seen an increasing number of unexpected economic and political changes that have interrupted a prolonged period of low market instability. This has challenged the predictability of macroeconomic leading indicators, due to delays in data being made available. As a result, economists have had a growing interest in considering present day written documents generated by media and government institutions, as a medium to develop new variables that are able to provide additional economic insights in real time.

This PhD thesis departs from this literature and contributes to enhancing the use of text as a valuable source of information for studying the behaviour of monthly crude oil prices. The work comprises three core chapters organised as follows.

In the first study I develop a set of text-based indexes capturing human sentiment and economic uncertainty in the oil market. The text analysis includes the titles and full articles of 138,797 oil related news items which featured in The Financial Times, Thompson-Reuters and The Independent. Empirical experiments show that sentiment indicators readily react to economic and geopolitical events affecting the price of oil, thereby enabling said indicators to accurately predict real oil prices. In contrast, measures of uncertainty hide structural weaknesses and thus yield unreliable oil price forecasts. This work results in a new text-based index that significantly improves the real oil price point forecasts, especially in periods of financial stress, when forecasting matters the most.

In the second essay I investigate the predictability of monthly real oil prices when daily and weekly text data are combined alongside the oil market fundamentals. Text data are retrieved from 140,096 full oil-related articles which featured in The Financial Times, Thomson Reuters and The Independent. I show that models containing high-frequency financial and commodity variables do not yield significant improvements

on the no-change forecast. In contrast, when text data are used along with commodity variables and oil market fundamentals, the preferred models reduce the MSPEs by 18%. However, despite this marginal improvement, gains are low. Indeed, the corresponding models with variables observed at homogeneous frequency, generate similar out-of-sample forecasts in terms of accuracy. I thus conclude that variables sampled at different frequencies do not significantly improve the predictability of monthly real oil prices. This is true for point and density forecasts.

In the final empirical chapter I highlight how oil studies typically assume the correct model specification and thus ignore the problem of estimating overly optimistic confidence sets. This implies that model uncertainty is pervasive in the empirical results. By relaxing this specification assumption, I revisit the role of (i) oil supply, (ii) aggregate demand and (iii) oil-specific demand shock, by proposing the Information Criterion model averaging as a strategy to address the problem of informational deficiency. In this analysis I consider a large macroeconomic panel, modelled with a structural vector autoregression model. The analysis is implemented with real and artificial data, and the non-orthogonalized impulse-response matrix shows that, in contrast to Kilian [2009], oil price response is less persistent after an aggregate demand shock, and more persistent following an oil specific demand shock.

Contents

List of Tables

List of Figures

This thesis is dedicated to my uncle Luigi Mele, 1964-2020. I miss him every day.

Acknowledgements

First and foremost, I owe my primary supervisor, Dimitris Korobilis, many thanks for his support and guidance since I took his class Advanced Bayesian Time Series Methods at the Barcelona GSE Summer School in July 2019. Dimitris has been a very patient mentor during my first steps in the academic world. He spent time to provide me with helpful advices in Bayesian econometrics and textual analysis. This work would have not been possible without Dimitris' continuous support and extremely valuable feedback.

I would also like to thank my second supervisor and head of the Department of Economics, John Tsoukalas, for being quick to provide any advice during my doctoral studies. Deep gratitude goes to Anthoulla Phella, who stepped up as my third supervisor in the last year of my PhD and, along with Zafer Kanik, provided any advice I required before going on the job market; thank you both. Many thanks also to Richard Dennis, Charles Nolan and Tatiana Kirsanova, who have been my discussants at the first, second and third annual progress review respectively.

I am very grateful to Marco Avarucci, Alberto Paloni and Oleg Kirsanov for the opportunity to GTA for them. More broadly, I would like to thank the admin staff of the Department of Economics for any support that I received. Many thanks to my office colleagues Daniel Mittendorf, Keegan Chisha and Yu Guo for their useful comments and stimulating discussions. I should also thank the University of Strathclyde, in which I was allowed to work as a Teaching Assistant during the last year of my studies. I could also not have written this thesis without the financial support of the University of Glasgow, which I gratefully acknowledge.

Finally, I have arrived at this point thanks to the support of my parents and my sister. They allowed me to write this work with passion and motivation. Thank you all, this is not an end but merely the beginning.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Luigi Gifuni

Introduction

1.1 Motivation

The oil industry plays a significant role in global economic growth. According to the International Energy Agency (IEA) recent report, oil meets 60 per cent of the world transportation demand and accounts for almost 3 per cent of global domestic product. What's more, the oil sector supports 11.9 million jobs worldwide, including 20 per cent of total energy employment. Both Covid-19 and the Russia-Ukraine war have increased the frequency of oil price spikes, potentially leading to long-term consequences for the global economy. It is not surprising, therefore, that policy makers have an increasing necessity to accurately forecast oil prices also in periods of economic instability. In this regard, it is worth mentioning that since 2010, due to a number of unexpected economic and political changes, the monthly price of oil has been extremely hard to forecast (see Baumeister et al. [2020]).

The reason behind such poor predictions is mainly twofold. Firstly, data commonly used to predict the monthly price of oil (i.e. global real economy, oil production and oil consumption at a world level) are made available by government authorities, usually with a delay. This implies that the econometric model used to forecast the price of oil is not informed in real time and the coefficient estimates cannot react promptly. Secondly, said indicators are by nature slow to respond to specific global events, such as political unrest and natural disasters. A potential solution to this problem can be found in a subfield of computer science, artificial intelligence and linguistics, known as natural language processing (NLP[1]).

NLP is a strategy that enables computers to extract structured data from written text, which can then be used for multiple purposes (e.g. sentiment analysis, speech recognition, topic modelling among many others).

[1] In this thesis I use the terms "text mining","natural language processing", "computational linguistics", and "text/textual analysis" interchangeably.

This technology field is providing outstanding insights to address policy questions in macroeconomics and finance that would have been hard to reach a few decades ago. Prominent contributions include Mosteller and Wallace [1963], Groseclose and Milyo [2005], Gentzkow and Shapiro [2010], Gentzkow et al. [2014], Baker et al. [2016], Hansen et al. [2018], Caldara and Iacoviello [2022], Shapiro et al. [2020]. These studies show that the information encoded in digital text is a valuable complement to the more structured data traditionally used in the economic research.

More specifically, in the context of macroeconomic forecasting and impulse response analysis, text mining provides timely news signals to low-frequency macroeconomic indicators (Larsen and Thorsrud [2019]). Such high-frequency information is useful when the economy evolves in extreme conditions, as traditional (hard) predictors can interact with many more variables that embed data in real time. This makes the text-cleaning, the feature selection and the assigning score to words/phrases key tasks that deserve extra attention if the analyst seeks to incorporate unbiased human knowledge into a purely data driven model. Some of these features have already shown to perform well when applied to a more general framework (see Baker et al. [2016]), but there are also other features that carry new information useful for understanding the behaviour of crude oil spot prices.

1.2 Thesis Contribution and Outline

Motivated by the premises outlined above and by departing from the literature on textual analysis, the goal of this thesis is to develop and apply new text based indicators to modelling and forecast the evolution of crude oil spot prices. This work is also linked to the literature on using machine learning tools in macroeconomics (Giannone et al. [2008], Fan et al. [2020]), and to the research aimed at discovering methodologies that perform well in periods of high economic instability (Chan [2022], Clark et al. [2022]). While these studies seek to modify off-the-shelf models in order to capture the time-varying volatility, I propose textual analysis as a strategy to develop new oil price drivers which incorporate information that was previously missed. Such contribution is provided in three core chapters.

Chapter 2 is titled "*Oil Price Forecasting: Gains and Weaknesses of Text Data*". In this essay I develop thirteen text-based indexes, nine of which are designed to capture the human sentiment and the remaining four aim to assess the economic uncertainty in the oil market. The textual analysis includes the titles and full body of 138,797 oil related news items that featured in The Financial Times, Thompson-Reuters and

The Independent from 1982M1 to 2021M11. The time series of such indicators demonstrates that human sentiment indexes promptly react to geoeconomics and geopolitical events affecting the price of oil. In contrast, uncertainty measures are not able to capture most of the events that have significantly affected the price of oil in the past decades. In the first experiment I show that including sentiment indexes into alternative vector autoregression (VAR) models yields more accurate out-of-sample forecasts in comparison to the corresponding non-text based model. Such results are 1%, 5% and 10% statistically significant as suggest by the Diebold and Mariano test. In the second experiment I develop a new text oil sentiment indicator (TOSI) by extracting the first principal component from a matrix that includes the best performing text indexes. I show that endogenizing TOSI in a Bayesian VAR with stochastic volatility improves the real oil price point forecasts in the short, middle and long run. Such good performance is observed in particular across periods of financial instability, when forecasting matters the most.

Chapter 3 is titled "*Do High Frequency Text Data Help Forecast Crude Oil Prices? MF-VAR vs. MIDAS*". This essay is motivated by the large body of literature assessing that models with variables sampled at different frequencies have the potential to release lower standard error estimates and more accurate predictions (Ghysels [2016]). In this regard, I investigate the predictability of monthly oil prices when daily/weekly text and financial data are combined along with the oil market fundamentals. The textual analysis builds on an updated version of the database used in Chapter 1. Specifically, text data are retrieved from 140,096 full oil-related articles featured in The Financial Times, Thomson Reuters and The Independent from 1982M1 to 2021M12. I show that models combining oil market fundamentals with both high-frequency financial and commodity variables do not yield significant improvements on the no-change forecast. In contrast, more accurate results in the short run are achieved when text data are used along with commodity variables and oil market fundamentals. In particular, including weekly observations of BERT and the Commodity Research Bureau index in a monthly-frequency VAR can reduce the minimum sum of prediction errors up to 18%. However, the effectiveness of using mixed-frequency models to forecast the monthly value of real oil prices is minimum. Indeed, the corresponding models with variables observed at the same frequency, generate similar out-of-sample forecasts in the short run, and perform significantly better in terms of accuracy in the middle and long run. I thus conclude that variables sampled at different frequencies do not improve the prediction accuracy of monthly real oil prices. This is true for point and density forecasts.

Chapter 4 is titled "*Oil Supply and Demand Shock under Model Uncertainty*". In this essay I address the problem of model uncertainty which is pervasive within empirical oil studies. In particular, I emphasise that previous research (Kilian [2009], Kilian and Murphy [2012], Alquist et al. [2013] and Baumeister and Hamilton [2019] are some examples) is built on interactions among a small number of oil market fundamentals, by implicitly assuming that the model is correctly specified. However, a body of research suggests that the oil prices are affected by a much larger number of variables (Jones and Kaul [1996], Carruth et al. [1998], Davis and Haltiwanger [2001]). This implies that in a framework of impulse response analysis, the existing literature proposes confidence sets that are too optimistic. I thus revisit the role of (i) oil supply, (ii) aggregate demand and (iii) oil-specific demand shock, by proposing the Information Criterion model averaging as a strategy to address the problem of informational deficiency. In this analysis I consider a large macroeconomic panel, modelled with a structural vector autoregression model. Unknown parameters are estimated through the random walk Metropolis-Hastings algorithm. The analysis is implemented with real and artificial data, and the non-orthogonalized impulse-response matrix shows that, in contrast to Kilian [2009], the oil price response is less persistent after an aggregate demand shock, and more persistent following an oil specific demand shock. Such results are robust across different VAR identification strategies.

2

Oil Price Forecasting: Gains and Weaknesses of Text Data

2.1 Introduction

Despite government efforts to reduce the consumption of heating oil in favour of green alternative solutions, oil in fuels is still the commodity in greatest demand due to its many uses. Fuel oil is commonly used for the propulsion of vehicles, the heating of buildings, the production of steam for industrial uses and also to generate electricity in power plants. Because of this crucial role, oil price structural shocks are still able to affect the inflation and the growth rate of a country. Therefore, predicting and controlling the volatility of oil prices is of the utmost importance for central banks as well as international organizations[1]. It is therefore imperative to make the correct choice of a model and its variables, in order to study the fluctuations of real oil prices.

For this purpose, a variety of models are suggested in the literature. Some economists have, for example, modelled the behaviour of oil prices through a random walk (RW) model, and have thus implicitly assumed that the future oil values could not be predicted on the basis of past history (Favero et al. [1994], Smith and McCardle [1998]). Many others, in contrast, have tried to contest such beliefs suggesting alternative econometric models that are able to forecast the price of oil better than a RW in the short and medium term (Knetsch [2007], Baumeister and Kilian [2012] Baumeister and Kilian [2014]). However, such empirical studies have two main limitations. Firstly, the econometric models have only been tested in periods of financial stability, whereas after 2010 oil prices have been particularly difficult to forecast (see Baumeister et al. [2020]). Secondly, the forecasting accuracy in the long run is still poor and falls far below the performance of a RW. What makes such results particularly discouraging is essentially (i) the inability of the model to anticipate structural economic shocks determined by oil price volatility, and (ii) the inability to rely on robust drivers of the oil market. Baumeister et al. [2020] provide a valuable contribution to the second problem by developing a global economic conditions (GECON) indicator, that improves the real oil price forecasts in both the short and long run. However, despite this good forecasting performance, Baumeister et al. [2020]'s methodology is still vulnerable to strong unexpected economic recessions (e.g. global financial crisis, the COVID pandemic) for two main reasons. Firstly, macroeconomic data running in Baumeister et al. [2020]'s dataset cannot inform the model in real time, as they are made available by government authorities, usually with a delay. Secondly, such variables are by nature slow to respond to

[1]Some companies that are particularly concerned about the future value of oil prices are: World Bank, Energy Information Administration (EIA), World Petroleum Council (WPC), International Energy Agency (IEA), World Energy Council (WEC), International Monetary Fund (IMF), Organization of Petroleum Exporting Companies (OPEC) and many others.

unexpected structural shocks[2].

In this work I propose a remedy for both of the aforementioned problems by developing a text-based oil-related human sentiment index, that relies on qualitative data retrieved from unbiased daily newspapers. In this way, information in real time can be used to inform a monthly-based vector autoregression (VAR) model, in order to anticipate periods of financial stress in the oil market.

Text analysis has become increasingly popular in economics and political science during recent years[3]. Written language is in fact the medium that central banks' executive board of governors periodically use to present a number of reports that account for multiple tasks, such as the country's financial position, economic results of the bank's maneuvers and monetary policy decisions. Additionally, there are myriads of newspaper articles and tweets that promote on a daily basis, the public understanding of the current economy and future trends. All these data are based on written words, and can be used to identify a number of economic parameters, such as the occurrence of financial market failures, investment and growth opportunities, economic priority challenges and so on. Following these premises, it is reasonable that in order to understand in which direction the economy is going, we also must be able to interpret what central bankers and economists are saying and writing.

Several studies in the past have already demonstrated the existence of a strong correlation between the content of economic newspapers and investor reactions (see for example Gentzkow and Shapiro [2010] and Engelberg and Parsons [2011]), but unfortunately the massive volume of data has always prevented the analyst from performing a meticulous study of written documents. Natural language processing (NLP) now offers many fast methodologies that, when used with modern operating systems, allow a computer to digest and elaborate strong volumes of text data. In this way, business companies and central institutions can analyse critical data in real time, prioritize urgent matters, and take quick actions.

[2]The authors use a 4-variable Bayesian vector autoregression with stochastic volatility originally developed by Carriero et al. [2016]. The dependent variables entering the econometric model are: (i) oil price, (ii) oil production, (iii) global real economy and (iv) oil inventories. Variables (ii), (iii) and (iv) react slowly to an oil price structural shock (see Baumeister and Hamilton [2019]).

[3]See for example, Acosta [2014], Hansen et al. [2018] and Nyman et al. [2021], where computational linguistics is applied in behavioural economics. Baker et al. [2016], Kalamara et al. [2020] and Mittermayer [2004] that compare the application of alternative text metrics to improve the out-of-sample forecast of macroeconomic and financial variables. Baumer et al. [2015], Johnson et al. [2017], Vafa et al. [2020] and Colladon [2020], where NLP is used to predict the outcome of presidential elections or classify and identify the political party an individual belongs to, based on Twitter and Facebook written posts. These are only few examples, NLP literature is growing up exponentially, and it is almost impossible to cite all works here individually. However, Bholat et al. [2015] and Xing et al. [2018] are good surveys that describe different NLP methodologies applied in economic studies.

In this work text data are retrieved from the titles and full bodies of 138,797 oil related daily news items, which featured in the Banking, Finance and Energy section of *The Financial Times* (FT), *Thomson Reuters* (TR) and *The Independent* (IND). The heuristic motivation behind the choice of using such newspapers lies on the least biased and highest factual reporting rating that Media Bias/Fact Check agency regularly gives regarding such data sources. Favouring one political party over others when reporting economic news, can indeed deviate the objective view of reality, and consequently the integrity of a text based indicator diminishes. Articles are selected based on the joint occurrence of the words "oil" and "price", and the analysis of written documents is performed as follows. Qualitative data are firstly retrieved from FT, since this is the longest data sample available in digital format on the LexisNexis database. Thirteen text indicators are developed and individually included in different econometric models, which are then used to forecast the price of oil. The analysis is then extended by including TR and IND articles, which are available for a shorter time horizon (Appendix A.1 provides additional details concerning the data sources). The text exercises previously investigated are then replicated for any possible combination between FT, TR and IND, in order to understand which combo generates the best performing text indicator relative to the forecast of each oil price measure[4]. After that, the first principal component is extracted from the best performing text indicators and I show that this text oil sentiment indicator (TOSI) not only improves the real oil price point forecast at any time horizon $h = 1, 3, 6, 12$ and 24, but in many cases empirical results are even 1%, 5% and 10% statistically significant based on the Diebold-Mariano test.

This paper contributes to enhancing the economic literature of oil price forecasts by providing a new text based indicator which captures the human sentiment in the oil market. I build upon the work of Baumeister et al. [2020] and show that VAR models with endogenous or exogenous text variables yield important and statistically significant forecasting gains. This work also provides some guidance for central banks, international organizations and oil companies by shedding light on the gains and structural weaknesses that different text metrics generate when economic information is retrieved from oil-related news items.

The gains are that human sentiment text based metrics depict time series indexes that are very responsive to historical geopolitical events (catastrophes, wars, terroristic attacks) that affect the price of oil (see Appendix A.2). Therefore, when such indicators are endogenised in a VAR model, statistically significant results are achieved in the short, medium and long term. Moreover, by plotting the difference between the cumulative sum of forecasting errors of a text based model and a RW, I also show that some sentiment

[4]Three main indicators are commonly used as a measure of crude oil prices: refiner acquisition cost (RAC) of oil imported in the US, West Texas Intermediate (WTI) index and Brent price. This study investigates all three of them.

based text metrics are able to anticipate economic downturns occurred in the oil market in the short run. In contrast, text indicators assessing economic uncertainty in the oil market hide structural weaknesses, such as incomprehensible peaks (see for example Figure A.10 in Appendix A.2), that prevent them from anticipating potential economic recessions, which is when forecasts matter the most.

This study is also related to many other works. Firstly, it contributes to the forecasting literature which aims to anticipate economic shocks (Townsend [1983], Evans and Lewis [1995], Cochrane and Piazzesi [2002], Christiano et al. [2014]), as I am including real time data in alternative econometric models in order to be able to predict unexpected periods of financial instability in the oil market. Then, it is linked to the rich literature on oil price analyses with the outstanding contributions of Kilian [2009], Baumeister and Hamilton [2019] and Baumeister et al. [2020]. Finally, this paper also seeks to exploit machine learning tools (e.g. NLP, BERT) to improve statistical models that have an economic fundamental (structural VAR, BVAR, SV-BVAR). Steps down this path have been taken by Hansen et al. [2018], Chernozhukov et al. [2018], Lamperti et al. [2018] and many others.

The outline of the paper is as follows. Section 2.2 reports the process by which qualitative data are transformed into numbers. Section 2.3 presents the methodologies used to construct sentiment and uncertainty text-based indicators. Section 2.4 offers several empirical applications where text variables are included in different VAR models, which are then used to forecast alternative oil price measures. Section 2.5 concludes.

2.2 From Text to Numbers

Cleaning text is imperative before analysing written documents. Articles in their original form have in fact a huge number of non informative punctuations such as quotation marks, apostrophes, commas and unnecessary white spaces, that affect the performance of any generic algorithm used for semantic analyses. Making text readable and unbiased mostly depends on the metric that the analyst is planning to work with. For example, the *unigram* count (thoroughly discussed in section 2.3.1) is only possible when words are spelled in lowercase. On the other hand, BERT (presented in section 2.3.4) requires only the document to be split into independent sections\sentences, as layers in BERT's network architecture are designed specifically to differentiate between uppercase and lowercase words, relative to the context in which they belong. However, it is beneficial to discuss some general steps that I commonly use in this work to transform text into numbers, regardless of the NLP procedure applied.

DOWNLOAD AND STRUCTURAL TRANSFORMATION. Oil-related articles are downloaded in `.rtf` format from the LexisNexis database. Items are selected from the Banking, Finance and Energy section of The Financial Times (FT), Thomson Reuters (TR) and The Independent (IND), based on the joint occurrence in each article of the words *oil* and *price*. The overall dataset is then manually checked in order to remove unrelated oil articles[5]. Similar to Kalamara et al. [2020], documents are also filtered by similarity, and observations are considered in real time, so that information leakage is avoided. After the download, documents are converted to `.txt` format by using a `bash` based macro.

PUNCTUATION AND STOP WORDS OMISSION. Single documents are removed of unnecessary non alphanumeric characters (i.e. more than two white spaces and redundant repetition of XML tags such as <>@”,’/&%[]#$ˆ_‘{}∼), and vocabulary words are then split by white spaces. For example, the sentence {`Member of the board’s capital committee, $250,000    per "consultant."` <|>}, becomes {`'Member'; 'of'; 'the'; 'board'; 's'; 'capital'; 'committee'; '250000'; 'per'; 'consultant';`}[6]. Stop words (e.g. the “s” featuring in the previous example) and high frequency text words (i.e. articles, conjunctions and prepositions) are then filtered out, as they do not convey meaningful information.

NORMALIZING, STEMMING AND LEMMATIZING. For unigram and geometrical text analysis, uppercase words are converted to their equivalent lowercase form, and the number of vocabularies is reduced by mapping similar terms to their single root. In this way, terms like `consultant`, `consulting`, `consultants`, `consulting`, `consultative` all became `consult`. This process is known as *stemming*, and contrary to *lemmatizing*, it does not seek to understand the context in order to convert a word to its base form, but it only removes (or stems) the suffix of a word. This means that `good`, `better` and `best` still remain unchanged after applying a stemming procedure, while a lemmatization would map all three of them to `good`, making a text variable slightly biased. For example, when oil prices turned negative for the first time in history, the Financial Times released the following two articles.

[5] *Oil* and *price* are in fact words that can also occur in articles discussing for example the “olive oil” market that, albeit correlated to petrol prices, do not affect the sentiment of oil investors directly. For this reason I also do a scanning-word analysis, in order to delete unnecessary articles and minimise the biasedness in the dataset.

[6] Text source: The Financial Times.

> **Why the oil price shock is nothing to celebrate? The collapse in prices will not deliver much of a boost to rich economies.** Lower revenues exacerbate the damage for poorer producer nations...Both countries appear to have badly miscalculated. Russia, opportunistically, saw the coronavirus as a chance to launch a broadside against American shale producers and the US economy. If demand was falling anyway, why not rely on others to make output cuts and hurt a geopolitical rival in the process? Saudi Arabia did not play ball. A shock-and-awe strategy of increasing its own production sent oil prices tumbling and wiped billions off the market value of western majors. The price of oil initially dropped at its fastest rate since the Gulf war in 1991; it went on to have its worst quarter on record. [*Financial Times, April 5 2020*]

> **A bailout of the oil industry is a fate worse than death.** Whenever things went wrong growing up, my dad would tell me, "What doesn't kill you only makes you stronger." Dying was supposed to be the worst-case scenario. He never mentioned zombies. But that's what a bailout for the oil industry would create: zombie companies that can't earn the cost of staying in business, kept afloat with taxpayer dollars......As debts rose along with oil supplies, a shake-out was inevitable. A bailout would throw good money after bad, propping up an industry desperately in need of productivity gains and consolidation. [*Financial Times, April 19 2020*]

If both pieces of text are lemmatised, words like *worse*, *worst* and *badly*, are all mapped to *bad* which is also used inside the text but for a different purpose. There are several stemming algorithms available in the computer science literature, but in the interest of brevity and simplicity, this work relies on Porter [1980]'s procedure. For VADER and BERT analysis words are lemmatised, as both methodologies aim to understand the context rather than the meaning of single words.

EXTRA CLEANING. Digital articles also present additional irrelevant written sections, such as copyrights, the author's name, the location where articles are written and many other notes that do not really influence the sentiment of a reader. Each newspaper places the "unnecessary" sections according to the company's style (they can be found at the beginning of the article, below the title, or at the end as a concluding remark). Therefore, supplementary newspapers-specific cleaning algorithms are run in order to ensure that the final article is produced solely with the title and full corpus. Removing futile sections is necessary to reduce structural bias further. Following the aforementioned steps, words are then ready to be transformed into numbers.

DATASET STRUCTURE. 138,797 articles are collected in monthly folders, sorted by year. Each folder includes a number of K articles, that change over time. Article k of a generic month m, resulting after applying the cleaning steps outlined above, is represented as a vector $N \times 1$ of unique words w_n, where N is the number of unique words running in document k. A collection of all vectors k of a generic month m generates a full (or sparsity) matrix X of dimension $N \times K$ expressing the occurrence of unique words used in month m. Text metrics presented in section 2.3 are all applied to the X matrix.

2.3 Text Mining Methodologies

This section presents the methodologies used for the semantic analysis of articles investigated in this paper. A deep description is given for each text mining methodology. The analysis starts with two basic statistical methods relying on the word count (*unigram* and *Boolean*). Thereafter I investigate some more advanced methodologies, five of which are dictionary based, two have a geometrical structure, and a final one has a neural network architecture. Some metrics are used to develop both a sentiment and an uncertainty indicator, whilst some others are used to develop only one of them. More details are given in each subsection. Overall, 14 different text based indicators are developed. The best performing ones are then used in section 2.4 to forecast the price of oil. However, the reader is encouraged to consult Appendix A.5 for the additional empirical results.

2.3.1 Statistical Models

The most straightforward text metric used in the semantic analysis of this study is the single word (or *unigram*) count probability model, which estimates the maximum likelihood of a specific word w_k by counting the number of times w occurs in document (article in our case) k. The result is then normalized by the number of words running in the document. Therefore, if X is a matrix of dimensions $N \times K$, with N representing the number of unique words running in a single document and K the number of documents analysed in month m, the unigram count probability of a generic word w_k is

$$p(w_k) = \frac{counts(w|k)}{\sum_{n=1}^{N} counts(w_n|k)}, \qquad \text{for } k = 1, \ldots, K. \tag{2.1}$$

The monthly score is the sum of K unigram probabilities, normalized by the number of monthly articles. Examples of works relying on the unigram probability model as a metric to analyse written documents are Antweiler and Frank [2004]; Schumaker and Chen [2009]; Wuthrich et al. [1998]; Zhai et al. [2007] and Schumaker et al. [2012]. Equation (2.1) is used to generate two time series variables. The first is based on the count of words "`economy`", "`economies`", "`economic`" and "`economics`", that are all stemmed to the single root "`econom`". Whereas the second relies on the count of words "`uncertain`", "`uncertainty`", "`uncertainties`" and "`uncertainly`", stemmed to "`uncert`". Following Kalamara et al. [2020], $p(\texttt{econom})$ is used to develop a human sentiment index, henceforth referred to as sentiment count in oil articles (**SentCO**), and $p(\texttt{uncert})$ is used to generate an index assessing the level of uncertainty in the oil market. The latter indicator is referred as uncertainty count in oil articles (**UnCO**). Figure A.1 and Figure A.2 display the behaviour of both time series variables. SentCO looks very responsive for some historical events such as the first Gulf War, OPEC's cut of 1.5 million barrels per day, Iraq War, the global financial crisis and COVID pandemic. The same applies for UnCO, which soars when Iraq increased missile attacks on Iran, on 9/11 terroristic attack, and when U.S. Marines invaded Iraq. However, for some events that affected the price of oil, both indexes do not provide a rational illustration. For example, 1.5 million barrels per day account for only 5% of global oil output, and generating a drop greater than COVID-19 is quite unlikely. Moreover, the uncertainty indicator is not responsive enough over the global financial crisis period. Such weakness is also reported by Baker et al. [2016], who show that the negative trend of their economic policy uncertainty index does not soar between 2007-2009 (with the exception of a short-period spike following the Lehman Brothers collapse). In any case, an unresponsive signal of uncertainty indicators across periods of financial stress is not surprising in the framework of crude oil markets. This is because uncertainty about oil prices is generally correlated to the fear of wars involving oil-exporting countries rather than financial market disruptions[7]. Wars in oil production countries can indeed cause the destruction of oil fields with subsequent shortfalls in crude oil production and oil price increases (Kilian [2009], Edelstein and Kilian [2009]). Therefore, the irrational response to geopolitical events can be considered as one of the main reasons why UnCO is not performing well when it is used to forecast the price of oil (see Appendix A.4).

A more complex statistical model investigated, which still relies on the word count, is the Boolean

[7]This peculiarity is also identified by other oil price uncertainty indexes developed through more structured data. See for example the conditional standard deviation indicator of Elder and Serletis [2010].

method used in Alexopoulos et al. [2009]; Pang et al. [2002]; Rachlin et al. [2007] and Baker et al. [2016]. In this case, I look for a sequence of two words that, unlike the *bigram*, need not be adjacent and can be placed anywhere within the document. Therefore, let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two sets of terms such that $\mathbf{S}_1$= {"economy", "economies", "economic", "economics"} and $\mathbf{S}_2$= {"'uncertain", "uncertainty", "uncertainties", "uncertainly"}. Given a set of unique words W defined in an N dimensional vector and running in article $k \in K$, a count is made if and only if

$$\left\{w_{n\in W|k} \in S_1\right\} \wedge \left\{w_{j\in W|k} \in S_2\right\}, \qquad \text{for any n } \neq j \in N \tag{2.2}$$

Namely, I make a count of one when at least one word inside S_1, jointly occurs with at least one word inside S_2. The sum of monthly counts is then normalized by the number of articles published in that month. Relation (2.2) is used to compute an index accounting for uncertainty in the oil market, henceforth referred to as **UnBool**. Figure A.3 depicts the time series behaviour of UnBool. In this case the index is much more responsive to negative events such as the first gulf war or Venezuelan protests, but unfortunately also for this metric the description of many historical events is not trustworthy. For example the COVID-19 shock is illustrated as half of the 9\11 attack's variation. However, in April 2020 oil prices turned negative for the first time in history, while after the terroristic attack the price of oil remained mostly unchanged. Appendix A.4 reports some empirical applications where UnBool is used to forecast the price of oil. The evidence demonstrates that this second uncertainty index is not a reliable predictor of oil prices.

2.3.2 Dictionary Models

Dictionary based text methods are designed to attribute a positive, negative or neutral value to a word, according to a predefined list of vocabularies with preallocated scores. It is a sort of key-value pairs methodology, where the word we look for in a dictionary is the key, and its definition is the value. There are several off-the-shelf dictionaries that can be used for text analysis. In this work I investigate five methodologies.

1. **FINANCIAL STABILITY** of Correa et al. [2017], with 391 words that can have a value of -1 (negative) or +1 (positive), and are calibrated to the language of financial stability reports used in central bank communications.

2. FINANCIAL LIABILITY of Loughran and McDonald [2011], with 4,150 words that can still have a value of -1 (negative) or +1 (positive), but they are selected from a more general financial context.

3. AFINN of Nielsen [2011], which is an improved version of ANEW dictionary trained for micro-blogs analysis (i.e. Identi.ca or Twitter), and accommodate 2,477 unique words. AFINN, like SentiStrength in Thelwall et al. [2010], assigns a number value to a text string between -5 (very negative) and +5 (very positive).

4. HARVARD-IV, which is a general-purpose dictionary developed by the Harvard University. It is used in Tetlock [2007], Tetlock et al. [2008] and Price et al. [2012], although for more specific economics frameworks it is sometimes not recommended (see Loughran and McDonald [2011]).

5. VALENCE AWARE DICTIONARY AND SENTIMENT REASONER (VADER), which rates words on a scale of "-4 Extremely Negative" and "$+4$ Extremely Positive" to evaluate sentiment in tokenized oil articles. This database has been developed by Hutto and Gilbert [2014] and, in contrast to previous dictionaries, words are not stemmed or in lower case. VADER in fact seeks to understand the context rather than merely classifying a single word.

Anyhow, the numerical score of a single article is based on a standard formula, regardless of the dictionary considered. Specifically, given the X matrix mentioned in section 2.3.1, if $w = (w_1, w_2, \dots, w_N)$ counts the occurrence of each word in document k (article in our case), and $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_N)$ is a set of weights of the corresponding words,

$$S_k = \frac{\sum_{n=1}^{N} \Theta_n w_n}{\sum_{n=1}^{N} w_n} \qquad \text{for } k = 1, \dots, K$$

describes the net score of document k, and

$$S_m = \frac{\sum_{k=1}^{K} S_k}{\text{Total No. Doc.}_m} \qquad \text{for } m = 1, \dots, M \tag{2.3}$$

is the monthly score in month m. Equation (2.3) is used across the overall data collection to develop five sentiment indicators reported in Figure A.4, A.5, A.6, A.7 and A.8. This time the proportion of positive and negative events is more realistic. In fact the COVID pandemic is correctly represented with the lowest value by each dictionary with the exception of Correa et al. [2017]. Indeed in April 2020 oil prices turned negative for the first time in history. Furthermore, the World Trade Center attack and Iraq invasion report a large drop in human sentiments as oil prices declined by 28% and 18% respectively[8]. Similarly, for the last two most negative events, namely the First Gulf War and the global financial crisis. The former generated a cut-off of Kuwaiti and Iraqi oil from the oil market, which increased the price of oil by 14.6 basis points in three months before a subsequent drop, following the decision of Saudi Arabia to increase oil production by more than 3 million barrels per day. The latter, in contrast, caused an oil price fall of 71% with respect to the previous year, since the negative aggregate shock yielded lower corporate earnings and more unemployment.

However, the evidence shows that VADER generates the best performing text indicator for oil price forecasts, among any other dictionary investigated. Therefore, the empirical application of VADER is reported in the main experiments of this paper, while the remaining results based on the other dictionaries are displayed in Appendix A.5. It is worth highlighting that this analysis could also be extrapolated to other individual dictionaries recently developed in the NLP literature (De Smedt and Daelemans [2012]), or else a combination of them (Shapiro et al. [2020]).

2.3.3 Geometrical Models

Still in the context of a "bag-of-words" representation[9], geometrical models are a more advanced information retrieval methodology. They have been frequently adopted in past (Salton and Buckley [1988], Joachims [1996], Joachims [1998]) as well as more recent papers (Fung et al. [2003], Mittermayer [2004], Groth and Muntermann [2011] and Hagenau et al. [2013]), and also look at the frequency of a specific word occurrence. However, contrary to unigram and Boolean, the relevance of a word does not increase proportionally with the term frequency. Such models exploit angle, distance, projection and vector space properties in order to derive a mathematical representation of a text document. They are very well-considered by many text analysts for their joint ability to be both straightforward to construct, and to generate accurate results.

The first geometrical text metric that I use is the term-document matrix, which incorporates information about the occurrence of a specific term in each document (newspaper article in my case) relative to the overall monthly data collection. In particular, given a sparse matrix X of dimension $N \times K$, the generic

[8]The former variation is not captured in the Harvard IV dictionary, most likely because this is a more general book of words, rather than an economic oriented one.

[9]Matrix X is constructed disregarding the word order in the articles.

binary vector k has dimensions $N \times 1$, where N is the total number of unique words running in a monthly data collection. This implies that cell $w_{n,k,m} = 1$ if the n^{th} word appears in article k, of month m, otherwise it is empty. This matrix, although simple to construct, has the powerful property to retrieve all words running in a specific document, as well as all documents containing the word I am looking for. The monthly score of a generic word w is then given by

$$TF_w = \frac{\sum_{k=1}^{K} w_{n,k}}{\text{Total No. Doc.}_{\cdot m}} \qquad \text{for } m = 1, \dots, M \tag{2.4}$$

where the numerator sums all binary cells in which the word w occurs, and the denominator reports the total number of articles running in month m. Equation (2.4) is used to generate a sentiment and an uncertainty index referred to as **SentOdx** and **UnOdx** respectively. SentOdx is developed by considering the set of words **S**={"economy", "economies", "economic", "economics"}, while UnOdx relies on **U**={"uncertain", "uncertainty", "uncertainties", "uncertainly"}. Figure A.9 and A.10 depict both time series indexes. It is clear that some historical events that have affected the price of oil are well illustrated by both series. For example, SentOdx goes up in August 1996 when the ARCO-led Petrolera Ameriven S.A. memorandum of understanding (MoU), set up the basis to produce 32,000 m3/d of crude oil in the Venezuelan region Hamaca, bringing a 17% surge in oil prices in fewer then two weeks. Or in June 2016, when SentOdx correctly turns negative, and UnOdx rises due to the Venezuelan protests. These occurrences prompted a collapse in Venezuelan oil prices, representing 95% of Venezuela's export revenue, and leading to severe shortages in food and other basic necessities. However, there are still some unexplained turnarounds in both series. The most evident is in May 2000 when unexpectedly UnOdx records a 13% increase in the level of uncertainty, even though the oil market is perfectly stable. I attempt to justify this deficiency by remembering that this vector-space model has a strong limitation. Namely, in a term-document matrix two different documents have the same vector-space representation if equivalent words (or set of vocabulary words) occur in both documents, regardless of the order. This problem is a direct consequence of the binary nature of the term-document matrix, since bag-of-words model representations are blind to distinguish between certain different documents. Anyhow, this weakness can be overcome by extending this methodology into a *weighted* vector-space model. By combining term frequency weights and inverse document frequency weights, it is possible to curb term-document matrix limitations and construct a new information retrieval strategy known as term-frequency inverse-document-frequency (TF-IDF) matrix.

TF-IDF is the second geometrical procedure used in the semantic analysis of this study. This methodology is based on a structural weighting scheme and exploits multidimensional vector spaces, where words constitute the orthogonal basis. There are two main intuitions behind the application of the weighting-scheme. The first is that not all words are equally important within the same document, and this is what the *term-frequency* scheme aims to model. The second is that the frequency distribution of a vocabulary word is not uniform across the data collection, and this issue is addressed by computing the *inverse-document-frequency*. In this way, the weighting scheme from one side focuses on those words that are common within a document, but rare across the data collection, and from the other it is also sensitive to those words that are (i) either rare within and across documents (very informative words), and (ii) common within and across documents (non informative words). To briefly illustrate TF-IDF, suppose we wish to start from a sparse $N \times K$ matrix, where N displays the total number of unique words running in K articles of month m. If the logarithm of word frequencies for each document is computed, as a natural consequence the term-frequency decreases as the word rank increases. Then, suppose the frequency of words at a row level is summed and the logarithm of the ratio between the value of K and the frequencies at a row level is computed. As a natural effect document-frequency exponentially increases as word rank increase, because the number of documents containing very frequent words is expected to be higher. Finally, by combining both effects we obtain a measure which is able to capture the occurrence of very informative words. In mathematical terms TF-IDF is expressed as

$$TF\text{-}IDF_w = \frac{log\left[1 + TF_w\right]}{log\left[1 + (K_m/n_m)\right]}, \tag{2.5}$$

where n_m is the number of documents in which the word w occurs, and K_m the overall data collection in month m[10]. As previously done with term-document matrix, I use TF-IDF to compute a sentiment and an uncertainty index, referred to as **SentOidf** and **UnOidf**, respectively depicted in Figure A.11 and A.12. While UnOidf still possesses the same weaknesses found in UnOdx (see Figure A.10), SentOidf perfectly matches the main historical events that affect the price of oil. In particular, there is strong evidence of a long negative drop caused by the global financial crisis as well as the First Gulf War. At this point, even though Venezuelan oil price reached the lowest value on summer 2016, the index is correctly negative for a larger period, as Venezuelan protests run from late 2015 through the end of 2016. Moreover, Iran-Iraq war and the dissolution of the Soviet Union are also correctly represented by negative human sentiments. The same

[10]For more details on geometrical models see Banchs [2012]

applies for COVID pandemic, where SentOidf reaches its minimum value since oil price turned negative in April 2020. It is worth noting, that this index is also very responsive to positive events such as ARCO-led Petrolera Ameriven S.A. MoU, the cease-fire decision in the Iran-Iraq war, and the first shipments of oil produced from Kazakhstan's Tengiz field that opened a new era for oil international exports. These reliable signals reflect the potential explanatory power when either SentOdx and SentOidf are used to predict future oil price values; especially in the short and medium term, as reported in section 2.4.

2.3.4 BERT

Bidirectional Encoder Representations from Transformers (BERT) is the last language representation model that I investigate in this study. The methodology is used to generate a human sentiment based index. BERT is based on deep neural networks, and has been recently presented by engineer researchers at Google AI Language (see Devlin et al. [2018]). Whilst initially this methodology was primarily designed to solve problems of language translation, it has been shown that BERT provides state-of-the art results in a wide variety of fields (e.g. question answering, sentiment analysis, text summarization and many other machine comprehension tasks[11]). The outstanding feature in BERT is the bidirectional encoder architecture that allows the machine to understand the context from both directions (left-to-right and right-to-left) simultaneously. This is in contrast to the bidirectional long short-term memory (LSTM) networks that first analyse left-to-right and right-to-left context separately, and then concatenate the two layers by slightly losing the true context (see Colón-Ruiz and Segura-Bedmar [2020] for a detailed comparison). The network architecture in BERT is fairly flexible. In my case I set a base network structured as follows.

- 2 input layers that accommodate a set of two sentences in order to investigate whether one sentence is correlated with the previous one,in a sort of binary classification problem. The maximum length for any input sentence is set to 512 words.
- 2 embedding layers with 23,041,535 parameters that transform words in tokens.
- 1 embedding position layer with 393,216 parameters that report where a single token was placed in the input sentence. As sentences are upper bounded to 512 words, a position can only take a value in between $[1, 512]$.

[11] See for example Yang et al. [2019], d'Hoffschmidt et al. [2020], Karpukhin et al. [2020], Liu and Lapata [2019], Singh et al. [2021], Pota et al. [2021] and Liu [2019].

- 12 bidirectional encoder layers with 85,648,129 parameters that understand the context of text data simultaneously.
- 1 fully connected softmax layer with 622,130 parameters, 30,000 of which are distinct tokens that break down single words to component words[12] by using a softmax activation function.

Overall, there are 109,705,010 parameters, but the structure can also accommodate a larger number of layers. Anyhow, the more parameters, the slower the algorithm. Because BERT works with a limited set of independent sentences, in order to make the resulting index unbiased, I firstly split each article k into n independent sentences, and then I run the first two input layers. Furthermore, in this case the monthly score is given by the arithmetic average between the results of each article belonging to the same month. Since this is a novel methodology, not many works in the economic literature have applied BERT to improve a variable forecast. However, there are already many private firms (such as Amazon, Microsoft, Google and Alibaba), where researchers are using BERT for multiple tasks (see for example Sun et al. [2019], Chang et al. [2019], Zhang et al. [2019] and Zhou et al. [2020]). This has motivated me to implement BERT in this study.

In this paper BERT is trained to learn human sentiment from oil news and the index that I develop is referred to as Bidirectional Text Representation of Crude Oil (BiTReCO). Figure A.13 plots the time series of BiTReCO from 1982M1 through 2021M11. It is evident that this index is very responsive to historical events that affect the price of oil. In particular, this is the first text based index that correctly depicts the strong decline of oil prices which started in late November 1997 and pushed Brent to a low of just $9.55 at the end of 1998. This crisis was jointly caused by a combination of the high level of oil inventories produced by OPEC countries from one side, and the stagnation of the Asian economy. Moreover, there is also evidence of an oil price plunge between 2014-2016, as caused by the exponential oversupply of US oil and a deteriorating demand in the mid 2015 by oil importing countries. These economic signals are well-fitted in BiTReCO. A reasonable explanation for this can be that BERT is trained to understand the context rather than counting the number of times a word occurs in a document. BiTReCO provides a powerful insight when the price of oil is forecast in the short run.

[12] For example the word `dropping` becomes (`drop`, `%%ing`), so if a sentence like "`oil price dropped about 6%`" occurs in future articles, the algorithm can infer the meaning of "`dropped`", as it splits the token into (`drop`,`%%ed`).

2.4 VAR Model Forecast

Since the seminal contribution of Kilian [2009], a wide variety of vector autoregression (VAR) models has been used in the economic literature to analyse and forecast the price of oil. Some prominent contributions are Baumeister and Kilian [2012], Lippi and Nobili [2012], Kilian and Murphy [2012], Inoue and Kilian [2013], Alquist et al. [2013], Baumeister and Kilian [2014], Baumeister et al. [2015], Baumeister and Kilian [2015], Hamilton [2019] and Baumeister et al. [2020]. This literature has consistently improved the original 3-variable VAR proposed in Kilian [2009]. In particular, it has been shown that the inclusion of additional oil market fundamentals in a VAR, such as oil inventories, improves the ability of the model to capture additional oil price volatilities that were previously missed (Baumeister and Hamilton [2019]). Moreover, Hamilton [2019] also shows that, in place of the dry cargo bulk freight rate, using the world industrial production index (originally developed in Baumeister and Hamilton [2019]) as a proxy of global real economy, is more efficient for predicting the price of oil. Finally, there is a consensus that real-time oil market fundamentals yield more accurate oil price forecasts (Baumeister and Kilian [2012], Garratt et al. [2019]).

Based on the aforementioned, for the purpose of this paper, I use text analysis to build upon the work of Baumeister et al. [2020] which uses Baumeister and Hamilton [2019]'s 4-variable VAR to forecast alternative measures of real oil prices. Such a model has been shown to produce very promising oil price forecasts. In this analysis, I am including in the benchmark model each text regressor developed in section 2.3, either as an exogenous, and as an endogenous factor. Alternative VAR models are investigated and the performance of each model is evaluated by comparing the minimum sum of prediction errors (MSPE) of the model itself to the ones generated by a random walk (RW) without drift[13]. MSPE values greater (less) than one indicate that VARs on average outperform (fall behind) a RW. The sample size runs from 1982M1 through 2021M11, unknown VAR parameters are estimated recursively, and real oil price is forecast for horizons $h = 1, 3, 6, 12$ and 24 months ahead. Namely, I first estimate a VAR from 1982M1 through 2000M12, then I forecast real oil price from 2001M1 through 2003M1. Thereafter, I re-estimate the VAR up to 2001M1, and forecast real oil price from 2001M2 through 2003M2 and so on. Crude oil predicted values are exponentiated, because real oil prices are entering the VAR in log-level. The empirical analysis

[13] This strategy has long been used among the forecasting literature (Meese and Rogoff [1983], MacDonald and Taylor [1994], Taylor [1995], Kilian and Taylor [2003] and Clarida et al. [2003]), especially when the variable is hard to forecast as it is the case of crude oil prices

proceeds as in Baumeister et al. [2020]. Namely, the comparison between *text* and *no-text* VAR is firstly investigated by using a frequentist approach, and then through alternative Bayesian methodologies. The evidence shows that, text regressors assessing human sentiment can not only improve real oil price forecasts, but in some cases it is also possible to anticipate periods of financial stress, particularly in the short run. Such promising results are achieved principally because qualitative data are now informing a VAR in real time, unlike previous models that rely exclusively on laggard macroeconomic data.

2.4.1 Frequentist Forecast

As remarked in the previous section, the benchmark model is Baumeister and Hamilton [2019]'s VAR, which can be expressed in the following reduced form:

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t, \tag{2.6}$$

where Y_t is a 4×1 vector of observed endogenous variables, and c is a 4×1 vector of intercepts; Φ_p is a 4×4 matrix of coefficients, with $p = 1, \ldots, 12$ indicating the number of lags[14], and ε_t is a white noise innovation vector. The out-of-sample forecasts resulting from equation (2.6) are then compared to the ones generated by the following two text-based VARs:

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + \gamma x_t + \varepsilon_t, \tag{2.7}$$

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t, \tag{2.8}$$

where, x_t in equation (2.7) is a 4×1 exogenous vector of text information, and γ is the associated 4×4 matrix of regression coefficient. Y_t in equation (2.8) is a 5×1 vector of dependent variables, in which x_t in equation (2.7) is now considered as an endogenous variable. Equation (2.6) can be seen as the unrestricted version of the structural VAR proposed by Kilian and Murphy [2014]. But in this case, I follow Baumeister et al. [2020] and I use alternative measures of real oil prices, and world industrial production in place of the dry cargo bulk freight rate.

[14] Baumeister and Kilian [2015] show that VAR with 12 autoregressive lags, generate the most accurate model estimation of real oil price.

Before discussing the empirical results, it is worth noting that VARs estimated through a frequentist approach generate accurate forecasts of real oil price until 2010, as oil price volatility was almost explained by crude oil commodity factors. Starting from 2011, oil price fluctuations are mostly explained by a crude oil financial attribute, which is directly affected by an additional number of financial factors. This trend variation makes oil price data difficult to forecast through traditional instruments (see Baumeister et al. [2020]). However, it is also vital to understand whether text variables can explain this trend variation, still by using a frequentist econometric approach. Table 2.1 and A.1 (appendix A.4) respectively report the out-of-sample forecast of text-based VARS, when text variables account for sentiment or uncertainty in the oil market. *Panel A*, *B* and *C* respectively display the forecasting results for WTI, RAC and Brent oil price, while *Model-1*, *2* and *3* stand for equation (2.6), (2.7) and (2.8) respectively. Any forecast that outperforms the benchmark's is displayed in red, whereas the lowest MSPE ratios of each specific time horizon h, relative to the oil price measure considered, are displayed in blue. The evidence shows that, consistently with Baumeister et al. [2020], the MSPEs are still greater than one at any time horizon considered. However, although a RW always outperforms both the benchmark and text models, some interesting results can be observed. First of all, *Model-3* with SentOdx and SentOidf endogenously included always outperforms *Model-1*. When VadOil is the text variable considered, *Model-3* outperforms *Model-1* for WTI and RAC oil price forecast, while for Brent price the performance is poor. The well performing TF-IDF and TF-MTX in the short and medium term is a recurrent evidence all over this empirical paper. Second, the marginal forecasting improvements for horizons $h = 1, 3, 6, 12, 24$, are 18%, 33%, 7%, 13% and 22% for WTI, 15%, 33%, 5%, 11% and 24% for RAC, 23%, 44%, 6%, 13% and 22% for Brent price. When text variables account for uncertainty in the oil market, forecasting gains are lower, but "blue" values still belong to VARs with text factors included (see appendix A.4 for additional details).

It is also interesting to compare the forecasting performance between *text* and *no-text* models at each point in time. For this purpose, I compute the difference between the cumulative sum of forecasting errors (CSFE) of equation (2.8) and (2.6). Values greater (less) than zero indicate that text based VARs outperform (fall behind) the benchmark. Results are depicted in Figure 2.1 and are fairly heterogeneous. In fact, from one side endogenising SentCO and SentOidf in a VAR is useful to anticipate periods of financial stress in the short run. From the other, VADER as well as BERT perform really poorly across the COVID period, but on average outperform the benchmark for almost any time horizon. A similar exercise is also performed in comparison to CSFE generated by a random walk. Figure A.19 in appendix A.5 reports such results, showing that at least with respect to short term predictions, equation (2.8) performs quite well, but unfortunately there are two drastic turnarounds at the time of the global financial crisis and the COVID pandemic. In contrast, for medium- and long-term forecasts, no VAR outperforms the RW, and this is clearly displayed in Table 2.1.

This exercise is also replicated when text variables account for uncertainty in the oil market. Results are displayed in Figure A.14 and clearly show the structural weaknesses that uncertainty text variables incorporate. Indeed, not only equation (2.8) is unable to predict unexpected periods of financial stress, but the amount of loss in terms of MSPE in comparison to the benchmark is significant, especially in the medium and long term. A similar outcome is visible in Figure A.20, where although text-based VARs outperform the random walk in the short run, the overall outcome in the medium and long term is very poor.

2.4.2 Bayesian Forecast

Estimating equation (2.8) implies uncovering $N(1+Np) = 305$ unknown VAR parameters, in a starting sample of 227 observations (1982M2-2000M12). This richly parameterized reduced form VAR generates high error values if estimated through a frequentist approach. As a direct consequence, beating the RW becomes even more complicated. A possible solution to this problem is to shrink the proliferation of unknown parameters in a Bayesian fashion. Many different methodologies are available in the literature. However, as remarked in section 2.4, one of the purposes of this work is to show whether the introduction of different text metrics in a VAR model improves the MSPEs produced in Baumeister et al. [2020]. Therefore, as in Baumeister et al. [2020], I also use the Bayesian shrinkage methodology suggested in Giannone et al. [2015]. Namely, I select the appropriate amount of shrinkage in each Bayesian VAR (BVAR), by choosing the hypervalues that maximise the marginal likelihood of data as a function of the unknown hyperparameters (additional details about this procedure are provided in appendix A.3). Anyhow, based on the choice of the prior, this experiment may also be implemented through alternative Bayesian methods (see the survey of Koop and Korobilis [2010] for a review). Empirical results are presented in Table 2.2, where again the out-of-sample forecasts of a BVAR(12) without text variables (*Model-1*) are compared to the out-of-sample forecasts of a BVAR(12) with exogenous (*Model-2*) and endogenous (*Model-3*) text variables expressing investors (or simply reader) sentiments. An equivalent exercise with text factors accounting for uncertainty in the oil market is given in Table A.2 (appendix A.4). The evidence shows that endogenising a text regressor in a stochastic system estimated in a Bayesian fashion, generates more convincing results in comparison to the previous experiment. In particular, *Model-3* with SentOdx or SentOidf as a text variable, not only yields the lowest MSPEs (see blue values), but for RAC short term forecasts the reduction is even 5% statistically significant based on the Diebold-Mariano test[15]. However, in the long run, recursive MSPE ratios are always greater than one, which highlights that overcoming a RW is still an onerous task, even when a system is estimated in a Bayesian fashion.

Figure 2.2 replicates the CSFE exercise presented in section 2.4.1, but this time the forecasting errors of equation (2.6) and (2.8) are generated from a text- and no-text-based BVAR(12). Results are very promising, as in the short run any text based model is now able to anticipate crucial periods of economic stress in the oil market. In fact, for $h = 1$ not only text models perform on average better than the benchmark, but there are also important marginal forecasting gains across the global financial crisis and COVID period. In particular, for the global crisis there are average improvements of 43%, 26% and 62%, while for the global pandemic they are 22%, 26% and 30% for WTI, RAC and Brent respectively. Less robust improvements are also achieved when text factors account for uncertainty. In fact, Figure A.15 shows that equation (2.8) does perform better than equation (2.6) even in periods of high volatility, as well as in the long term.

In regards to the comparison between text based BVARs and RWs, the evidence shows that for horizons $h = 1$ and 3, BVAR models on average outperform RW, but the performance across the global financial crisis and COVID period is poor. The unanticipated falls are thus not captured yet, but the downward jump is now lower in comparison to the frequentist experiment. For SentOidf variable, the financial crisis downturn is quite unobservable (see Figure A.21 and A.22 in appendix A.5). As highlighted in Table 2.2 and A.2, RW performs really well in the long run, and this is reflected in the CSFE differences, as they are constantly negative for $h = 12$ and 24, regardless of the sentiment or uncertainty measure used.

2.4.3 Bayesian Forecast with Stochastic Volatility

Crude oil prices have long been subject to drastic changes determined by external events (e.g. wars, 9\11 terroristic attack, financial crisis, COVID) and internal decisions (e.g. OPEC unilateral changes on the quantity of oil produced, and on the stock of oil inventories), that have affected the aggregate demand of industrialized economies. This has made the price of oil a highly volatile variable, which is hard to forecast especially from 2010 onwards. A possible solution to address the problem of structural changes in the volatility of a variable, is to define a parsimonious law of motion for the error term structure in order to estimate the dynamic path (see, for example, Primiceri [2005], Cogley and Sargent [2005],

[15]This is consistent with Kalamara et al. [2020]'s evidence, where sentiment *TF-IDF* also produce good results.

Sims and Zha [2006] and Mumtaz and Zanetti [2013]). Following Baumeister et al. [2020], in this section I investigate the exogenous and endogenous inclusion of text variables developed in section 2.3, inside the stochastic volatility BVAR (SV-BVAR) proposed in Carriero et al. [2016]. The benchmark model is

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \Sigma_t) \tag{2.9}$$

where Y_t still denotes the vector of N dependent variables, while Σ_t is a full covariance matrix with elements changing over time. Without loss of generality, we can write

$$\Sigma_t = \left(A^{-1}\right) \Lambda_t \left(A^{-1}\right)' \qquad \text{and} \qquad \Sigma_t^{1/2} = \left(A^{-1}\right) \Lambda_t^{1/2}$$

with Λ_t being a diagonal matrix and $\lambda_{i,t}$ a generic diagonal element, for $i = 1, \ldots, N$ representing the i^{th} dependent variable. A^{-1} is a lower diagonal matrix with ones on the main diagonal. The system in equation (2.9) can be then rewritten as

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + A^{-1} \Lambda_t^{1/2} \varepsilon_t, \qquad \varepsilon_t \sim iid\, N(0, I_N) \tag{2.10}$$

In the benchmark case, $N = 4$ and the structure of Λ_t is

$$\Lambda_t = \begin{pmatrix} \exp(\lambda_{1,t}) & 0 & \cdots & 0 \\ 0 & \exp(\lambda_{2,t}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp(\lambda_{4,t}) \end{pmatrix} \tag{2.11}$$

where $\lambda_{1,t}, \lambda_{2,t}, \ldots, \lambda_{4,t}$ refer to the log volatility of the $N = 4$ structural shocks in equation (2.9), and the law of motion for the stochastic volatility is given by

$$\tilde{\lambda}_t = \tilde{\lambda}_{t-1} + v_t, \qquad \text{with} \quad v_t \sim N(0, Q) \quad \text{and} \quad \mathbb{E}(\varepsilon_t, v_t) = 0$$

where $\tilde{\lambda}_t = [\lambda_{1,t}, \lambda_{2,t}, \ldots, \lambda_{4,t}]$, and by considering ε_t as a diagonal matrix with ones on the main diagonal, we implicitly suppose that the disturbance terms are uncorrelated across the dependent variables. In this way the system can be estimated equation-by-equation.

As a first exercise, consistently with the previous experiments, I compare the benchmark forecasts of eq (2.9) to the ones generated from a text based SV-BAR, with exogenous (*Model-2*) and endogenous (*Model-3*) text variables. Results are reported in Table 2.3. For almost any time horizon h, endogenizing a text factor in a SV-BVAR generates more accurate results. In particular, SentOidf forecasts are 1% statistically significant for WTI, and 1% for RAC in the short run. Moreover, SentOdx forecasts in the medium and long term ($h = 6$ and 12) are also 10% statistically significant. Based on the relevant gains achieved in this exercise, it is also important to compare the forecasting performance at each point in time. The difference between the CSFE of a text and no-text SV-BVAR is displayed in Figure 2.3. The results show that the good performance of some text metrics is consistent across the short, medium and long term (see for example VadOil, SentOdx and BiTReCO). There is still evidence that text models perform better in periods of high volatility, especially at the beginning of the COVID pandemic. Similar conclusions cannot be claimed when text regressors account for uncertainty in the oil market. In fact according to Table A.3 the benchmark model sometimes outperforms the text based models. Such results are also visible in Figure A.16, where the CSFE is negative for almost any time horizon h. This experiment emphasizes, once again, the structural limitations of text measures accounting for uncertainty in the oil market.

In any case, it is worth pointing out that SentOidf can efficiently anticipate RAC future values before the global financial crisis and the COVID pandemic. And can do so even better than a RW. Although for WIT and Brent there are still few negative downturns, on average the forecasting errors generated by a text model are far lower than a RW's (see Figure A.23 in appendix A.5). Similar conclusions can be drawn when text measures account for uncertainty, despite the lower marginal gains in comparison to sentiment variables (see Figure A.24 in appendix A.5).

2.4.4 Are More Text Data Better?

In many recent empirical studies, as common practice text experiments are performed by relying on only one source of text data (Diederich et al. [2003], Ming et al. [2014], Bybee et al. [2021]). This choice implies that the analyst is implicitly assuming that the readers\investors sentiment can only be affected by the content of a specific newspaper, thereby ignoring the possibility that people can read alternative newspapers, which also provide meaningful information able to affect the human sentiment. In this way, not only is uncertainty pervasive within empirical results, but the evidence is also biased by the analyst's subjective choice regarding the newspaper selected for text analysis. This problem is addressed in this section, as I consider two additional sources of text data: The Independent and Thomson Reuters. For each newspaper, I firstly generate the same text indicators developed with information retrieved from The Financial Times, and then I investigate any possible combination of text data, relative to each oil price measure forecast.

Such additional data sources are now including types of information that are not generally found in The Financial Times, such as oil company's CEO statements, information for shareholders and future plans of energy sector leading firms, like Laredo Petroleum, KBR, Transocean, Nordic American Tankers, Seadrill, Statoil, Marathon Petroleum and many others. Although it would have been interesting to compare the individual outcomes generated by each text source at each point in time, the different availability of digital archives[16] prevented me from being able to investigate this comparison. Therefore, this experiment proceeds as follows. Human sentiment data are firstly extracted from The Financial Times newspapers from 1982M01 to 1988M08, as they are the only items available across this time span. Then, from 1988M09 through 2002M10, the analysis also includes daily oil related articles featured in The Independent. Finally, from 2002M11 up to 2021M11 also Thomson Reuters' articles are incorporated. It is worth emphasizing that I do not compute the sentiment (or uncertainty) score for each newspaper individually, and then average the three different metrics. Articles are instead selected as they belonged to the same folder.

When transforming words from text into numbers, articles are often cleaned in a "personalised" way. For example, articles featured in Thomson Reuters have been manually checked, because hundreds of documents reported a section with sensitive information about the company (e.g. headquarter position, telephone numbers and so on), which do not tend to affect the sentiment of a reader, and consequently no relevant information can be retrieved from this section. On the contrary, keeping such information in the dataset can only generate bias in the empirical results.

Table 2.4, reports the forecasting outcome when news items of multiple sources are jointly combined and the resulting index is endogenised in a SV-BVAR. The evidence shows that additional text data not only generate marginal gains in comparison to the equivalent methodology that only relies on one text source, but for the short and long run forecasts such improvements, in some cases are even statistically significant based on the Diebold-Mariano test. For example, the use of multiple text sources drastically improves the one-month ahead forecast of Brent oil, which is now 10% statistically significant. Then, although the

[16]The Independent articles are available from 1988M9, whereas Thomson Reuters from 2002M11.

short-run forecasts for WTI and RAC are statistically equivalent to the results reported in Table 2.3, there is however a 1% and 3% of marginal gain. Finally, for medium- and long-term forecasts there is evidence of a 10% statistical significance (see e.g. WTI and RAC in Table 2.4 panel C), and overall a marginal improvement between 1%- 6%. Also the CSFEs associated with multiple sources, on average outperforms the correspondent one source based CSFEs in the short, medium and long run, other than showing a better performance in periods of financial stress too (see Figure A.27 appendix A.5).

This remarks the importance of including additional - and politically unbiased - text information in the empirical experiment, as a strategy to incorporate the human sentiment of people reading alternative newspapers.

2.4.5 A New Text Based Sentiment Indicator

So far empirical evidence has shown that on average sentiment indicators based on economic information retrieved from daily newspapers can drastically improve the point forecasts of alternative oil price measures, regardless of the econometric model performed. On the other hand text based uncertainty indicators hide structural weaknesses and as a result inaccurate oil price forecasts are generated. However, the outcomes resulting from the previous exercises can only assess that some sentiment metrics are more reliable than others, although no final solution has been reached yet. When analysts and policy makers forecast the price of oil, in place of running multiple sentiment indicators, and compare the related results, they might want to use one econometric model that incorporates the benefits of any best performing text indicators previously investigated. For this reason, in this section I extract the first principal component from the time series sentiment measures that have produced the lowest MSPEs in the last experiment (i.e. VadOil, SentOdx and BiTReCO extracted from multiple text sources[17]). The resulting index is referred to as text oil sentiment indicator (TOSI[18]), and the related series is displayed in figure 2.5. It is clear that, the normalized index tracks pretty well the main geopolitical events that affected the price of oil over the period 1982M1-2021M11. In fact, COVID-19 and global financial crisis correctly identify the main economic drops that have negatively affected the economy at a world level and consequently the price of oil as well. Then, in

[17] According to Table 2.4 panel D, even though TF-IDF yields the lowest MSPEs for WTI short-run forecast, SentOidf is however excluded from the non parameterized combination. The reasons behind this choice can be explained by highlighting the evidence that not only TF-IDF used for WTI short-run forecast does not generate statistically significant improvements, but the remaining point forecasts are even worse compared to the BERT based ones. Therefore, including SentOidf in the PCA analysis would decrease the performance the new index.

[18] Text variables are standardized before conducting the principal component analysis. Overall TOSI explains 86% variation of the original dataset.

the interval $[0,-2]$ we can also observe other important negative shocks, such as the Gulf War, the oil price crisis of 1998 where Brent price fell below a level of £18 per barrel, the Twin Towers attack and the vibrant Venezuelan protests of 2016 demanding president Maduro resign. On the other hand, TOSI also tracks the positive events that have stabilized the oil price level of both WTI and Brent following an increase in oil production.

The choice of using the first principal component is also supported by the goodness-of-fit that TOSI provides as opposed to the remaining common factors. Specifically, after running the principal component analysis, I follow McCracken and Ng [2016] by regressing the i_{th} best performing text series on each common component k. This yields $R_i(k)^2$ for text series i. The average importance of factor-k is then given by $\frac{1}{N}\sum_{i=1}^{N} R_i(k)^2$. Results are reported in Table 2.7. Not only TOSI provides more information across each text series, but it also yields the highest average importance.

In order to assess the forecasting performance of TOSI in predicting the price of oil and to anticipate periods of high volatility, I run a SV-BVAR as in equation (2.9) and use TOSI as an endogenous variable. For the sake of robustness, results are now compared to alternative proxy SV-BVARs. Namely, in the previous experiments I have implicitly supposed that uncertainty and human sentiment indicators can directly affect a commodity price. In fact, text variables were entering the benchmark model of equation (2.6) and (2.10) as exogenous or endogenous variables. Now, I investigate the possibility to consider a text indicator as a proxy of the stochastic volatility. Hence, suppose I wish to apply the Kalman filter to estimate the law of motion of each λ_t in (2.11), I then consider the stochastic volatility vector related to oil prices $\tilde{\lambda}_t^p$ and run the following regression

$$\tilde{\lambda}_t^p = \beta_0 + \beta_1 Sent_t + u_t \tag{2.12}$$

where $Sent_t$ is a text indicator that in the previous experiments was entering the main VAR expression. The resulting $\hat{\tilde{\lambda}}_t$ is then used in place of $\tilde{\lambda}_t^p$ in the $N \times 4$ matrix of volatility states. Such comparison is reported in Table 2.6, and the evidence shows that although the proxy SV-BVAR does perform slightly better than a TOSI based SV-BVAR in WTI one-step-ahead forecast, such gains are not statistically significant. Other than that, in any other step-ahead forecast, TOSI based SV-BVARs outperform any alternative model, and in several cases the results are also 1% and 10% statistically significant for the short, medium and long run[19] (see RAC panel). Moreover, by looking at the cumulative sum of prediction errors of a TOSI based

[19] Figure 2.5 displays the overall first principal component series. However, in this exercise TOSI is generated recursively. This means that for each iteration a first principal component is extracted and endogenously included in the stochastic model. This is important as text factors can only incorporate information up to the time in which the forecast is performed and no future signal

SV-BAVR, I can assess that (i) the use of this text index generates on average better results in comparison to a random walk (line 1, figure 2.4) and to the benchmark model of equation (2.10) (line 2, figure 2.4), especially in the short and long run. (ii) In periods of high volatility (global financial crisis and COVID-19 among all), text variables do improve point forecast accuracy drastically.

Before concluding with a summary of the empirical results, to improve the validity of the evidence, I also run 4 simplified versions of a SV-BVAR by comparing the case of *TOSI* and *no-TOSI*. Specifically, I start from a univariate autoregression model where the price of oil is the unique dependent variable and then I compare the forecasting outcomes of this model to a two-variable SV-BVAR's, where TOSI is considered as a second endogenous variable. After that, I run a two-variable SV-BVAR with oil price and global real economy as dependent variables, and compare the forecasting outcomes to a three-variable SV-BVAR's with TOSI endogenised, and so on. The purpose of this exercise is to show that the forecasting gains achieved by endogenising TOSI are robust to different model specifications. Results are displayed in figure 2.6, and effectively show that TOSI does improve the performance accuracy of any SV-BVAR used to forecast the price of oil 1-, 3- and 6-months ahead. In particular, other than the outcome produced by a 3-variable SV-BVAR for RAC 1-month ahead forecast, TOSI generates better results for any time horizon h considered, and such outcome is often even statistically significant according to the Diebold-Mariano test. Therefore, empirical results achieved so far can definitely motivate the choice of using TOSI to improve actual econometric models used to forecast real oil prices.

Summary of Empirical Results. In short, according to the evidence provided in Table 2.1, 2.2, 2.3, 2.4 and 2.6, and with the support of Figure 2.1, 2.2, 2.3 and 2.4, the following can be inferred. 5 variable SV-BVARs with endogenous text variables outperform any other model investigated in this study, and it is helpful to combine multiple text sources in order to minimize the bias in general (as people read alternative newspapers), and in doing so drastically improve the oil forecasts in particular. For $h = 1$, the Financial Times and The Independent incorporate more useful information to predict WTI, and TF-IDF is used as a text metric. Whereas, for RAC and Brent forecasts, BERT is the text metric used to retrieve relevant information from the Financial Times and Thomson Reuters. The latter strategy is also suggested for a 3-months ahead forecast, and just for WTI and Brent when oil price is forecast 6-months ahead. For long term forecasts ($h = 12$), the evidence suggests that incorporating any text source is better off, TF-MTX is useful for WTI and RAC, while BERT for Brent. Lastly, when oil prices are predicted two-years ahead, a

is included.

dictionary based model is still the best performing text metric and the newspaper combination suggested is the Financial Times and Thomson Reuters. Table 2.5 summarizes the aforementioned results.

Many other experiments have been performed in this paper. Only a few of them are mentioned here, but the reader is addressed to consult appendix A.5 for additional results. Firstly, I have used the global economic conditions indicator (GECON) developed in Baumeister et al. [2020], in place of world industrial production in a SV-BVAR(12), and have investigated the forecasting performance with both sentiment and uncertainty text based indicators (see Table A.4 and A.5 in appendix A.5). I have also used a measure of world oil consumption in place of oil produced at a global level with sentiment and uncertainty text variables. Finally, I have replicated the same exercise by considering the global steel production factor as a measure of real economy. Such results are provided in appendix A.5 and still show that text variables can drastically improve a no text based VAR.

2.5 Conclusions

This work develops a set of text based time series variables assessing (i) human sentiment and (ii) economic uncertainty in the oil market. The empirical application of such measures in multiple vector autoregression models shows the gains and weaknesses of the text metrics used to retrieve meaningful information from oil related news items. The weaknesses are mostly associated with uncertainty measures, as their empirical application does not generate statistically significant improvements when alternative oil price measures are forecast, and the related time series are not prone to react to the main geopolitical events affecting the price of oil. In contrast, sentiment indicators can track such episodes reasonably well, and consequently the econometric models relying on such indicators can significantly enhance the forecasting accuracy of real oil prices.

The empirical evidence also shows that by endogenising VadOil, SentOdx and BiTReCO in a 5-variable SV-BVAR, and by taking into consideration multiple text sources, the forecasting performance of a SV-BVAR model not only outperforms the random walk's in the short, medium and long run, but the MSPEs associated with such text-based SV-BVARs are even significantly lower than any other benchmark model considered in this study. The first principal component is extracted from the best performing text indicators, and the work results in a new text-based index that significantly improves the real oil price point forecasts, especially in periods of financial instability. A number of improvements could be made to this research. For example, a considered, proportionate, incorporation of a wider range of newspaper types (red top tabloids versus left-wing publications) would allow the collection and analysis of more data with the potential to enhance TOSI's performance. This task is to be left for future research.

Thus, in conclusion, this work provides a state-of-the art guidance for energy companies, central banks and international organizations that need accurate forecasts of real oil prices to make informative policy and strategic decisions, in order to anticipate potential periods of economic downturn.

Table 2.1: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil price in a VAR(12), estimated through a frequentist approach. World industrial production is used as a measure of global real economy, and text regressors account for human sentiment about oil news.

Monthly horizon		**SentCO**		**VadOil**		**SentOdx**		**SentOidf**		**BiTReCO**	
	Model-1	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3
					A. WTI based VAR						
1	1.404	1.380	1.449	1.455	1.365	1.411	1.310	1.336	1.246	1.478	1.524
3	1.407	1.688	1.448	1.557	1.429	1.581	1.201	1.347	1.156	1.772	2.202
6	1.108	1.437	1.134	1.157	1.097	1.681	1.050	1.374	1.091	1.137	1.074
12	1.329	1.599	1.322	1.532	1.322	3.618	1.227	1.899	1.264	1.432	1.377
24	1.319	1.896	1.160	1.401	1.202	6.612	1.277	2.720	1.077	1.695	1.373
					B. RAC based VAR						
1	1.183	1.169	1.232	1.238	1.158	1.181	1.100	1.141	1.084	1.255	1.297
3	1.391	1.578	1.454	1.616	1.345	1.563	1.187	1.289	1.140	1.779	1.983
6	1.126	1.297	1.148	1.243	1.105	1.768	1.083	1.272	1.112	1.152	1.096
12	1.406	1.473	1.411	1.713	1.390	4.451	1.318	1.816	1.344	1.537	1.464
24	1.302	1.576	1.142	1.466	1.180	7.639	1.271	2.226	1.029	1.722	1.347
					C. Brent based VAR						
1	1.637	1.606	1.714	1.682	1.626	1.641	1.452	1.563	1.451	1.756	1.861
3	1.855	2.118	1.850	2.018	1.865	2.119	1.446	1.679	1.445	2.642	2.973
6	1.202	1.554	1.227	1.235	1.202	1.768	1.160	1.350	1.178	1.278	1.187
12	1.378	1.591	1.368	1.528	1.359	3.615	1.291	1.750	1.293	1.620	1.427
24	1.277	1.677	1.104	1.278	1.153	5.577	1.256	2.109	1.020	1.910	1.329

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*.

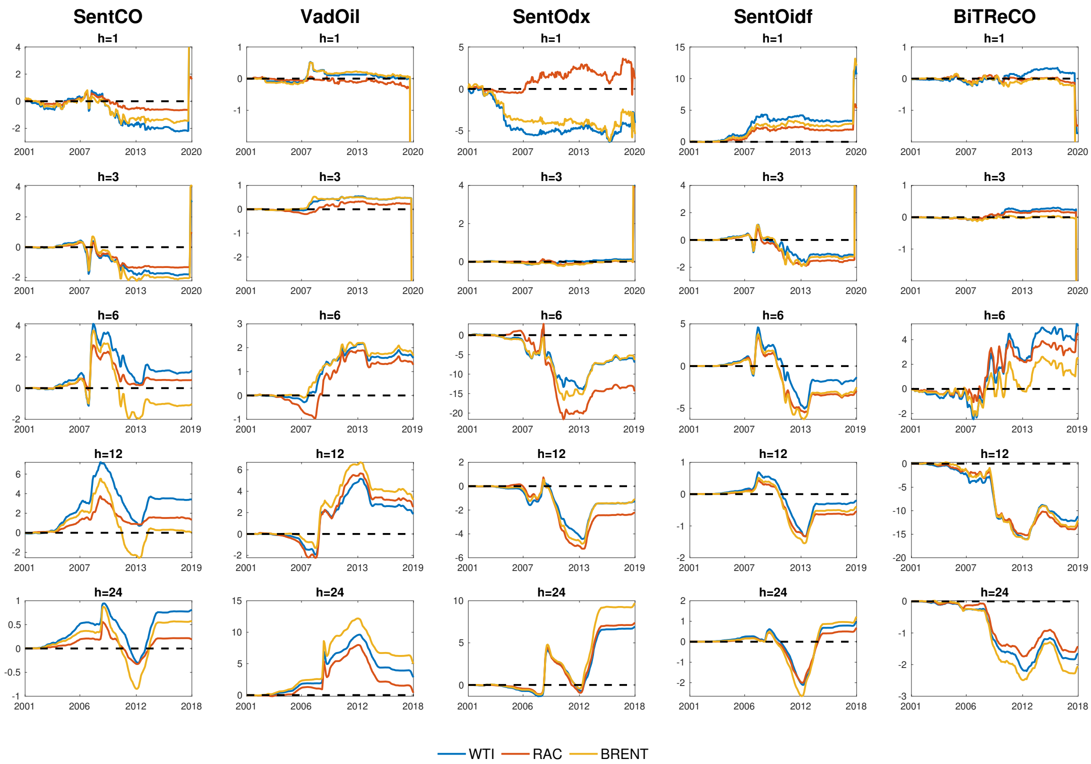


Figure 2.1: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and the benchmark model in (2.6). VAR parameters are estimated through a frequentist approach and text variables account for human sentiment about oil news.

Table 2.2: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a BVAR(12). World industrial production is used as a measure of global real economy, and text regressors account for human sentiment about oil news.

Monthly horizon		**SentCO**		**VadOil**		**SentOdx**		**SentOidf**		**BiTReCO**	
	Model-1	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3
A. WTI based VAR											
1	0.996	0.991	0.973	0.983	0.975	0.976	0.980	0.978	0.936	0.976	0.980
3	0.951	0.951	0.948	0.976	0.944	1.021	0.928	0.957	0.897	0.944	0.942
6	1.001	0.998	1.004	1.043	1.009	1.293	0.974	1.068	0.971	0.998	0.978
12	1.107	1.107	1.107	1.174	1.105	1.902	1.062	1.269	1.071	1.106	1.096
24	1.210	1.204	1.135	1.345	1.133	3.186	1.178	1.527	1.110	1.206	1.205
B. RAC based VAR											
1	0.846	0.846	0.831*	0.845	0.823*	0.838*	0.803**	0.826*	0.803**	0.843	0.845
3	0.934	0.937	0.925	0.968	0.914	1.071	0.887	0.933	0.875	0.931	0.931
6	1.001	1.000	1.018	1.061	1.009	1.490	0.977	1.064	0.970	0.997	0.993
12	1.151	1.142	1.141	1.257	1.129	2.516	1.093	1.317	1.096	1.138	1.137
24	1.227	1.222	1.147	1.413	1.138	4.791	1.191	1.569	1.105	1.222	1.220
C. Brent based VAR											
1	1.048	1.014	1.000	1.013	1.007	0.998	0.980	1.000	0.964	1.023	1.015
3	1.039	1.008	0.991	1.040	1.013	1.258	0.970	1.032	0.943	1.027	1.006
6	1.040	1.042	1.047	1.104	1.045	1.896	1.010	1.146	0.996	1.040	1.022
12	1.141	1.137	1.132	1.236	1.127	3.456	1.094	1.372	1.078	1.142	1.127
24	1.202	1.198	1.119	1.359	1.127	7.675	1.175	1.663	1.086	1.203	1.200

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*. *, **, *** respectively denote 10%, 5% and 1% level of significance of Diebold-Mariano test.

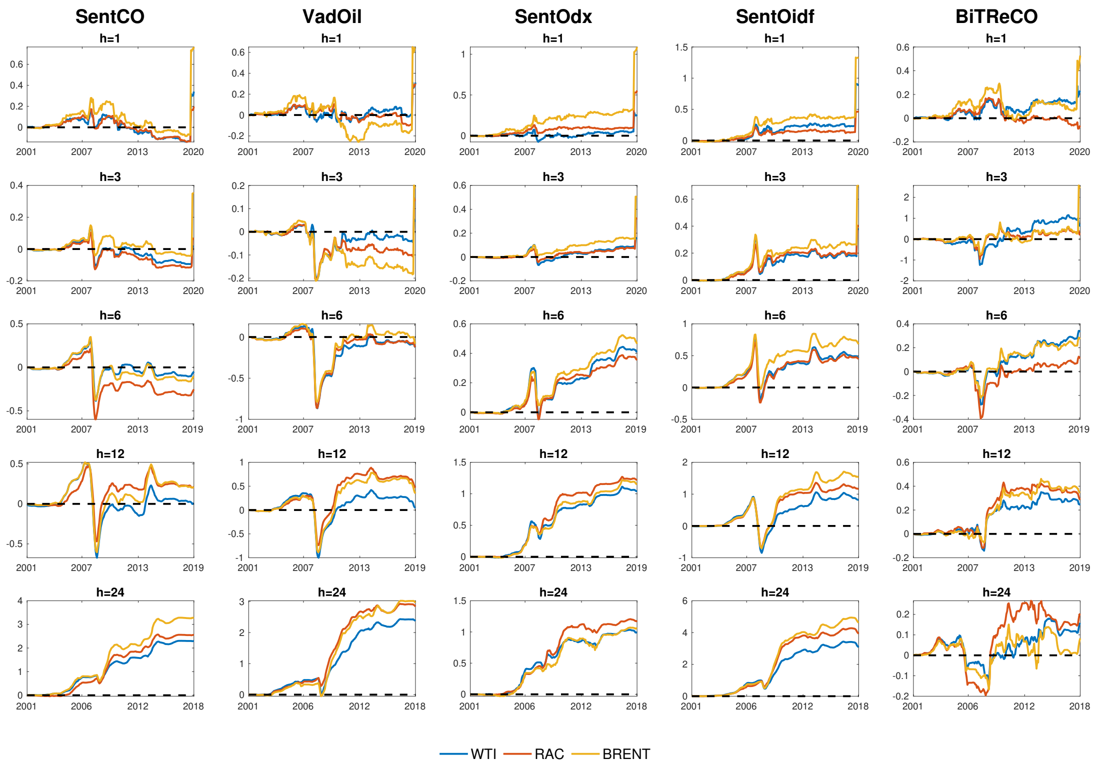


Figure 2.2: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and the benchmark model in (2.6). VAR parameters are estimated in a Bayesian fashion and text variables account for human sentiment about oil news.

Table 2.3: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a SV-BVAR(12). World industrial production is used as a measure of global real economy, and text regressors account for human sentiment about oil news.

Monthly horizon	*Model*-1	**SentCO** *Model*-2	**SentCO** *Model*-3	**VadOil** *Model*-2	**VadOil** *Model*-3	**SentOdx** *Model*-2	**SentOdx** *Model*-3	**SentOidf** *Model*-2	**SentOidf** *Model*-3	**BiTReCO** *Model*-2	**BiTReCO** *Model*-3
					A. WTI based VAR						
1	0.976	0.978	0.970	0.982	0.946	0.977	0.958	0.967*	0.924*	0.960	0.955
3	0.962	0.945	0.951	0.987	0.929	0.954	0.933	0.934	0.913	1.015	0.923
6	0.918	0.902	0.944	0.944	0.916	0.915	0.903	0.910	0.943	1.010	0.896
12	0.946	0.943	0.983	0.985	0.959	0.945	0.915	0.940	1.009	1.194	0.939
24	0.925	0.919	0.918	1.034	0.892	0.925	0.927	0.910	1.013	1.604	0.936
					B. RAC based VAR						
1	0.818**	0.819**	0.809**	0.821**	0.799**	0.822**	0.813**	0.824**	0.793***	0.796**	0.802**
3	0.909	0.901	0.907	0.936	0.890	0.907	0.875	0.907	0.875*	0.959	0.872
6	0.971	0.876	0.912	0.898	0.882	0.872	0.841*	0.868	0.911*	0.949	0.842
12	0.938	0.927	0.971	0.964	0.926	0.906	0.873*	0.917	0.982	1.119	0.899
24	0.847	0.852	0.845	0.973	0.811	0.830	0.828*	0.838	0.911	1.339	0.841
					C. Brent based VAR						
1	0.980	0.975	0.966	0.982	0.952	0.980	0.980	0.989	0.933	0.948	0.947
3	1.032	1.009	1.011	1.041	0.991	1.044	1.016	1.031	0.975	1.086	0.973
6	0.910	0.913	0.957	0.944	0.906	0.912	0.890	0.915	0.934	0.987	0.873
12	0.937	0.948	0.977	0.992	0.938	0.928	0.909	0.944	0.985	1.141	0.903
24	0.908	0.928	0.905	1.019	0.884	0.898	0.908	0.913	0.956	1.431	0.921

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*. *, **, *** respectively denote 10%, 5% and 1% level of significance of Diebold-Mariano test.

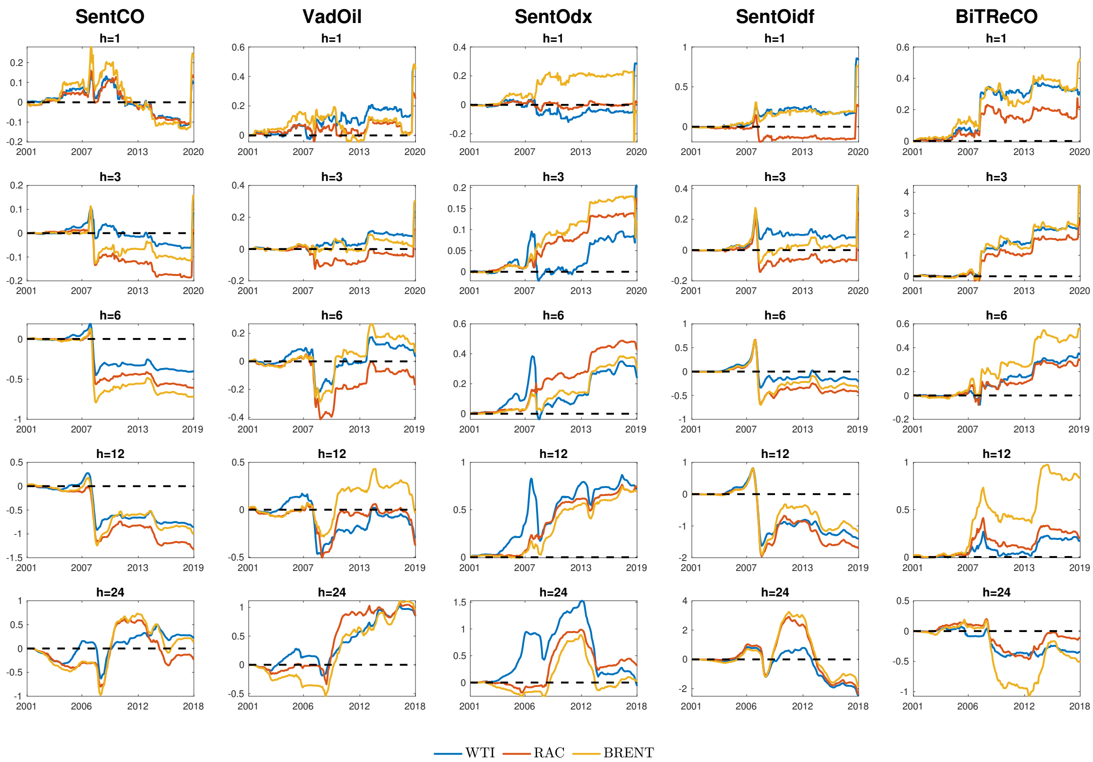


Figure 2.3: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.10) with and without text indicators. VAR parameters are estimated in a Bayesian fashion by assuming stochastic volatility in the error term, and text variables account for human sentiment about oil news.

Table 2.4: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices. Combination between different text data sources

Model	1-month			3-months			6-months			12-months			24-months		
	WTI	RAC	BRENT	WTI	RAC	BRENT	WTI	RAC	BRENT	WTI	RAC	BRENT	WTI	RAC	BRENT
A. SentCO															
FT-TR	0.970	0.809**	0.965	0.950	0.907	1.010	0.934	0.907	0.949	0.964	0.938	0.962	0.915	0.836	0.901
FT-IND	0.973	0.806**	0.962	0.965	0.917	1.020	0.948	0.920	0.960	0.982	0.958	0.977	0.919	0.845	0.907
FT-IND-TR	0.973	0.806**	0.962	0.965	0.917	1.02	0.948	0.920	0.960	0.982	0.958	0.977	0.919	0.845	0.907
B. VadOil															
FT-TR	0.945	0.797**	0.952	0.926	0.882	0.999	0.911	0.870*	0.906	0.946	0.901*	0.923	0.882	0.783*	0.870
FT-IND	0.949	0.806**	0.948	0.938	0.889	1.008	0.925	0.886	0.929	0.958	0.923	0.938	0.890	0.797*	0.876
FT-IND-TR	0.954	0.800**	0.954	0.934	0.893	1.009	0.920	0.882	0.915	0.953	0.904	0.921	0.888	0.784*	0.874
C. SentOdx															
FT-TR	0.961	0.812**	0.980	0.929	0.885	1.021	0.897*	0.839*	0.880	0.912*	0.873*	0.909	0.920	0.822*	0.908
FT-IND	0.954	0.811**	0.967	0.916	0.879	0.994	0.881*	0.835*	0.880	0.899*	0.864*	0.900	0.918	0.835	0.907
FT-IND-TR	0.953	0.807**	0.971	0.924	0.881	0.998	0.885	0.830*	0.876	0.892*	0.862*	0.897	0.916	0.833	0.909
D. SentOidf															
FT-TR	0.920	0.790***	0.949	0.913	0.871	0.988	0.915	0.889**	0.919	0.962	0.932	0.937	0.924	0.818*	0.898
FT-IND	0.914	0.788***	0.925*	0.902	0.861*	0.954	0.917	0.881*	0.911	0.969	0.943	0.937	0.968	0.871	0.928
FT-IND-TR	0.923	0.790***	0.926	0.901	0.860	0.952	0.900	0.862*	0.889*	0.939	0.903*	0.901*	0.915	0.807*	0.882
E. BiTReCO															
FT-TR	0.920	0.772***	0.891*	0.891	0.838	0.916	0.887	0.832	0.869	0.926	0.881	0.891	0.924	0.841	0.914
FT-IND	0.950	0.795**	0.949	0.928	0.877	0.983	0.887	0.848	0.879	0.929	0.885	0.898	0.929	0.840	0.919
FT-IND-TR	0.947	0.796**	0.940	0.921	0.864	0.978	0.891	0.835	0.875	0.931	0.875	0.897	0.938	0.837	0.928

Note: In column 1, FT: Financial Times, TR: Thomson Reuters, IND: Independent. Blue values report the lowest MSPE result relative to each horizon h and oil price measure (WTI, RAC and Brent). *, **, *** respectively denote 10%, 5% and 1% level of significance of Diebold-Mariano test.

Table 2.5: Summary of Empirical Results

	Oil Variable	Table	Model Suggested	Source
h=1				
	WTI	Table 2.4 Panel D	SentOidf based SV-BVAR	FT-IND
	RAC	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
	BRENT	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
h=3				
	WTI	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
	RAC	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
	BRENT	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
h=6				
	WTI	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
	RAC	Table 2.4 Panel C	SentOdx based SV-BVAR	FT-IND-TR
	BRENT	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
h=12				
	WTI	Table 2.4 Panel C	SentOdx based SV-BVAR	FT-IND-TR
	RAC	Table 2.4 Panel C	SentOdx based SV-BVAR	FT-IND-TR
	BRENT	Table 2.4 Panel E	BiTReCO based SV-BVAR	FT-TR
h=24				
	WTI	Table 2.4 Panel B	VadOil based SV-BVAR	FT-TR
	RAC	Table 2.4 Panel B	VadOil based SV-BVAR	FT-TR
	BRENT	Table 2.4 Panel B	VadOil based SV-BVAR	FT-TR

Note: In column 5, FT: Financial Times, TR: Thomson Reuters, IND: Independent.

Table 2.6: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a SV-BVAR(12). World industrial production is used as a measure of global real economy, and text regressors accounting for human sentiment about oil news are used as instrumental variables of stochastic errors of oil prices.

Monthly horizon	**TOSI**	**SentCO** *IV*-SV-BVAR	**VadOil** *IV*-SV-BVAR	**SentOdx** *IV*-SV-BVAR	**SentOidf** *IV*-SV-BVAR	**BiTReCO** *IV*-SV-BVAR
			A. WTI based VAR			
1	0.925*	0.919*	0.911*	0.909*	0.909*	0.909*
3	0.905	0.910	0.909	0.905	0.908	0.910
6	0.894	0.925	0.925	0.925	0.923	0.924
12	0.935	0.980	0.977	0.978	0.975	0.977
24	0.883	1.052	1.041	1.043	1.048	1.040
			B. RAC based VAR			
1	0.782***	0.804**	0.800**	0.803**	0.805**	0.807**
3	0.862	0.889	0.882	0.886	0.885	0.886
6	0.849*	0.919	0.912	0.919	0.915	0.917
12	0.887*	1.001	0.989	0.994	0.996	0.999
24	0.786*	1.053	1.051	1.051	1.053	1.053
			C. Brent based VAR			
1	0.928*	0.963	0.942	0.948*	0.950	0.944*
3	0.960	0.965	0.954	0.956	0.951	0.952
6	0.894	0.948	0.942	0.942	0.939	0.943
12	0.912	1.002	0.993	0.997	0.998	1.001
24	0.882	1.046	1.046	1.046	1.048	1.045

Note: Blue values show the lowest MSPE results relative to a specific time horizon *h*. *, **, *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

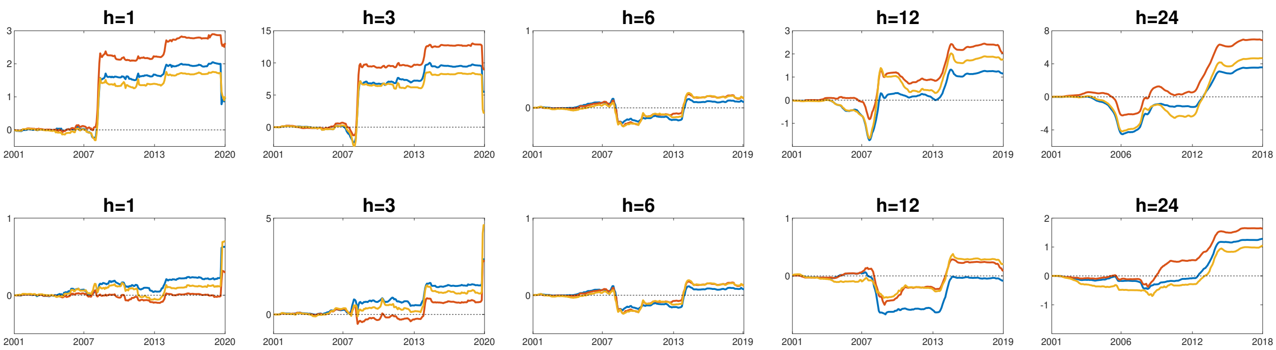


Figure 2.4: First line plots the difference between the cumulative sum of forecasting errors of equation (2.10) with TOSI endogenised and a random walk. Second line plots the difference between the cumulative sum of forecasting errors of equation (2.10) with and without TOSI. In both cases VARs are estimated in a Bayesian fashion by assuming stochastic volatility in the error term. Blue, red and yellow lines describe WTI, RAC and Brent crude oil respectively.

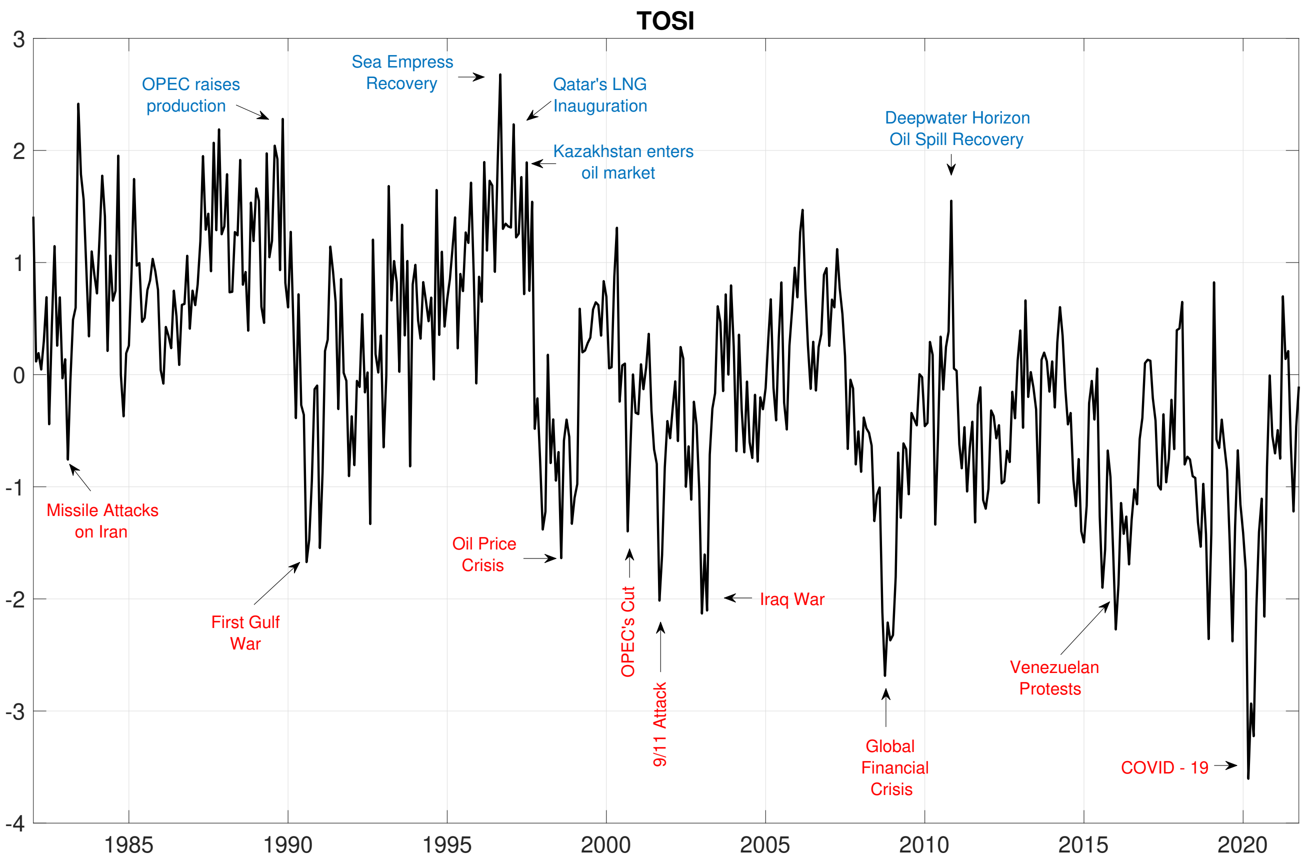


Figure 2.5: First principal component factor extracted from a dataset made up by the following text indicators: VadOil, SentOdx and BiTReCO. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil. Sample period runs from 1982M1 through 2021M11 and articles are drawn from The Financial Times, Thomson Reuters and The Independent.

Table 2.7: Total variation explained by individual common components.

PCA factors	R^2 **VadOil**	R^2 **SentOdx**	R^2 **BiTReCO**	**Overall dataset**	**Avg importance**
TOSI	0.53	0.99	0.90	0.86	0.81
Factor 1	0.47	0.11	0.01	0.14	0.19
Factor 2	0.01	0.10	0.01	0.01	0.04

Note: Column 1 lists the 3 principal component factors. Columns 2, 3 and 4 reports the R^2 generated by regressing the relative text series on each component. For example, TOSI explains 53% variation of VadOil, 99% variation of SentOdx and 90% variation of BiTReCO. The fraction of variance in the data explained by individual components is listed in column 5, whereas the last column reports the average importance of each principal component variable.

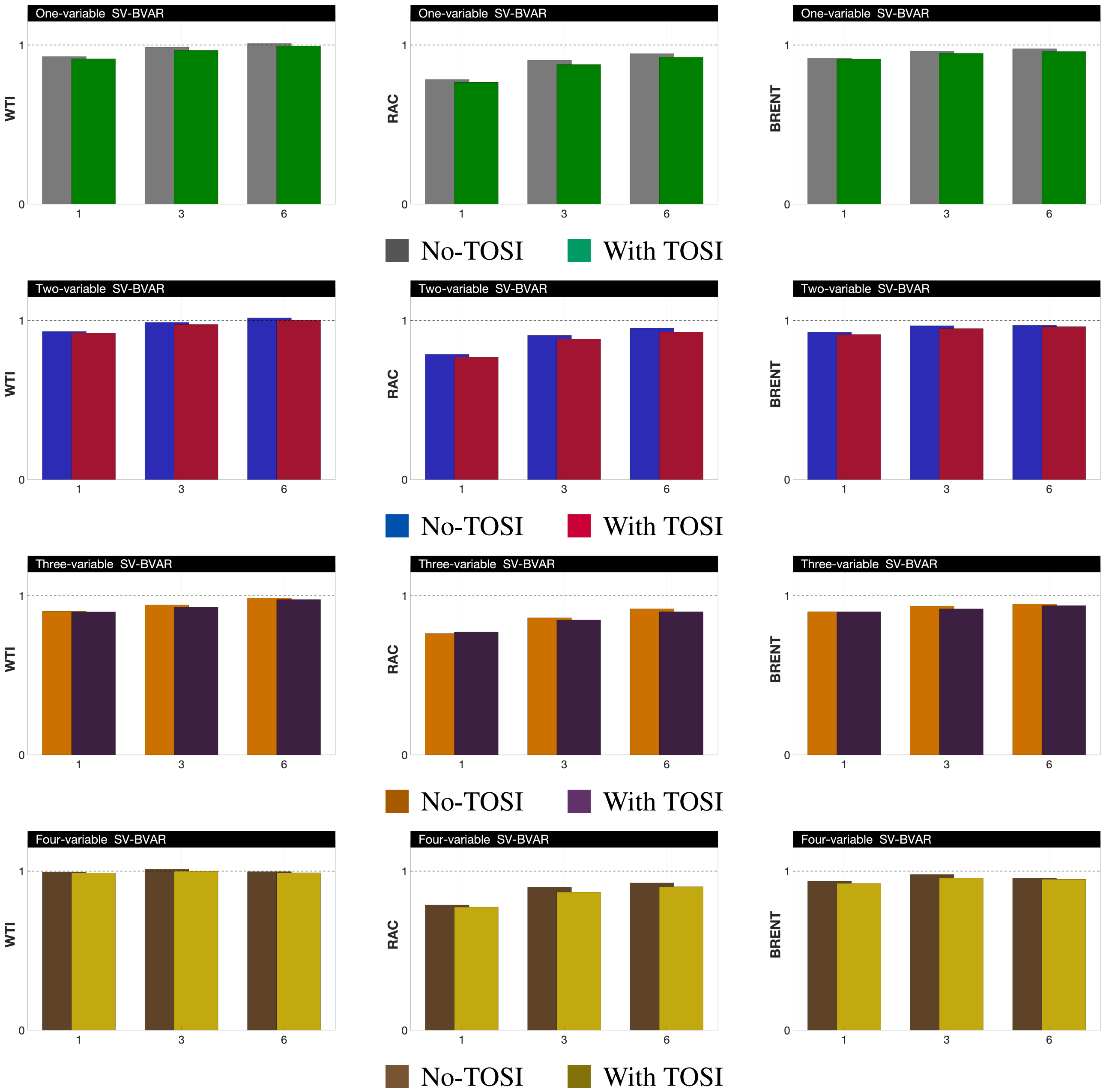


Figure 2.6: Comparison between the cumulative sum of forecasting errors among four alternative SV-BVAR models with and without TOSI. The oil measures are reported on the y-axis, while the month-ahead forecast is displayed on the x-axis. Models beat the random walk when bars lay below 1.

3

Do High Frequency Text Data Help Forecast Crude Oil Prices?

MF-VAR vs. MIDAS

3.1 Introduction

Following Russia's invasion of Ukraine and the unstable inflation rates over the past few years, investors and central bankers are apprehensive about the future prices of oil. The idea that crude barrel prices could hit new record highs is concerning to firms, households, and governments. But how far can oil prices go in the short run? And what is the probability that they will turn negative again? Such questions are of crucial importance, as central banks can only make informed policy decisions if they can rely on accurate forecasts of inputs which could have far-reaching impacts on the gross domestic product and inflation.

In Gifuni [2021] I emphasize that oil market fundamentals commonly used to forecast future prices of oil are vulnerable to unexpected economic shocks, due to delays in data being made available. For this reason, I provide a solution by developing a new and highly informative text oil sentiment indicator (TOSI). This index is based on written information retrieved in real time and is designed to capture the human sentiment in the oil market. TOSI is calculated on a monthly basis and is used as an endogenous variable in a vector autoregression (VAR) model. In Gifuni [2021] I show that text data yield significant forecasting gains, especially across periods of financial instability. This is mostly because not only is TOSI available at high frequency, but it is also not subject to revisions. However, daily data combined on a monthly basis are unable to provide a truly accurate reflection of specific, global and economic events. Albeit infrequent, occurrences (for example wars and natural disasters) still have the potential to evoke widespread and profound economic changes. What's more, information available at the end of the month is likely to be more informative than older news (Ghysels et al. [2004], Ghysels et al. [2007]).

For this reason, in this work I aim to show whether information retrieved from daily newspapers, combined on a weekly and daily basis, is useful for improving how accurately monthly real oil prices can be predicted. Specifically, I develop alternative weekly and daily text-based series and examine whether said indicators can improve the out-of-sample forecasts of monthly real oil prices. Both tasks require the use of text mining strategies[1] and mixed-frequency (MF) models.

This study builds on the literature aiming to explain the fluctuations in crude oil prices through variables sampled at different frequencies. Prominent contributions include Baumeister et al. [2015], Degiannakis and Filis [2018], Ma et al. [2019] and Gong et al. [2022]. I return to this question using text data, which has

[1]In his paper I use the terms "text mining", "natural language processing", "computational linguistics" and "text analysis" are used interchangeably.

long been considered as an alternative source of knowledge with potentially informative content. Text as data is indeed used in many areas of the literature (see for example Baker et al. [2016], Hansen et al. [2018], Shapiro et al. [2022] and Shapiro and Wilson [2022]). This paper is also closely related to the fast-growing body of work aimed at discovering methodologies that perform particularly well in periods of economic instability; see Clark et al. [2022], Chan [2022] and Carriero et al. [2022]. In the context of my framework, I develop a set of weekly and daily text indicators that incorporate new information available in real time. Said indicators are further included in alternative models with variables sampled at different frequencies. In this way immediate changes can be promptly captured and the reaction of low frequency variables can be identified accordingly.

While the use of models accommodating MF data finds a broad empirical application among several areas in the economic literature[2], the idea of forecasting the monthly price of oil with data sampled at different frequencies is relatively unexplored. To the best of my knowledge there are only two notable exceptions: Baumeister et al. [2015] and Degiannakis and Filis [2018]. In the former study, the authors use financial data available at a daily and weekly frequency to fit Mixed-data sampling (MIDAS) and MF-VAR models. Findings show that despite the predictability of oil prices through MF models, overall not much information is lost by ignoring high frequency data when forecasting the monthly price of oil. In contrast, Degiannakis and Filis [2018] point out that there are several informative financial and commodity variables not explored in Baumeister et al. [2015], the variation of which can impact on crude oil spot prices. Said variables also provide a large amount of ultra-high frequency data[3], that the authors use to construct variable-specific realised volatilities, which are then incorporated into alternative MIDAS. Results show that realised volatilities of financial variables radically improve the MIDAS performance on a no-change forecast at short, medium- and long-term horizons.

However, a number of pitfalls can be observed from both empirical works. Firstly, high-frequency variables only rely on financial and commodity data, while there are several recent studies showing the importance of combining daily and weekly text data along with financial variables and oil market fundamentals (Li et al. [2019], Gifuni [2021], Bai et al. [2022]). Secondly, the evaluation period used in Baumeister et al.

[2] A comprehensive review concerning the use of MF models is beyond the purpose of this paper. However, Foroni and Marcellino [2013] and Foroni et al. [2013] survey the most common techniques for sampling variables at different frequencies, alongside many empirical applications.

[3] This expression refers to the availability of a large number of intraday observations. These were originally determined by massing tick-by-tick market data, where each tick is one logical unit of information. In financial markets it is common practice to observe thousand of ticks or transactions per business day.

[2015] and Degiannakis and Filis [2018] does not cover the most recent events that have made crude spot oil prices hard to predict (e.g. the Covid pandemic and the Russia-Ukraine war). If indeed the sample period runs until recent times, high frequency financial data do not produce similar gains in terms of accuracy (see section 3.5). Such drawbacks are addressed in this study.

Specifically, I use Gifuni [2021]'s dataset to construct a set of real-time text variables. Written information is retrieved from the most widely read newspapers, such as The Financial Times, Thompson Reuters and The Independent. Text data run from 1982M1 to 2021M12 for a total of 140,096 oil related articles downloaded from the LexisNexis database. Text variables are designed to capture the human sentiment and the uncertainty in the oil market. Said indicators are included into alternative MF models alongside several financial and commodity variables that are commonly used to forecast the real oil prices. The analysis starts with a comparison between homogeneous- and mixed-frequency models, with and without text variables. I demonstrate that, endogenising weekly text data in a monthly based mixed-frequency stochastic volatility Bayesian VAR does improve the forecasting performance of the model in the short run. However, such marginal gains are low and negligible.

In a second experiment I follow Degiannakis and Filis [2018] and use intraday data to compute the daily returns from financial and commodity markets. MIDAS models with daily returns and oil market fundamentals are then used to forecast WTI crude oil spot prices. The evidence shows that intraday returns marginally improve the out-of-sample 1-month ahead prediction on the no-change forecast. But at any other time horizon, random walks generate better forecasts in terms of accuracy. In any case, statistically significant improvements are observable when MIDAS models include both text data and intraday returns. In particular, I find that the Commodity Research Bureau (CRB) index, the GBP/USD exchange rate and the natural gas spot price generate the most informative daily returns. When such information is used in a MIDAS model along with text data, it is possible to observe a marginal improvement on the no-change forecast up to 18%, and a statistically significant gain up to 1%. For 1-step ahead forecasts, results are even more accurate than the outcome generated by the corresponding model incorporating variables sampled at homogeneous frequency. However, improvements are minimal and the evidence still supports Baumeister et al. [2015]'s empirical findings. This is true for weekly and daily data, as well as point and density forecasts.

The remainder of this paper is organised as follows. Section 3.2 illustrates the methodology used to manipulate daily text data at different frequencies. Text metrics are described in section 3.3, while section

3.4 outlines the mixed-frequency models used to forecast the price of oil. Section 3.5 reports the empirical results and section 3.6 concludes.

3.2 Text-Data Transformation

Text documents rely on written articles that featured in the banking, finance and energy section of The Financial Times, Thomson Reuters and The Independent from 1982M1 to 2021M12. In order to retrieve articles and transform text into structured data, a number of procedures were required. The main steps are outlined as follows:

1-Download and Storage. Articles are selected based on the joint occurrence of the words *oil* and *price*. Documents are firstly downloaded in `.rtf` format from the LexisNexis database, then transformed into `.txt` files, and finally stored in "newspaper-specific" monthly folders. Before transforming text into numbers, I have also done a word search with the purpose of removing articles not discussing oil-related topics. For example, the words *oil* and *price* might also occur in articles discussing topics related to the olive oil market, or the biography of an oil-company worker. Said articles are all removed from the database, as they do not include informative data.

2-Cleaning. After storing `.txt` files in monthly folders, documents have unnecessary white spaces, punctuation and numbers removed. I have then retrieved the title and full text by removing non informative written sections, such as copyrights, publication name, location, editor and additional notes placed at the end of the body. Following this, words are lowercased and then stemmed or lemmatised[4], based on the text processing methodology performed. In particular, for *bag-of-words* methodologies, I apply the stemming procedure, as converting a word to its lemma might generate bias when tokens are counted. For example, when oil prices turned negative for the first time in history, the Independent released two articles with the following content:

> **Firms have an opportunity to change for the better.** ...They should not enter into sham schemes to reduce their taxes. They should not pollute or treat their staff and customers badly. And they should be seen to mean it. In essence, they should behave as good citizens...There will be a reckoning after this. Somebody is going to have to pay for the mind-boggling, budget-

[4] Stemming procedure is used to cut off the suffix of a word by preserving its original form. Differently from the lemmatization that works on the morphological analysis and reduces words to their original root.

> busting rescue packages and handouts. Governments do not make money; they collect and spend other people's. It's inevitable then that attention will turn to those who can best fork out. That is bound to point to the wealthiest in society, to the richest individuals and corporations. (*The Independent, April 11 2020*)
>
> **British Airways job cuts: what will the redundancies mean for passengers? It is certain that BA's intercontinental fleet will become significantly smaller.** ...Balpa says it is baffled by the decision. The general secretary, Brian Strutton, told The Independent: "I would have thought British Airways would wait to see what was happening to its competitors. "I wonder also if it's a bit opportunistic? Maybe a bit of a land grab, because they're also talking about our terms and conditions." A phrase that is widely repeated in aviation is: "Never waste a good crisis." ...A reduction in capacity generally means an increase in prices. If procedures such as leaving middle seats empty are introduced, that alone will trigger a 50 per cent fare rise. I predict that for years we will look back on the summer of 2019 as "peak mobility" and best value for airline passengers. (*The Independent, April 29 2020*)

If both pieces of text are lemmatised, words like *good*, *better* and *best*, are all mapped to *good* which is also used inside the text but with a different meaning. Therefore for *unigram* and geometrical models, tokens are stemmed, as the purpose of both strategies is to count a preselected root word (in my case "`econom`" for human sentiment and "`uncertain`" for oil market uncertainty). In contrast, for dictionary and Bert model words are lemmatised in order to retrieve the meaning from each stand-alone token.

3-Transformation. The procedure for mapping words into numbers changes across the text metric used. For this reason the reader is advised to read section 3.3, where each methodology is discussed thoroughly. However, it is important to outline here the general procedure employed to generate daily, weekly and monthly text indicators. As previously remarked, the database consists of daily articles. A numerical score is computed for each article and averaged across the remaining daily scores (if any). For weekly/monthly indicators, daily scores are then averaged across the values belonging to the corresponding week/month.

3.3 Text Metrics

NLP methodologies investigated in this study build on Gifuni [2021]. However, while Gifuni [2021] only uses computational linguistics to develop monthly indicators, in this paper text metrics are also used to compute daily and weekly indexes. Text indicators are designed to capture the human sentiment and the uncertainty in the oil market, and the text metrics are outlined as follows.

I start with the *unigram* count probability model, where I use the word-specific maximum likelihood estimation

$$p(\omega_k) = \frac{counts(\omega|k)}{\sum_{n=1}^{N} counts(\omega_n|k)}, \qquad \text{for } k = 1, \ldots, K. \tag{3.1}$$

for tokens ω = {"`economy`", "`economies`", "`economic`" and "`economics`"} occurring in document k. N represents the number of words running in article k. Document-specific *unigram* scores are then averaged across the number of articles released on day d or week w as follows

$$\omega_{d/w} = \frac{\sum_{k=1}^{K} p(\omega_k|d/w)}{\text{Total No. Doc.}_{d/w}}$$

It is common practice to use equation (3.1) for constructing text series aiming to capture the human sentiment. Abrahams et al. [2015], Renault [2017], McGurk et al. [2020] and Shapiro et al. [2020] are some examples of empirical works using such a procedure for improving the forecasting accuracy of alternative econometric models. This measure, albeit not statistically well-performing when considered at a monthly frequency (see Gifuni [2021]), is in this work used for the first time to develop an oil related human sentiment index at a daily and weekly frequency. The time series are depicted in Figure 3.1. An equivalent procedure is used to construct uncertainty based indicators. Tokens ω = {"`uncertain`", "`uncertainty`", "`uncertainties`" and "`uncertainly`"} are stemmed to "`uncert`". For expository purposes, the empirical application of uncertainty measures is provided in Appendix B.3.

In a second analysis, I use Valence Aware Dictionary and sEntiment Reasoner (VADER) of Hutto and Gilbert [2014] to classify tokens into positive, neutral and negative classes. The mapping of words to numbers follows the following equation

$$S_k = \frac{\sum_{n=1}^{N} \Theta_n \omega_n}{\sum_{n=1}^{N} \omega_n} \qquad \text{for } k = 1, \ldots, K \tag{3.2}$$

and the daily/weekly score is

$$S_m = \frac{\sum_{k=1}^{K} S_k}{\text{Total No. Doc.}_{d/w}}$$

Θ is the set of pre-assigned weights with values running from -4 (for very negative tokens) to +4 (for very positive tokens), which is then normalised into a more narrowed interval [-1; +1]. Figure 3.2 shows the weekly and daily series of VADER. This methodology has also been applied in several recent works; see for example Shu et al. [2020], Baylis [2020] and Suardi et al. [2022].

Furthermore, human sentiment and oil market uncertainty information is retrieved through the application of two geometrical models: (i) term frequency matrix (tfm) and (ii) term frequency - inverse document frequency matrix (tf-idf). Both strategies rely on the frequency of occurrences of preselected *unigrams*. However, in contrast to equation (3.1), geometrical models are not proportionally sensitive to term frequency changes. Specifically, tfm is given by

$$tfm_{\omega} = \frac{\sum_{k=1}^{K} \omega_{n,k}}{\text{Total No. Doc.}_{d/w}} \tag{3.3}$$

and incorporates information about the frequency of a token in each article for the overall daily or weekly data collection K. While tf-idf is given by

$$tf\text{-}idf_{\omega} = \frac{log\left[1 + tfm_{\omega}\right]}{log\left[1 + \left(K_{d/w}/m_{d/w}\right)\right]} \tag{3.4}$$

and helps to overtake the limits associated with equation (3.3). In particular, tf-idf normalizes the tfm value over the inverse ratio between the overall data collection K and the number of (daily or weekly) documents m where a generic token ω occurs. This particular weighting scheme allows the methodology expressed by equation (3.4) to map words that are rare within and across the dataset as very informative tokens, and words common within and across the dataset as non informative (see Banchs [2012] for a review, or Kalamara et al. [2020] and Li et al. [2021] for an empirical application). For expository purposes only tfm series are reported (see Figure 3.3).

Finally, I also use Bidirectional Encoder Representations from Transformers (BERT) of Devlin et al. [2018] to process words in a deep learning structure of 12 layers, 768 hiddens, 12 heads and 110M parameters. Daily and weekly series are displayed in Figure 3.4. The methodologies briefly described in this section are used to develop human sentiment and oil market uncertainty indexes at a daily and weekly frequency, which are then included in alternative econometric models explained in section 3.4. For additional details of such methodologies, see Gifuni [2021].

3.4 Real-Time Forecasting Models

In this section I discuss the econometric models used to forecast alternative measures of monthly real oil prices. Two macro models are explored: (i) polynomials with variables sampled at homogenous frequency and (ii) mixed-frequency models.

The univariate MIDAS regression is the first mixed frequency model examined. The structure of this polynomial is also discussed in Ghysels et al. [2007], Ghysels and Wright [2009] and Clements and Galvão [2008, 2009], and involves the use of a Beta lag weight function. Results are reported in Appendix B.3, where I show that exponential Almond, equal-weighted lag and unrestricted MIDAS generate approximately similar results in terms of accuracy. I then estimate a mixed-frequency VAR (MF-VAR) model for each text variable examined. For this polynomial, I use a parsimonious data-driven methodology, where low and high frequency variables are stacked in one matrix and the model is estimated at its lowest observed frequency (see Baumeister et al. [2015] and Ghysels [2016] for more details about this methodology)[5].

3.4.1 Non Text Based Models

The most basic methodology used in this work to forecast the monthly value of real oil prices is the univariate autoregressive (AR) model. The reason behind this choice is because AR polynomials have long been used as a model benchmark across the literature, as they provide robust forecasts when estimated with consistent estimators such as OLS (Stock and Watson [2003], Banerjee et al. [2005]). Therefore, I first assume that oil prices evolve according to the following stochastic equation:

$$y_t = c + \rho_1 y_{t-1} + \ldots + \rho_p y_{t-p} + \varepsilon_t, \qquad \varepsilon \sim N\left(0, \sigma^2\right) \tag{3.5}$$

[5]There is a second approach that is widely used in the literature, where the VAR is modelled in a state space representation form (see Giannone et al. [2008], Ghysels and Wright [2009], Schorfheide and Song [2015] and Schorfheide et al. [2018]). Albeit efficient, this methodology is useful to estimate and then now/forecast the missing values of a low-frequency variable (i.e. quarterly GDP). In contrast, the purpose of this work is to forecast the monthly value of an ultra-high-frequency variable as the price of oil. Using a state space model would only create bias, as there is no missing information to estimate.

where the dependent variable y represents the price of oil in log-level at time t, c is the intercept, p is the number of lags, ρ are the unknown coefficients estimated through least square, and ε is an exogenous shock. In the empirical application, the number of lags is determined by using the Bayesian Information Criterion (BIC) methodology, and it changes across each recursive forecast.

Subsequently, the univariate model in equation (3.5) is augmented with exogenous lagged factors representing the main oil market fundamentals. Such variables are: (i) oil produced at a world level (q^{oil}), (ii) industrial production as a proxy of global real economy (wip) and (iii) oil inventories (inv)[6]. In this case the dependent variable y_t is also estimated through OLS, and BIC is used to determine the most likely number of lags p.

In conjunction with the estimation through OLS, oil prices and market fundamentals are then included into a 4-variable VAR specified through the following reduced form:

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + u_t \qquad u \sim iid\ N(0, \Sigma) \tag{3.6}$$

where Y_t is a 4×1 vector comprising $\left[q_t^{oil}, wip_t, p_t^{oil}, inv_t\right]$, c is a vector of intercepts, Φ are 4×4 matrices of unknown coefficients and u is a normally distributed error vector with mean 0 and full covariance matrix Σ. In this case, the number of lags is not recursively determined by BIC, but it is set equivalent to 12 for each out-of-sample forecast, as suggested by Baumeister and Kilian [2015]. Equation (3.6), as well as its structural representation, has been used in several influential works in the oil literature, such as Baumeister and Kilian [2012], Kilian and Murphy [2014], Alquist et al. [2013], Degiannakis and Filis [2018] and many others. A natural concern of estimating a model through OLS across periods of financial turbulences is that the model might overfit the data. Indeed, since 2010 oil prices have become quite unstable due to their high uncertainty principally linked to unexpected economic/political changes and natural disasters (e.g. Great Recession, Covid pandemic, OPEC disagreements, Russia-Ukraine war). This has made the price of oil hard to forecast with traditional statistical methods, especially when the system is richly parameterised (see Baumeister et al. [2020] and Gifuni [2021]). For this reason, I regularize the overfitting problem through the following objective functions:

[6] Lippi and Nobili [2012], Caldara et al. [2019] and Baumeister and Kilian [2015] are a few examples of works where oil market fundamentals (along with additional oil related variables) are used to analyse the behaviour of real oil prices over time

- Ridge (L_2 penalty): $\sum_t \left(y_{t+h} - \hat{y}_{t+h|t}\right)^2 + \lambda_R \sum_{j=1}^{p} \rho_j^2$;

- Lasso (L_1 penalty): $\sum_t \left(y_{t+h} - \hat{y}_{t+h|t}\right)^2 + \lambda_L \sum_{j=1}^{p} |\rho_j|$;

- ElasticNet (L_1 + L_2 penalty): $\sum_t \left(y_{t+h} - \hat{y}_{t+h|t}\right)^2 + \lambda_R \sum_{j=1}^{p} \rho_j^2 + \lambda_L \sum_{j=1}^{p} |\rho_j|$.

where $\hat{y}_{t+h|t}$ denotes the forecast value of variable y_t at horizon h, while the parameters λ_R and λ_L control the shrinkage of coefficients. For λ_R and/or $\lambda_N = 0$, the coefficient estimates converge to ordinary least square estimators. In contrast, larger values of λ imply a more aggressive penalty on parameters which shrink towards zero. Ridge, Lasso and ElasticNet are all solutions that prevent the parameter proliferation by using a different penalty factor. In particular, Ridge uses the squared magnitude of coefficients as a penalty term in order to limit the size of the coefficient vector. The larger λ_R is, the closer the coefficients get to zero. Lasso uses the absolute operator to eliminate variables with coefficients that zero. If from one side this can be seen as an improvement (since Ridge model does not shrink coefficients to zero), in case of multicollinearity it can happen that Lasso might rule out relevant independent variables. ElasticNet uses both L_1 and L_2 penalties in order to combine the characteristic of both Ridge and Lasso. In this way the total effect determined by all coefficients is reduced without eliminating all the features.

Another approch that I use to reduce the number of parameters is the Bayesian shrinkage, where a reasonable choice of prior not only shrinks a rich set of coefficient estimates,but in many cases also yields a better inference. Estimating equation (3.6) in a Bayesian Fashion does not preclude any structural change of the VAR model, other than specifying a prior distribution for the unknown parameters Φ and Σ. In my case, I follow Giannone et al. [2015] and specify the following priors:

$$\Sigma|\xi \sim IW(\psi, d)$$

$$\Phi|\Sigma \sim N(\phi, \Sigma \otimes \Omega\xi)$$

where Ω is a predefined Minnesota shrinkage rule, and ξ is an unknown parameter used to make inference on the informativeness of the prior. Specifically, a value for ξ is chosen whenever the marginal density of data as a function of all possible values of ξ is maximized. Under the aforementioned assumptions, the posterior densities are shown to be:

$$
\begin{aligned}
\Phi|\Sigma, Y &\sim N\left(\hat{\Phi}(\xi), \Sigma \otimes \hat{V}(\xi)\right) \\
\hat{\Phi}(\xi) &= \operatorname{vec}\left(\hat{\phi}(\xi)\right) \\
\hat{\phi}(\xi) &= \left(X'X + (\Omega\xi)^{-1}\right)^{-1}\left(X'Y + (\Omega\xi)^{-1}\hat{\phi}\right) \\
\hat{V}(\xi) &= \Sigma \otimes \left(X'X + (\Omega\xi)^{-1}\right)^{-1}
\end{aligned}
$$

However, as previously mentioned, oil prices can be particularly hard to forecast in the long run, especially when unexpected shocks occur. This methodology, although well performing across different kinds of dataset[7], is not a well-suited strategy for oil price long-run forecasts (see Baumeister et al. [2020] and Gifuni [2021]). In contrast, Bayesian VAR allowing for stochastic volatility (SVBVAR) in the error structure are shown to be a valuable alternative for solving the issues of their homoskedastic counterpart[8].

For this reason, in a first application I implement the SVBVAR developed in Carriero et al. [2015], which is further extended to a text based mixed-frequency BVAR (MF-SVBVAR). The model is specified as follows:

$$
Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_p Y_{t-p} + u_t, \qquad u_t \sim N(0, \Sigma_t) \tag{3.7}
$$

where $\Sigma_t^{-1} = A' \Lambda_t^{-1} A$, and A is $N \times N$ lower triangular, and Λ_t is a diagonal matrix:

$$
A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \alpha_{1,1} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \alpha_{N,1} & \cdots & \alpha_{N,N-1} & 1 \end{bmatrix} \qquad \Lambda_t = \begin{bmatrix} \exp(\lambda_{1,t}) & 0 & \cdots & 0 \\ 0 & \exp(\lambda_{2,t}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp(\lambda_{N,t}) \end{bmatrix}
$$

The log-volatilities in Λ_t evolve according to the following random walk:

$$
\tilde{\lambda}_t = \tilde{\lambda}_{t-1} + v_t, \qquad \text{with } v_t \sim N(0, Q),\ Q = diag\left(q_1^2, \ldots, q_N^2\right) \text{ and } \mathbb{E}(u_t, v_t) = 0
$$

Models presented in this section have all been extended to incorporate monthly, weekly and daily text data. The latter two cases identify the so called mixed-frequency models, which are thoroughly discussed in the

[7] See Alquist et al. [2013], Weale and Wieladek [2016], Lenza and Primiceri [2020] and Miranda-Agrippino and Ricco [2021].
[8] See Koop and Korobilis [2013] and Carriero et al. [2019] for a comparison.

next section.

3.4.2 Univariate Text Based Models

Time series regression models presented so far involve the use of data sampled at the same frequency. These models are inadequate when weekly or daily text data are used to explain the behaviour of a lower-frequency variable, such as the monthly real price of oil. In this case a new class of models needs to be introduced.

For univariate regressions, since the seminal work of Ghysels et al. [2004], mixed-data sampling (MIDAS) models have made a great contribution to modelling and forecasting macroeconomic and financial variables[9]. MIDAS models are parsimonious and flexible one-sided polynomials, in which high-frequency explanatory variables are incorporated into a lower frequency regression. In my framework, I propose the following general model structure:

$$y_t = c + \sum_{j=1}^{p} \left\{ X_{t-j} B_j + txt_{t-j}^{(w)} \Theta_j \left[D\left(L^{1/w}; \tau \right) \right] \right\} + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \sigma^2\right) \tag{3.8}$$

where, y_t is the dependent variable assessing the monthly value for alternative measures of real oil prices. $X_t = [y_{t-1},\, x_{t-1}]$ is the matrix of low-frequency observations, and x_{t-1} incorporates the lagged values of oil production, global real economy and oil inventories. $txt_{t-j}^{(w)} = \left[x_t^{w,\,1},\, x_t^{w,\,2}, \ldots, x_t^{w,\,n} \right]$ stacks the weekly text-data released in month $t-j$, where n is the number of weeks in a given month, and ε_t is the vector of error terms. B and Θ are unknown parameters to be estimated, $L^{1/w}$ is a lag operator, and $D\left(L^{1/w}; \tau\right)$ is the weighting scheme that is parameterized in three different ways:

Beta Lag. I follow Ghysels et al. [2007], Rodriguez and Puggioni [2010] and Foroni and Marcellino [2013], and use the Beta lag estimation to define the daily weights. In particular,

$$D(k;\, \tau) = \frac{f\left(k/n;\, \tau\right)}{\sum_{k=0}^{n-1} f\left(k/n;\, \tau\right)},$$

$$\text{where} \qquad f(x,a,b) = \frac{x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)} \qquad \text{and} \qquad \Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$$

where I set $\tau = \left[\tau_{(1)},\, \tau_{(2)}\right]$, with $\tau_{(1)} = 1$ and $\tau_{(2)} = 5$ in order to allow for a declining weighting scheme.

[9] Some notable contributions are Ghysels et al. [2007], Ghysels et al. [2005], Ghysels et al. [2006], Clements and Galvão [2009], and more recently Guérin and Marcellino [2013], Foroni et al. [2015].

The greater $\tau_{(2)}$ is, the faster the declining rate will be. B and Θ are then estimated through a non linear least square method.

EXPONENTIAL ALMON LAG. This refers to the interpolation distribution originally developed by Almon [1965] and pursued in several prominent works, such as Modigliani and Sutch [1966], Griliches [1967], Ashenfelter and Pencavel [1969], Feldstein and Eckstein [1970], Laffont and Garcia [1977], Blinder [1981] and many others. In which:

$$D\left(L^{1/w};\ \tau\right) = \sum_{k=0}^{3} w\left(k;\tau\right) L^{k/w},$$

$$L^{k/w}\left(X_t^{(w)}\right) = X_{t-k/w}^{(w)},$$

$$w\left(k;\ \tau\right) = \frac{e^{\tau_1 k + \ldots + \tau_z k^z}}{\sum_{k=0}^{3} e^{\tau_1 k + \ldots + \tau_z k^z}}, \qquad \text{and } z = 2.$$

I set $\tau_{(2)} \leq 0$ to guarantee a declining weight. The number of lagged weeks p is defined by the minimum value that the Bayesian information criterion determines for each model estimation. The unknown parameters B and Θ are then estimated through a non linear least square methodology.

EQUAL-WEIGHTED LAG. This is a more parsimonious univariate regression representation, as the weekly variables $x_t^1, x_t^2, x_t^3, x_t^4$ all have the same weight. While Almond and Beta approaches are adopted to guarantee higher weights to data observed at the end of each month, the use of equal weights is useful to avoid any misspecification bias in the variance of the out-of-sample forecasts. It is worth noting that equal weights may also be obtained in the exponential Almond case by setting $\tau_1 = \tau_2 = 0$, or with the Beta function as long as $\tau_1 = \tau_2 = 1$. In any case, in this case equation (3.8) is linear in B and Θ, and can be expressed as:

$$y_t = c + \sum_{j=1}^{p} \left\{ X_{t-j} B_j + \sum_{k=0}^{3} \frac{1}{4} X_{t-k/4-j}^{(w)}\, \Theta_j \right\} + \varepsilon_t, \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

where B and Θ can be now recursively estimated through linear least square.

U-MIDAS. The weighting scheme is left unrestricted, the unknown coefficients of weekly text data linearly enter the univariate regression and then are estimated through ordinary least square. Namely,

$$y_t = c + \sum_{j=1}^{p} \left\{ X_{t-j} B_j + \sum_{k=0}^{3} \Theta_{k,j} X_{t-k/4-j}^{(w)} \right\} + \varepsilon_t, \qquad \varepsilon \sim N\left(0, \sigma^2\right)$$

For expository reasons, only the outcome of MIDAS regressions with weighting scheme estimated through the Beta function is reported in the main section of empirical results. However, the reader is directed to consult Appendix B.3, where the alternative parametrization methods are examined.

3.4.3 MF-VAR

By assuming that oil prices still evolve at a weekly frequency, an alternative dynamic representation can be outlined through the following MF-VAR:

$$y_t = c + \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim iid\, N(0,\, \Sigma) \tag{3.9}$$

where $y_t = \left[y'_{w,t}\,,\, y'_{m,t}\right]'$ is the $n \times 1$ vector of mixed-frequency observations, in which $y_{w,t}$ includes the weekly observable data, such as text variables discussed in section 3.3. $y_{m,t}$ collects the remaining monthly variables, such as oil production, global real economy and oil inventories. c is the vector of intercepts, Φ is the matrix of unknown coefficients and ε is the error term.

The advantage of this methodology is that there are no missing observations that need to be estimated as in Durbin and Koopman [2012]. In fact, by denoting with $y_t^{(d)}$ the vector of daily data retrieved from oil related articles, weekly data can be stacked into four different vectors $y_{w,t} = \left[y_t^{w,\,1},\, y_t^{w,\,2},\, y_t^{w,\,3},\, y_t^{w,\,4}\right]$ based on the daily release of each news item. Φ and Σ are estimated in a frequentist and Bayesian approach. All variables are taken in log-differences, other than oil prices which are considered in log-levels, and lags p are set equivalent to 12[10]. The out-of-sample forecasts of real oil prices for $h = 1,\, 3,\, 6,\, 12$ and 24 months ahead are performed by iterating the estimated BVAR models recursively, conditionally on the date t of the information set. Empirical results are then compared to the outcome generated by the corresponding BVAR model with variables all sampled at a monthly frequency.

3.5 Empirical Results

The empirical analysis begins by comparing two classes of univariate models, such as autoregressive (AR) and MIDAS models. After that, I estimate different VARs in a frequentist and Bayesian fashion. For the former methodology, VARs are estimated through standard ordinary least square, Ridge, Lasso and Elastic-Net regression techniques. For the latter I firstly use a standard Bayesian shrinkage, and then I also estimate

[10]BVAR with the aforementioned data transformation, as well as number of lags has been proven to generate superior out-of-sample forecasts (see Baumeister and Kilian [2015] for additional details)

the stochastic system by assuming the heteroskedasticity in the error structure. Finally, I estimate different MIDAS models to assess the contribution of text data for forecasting the price of oil.

3.5.1 HF vs. MF Models

The first part of the analysis presents a comparison across a broad set of univariate and multivariate models, both used to forecast the monthly real oil prices. Results are reported in table 3.1 and are based on the ratio between the minimum sum of prediction errors (MSPEs) of models specified in column 1, over the MSPEs of a no-change forecast. Panel A includes models that incorporate oil market fundamentals as explanatory variables, while Panel B, C, D and E include models that also accommodate the weekly text variables[11] computed in section 3.3. The evidence shows that on average Bayesian estimation and text based VARs generate lower MSPEs. In particular, for a short-run forecast, MF models (i.e. MF-SVBVAR-TXT) outperform the corresponding models having variables sampled at the same frequency. However, for medium- and long-term forecasts, SVBVAR-TXT yield the lowest MSPEs (the only exception is MF-SVBVAR-TXT, when text series are generated through the VADER dictionary model). In this exercise, there is thus evidence in favour of Baumeister et al. [2015]. Namely, mixed frequency models slightly improve a 1-step ahead forecast of monthly oil prices. This improvement is however negligible, as SVBVAR-TXT models on average yield 18% of marginal gains in comparison to a no-change forecast, while MF-SVBVAR-TXT show 19% of marginal gains. In the medium and long term, monthly based models always yield lower MSPEs. It is worth stressing that MIDAS models in table 3.1 incorporate text data collected on a weekly frequency and the Beta lag function is used to estimate the unknown coefficients. This exercise is also replicated with different weighting schemes. Details are provided in Appendix B.3.

3.5.2 High-Frequency Data

Several works in the literature have tried to explain the behaviour of monthly real oil prices by using high-frequency financial data (Sari et al. [2011], Baumeister et al. [2015], Miao et al. [2017] and Degiannakis and Filis [2018]). Degiannakis and Filis [2018] have concluded that ultra-high frequency financial data have the potential to generate more accurate real oil price forecasts. In contrast, no robust evidence has been found by Baumeister et al. [2015] regarding the possibility to improve the out-of-sample forecasts of real oil prices through high-frequency financial data. However, in both empirical works the performance of each

[11]The title of each panel reports the nature of text variable used to forecast the price of oil.

forecasting model is tested by using a limited time horizon not covering the most recent economic/political changes. It is well know that after 2010, oil prices became harder to forecast. Therefore, a good forecasting model might be challenged if the dataset was extended to include the most recent data. Remember that the Covid pandemic and the war between Russia and Ukraine have increased the number of volatility spikes in the oil market significantly.

In this section I use MF models that incorporate weekly financial variables that have been shown to outperform traditional models commonly used to forecast the price of oil. MF models are tested on a larger time-span, and each model is then compared to the case in which a text variable is also included. This experiment starts by using oil future prices as a medium to forecast the monthly price of oil. As pointed out in Alquist and Kilian [2010], in the absence of a risk premium, arbitrage implies that the oil futures price is the conditional expectation of the spot price of oil. This implies that

$$E_t\left(\Delta s_{t+h}\right) = f_{t,h} - s_t \tag{3.10}$$

where s_t is the WTI spot price and f_h is the futures prices with maturity on month h. The spread reported in equation (3.10) enters in the MIDAS regression (3.8) as a polynomial in $Fin_{t,h,w} = f_{t,h,w} - s_{t,h,w}$ where w refers to week 1, 2, 3, 4, and/or 5 of a given month t. Equation (3.8) is therefore rewritten as:

$$y_t = c + \sum_{j=1}^{p} \left\{ y_{t-j}B_j + Fin_{t-j}^{(w)}\Theta_j \left[D\left(L^{1/w};\tau\right)\right]\right\} + \varepsilon_t, \tag{3.11}$$

and it is compared to the case in which also text data enter the linear regression:

$$y_t = c + \sum_{j=1}^{p} \left\{ y_{t-j}B_j + \left(Fin_{t-j}^{(w)}\Theta_j + Txt_{t-j}^{(w)}\Gamma_j\right) \left[D\left(L^{1/w};\tau\right)\right]\right\} + \varepsilon_t, \tag{3.12}$$

Results in Table 3.2 show that, for a 1-month ahead oil price forecast, weekly oil futures yield 9% of marginal improvement, which is also 10% statistically significant according to the Diebold and Mariano (DM) test. However, when text data also fit the polynomial, there is not only a marginal improvement equivalent to 11%, but results are even 5% statistically significant when Bert series is included in the polynomial. For any other time horizons $h > 1$ month, random walk generates lower MSPEs.

In a second experiment, I follow the literature aiming to explain the oil price volatility through petroleum product data (Karrenbrock et al. [1991] , Borenstein et al. [1997], Bachmeier and Griffin [2003]). Specifically, I use the spread between gasoline price and WTI spot price to forecast the monthly crude oil spot price. Gasoline data are available since April 1990. This implies that equations (3.11) and (3.12) are trained until October 2014 and tested on the remaining period up to December 2021. Results show that also in this experiment weekly observations of gasoline-crude oil spread generate a 10% statistically significant gain for short-run forecasts. However, there is still evidence of more accurate forecasts when weekly text data are used alongside this commodity variable. In particular, Bert-MIDAS yields 17% of marginal gains for 1-month ahead forecast. This result is 5% statistically significant. As in the previous experiment, the random walk again outperforms the forecasts of text and no-text based MIDAS at time horizons $h > 1$ month.

Panel C, in Table 3.2, explores the case in which the Commodity Research Bureau (CRB) index is used to forecast the oil spot prices. The CRB index is based on the arithmetic average of 19 commodity futures prices and is designed to capture the directional movement of such industrial commodities. The main idea behind the use of this indicator is that variations of non-oil industrial raw materials Granger cause variations in the crude oil prices (see Alquist et al. [2013], Baumeister and Kilian [2012], Baumeister et al. [2015] and more recently Aastveit et al. [2022]). In this experiment there is stronger evidence in favour of using text data. In fact, when only the CRB index fits the MIDAS model, no marginal gain is achieved at any time horizon h. In contrast, when this indicator is included along with text based series, there is a marginal improvement between 1% and 18% up to 3-months ahead forecast. Outstanding results are achieved with Bert, where 1-month ahead forecasts are even 1% statistically significant as suggested by the DM test.

In another experiment I examine the long-term relationship between the weekly Baltic Dry (BD) index and the monthly real oil prices. The BD indicator measures the daily change in the cost of moving alternative dry bulk cargoes containing raw materials, such as coal, steel, grain, iron and other commodities. As such, this composite index can be considered as a good approximation of the economic activity in real time, and thus a good predictor of future world economy (see Bakshi et al. [2010]). The evidence shows that with the BD series, MIDAS beat the random walk only when oil prices are forecast one month ahead. Results are 10% statistically significant. However, including text data does generate up to 10% of marginal gains.

The last exercise builds on the dated diatribe between some economists assessing that high crude oil prices raise real interest rates (Barro and Sala-i Martin [1990], Barro [1991]), and some others showing that oil prices rise significantly following a reduction of interest rates (Hotelling [1931], Frankel [1986]). Both cases imply a relationship either positive or negative between monetary policy and crude oil prices. This has motivated my choice to investigate the correlation between the weekly US short-term interest rates and crude oil spot prices. Results are displayed in panel E and the evidence shows that USi-MIDAS outperforms the random walk when oil prices are forecast one month ahead. However, it is still worth fitting a MIDAS with text data. In particular, Bert series yields 15% of marginal gains and results are also 5% statistically significant according to the DM test.

In summary, this section demonstrates that for a 1-month ahead forecast, it is possible to outperform a random walk by fitting a MIDAS with weekly financial and commodity data. However, Bert based SVBVARs generate lower MSPEs in comparison to MIDAS models at any time horizon (see Table 3.1). Nevertheless, if text data fit a MIDAS alongside a weekly financial or commodity factor, there is not only a marginal gain up to 18% for $h = 1$- and 3-months ahead forecast, but results are even 1%, 5% and 10% statistically significant as suggested by the DM test. Furthermore, when Bert and the CRB index are incorporated into a MIDAS, the MSPEs of 1-month ahead forecast is lower than SVBVAR-TXT's, with 2% of marginal gains. Thus, it can be assessed that for short-term forecasts MIDAS models can be slightly preferred to SVBVARs, but for longer horizons SVBVARs yield lower forecasting errors.

3.5.3 Financial Intra-Day Returns

The increased availability of high-frequency data has also motivated a wide range of researchers to use intraday data in order to improve the in- and out-of-sample volatility forecasts. For example, Andersen and Bollerslev [1998] and Martens [2001] demonstrate that intraday returns can be used to construct well performing ex-post interdaily volatility measurements. Such factors are shown to provide a radical reduction in noise, and more accurate daily volatility forecasts of alternative exchange rates. Sévi [2014] and Ghysels et al. [2006] exploits the information in intraday data to forecast the volatility of crude oil futures and the Dow Jones Index. Furthermore, there are also several studies showing that the covariance between crude oil prices and country-specific stock markets is affected by geopolitical shocks (Kollias et al. [2013], Aloui and Jammazi [2009], Mugaloglu et al. [2021]), and financial data yield more efficient oil price forecasts (Nguyen and Walther [2020]).

By departing from this literature, I use high-frequency financial and commodity data to construct intraday returns and forecast the monthly WTI crude oil prices. The analysis starts by fitting a MIDAS with intraday returns of FTSE 100, which is a share index of the top 100 companies by market capitalisation that trade on the London Stock Exchange. The resulting model is referred to as FT-IR-MIDAS. Data are considered on a weekly average, and results are compared to the case in which alternative text variables are included into

FT-IR-MIDAS model. Results are displayed in Table 3.3 and are consistent with the previous experiments. Indeed, despite a 10% of significant improvement on the random walk for 1-month ahead forecast, overall text data provide lower MSPEs. In particular, Bert-IR-MIDAS beats FT-IR-MIDAS at any time horizon with a marginal improvement up to 17%.

By applying a similar approach I also use the intraday returns of S&P 500, the Euro Stoxx 50 index and the Hang Seng index, which are the market-capitalization-weighted indexes of the largest public companies of the US, Europe and Hong Kong respectively. Said indicators are included into a MIDAS and results are still consistent with the previous exercises. Namely, using Bert alongside a financial index generates a radical improvement on a no-text based model. The lowest MSPEs are achieved in 1-step ahead forecasts by using Bert and the Euro Stoxx 50 index. Such results are also 1% statistically significant.

Another popular approach widely used among many empiricists is to forecast the real oil prices through alternative currency exchange rates. Some observers have in fact noted a divergent, but existent reaction of foreign exchange markets to oil price fluctuations (Golub [1983], Akram [2004], Ferraro et al. [2015]). This has motivated my interest to determine whether text data and the intraday returns of the most liquid trading pairs can fit a MIDAS and improve the out-of-sample forecasts of real oil prices. The trading pairs investigated are GBP/USD, EUR/USD and CAD/USD. Based on Table 3.3 it is possible to observe that intraday returns of GBP/USD generate the most accurate forecasts with a marginal improvement of 12% in the short term. Once again, Bert series radically improves the financial based MIDAS forecasting performance for one and three months ahead. The marginal improvement is equivalent to 18% with a statistical significance of 1%. It is worth pointing out that GBP/USD and Bert based MIDAS overall generate the lowest MSPE ratios when the oil prices are forecast one month ahead. Table 3.3 also shows that improvements on a random walk can only be observed up to $h = 3$-months ahead forecast.

3.5.4 Commodity Intra-Day Returns

Following the previous experiment, in this section MIDAS models are fit by intraday returns of alternative commodity market prices. This procedure still aims to reduce the noise that daily spot prices incorporate, by employing the additional information that intraday returns enclose. Sévi [2014], Degiannakis and Filis [2018] and Degiannakis and Filis [2022] are examples of studies that advocate the use of intraday returns of commodity prices as a strategy to improve the out-of-sample forecasts. Intraday returns are computed for: WTI crude oil, gold, copper, natural gas, palladium and silver. Results are reported in Table 3.4. As per in the previous experiments, the evidence shows that for 1-month ahead forecast, MIDAS with exogenous intraday returns outperform the random walk. The outcome is 10% statistically significant, with a marginal improvement up to 9%. However, results are further improved when text based series are used alongside commodity intraday returns. In particular, including Bert in NG-IR-MIDAS or PL-IR-MIDAS improves the results significantly. Despite the valuable contribution of text data and intraday returns, for medium- and long-term forecasts MIDAS tend to have higher MSPEs in comparison to a random walk.

Based on the empirical findings provided above, the following can be assessed. High frequency financial and commodity data can be modelled and included into mixed data sampling models in order to forecast the price of oil. However, when MIDAS are tested over the recent periods of high volatility (the Covid pandemic or in general from 2010 onwards), a marginal improvement on a no-change forecast is observable only for 1-month ahead forecast. In any case, such results are still weak in comparison to the outcome achieved by using SVBVARs. A potential workaround is to incorporate weekly text based data into a MIDAS. In particular, for 1-month ahead forecast, using Bert alongside other high-frequency indicators (i.e. CRB index, GBP/USD exchange rate, natural gas spot price) yields a marginal improvement on both SVBVAR and MF-SVBVAR. But for time horizons greater than one month, SVBVARs have on average lower MSPEs, reinforcing the conclusion of Baumeister et al. [2015] about the negligible contribution of high-frequency data when used to forecast the monthly real price of oil.

3.5.5 Sensitivity Analysis

Despite the high frequency availability of data, so far observations have been modelled on a weekly frequency and then included into alternative MF models. I now show that the main results are still robust to a number of modifications and experiments.

Daily Predictors Fitting the MIDAS.

A rational exercise is to fit MIDAS models by using daily observations. Equation (3.8) then becomes:

$$y_t = c + \sum_{j=1}^{p} \left\{ X_{t-j} B_j + txt_{t-j}^{(d)} \Theta_j \left[D\left(L^{1/d}; \tau \right) \right] \right\} + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \sigma^2\right) \tag{3.13}$$

where, $txt_{t-j}^{(d)} = \left[x_t^{d,1}, x_t^{d,2}, \ldots, x_t^{d,n} \right]$ stacks the daily text-data released in month $t-j$, n is the monthly length and the lag operator and the weighting scheme are $L^{1/d}$ and $D\left(L^{1/d}; \tau\right)$ respectively. Based on this, I replicate the previous exercises and use daily data to forecast monthly WTI, RAC and Brent crude oil prices. Results are reported in Table 3.5 and, like in the previous experiments, the evidence shows that the CRB index used alongside the Bert series yields the highest marginal gains for each oil price measure, with a statistical significance between 1% and 5%. It is worth noting that for RAC and Brent price forecasts, the MSPEs are even lower than monthly based VAR's (see Table 3.1, panel E). It is also possible to observe additional important points. Firstly, incorporating the CRB index and text data into a MIDAS yields the lowest MSPEs up to three months ahead. Secondly, and perhaps most importantly, daily observations make a remarkable contribution to medium- and long-term forecasts. This evidence diverges from that of weekly data, where MIDAS do not beat the no-change forecast for time horizons longer than three months.

Oil Prices Density Forecasts.

Forecasting densities has the advantage of providing information on the likelihood of any future quantile, as well as on point forecasts through the mean of density forecasts. This strategy also takes into account the uncertainty that is pervasive in point forecasts[12]. As previously remarked, one of the goals of this paper is to show that low frequency based models incorporating high-frequency text data yield negligible improvements of monthly real oil price forecasts. For this reason, in this section I evaluate the density forecasts in order to ensure the reliability of the empirical findings discussed so far. In particular, I compare the forecasting outcome of SVBVARs and MF-SVBVARs, as well as MIDAS with and without weekly text variables.

The quality of density forecasts is evaluated through the averages of log predictive likelihoods (ALPLs). Higher values of ALPLs indicate a more accurate forecast. Table 3.6 reports the ALPLs for the subset of the best performing models at each time horizon h. Results are consistent with the outcome shown in the previous experiments. Namely, including the CRB index alongside text data yields the highest ALPLs among any MIDAS model. However, by observing the BVARs outcome, there is evidence that the improvement of MF models on SVBVAR-TXT is negligible for each oil price measure. Indeed, SVBVAR with dictionary or Bert based variables are always the best performing models, except for horizon 12 where MF-SVBVAR-TXT yield slightly higher, but still imperceptible, ALPLs.

Therefore, according to the density forecasts it is possible to conclude that high-frequency text data provide a remarkable contribution to MIDAS for monthly oil price forecasts. However, this improvement becomes negligible in comparison to the performance of a low frequency based SVBVAR-TXT model. The latter does in fact generate the highest log predictive likelihoods at almost every time horizon.

[12] See Rossi [2014] for a more practical illustration of point and density forecasts.

3.6 Conclusions

Data available or combined on a monthly basis are unable to provide an accurate reflection of global, specific economic events. Despite their infrequency, occurrences like natural disasters and political unrest in oil-producing countries have the potential to affect the price of oil and bring about profound economic changes. A natural question, therefore, is whether combining high-frequency data alongside the oil market fundamentals can improve the forecast of monthly real oil prices. In this paper I answer this question by using text mining strategies and mixed-frequency models.

Empirical findings show that MF-SVBVAR accommodating weekly text data yields more accurate 1-step ahead forecasts in comparison to the corresponding model with variables observed at the same frequency. The best performing text indicator is Bert. However, this improvement is low. Indeed, for Brent crude spot price, MSPEs are improved by 2%, while for WTI and RAC marginal gains are even lower than 1%. In a second experiment, I follow Degiannakis and Filis [2018] by fitting alternative MIDAS models with high-frequency financial and commodity variables. I show that financial and commodity variables alone do not generate significant improvements on a no-change forecast. In contrast, when text data are used along with commodity variables the MSPEs are reduced significantly. The best performing model is a MIDAS including oil market fundamentals, available on a monthly basis, and weekly observations of the CRB index and Bert series. The latter model displays the lowest 1-month ahead MSPEs among any model investigated in this study. However, such improvement is marginal and negligible. Indeed, no statistically significant gain is achieved on the corresponding model based on variables available at a low frequency.

I thus conclude that despite a marginal improvement, on average, combining text, financial and commodity variables along with oil market fundamentals, does not significantly improve the out-of-sample forecasts of monthly real oil prices. This is true for point and density forecasts. A natural extension of this research is to investigate whether daily text-based information is able to improve the predictability of weekly oil prices. This task is left for future research.

Table 3.1: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices. Homogeneous vs. mixed frequency models.

Model	Freq	1-month			3-months			6-months			12-months			24-months		
		WTI	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*
A. No text-based models																
AR	HF	**0.932**	**0.805****	**0.954**	1.022	**0.973**	1.039	1.073	1.056	1.088	1.086	1.095	1.105	1.166	1.173	1.154
AR-X	HF	**0.862****	**0.760*****	**0.877****	**0.997**	**0.962**	**0.948**	**0.946**	**0.946**	**0.972**	**0.975**	1.022	1.033	1.161	1.164	1.162
VAR	HF	1.015	**0.824***	1.061	1.115	1.038	1.177	1.100	1.124	1.195	1.288	1.360	1.341	1.303	1.291	1.263
Ridge	HF	1.032	**0.864**	1.065	1.070	1.011	1.136	1.016	1.046	1.110	1.168	1.246	1.234	1.187	1.205	1.184
Lasso	HF	1.848	2.042	2.153	2.005	2.265	2.448	1.906	2.185	2.421	1.880	2.175	2.251	1.748	1.796	1.784
ElasticNet	HF	2.090	2.373	2.462	2.086	2.371	2.525	1.925	2.207	2.418	1.877	2.175	2.234	1.749	1.797	1.781
BVAR	HF	**0.898**	**0.752*****	**0.912**	**0.938**	**0.897**	**0.977**	**0.994**	**0.999**	1.032	1.087	1.125	1.120	1.195	1.216	1.191
SV-BVAR	HF	**0.884****	**0.776*****	**0.893****	**0.915**	**0.864**	**0.925**	**0.933**	**0.913**	**0.930**	1.007	1.132	1.019	**0.934**	**0.837**	**0.908**
B. Unigram based variable																
AR-X-TXT	HF	**0.863****	**0.761*****	**0.882****	1.014	**0.961**	**0.950**	**0.956**	**0.956**	**0.981**	**0.993**	1.035	1.049	1.049	1.042	1.045
Ridge	HF	1.030	**0.863**	1.063	1.071	1.012	1.143	1.017	1.045	1.108	1.166	1.250	1.232	1.183	1.203	1.182
Lasso	HF	1.885	2.057	2.190	2.022	2.277	2.482	1.901	2.185	2.430	1.853	2.162	2.227	1.722	1.776	1.762
ElasticNet	HF	2.131	2.388	2.502	2.100	2.381	2.556	1.916	2.203	2.422	1.851	2.163	2.211	1.724	1.778	1.760
MIDAS-TXT	MF	**0.902***	**0.765*****	**0.931***	**0.959**	**0.992**	**0.970**	**0.953**	**0.949**	**0.980**	**0.980**	1.016	1.047	1.101	1.101	1.092
BVAR-TXT	HF	**0.896**	**0.765***	**0.908**	**0.928**	**0.903**	**0.962**	**0.998**	1.019	1.039	1.083	1.126	1.115	1.134	1.154	1.118
MF-BVAR-TXT	MF	0.919	**0.790*****	**0.945**	**0.950**	**0.923**	**0.986**	1.019	1.028	1.056	1.077	1.104	1.107	1.131	1.149	1.125
SV-BVAR-TXT	HF	**0.887****	**0.770*****	**0.904****	**0.926**	**0.892**	**0.946**	**0.956**	**0.975**	**0.986**	1.044	1.218	1.100	**0.936**	**0.878**	**0.929**
MF-SV-BVAR-TXT	MF	**0.898****	**0.778*****	**0.919***	**0.945**	**0.894**	**0.966**	**0.983**	**0.978**	1.012	1.075	1.234	1.080	**0.935**	**0.880**	**0.914**
C. Dictionary based variable																
AR-X-TXT	HF	**0.850****	**0.757*****	**0.879***	1.013	**0.988**	**0.968**	**0.963**	**0.962**	**0.987**	**0.980**	1.012	1.027	1.050	1.049	1.048
Ridge	HF	1.017	**0.869**	1.079	1.056	1.003	1.142	1.013	1.035	1.109	1.176	1.251	1.234	1.087	1.105	1.083
Lasso	HF	1.819	2.013	2.123	1.962	2.220	2.406	1.859	2.134	2.367	1.847	2.148	2.214	1.718	1.771	1.755
ElasticNet	HF	2.058	2.339	2.426	2.043	2.327	2.484	1.879	2.159	2.367	1.847	2.152	2.200	1.721	1.774	1.755
MIDAS-TXT	MF	**0.896***	**0.759*****	**0.926**	**0.955**	**0.986**	**0.965**	**0.960**	**0.956**	**0.996**	**0.988**	1.013	1.038	1.168	1.161	1.155
BVAR-TXT	HF	**0.877***	**0.748*****	**0.903**	**0.926**	**0.892**	**0.965**	**0.994**	**0.990**	1.027	1.079	1.104	1.102	1.129	1.134	1.122
MF-BVAR-TXT	MF	**0.890***	**0.767*****	**0.902***	**0.927**	**0.897**	**0.955**	1.002	1.002	1.027	1.068	1.092	1.094	1.131	1.139	1.124
SV-BVAR-TXT	HF	**0.860****	**0.763***	**0.887****	**0.894**	**0.861**	**0.925**	**0.931**	**0.930**	**0.940**	1.053	1.284	1.112	**0.908**	**0.833**	**0.893**
MF-SV-BVAR-TXT	MF	**0.859****	**0.751***	**0.871****	**0.892**	**0.846**	**0.920**	**0.950**	**0.940**	**0.947**	1.221	1.518	1.155	**0.898**	**0.848**	**0.897**
D. Geometrical based variables																
(tfm) AR-X-TXT	HF	**0.854****	**0.755***	**0.866****	**0.993**	**0.958**	**0.942**	**0.944**	**0.943**	**0.966**	**0.937**	**0.979**	**0.997**	1.142	1.143	1.143
(idf) AR-X-TXT	HF	**0.862****	**0.760***	**0.870****	1.002	**0.944**	**0.935**	**0.939**	**0.933**	**0.954**	**0.936**	**0.960**	**0.981**	1.015	**0.993**	**0.996**
(tfm) Ridge	HF	1.040	**0.873**	1.069	1.060	1.007	1.136	**0.984**	1.020	1.085	1.113	1.194	1.185	1.166	1.188	1.178
(idf) Ridge	HF	1.028	**0.866**	1.060	1.066	1.001	1.128	1.017	1.047	1.096	1.172	1.243	1.209	1.041	1.009	1.007
(tfm) Lasso	HF	1.831	2.017	2.121	1.979	2.239	2.431	1.869	2.148	2.398	1.841	2.140	2.219	1.732	1.780	1.771
(idf) Lasso	HF	1.812	2.003	2.119	1.932	2.190	2.388	1.815	2.093	2.333	1.798	2.096	2.166	1.682	1.732	1.722
(tfm) ElasticNet	HF	2.072	2.344	2.427	2.059	2.344	2.508	1.888	2.170	2.394	1.840	2.142	2.204	1.733	1.782	1.768
(idf) ElasticNet	HF	2.050	2.327	2.424	2.008	2.291	2.459	1.832	2.115	2.329	1.799	2.099	2.153	1.686	1.737	1.723
(tfm) MIDAS-TXT	MF	**0.906****	**0.755***	**0.929****	**0.946**	**0.981**	**0.959**	**0.947**	**0.950**	**0.974**	**0.976**	1.016	1.034	1.155	1.158	1.158
(idf) MIDAS-TXT	MF	0.918	**0.761*****	**0.935**	**0.967**	1.005	**0.979**	**0.948**	**0.946**	**0.971**	**0.980**	1.015	1.037	1.160	1.171	1.176
(tfm) BVAR-TXT	HF	**0.887***	**0.749*****	**0.897***	**0.911**	**0.882**	**0.942**	**0.968**	**0.967**	1.003	1.042	1.075	1.076	1.168	1.182	1.162
(idf) BVAR-TXT	HF	**0.879***	**0.744*****	**0.884***	**0.887**	**0.849**	**0.923**	**0.943**	**0.939**	**0.982**	1.045	1.069	1.057	1.106	1.108	1.090
(tfm) MF-BVAR-TXT	MF	**0.911***	**0.779****	**0.913***	**0.958**	**0.918**	**0.974**	1.021	1.019	1.050	1.062	1.082	1.093	1.161	1.169	1.152
(idf) MF-BVAR-TXT	MF	**0.911***	**0.776****	**0.902***	**0.941**	**0.905**	**0.959**	1.018	1.014	1.038	1.064	1.087	1.092	1.147	1.162	1.144
(tfm) SV-BVAR-TXT	HF	**0.885****	**0.767***	**0.886****	**0.894**	**0.849**	**0.914**	**0.918**	**0.904**	**0.922**	1.005	1.263	1.035	**0.941**	**0.847**	**0.917**
(idf) SV-BVAR-TXT	HF	**0.873****	**0.788***	**0.900****	**0.905**	**0.887**	**0.936**	**0.945**	**0.970**	**0.954**	1.084	1.417	1.110	1.015	**0.922**	**0.977**
(tfm) MF-SV-BVAR-TXT	MF	**0.900****	**0.782***	**0.907****	**0.917**	**0.860**	**0.934**	**0.938**	**0.899**	**0.945**	1.033	1.041	**0.984**	**0.941**	**0.852**	**0.926**
(idf) MF-SV-BVAR-TXT	MF	**0.924***	**0.806*****	**0.922***	**0.934**	**0.885**	**0.946**	**0.977**	**0.933**	**0.968**	1.070	1.072	1.047	1.034	1.022	1.075
E. Network based variable																
AR-X-TXT	HF	**0.844****	**0.743***	**0.850****	**0.975**	**0.956**	**0.938**	**0.930**	**0.933**	**0.954**	**0.986**	1.041	1.039	1.195	1.200	1.197
Ridge	HF	**0.961**	**0.826***	1.016	**0.999**	**0.957**	1.098	**0.987**	1.022	1.094	1.185	1.276	1.254	1.223	1.238	1.220
Lasso	HF	1.837	2.048	2.150	1.993	2.257	2.448	1.908	2.182	2.427	1.894	2.193	2.264	1.749	1.796	1.786
ElasticNet	HF	2.079	2.378	2.457	2.074	2.363	2.525	1.926	2.204	2.422	1.891	2.191	2.244	1.749	1.797	1.782
MIDAS-TXT	MF	**0.880****	**0.745***	**0.909***	**0.931**	**0.942**	**0.933**	**0.958**	**0.933**	**0.971**	**0.995**	1.032	1.054	1.184	1.265	1.207
BVAR-TXT	HF	**0.867***	**0.745*****	**0.877***	**0.909**	**0.875**	**0.942**	**0.976**	**0.989**	1.017	1.082	1.113	1.109	1.195	1.209	1.181
MF-BVAR-TXT	MF	**0.868****	**0.751*****	**0.888***	**0.920**	**0.891**	**0.948**	**0.989**	**0.993**	1.017	1.069	1.099	1.109	1.158	1.171	1.156
SV-BVAR-TXT	HF	**0.846*****	**0.747*****	**0.868****	**0.875**	**0.830**	**0.890**	**0.897**	**0.876**	**0.890**	**0.950**	1.056	**0.941**	**0.932**	**0.856**	**0.925**
MF-SV-BVAR-TXT	MF	**0.842*****	**0.739*****	**0.848****	**0.885**	**0.840**	**0.900**	**0.929**	**0.890**	**0.919**	1.023	1.072	**0.995**	**0.960**	**0.913**	**0.967**

Note: In column 1 AR: autoregression model, ARX: autoregression augmented with (no text-based) explanatory variables, VAR: vector autoregression, BVAR: Bayesian vector autoregression, SV-BVAR: Bayesian vector autoregression assuming stochastic volatility of the error term, AR-X-TXT, autoregression model augmented with text and no text-based explanatory variables, MIDAS-TXT: mixed frequency model, BVAR-TXT: text-based Bayesian vector autoregression model, MF-BVAR-TXT: mixed frequency text-based vector autoregression, SV-BVAR-TXT: stochastic volatility text-based Bayesian vector autoregression, MF-SV-BVAR-TXT: mixed frequency stochastic volatility text-based Bayesian vector autoregression, tfm: term-frequency matrix, idf: term-frequency inverse-document-frequency matrix. In column 2, HF and MF respectively denote homogeneous and mixed frequency models. Bold values indicate improvements on the no-change forecast. Blue entries stand for the lowest MSPE for a given time horizon, relative to a specific oil price measure. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table 3.2: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with high-frequency financial variables; text vs. no-text.

Monthly	**TXT-FIN-MIDAS**					**FIN-MIDAS**
horizon	Uc-MIDAS	Dc-MIDAS	idx-MIDAS	idf-MIDAS	Bert-MIDAS	OFs-MIDAS
1	**0.903***	**0.907***	**0.907***	**0.901***	**0.887****	**0.906***
3	1.079	1.079	1.094	1.090	1.040	1.082
6	1.105	1.112	1.113	1.105	1.089	1.101
12	1.168	1.151	1.197	1.193	1.192	1.186
24	1.525	1.705	1.545	1.605	1.560	1.577
						GLs-MIDAS
1	**0.899****	**0.868****	**0.909****	**0.910****	**0.835****	**0.911****
3	1.078	1.067	1.110	1.109	0.997	1.102
6	1.139	1.141	1.178	1.158	1.072	1.147
12	1.213	1.239	1.221	1.212	1.178	1.212
24	1.481	1.504	1.462	1.450	1.505	1.430
						CRB-MIDAS
1	**0.856****	**0.850****	**0.861****	**0.868****	**0.821*****	1.718
3	**0.946**	**0.962**	**0.994**	**0.918**	**0.928**	2.566
6	**0.988**	1.002	**0.993**	1.030	**0.964**	1.538
12	1.078	1.113	1.093	1.088	1.089	1.338
24	1.379	1.425	1.481	1.494	1.516	6.618
						BDi-MIDAS
1	**0.908***	**0.908***	**0.909***	**0.919***	**0.898***	**0.914***
3	1.067	1.078	1.077	1.094	1.036	1.087
6	1.151	1.137	1.152	1.107	1.149	1.107
12	1.250	1.321	1.316	1.285	1.281	1.287
24	1.399	1.342	1.354	1.352	1.346	1.342
						USi-MIDAS
1	**0.904***	**0.896***	**0.908***	**0.912***	**0.853****	**0.909***
3	1.084	1.059	1.051	1.075	**0.983**	1.046
6	1.069	1.081	1.066	1.062	1.022	1.067
12	1.080	1.098	1.076	1.078	1.079	1.078
24	1.461	1.515	1.487	1.518	1.537	1.512

Note: For column headers OFs-MIDAS, GLs-MIDAS, CRB-MIDAS, BDi-MIDAS, USi-MIDAS denote MIDAS models where the high-frequency financial variables fitting the polynomial are (i) crude oil-futures prices spread, (ii) crude oil-gasoline spread, (iii) CRB spot price index, (iv) Baltic Dry index and (v) the federal funds rate. Each outcome is then compared to the case in which text data are included in the model, in addition to the financial variable. In particular, Uc-MIDAS, Dc-MIDAS, idx-MIDAS, idf-MIDAS Bert-MIDAS denote MIDAS models where the text variable fitting the polynomial is respectively developed through (i) unigram word-count, (ii) dictionary, (iii) term-frequency matrix, (iv) term-frequency inverse-document frequency matrix and (v) BERT. Black bold values indicate improvements on the no-change forecast. Green bold values indicate improvements of TXT-FIN-MIDAS on FIN-MIDAS. Blue entries stand for the lowest MSPE for a given time horizon. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table 3.3: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with realized volatility of ultra-high-frequency financial variables; text vs. no-text.

Monthly	**TXT-FIN-MIDAS**					**FIN-MIDAS**
horizon	Uc-IR-MIDAS	Dc-IR-MIDAS	idx-IR-MIDAS	idf-IR-MIDAS	Bert-IR-MIDAS	FT-IR-MIDAS
1	**0.883****	**0.870****	**0.879****	**0.886****	**0.830***	**0.884***
3	1.064	1.066	1.059	1.085	**0.986**	1.072
6	1.073	1.085	1.081	1.074	1.015	1.074
12	1.091	1.127	1.090	1.085	1.077	1.080
24	1.339	1.405	1.376	1.366	1.370	1.373
						SP-IR-MIDAS
1	**0.886***	**0.868***	**0.873***	**0.892***	**0.832****	**0.876***
3	1.047	1.042	1.049	1.052	**0.987**	1.050
6	1.066	1.080	1.071	1.066	1.013	1.063
12	1.088	1.133	1.068	1.085	1.087	1.083
24	1.356	1.405	1.447	1.418	1.483	1.428
						ES-IR-MIDAS
1	**0.873****	**0.868****	**0.875****	**0.880****	**0.821*****	**0.876****
3	1.056	1.051	1.069	1.076	**0.984**	1.071
6	1.123	1.112	1.148	1.111	1.025	1.111
12	1.146	1.192	1.176	1.136	1.097	1.139
24	1.360	1.456	1.397	1.336	1.420	1.352
						HS-IR-MIDAS
1	**0.887***	**0.875****	**0.887***	**0.895***	**0.829****	**0.889***
3	1.051	1.055	1.048	1.073	**0.988**	1.068
6	1.107	1.104	1.082	1.092	1.052	1.091
12	1.098	1.102	1.094	1.092	1.095	1.098
24	1.433	1.471	1.488	1.520	1.584	1.494
						PD-IR-MIDAS
1	**0.879****	**0.865****	**0.880****	**0.881****	**0.820*****	**0.882****
3	1.026	1.020	1.042	1.048	**0.953**	1.036
6	1.066	1.078	1.065	1.072	1.028	1.062
12	1.143	1.192	1.149	1.148	1.162	1.141
24	1.563	1.676	1.590	1.609	1.684	1.601
						CD-IR-MIDAS
1	**0.929***	**0.893****	**0.918****	**0.919***	**0.853*****	**0.916****
3	1.184	1.181	1.182	1.218	1.076	1.205
6	1.397	1.299	1.361	1.502	1.328	1.419
12	2.121	1.441	1.378	1.368	1.263	1.387
24	3.873	2.410	3.113	3.951	4.433	6.278
						ED-IR-MIDAS
1	**0.886****	**0.873****	**0.884****	**0.910***	**0.835****	**0.904***
3	1.039	1.016	1.035	1.017	**0.964**	1.020
6	1.067	1.075	1.064	1.081	1.019	1.060
12	1.126	1.163	1.114	1.118	1.124	1.113
24	1.489	1.451	1.539	1.542	1.613	1.543

Note: For column headers FT-IR-MIDAS, SP-IR-MIDAS, ES-IR-MIDAS, HS-IR-MIDAS, PD-IR-MIDAS, CD-IR-MIDAS, ED-IR-MIDAS denote MIDAS models where the ultra-high-frequency financial variables fitting the polynomial are (i) intraday returns of FTSE100 index , (ii) intraday returns of S&P500 index, (iii) intraday returns of Euro Stoxx 50 index, (iv) intraday returns of Hang Seng index, (v) intraday returns of GBP/USD exchange rate, (vi) intraday returns of CAD/USD exchange rate and (vii) intraday returns of EUR/USD exchange rate. Each outcome is then compared to the case in which text data are included in the model, in addition to the financial variable. In particular, Uc-MIDAS, Dc-MIDAS, idx-MIDAS, idf-MIDAS Bert-MIDAS denote MIDAS models where the text variable fitting the polynomial is respectively developed through (i) unigram word-count, (ii) dictionary, (iii) term-frequency matrix, (iv) term-frequency inverse-document frequency matrix and (v) BERT. Black bold values indicate improvements on the no-change forecast. Green bold values indicate improvements of TXT-FIN-MIDAS on FIN-MIDAS. Blue entries stand for the lowest MSPE for a given time horizon. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table 3.4: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with realized volatility of ultra-high-frequency commodity prices; text vs. no-text.

Monthly	**TXT-CMDTY-MIDAS**					**CMDTY-MIDAS**
horizon	Uc-IR-MIDAS	Dc-IR-MIDAS	idx-IR-MIDAS	idf-IR-MIDAS	Bert-IR-MIDAS	CO-IR-MIDAS
1	**0.902***	**0.897***	**0.894***	**0.909***	**0.866****	**0.895***
3	1.103	1.104	1.091	1.077	1.135	1.081
6	1.070	1.074	1.102	1.061	1.097	1.063
12	1.059	1.040	1.079	1.060	1.079	1.057
24	1.241	1.425	1.289	1.278	1.295	1.280
						GL-IR-MIDAS
1	**0.893***	**0.878****	**0.893****	**0.898***	**0.832****	**0.896***
3	1.053	1.064	1.058	1.079	**0.977**	1.085
6	1.061	1.076	1.066	1.074	1.009	1.061
12	1.090	1.096	1.064	1.101	1.066	1.087
24	1.281	1.345	1.374	1.426	1.434	1.394
						CP-IR-MIDAS
1	**0.927***	**0.930***	**0.940***	**0.936***	**0.923****	**0.941***
3	1.120	1.158	1.139	1.106	1.144	1.135
6	1.097	1.090	1.103	1.118	1.134	1.104
12	1.101	1.143	1.079	1.101	1.094	1.172
24	1.267	1.424	1.315	1.305	1.292	1.316
						NG-IR-MIDAS
1	**0.913***	**0.882****	**0.918***	**0.920***	**0.821*****	**0.922***
3	1.078	1.041	1.106	1.149	**0.933**	1.148
6	1.163	1.143	1.198	1.207	1.020	1.159
12	1.169	1.239	1.197	1.172	1.120	1.165
24	1.467	1.704	1.402	1.403	1.485	1.402
						PL-IR-MIDAS
1	**0.907****	**0.876****	**0.901****	**0.910***	**0.825*****	**0.914***
3	1.056	1.049	1.081	1.089	**0.951**	1.081
6	1.143	1.123	1.189	1.115	1.002	1.121
12	1.173	1.190	1.178	1.149	1.086	1.155
24	1.429	1.560	1.420	1.428	1.537	1.411
						SL-IR-MIDAS
1	**0.892***	**0.880***	**0.891***	**0.898***	**0.833****	**0.893***
3	1.066	1.064	1.073	1.085	**0.989**	1.088
6	1.075	1.088	1.052	1.071	1.010	1.059
12	1.088	1.107	1.073	1.074	1.087	1.074
24	1.253	1.512	1.388	1.383	1.435	1.393

Note: For column headers CO-IR-MIDAS, GL-IR-MIDAS, CP-IR-MIDAS, NG-IR-MIDAS, PL-IR-MIDAS, SL-IR-MIDAS, denote MIDAS models where the ultra-high-frequency commodity variables fitting the polynomial are (i) intraday returns of WTI index, (ii) intraday returns of Gold index, (iii) intraday returns of Copper index, (iv) intraday returns of Natural Gas index, (v) intraday returns of Palladium index and (vi) intraday returns of Silver index. Each outcome is then compared to the case in which text data are included in the model, in addition to the commodity variable. In particular, Uc-MIDAS, Dc-MIDAS, idx-MIDAS, idf-MIDAS Bert-MIDAS denote MIDAS models where the text variable fitting the polynomial is respectively developed through (i) unigram word-count, (ii) dictionary, (iii) term-frequency matrix, (iv) term-frequency inverse-document frequency matrix and (v) BERT. Black bold values indicate improvements on the no-change forecast. Green bold values indicate improvements of TXT-CMDTY-MIDAS on CMDTY-MIDAS. Blue entries stand for the lowest MSPE for a given time horizon. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table 3.5: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. Sensitivity analysis with daily values fitting the MIDAS models.

Model	Text Variable	Financial Variable	1-month			3-months			6-months			12-months			24-months		
			WTI	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*
TXT-MIDAS	U	-	0.903*	0.767***	0.930*	0.954	0.987	0.966	0.953	0.955	0.981	0.979	1.014	1.031	1.126	1.169	1.095
TXT-MIDAS	D	-	0.895*	0.757***	0.919*	0.949	0.971	0.957	0.965	0.961	0.982	0.984	1.013	1.039	1.149	1.159	1.148
TXT-MIDAS	G	-	0.910*	0.761***	0.929*	0.951	0.977	0.961	0.951	0.949	0.973	0.970	1.003	1.034	1.151	1.167	1.167
TXT-MIDAS	B	-	0.869**	0.725*	0.881**	0.892	0.884	0.882	0.927	0.913	0.938	0.981	1.020	1.032	1.211	1.208	1.209
FIN-MIDAS	-	OFs	1.086	-	1.104	0.981	-	0.979	0.941	-	0.965	0.972	-	1.006	1.350	-	1.312
TXT-FIN-MIDAS	U	OFs	1.034	-	1.065	0.945	-	0.948	0.989	-	0.965	0.979	-	1.017	1.180	-	1.191
TXT-FIN-MIDAS	D	OFs	0.962	-	0.998	0.977	-	0.970	0.969	-	0.992	0.924	-	0.991	1.324	-	1.297
TXT-FIN-MIDAS	G	OFs	1.000	-	1.071	0.982	-	0.959	0.963	-	0.982	0.966	-	1.017	1.352	-	1.317
TXT-FIN-MIDAS	B	OFs	0.978	-	1.031	0.953	-	0.940	0.954	-	0.979	0.982	-	1.012	1.350	-	1.299
FIN-MIDAS	-	GLs	0.962	0.785***	0.989	1.027	1.089	1.033	1.056	1.061	1.092	1.184	1.240	1.237	1.435	1.449	1.416
TXT-FIN-MIDAS	U	GLs	0.957	0.780***	0.993	1.021	1.068	1.030	1.068	1.068	1.106	1.182	1.243	1.236	1.498	1.496	1.476
TXT-FIN-MIDAS	D	GLs	0.942	0.773***	0.969	1.015	1.066	1.023	1.077	1.078	1.123	1.215	1.268	1.266	1.446	1.437	1.419
TXT-FIN-MIDAS	G	GLs	0.953	0.781***	0.990	1.023	1.087	1.036	1.077	1.081	1.113	1.198	1.254	1.249	1.427	1.447	1.417
TXT-FIN-MIDAS	B	GLs	0.933	0.770***	0.968	0.992	1.011	0.994	1.051	1.055	1.079	1.188	1.241	1.241	1.478	1.487	1.462
FIN-MIDAS	-	CRB	0.890*	0.725***	0.868**	0.902	0.919	0.906	0.897	0.876	0.916	0.962	1.006	1.020	1.152	1.145	1.168
TXT-FIN-MIDAS	U	CRB	0.890*	0.730***	0.872**	0.906	0.901	0.904	0.903	0.864	0.909	0.984	1.007	1.031	1.094	1.090	1.084
TXT-FIN-MIDAS	D	CRB	0.872**	0.730***	0.864**	0.903	0.895	0.897	0.908	0.892	0.931	0.979	1.010	1.023	1.172	1.165	1.148
TXT-FIN-MIDAS	G	CRB	0.876**	0.716***	0.873**	0.886	0.878	0.882	0.891	0.874	0.919	0.973	1.012	1.023	1.152	1.152	1.145
TXT-FIN-MIDAS	B	CRB	0.847**	0.713***	0.835***	0.886	0.862	0.864	0.863	0.885	0.926	0.978	1.020	1.031	1.171	1.192	1.180
FIN-MIDAS	-	BDi	0.964	0.790**	0.984	1.049	1.061	1.058	1.052	1.055	1.090	1.205	1.248	1.237	1.367	1.368	1.348
TXT-FIN-MIDAS	U	BDi	0.952	0.804**	0.975	1.059	1.076	1.062	1.049	1.045	1.088	1.164	1.242	1.212	1.322	1.315	1.300
TXT-FIN-MIDAS	D	BDi	0.966	0.794**	0.969	1.066	1.071	1.060	1.059	1.054	1.094	1.184	1.231	1.236	1.352	1.350	1.355
TXT-FIN-MIDAS	G	BDi	0.949	0.787**	0.975	1.039	1.063	1.056	1.076	1.069	1.107	1.188	1.245	1.238	1.363	1.362	1.347
TXT-FIN-MIDAS	B	BDi	0.934	0.779**	0.959	1.033	1.049	1.037	1.030	1.067	1.090	1.185	1.246	1.234	1.386	1.402	1.381
FIN-MIDAS	-	USi	0.940	0.761***	0.945*	0.966	1.005	0.968	0.960	0.958	0.994	1.046	1.094	1.102	1.380	1.385	1.354
TXT-FIN-MIDAS	U	USi	0.929	0.766***	0.939	0.958	0.992	0.963	0.967	0.963	0.997	1.047	1.094	1.105	1.333	1.328	1.305
TXT-FIN-MIDAS	D	USi	0.925	0.766**	0.942	0.956	0.992	0.960	0.976	0.971	1.005	1.056	1.093	1.107	1.338	1.324	1.345
TXT-FIN-MIDAS	G	USi	0.939	0.759***	0.934*	0.965	1.009	0.970	0.966	0.963	0.998	1.049	1.098	1.106	1.384	1.385	1.350
TXT-FIN-MIDAS	B	USi	0.918	0.745***	0.938	0.933	0.965	0.937	0.955	0.953	0.986	1.055	1.101	1.110	1.412	1.410	1.379
FIN-IR-MIDAS	-	FT	0.922	0.758**	0.922	1.017	1.035	1.001	1.008	1.003	1.040	1.063	1.118	1.115	1.301	1.304	1.274
TXT-FIN-IR-MIDAS	U	FT	0.929	0.759**	0.925	1.013	1.043	0.998	1.040	1.010	1.049	1.077	1.118	1.126	1.286	1.277	1.242
TXT-FIN-IR-MIDAS	D	FT	0.953	0.759**	0.936	1.014	1.039	1.003	1.046	1.026	1.065	1.104	1.142	1.151	1.320	1.275	1.251
TXT-FIN-IR-MIDAS	G	FT	0.922	0.751**	0.922*	0.998	1.022	0.994	1.011	1.010	1.045	1.065	1.120	1.123	1.315	1.309	1.276
TXT-FIN-IR-MIDAS	B	FT	0.915	0.749**	0.924	0.987	1.003	0.973	1.037	1.016	1.040	1.086	1.139	1.141	1.341	1.331	1.306
FIN-IR-MIDAS	-	SP	0.899*	0.754***	0.931*	0.956	0.978	0.959	0.931	0.945	0.971	0.974	1.011	1.026	1.117	1.109	1.103
TXT-FIN-IR-MIDAS	U	SP	0.910*	0.760***	0.938	0.970	0.983	0.973	0.962	0.946	0.986	0.978	1.018	1.034	1.079	1.069	1.068
TXT-FIN-IR-MIDAS	D	SP	0.918	0.760**	0.933	0.967	0.982	0.954	0.957	0.956	0.997	0.987	1.022	1.042	1.157	1.130	1.121
TXT-FIN-IR-MIDAS	G	SP	0.910*	0.751***	0.931*	0.951	0.976	0.951	0.950	0.950	0.982	0.972	1.018	1.023	1.111	1.103	1.097
TXT-FIN-IR-MIDAS	B	SP	0.892*	0.742***	0.908*	0.964	0.965	0.936	0.945	0.937	0.960	0.989	1.031	1.041	1.142	1.128	1.125
FIN-IR-MIDAS	-	ES	0.943	0.769**	0.979	1.012	1.032	1.009	1.019	1.022	1.051	1.090	1.130	1.121	1.377	1.353	1.336
TXT-FIN-IR-MIDAS	U	ES	0.943	0.769***	0.972	1.011	1.019	1.010	1.033	1.029	1.060	1.103	1.142	1.139	1.367	1.352	1.316
TXT-FIN-IR-MIDAS	D	ES	0.946	0.772**	0.973	1.011	1.034	1.007	1.052	1.044	1.070	1.135	1.164	1.162	1.482	1.375	1.320
TXT-FIN-IR-MIDAS	G	ES	0.938	0.768***	0.977	0.997	1.020	1.002	1.043	1.042	1.070	1.129	1.163	1.166	1.360	1.332	1.302
TXT-FIN-IR-MIDAS	B	ES	0.921	0.761**	0.937	0.982	0.993	0.979	1.010	1.016	1.037	1.092	1.143	1.118	1.372	1.363	1.320
FIN-IR-MIDAS	-	HS	0.884*	0.756***	0.906*	0.966	1.000	0.969	0.965	0.967	1.001	0.978	1.024	1.035	1.190	1.192	1.191
TXT-FIN-IR-MIDAS	U	HS	0.880*	0.758***	0.907*	0.961	0.999	0.967	0.970	0.974	1.008	0.964	1.023	1.021	1.127	1.101	1.111
TXT-FIN-IR-MIDAS	D	HS	0.881*	0.753***	0.904*	0.967	0.991	0.963	0.986	0.985	1.017	0.978	1.018	1.030	1.200	1.152	1.160
TXT-FIN-IR-MIDAS	G	HS	0.894*	0.751***	0.903*	0.950	0.985	0.955	0.962	0.967	1.001	0.984	1.027	1.038	1.176	1.170	1.169
TXT-FIN-IR-MIDAS	B	HS	0.880**	0.740***	0.897*	0.940	0.943	0.929	0.974	0.960	0.993	0.996	1.047	1.037	1.242	1.249	1.230
FIN-IR-MIDAS	-	PD	0.906*	0.764***	0.925*	0.958	0.980	0.962	0.961	0.966	1.001	1.041	1.087	1.105	1.206	1.202	1.202
TXT-FIN-IR-MIDAS	U	PD	0.904*	0.762***	0.924*	0.959	0.978	0.971	0.961	0.966	1.000	1.032	1.096	1.098	1.170	1.163	1.159
TXT-FIN-IR-MIDAS	D	PD	0.903*	0.757***	0.932	0.962	0.974	0.962	0.986	0.976	1.006	1.053	1.088	1.112	1.225	1.215	1.222
TXT-FIN-IR-MIDAS	G	PD	0.898**	0.758***	0.925*	0.948	0.966	0.960	0.955	0.964	0.999	1.032	1.076	1.088	1.204	1.205	1.202
TXT-FIN-IR-MIDAS	B	PD	0.883**	0.748***	0.909	0.932	0.931	0.932	0.956	0.956	0.986	1.061	1.110	1.116	1.262	1.323	1.300
FIN-IR-MIDAS	-	CD	0.932	0.768**	0.961	1.018	1.067	0.976	0.967	0.940	0.957	1.025	0.976	1.048	2.646	2.320	2.496
TXT-FIN-IR-MIDAS	U	CD	0.928*	0.769***	0.961	1.014	1.048	0.981	0.966	0.932	0.943	0.993	0.965	1.014	2.503	2.129	2.309
TXT-FIN-IR-MIDAS	D	CD	0.916	0.763**	0.950	0.997	1.031	0.965	0.911	0.929	0.939	1.168	1.093	1.172	2.935	2.543	2.702
TXT-FIN-IR-MIDAS	G	CD	0.941	0.767**	0.961	1.002	1.044	0.976	0.940	0.936	0.944	1.025	0.984	1.044	2.977	2.479	2.731
TXT-FIN-IR-MIDAS	B	CD	0.920	0.750***	0.930	0.975	0.970	0.935	0.942	0.838	0.880	1.043	1.040	1.041	2.976	2.553	2.806
FIN-IR-MIDAS	-	ED	0.908*	0.759***	0.926*	0.970	0.988	0.970	0.955	0.960	0.991	1.013	1.073	1.088	1.215	1.219	1.214
TXT-FIN-IR-MIDAS	U	ED	0.899*	0.766***	0.926*	0.967	0.981	0.973	0.965	0.966	0.996	1.024	1.076	1.087	1.166	1.160	1.164
TXT-FIN-IR-MIDAS	D	ED	0.901*	0.757***	0.936	0.967	0.974	0.970	0.980	0.978	1.008	1.040	1.080	1.087	1.244	1.224	1.217
TXT-FIN-IR-MIDAS	G	ED	0.897*	0.762***	0.927*	0.954	0.972	0.963	0.956	0.963	0.994	1.020	1.070	1.082	1.196	1.195	1.193
TXT-FIN-IR-MIDAS	B	ED	0.891*	0.746***	0.907	0.945	0.938	0.937	0.957	0.955	0.984	1.038	1.088	1.092	1.249	1.251	1.249
CO-IR-MIDAS	-	-	0.937	0.765**	0.935	0.983	1.041	0.981	0.955	0.957	0.994	0.936	0.971	0.989	1.013	1.009	1.004
TXT-CO-IR-MIDAS	U	-	0.925	0.761**	0.931	0.973	1.031	0.975	0.952	0.952	0.988	0.940	0.976	0.992	1.053	1.039	1.019
TXT-CO-IR-MIDAS	D	-	0.923	0.762**	0.929	0.975	1.024	0.974	0.967	0.961	1.002	0.961	0.988	1.006	1.103	1.121	1.107
TXT-CO-IR-MIDAS	G	-	0.927	0.757***	0.934	0.963	1.018	0.967	0.959	0.956	0.988	0.944	0.979	0.992	1.011	1.009	1.004
TXT-CO-IR-MIDAS	B	-	0.905*	0.752***	0.931	0.947	0.953	0.945	0.959	0.966	0.993	0.949	0.979	0.995	1.370	1.049	1.047
GL-IR-MIDAS	-	-	0.925	0.777**	0.943	1.001	1.031	1.004	1.007	1.005	1.037	1.079	1.134	1.141	1.321	1.313	1.287
TXT-GL-IR-MIDAS	U	-	0.927	0.775**	0.940	0.998	1.025	0.987	1.032	0.997	1.040	1.087	1.152	1.187	1.240	1.256	1.218
TXT-GL-IR-MIDAS	D	-	0.928	0.774**	0.940	0.993	1.026	0.993	1.036	1.012	1.053	1.092	1.136	1.157	1.326	1.300	1.293
TXT-GL-IR-MIDAS	G	-	0.918	0.768**	0.939	0.986	1.017	0.983	1.008	1.007	1.041	1.082	1.139	1.149	1.289	1.288	1.250
TXT-GL-IR-MIDAS	B	-	0.927	0.764**	0.937	0.966	0.987	0.964	1.029	1.002	1.038	1.102	1.162	1.169	1.363	1.353	1.301
CP-IR-MIDAS	-	-	0.935	0.770**	0.970	1.002	1.059	0.996	1.058	1.065	1.092	1.069	1.114	1.099	1.385	1.368	1.331
TXT-CP-IR-MIDAS	U	-	0.939	0.768**	0.975	1.000	1.061	0.996	1.080	1.090	1.109	1.081	1.110	1.114	1.402	1.383	1.350
TXT-CP-IR-MIDAS	D	-	0.930	0.771**	0.960	0.991	1.055	0.986	1.084	1.090	1.098	1.103	1.134	1.133	1.413	1.366	1.330
TXT-CP-IR-MIDAS	G	-	0.925	0.767**	0.977	0.984	1.047	0.990	1.081	1.079	1.112	1.073	1.102	1.091	1.383	1.348	1.312
TXT-CP-IR-MIDAS	B	-	0.905*	0.756**	0.914*	0.953	0.991	0.952	1.009	1.030	1.050	1.067	1.109	1.092	1.385	1.367	1.318
NG-IR-MIDAS	-	-	0.960	0.791**	0.955	1.037	1.035	1.077	1.069	1.081	0.996	1.068	1.107	1.100	1.439	1.421	1.376
TXT-NG-IR-MIDAS	U	-	0.953	0.787**	0.961	1.033	1.039	1.064	1.115	1.139	1.118	1.085	1.131	1.132	1.444	1.437	1.382
TXT-NG-IR-MIDAS	D	-	0.953	0.788**	0.946	1.001	1.033	1.071	1.073	1.065	0.998	1.095	1.135	1.131	1.604	1.536	1.484
TXT-NG-IR-MIDAS	G	-	0.959	0.798**	0.967	1.008	1.041	1.076	1.071	1.098	1.088	1.077	1.117	1.105	1.388	1.387	1.349
TXT-NG-IR-MIDAS	B	-	0.911	0.778**	0.928	0.987	1.014	1.034	1.025	1.020	1.044	1.081	1.116	1.101	1.651	1.593	1.539
PL-IR-MIDAS	-	-	0.985	0.795***	1.010	1.006	1.070	1.003	1.079	1.067	1.100	1.114	1.135	1.132	1.369	1.390	1.343
TXT-PL-IR-MIDAS	U	-	0.981	0.793***	1.002	1.009	1.075	1.002	1.095	1.093	1.131	1.119	1.147	1.150	1.423	1.388	1.334
TXT-PL-IR-MIDAS	D	-	0.970	0.791***	0.998	0.998	1.056	0.995	1.091	1.092	1.126	1.125	1.155	1.153	1.494	1.426	1.399
TXT-PL-IR-MIDAS	G	-	0.971	0.791***	0.999	0.988	1.052	0.992	1.092	1.089	1.120	1.112	1.143	1.147	1.333	1.318	1.312
TXT-PL-IR-MIDAS	B	-	0.938	0.777***	0.964	0.959	1.019	0.958	1.018	1.045	1.071	1.089	1.130	1.113	1.475	1.434	1.377
SL-IR-MIDAS	-	-	0.944	0.767***	0.942	1.023	1.041	1.017	0.998	0.993	1.037	1.058	1.103	1.112	1.267	1.299	1.263
TXT-SL-IR-MIDAS	U	-	0.936	0.772***	0.942	1.010	1.036	1.003	1.010	0.997	1.040	1.077	1.116	1.137	1.255	1.204	1.195
TXT-SL-IR-MIDAS	D	-	0.938	0.773***	0.944	1.011	1.032	1.004	1.021	1.002	1.043	1.070	1.098	1.118	1.309	1.291	1.275
TXT-SL-IR-MIDAS	G	-	0.936	0.759***	0.937	0.990	1.021	0.992	0.993	0.993	1.028	1.056	1.113	1.116	1.282	1.279	1.259
TXT-SL-IR-MIDAS	B	-	0.913	0.759**	0.931	0.988	1.005	0.975	1.001	0.986	1.018	1.078	1.119	1.133	1.319	1.278	1.253

Note: In column 1 TXT: text data defined in column 2, FIN: financial variable defined in column 3, IR: intraday returns, CO: crude oil, GL: gold, CP: copper, NG: natural gas, PL: palladium, SL: silver. In column 2 U: unigram count, D: dictionary method, G: geometrical model, B: Bert. In column 3 OFs: crude oil-futures prices spread, GLs: crude oil-gasoline spread, CRB: Commodity research bureau index, BDi: Baltic Dry index, USi: federal funds rate, FT: FTSE100 index, SP: S&P500 index, ES: Euro Stoxx 50 index, HS: Hang Seng index, PD: GBP/USD exchange rate, CD: CAD/USD exchange rate, ED: UER/USD exchange rate. Black bold values indicate improvements on the no-change forecast. Red bold values indicate the lowest MSPE achieved with no text variable. Blue entries stand for the lowest MSPE for a given time horizon and relative to a specific oil price measure. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table 3.6: Recursive ALPL ratios relative to a random walk density forecast of alternative monthly indicators of real oil prices; text vs. no-text.

Model	Text Variable	1-month			3-months			6-months			12-months			24-months		
		WTI	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*
SV-BVAR	-	0.961***	1.009***	0.947***	0.123***	0.032***	0.122***	-0.285***	-0.326***	-0.339***	-0.601***	-0.544***	-0.778***	-0.971***	-0.994***	-1.331***
SV-BVAR-TXT	U	0.957***	1.004***	0.938***	0.106***	0.012***	0.126***	-0.298***	-0.344***	-0.338***	-0.618***	-0.563***	-0.762***	-0.953***	-0.984***	-1.255***
SV-BVAR-TXT	D	0.957***	0.988***	0.942***	0.105***	-0.001***	0.093***	-0.290***	-0.330***	-0.343***	-0.623***	-0.551***	-0.787***	-0.952***	-0.954***	-1.301***
SV-BVAR-TXT	G	**0.963*****	**1.010*****	0.957***	**0.135*****	**0.046*****	**0.139*****	**-0.260*****	**-0.280*****	**-0.289*****	-0.593***	**-0.531*****	-0.733***	-1.015***	-1.043***	-1.327***
SV-BVAR-TXT	B	**0.963*****	1.005***	**0.958*****	0.105***	0.010***	0.100***	-0.281***	-0.327***	-0.334***	-0.602***	-0.538***	-0.736***	-0.939***	**-0.931*****	-1.224***
MF-SV-BVAR-TXT	U	0.956***	0.996***	0.933***	0.104***	0.013***	0.117***	-0.272***	-0.316***	-0.301***	**-0.572*****	**-0.531*****	**-0.679*****	**-0.907*****	-0.942***	**-1.133*****
MF-SV-BVAR-TXT	D	0.899***	0.989***	0.942***	0.079***	0.027***	0.087***	-0.305***	-0.288***	-0.328***	-0.631***	-0.552***	-0.757***	-0.946***	-0.986***	-1.289***
MF-SV-BVAR-TXT	G	0.942***	0.996***	0.944***	0.091***	0.003***	0.096***	-0.295***	-0.294***	-0.296***	-0.614***	-0.549***	-0.734***	-0.984***	-1.039***	-1.341***
MF-SV-BVAR-TXT	B	0.929***	0.991***	0.943***	0.087***	-0.048***	0.050***	-0.310***	-0.355***	-0.362***	-0.636***	-0.580***	-0.734***	-0.970***	-0.949***	-1.212***
CRB index																
MIDAS	-	-3.928**	-3.890*	-3.821***	-4.923	-4.326	-4.553	-5.578	-4.881	-5.579	-7.653	-6.269	-7.364*	-12.874*	-11.194*	-12.720*
TXT-MIDAS	U	-3.950**	-3.716**	-3.762**	-4.858	-4.091	-4.593	-5.805	-4.701	-5.239	-7.373	-6.302	-7.231*	-10.691*	-8.748*	-10.203*
TXT-MIDAS	D	**-3.838****	-3.632**	-3.673**	-4.865	-4.128	-4.628	-5.757	-4.757	-5.326	**-6.586**	**-5.561**	**-6.308***	**-6.695****	**-5.743***	**-6.658****
TXT-MIDAS	G	-3.901**	**-3.387****	**-3.526*****	**-3.859**	**-3.464**	**-3.312**	**-5.131**	**-4.010**	**-4.451**	-7.158	-6.169	-7.061*	-10.135*	-9.211*	-10.259**
TXT-MIDAS	B	-3.874**	-3.694**	-3.792**	-4.685	-4.039	-4.515	-5.941	-4.899	-5.243	-7.763	-6.500	-7.257*	-12.682*	-10.386	-12.158*
Euro Stoxx 50 index																
MIDAS	-	-4.838*	-4.165**	-4.160**	-6.888	-5.720	-6.160	-8.843	-7.049	-7.558	-9.114*	-7.630*	-8.708**	-10.199*	-9.218*	-10.654*
TXT-MIDAS	U	-4.759*	-4.088**	-4.107**	-6.200	-5.433	-5.447	-8.666	-6.984	-7.302	-8.970*	-7.508*	-8.520**	-10.187*	-9.215*	-10.623**
TXT-MIDAS	D	-4.693**	-4.046**	-4.096**	-6.652	-5.483	-5.791	-8.785	-7.028	-7.441*	-8.670*	-7.165**	-8.233	**-7.679***	**-7.045***	**-8.062***
TXT-MIDAS	G	-4.673*	**-3.709****	**-3.847****	**-4.907**	**-4.622**	**-3.989**	**-6.183**	**-4.942**	**-5.246**	**-7.755***	**-6.652****	**-7.596****	-9.316*	-8.079*	-9.424**
TXT-MIDAS	B	**-4.548****	-3.920**	-4.011**	-6.269	-5.245	-5.529	-8.097	-6.518	-6.859	-8.726*	-7.313*	-8.290**	-10.861*	-9.561*	-10.811*
GBP/USD index																
MIDAS	-	-4.415**	-4.192**	-4.487**	-5.894	-5.094	-5.757	-7.134	-5.972	-6.554	-7.602*	-6.541**	-7.622**	-13.630*	-13.521**	-15.673**
TXT-MIDAS	U	-4.331**	-4.042**	-4.222**	-5.417	-4.706	-5.173	-7.088	-5.862	-6.442	-7.658*	-6.548**	-7.622**	-11.738*	-9.240*	-10.580*
TXT-MIDAS	D	-4.235**	-3.975**	-4.136**	-5.745	-4.898	-5.491	-7.226	-6.082	-6.615	**-6.949***	**-6.040****	**-7.024****	**-7.869****	**-6.875****	**-7.228****
TXT-MIDAS	G	-4.355**	-3.903**	-4.170**	**-4.881**	**-4.238**	**-4.308**	**-5.810**	**-4.830**	**-5.229**	-7.383*	-6.296*	-7.435**	-11.668*	-11.418**	-13.738**
TXT-MIDAS	B	**-4.081****	**-3.867****	**-4.085****	-5.364	-4.580	-5.155	-6.848	-5.668	-6.174	-7.705*	-6.625**	-7.676**	-13.887*	-13.491**	-15.071**
Natural Gas index																
MIDAS	-	-5.201	-4.431	-4.709*	-8.605	-6.198	-8.959	-9.444	-7.467	-9.651	-8.929*	-7.283*	-8.496**	-11.247**	-10.165**	-11.722**
TXT-MIDAS	U	-4.815	-4.172*	-4.221*	-6.639	-5.546	-7.827	-8.908	-7.261	-9.594	-8.472*	-7.000*	-8.103**	-9.064*	-8.176*	-9.668**
TXT-MIDAS	D	-4.586*	-3.988*	-4.067*	-7.211	-5.651	-7.995	-9.200	-7.441	-9.601	-8.590*	-7.029*	-8.341*	**-7.210****	**-6.766****	**-7.554****
TXT-MIDAS	G	**-4.360****	**-3.616****	**-3.612****	**-5.783**	**-4.853**	**-6.302**	**-7.238**	**-6.084**	**-8.518**	**-8.120***	**-6.603***	**-7.815****	-8.860*	-8.223*	-9.818**
TXT-MIDAS	B	-4.407*	-3.775**	-4.026**	-6.275	-4.944	-6.684	-7.965	-6.380	-8.529	-8.508*	-7.114*	-8.221**	-11.024**	-10.041**	-11.510**
Palladium index																
MIDAS	-	-4.803*	-4.087*	-4.257*	-6.490	-5.454	-5.759	-8.474	-6.674	-7.130	-8.898*	-7.360*	-8.449**	-11.197**	-9.835*	-11.389**
TXT-MIDAS	U	-4.614*	-3.916*	-3.953*	-5.837	-5.230	-5.135	-8.173	-6.579	-6.923	-8.611*	-7.146*	-8.163**	-10.066**	-8.886*	-10.269**
TXT-MIDAS	D	-4.398*	-3.789*	-3.885*	-6.061	-4.874	-5.336	-8.272	-6.681	-7.119	-8.726*	-7.048*	-8.429**	**-7.059****	**-6.533****	**-7.346****
TXT-MIDAS	G	-4.554	**-3.682****	**-3.567****	**-4.659**	**-4.316**	**-3.787**	**-5.942**	**-4.711**	**-4.911**	-8.337*	**-6.926***	-8.089*	-8.638**	-7.948**	-9.371**
TXT-MIDAS	B	**-4.344***	-3.704**	-3.821**	-5.658	-4.690	-5.025	-7.551	-6.103	-6.522	**-8.262***	-6.956*	**-7.750****	-11.909**	-10.625**	-12.586**

Note: In column 1 SV: stochastic volatility, BVAR: Bayesian vector autoregression, MF: mixed frequency, TXT: text data defined in column 2. In column 2 U: unigram count, D: dictionary method, G: geometrical model, B: Bert. Bold values indicate the highest ALPL improvements on the no-text based model, for a given time horizon and relative to a specific oil price measure. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

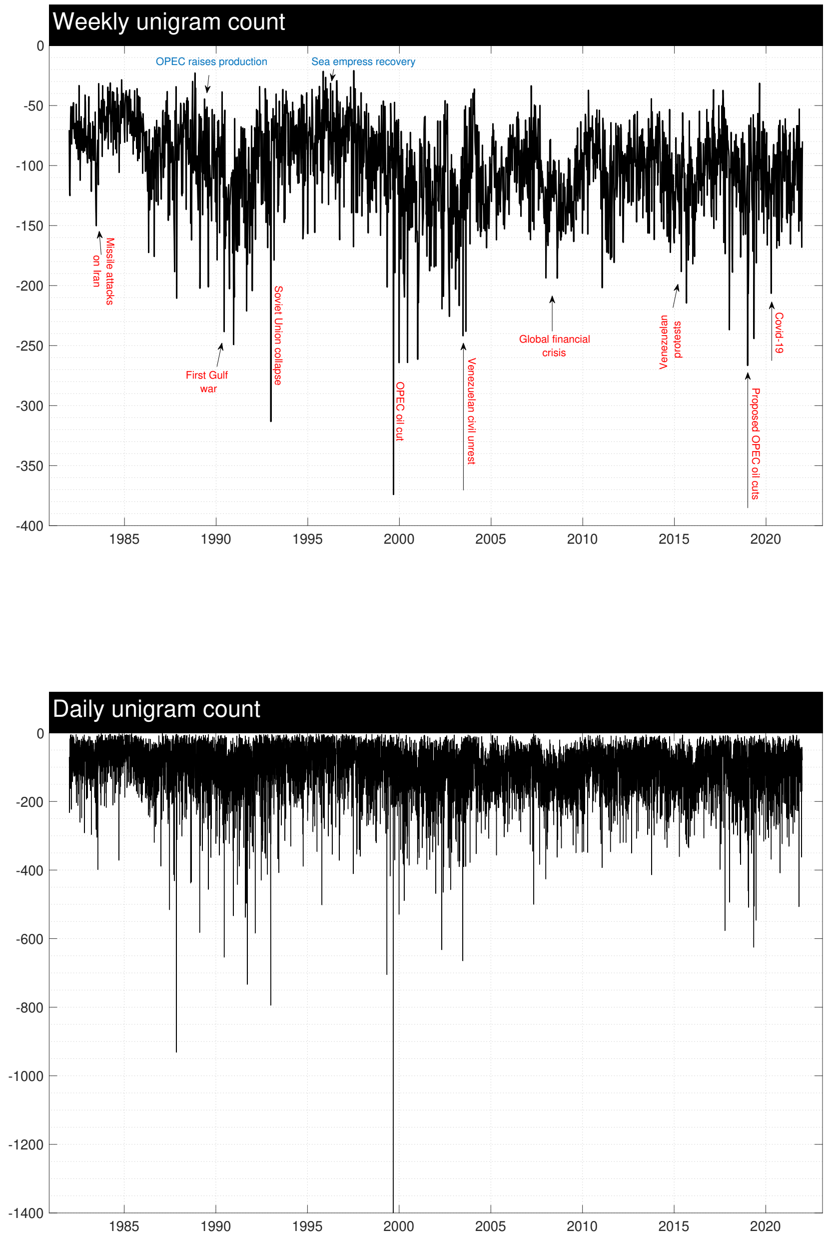
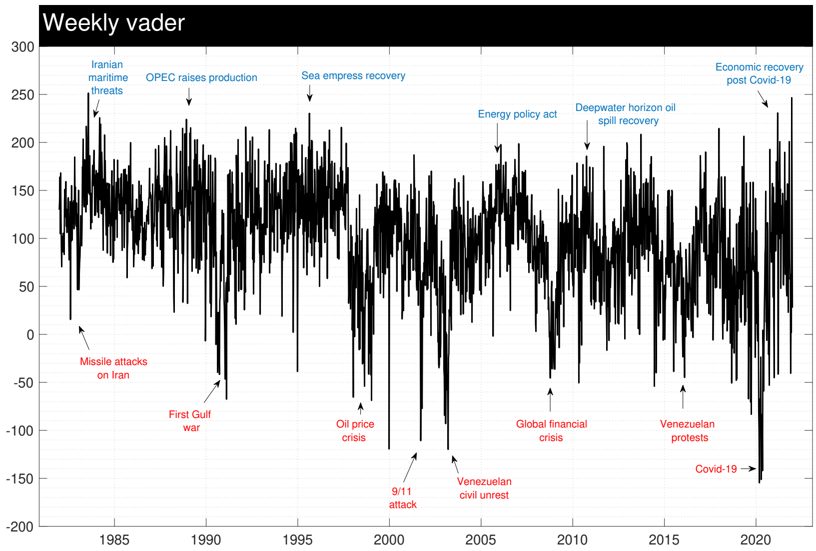


Figure 3.1: The top panel reports the unigram probability count indicator, developed by counting the number of words "economy", "economies", "economic", "economics" occurring in each article, normalized by the total number of running words. The figure plots the time series weekly score from 1982M1 through 2021M12. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil. The bottom panel replicates the equivalent time series by using daily scores (labels are omitted on the bottom panel as in several cases daily variations are not informative).



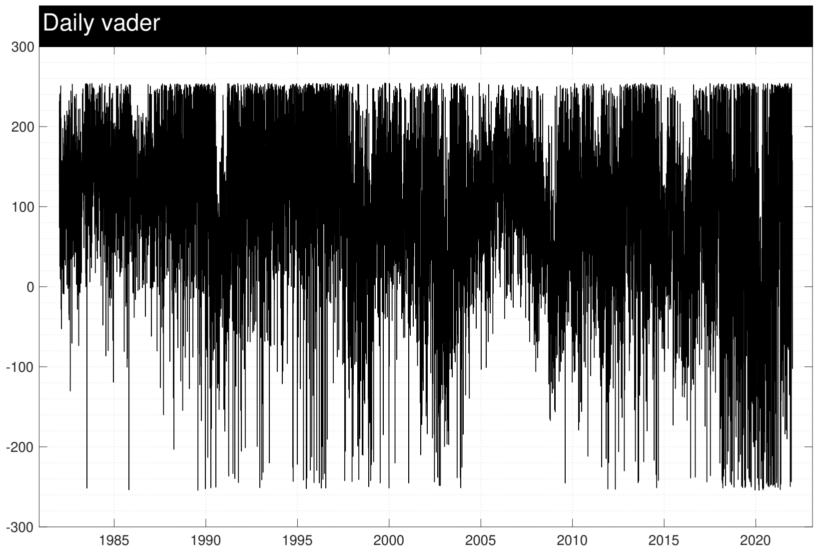


Figure 3.2: The top panel reports the VADER sentiment indicator, where words are assigned a value based on Hutto and Gilbert [2014]'s dictionary. The figure plots the time series weekly score from 1982M1 through 2021M12. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil. The bottom panel replicates the equivalent time series by using daily scores (labels are omitted on the bottom panel as in several cases daily variations are not informative).

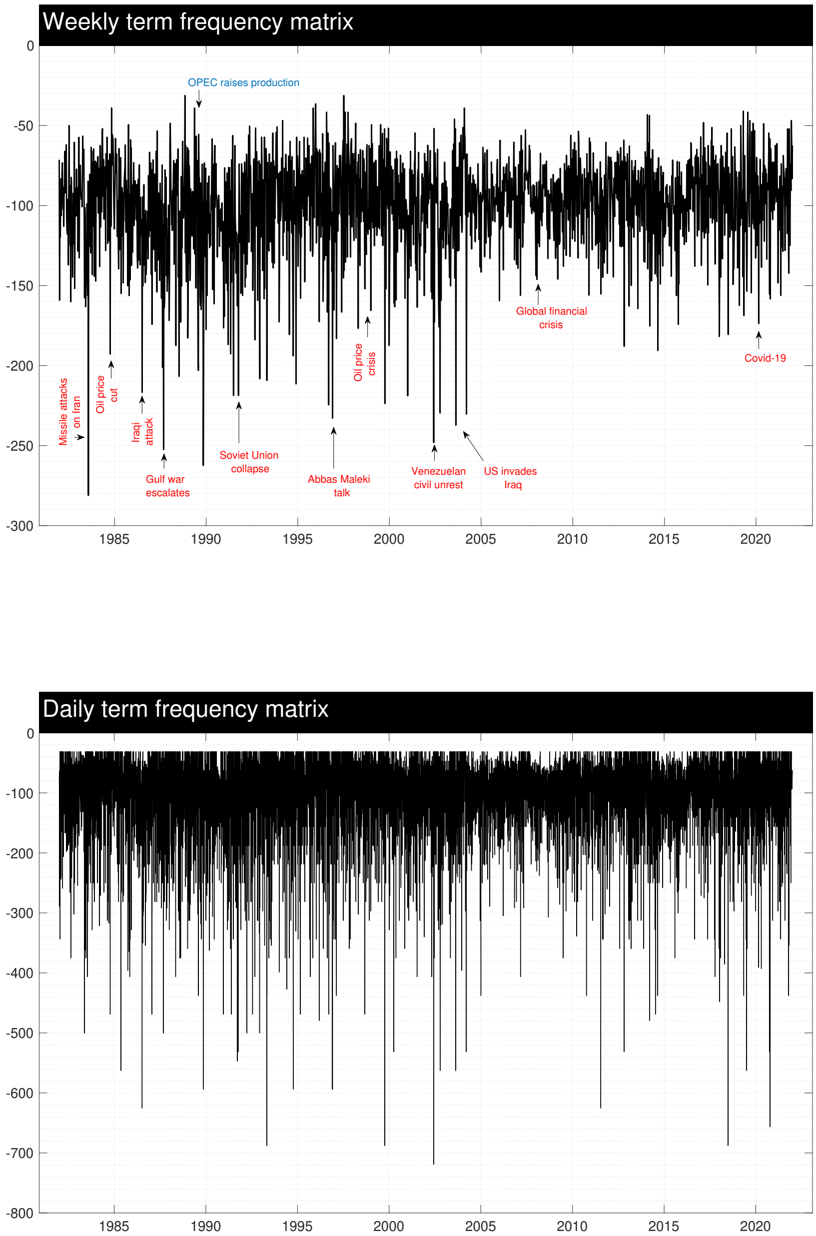
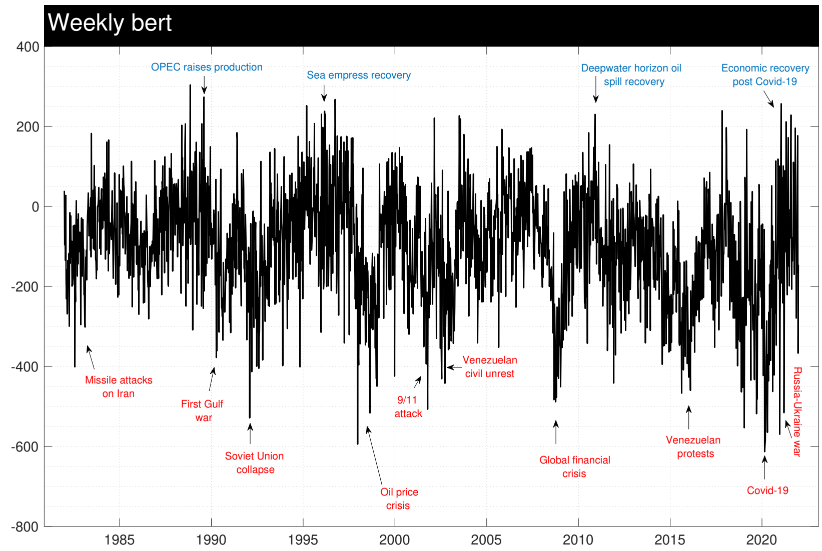


Figure 3.3: The top panel reports the term-frequency matrix indicator, developed by counting the number of words "economy", "economies", "economic", "economics" occurring in each article, normalized by the total number of words running in a weekly/daily dataset. The figure plots the time series weekly score from 1982M1 through 2021M12. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil. The bottom panel replicates the equivalent time series by using daily scores (labels are omitted on the bottom panel as in several cases daily variations are not informative).



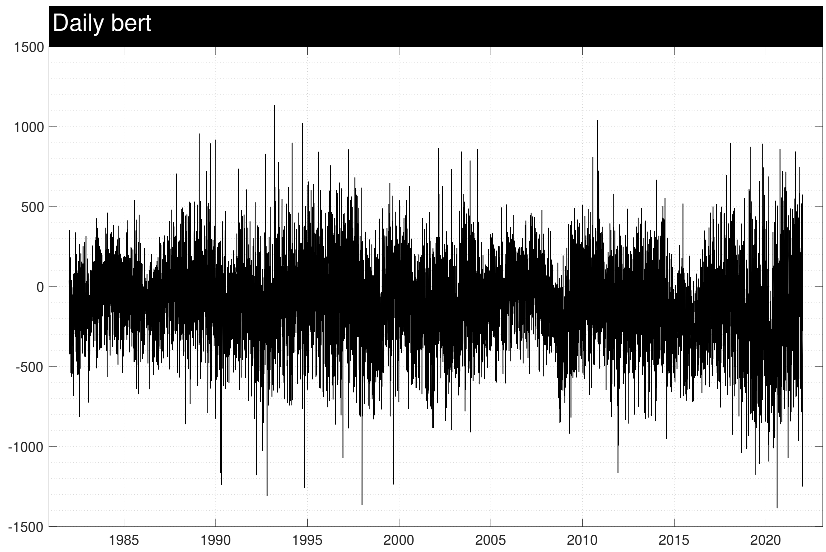


Figure 3.4: The top panel reports the sentiment indicator extracted from oil related articles by using BERT. The figure plots the time series weekly score from 1982M1 through 2021M12. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil. The bottom panel replicates the equivalent time series by using daily scores (labels are omitted on the bottom panel as in several cases daily variations are not informative).

4

Oil Supply and Demand Shock under Model Uncertainty

4.1 Introduction

The instability over time between oil price fluctuations and the response of macroeconomic variables has long been considered a case of study of primary interest for academics and central bankers. Economists are particularly concerned regarding the spike in crude oil prices, as this results in rising inflation as seen in the deep recessions in the mid-1970s and the early 1980s, and following the Russia-Ukraine war outbreak.

A considerable number of academic studies assessing the macroeconomic effects following oil disturbances has been proposed in the economic literature. Some influential works are Hamilton [1983], Mork [1989], Blanchard and Gali [2007], Kilian [2009] and many others. Whilst these papers offer no settled consensus regarding a specific strategy related to the correct identification of the economic disturbances, they all agree on the qualitative effects following a change in crude oil prices. However, as common practice, inference is made by using a pre-determined small set of variables, as if this model had been given in advance. This strategy generates uncertainty in the empirical results, and this problem has never been considered in the economic literature. In this paper I use model averaging as a strategy to address this informational problem.

MODEL AVERAGING and **MODEL SELECTION** are two valid approaches that can partially solve the problem of informational deficiency. The latter refers to a methodology in which, given a number of "reasonable" datasets, the analyst assigns weight 1 to one model (or even more) and 0 to the remaining candidate models. Several model selection criteria have been proposed in the literature and listing them all here is quite impossible. However, there are two main reasons that have led me to choose model averaging as opposed to model selection. Firstly, the process of selecting one optimal model and discarding the remaining candidates would under-report the overall variance. This generates additional uncertainty, or even bias (see Madigan and Raftery [1994]). Secondly, alternative selection criteria usually result in offering different best models for the same dataset. This implies that more candidate models do almost as well as the optimal model. Hence, estimation post selection can generate very misleading results, whereas combining the output across all possible solutions eliminates both the aforementioned problems all at once (see Claeskens et al. [2008]).

Model averaging is thus a valid alternative to model selection for the purposes of this paper, because no specific "winner model" is first selected and then estimated, but inference is made on all candidate models and the results are averaged according to how likely each model is. Several statistical arguments have been made in support of model averaging (i.e. Leamer [1978], Draper [1995], Raftery et al. [1997], Hoeting et al. [1999], Claeskens et al. [2008] and many others) and myriads of macro and microeconometric empirical studies have used this procedure (see for example Min and Zellner [1993], Raftery et al. [1997], Fernandez et al. [2001], Koop and Korobilis [2012] and Strachan and Van Dijk [2013]). There are many advantages to averaging over a number of weighted models and only a few are mentioned here. Firstly, inference on several weighted models implicitly incorporates model uncertainty and this allows the analyst to obtain better coefficient estimates. Secondly, it leads to a significant decrease of error predictions in high-dimensional regression problems (see Ando and Li [2014]). Third, this methodology provides useful empirical results also when applied in some machine learning techniques such as computational linguistic for speech recognition, and decision trees (see McAllester [1999]).

This study considers a strategy originally developed in Kass and Raftery [1995], known as Bayesian Information Criterion (BIC) model averaging[1], which is formalized as follows. Assume we wish to start with a dataset of K variables and N observations, where $K < N$. If we assume that the set of reasonable models is designed by whether a variable is included or excluded, the total number of candidate models is 2^k. In this paper, the information set is made up by 10 dependent variables, in which, the first three are kept fixed, and the remaining $K = 7$ predictors are combined in order to define the set of all admissible candidates. This means that, $2^7 = 128$ different models are estimated. Coefficient estimates are then averaged over the goodness-of-fit of each candidate model. Namely, weights are set proportional to $exp^{\frac{1}{2}(BIC_k)}$, where BIC_k describes the goodness-of-fit of model $k = 1, \ldots, K$. Buckland et al. [1997] use a similar methodology, but they express the quality of each candidate model with the Akaike Information Criterion in place of BIC.

The contribution of this paper is twofold. Firstly, this strategy is used to revisit the role of (i) oil supply, (ii) aggregate demand and (iii) oil-specific demand shock and explain the macroeconomic effects determined by oil price fluctuations. In particular, I revisit, in face of model uncertainty, Kilian [2009] and Kilian and Murphy [2012]. In fact, in both papers inference is made by using a 3-variable vector autoregression (VAR) model, that is implicitly assumed to be correctly specified. In contrast, I use a set of 10 macroeconomic variables, modelled in the form of a structural VAR, in which the first three (oil production $\{q^{oil}\}$, real global economy $\{y^{GDP}\}$ and real oil price $\{p^{oil}\}$) are kept fixed for any candidate model, and the remain-

[1] In this work the expressions BIC model averaging (BICMA) and information criterion model averaging (ICMA) are used interchangeably.

ing seven predictors are combined in order to generate 128 different datasets. Even though the analysis estimates all model combinations, the reduced-form coefficients are used to compute the impulse response functions (IRFs) only related to the first three dependent variables. Namely, the dynamic response of the remaining factors is disregarded. I compare IRFs achieved through model averaging to IRFs based on the 3-variable VAR model used in Kilian [2009] and Kilian and Murphy [2012]. The latter model is referred to as "*misspecified*" VAR. Empirical results show that following a demand-specific oil shock, the price of oil increases on the impact as in Kilian [2009], but in the long term the positive variation is more persistent and does not converge to zero as in the misspecified VAR. Moreover, the oil price response resulting from an aggregate demand shock, is more stable in the long run and does not converge to infinite as in the informationally deficient VAR. The performance of this methodology is also assessed through a simulation study. In particular I show that the median response of the true VAR falls inside the 100% posterior of credible set of the artificial VAR responses. The results are robust to different VARs specification.

The second major contribution is that I propose the "oil-news shock" as a novel oil-related structural disturbance. In particular, I show that other than world industrial production, TOSI is the variable that best explains oil price fluctuations. TOSI is a text based factor that captures human sentiment related to oil news items (see Gifuni [2021]). For this experiment I use the same dataset used in the previous exercise and the variables that are kept fixed for any candidate model are: q^{oil}, y^{GDP}, p^{oil}, oil inventories $\{oil^{inv}\}$ and oil news $\{oil^{news}\}$. The evidence shows that oil price stability is considerably affected by oil news shocks, and this variation generates a persistent drop in the quantity of oil produced. Moreover, the difference between oil-news IRFs in the misspecified and full-specified VAR is clear. This highlights that the content of newspapers is prone to react to any additional information.

The informational problem addressed in this paper is related to several other works in the literature. Bernanke and Boivin [2003] use a factor augmented VAR and show that if factors are generated from a rich dataset through principal component analysis, the model comes up with better predictions in comparison to Fed's forecasts. Forni and Gambetti [2014] demonstrate the necessary and sufficient conditions under which a VAR is not informationally deficient. They also assume that the economy is described by a state-space model and the hidden factors in the measurement equation are the principal components of a large dataset, which Granger cause the dependent variables. However, using principal components to inform a VAR does not allow the researcher to understand which factor is useful in estimating valid IRFs. This, in a nutshell, is instead what model averaging does. Fernández-Villaverde et al. [2007] discuss the information accuracy of unrestricted VARs, but they assume that the DSGE structure of the economy is known *a priori*. But what about if the economic model is unknown? It is when we hypothesize how the economy evolves that structural VARs come into play. Model averaging can minimize the information deficiency of a pre-specified VAR, by combining the inference output resulting from any candidate model.

The remainder of the paper is organized as follows. Section 4.2 describes the main idea of BIC model averaging. Section 4.3 presents the identification methodologies and the Bayesian approach used to revisit Kilian [2009] and Kilian and Murphy [2012]. Section 4.4 illustrates through a simulation study the finite sample properties of Information Criterion model averaging (ICMA)[2]. Section 4.5 provides the empirical results and Section 4.6 concludes.

4.2 Information Criterion Model Averaging

The absence of a completely defined model and the related informational deficiency issue was firstly described in Leamer [1978]. In this book, the author suggests that averaging a number of competing models over a set of asymptotically consistent weights, allows inference output to account for model uncertainty[3]. In response to this view, Sala-i Martin [1997] runs four million regressions and average the coefficient estimates over the integrated likelihood of each model with the purpose of finding as many good economic growth predictors as possible. After this work, model averaging has begun to find an application amongst several disciplines such as management science, meteorology and medicine (see for example Raftery et al. [2005], Gneiting et al. [2005], Vrugt and Robinson [2007], Vrugt et al. [2008], Viallefont et al. [2001], Raftery et al. [2005] and Yin and Yuan [2009]). Myriads of different data-driven averaging techniques have been developed so far. For example, Granger and Ramanathan [1984] regress the observed values on the R matrix of forecasted points (where R is the number of all model combinations) and provide OLS-estimated weights. Bates and Granger [1969] weight each model over its forecasted variance and Wright [2008] forecasts a class of bilateral exchange rates by considering the weights as free parameters to estimate in a Bayesian fashion.

[2] In this work the expressions BIC model averaging (or BICMA) and information criterion model averaging (or ICMA) are used interchangeably.

[3] In theory, weights are shown to be asymptotically optimal if they all sum up to one. However, recent studies have relaxed this constraint, showing that in some specific circumstances, although the sum of weights is greater than one, asymptotic consistency still holds (see Ando and Li [2014]).

This paper suggests the Information Criterion averaging procedure, firstly proposed in Buckland et al. [1997], as a method to compute the posterior weights ψ expressed as:

$$\psi_k = \frac{exp^{-\frac{1}{2}IC_k}}{\sum_{k=1}^{2K} exp^{-\frac{1}{2}IC_k}} \qquad \forall\ k = 1, \cdots, 128 \tag{4.1}$$

where IC_k indicates the Information Criterion used to select the best lag order relative to VAR_k, for $k = 1, \cdots, 128$. In my case, IC_k is the Bayesian Information Criterion, which is based on the following equation:

$$\text{BIC} = -2ln(L_T(\hat{\theta})) + \tfrac{ln(T)r}{T}$$

where $ln(L_T(\hat{\theta}))$ is the normalized log-likelihood evaluated at the estimated vector of parameters $\hat{\theta}$, $\frac{ln(T)k}{T}$ is a penalty term which indicates the loss of degrees of freedom resulting from increasing the lag order of model k, and r defines the number of VAR parameters[4]. However, empirical results are robust to different Information Criteria[5].

Now, turning to the model specification, each VAR is estimated in a Bayesian fashion illustrated in section 4.3.6, and the structural response relative to a variable j after a shock i in model k is denoted as $\text{I}\tilde{\text{R}}\text{F}_{j,i,k}$. Hence, the weighted impulse-response of variable j after a shock i is given by:

$$\text{I}\tilde{\text{R}}\text{F}_{j,i} = \sum_{k=1}^{2^7} \psi_k \, \text{I}\tilde{\text{R}}\text{F}_{j,i,k} \tag{4.2}$$

in which, k =1,...,128, j and i =1,2,3 since we are averaging 128 structural VARs and only the upper left 3×3 matrix response in investigated.

[4] The lag length p can assume a minimum value of 12 and a maximum of 36 in each model combination. Reasons behind this choice can be found in Baumeister and Kilian [2015]

[5] Results do not deviate in case Akaike Information Criterion or Hannan Information Criterion are used in place of BIC.

4.3 SVAR Identification Methodologies

As remarked in the introduction, this article evaluates, in face of model uncertainty, the importance of oil supply and aggregate demand shock by revisiting Kilian [2009]'s and Kilian and Murphy [2012]'s empirical results. Both papers investigate the identification of oil structural disturbances based on a small dynamic simultaneous equation model in the form of a structural VAR. Kilian [2009] assumes a zero price elasticity of oil supply and global economy, implying a contemporaneous impact response of real price of oil to oil supply and aggregate demand shock. Based on this hypothesis the author suggests to identify the structural covariance matrix through a short-run recursive identification procedure. On the contrary, Kilian and Murphy [2012] allow the oil production to react after an oil-specific demand shock, provided that the size effect must be positive and fall in the interval $[0, 0.0258]$. According to this assumption, the authors propose to identify the matrix of structural disturbances by imposing sign restrictions on the impact effect of each shock.

Since I aim to evaluate the substantial contribution of Information Criterion model averaging (ICMA) on providing accurate IRFs which address the problem of model uncertainty, it is important to describe how the identification strategies of both papers have been adapted in this model averaging experiment. In addition, this section also reports the description of SVAR identification via heteroskedasticity that I use to capture the changes in variance regimes when additional observations are included in the original dataset. The last two subsections provide the Bayesian methodology that I use to estimate the reduced-form VAR coefficients for each experiment. Additional details about the Bayesian sampling method, are included in Appendix C.2 and C.3

4.3.1 Monthly Dataset

Before presenting the identification methodologies, it is useful to show the credibility of the dataset ICMA relies on. Namely, it is important to demonstrate whether all variables are statistically significant in predicting crude oil prices. Hence, in this subsection, I show in a two steps procedure that the macroeconomic panel used in this study produces consistently reliable predictors. The plan follows Hamilton [2019]. In particular, I firstly regress each measure of crude oil price on three own lags, plus three lags of a single predictor selected from the panel. The linear regression is expressed in the form of an autoregressive (AR) process written as:

$$\Delta p_t^{oil} = c + \sum_{j=1}^{3} \beta_j \Delta p_{t-j}^{oil} + \sum_{j=1}^{3} \phi_j x_{t-j} + \varepsilon_t, \qquad \varepsilon \sim N(0, \Omega) \tag{4.3}$$

where x_t is the time series of the growth/index variable selected from the information set (see the last column on the right of Table 4.1 for the description of all factors considered in this analysis). Secondly, I estimate the joint significance of each candidate x_{t-1}, x_{t-2} and x_{t-3} through the F-statistic test. The null hypothesis is $H0$: $\phi_1 = \phi_2 = \phi_3 = 0$, with a p-value of $\alpha = 0.01$ and a relative F-critical value of $F = 2.89$. Columns 2 in Table 4.1 reports the F-test results of all potential predictors. It can be noticed that the F-statistics of each variable is always greater than the F critical value, and thus the null hypothesis that the factors are of no help for predicting the price of oil can be rejected.

Turning to the information set used in this work, 10 monthly variables are considered. As common practice, all factors have been differenced once before rendering the time series stationary. Data cover the period January 1973 - December 2007 in experiment 1 (revisiting of Kilian [2009]); January 1973 - September 2008 in experiment 2 (revisiting of Kilian and Murphy [2012]); January 1974 - December 2019 in experiment 3 (designed to capture the change of variance regimes when more observations are considered). Section 4.5 provides additional empirical results, where world industrial production is plugged in place of dry cargo index rate, and WTI and Brent are alternatively used in place of oil price based on refiner acquisition costs. In these case of studies observations run from January 1982 to December 2019. Appendix C.1 provides useful information regarding the source of each variable.

Table 4.1: Monthly dataset and relative F-test

Acronym	**Code**	**F-test**	**Description**
WoP	1	3.56	World Oil Production
WiP	1	7.61	World Industrial Production
Inv	2	4.07	Oil Inventories
TOSI	2	6.62	Text Oil Sentiment Indicator
PPI-M	1	6.41	Producer Price Index Metal
PPI-C	1	4.35	Producer Price Index Coal
PPI-IS	1	6.04	Producer Price Index Iron and Steel
Gold	1	4.19	Gold Fixing Price (London)
EX	1	4.57	U.S. / U.K. Foreign Exchange Rate

Note: Column 2 reports the strategy used to transform the variable original value. In particular, if the code equals 1, the series is detrended by considering 100 times the log-difference of the original value. No transformation is applied when the code is 2.

4.3.2 Short-Run Restrictions

Much of the literature regarding identifying oil supply and demand shocks relies on short-run exclusion restrictions. The object of this methodology concerns the identification of a covariance matrix in which the first shock only affects the first dependent variable, the first and the second shock impact on the second dependent variable, and so on up to the last dependent variable, which is affected by all shocks. Kilian [2009], Apergis and Miller [2009] and Bernanke et al. [1997] are relevant examples that apply this identification procedure to study oil price fluctuations. In this work, the first ICMA experiment revisits Kilian [2009] empirical results. I start by generating all possible model combinations and subsequently identify each SVAR by imposing contemporaneous zero restrictions on the covariance matrix. After that, I compute the impulse-response functions related to the first three dependent variables, regardless of the number of factors in each candidate model. Finally, I average 128 3×3 impulse response matrices over the goodness-of-fit of each VAR. Specifically, suppose to start with the following generic reduced form VAR:

$$y_t = c + \sum_{j=1}^{p} \beta_j y_{t-j} + \varepsilon_t \qquad \forall\ t = 1, \ldots, T \qquad \text{and} \quad \varepsilon \sim N(0, \Sigma) \tag{4.4}$$

where y_t is a $N \times 1$ vector of endogenous variables of interest, c is a $N \times 1$ vector of intercepts, β_j describes the $N \times N$ matrices of coefficients, y_{t-j} is a $N \times 1$ matrix of lagged observations, and ε_t is the $N \times 1$ error vector with zero mean and full variance-covariance matrix. $N = 3, \ldots, 10$ depending on which model combination is taken into consideration and t refers to the monthly observations of Kilian [2009], which run from 1973M1 to 2007M12. In order to derive the IRFs, the VAR(p) system in (4.4) needs to be re-specified in a *companion* VAR(1) form. In particular, it is possible to set $\mathbf{e_t'} = (\varepsilon', 0, \cdots, 0)$, $\mathbf{y_t'} = \left(y_t', y_{t-1}', \cdots, y_{t-p+1}'\right)$ and define

$$\mathbf{B} = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_{p-1} & \beta_p \\ I_N & 0 & \ldots & 0 & 0 \\ 0 & I_N & \ldots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \ldots & \ldots & I_N & 0 \end{bmatrix}$$

Therefore, the *companion* form of (4.4) may be written as $\mathbf{y_t} = \mathbf{B y_{t-1}} + \mathbf{e_t}$. By substituting backward for infinite periods (under the hypothesis that the eigenvalues of $\mathbf{B}$ are less than one in absolute value) the system is re-parameterized as: $\mathbf{y_t} = \mathbf{e_t} + \mathbf{B}\mathbf{e_{t-1}} + \mathbf{B^2}\mathbf{e_{t-2}} + \ldots$.

By writing $\mathbf{y_t} = \mathbf{B}(\mathbf{L})\mathbf{e_t}$, where $\mathbf{L}$ is the lag operator and $\mathbf{e_t} \sim N(0, \Sigma)$, the first upper-left $N \times N$ matrices of $\mathbf{B^j}$ describe the effects of ε_t on $\mathbf{y_{t+j}}$. Now, by applying a Cholesky factorization on Σ, we can find the unique lower triangular matrix $\mathbf{A_0}$ such that $\mathbf{A_0}\mathbf{A_0'} = \Sigma$. Since $\mathbf{A_0}$ is lower triangular, the system ends up in a recursive form as follows:

$$\mathbf{y_t} = \mathbf{B}(\mathbf{L})\mathbf{A_0}\nu_\mathbf{t} \qquad \nu_t \sim N(0, I) \tag{4.5}$$

where, in case $N = 10$,

$$\mathbf{A_0}\nu_\mathbf{t} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ \alpha^0_{2,1} & 1 & 0 & 0 & \cdots & 0 \\ \alpha^0_{3,1} & \alpha^0_{3,2} & 1 & 0 & \cdots & 0 \\ \alpha^0_{4,1} & \alpha^0_{4,2} & \alpha^0_{4,3} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \alpha^0_{10,1} & \alpha^0_{10,2} & \alpha^0_{10,3} & \alpha^0_{10,4} & \cdots & 1 \end{bmatrix} \begin{pmatrix} {\nu_t}^{oil\ produced} \\ {\nu_t}^{GDP} \\ {\nu_t}^{oil\ price} \\ {\nu_t}^{4} \\ \vdots \\ {\nu_t}^{10} \end{pmatrix}$$

and ν_t is the independent shock. Without loss of generality, it is assumed that ν_t follows a Gaussian distribution with zero mean and diagonal covariance matrix Λ. Please note, structural shocks related to the seven additional factors, have been marked as ${\nu_t}^4, \ldots, {\nu_t}^{10}$, since in this work I am only considering the structural disturbances generated by the first three dependent variables. Thus, it is not necessary to specify the recursive order, as well as the name, of the remaining predictors.

Identification of structural shocks follows Kilian [2009]. This implies, from a Bayesian perspective, that I assume with certainty that the short-run supply curve is vertical and the price elasticity of oil supply is zero. On the contrary, oil price is assumed to contemporaneously respond to movements of oil production and global economy, although I disregard the values that the parameters α_{ij} may have. This hypothesis is formalized by using a flat prior on $\mathbf{A_0}$ and a weak weighted random walk prior on lagged coefficients. In particular α_{ij} are supposed to follow a t-Student distribution of the form

$$p\left(\alpha^0_{ij}\right) = \frac{\Gamma\left(\frac{\phi+1}{2}\right)}{\Gamma\left(\frac{\phi}{2}\right)\sqrt{\pi\phi\sigma}} \left[1 + \frac{1}{\phi}\left(\frac{\alpha_{ij} - c}{\sigma}\right)^2\right]^{-\frac{\phi+1}{2}}$$

in which, $c = 0$ is the location parameter, $\sigma = 100$ is the scale parameter, and $\phi = 3$ represents the degrees of freedom. Lagged coefficients shrink to zero as the lag p becomes higher with a weak weight of $\lambda_0 = 10^9$ (see Doan [2013] for the choice of Minnesota prior). A general overview of prior and sampling posterior is presented in subsections 4.3.5 and 4.3.6, whereas Appendix C.2 provides more specific details.

4.3.3 Identification through Sign Restrictions

In recent studies, such as Baumeister and Peersman [2013], Lippi and Nobili [2012], Kilian and Murphy [2012] and Peersman [2005], sign restrictions have been adopted as an alternative approach to the exclusion restriction method discussed above for the study of oil price fluctuations. The main difference with respect to this identification methodology lies in the fact that the analyst imposes restrictions on the signs of impact responses, without bounding any parameter of the covariance matrix to have value zero. Therefore, while the lower triangular system just-identifies the reduced form VAR, the inequality restrictions do not identify a sequence of IRFs uniquely, rather, they give rise to a large interval of admissible shocks, in which all results respect the sign rule. The second experiment of this paper revisits Kilian and Murphy [2012] and shows how dynamic effects of a standard 3-variable VAR, identified through sign restrictions, change when the weighted value of all possible impulse-response functions are combined. The restriction methodology follows Rubio-Ramirez et al. [2010]. In particular, assuming that the economy evolves according to equation (4.4), which can be re-parameterized in the following structural forms:

$$A_0 y_t = d + \sum_{j=1}^{p} A_j y_{t-j} + v_t \qquad v \sim N(0, \Lambda) \tag{4.6}$$

where A_0 is a lower triangular $N \times N$ matrix (obtained factorizing Σ in a Cholesky way) displaying the simultaneous relationship among exogenous and endogenous variables, A_j is a $N \times N$ matrix of coefficients related to the lag j operator, d is a $N \times 1$ vector of constants and v_t is the independent shock with zero mean and diagonal variance matrix $\left(\Lambda = diag\left[\lambda_1^2, \lambda_2^2, ..., \lambda_N^2\right]\right)$. In order to uniquely identify all the shocks over the sample 1973M1-2008M9, at least $n(n-1)/2$ additional restrictions are needed. Such "extra" (sign) restrictions are directly imposed when sampling from the contemporaneous structural matrix A_0 as in Baumeister and Hamilton [2015]. Then, the reduced form coefficients $A_0^{-1} A_j$ are estimated by imposing a Minnesota prior similar to the recursive experiment analysed so far.

The identification of the 3×3 nonorthogonalized impulse-response matrix follows Baumeister and Hamilton [2019] and Kilian and Murphy [2012]. This implies that the algorithm identifies (i) a supply shock when q^{oil} and p^{oil} move in opposite directions and an unfavorable oil productivity shock (decrease of $v_t^{oil\ produced}$) drops the level of global real economy; (ii) an aggregate demand shock when q^{oil}, GDP and p^{oil} move in the same direction; (iii) a positive oil-specific demand shock when an increase of $v_t^{oil\ price}$ leads to a decrease of global economy. Signs of impact effects on matrix A_0^{-1} are summarized in Table 4.2 below.

Table 4.2: Inequality constraints imposed to the first three variable responses

q^{oil}	GDP	p^{oil}	Structural shocks
$-$	$-$	$+$	**Oil Supply Shock**
$+$	$+$	$+$	**Aggregate Demand Shock**
$+$	$-$	$+$	**Oil-Specific Demand Shock**

Kilian and Murphy [2012] argued that sign restrictions displayed in Table 4.2 alone, are not enough to identify a structural VAR, because implausible inference outcomes may have the same consistency as more credible empirical results. Therefore, they suggest to impose a low impact effect of oil supply shock on real economy and of aggregate demand shock on oil production; I set both as equivalent to zero as in Baumeister and Hamilton [2019]. Despite the addition of these restrictions, the SVAR is still not identified, because there are infinite ways to achieve the same maximum value of the likelihood function. Therefore, as in Kilian and Murphy [2012], I assume that (i) the positive value of price elasticity of oil supply is bounded with a range of values between $[0, 0.258]$ and (iii) the size effect of an oil-specific demand shock on global real activity must fall in the interval $[-1.5, 0]$. According to the sign rule outlined in Table 4.2 and the additional constraints further imposed, it is now possible to estimate each individual SVAR resulting from the 128 combinations based on the structural covariance matrix

$$\mathbf{A_0} = \begin{bmatrix} 1 & 0 & [0, 0.0258] & 0 & \cdots & 0 \\ 0 & 1 & [-1.5, 0] & 0 & \cdots & 0 \\ - & + & 1 & 0 & \cdots & 0 \\ \alpha_{4,1}^0 & \alpha_{4,2}^0 & \alpha_{4,3}^0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \alpha_{10,1}^0 & \alpha_{10,2}^0 & \alpha_{10,3}^0 & \alpha_{10,4}^0 & \cdots & 1 \end{bmatrix}$$

It is worth stressing that this experiment only considers the dynamic responses of the first three dependent variables of each model; I impose the exclusion restriction methodology on the remaining predictors in order to leave the structural model still partially identified. Reduced-form VAR coefficients are estimated as in the previous identification methodology. Specifically, I assume to have no information regarding $S = \left\{\alpha^0_{1,3}, \alpha^0_{2,3}, \alpha^0_{3,1}, \alpha^0_{3,2}\right\}$, but this time they are sampled form a truncated t-Student distribution in order to respect the sign rule imposed in $\mathbf{A_0}$. For all VAR combinations where $3 < N \leq 10$, any α_{ij} not included in S is sampled from a standard t-Student distribution with $c = 0$, $\sigma = 100$ and $\phi = 3$, as in the short-run recursive case.

4.3.4 Identification via Heteroskedasticity

In the third experiment of this paper I analyse the same three structural shocks, but with a larger sample. The observation period runs from 1974M1 to 2019M12. In this exercise is it important to understand whether oil price fluctuations across the financial crisis affects the standard dynamic responses path outlined in earlier studies. For this purpose, it is convenient to identify each SVAR combination via heteroskedasticity, thereby capturing any change in volatility. This methodology was initially suggested in Rigobon [2003], and then proposed in further studies like Rigobon and Sack [2003], Lanne and Lütkepohl [2008], Lanne et al. [2010] and many others. However, even though the standard approach entails the variance to switch regime according to a hypothetical probability state matrix, this paper adopts a different approach, parallel with Brunnermeier et al. [2017]. Namely, all regimes are *a priori* exogenously identified. In particular, the whole sample period is split in two different sub-samples (see Table 4.3) according to the observed variation in the time series variables. The first sample period (1974M1-2005M12) covers the main economic events of Kilian [2009] and Kilian and Murphy [2012], thus I do not expect any change in the dynamic impulse-response paths. In contrast, the second sample starts one year before the opening period of the financial crisis and ends in December 2019. This choice is based on the assumption that if the new sample was observed in January 2007, by considering the effects of the first-difference operator and by including a large number of lags (between 12 and 24) on each time series, in the best case scenario the first observed value of the dependent variable would be in January 2008. This would imply to inevitably lose the economic response in 2007 and 2008. For this reason, the second sample periods is set to start in January 2006.

The basic model partially follows the stochastic system described in section 4.3.1. In particular, by taking into consideration the structural re-parametrization of the reduced form VAR depicted in equation (4.6):

Table 4.3: Sub-periods when variance switches regime

Start	End	Description
Jan 1982	Dec 2005	**Period Pre-Crisis**
Jan 2006	Dec 2019	**Period Post-Crisis**

$$A_0 y_t = d + \sum_{j=1}^{p} A_j y_{t-j} + \nu_t \qquad \nu \sim N(0, \Lambda) \tag{4.7}$$

Where A_0 is a lower triangular $N \times N$ Cholesky factorized matrix of Σ of equation (4.4) displaying the simultaneous relationship among exogenous and endogenous variables, A_j is a $N \times N$ matrix of coefficients related to the lag j operator, d is a $N \times 1$ vector of constants and ν_t is the independent shock with zero mean and diagonal variance matrix $\left(\Lambda = diag\left[\lambda_1^2, \lambda_2^2, ..., \lambda_N^2\right]\right)$. As stated before, the whole time span $t = \{1 \ldots T\}$ is exogenously separated into $M = 2$ sub-periods and $\mathbb{E}\left(\nu_t \nu_t'\right) = \Lambda_m$ if and only if t is in period $m \in M$; where Λ_m is diagonal. Estimation of IRFs follows the following procedure:

- First, estimate the dynamic relationship determined by A_0 and A_j of each structural VAR combination based on the entire sample and store the resulting coefficient estimates $\hat{A}_0$ and $\hat{A}_j$;
- Second, perform ICMA on both reduced samples and store the variance estimates $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$;
- Then, compute IRFs$_1$ and IRFs$_2$ of the first and second observation segment respectively, by considering the full sample $\hat{A}_0$ and $\hat{A}_j$ coefficients and the related variances $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$;
- Finally, average IRFs$_1$ and IRFs$_2$ in order to have full sample IRFs.

This procedure implies that the variance of structural shocks changes across samples, but the dynamic relationship determined by A_0 and A_j, remains fixed. In this way, IRFs of structural shocks will have the same shapes across the periods, but with a different scale. Since each interval is well defined and continuous, the results will not differ from the endogeneity switching regimes case. An endogenous switch of the variance could have been possible, but since the regimes are few in the data, this choice would have complicated the model, with the possibility of erroneously determining the number of regime switches. This restriction methodology allows for point identification as in the short-run recursive case, in contrast to sign restrictions which do not provide point estimates.

4.3.5 Prior Densities

Prior $p(A_0)$. As mentioned in the last part of each identification strategy, the parameters of structural VARs are estimated in a Bayesian fashion. Starting from the contemporaneous relation matrix A_0, informative priors on single elements $\alpha_{i,j}^0$ are represented in the form of a density function $p\left(\alpha_{i,j}^0\right)$, where a high value of $p\left(\alpha_{i,j}^0\right)$ implies a strong information about the generic i,j element of A_0, whereas $p\left(\alpha_{i,j}^0\right) = 0$ when no useful prior information is available. Single elements inside A_0 are supposed to follow a Student t-distribution, in which the scale parameter values are chosen according to the prior belief that is assumed on the specific elasticity. In case of lower triangular and heteroskedasticity identification all α_{ij} have $c = 0$, $\sigma = 0$ and $\phi = 3$, because it is assumed that nothing is known about the potential value of the parameters. This hypothesis is relaxed in the second experiment, where an initial value for α_{ij} is guessed, according to the sign rule imposed.

Prior $p(\Lambda|A_0)$. Prior information about Λ conditional on A_0 is described by a $\Gamma(\underline{\kappa}_i, \underline{\tau}_i)$ distribution for each λ_{ii}^{-1} reciprocal of element in row i and column i of matrix Λ. Namely:

$$p(\Lambda|A_0) = \prod_{i=1}^{n} p(\lambda_{ii}|A_0)$$

$$\Rightarrow \; p\left(\lambda_{ii}^{-1}|A_0\right) = \begin{cases} \frac{\underline{\tau}_i^{\underline{\kappa}_i}}{\Gamma(\underline{\kappa}_i)} \left(\lambda_{ii}^{-1}\right)^{\underline{\kappa}_i - 1} exp\left(-\underline{\tau}_i \lambda_{ii}^{-1}\right) & \text{for} \lambda_{ii}^{-1} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where, $(\underline{\kappa}_i/\underline{\tau}_i)$ is the prior mean of λ_{ii}^{-1} and $\left(\underline{\kappa}_i/\underline{\tau}_i^2\right)$ is its variance. We set $\underline{\kappa}_i = 0.5$ and $\underline{\tau}_i = \underline{\kappa}_i A_0 S A_0$, where S is the $N \times N$ OLS variance of y_t. It is assumed that $\underline{\tau}_i$ depends on (A_0), while $(\underline{\kappa}_i)$ does not. Appendix C.2 provides a detailed illustration and some suggestions for the choice of values $\underline{\kappa}_i$ and $\underline{\tau}_i$.

Prior $p\left(A_j|\Lambda, A_0\right)$. Prior information regarding the lagged structural coefficients A_j are represented with a Normal conditional probability density function $p\left(A_j|\Lambda, A_0\right)$ where $A_j \sim N\left(\underline{m}_i, \lambda_{ii}^{-1}\underline{M}_i\right)$, which can be summarized as follows:

$$p\left(A_j|\Lambda, A_0\right) = \prod_{i=1}^{n} p(\alpha_i^j|\Lambda, A_0)$$

$$\Rightarrow \; p\left(\alpha_i^j|\Lambda, A_0\right) = \frac{1}{(\pi)^{\frac{\kappa}{2}} |\lambda_{ii}\underline{M}_i|^{\frac{1}{2}}} exp\left[-\left(\frac{1}{2}\right)\left(\alpha_i^j - \underline{m}_i\right)' (\lambda_{ii}\underline{M}_i)^{-1}\left(\alpha_i^j - \underline{m}_i\right)\right]$$

$\underline{M}_i$ incorporates information regarding the Minnesota structure, whose hyperparameters are $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$.

I follow Doan [2013] and set $\lambda_0 = 10^9$ to express the weight on prior values; $\lambda_1 = 1$, which implies that lagged coefficients shrink to zero as the lag order increases; $\lambda_2 = 1$ governs the confidence in other-than-own lags; $\lambda_3 = 100$ makes the constant term essentially irrelevant. In this way parameters related to higher lags shrink to zero and prior information about the intercept is essentially irrelevant. Vector m_i indicates the best guess of value α_i^j before seeing the data, where i denotes the i^{th} structural equation of matrix A_j. Appendix C.2 includes additional details in the choice of hyperparameters $\{\lambda_0, \lambda_1, \lambda_2, \lambda_3\}$.

4.3.6 Posterior Sampling

Posterior $p(A_0|Y_t)$. Based on the prior densities outlined in the previous section, the posterior distribution is given by the product of prior densities conditional on having observed the sample Y_T. More specifically, according to Baumeister and Hamilton [2015] the posterior of $p(A_0)$ can be expressed as:

$$p(A_0|Y_T) = \frac{K_T p(A) \left[\det\left(A\hat{\Omega}_T A'\right)\right]^{\frac{T}{2}}}{\prod_{i=1}^{n} \left[\left(\frac{2}{T}\right)\overline{\tau}_i(A)\right]^{\overline{\kappa}_i}} \prod_{i=1}^{n} \underline{\tau}_i(A)^{\underline{\kappa}_i}$$

where K_T is a function of the data and prior parameters, that allows the posterior density to integrate to unity. It does not depend upon A_0, A_j or Λ and does not need to be calculated to determine the posterior. $\overline{\kappa}$ and $\overline{\tau}$ are the posteriors of $\underline{\kappa}$ and $\underline{\tau}$ respectively (see below for the specific value). $p(A_0)$ is the prior density of matrix A_0 and $\hat{\Omega}_T$ is the variance matrix of reduced-form VAR residuals:

$$\hat{\Omega}_T = T^{-1}\left\{\sum_{t=1}^{T} y_t y_t' - \left(\sum_{t=1}^{T} y_t x_{t-1}'\right)\left(\sum_{t=1}^{T} x_{t-1}x_{t-1}'\right)^{-1}\left(\sum_{t=1}^{T} x_{t-1}x_{t-1}'\right)\right\}$$

for X_{t-1} the matrix of lagged observations.

Posterior $p(\Lambda|A_0, Y_t)$. With the same logic, if the prior of λ_{ii}^{-1} given A_0 is $\Gamma(\underline{\kappa}_i, \underline{\tau}_i(A_0))$, the related posterior is shown to be $\Gamma(\overline{\kappa}_i, \overline{\tau}_i(A_0))$, in which:

$$\overline{\kappa}_i = \underline{\kappa}_i + T/2 \tag{4.8}$$

$$\overline{\tau}_i(A_0) = \underline{\tau}_i(A_0) + 1/2\overline{\zeta}(A_0) \tag{4.9}$$

for $\overline{\zeta}(A_0)$ being the sum of squared residuals resulting from the regression of $\tilde{Y}_i(A_0)$ over $\tilde{X}_i$. In which:

$$\tilde{Y}_i(A_0) = \left[\alpha_i' y_1 \cdots \alpha_i' y_T \quad m_i(A_0)' P_i\right]$$
$$\tilde{X}_i = \left[x_0 \cdots x_{T-1}' \quad P_i\right]$$

for P_i the Cholesky factorization of $\underline{M}_i = P_i P_i'$

Posterior $p\left(A_j|\Lambda, A_0, Y_t\right)$. Finally, the posterior of $p\left(A_j|A_0,\Lambda\right)$ turns out to be a Normal density with the following parameters $\left(\overline{m_i}(A_0), \lambda_{ii}\overline{M_i}\right)$. In which:

$$\overline{m}_i(A_0) = \left(\tilde{X}_i'\tilde{X}_i\right)^{-1}\left(\tilde{X}_i'\tilde{Y}_i A_0\right)$$
$$\overline{M}_i = \left(\tilde{X}_i'\tilde{X}_i\right)^{-1}$$

In summary, the posterior distribution can be expressed in a closed-form expression, since we are assuming that priors follow a proper parametric distribution. This implies that:

$$p\left(A_0, A_j, \Lambda|Y_T\right) = p\left(A_0|Y_T\right) p\left(\Lambda|Y_T, A_0\right) p\left(A_j|Y_T, A_0, \Lambda\right)$$

In each experiment, regardless of the structural identification methodology adopted, the posterior values of A_0 are sampled by using a random-walk Metropolis Hastings algorithm, with a total of 20,000 draws for each SVAR combination, 10,000 of which are discarded. The remaining 10,000 draws of A_0 are used to generate candidate estimates of λ_{ii} from $\Gamma\left(\overline{\kappa}_i, \overline{\tau}_i\left(A\left(\alpha_{ij}^0\right)\right)\right)$ and estimates of α_{ij}^j from $N\left(\overline{m}_i\left(A\left(\alpha_{ij}^0\right)\right), \lambda_{ii}\overline{M}_i\right)$. This means that each exercise runs a total of 1,280,000 draws. Appendix C.3 provides a detailed guide of the posterior sampling procedure.

4.4 Simulation Experiment

This section designs a simulation study that gives an empirical illustration of the finite sample properties of the ICMA methods discussed so far. In doing that, I generate an artificial dataset from a VAR(12), with N=10, which is based on real observations, and analyze the dynamic response of 5 different artificial VARs. Real observations are manipulated making sure that the stochastic system they originate from is described by equation (4.6). The purpose of this exercise is twofold. Firstly, it demonstrates the estimation accuracy of ICMA by comparing IRFs generated with artificial VARs and IRFs based on real observations. Secondly, that ICMA can address the problem of estimation uncertainty with different natures of data specifications.

The simulation experiment proceeds as follows. I first generate artificial data from a VAR(12), with N=10. Second, I set up 5 different artificial VARs which are described by the following structures:

$$\textit{Model 1.}\quad A_0 y_t = d + \sum_{j=1}^{12} A_j y_{t-j} + v_t \qquad \text{with } N = 10,$$

$$\textit{Model 2.}\quad A_0 y_t = d + \sum_{j=1}^{24} A_j y_{t-j} + v_t \qquad \text{with } N = 10,$$

$$\textit{Model 3.}\quad A_0 y_t = d + \sum_{j=1}^{12} A_j y_{t-j} + v_t \qquad \text{with } N = 6,$$

$$\textit{Model 4.}\quad A_0 y_t = d + \sum_{j=1}^{36} A_j y_{t-j} + v_t \qquad \text{with } N = 10,$$

$$\textit{Model 5.}\quad A_0 y_t = d + \sum_{j=1}^{12} A_j y_{t-j} + v_t \qquad \text{with } N = 3,$$

in which y_t includes the artificial generated data, A_0 and A_j are the parameters to be estimated, and v_t is the structural disturbance. Although in reality the researcher has no prior belief about which particular model generates the best estimates, here I assume that *Model 1* is the correct specified VAR, and the remaining models have some kinds of misspecifications. Third, I estimate the reduced-form coefficients and IRFs of each model by applying the ICMA procedure discussed above. Finally, I compare IRFs of each model with IRFs of the *true* VAR.

In short, it is evident that *Model 2* and *Model 4* incorporate more information, since I allow both models to have a (larger) number of lagged values equivalent to 24 and 36 respectively. *Model 3* omits the last four factors, and *Model 5* keeps only the first three variables. Under these assumptions, *Model 1* is qualified as being the correct estimated VAR.

This Monte Carlo (MC) simulation exercise, runs 500 parallel chains over 20,000 draws (10,000 of which are discarded) for each case of study. This means that 500 different datasets are generated, and thus 25,000,000 distinct posterior coefficients are estimated. Listing all possible configurations of $\{\hat{A}_0^{-1}\hat{A}_j\}$

and $\hat{\Sigma}$ is futile, as well as impractical. The goal here concerns the evaluation of structural responses. For this reason, and for the sake of brevity and simplicity, it is important to compare the impulse-response functions of *Model 1*, which is the correct specified model, with the dynamic response of the *true* VAR for all three experiments discussed in section 4.3. The reader is addressed to Appendix C.4 if interested to see the dynamic response of the remaining models. I only anticipate that, in conformity with the assumptions, misspecified VARs generate incorrect impulse-response functions in comparison to *Model 1* and consequently to the *true* VAR.

Before plunging into any empirical example, it is also important to compare the goodness-of-fit, in face of model complexity, of each case of study. Even though ICMA embeds all possible combinations of a given VAR, it is impossible to list the goodness-of-fit of each model for every case of study. However, it is worth noting which starting model displays the most appropriate dimensionality. I do that, by computing (i) Bayesian Information Criterion (BIC), (ii) Akaike Information Criterion (AIC) and (iii) Hannan Information Criterion (HIC) of the five models proposed in this simulation study.

Firstly, it is important to consider, that *Model 2* and *Model 5* admit a smaller number of variables, and comparing the BIC values with the remaining VARs would not make much statistical sense. Therefore, I first report BIC of all cases out of the second and the last, and then i estimate the goodness-of-fit of each model taking into consideration just the first three variables. Please note, that in the latter experiment I would have to estimate three equivalent models if I did not modify at least the number of lags. For this reason, I assume that *Model 3* and *Model 5* have 20 and 30 lags respectively. The results are reported in Table 4.4. Other than AIC in *Model 4*, the outcomes appear to prefer *Model 1*, as it displays the lowest goodness-of-fit value.

Before showing the graphical results, please note that this analysis is using artificial generated data, and therefore, the path displayed by the IRFs does not necessarily need to have an economic intuition. The object of this discussion is merely to demonstrate the estimation accuracy of the methodology presented in this paper.

Figure 4.1 displays a comparison between IRFs of *Model 1* and the median response of *true* VAR (red dotted lines). It should be considered that the dynamic response of the *true* VAR is based on OLS coefficient estimates of a VAR(12) with 10 dependent variables. In contrast, IRFs of *Model 1* (and of the remaining artificial VARs) are generated from the reduced-form coefficients of any possible model combination of the

Table 4.4: BIC, AIC and HIC for **Model 1-5** when n=10 and n=3

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
***n*=10**					
BIC	-0.9457	0.1412	*NA*	-0.4325	*NA*
AIC	-2.0593	-2.0020	*NA*	-3.1811	*NA*
HIC	-1.5877	-1.0402	*NA*	-1.3877	*NA*
***n*=3**					
BIC	2.0710	2.4684	2.2890	2.9349	2.7087
AIC	1.7065	1.7330	1.7169	1.8113	1.7815
HIC	1.8508	2.0244	1.9207	2.2570	2.1491

NB: Each column shows the quality of each case of study computed through BIC, AIC and HIC. *NA* indicates that the experiment has not been done for that specific case.

starting artificially generated VAR(12) with $N = 10$. Therefore, the median path of true IRFs and artificial IRFs can never perfectly match. Put differently, suppose I estimate the artificial coefficients of a model combination in a VAR(12) with $N = 10$, when $N = 4$. The explanatory power of the $\hat{A}_j$ coefficients in the upper left 4×4 matrix of the companion VAR, is inevitably different from the one reported in the first 4×4 upper left companion matrix of the *true* VAR. This implies, that the asymptotic accuracy of ICMA is tested positive if the median response of *true* VAR (red dotted line) falls inside the 68/95% of credible set of the artificial VAR responses. Looking at Figure 4.1 it is evident that this condition is satisfied and thus it can be inferred that *Model 1* generates IRFs that are asymptotically efficient when the SVAR is identified with short-run recursive restrictions.

Figure 4.2 and Figure 4.3 report the same exercise, when the artificial generated SVARs are identified through sign restrictions and heteroskedasticity respectively. Appendix C.4 provides a detailed representation of all IRFs for each simulation. It can be inferred that among all the misspecified models, *Model 2* exhibits the worst IRFs. Hence, the loss of information related to the lag length, weights more than dropping a specific number of dependent variables.

4.5 Empirical Results

Before analysing the empirical evidence, it is useful to formulate a premise. Kilian [2009] and Kilian and Murphy [2012] both use a proxy of global economic activity based on dry cargo shipping rate. This measure has been revised following Hamilton [2019] and it is now based on the residuals of the regression

$$x_t - log(CPI) = \alpha + \beta t + \varepsilon_t$$

where x_t is the nominal cost of bulk dry cargo shipping, CPI is the US consumer price index, t is a linear time trend and the residuals ε are the index of real economic activity (see Hamilton [2019]). Even though the standard shape of structural shocks remains unchanged, following this update, the impact size of some disturbances is now slightly different and the reader is addressed to check Figure C.11 and C.12 in Appendix C.5, because the first two experiments are investigated with Kilian's original and updated index.

4.5.1 ICMA with Original Data

For each experiment, impulse-response functions are computed in two ways. Firstly, I identify a misspecified 3-variables VAR whose posterior median response matches with Kilian [2009]'s and Kilian and Murphy [2012]'s, and store the results. Then, I extract 128 different VARs from my dataset, estimate all reduced-form coefficients, compute 128 different impulse-response functions, weight the posterior estimates according to equation (4.2), and finally compare ICMA response (gray shaded regions) with the misspecified VAR response (red dotted lines). The 3×3 nonorthogonalized impulse-response matrix is formalized as follows:

$$\begin{bmatrix} \nu^q & \alpha_{q,y} & \alpha_{q,p} \\ \alpha_{y,q} & \nu^y & \alpha_{y,p} \\ \alpha_{p,q} & \alpha_{p,y} & \nu^p \end{bmatrix}$$

Figure 4.4 reports IRFs of experiment 1, where Kilian [2009] in revisited in face of model uncertainty. Even though both posterior credible sets have a similar general path within the first year, in some cases the remaining five months display a different response. For example, after an aggregate demand shock, real oil prices in the misspecified model show a double increase after one year with respect to the equivalent disturbance under model uncertainty. Similarly, but in the opposite direction, after an oil-specific demand shock the misspecified VAR shows a decrease in the variation right after the 17^{th} month. Whereas, under model uncertainty, the shock is more persistent and does not show any sign of future decrease. In regards to the magnitude of each shock, all variables have an impact size equivalent to the misspecified model.

Figure 4.5 reports IRFs of experiment 2, where Kilian and Murphy [2012] is the study revisited. Results are consistent with the previous experiment, although more exogenous disturbances show a different size response when ICMA effects are compared to those of the misspecified VAR. In particular, in case of a 3-variable VAR, real oil prices after an aggregate demand shock show a greater response than the equivalent disturbance generated under model uncertainty. In contrast, after an oil-specific demand shock, the real oil prices response in Kilian and Murphy [2012] goes to zero at the end of the period, while under model uncertainty it does not. In conclusion, the difference between ICMA-VAR and misspecified VAR structural responses in experiment 2 is larger in comparison to the previous experiment. A logical explanation can be found in the identification strategy used. In fact, as remarked in section 4.3, sign restrictions do not identify a structural response uniquely, but different results can maximize the vale of the likelihood function as long as the sign rule imposed in matrix A_0 is true.

Figure 4.6 reports IRFs of experiment 3, where observed data running in Kilian [2009] and Kilian and Murphy [2012] are extended up to 2021M12. In this exercise, other then addressing the problem of model uncertainty, the identification method is also useful in order to take into consideration two different levels of variance occurring pre and post the global financial crisis. The differences between misspecified and ICMA model are much more clear in this exercise. In fact, Figure 4.6 shows that the impact of oil supply on oil production is stronger when additional variables are considered. The opposite applies at the same dependent variable after an aggregate demand and oil-specific shock. Differences on dynamic effects on real economy are mostly negligible, whereas for real oil prices it is possible to see that the size of the structural shock completely differs between misspecified and ICMA model. In particular, after an oil-specific shock a standard 3-variable VAR would suggest a positive increasing impact on oil prices up to 0.84% on the 18^{th} month. Whereas, under model uncertainty, oil prices go up to 1.77 percentage points. Moreover, following an oil supply and aggregate demand shock the misspecified model is completely missing half of the 95% posterior credible set. Therefore, by considering additional variables who do have an explanatory power to predict real oil prices, it is clear that the reaction of oil price following a structural shock is larger in comparison to what was previously assessed in the economic literature. A concrete evidence in favour of this experiment is also given by looking at the economic recovery after the Covid pandemic between April and July 2020. In this period, global real economy registered a growth of almost 2.9%, after a drop of 1.5% in 2020Q1. This positive variation yielded a raise of refiner acquisition costs of imported oil in the US of about 28.5%. Such change is very close to the empirical results reported in Figure 4.6 for $\alpha_{y,p}$, which raises by 27.8% in the first quarter after the shock under model uncertainty, in comparison to the misspecified model that registered a raise in oil prices of about 12.5%.

4.5.2 ICMA with Updated Variables and Data

Although Kilian [2009] and Kilian and Murphy [2012] represent two important contributions to the academic literature related to oil shocks, recent studies suggest the use of alternative measures to better understand the role of oil supply and demand shock. In particular, Hamilton [2019] shows that world industrial production (WIP) offers a better monthly measure of global real economy than Kilian's index, which is based on shipping costs. Baumeister and Hamilton [2019] is another good example whereby the authors point out the limits of expressing the real price of oil in terms of the U.S. refiner acquisition costs and suggest the growth rate of West Texas Intermediate (WTI) price as a valid alternative.

In this section I replicate the previous exercises by expressing y^{GDP} and p^{oil} with alternative measures. Two experiments are investigated. First, following Baumeister and Hamilton [2019] I consider a 4×4 nonorthogonalized impulse-response matrix, where also oil inventories are included. The misspecified VAR is then compared with ICMA empirical results when WIP and WTI proxy for y^{GDP} and p^{oil} respectively. In this way the dataset can be extended up to December 2021, and take into considerations the devastating effects of Covid pandemic on real economy in general, as well as real oil prices in particular[6]. Moreover, the experiment is also replicated by using Baumeister et al. [2020]'s index GECON in place of WIP. For both experiments a larger dataset is used, where data run from 1982M1 to 2021M12.

Figure 4.7 and Figure 4.8 report the structural responses of ICMA (black lines) vs the 4-variables VAR (red dotted lines). Both experiments display the same peculiarities found in the first three exercises. In fact, in Figure 4.7, an oil-specific demand shock on real oil price approaches to zero at the end of the horizon considered when the model is misspecified, whereas under model uncertainty it does not. The same characteristic holds for oil supply shock on real oil price. Then, an aggregate shock on oil price increases after one year and half when the model is misspecified, while under model uncertainty the disturbance is more persistent. What is odd in this experiment is the real economy reaction after each shock, which shows a less significant response than Kilian's index. If GECON proxies for real economy, quite surprisingly, the behaviour of real oil prices following an aggregate demand shock when the VAR is informationally deficient, does not change under model uncertainty. But, when the shock is determined by oil demand, the

[6] After COVID-19 outbreak, following government dispositions in most of the countries, non essential businesses shut and industrial activities slowed. This generated a drop in the consumption of energy, which led oil prices in April 2020 to turn negative for the first time in history.

real oil price response converges again to zero in the misspecified VAR, while under model uncertainty it does not (see Figure 4.8).

4.5.3 Oil Sentiment Shock

Based on Table 4.1, excluding real activity, TOSI is the factor that best explains the behaviour of real oil prices. This is because news are able to affect individual thinking and rational decisions in financial markets (Fang and Peress [2009], Peress [2014]), especially when the subject is the price of oil, which is by nature characterised by a high level of uncertainty and volatility. It is therefore essential to understand what happens when the economy is disturbed by an oil-related news shock. Namely, how oil production, real economy, oil prices and oil inventories react following a variation determined by oil articles. To the extent of my knowledge, the impact that oil news have on oil market fundamentals has never been demonstrated. In this section I investigate this experiment firstly through a misspecified 5-variable VAR, and then under model uncertainty. In this experiment West Texas Intermediate index is used as a proxy of the price of oil, while world industrial production proxies the global real economy. The structural VAR is identified through short-run restrictions and the factor that captures human sentiment from oil related news items is placed at the bottom of the dependent variables matrix. The motivation behind this decision is based on the assumption that newspapers are delivered in real time, and therefore they react immediately to any shock. In contrast, individual decisions that are influenced by the content of news items, will affect oil market fundamental with a lag.

Figure 4.9 depicts a misspecified 5-variable VAR and the evidence shows that following an oil news shock, as expected, the price of oil drops by almost 1% after one month, and then the variable follow a highly volatile trajectory in the following months. This might be interpreted as the impetus to realize profits by selling Crude Oil contracts, which shifts the price of oil downward, thereby decreasing oil production. Are the results robust to different variable specification? This question is addressed in a second experiment in which the 5 dependent variables used in the previous exercise remain fix, and then I consider any possible variable combination based on factors listed in Table 4.1.

Empirical evidence is reported in Figure 4.10. On average, for oil market fundamentals both the misspecified and the complete model depict similar impulse response functions. However, accounting for model uncertainty is beneficial to understand how news react following an oil supply, aggregate demand and oil-specific demand shock. In fact, based on the last column of Figure 4.10 it can be inferred that the misspecified model is missing 50% of the posterior credible set.

4.6 Conclusions

In this study the usefulness of Bayesian Information Criterion model averaging proposed by Kass and Raftery [1995] is used to revisit the role of (i) oil supply, (ii) aggregate demand and (iii) oil-specific demand shock under model uncertainty. The evidence shows that following an oil-specific demand shock, oil price response converges to zero after eighteen months when the model is misspecified, whereas under model uncertainty the disturbance is shown to be more persistent. In addition, when uncertainty is pervasive in the empirical results, after an aggregate demand shock, the rise of oil constantly increases even after one year and half. In contrast, under model uncertainty, oil prices rise at the impact but the positive response is constant over the months which follow. Moreover this study highlights the potential explanatory power of oil news. In particular, it is shown that variations caused by the content of news items generate considerable effects on oil prices, which lead to a reduction of oil production.

Overall this analysis highlights that model uncertainty is pervasive in empirical results. In particular, there are additional macroeconomic variables which affect the volatility of oil prices that have never been considered before. Thus I suggest that averaging across all possible outputs is useful to generate more accurate impulse-response functions. A number of extensions of this research are possible, of which only a few are mentioned here. Firstly, I have imposed a flat prior on the additional 7 factors, but it would be interesting to inform the parameter c (based on historical data) of each variable and generate more reliable impulse responses. Secondly, for the sake of robustness, more weighting schemes and alternative variables should be investigated, in order to make a more substantial comparison between a misspecified and full-specified model. This task is left for future research.

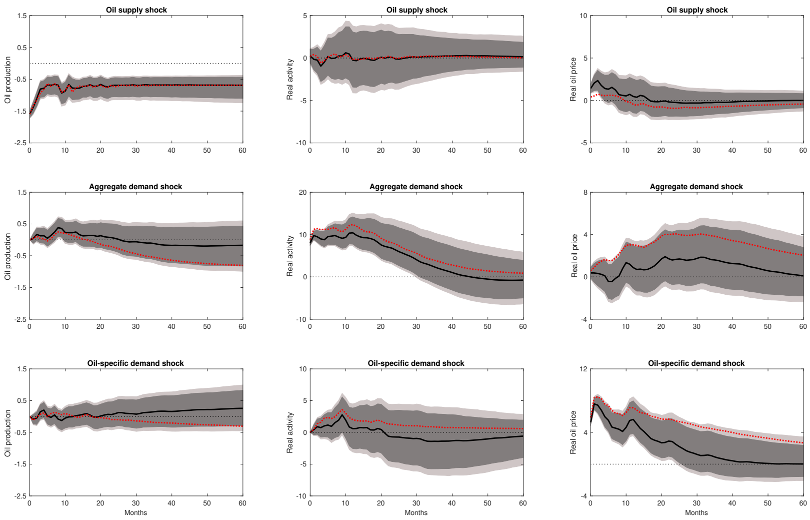


Figure 4.1: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through short-run restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 98% posterior credible set. Red dotted lines show the response of the true model.

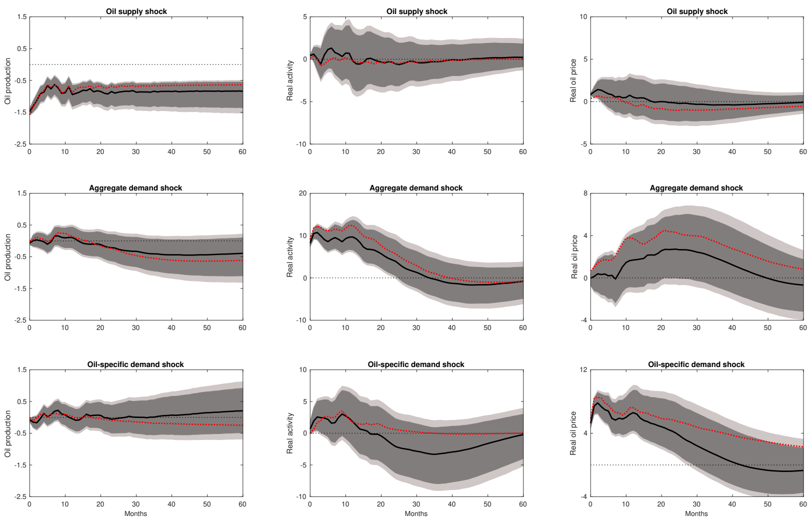


Figure 4.2: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through sign restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 98% posterior credible set. Red dotted lines show the response of the true model.

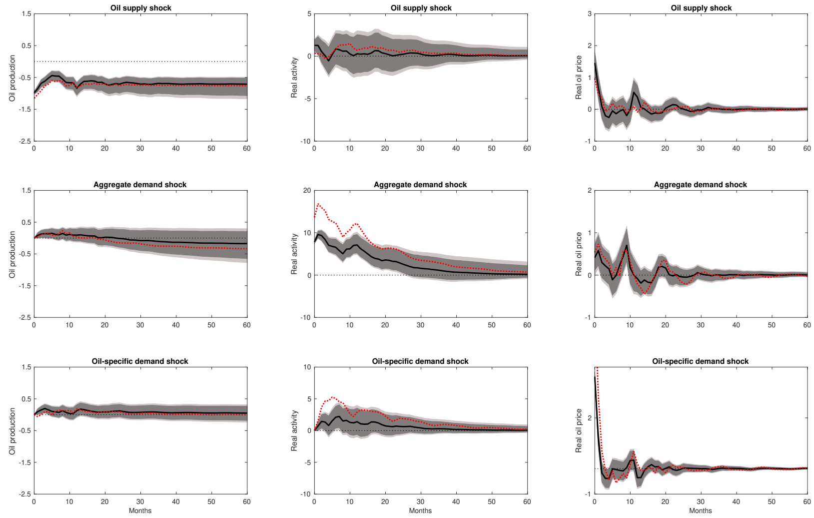


Figure 4.3: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified via heteroskedasticity restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 98% posterior credible set. Red dotted lines show the response of the true model.

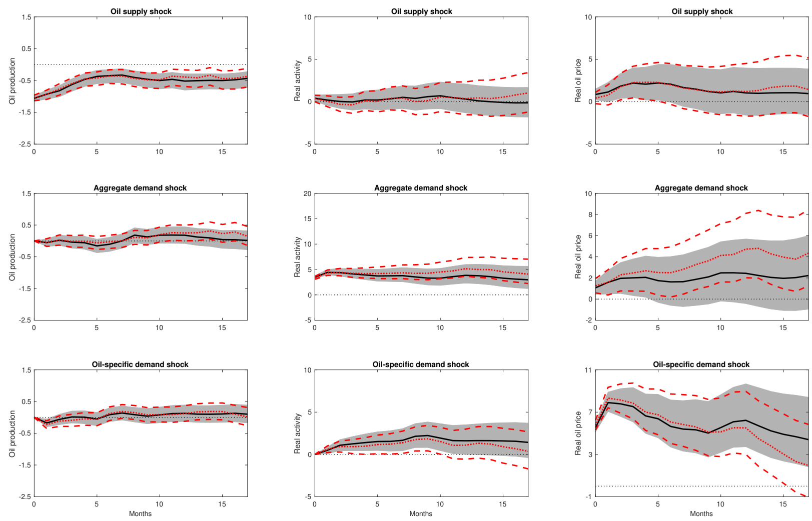


Figure 4.4: IRFs of the fixed 3×3 matrix, where SVARs are identified through short-run restrictions. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 3-variables misspecified VAR.

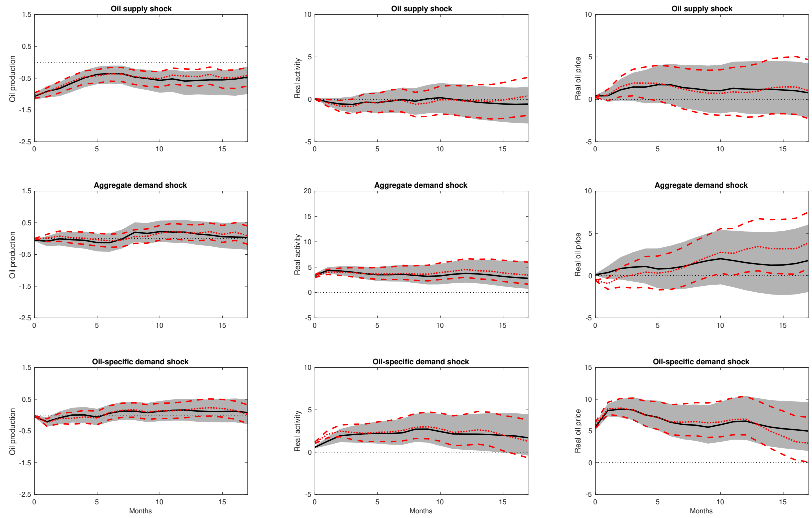


Figure 4.5: IRFs of the fixed 3×3 matrix, where SVARs are identified through sign restrictions. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 3-variables misspecified VAR.

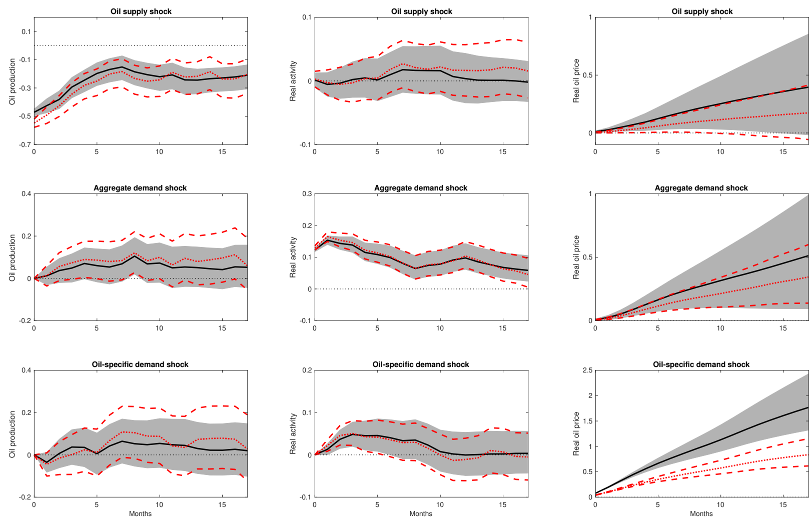


Figure 4.6: IRFs of the fixed 3×3 matrix, where SVARs are identified via heteroskedasticity restrictions. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 3-variables misspecified VAR.

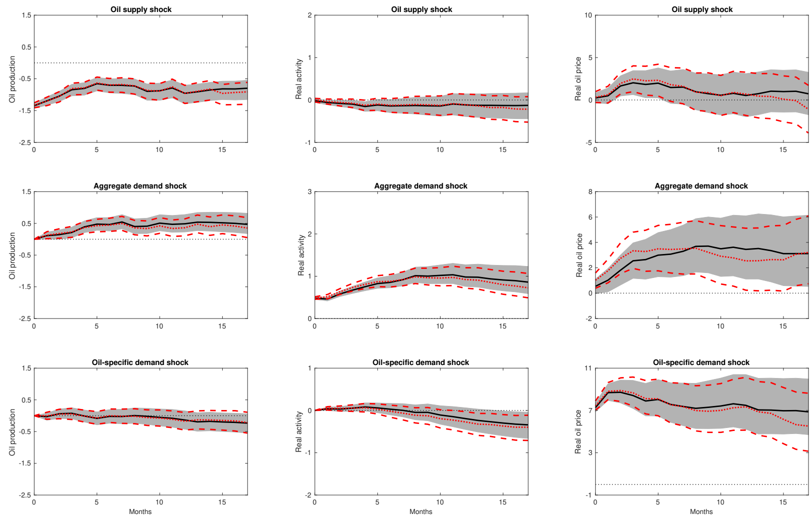


Figure 4.7: IRFs of the fixed 3×3 matrix, with WIP and WTI describing global economy and real oil price. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 4-variables misspecified VAR.

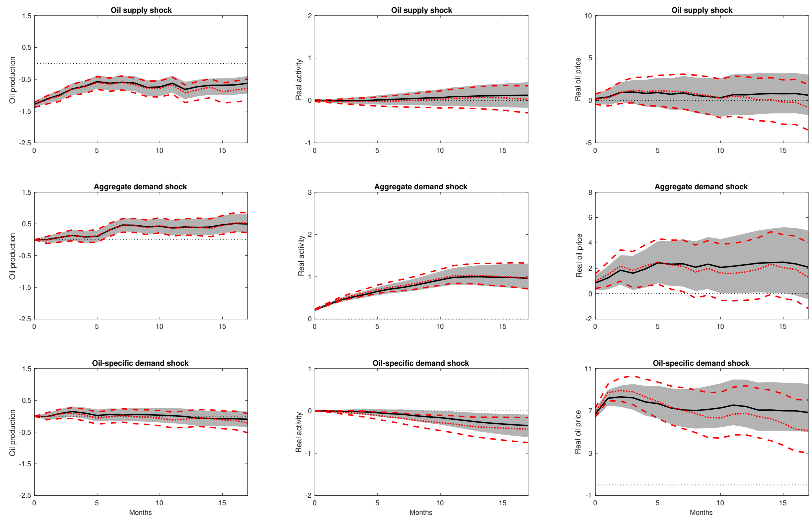


Figure 4.8: IRFs of the fixed 3×3 matrix, with GECON and WTI describing global economy and real oil price. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 4-variables misspecified VAR.

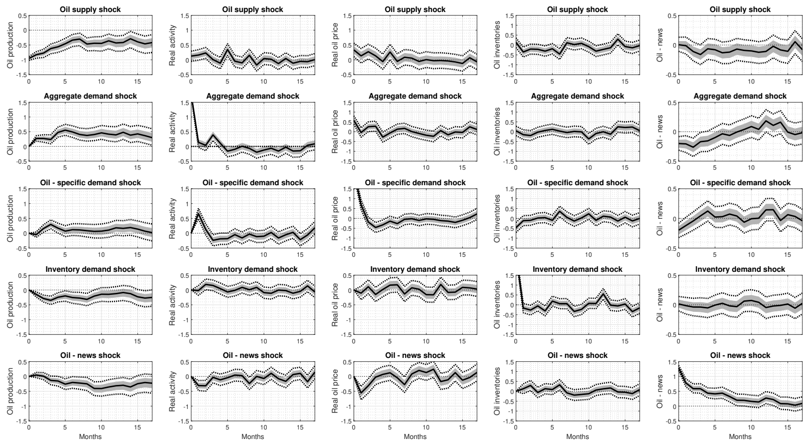


Figure 4.9: IRFs of the misspecified 5×5 matrix, with WIP and WTI describing global economy and real oil price. Black solid lines show the median responses, shaded regions and black dotted lines describe the relative 68% and 95% posterior credible set.

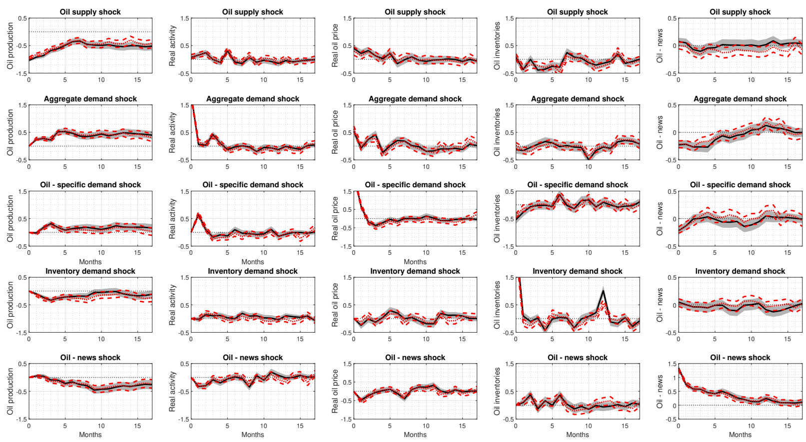


Figure 4.10: IRFs of the fixed 5×5 matrix, with WIP and WTI describing global economy and real oil price. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 68% posterior credible set. Red dotted lines show the response of the 5-variables misspecified VAR.

Concluding Remarks

5.1 Summary of Contributions

The empirical literature focusing on analysing the dynamic behaviour of crude oil prices is built on models where only a small number of low-frequency variables interact. Said indicators are made available by government agencies, usually with a delay of weeks. Some of them are even subject to revisions that can take up to two years. Furthermore, and perhaps most importantly, they are by nature slow to respond to global specific events, such as political unrest and natural disasters.

In this thesis I take the literature on natural language processing as a medium to develop non-traditional economic variables as my starting point. I demonstrate that digital text is a valuable source of information, able to capture the dynamics of crude oil prices. The textual database includes oil related articles retrieved from the Banking, Finance and Energy section of The Financial Times, Thomson Reuters and The Independent. I use the title and full body of roughly 140,000 news items to develop thirteen text based indicators. Nine of these are designed to capture the human sentiment and the remaining four aim to assess the uncertainty in the oil market. Overall, what I find can be summarized in the following three points.

First, sentiment indicators readily react to economic and geopolitical events affecting the price of oil. This enables said indicators to interact with oil market fundamentals and improve the predictions of crude oil prices. In contrast, uncertainty measures show structural weaknesses in the time series which generate unreliable oil price forecasts. A new text oil sentiment indicator (TOSI) is developed through a non-parametric combination between the best performing human sentiment indexes. I show that, by endogenizing TOSI, oil production, global real economy, real oil prices and oil inventories in a stochastic volatility Bayesian vector autoregression (SV-BVAR) model, the out-of-sample forecasts of the monthly price of oil improve significantly. Indeed, as suggested by the Diebold-Mariano (DM) test, results are 1%, 5% and 10% statistically significant for short-, medium- and long-term forecasts respectively. Such improvements on the no-change forecast are particularly evident when the economy runs periods of financial instability, which is when forecasting matters the most.

Second, daily and weekly text data can be combined with commodity and financial variables, as well as with oil market fundamentals observed at a monthly frequency. The resulting mixed-frequency models can be used to predict the monthly price of oil. However, neither mixed-data sampling (MIDAS) or mixed-frequency VAR (MF-VAR) models yield significant forecasting improvements in comparison to the corresponding model with variables sampled at the same frequency. This is true for point and density forecasts. In particular, the preferred mixed-frequency model reduces the minimum sum of prediction errors by 18% in the short run, but according to the DM test such improvements are not statistically significant.

Third, in a framework of impulse response analysis modelled with a large Bayesian VAR (BVAR), and by taking the model uncertainty into consideration, I show that Kilian [2009]'s confidence sets are too optimistic. Specifically, I use the information criteria model averaging (ICMA) to address the problem of information deficiency in a BVAR. I show that a demand-specific oil shock increases the price of oil on impact, yet in the long term the positive variation is more persistent and does not converge to zero as in the commonly used 3-variable VAR model. Moreover, the oil price response resulting from an aggregate demand shock, is more stable in the long run and does not converge to infinite as reported in Kilian [2009]. Finally, I also show that the dynamic stability of crude oil prices is considerably affected by oil news shocks. Such variation generates a persistent drop in the quantity of oil produced at a world level.

5.2 Avenues for Future Research

There are still many challenges to overcome when studying the dynamics of crude oil prices through text data. In the future, it would be interesting to extend the textual database used in this thesis to incorporate oil related articles retrieved from additional newspapers (e.g. The Wall Street Journal, The New York Times, and The Economist to name a few). Text indicators presented in this work implicitly assume that agents can only read The Financial Times, Thomson Reuters and The Independent. This implies that some bias is pervasive in the empirical results. By including the most widely read newspapers in the text analysis, it would be possible to decrease the bias even further.

The literature focusing on forecasting oil prices through variables sampled at different frequencies also requires further research. In Chapter 2, I show that not much information is lost when high-frequency variables are not used to forecast the monthly value of real oil prices. However, it would be interesting to understand whether daily text data can yield more accurate forecasts of the weekly crude oil prices. This research question can also be extended to the weekly forecast of other commodity variables that have played a major role in inflation's surge in the last few years (e.g. natural gas).

Furthermore, there is a rapidly growing area of research aimed at converting audio and videos to text data in order to extract human sentiment and emotion recognition (Poria et al. [2016], Li et al. [2022]). To date, the idea of using audio and videos as a source of information for examining the behaviour of crude oil prices is a relatively unexplored area. It would be therefore interesting to understand whether audio signals bring new knowledge and can improve the prediction performance of existing models. I plan to further contribute to this line of research.

Appendix
Chapter 1

A.1 Data Sources

As remarked in section 2.2, 138,797 articles that featured in the Banking, Finance and Energy section of the following newspapers:

Source	No. of Articles	Period
The Financial Times	103,966	1982M01-2021M11
The Independent	28,103	1988M09-2021M11
Thomson Reuters	6,728	2002M11-2021M11

are retrieved from the LexisNexis database. Articles are selected based on the joint occurrence of the words *oil* and *price*. Global oil production (Qoil) is considered in millions barrels per day, and world industrial production (WIP) is used as a proxy of global real economy. Oil inventories (Oinv) rely on the ratio between OECD petroleum inventories and US petroleum inventories. The result is then multiplied by US crude oil inventories and normalized by the monthly production of oil (see Baumeister and Hamilton [2019] for more details).

The out-of-sample forecast of three oil price measures is investigated. West Texas Intermediate and Brent crude oil price are downloaded in nominal values from the Federal Reserve Economic Data (FRED) database maintained by the St. Louis FED. Data are then normalized by the US consumer price index in order to obtain the respective real values. The third measure of oil price is based on the refiner acquisition cost (RAC) of imported oil, available on the Energy Information Administration website. Data transformation is reported in the table below.

Variable	Transformation
Qoil	100*Log-diff
WIP	100*Log-diff
Oinv	100*Log-diff
Text-Data	Raw
Brent	Raw
WTI	Raw
RAC	Raw

A.2 Text Metrics Indicators

This section provides the time series of the text based indicators developed in section 2.3. Monthly data run from January 1982 to November 2011.

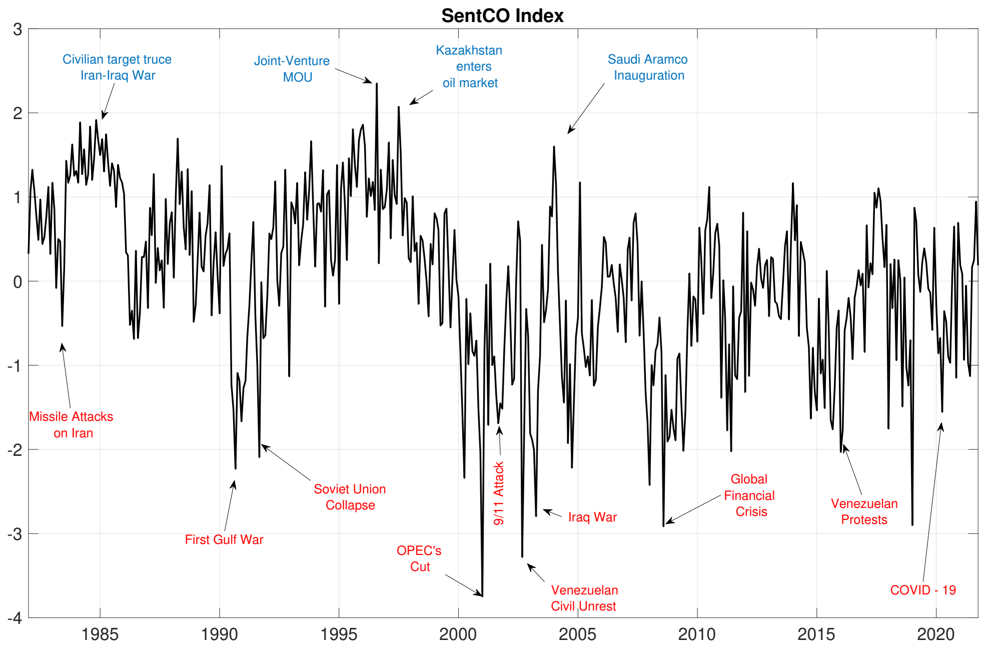


Figure A.1: SentCO is a unigram human sentiment based index, developed by counting the number of words “economy”, “economies”, “economic”, “economics” occurring in each article, normalized by the total number of running words. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

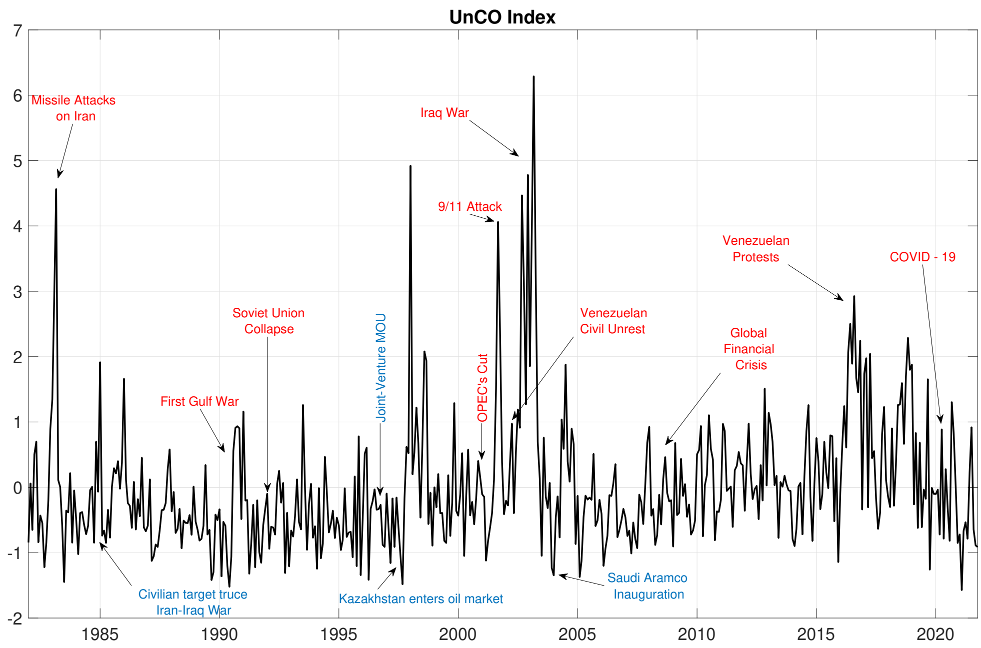


Figure A.2: Unigram UnCO accounts for uncertainty in oil market by counting the number of words "uncertain", "uncertainty", "uncertainties", "uncertainly" occurring in each article, normalized by the total number of running words. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

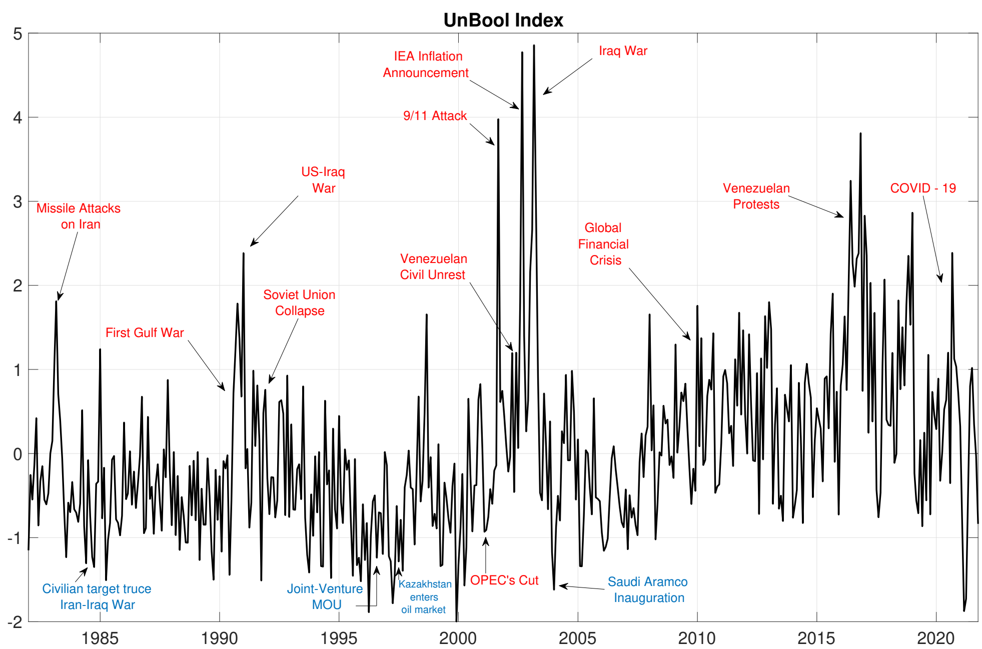


Figure A.3: UnBool index accounts for uncertainty in oil market and is based on the Boolean count as in Baker et al. [2016]. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

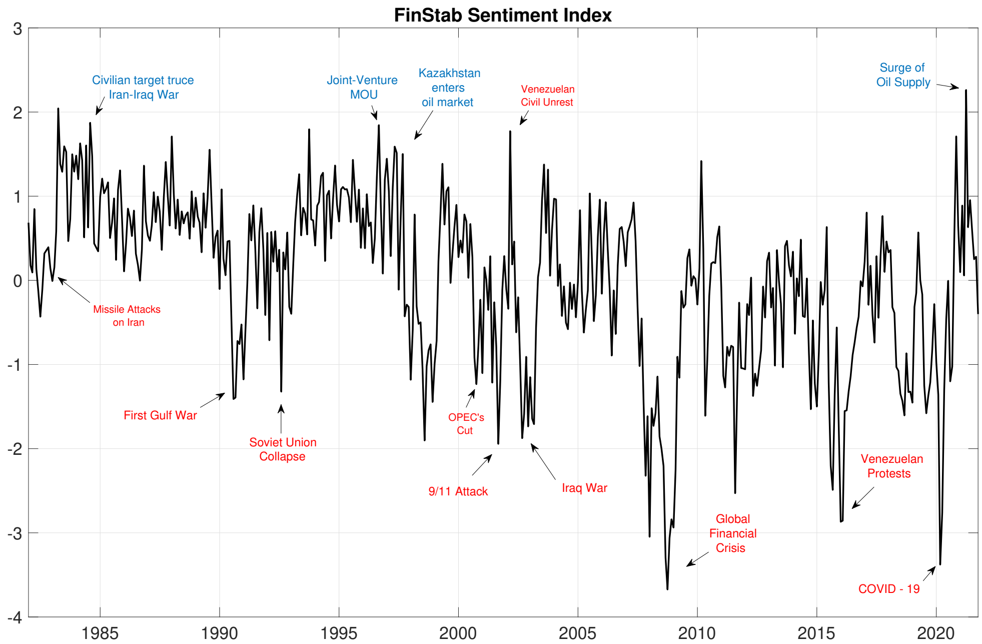


Figure A.4: FinStab is a human sentiment index where words are assigned a numeric value between $[-1;1]$ based on Correa et al. [2017]'s dictionary. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

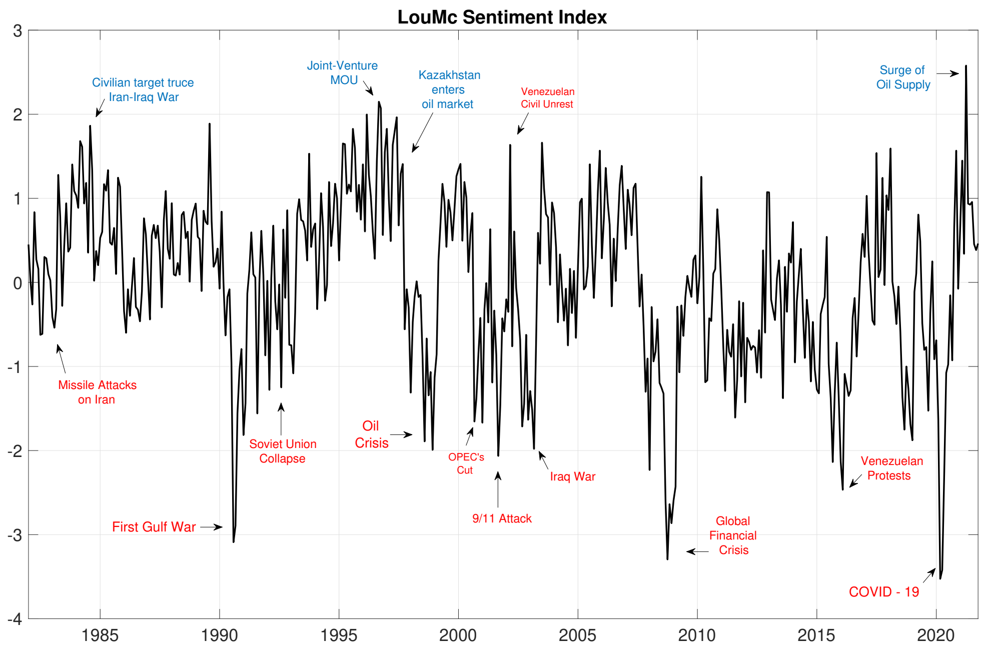


Figure A.5: LouMc is a human sentiment based index where words are assigned a numeric value between $[-1;1]$ based on Loughran and McDonald [2011]'s dictionary. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

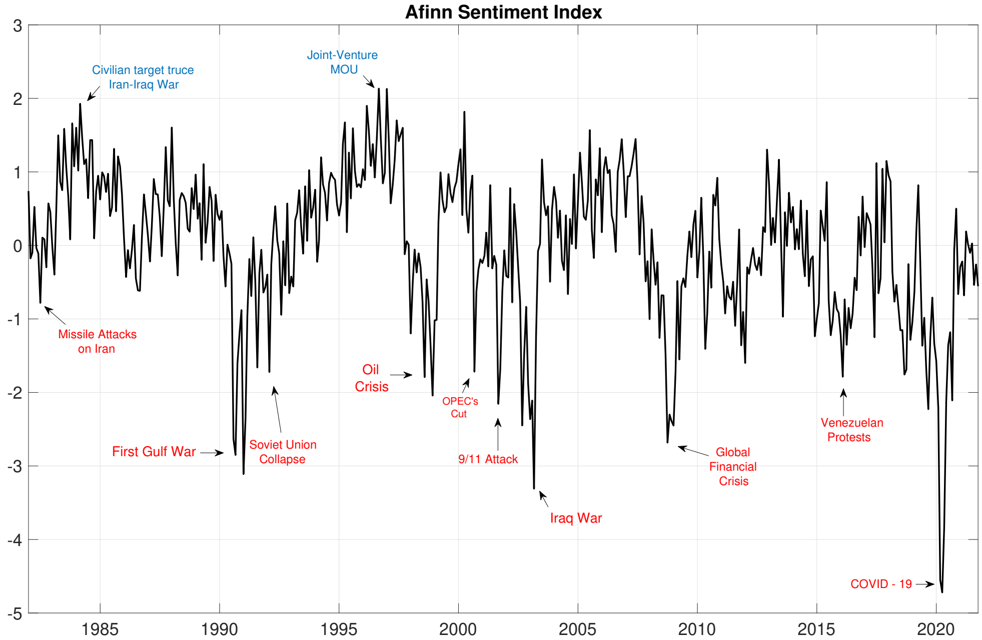


Figure A.6: Afinn is a human sentiment based index where words are assigned a numeric value between $[-5;5]$ based on Nielsen [2011]'s dictionary. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

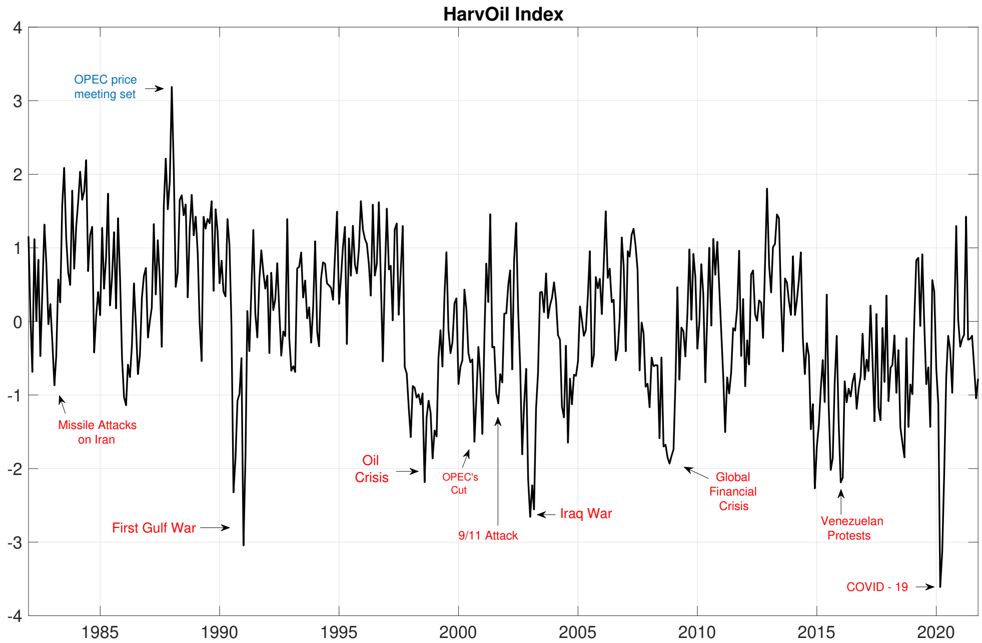


Figure A.7: HarvOil is a human sentiment based index where words are assigned a numeric value between $[-1;1]$ based on Harvard IV dictionary. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

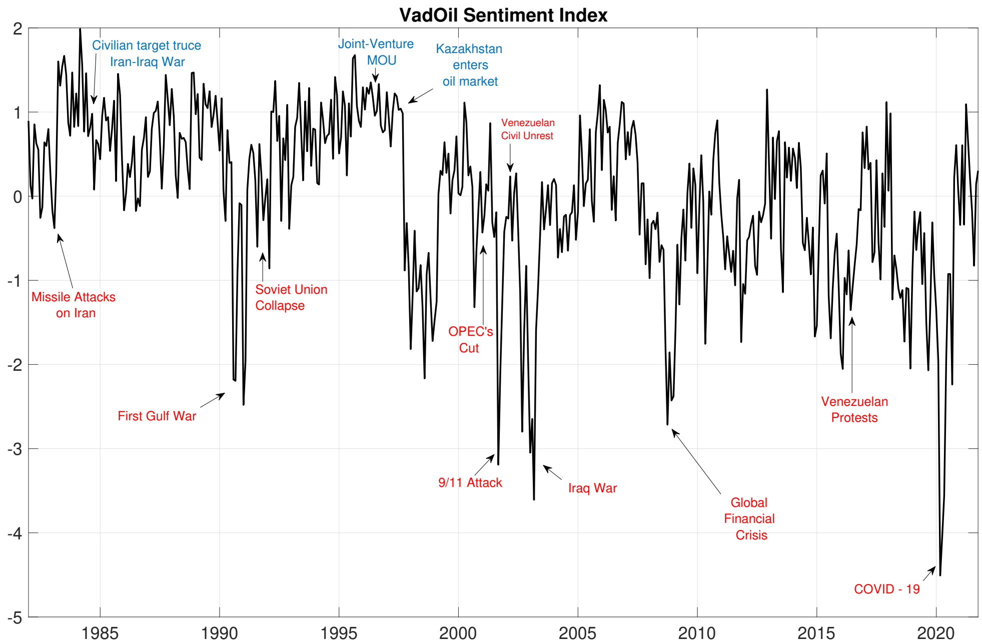


Figure A.8: VadOil is a human sentiment based index where words are assigned a numeric value between $[-4;4]$ based on VADER dictionary (see Hutto and Gilbert [2014]). The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

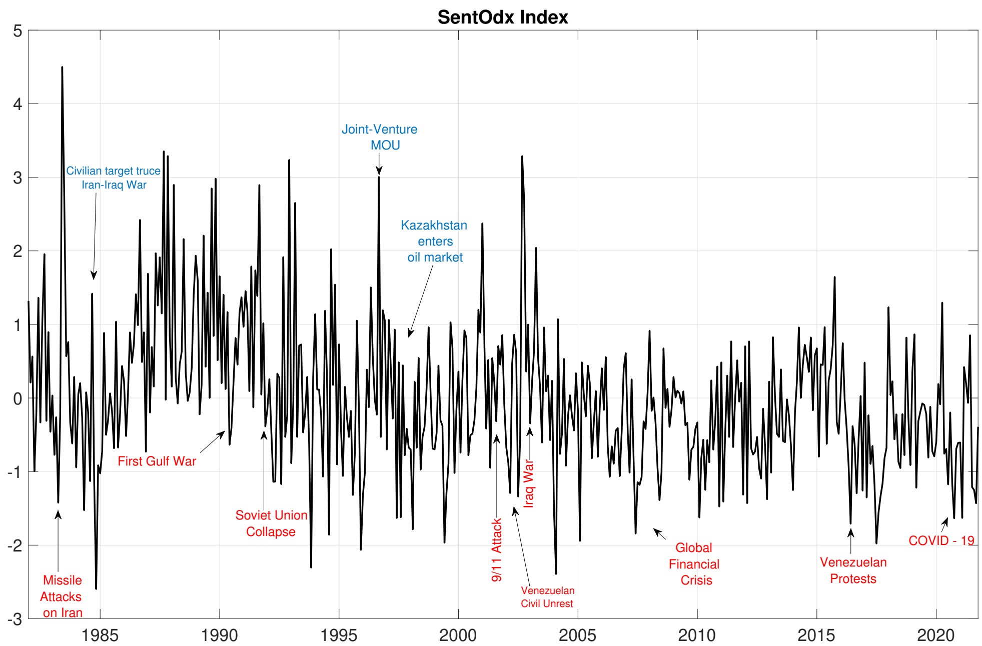


Figure A.9: SentOdx is a human sentiment based index, where words “economy”, “economies”, “economic”, “economics” are considered in a term-document matrix. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

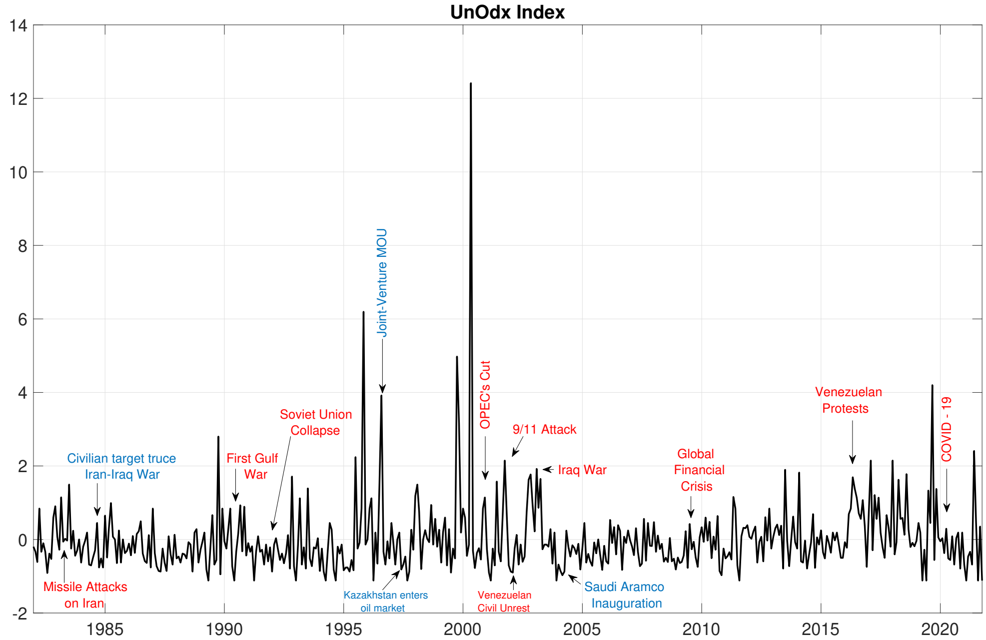


Figure A.10: UnOdx accounts for uncertainty in oil market, where words "uncertain", "uncertainty", "uncertainties" and "uncertainly" are analysed in a document-term matrix. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

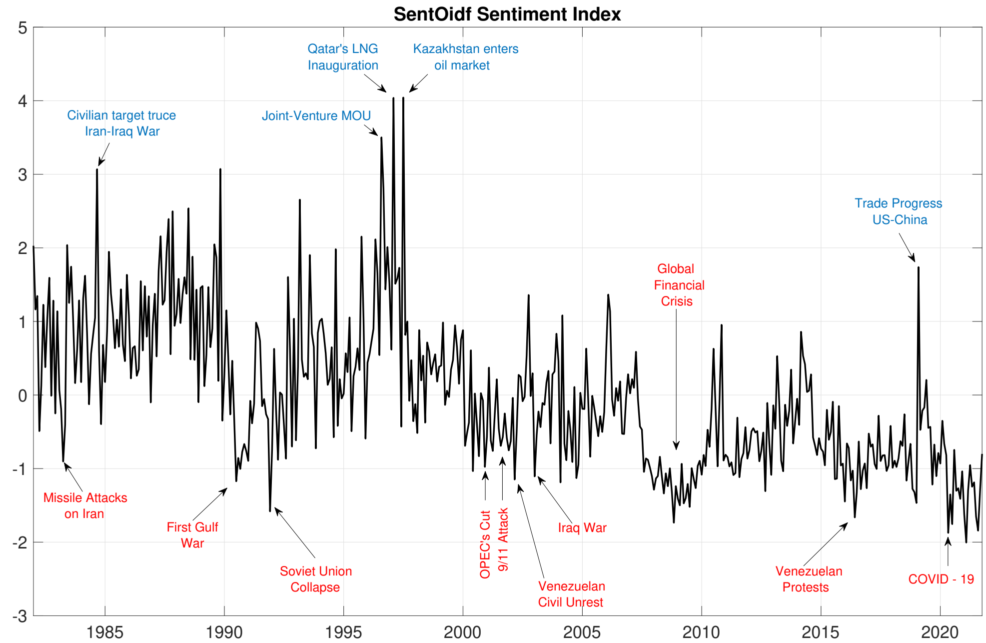


Figure A.11: SentOidf is a human sentiment based index, where words "economy", "economies", "economic", "economics" are considered in a TF-IDF matrix. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

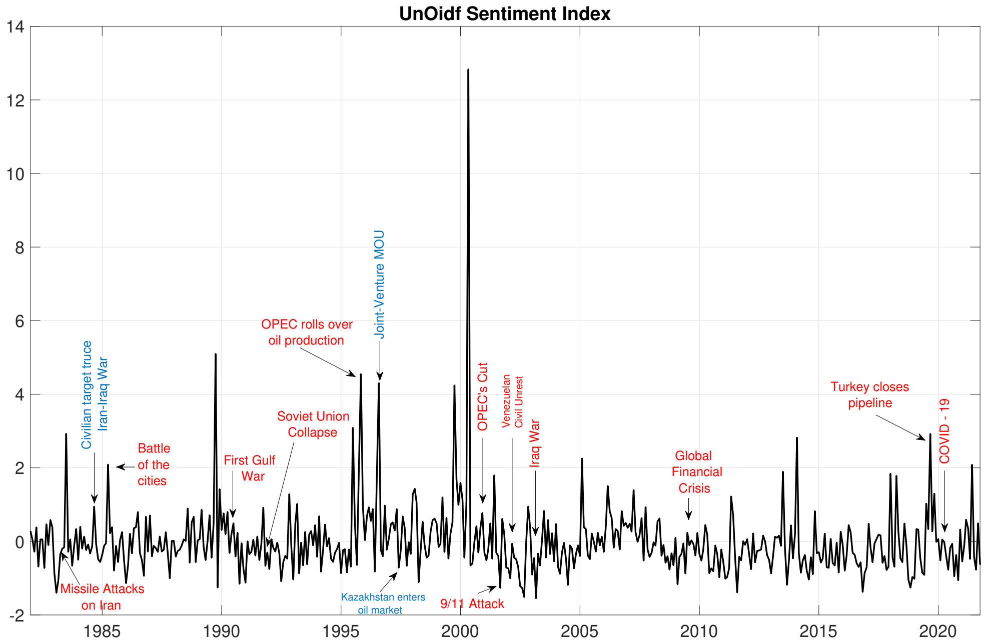


Figure A.12: UnOidf accounts for uncertainty in oil market, where words "uncertain", "uncertainty", "uncertainties" and "uncertainly" are analysed in a TF-IDF matrix. The figure plots the time series monthly score from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

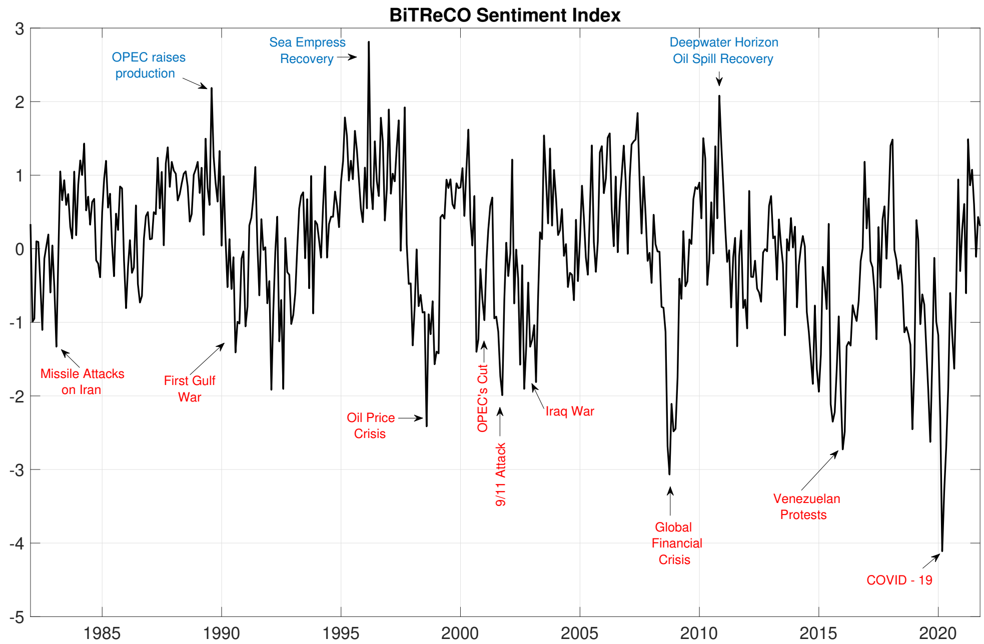


Figure A.13: BiTReCO Sentiment index is based on BERT methodology and covers a period running from 1982M1 through 2021M11. Blue and red notes describe the main historical events that have positively or negatively affected the price of oil.

A.3 Prior Selection in Bayesian Estimation

Let Y_t denote a set of n endogenous variables, and consider the following reduced form VAR(p):

$$Y_t = c + \Phi_1 Y_{t-1} + ... + \Phi_p Y_{t-p} + \varepsilon_t \tag{A.1}$$

where p = 12, c is the vector of intercepts with dimension $n \times 1$, $\Phi_i \, (i = 1, \ldots, 12)$ are $n \times n$ matrices of coefficients and ε_t is a white noise innovation vector. I follow Giannone et al. [2015] and multiply the likelihood of data by a subjective prior belief assessing the distribution of the reduced form VAR coefficients (i.e. a random walk with drift). In particular I assume a Normal-Inverse Wishart distribution of my prior, which takes the following form:

$$\begin{aligned} \Sigma &\sim W^{-1}(\psi, \nu) \\ \Phi|\Sigma &\sim N\left(\hat{\phi}, \Sigma \otimes (X'X)^{-1} \zeta\right), \end{aligned}$$

where $\hat{\Phi} = (X'X)^{-1} X'Y$ is the OLS estimate of Φ, $\hat{\phi} = \text{vec}(\hat{\Phi})$ and X is a $n \times k$ matrix containing the lagged values of Y. The degrees of freedom of the Inverse-Wishart are set such that $\nu = n + 2$, whereas ψ is diagonal and elements ψ_i are function of the residual generated by regressing each variable on its own first 12 lags. ζ is the hyperparameter controlling the overall tightness of the prior distribution.
By assuming that $\Sigma \otimes (X'X)^{-1} \zeta = \Omega$, without loss of generality, Φ takes the following posterior probability form

$$\begin{aligned} \Phi|\Sigma, Y &\sim N\left(\hat{\Phi}(\zeta), \Sigma \otimes \hat{V}(\zeta)\right) \\ \hat{\Phi}(\zeta) &= \text{vec}\left(\hat{\phi}(\zeta)\right) \\ \hat{\phi}(\zeta) &= \hat{V}(\zeta)\left(x'y + \Omega^{-1}\hat{\phi}\right) \\ \hat{V}(\zeta) &= \left(x'x + \Omega^{-1}\right)^{-1} \end{aligned}$$

where variables in lowercase are vectorized. Based on the Bayesian theory, a small value of ζ makes the prior highly informative, whilst in contrast high values of ζ would yield an uninformative prior. It is evident that neither a small, nor a high value of ζ is convenient. However this issue may be seen as a selection problem of the best tightness among all possible hyperparameters. In particular, if $p(Y|\theta)$ is the likelihood of data as function of unknown parameters and $p(\theta)_\gamma$ describes the belief of prior distributions as a function of specific hyperparameters, the product between likelihood and prior can be solved through a hierarchical model. In this way the starting likelihood can be written as $p(Y|\theta)_\gamma$ because such density is functional to the choice of hyperparameters and the simple expression $p(Y|\theta)$ would implicitly suppose that all hyperparameter values have been marginalised. Assuming an uninformative hyperprior, maximizing the marginal likelihood with this procedure is equivalent to maximizing the one step ahead out-of-sample forecasting ability of the model. Therefore, the likelihood form of $p(Y|\theta)_\gamma$ can be obtained from

$$p(\gamma|Y) \propto p(Y|\gamma)\, p(\gamma),$$

where p(γ) is the prior distribution of hyperparameters, while $p(Y|\gamma)_\gamma$ is the marginal likelihood of data as a function of hyperparameters. Under the hypothesis of Natural conjugate priors, the marginal likelihood has the following closed form expression

$$p(Y|\gamma) = \int p(y|\theta,\gamma)\, p(\theta|\gamma)\, d\theta \tag{A.2}$$

Giannone et al. [2015] use a Monte Carlo Markov Chain strategy to maximize equation (A.2).

A.4 Uncertainty Based Empirical Results

Table A.1: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil price in a VAR(12), estimated through a frequentist approach. Text regressors account for uncertainty in oil market.

Monthly horizon	*Model*-1	**UnCO** *Model*-2	**UnCO** *Model*-3	**UnBool** *Model*-2	**UnBool** *Model*-3	**UnOdx** *Model*-2	**UnOdx** *Model*-3	**UnOidf** *Model*-2	**UnOidf** *Model*-3
				A. WTI based VAR					
1	1.404	1.410	1.403	1.403	1.348	1.427	1.468	1.451	1.434
3	1.407	1.423	1.371	1.398	1.321	1.390	1.527	1.489	1.472
6	1.108	1.116	1.148	1.115	1.171	1.107	1.145	1.116	1.147
12	1.329	1.337	1.348	1.335	1.395	1.325	1.338	1.325	1.343
24	1.319	1.322	1.300	1.299	1.342	1.326	1.311	1.327	1.356
				B. RAC based VAR					
1	1.183	1.187	1.190	1.182	1.182	1.201	1.244	1.214	1.184
3	1.391	1.400	1.353	1.370	1.368	1.381	1.495	1.466	1.437
6	1.126	1.133	1.174	1.133	1.197	1.124	1.168	1.132	1.152
12	1.406	1.414	1.440	1.414	1.487	1.405	1.423	1.403	1.421
24	1.302	1.307	1.301	1.288	1.334	1.314	1.293	1.313	1.338
				C. Brent based VAR					
1	1.637	1.642	1.620	1.638	1.588	1.662	1.849	1.703	1.753
3	1.855	1.886	1.642	1.859	1.598	1.789	2.070	2.010	1.937
6	1.202	1.207	1.262	1.207	1.275	1.195	1.248	1.217	1.237
12	1.378	1.383	1.403	1.384	1.438	1.361	1.395	1.371	1.395
24	1.277	1.281	1.276	1.263	1.306	1.267	1.274	1.280	1.331

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*.

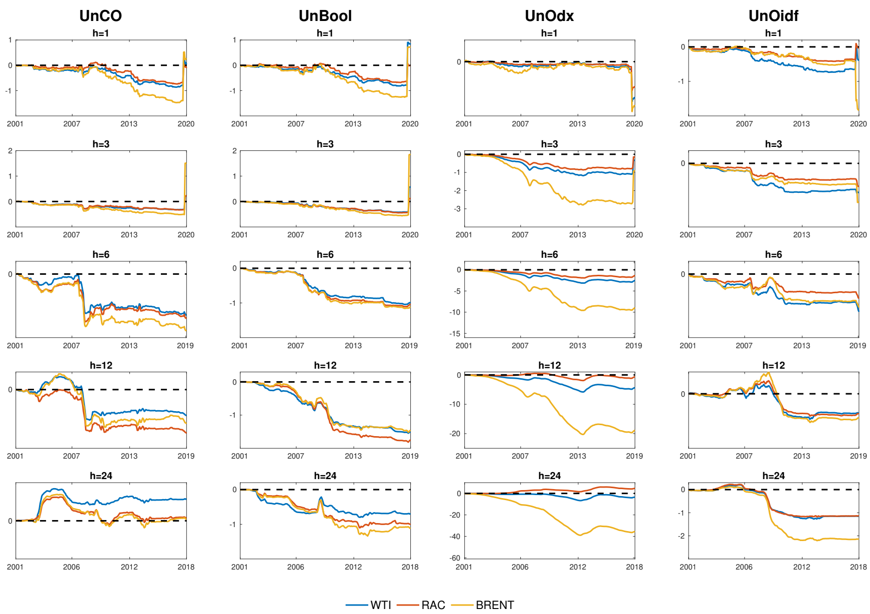


Figure A.14: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and the benchmark model in (2.6). VAR parameters are estimated through a frequentist approach and text variables account for uncertainty in the oil market.

Table A.2: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil price in a BVAR(12). Text regressors account for uncertainty in oil market.

Monthly horizon		UnCO		UnBool		UnOdx		UnOidf	
	Model-1	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3
A. WTI based VAR									
1	0.996	0.993	0.970	0.997	0.973	0.981	0.955	0.975	0.962
3	0.951	0.951	0.961	0.957	0.960	0.973	0.949	1.181	0.944
6	1.001	1.000	1.022	0.998	1.013	1.050	1.010	1.480	1.015
12	1.107	1.113	1.105	1.108	1.101	1.198	1.103	1.978	1.098
24	1.210	1.212	1.193	1.208	1.194	1.388	1.190	2.681	1.195
B. RAC based VAR									
1	0.846	0.851	0.830*	0.852	0.831*	0.833*	0.810**	0.828*	0.813**
3	0.934	0.940	0.947	0.942	0.932	0.935	0.916	1.155	0.914
6	1.001	1.000	1.035	1.008	1.023	1.021	1.011	1.517	1.011
12	1.151	1.144	1.141	1.146	1.137	1.175	1.124	2.088	1.127
24	1.227	1.226	1.205	1.218	1.216	1.273	1.206	2.585	1.205
C. Brent based VAR									
1	1.048	1.020	1.000	1.021	1.014	1.030	0.989	1.016	0.964
3	1.039	1.015	1.013	1.021	1.008	1.020	1.000	1.197	0.976
6	1.040	1.038	1.072	1.040	1.058	1.044	1.047	1.463	1.047
12	1.141	1.140	1.141	1.137	1.132	1.132	1.127	1.846	1.129
24	1.202	1.204	1.186	1.196	1.198	1.190	1.182	1.846	1.192

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*. *, **, *** respectively denote 10%, 5% and 1% level of significance of Diebold-Mariano test.

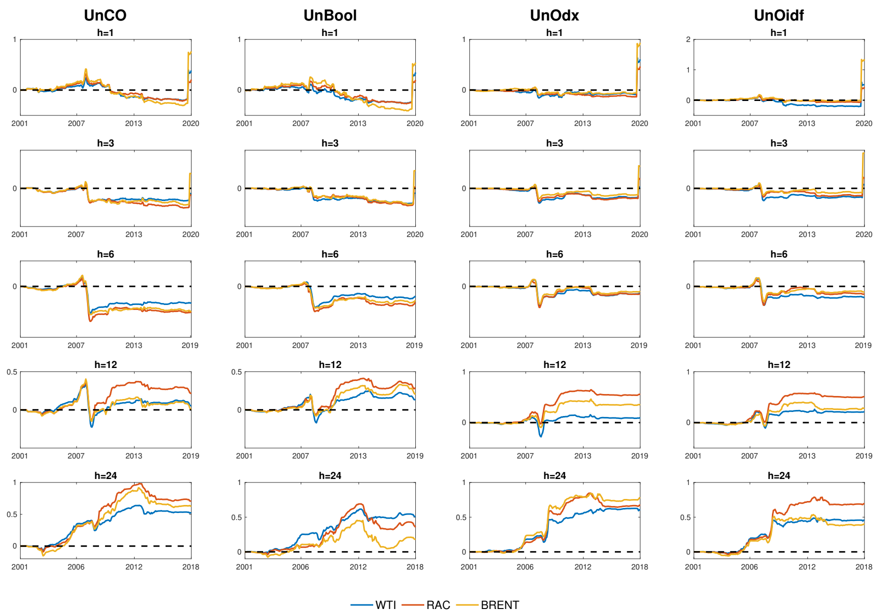


Figure A.15: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and the benchmark model in (2.6). VAR parameters are estimated in a Bayesian fashion and text variables account for uncertainty in the oil market.

Table A.3: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a SV-BVAR(12). Text regressors account for uncertainty in oil market.

Monthly horizon		**UnCO**		**UnBool**		**UnOdx**		**UnOidf**	
	Model-1	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3
A. WTI based VAR									
1	0.976	0.982	0.981	0.980	0.981	0.984	0.973	0.990	0.957
3	0.962	0.968	0.967	0.970	0.956	0.953	0.955	0.970	0.942
6	0.918	0.939	0.938	0.938	0.928	0.918	0.913	0.914	0.911
12	0.946	0.958	0.966	0.957	0.953	0.946	0.945	0.944	0.947
24	0.925	0.942	0.931	0.951	0.928	0.933	0.939	0.936	0.954
B. RAC based VAR									
1	0.818**	0.833**	0.833**	0.819**	0.823**	0.811**	0.816**	0.813**	0.808**
3	0.909	0.933	0.934	0.927	0.914	0.905	0.901	0.909	0.899
6	0.871	0.898	0.924	0.894	0.895	0.872	0.867	0.881	0.885
12	0.908	0.929	0.944	0.932	0.926	0.909	0.912	0.910	0.933
24	0.837	0.842	0.845	0.873	0.835	0.833	0.850	0.843	0.892
C. Brent based VAR									
1	0.980	0.979	0.993	0.987	0.997	0.970	0.975	0.973	0.974
3	1.032	1.045	1.052	1.053	1.031	1.022	1.024	1.025	1.019
6	0.910	0.933	0.946	0.928	0.933	0.910	0.917	0.918	0.927
12	0.937	0.943	0.962	0.951	0.942	0.930	0.939	0.944	0.956
24	0.908	0.916	0.905	0.937	0.911	0.911	0.922	0.917	0.948

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*. *, **, *** respectively denote 10%, 5% and 1% level of significance of Diebold-Mariano test.

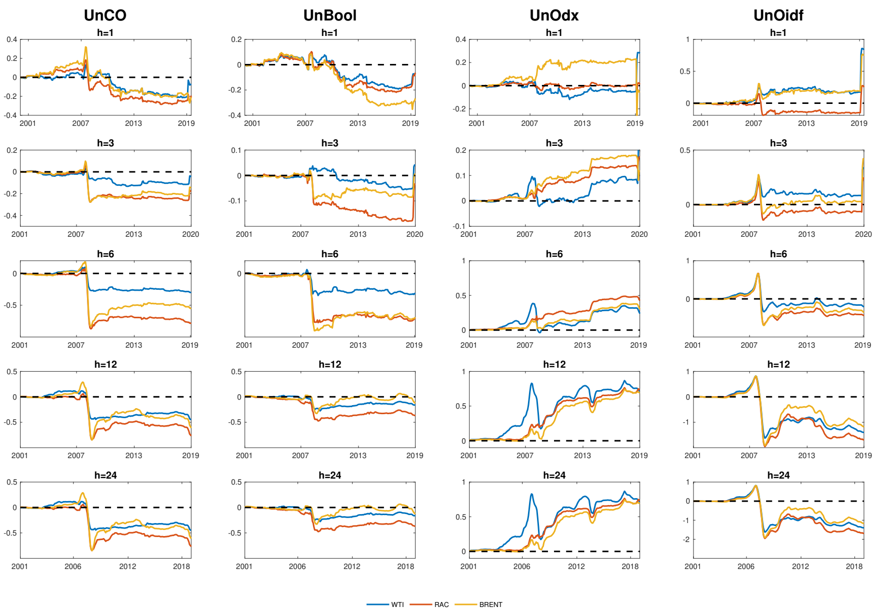


Figure A.16: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and the benchmark model in (2.6). VAR parameters are estimated in a Bayesian fashion by assuming stochastic volatility of the error term, and text variables account for uncertainty in the oil market.

A.5 Additional Empirical Results

Table A.4: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a SV-BVAR(12). GECON is used as a measure of global real economy, and text regressors account for human sentiment about oil news.

Monthly horizon		**SentCO**		**VadOil**		**SentOdx**		**SentOidf**		**BiTReCO**	
	Model-1	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3
A. WTI based VAR											
1	1.002	0.994	1.000	0.994	0.981	0.992	0.984	0.997	0.968	0.952	0.973
3	1.032	1.028	1.033	1.088	1.004	1.093	0.998	1.120	1.009	1.048	0.981
6	0.968	0.969	0.990	1.118	0.974	1.194	0.948	1.176	0.978	1.072	0.934
12	0.945	0.938	0.970	1.310	0.962	1.649	0.901	1.372	0.980	1.261	0.930
24	0.948	0.944	0.968	1.742	0.925	3.084	0.945	1.922	0.994	1.837	0.941
B. RAC based VAR											
1	0.830**	0.835**	0.833**	0.830**	0.819**	0.835**	0.828**	0.841**	0.824**	0.792***	0.803**
3	0.968	0.966	0.985	1.047	0.951	1.005	0.942	1.025	0.956	0.994	0.919
6	0.926	0.930	0.991	1.109	0.957	1.014	0.894	1.045	0.974	1.032	0.907
12	0.912	0.906	0.987	1.377	0.958	1.198	0.860*	1.170	0.978	1.231	0.919
24	0.863	0.867	0.922	1.775	0.863	1.640	0.888	1.366	0.946	1.676	0.891
C. Brent based VAR											
1	0.998	1.006	0.997	1.009	0.994	1.015	1.010	1.021	0.995	0.943	0.947
3	1.091	1.092	1.095	1.168	1.077	1.157	1.083	1.189	1.073	1.122	0.990
6	0.970	0.968	1.034	1.134	1.004	1.111	0.942	1.165	1.018	1.095	0.918
12	0.935	0.933	1.000	1.317	0.979	1.344	0.899	1.329	0.981	1.293	0.913
24	0.943	0.938	0.976	1.599	0.938	1.945	0.955	1.712	0.993	1.833	0.946

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*.

Figure A.17: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.10) and the cumulative sum of forecasting errors generated by a random walk. GECON based VAR parameters are estimated in a Bayesian fashion by assuming the stochastic volatility of the error term, and text variables account for human sentiment about oil news.

Table A.5: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a SV-BVAR(12). Text regressors account for uncertainty in oil market, and GECON is now used in place of world industrial production.

Monthly horizon	*Model*-1	**UnCO** *Model*-2	**UnCO** *Model*-3	**UnBool** *Model*-2	**UnBool** *Model*-3	**UnOdx** *Model*-2	**UnOdx** *Model*-3	**UnOidf** *Model*-2	**UnOidf** *Model*-3
A. WTI based VAR									
1	1.002	1.005	0.995	1.003	1.005	0.999	0.998	0.996	0.989
3	1.032	1.035	1.026	1.039	1.041	1.023	1.038	1.029	1.019
6	0.968	0.983	0.992	0.988	0.997	0.971	0.971	0.970	0.971
12	0.945	0.958	0.971	0.967	0.957	0.943	0.952	0.936	0.950
24	0.948	0.963	0.945	0.962	0.951	0.950	0.983	0.962	0.987
B. RAC based VAR									
1	0.830**	0.840**	0.837**	0.845**	0.848**	0.836**	0.838**	0.829**	0.830**
3	0.968	0.976	0.972	0.981	0.980	0.965	0.970	0.963	0.959
6	0.926	0.940	0.970	0.938	0.958	0.930	0.934	0.929	0.943
12	0.912	0.925	0.951	0.929	0.932	0.909	0.922	0.908	0.934
24	0.863	0.870	0.885	0.898	0.900	0.863	0.896	0.880	0.929
C. Brent based VAR									
1	0.998	1.014	1.004	1.009	1.017	0.996	1.003	1.001	0.999
3	1.091	1.099	1.074	1.097	1.099	1.080	1.085	1.083	1.072
6	0.970	0.975	0.991	0.977	0.986	0.973	0.982	0.979	0.984
12	0.935	0.945	0.957	0.958	0.942	0.938	0.948	0.939	0.953
24	0.943	0.937	0.924	0.960	0.958	0.937	0.961	0.946	0.975

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon h.

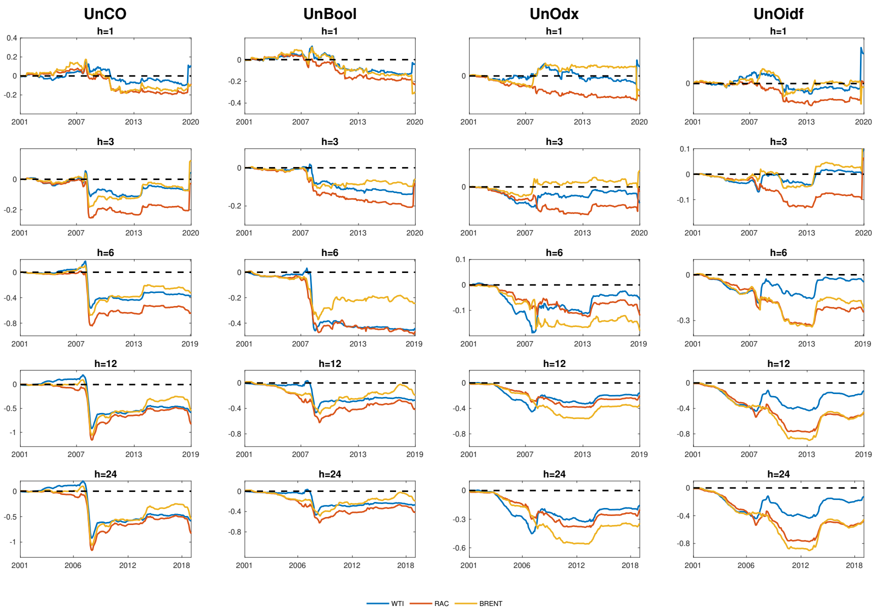


Figure A.18: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.10) and the cumulative sum of forecasting errors generated by a random walk. VAR parameters are estimated in a Bayesian fashion by assuming the stochastic volatility of the error term, and text variables account for uncertainty in oil market.

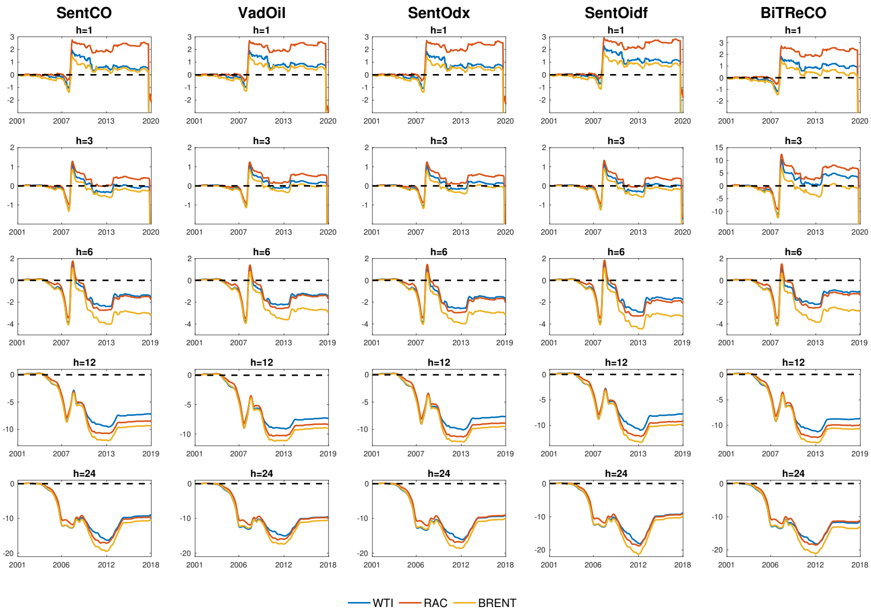


Figure A.19: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated through a frequentist approach and text variables account for human sentiment about oil news.

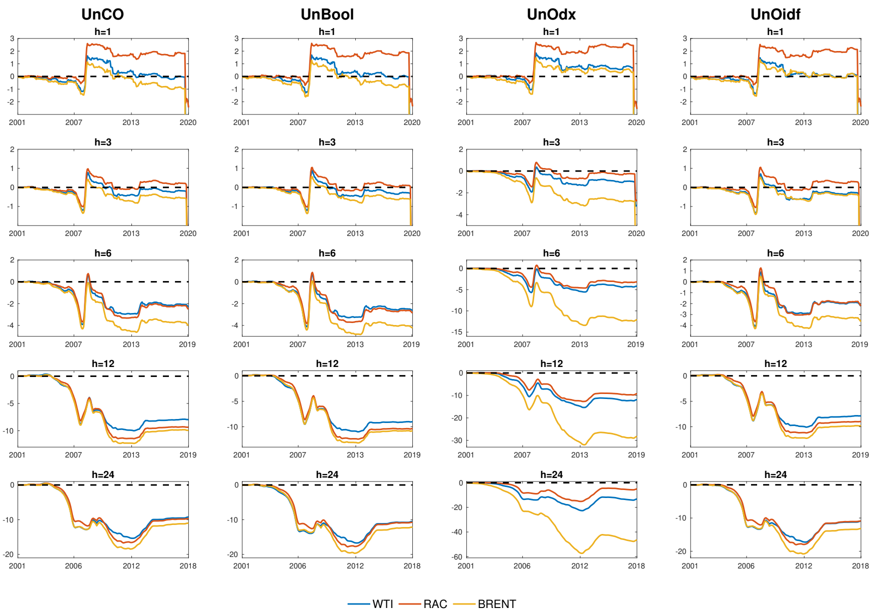


Figure A.20: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated through a frequentist approach and text variables account for oil market uncertainty.

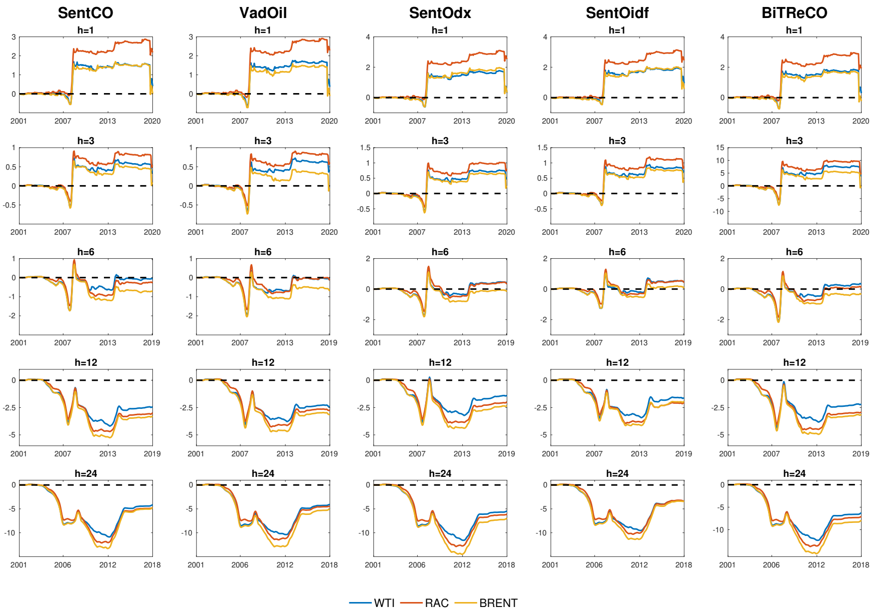


Figure A.21: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated in a Bayesian fashion and text variables account for human sentiment about oil news.

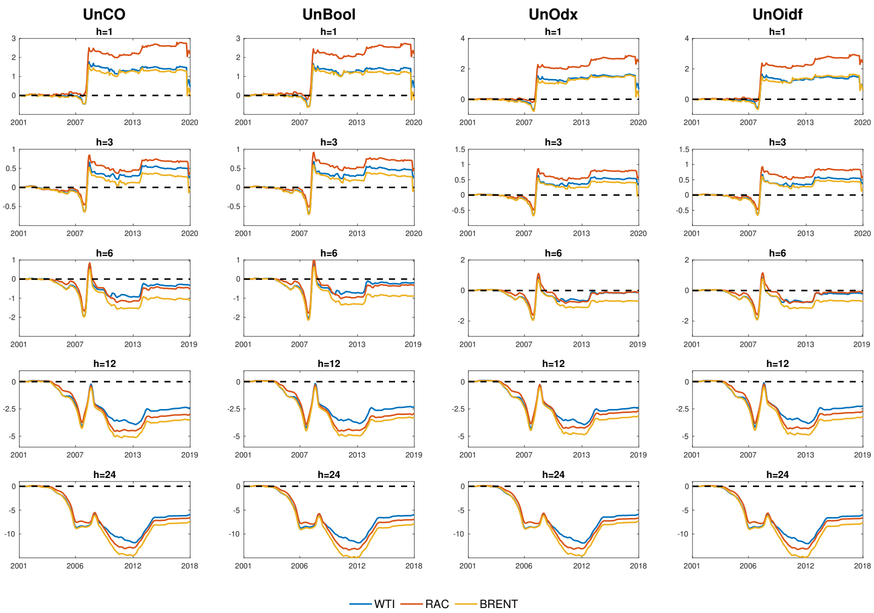


Figure A.22: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated in a Bayesian fashion and text variables account for oil market uncertainty.

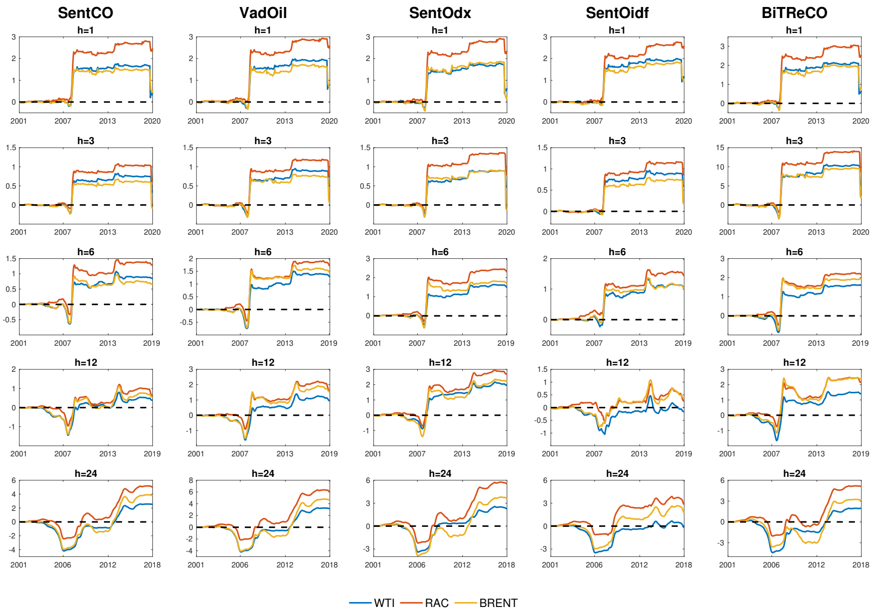


Figure A.23: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated in a Bayesian fashion by assuming stochastic volatility in the error term, and text variables account for human sentiment about oil news.

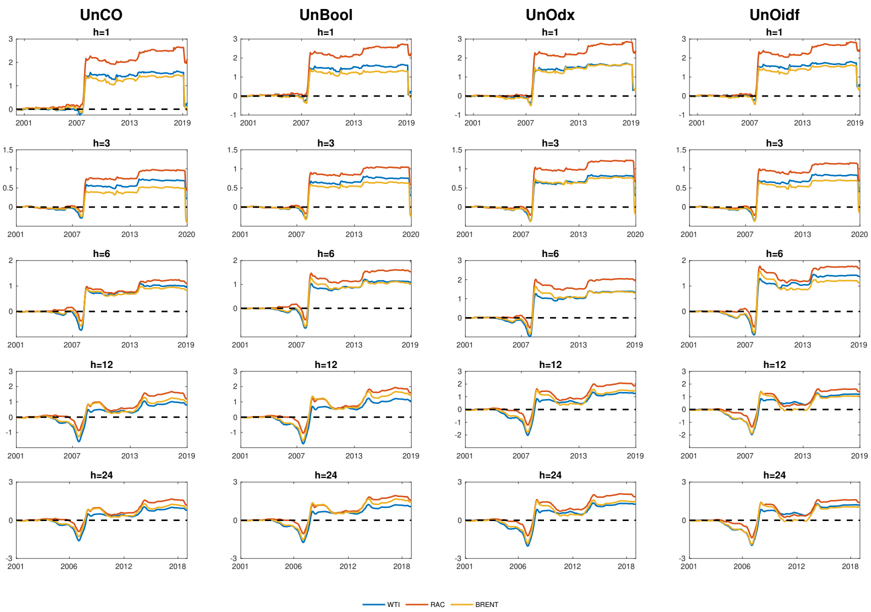


Figure A.24: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated in a Bayesian fashion by assuming stochastic volatility in the error term, and text variables account for oil market uncertainty.

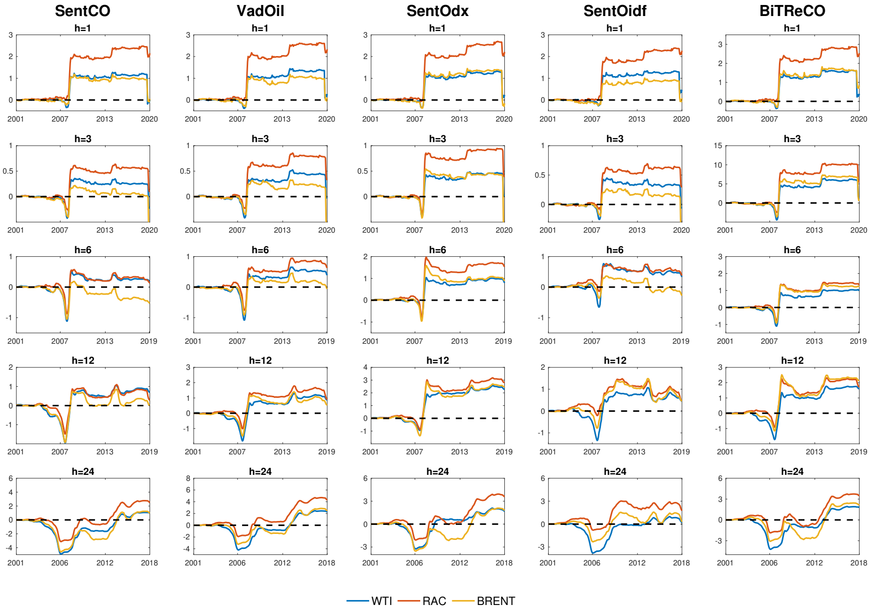


Figure A.25: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated in a Bayesian fashion by assuming stochastic volatility in the error term, and text variables account for human sentiment about oil news. GECON is used in place of WIP

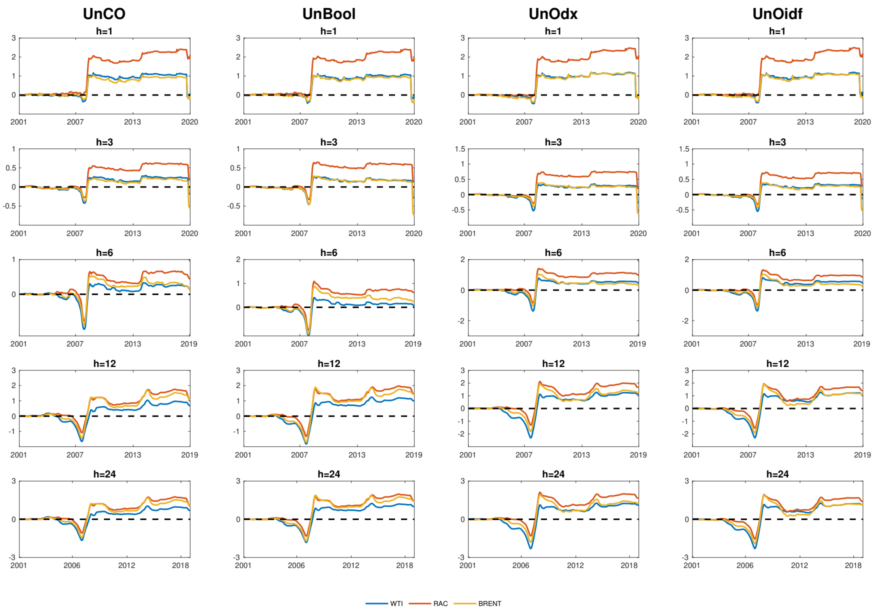


Figure A.26: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and of a random walk. VAR parameters are estimated in a Bayesian fashion by assuming stochastic volatility in the error term, and text variables account for oil market uncertainty. GECON is used in place of WIP

Table A.6: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices in a SV-BVAR(12). World industrial production is used as a measure of global real economy, text regressors are dictionary based and account for human sentiment.

Monthly horizon		**FinStab**		**LouMc**		**Afinn**		**HarvOil**		**VadOil**	
	Model-1	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3	*Model*-2	*Model*-3
A. WTI based VAR											
1	0.976	0.976	0.975	0.968	0.965	0.986	0.956	0.975	0.950	0.982	0.946
3	0.962	0.961	0.950	0.954	0.933	0.959	0.950	0.949	0.939	0.987	0.929
6	0.918	0.917	0.919	0.917	0.992	0.921	0.939	0.919	0.938	0.944	0.916
12	0.946	0.962	0.982	0.966	1.083	0.969	1.090	0.967	0.969	0.985	0.959
24	0.925	0.927	0.920	0.927	0.983	0.928	0.932	0.925	0.945	1.034	0.892
B. RAC based VAR											
1	0.818**	0.820**	0.815**	0.818**	0.816**	0.819**	0.817**	0.818**	0.818**	0.821**	0.799**
3	0.909	0.914	0.909	0.914	0.907	0.909	0.905	0.919	0.915	0.936	0.890
6	0.971	0.964	0.931	0.970	1.010	0.967	0.946	0.900	0.928	0.925	0.882
12	0.938	0.939	1.236	0.936	1.236	0.909	1.338	0.931	1.044	0.964	0.926
24	0.847	0.838	0.846	0.826	0.929	0.833	0.861	0.832	0.872	0.973	0.811
C. Brent based VAR											
1	0.980	0.977	0.966	0.979	0.978	0.972	0.966	0.976	0.970	0.982	0.952
3	1.032	1.025	1.019	1.030	1.016	1.031	1.026	1.035	1.027	1.041	0.991
6	0.910	0.910	0.934	0.913	1.015	0.908	0.944	0.914	0.955	0.944	0.906
12	0.939	1.004	1.005	0.940	1.090	0.939	1.041	0.939	0.998	0.992	0.938
24	0.908	0.912	0.934	0.906	0.972	0.909	0.933	0.913	0.949	1.019	0.884

Note: Red values report MSPE ratio results lower than recursive MSPE ratios based on equation (2.6). Blue values are the lowest MSPE results relative to a specific horizon *h*. *, **, *** respectively denote 10%, 5% and 1% level of significance of Diebold-Mariano test.

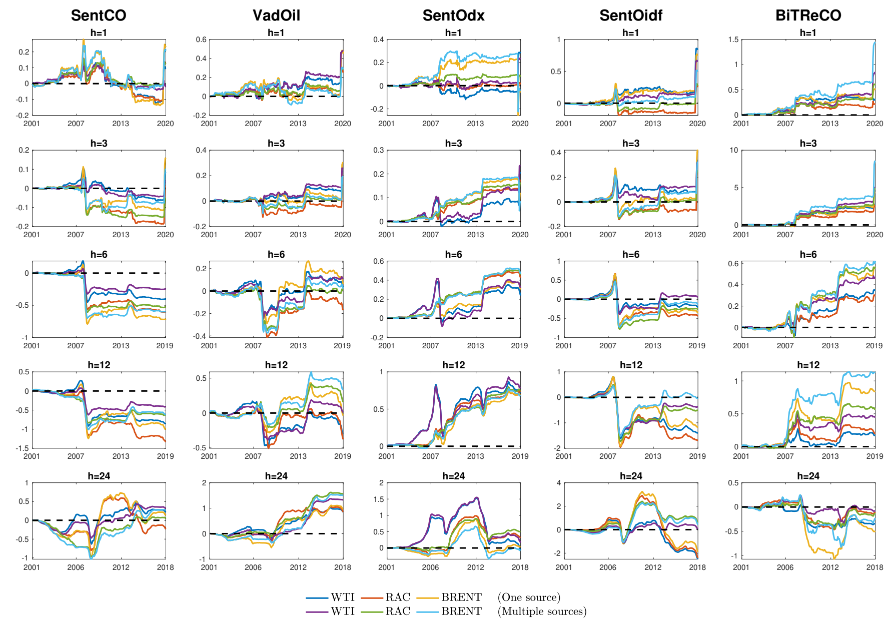


Figure A.27: The figure plots the difference between the cumulative sum of forecasting errors of equation (2.8) and the benchmark model in (2.6). A comparison between one and multiple text data sources is presented.

Appendix
Chapter 2

B.1 Data Sources

In this paper I use 140,096 oil related daily articles which featured in the Banking, Finance and Energy section of the following newspapers:

Source	No. of Articles	Period
The Financial Times	103,966	1982M01-2021M12
The Independent	29,377	1988M09-2021M12
Thomson Reuters	6,753	2002M11-2021M12

Articles are retrieved from the LexisNexis database, and are selected based on the joint occurrence of the words *oil* and *price*. Table B.1 reports the definition, the data transformation and the source of the macroeconomic, financial and commodity variables used in this empirical analysis. Intraday returns are computed by subtracting the closing price at which a stock/commodity variable has traded during a regular trading session, from its opening. This measure captures the return generated by a variable during regular business hours.

Table B.1: Variable definition

Variable	Transformation	Period	Frequency	Source
World industrial production	100*log-difference	1982M01-2021M12	M	Baumeister and Hamilton [2019]
WTI index	100*log-levels	1982M01-2021M12	M/W/D	FRED & Bloomberg
Brent crude oil index	100*log-levels	1982M01-2021M12	M/W/D	FRED & Bloomberg
Refiner acquisition cost of crude oil	100*log-levels	1982M01-2021M12	M	EIA
World oil production	100*log-difference	1982M01-2021M12	M	EIA
Oil inventories	log-difference	1982M01-2021M12	M	EIA
WTI future prices	growth rate	1995M09-2021M12	M/W/D	Bloomberg
Brent future prices	growth rate	1995M09-2021M12	M/W/D	Bloomberg
Gasoline	growth rate	1986M06-2021M12	M/W/D	Bloomberg
CRB index	growth rate	1982M01-2021M12	M/W/D	Bloomberg
Baltic Dry index	growth rate	1985M01-2021M12	M/W/D	Bloomberg
Federal Funds rate	growth rate	1982M01-2021M12	M/W/D	Bloomberg
FTSE100	growth rate	1984M01-2021M12	M/W/D	Bloomberg
S&P500	growth rate	1982M01-2021M12	M/W/D	Bloomberg
Euro Stoxx 50	growth rate	1987M01-2021M12	M/W/D	Bloomberg
Hang Seng index	growth rate	1982M01-2021M12	M/W/D	Bloomberg
GBP/USD	growth rate	1982M01-2021M12	M/W/D	Bloomberg
CAD/USD	growth rate	1982M01-2021M12	M/W/D	Bloomberg
EUR/USD	growth rate	1982M01-2021M12	M/W/D	Bloomberg
Gold	growth rate	1982M01-2021M12	M/W/D	Bloomberg
Copper	growth rate	1988M12-2021M12	M/W/D	Bloomberg
Natural Gas	growth rate	1990M04-2021M12	M/W/D	Bloomberg
Palladium	growth rate	1987M01-2021M12	M/W/D	Bloomberg
Silver	growth rate	1982M01-2021M12	M/W/D	Bloomberg

Note: In column 4, M = monthly, W = weekly and D = daily.

B.2 Bayesian Shrinkage with Hierarchical Minnesota Priors

Suppose oil prices evolve according to the following reduced form VAR(12):

$$Y_t = c + \Phi_1 Y_{t-1} + \ldots + \Phi_{12} Y_{t-12} + \varepsilon_t$$

where c is the $n \times 1$ vector of intercepts, Φ_i $(i = 1, \ldots, 12)$ are $n \times n$ matrices of coefficients and ε is the covariance matrix of innovations with mean zero and variance Σ. In the spirit of Giannone et al. [2015] I assume that parameters are distributed as follows:

$$\Sigma|\xi \sim IW(\psi, d)$$
$$\Phi|\Sigma \sim N(\phi, \Sigma \otimes \Omega\xi)$$

where Ω is a predefined Minnesota shrinkage rule, and ξ is an unknown parameter used to make inference on the informativeness of the prior. In comparison to the standard versions of Minnesota prior where the hyperparameters are fixed at some subjective values (see Doan et al. [1984] and Litterman [1986]), in this data-driven variant the overall shrinkage hyperparameters are treated as coefficients to be estimated from the data. Specifically, if $p(Y|\theta)$ is the likelihood of data as function of unknown parameters and $p(\theta)_\gamma$

describes the belief of prior distributions as a function of specific hyperparameters, the product between likelihood and prior is solved through a hierarchical model. In this way the starting likelihood is written as $p(Y|\theta)_\gamma$ because such density is functional to the choice of hyperparameters and the simple expression $p(Y|\theta)$ would implicitly suppose that all hyperparameter values have been marginalised. Assuming an uninformative hyperprior, maximizing the marginal likelihood with this procedure is equivalent to maximizing the one step ahead out-of-sample forecasting ability of the model. Therefore, the likelihood form of $p(Y|\theta)_\gamma$ can be obtained from

$$p(\gamma|Y) \propto p(Y|\gamma)\, p(\gamma),$$

where p(γ) is the prior distribution of hyperparameters, while $p(Y|\gamma)_\gamma$ is the marginal likelihood of data as a function of hyperparameters. Under the hypothesis of Natural conjugate priors, the marginal likelihood has the following closed form expression

$$p(Y|\gamma) = \int p(y|\theta,\gamma)\, p(\theta|\gamma)\, d\theta.$$

In my empirical application, for data in growth rates I set $\hat{\phi} = 0$ to shrink all coefficients towards zero. In contrast, for data in levels, I set $\hat{\phi} = 0.99$ by expressing a belief that coefficients evolve as a random walk. Furthermore, I set $\nu = n+2$, ψ is diagonal with elements ψ_i that are function of the residual generated by regressing each variable on its own first 12 lags. By assuming that $\Sigma \otimes (X'X)^{-1}\zeta = \Omega$, without loss of generality, Φ takes the following posterior probability form

$$\begin{aligned}
\Phi|\Sigma, Y &\sim N\left(\hat{\Phi}(\zeta), \Sigma \otimes \hat{V}(\zeta)\right) \\
\hat{\Phi}(\zeta) &= \mathrm{vec}\left(\hat{\phi}(\zeta)\right) \\
\hat{\phi}(\zeta) &= \hat{V}(\zeta)\left(x'y + \Omega^{-1}\hat{\phi}\right) \\
\hat{V}(\zeta) &= \left(x'x + \Omega^{-1}\right)^{-1}
\end{aligned}$$

where $\hat{\Phi} = (X'X)^{-1}X'Y$ is the OLS estimate of Φ, $\hat{\phi} = \mathrm{vec}(\hat{\Phi})$ and X is a $n \times k$ matrix containing the lagged values of Y. This methodology is shown to yield accurate out-of-sample forecast performance for a variety of datasets (see Carriero et al. [2015], Lenza and Primiceri [2020] and Miranda-Agrippino and Ricco [2021]).

B.3 Additional empirical results

Table B.2: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with high-frequency financial variables; text vs. no-text. Alternative MIDAS parametrisations.

	Model	Weighting Scheme	1-month			3-months			6-months			12-months			24-months		
			WTI	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*
			A. Unigram based variable														
	MIDAS-TXT	Almond	**0.930**	**0.778*****	**0.949**	**0.979**	1.021	**0.985**	**0.944**	**0.945**	**0.976**	**0.953**	**0.985**	**0.992**	1.218	1.226	1.245
	MIDAS-TXT	Eq-Weights	**0.919**	**0.763*****	**0.932***	**0.973**	1.006	**0.974**	**0.948**	**0.943**	**0.965**	**0.981**	1.010	1.027	1.036	1.035	1.036
	MIDAS-TXT	U-MIDAS	**0.915***	**0.763*****	**0.929***	**0.978**	**0.964**	**0.954**	**0.954**	**0.949**	**0.977**	**0.985**	1.011	1.041	**0.997**	**0.965**	**0.983**
			B. Dictionary based variable														
	MIDAS-TXT	Almond	**0.909***	**0.789****	**0.957**	**0.970**	**0.973**	**0.962**	**0.955**	**0.952**	**0.977**	1.030	1.057	1.089	1.370	1.375	1.418
	MIDAS-TXT	Eq-Weights	**0.881****	**0.754*****	**0.913***	**0.932**	**0.937**	**0.930**	**0.930**	**0.894**	**0.914**	**0.987**	1.011	1.034	1.123	1.123	1.122
	MIDAS-TXT	U-MIDAS	**0.879****	**0.761*****	**0.919***	**0.915**	**0.941**	**0.911**	**0.969**	**0.949**	**0.961**	**0.997**	1.014	1.041	1.088	1.080	1.087
			C. Geometrical based variables														
(tfm)	MIDAS-TXT	Almond	**0.923**	**0.767*****	**0.986**	**0.965**	1.003	**0.982**	**0.948**	**0.946**	**0.974**	**0.977**	1.014	1.034	1.103	1.122	1.132
(idf)	MIDAS-TXT	Almond	**0.918**	**0.766*****	**0.931***	**0.965**	**0.997**	**0.971**	**0.961**	**0.950**	**0.982**	**0.978**	1.019	1.032	1.243	1.232	1.256
(tfm)	MIDAS-TXT	Eq-Weights	**0.910***	**0.760*****	**0.928***	**0.957**	**0.992**	**0.960**	**0.973**	**0.973**	**0.973**	**0.928**	**0.959**	**0.937**	1.023	1.011	1.029
(idf)	MIDAS-TXT	Eq-Weights	**0.924**	**0.759*****	**0.928***	**0.967**	1.003	**0.972**	**0.968**	**0.961**	**0.981**	**0.944**	**0.920**	**0.956**	1.077	1.083	1.093
(tfm)	MIDAS-TXT	U-MIDAS	**0.914***	**0.755*****	**0.929***	**0.957**	**0.989**	**0.959**	**0.951**	**0.951**	**0.978**	**0.920**	**0.933**	**0.946**	1.003	1.002	1.036
(idf)	MIDAS-TXT	U-MIDAS	**0.925**	**0.761*****	**0.930***	**0.962**	**0.996**	**0.962**	**0.952**	**0.951**	**0.973**	**0.931**	**0.932**	**0.969**	1.093	1.093	1.115
			D. Network based variable														
	MIDAS-TXT	Almond	**0.897**	**0.816****	1.035	**0.970**	1.061	1.001	**0.983**	1.032	1.125	1.992	1.606	2.001	1.336	1.463	1.423
	MIDAS-TXT	Eq-Weights	**0.879****	**0.737*****	**0.897***	**0.925**	**0.921**	**0.927**	**0.923**	**0.924**	**0.951**	**0.993**	1.032	1.036	1.195	1.203	1.213
	MIDAS-TXT	U-MIDAS	**0.891****	**0.740*****	**0.909***	**0.922**	**0.929**	**0.936**	**0.954**	1.152	1.280	1.923	3.471	2.559	1.179	1.233	1.227

Note: In column 1 MIDAS: mixed data sampling, TXT indicates models including a text variable, tfm: term-frequency matrix, idf: term-frequency inverse-document-frequency matrix. Column 2 displays the weighting scheme of MIDAS parameters. Bold values indicate improvements on the no-change forecast. Blue entries stand for the lowest MSPE for a given time horizon, relative to a specific oil price measure. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table B.3: Recursive MSPE ratios relative to a random walk forecast of alternative monthly indicators of real oil prices. Homogeneous vs. mixed frequency models with uncertainty text indicators.

	Model	Freq	Weighting Scheme	1-month			3-months			6-months			12-months			24-months		
				WTI	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*
				A. Unigram based variable														
	AR-X-TXT	HF	-	**0.871****	**0.771*****	**0.891****	**0.997**	**0.973**	**0.958**	**0.956**	**0.961**	**0.988**	**0.981**	1.032	1.039	1.174	1.176	1.175
	MIDAS-TXT	MF	Beta	**0.926***	**0.763*****	**0.917***	**0.962**	**0.989**	**0.952**	**0.951**	**0.951**	**0.984**	**0.977**	1.012	1.035	1.155	1.159	1.157
	MIDAS-TXT	MF	Almond	**0.942**	**0.787****	**0.955**	**0.975**	1.015	**0.991**	**0.947**	**0.961**	**0.992**	**0.966**	1.414	1.030	1.356	1.270	1.085
	MIDAS-TXT	MF	Eq-Weights	**0.928**	**0.771*****	**0.929**	**0.963**	**0.998**	**0.972**	**0.954**	**0.957**	**0.984**	**0.967**	1.009	1.032	1.131	1.142	1.138
	MIDAS-TXT	MF	U-MIDAS	**0.936**	**0.802****	**0.970**	**0.964**	**0.991**	**0.965**	**0.948**	**0.951**	**0.980**	1.215	1.025	1.075	1.116	1.125	1.118
	BVAR-TXT	HF	-	**0.906**	**0.771*****	**0.920**	**0.950**	**0.911**	**0.974**	1.021	1.025	1.058	1.089	1.119	1.126	1.181	1.198	1.177
	MF-BVAR-TXT	MF	-	**0.951**	**0.807**	**0.943*****	**0.973**	**0.934**	**0.981**	1.025	1.032	1.062	1.063	1.096	1.102	1.162	1.180	1.161
	SV-BVAR-TXT	HF	-	**0.892****	**0.792*****	**0.912***	**0.936**	**0.902**	**0.949**	**0.956**	1.021	1.011	1.089	1.631	1.264	**0.940**	1.154	1.136
	MF-SV-BVAR-TXT	MF	-	**0.905***	**0.805*****	**0.922***	**0.931**	**0.914**	**0.950**	**0.950**	**0.961**	**0.968**	1.047	1.238	1.075	**0.926**	**0.872**	**0.926**
				B. Geometrical based variables														
(tfm)	AR-X-TXT	HF	-	**0.866****	**0.762*****	**0.877****	**0.999**	**0.964**	**0.950**	**0.948**	**0.948**	**0.974**	**0.979**	1.025	1.037	1.158	1.159	1.159
(idf)	AR-X-TXT	HF	-	**0.866****	**0.765*****	**0.879****	**0.997**	**0.970**	**0.953**	**0.951**	**0.953**	**0.980**	**0.979**	1.027	1.039	1.159	1.159	1.158
(tfm)	MIDAS-TXT	MF	Beta	**0.912***	**0.749*****	**0.906****	**0.952**	**0.983**	**0.955**	**0.949**	**0.950**	**0.974**	**0.978**	1.012	1.034	1.125	1.129	1.118
(idf)	MIDAS-TXT	MF	Beta	**0.919***	**0.761*****	**0.935***	**0.968**	1.003	**0.975**	**0.948**	**0.946**	**0.973**	**0.980**	1.015	1.038	1.157	1.167	1.177
(tfm)	MIDAS-TXT	MF	Almond	**0.964**	**0.823**	**0.961**	**0.974**	1.037	1.020	**0.946**	**0.957**	**0.970**	**0.965**	1.022	1.025	1.364	1.518	1.350
(idf)	MIDAS-TXT	MF	Almond	**0.918***	**0.766*****	**0.932***	**0.965**	**0.997**	**0.971**	**0.958**	**0.959**	**0.983**	**0.957**	1.011	1.029	1.221	1.223	1.231
(tfm)	MIDAS-TXT	MF	Eq-Weights	**0.912***	**0.755*****	**0.940**	**0.955**	**0.988**	**0.970**	**0.962**	**0.971**	**0.991**	**0.970**	**0.980**	**0.981**	**0.889**	**0.920**	**0.894**
(idf)	MIDAS-TXT	MF	Eq-Weights	**0.921***	**0.758*****	**0.927***	**0.964**	**0.999**	**0.970**	**0.963**	**0.953**	**0.978**	**0.894**	**0.896**	**0.932**	1.079	1.079	1.090
(tfm)	MIDAS-TXT	MF	U-MIDAS	**0.911***	**0.762*****	**0.940**	**0.949**	**0.987**	**0.965**	**0.963**	**0.948**	**0.974**	**0.914**	**0.887**	**0.913**	**0.890**	**0.875**	**0.895**
(idf)	MIDAS-TXT	MF	U-MIDAS	**0.911***	**0.760*****	**0.929***	**0.961**	**0.986**	**0.963**	**0.953**	**0.951**	**0.973**	**0.922**	**0.863**	**0.892**	1.097	1.083	1.114
(tfm)	BVAR-TXT	HF	-	**0.894***	**0.759*****	**0.903***	**0.937**	**0.901**	**0.959**	1.013	1.011	1.042	1.084	1.105	1.109	1.184	1.200	1.172
(idf)	BVAR-TXT	HF	-	**0.899***	**0.764*****	**0.908***	**0.935**	**0.904**	**0.964**	1.003	1.012	1.043	1.081	1.113	1.112	1.184	1.198	1.178
(tfm)	MF-BVAR-TXT	MF	-	**0.858****	**0.749*****	**0.881***	**0.909**	**0.889**	**0.949**	**0.977**	**0.987**	1.017	1.065	1.100	1.102	1.172	1.193	1.174
(idf)	MF-BVAR-TXT	MF	-	**0.909***	**0.780*****	**0.904***	**0.949**	**0.911**	**0.963**	1.016	1.020	1.052	1.063	1.095	1.098	1.148	1.166	1.147
(tfm)	SV-BVAR-TXT	HF	-	**0.886****	**0.779***	**0.898****	**0.920**	**0.878**	**0.934**	**0.934**	**0.933**	**0.951**	1.078	1.249	1.159	**0.945**	**0.866**	**0.927**
(idf)	SV-BVAR-TXT	HF	-	**0.887****	**0.784*****	**0.899****	**0.914**	**0.888**	**0.925**	**0.942**	**0.958**	**0.945**	1.050	1.265	1.018	**0.956**	**0.892**	**0.952**
(tfm)	MF-SV-BVAR-TXT	MF	-	**0.848****	**0.739*****	**0.852****	**0.892**	**0.844**	**0.896**	**0.933**	**0.881**	**0.913**	1.048	1.091	1.005	**0.968**	**0.929**	**0.978**
(idf)	MF-SV-BVAR-TXT	MF	-	**0.930***	**0.805*****	**0.925***	**0.950**	**0.899**	**0.955**	**0.980**	**0.964**	**0.973**	1.106	1.164	1.074	1.017	1.006	1.073

Note: In column 1 AR: autoregression model, ARX: autoregression augmented with (no text-based) explanatory variables, VAR: vector autoregression, BVAR: Bayesian vector autoregression, SV-BVAR: Bayesian vector autoregression assuming stochastic volatility of the error term, AR-X-TXT, autoregression model augmented with text and no text-based explanatory variables, MIDAS-TXT: mixed frequency model, BVAR-TXT: text-based Bayesian vector autoregression model, MF-BVAR-TXT: mixed frequency text-based vector autoregression, SV-BVAR-TXT: stochastic volatility text-based Bayesian vector autoregression, MF-SV-BVAR-TXT: mixed frequency stochastic volatility text-based Bayesian vector autoregression, tfm: term-frequency matrix, idf: term-frequency inverse-document-frequency matrix. In column 2, HF and MF respectively denote homogeneous and mixed frequency models. Column 3 displays the weighting scheme of MIDAS parameters. Bold values indicate improvements on the no-change forecast. Blue entries stand for the lowest MSPE for a given time horizon, relative to a specific oil price measure. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table B.4: Recursive ALPL ratios relative to a random walk density forecast of alternative monthly indicators of real oil prices; text vs. no-text with uncertainty text indicators.

Model	Text Variable	1-month			3-months			6-months			12-months			24-months		
		WTI	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*	*WTI*	*RAC*	*BRENT*
SV-BVAR	-	**0.961*****	1.009***	0.947***	0.123***	0.032***	0.122***	-0.285***	-0.326***	-0.339***	-0.601***	-0.544***	-0.778***	-0.971***	-0.994***	-1.331***
SV-BVAR-TXT	U	0.939***	0.996***	0.941***	0.067***	0.018***	**0.133*****	-0.295***	-0.343***	**-0.282*****	**-0.499*****	-0.710***	**-0.754*****	**-0.949*****	**-0.962*****	-1.414***
SV-BVAR-TXT	G	0.939***	**1.009*****	**0.949*****	0.027***	0.029***	0.110***	-0.286***	-0.331***	-0.367***	**-0.593*****	**-0.539*****	**-0.769*****	**-0.964*****	**-0.994*****	**-1.312*****
SV-BVAR-TXT	F	0.926***	0.993***	0.938***	0.019***	**0.053*****	0.110***	-0.301***	**-0.317*****	-0.341***	-0.615***	-0.559***	**-0.772*****	-0.998***	-1.017***	**-1.330*****
MF-SV-BVAR	-	0.931***	1.009***	0.947***	0.023***	0.032***	0.122***	-0.285***	-0.326***	-0.339***	-0.601***	-0.544***	-0.778***	-0.971***	-0.994***	-1.331***
MF-SV-BVAR-TXT	U	0.925***	0.970***	0.935***	**0.044*****	**0.038*****	**0.131*****	**-0.280*****	**-0.321*****	**-0.293*****	-0.665***	-0.679***	**-0.707*****	**-0.917*****	**-0.952*****	-1.349***
MF-SV-BVAR-TXT	G	0.924***	0.993***	0.940***	**0.073*****	**0.039*****	0.050***	-0.313***	-0.352***	-0.365***	-0.640***	-0.576***	**-0.737*****	**-0.963*****	**-0.956*****	**-1.320*****
MF-SV-BVAR-TXT	F	0.923***	1.004***	**0.952*****	**0.030*****	**0.050*****	**0.138*****	-0.316***	**-0.311*****	**-0.336*****	-0.664***	-0.599***	-0.841***	-1.110***	-1.228***	-1.598***
CRB index																
MIDAS	-	-3.928**	-3.890*	-3.821***	-4.923	-4.326	-4.553	-5.578	-4.881	-5.579	-7.653	-6.269	-7.364*	-12.874*	-11.194*	-12.720*
TXT-MIDAS	U	-4.033**	**-3.419****	**-3.562*****	**-4.869**	**-4.136**	-4.869	**-5.424**	**-4.328**	-5.945	**-7.083**	-6.662	-7.506*	**-11.580***	-11.867*	-13.640*
TXT-MIDAS	G	-3.968**	**-3.409****	**-3.587*****	**-4.806**	-4.562	**-4.302**	**-5.201**	-4.968	-5.656	**-7.335**	-6.715	-7.558*	**-11.050***	**-9.417***	**-11.734***
TXT-MIDAS	F	-3.984**	**-3.528****	**-3.564*****	**-4.508**	-4.904	-4.899	-6.061	**-4.790**	-5.882	-7.890	-6.420	**-7.171***	-12.900*	**-9.700****	**-11.735***
Euro Stoxx 50 index																
MIDAS	-	-4.838*	-4.165**	-4.160**	-6.888	-5.720	-6.160	-8.843	-7.049	-7.558	-9.114*	-7.630*	-8.708**	-10.199*	-9.218*	-10.654*
TXT-MIDAS	U	**-4.782***	-4.933**	-4.662**	**-6.072**	**-5.660**	-6.258	**-8.496**	-7.755	-7.808	**-8.177***	-7.793*	**-8.383****	-10.560*	**-9.205***	**-10.357****
TXT-MIDAS	G	**-4.600***	**-4.096****	-4.485**	**-6.796**	**-5.478**	-6.297	**-8.418**	**-6.630**	-7.860	**-8.184***	**-7.125***	**-8.059****	-10.764*	-9.457*	-10.931*
TXT-MIDAS	F	**-4.659***	-4.363**	**-4.122****	-6.999	**-5.369**	-6.836	-8.933	**-6.494**	-7.695	**-8.679***	-7.711*	**-8.701****	-10.580*	-9.778*	**-10.405***
GBP/USD index																
MIDAS	-	-4.415**	-4.192**	-4.487**	-5.894	-5.094	-5.757	-7.134	-5.972	-6.554	-7.602*	-6.541**	-7.622**	-13.630*	-13.521**	-15.673**
TXT-MIDAS	U	-4.840**	-4.699**	**-4.305****	**-5.244**	**-4.531**	-5.875	-7.851	**-5.217**	-6.890	-7.950*	**-6.052****	**-7.078****	**-12.482***	**-13.461***	**-15.603****
TXT-MIDAS	G	-4.496**	-4.983**	-4.909**	-5.973	**-4.494**	**-5.604**	-7.657	**-5.742**	**-6.298**	**-7.515***	-6.729*	-7.641**	**-11.045***	**-11.164****	**-13.565****
TXT-MIDAS	F	**-4.118****	-4.664**	-4.630**	**-5.725**	-5.987	-5.987	-7.472	**-5.083**	**-6.269**	-7.652*	**-6.343***	-7.688**	**-12.720***	**-12.506***	**-13.911****
Natural Gas index																
MIDAS	-	-5.201	-4.431	-4.709*	-8.605	-6.198	-8.959	-9.444	-7.467	-9.651	-8.929*	-7.283*	-8.496**	-11.247**	-10.165**	-11.722**
TXT-MIDAS	U	**-5.155**	**-4.205**	**-4.352***	**-8.556**	-6.848	**-8.319**	**-9.368**	-7.897	**-9.178**	**-8.637***	**-7.011***	**-8.402****	**-10.904****	**-9.633****	**-10.813****
TXT-MIDAS	G	-5.986	-4.929*	**-4.576***	**-8.339**	-6.592	**-8.882**	**-9.301**	-7.841	**-9.379**	**-8.612***	**-7.007****	**-8.049****	**-9.718****	**-8.853****	**-10.724****
TXT-MIDAS	F	**-5.100***	-4.526*	**-4.218***	**-8.027**	**-5.663**	**-8.901**	**-9.237**	-7.540	**-9.099**	**-8.546***	**-7.106***	**-8.154****	**-10.279****	**-9.282****	**-11.018****
Palladium index																
MIDAS	-	-4.803*	-4.087*	-4.257*	-6.490	-5.454	-5.759	-8.474	-6.674	-7.130	-8.898*	-7.360*	-8.449**	-11.197**	-9.835*	-11.389**
TXT-MIDAS	U	**-4.448***	-4.216*	-4.862*	**-6.343**	-5.960	-5.799	-8.807	**-6.631**	-7.555	**-8.873***	-7.517*	-8.721**	**-11.094****	**-9.623***	-11.929**
TXT-MIDAS	G	**-4.473***	**-3.825***	-4.409*	-6.988	**-5.032**	**-5.552**	**-8.401**	**-6.674**	-7.146	**-8.516***	**-6.311****	-8.840**	**-10.921****	-9.922*	**-11.220****
TXT-MIDAS	F	**-4.445***	-4.748*	-4.364*	**-6.340**	-5.679	**-5.439**	**-8.329**	**-6.034**	-7.905	**-8.272***	**-7.275***	**-8.159****	-11.277**	-9.836**	-11.493**

Note: In column 1 SV: stochastic volatility, BVAR: Bayesian vector autoregression, MF: mixed frequency, TXT: text data defined in column 2. In column 2 U: unigram count, G: geometrical term-frequency model, F: geometrical term-frequency inverse-document frequency model. Bold values indicate the highest ALPL improvements on the no-text based model, for a given time horizon and relative to a specific oil price measure. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table B.5: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with high-frequency financial variables; text vs. no-text with uncertainty text indicators.

Monthly horizon	TXT-FIN-MIDAS Uc-MIDAS	idx-MIDAS	idf-MIDAS	FIN-MIDAS OFs-MIDAS
1	**0.914***	**0.892****	**0.909***	**0.906***
3	1.056	1.087	1.084	1.082
6	1.073	1.094	1.099	1.101
12	1.295	1.213	1.175	1.186
24	3.737	1.768	1.619	1.577
				GLs-MIDAS
1	**0.909****	**0.906****	**0.909****	**0.911****
3	1.099	1.101	1.109	1.102
6	1.144	1.155	1.158	1.147
12	1.216	1.219	1.211	1.212
24	1.447	1.438	1.452	1.430
				CRB-MIDAS
1	**0.868****	**0.858****	**0.867****	1.718
3	**0.944**	**0.938**	**0.941**	2.566
6	**0.988**	**0.990**	1.025	1.538
12	1.103	1.121	1.084	1.338
24	1.466	1.389	1.489	6.618
				BDi-MIDAS
1	**0.932**	**0.917**	**0.919**	**0.914***
3	1.120	1.108	1.100	1.087
6	1.184	1.111	1.106	1.107
12	1.298	1.302	1.288	1.287
24	1.374	1.304	1.351	1.342
				USi-MIDAS
1	**0.907***	**0.905***	**0.912***	**0.909***
3	1.067	1.061	1.074	1.046
6	1.062	1.060	1.061	1.067
12	1.089	1.077	1.092	1.078
24	1.506	1.456	1.516	1.512

Note: For column headers OFs-MIDAS, GLs-MIDAS, CRB-MIDAS, BDi-MIDAS, USi-MIDAS denote MIDAS models where the high-frequency financial variables fitting the polynomial are (i) crude oil-futures prices spread, (ii) crude oil-gasoline spread, (iii) CRB spot price index, (iv) Baltic Dry index and (v) the federal funds rate. Each outcome is then compared to the case in which text data are included in the model, in addition to the financial variable. In particular, Uc-MIDAS, idx-MIDAS, idf-MIDAS denote MIDAS models where the text variable fitting the polynomial is respectively developed through (i) unigram word-count, (ii) term-frequency matrix and (iii) term-frequency inverse-document frequency matrix. Black bold values indicate improvements on the no-change forecast. Green bold values indicate improvements of TXT-FIN-MIDAS on FIN-MIDAS. Blue entries stand for the lowest MSPE for a given time horizon. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table B.6: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with realized volatility of ultra-high-frequency financial variables; text vs. no-text with uncertainty text indicators.

Monthly	**TXT-FIN-MIDAS**			**FIN-MIDAS**
horizon	Uc-IR-MIDAS	idx-IR-MIDAS	idf-IR-MIDAS	FT-IR-MIDAS
1	**0.887***	**0.888***	**0.884***	**0.884***
3	1.079	1.064	1.081	1.072
6	1.081	1.079	1.074	1.074
12	1.086	1.093	1.085	1.080
24	1.394	1.371	1.366	1.373
				SP-IR-MIDAS
1	**0.891***	**0.889****	**0.882***	**0.876***
3	1.053	1.052	1.052	1.050
6	1.072	1.059	1.064	1.063
12	1.091	1.096	1.086	1.083
24	1.447	1.425	1.420	1.428
				ES-IR-MIDAS
1	**0.881****	**0.879****	**0.879****	**0.876****
3	1.073	1.072	1.075	1.071
6	1.115	1.103	1.110	1.111
12	1.133	1.165	1.136	1.139
24	1.348	1.337	1.337	1.352
				HS-IR-MIDAS
1	**0.887***	**0.890***	**0.895***	**0.889***
3	1.063	1.060	1.075	1.068
6	1.092	1.097	1.090	1.091
12	1.084	1.115	1.091	1.098
24	1.506	1.478	1.519	1.494
				PD-IR-MIDAS
1	**0.883***	**0.881****	**0.882****	**0.882****
3	1.037	1.029	1.049	1.036
6	1.067	1.072	1.072	1.062
12	1.148	1.148	1.147	1.141
24	1.610	1.565	1.607	1.601
				CD-IR-MIDAS
1	**0.922****	**0.923***	**0.921***	**0.916****
3	1.198	1.244	1.218	1.205
6	1.290	1.396	1.505	1.419
12	1.386	1.414	1.359	1.387
24	8.013	3.482	4.072	6.278
				ED-IR-MIDAS
1	**0.901***	**0.895***	**0.907***	**0.904***
3	1.035	1.016	1.004	1.020
6	1.068	1.062	1.077	1.060
12	1.131	1.122	1.117	1.113
24	1.551	1.501	1.542	1.543

Note: For column headers FT-IR-MIDAS, SP-IR-MIDAS, ES-IR-MIDAS, HS-IR-MIDAS, PD-IR-MIDAS, CD-IR-MIDAS, ED-IR-MIDAS denote MIDAS models where the ultra-high-frequency financial variables fitting the polynomial are (i) intraday returns of FTSE100 index , (ii) intraday returns of S&P500 index, (iii) intraday returns of Euro Stoxx 50 index, (iv) intraday returns of Hang Seng index, (v) intraday returns of GBP/USD exchange rate, (vi) intraday returns of CAD/USD exchange rate and (vii) intraday returns of EUR/USD exchange rate. Each outcome is then compared to the case in which text data are included in the model, in addition to the financial variable. In particular, Uc-MIDAS, idx-MIDAS and idf-MIDAS denote MIDAS models where the text variable fitting the polynomial is respectively developed through (i) unigram word-count, (ii) term-frequency matrix and (iii) term-frequency inverse-document frequency matrix. Black bold values indicate improvements on the no-change forecast. Green bold values indicate improvements of TXT-FIN-MIDAS on FIN-MIDAS. Blue entries stand for the lowest MSPE for a given time horizon. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Table B.7: Recursive MSPE ratios relative to a random walk forecast of monthly real oil prices. MIDAS with realized volatility of ultra-high-frequency commodity prices; text vs. no-text with uncertainty text indicators.

Monthly	**TXT-CMDTY-MIDAS**			**CMDTY-MIDAS**
horizon	Uc-IR-MIDAS	idx-IR-MIDAS	idf-IR-MIDAS	CO-IR-MIDAS
1	**0.897***	**0.899***	**0.902***	**0.895***
3	1.053	1.078	1.077	1.081
6	1.061	1.060	1.060	1.063
12	1.052	1.050	1.060	1.057
24	1.310	1.326	1.287	1.280
				GL-IR-MIDAS
1	**0.897***	**0.894***	**0.897***	**0.896***
3	1.068	1.066	1.082	1.085
6	1.060	1.053	1.070	1.061
12	1.098	1.094	1.088	1.087
24	1.432	1.392	1.419	1.394
				CP-IR-MIDAS
1	**0.918****	**0.926***	**0.926***	**0.941***
3	1.176	1.126	1.098	1.135
6	1.125	1.106	1.125	1.104
12	1.120	1.098	1.087	1.172
24	1.320	1.276	1.322	1.316
				NG-IR-MIDAS
1	**0.927***	**0.931***	**0.923***	**0.922***
3	1.140	1.138	1.149	1.148
6	1.166	1.175	1.174	1.159
12	1.168	1.221	1.163	1.165
24	1.440	1.398	1.403	1.402
				PL-IR-MIDAS
1	**0.918***	**0.908****	**0.902****	**0.914***
3	1.095	1.088	1.091	1.081
6	1.089	1.110	1.120	1.121
12	1.155	1.186	1.145	1.155
24	1.448	1.443	1.574	1.411
				SL-IR-MIDAS
1	**0.897***	**0.892***	**0.893***	**0.893***
3	1.074	1.075	1.085	1.088
6	1.060	1.061	1.072	1.059
12	1.075	1.086	1.073	1.074
24	1.373	1.359	1.384	1.393

Note: For column headers CO-IR-MIDAS, GL-IR-MIDAS, CP-IR-MIDAS, NG-IR-MIDAS, PL-IR-MIDAS, SL-IR-MIDAS, denote MIDAS models where the ultra-high-frequency commodity variables fitting the polynomial are (i) intraday returns of WTI index, (ii) intraday returns of Gold index, (iii) intraday returns of Copper index, (iv) intraday returns of Natural Gas index, (v) intraday returns of Palladium index and (vi) intraday returns of Silver index. Each outcome is then compared to the case in which text data are included in the model, in addition to the commodity variable. In particular, Uc-MIDAS, idx-MIDAS and idf-MIDAS denote MIDAS models where the text variable fitting the polynomial is respectively developed through (i) unigram word-count, (ii) term-frequency matrix and (iii) term-frequency inverse-document frequency matrix. Black bold values indicate improvements on the no-change forecast. Green bold values indicate improvements of TXT-CMDTY-MIDAS on CMDTY-MIDAS. Blue entries stand for the lowest MSPE for a given time horizon. *, ** and *** respectively denote 10%, 5% and 1% level of significance as suggested by the Diebold-Mariano test.

Appendix

Chapter 3

C.1 Data Sources

The time series of the first three dependent variables ($q^{oil}, y^{GDP}, p^{oil}$) in experiment 1 and 2 are based on Kilian [2009]'s and Kilian and Murphy [2012]'s original data, made available on the public library of *Journal of the European Economic Association* (http://onlinelibrary.wiley.com/doi/10.1111/j.1542-4774.2012.01080.x/suppinfo). q^{oil} is expressed as 100 times the log difference of oil productivity growth, y^{GDP} relies on shipping costs, and p^{oil} is based on the U.S. refiner acquisition costs (RAC). the latter proxy can be easily downloaded from the U.S. Energy Information Administration page in the section *Imported*, and than deflated by the U.S. consumer price index (CPI) available from the Federal Reserve Economic Data (FRED) database maintained by the St. Louis FED (CPIAUCSL).

In experiment 3, data are observed from 1974M1 to 2019M12, therefore y^{GDP} has been downloaded from Lutz Kilian's official page, while oil production and RAC are made available on Christiane Baumeister's official page (https://sites.google.com/site/cjsbaumeister/research). However, FRED is a valid alternative solution for the last two variables.

In Appendix C.5, experiment 1 and 2 are proposed again with Kilian's updated index, downloaded from his official page (https://sites.google.com/site/lkilian2019/research/data-sets).

In regard to the remaining seven time series, oil inventories data come from Christiane Baumeister's official page, and the residual six factors from FRED, coded as follows. (i) Producer Price Index by Commodity: Metals and Metal Products (WPU10); (ii) Gold Fixing Price 10:30 A.M. (London time) in London Bullion Market, based in U.S. Dollars (GOLDAMGBD228NLBM); (iii) Producer Price Index by Commodity: Fuels and Related Products and Power: Coal (WPU051); (iv) Producer Price Index by Commodity: Metals and Metal Products: Iron and Steel (WPU101); (v) U.S./U.K. Foreign Exchange Rate (DEXUSUK); (vi) Consumer Price Index: OECD Groups: All Items Non-Food and Non-Energy for Japan (CPGRLE01JPM657N).

C.2 Informative Prior

In sections 4.3.5 and 4.3.6 I have outlined the structure of prior distributions for parameters A_0, A_j and Λ, addressing the reader to consult this section for a detailed description of the methodology implemented across the different identification strategies. It is important, in fact, to illustrate the specific steps that form the basis for the posterior sampling presented in the next section.

C.2.1 Prior for A_0

Short-run recursive. Suppose to identify the structural VAR as expressed in equation (4.7) by imposing zero contemporaneous restrictions on the covariance matrix A_0^{-1}. Prior information regarding the expected value of elements α_{ij}^0 of the lower triangular matrix A_0 are described by an uninformed t-Student distribution with mode at $c = 0$, scale parameter $\sigma = 100$, and $\phi = 3$ degrees of freedom. This distribution is expressed as:

$$p\left(\alpha_{ij}^0\right) = \frac{\Gamma\left(\frac{\phi+1}{2}\right)}{\Gamma\left(\frac{\phi}{2}\right)\sqrt{\pi\phi\sigma}}\left[1 + \frac{1}{\phi}\left(\frac{\alpha_{ij} - c}{\sigma}\right)^2\right]^{-\frac{\phi+1}{2}}$$

where i and j are rows and columns of the Cholesky matrix A_0.

Sign restrictions. In this case I assume that structural shocks at time zero generate an impact effect, whose sign is known and follows Kilian and Murphy [2012]'s rule (see section 4.3.3). This hypothesis is used to inform my prior $p(A_0)$ by imposing a positive truncated t-Student distribution for values α_{ij}^0 which positively respond to a structural disturbance, and a negative truncated t-Student distribution in the opposite case. In this way candidate draws of α_{ij}^0 are sampled from the appropriate distribution. When the SVAR is identified via heteroskedasticity, the same rules of lower triangular identification are applied.

C.2.2 Prior for $\Lambda|A_0$

For this prior, and even for $p\left(A_j|\Lambda, A_0\right)$, there is no difference across the different experiments investigated. Therefore, as remarked in section 4.3.5, I assume that λ_{ii}^{-1} follow a $\Gamma\left(\underline{\kappa}_i, \underline{\tau}_i(A_0)\right)$ distribution, where $\underline{\tau}$ depends on A_0, while $\underline{\kappa}$ does not. Baumeister and Hamilton [2015] show that $\underline{\kappa}/\underline{\tau}$ and $\underline{\kappa}/\underline{\tau}^2$ are the prior mean and the prior variance of λ_{ii}^{-1}, which is set equal to $A_0'\hat{S}A_0$. This implies that $\underline{\tau}_i(A_0) = \underline{\kappa}_i A_0'\hat{S}A_0$, with $\underline{\kappa}_i = 0.5$ and $\hat{S}$ the $N \times N$ OLS variance matrix of y_t.

C.2.3 Prior for $A_j|\Lambda, A_0$

Based on Baumeister and Hamilton [2019], individual lagged coefficients of matrix A_j have a conditional Normal distribution of the form $\left(\alpha_{ij}^{j}|A_0,\Lambda\right) \sim N\left(\underline{m}_i, \lambda_{ii}\underline{M}\right)$, where α_{ij}^{j} are supposed to behave like random walks, and thus $m_i = 0$. $\underline{M}$ is the prior variance, and as in Doan [2013] the matrix is set diagonal, with hyperparameters that reflect a Minnesota structure of the form:

$$\underline{M}_{i,rr} = \begin{cases} \lambda_0^2\left(\frac{1}{\hat{S}p^{\lambda_1}}\right) & \text{for coefficients on own lag } j,\ \forall\ j = 1,\ldots,24 \\ \lambda_0^2\left(\frac{1}{\hat{S}p^{\lambda_1}\lambda_2}\right) & \text{for lagged coefficients } j \text{ of variable } r \neq i,\ \forall\ j = 1,\ldots,24 \\ \lambda_0^2\left(100\lambda_3\right) & \text{for the intercept} \end{cases}$$

λ_0 represents the overall confidence of prior beliefs; bigger values of λ_0 imply a lower weight for the random walk behavior. λ_1 describes how fast lagged coefficients shrink to zero when j increases; when the hyperparameter is set equal to zero, all lags have the same weight. λ_2 is the confidence in other-than-own coefficient lags. λ_3 is the variance of the intercept; high values of λ_3 imply that the constant term is not relevant. In the first three experiments $\left\{\lambda_0 = 10^9, \lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 100\right\}$, while in the last exercise $\lambda_0 = 0.5$ and the remaining hyperparameters do not change.

C.3 Posterior Sampling

The sampling procedure is equivalent in all experiments, even though we make two additional steps when the SVAR is identified via heteroskedasticity.

C.3.1 Posterior for $A_0|Y_T$

Given the prior value c, we first compute the posterior mode by maximizing the likelihood value of the lower triangular matrix A_0, then for any numerical value of A_0 we find the log of the target

$$\Upsilon\left(A_0^{max}\right) = log\left(p\left(A_0\right)\right) + \left(\frac{T}{2}\right) log\left[det\left(A_0\hat{\Omega_T}A_0'\right)\right] - \sum_{i=1}^{N}\overline{\kappa}_i log\left[\left(\frac{2}{T}\right)\overline{\tau}_i\left(A_0\right)\right] \tag{C.1}$$

which is used to inform the random walk Metropolis Algorithm and generate candidate draws of α_{ij}^0. $\overline{\kappa}$ and $\overline{\tau}$ are calculated with equation (4.8) and (4.9). Draws of α_{ij}^0 are progressively identified in each sampling step ρ as following. Suppose to generate a first draw for A_0 which we call A_0^{ρ} from a random walk expressed in the following form:

$$\hat{A}_0^{\rho} = A_0^{max} + \zeta \left(\hat{Q}^{-1}\right)' v_t \tag{C.2}$$

where A_0^{max} is the optimized value of the lower triangular matrix A_0 done before computing equation (C.1), ζ is a parameter scalar chosen so that 30% of draws are retained, and $\hat{Q}^{-1}$ is the Cholesky of second derivatives of A_0 used to improve the efficiency of the algorithm (see Baumeister and Hamilton [2019]). If $\Upsilon\left(\hat{A}_0^{\rho}\right) < \Upsilon\left(A_0^{max}\right)$, we set $A_0^{\rho} = A_0^{max}$ with probability $1 - exp\left[\hat{A}_0^{\rho} - A_0^{max}\right]$; otherwise, we set $A_0^{\rho} = \hat{A}_0^{\rho}$. In a second step, we draw $A_0^{\rho+1}$ from equation (C.2), but plug A_0^{ρ} in place of A_0^{max}, compare which parameter between $\hat{A}_0^{\rho+1}$ and A_0^{ρ} best describe $A_0^{\rho+1}$, and do that iteratively for $P = 20{,}000$ times.

C.3.2 Posterior for $\Lambda|A_0, Y_T$ and $A_j|\Lambda, A_0, Y_T$

For each of these P final values for α_{ij}^{0} we generate λ_{ii}^{-1} posterior candidates from a $\Gamma\left(\overline{\kappa}_i, \overline{\tau}_i\left(A_0^{\rho}\right)\right)$ distribution, and next we generate α_i^{j} posterior candidates from a $N\left(\overline{m}_i\left(A_0^{\rho}\right), \lambda_{ii}^{-1}\overline{M}_i\right)$. The process is repeated 10,000 times for each VAR combination, implying a total of 1,280,000 draws. The reduced form coefficients come up as the weighted average across the goodness-of-fit of each VAR, and then they are used in the dynamic analysis.

C.4 Supplementary Results of DGP

This section reports the remaining results generated through a Monte Carlo simulation exercise for model 2-5 of section 4.4 in the main paper. Structural VARs of each model are identified through (i) short-run restrictions, (ii) sign restrictions and (iii) heteroscedasticity.

C.5 Additional Empirical Evidence

It might be of interest to understand whether the empirical evidence provided in section 4.5 still holds when we use the updated version of Kilian's index. Figure C.13 and C.14 show the different size impact between old and revised index which may help to better understand the diversity between Figure C.1 and C.2, with experiment 1 and 2.

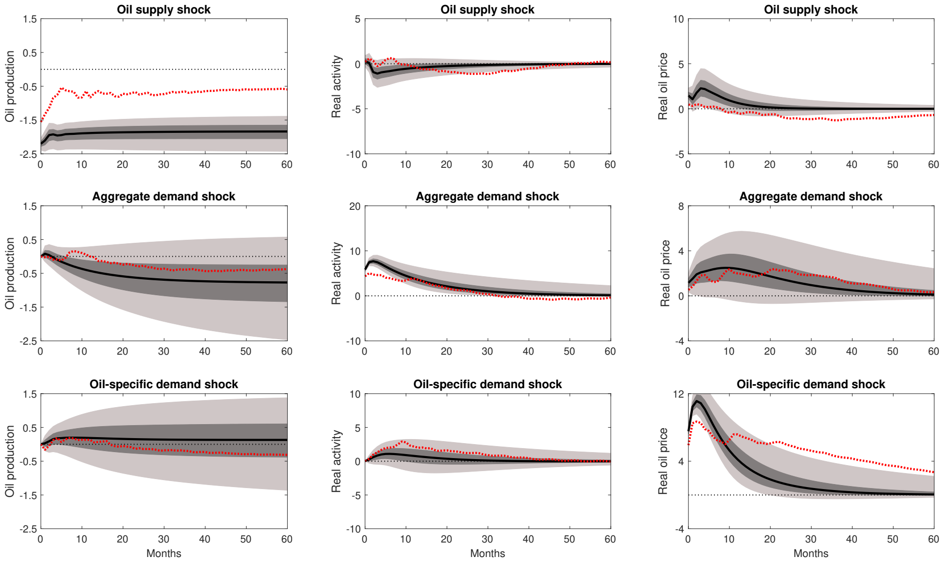


Figure C.1: Model 2: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through short-run restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

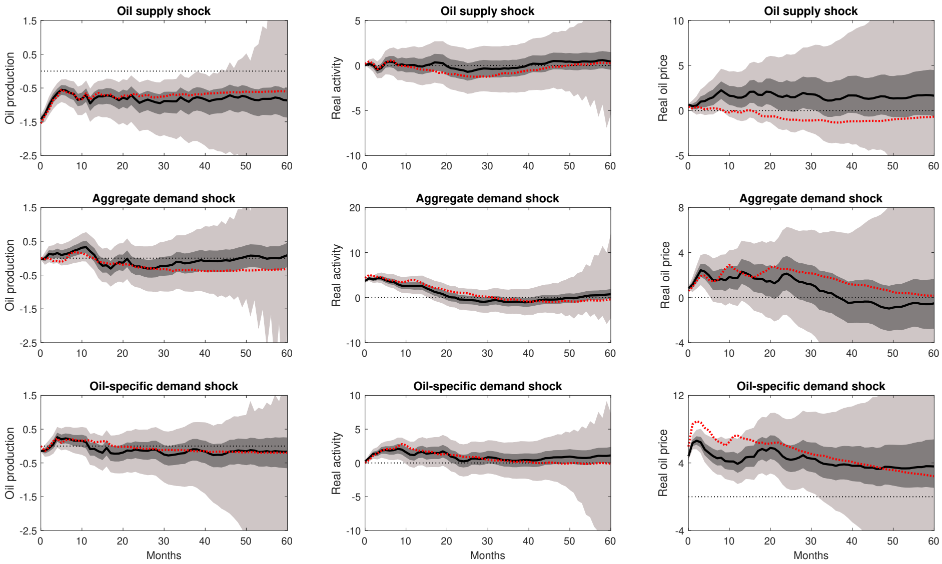


Figure C.2: Model 2: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through sign restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

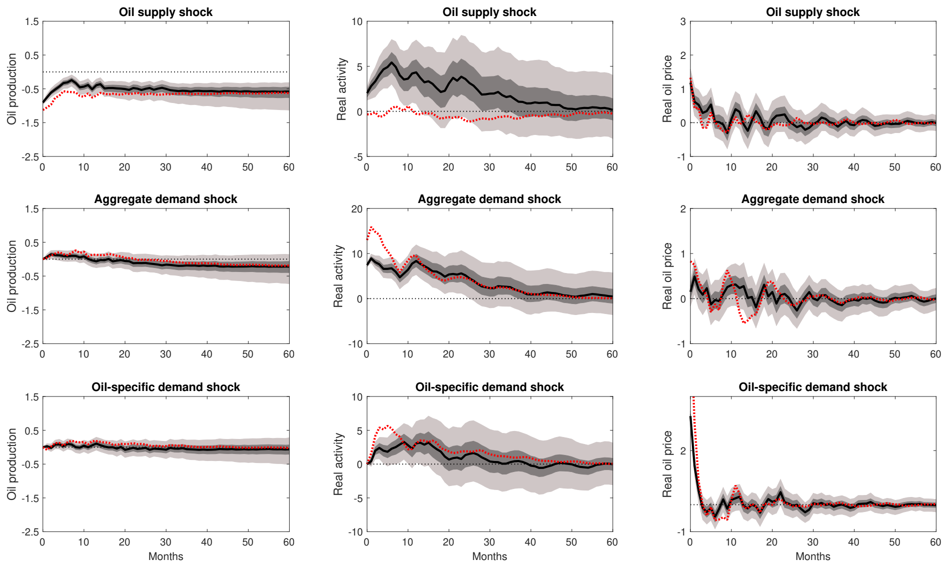


Figure C.3: Model 2: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified via heteroscedasticity. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

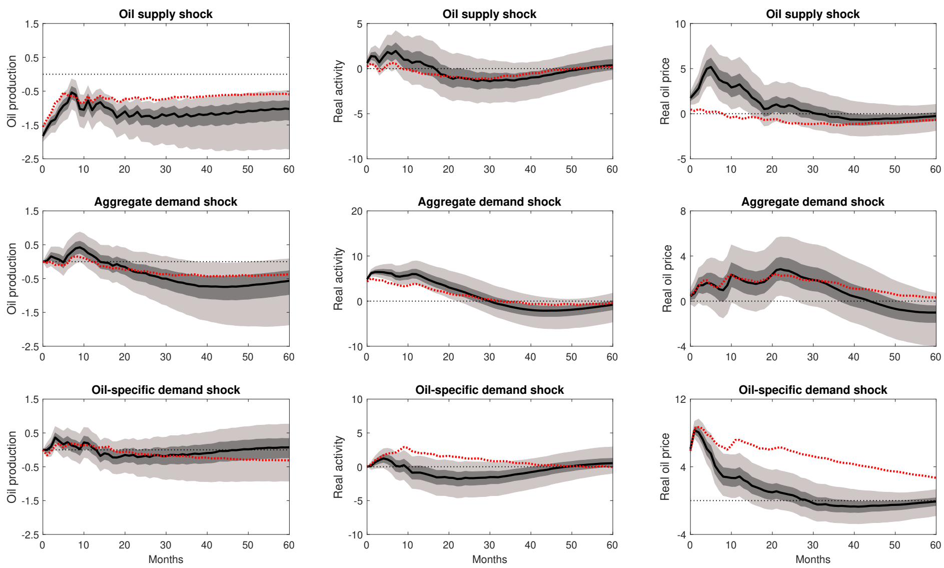


Figure C.4: Model 3: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through short-run restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

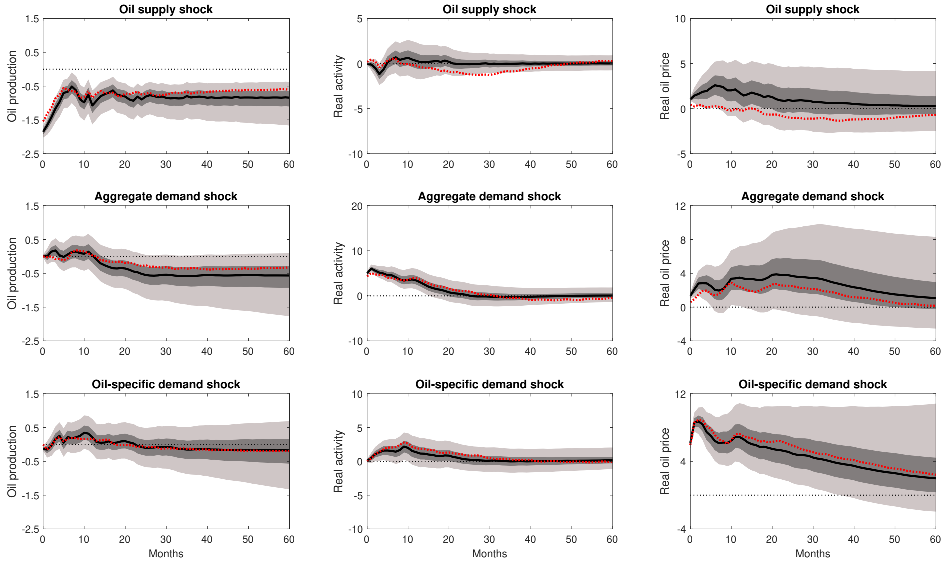


Figure C.5: Model 3: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through sign restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

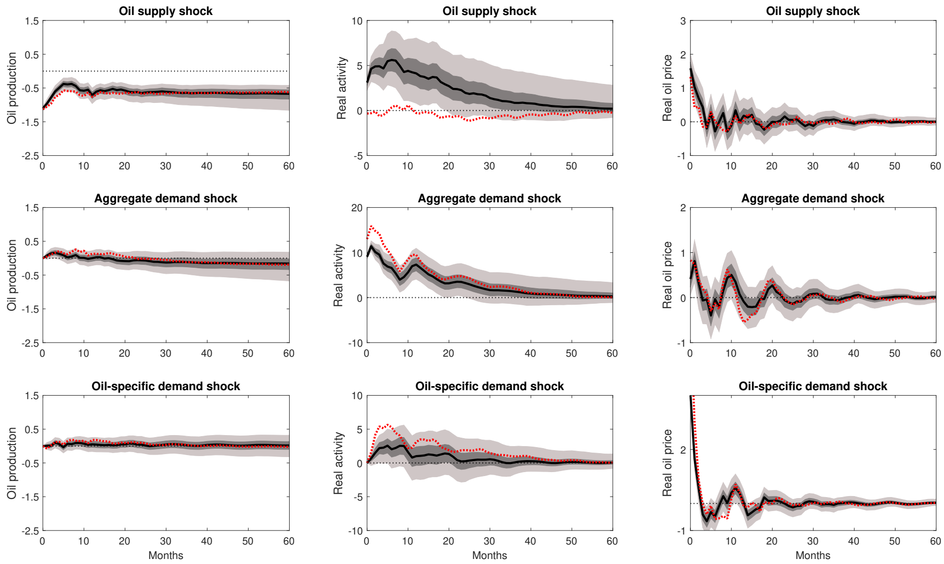


Figure C.6: Model 3: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified via heteroscedasticity. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

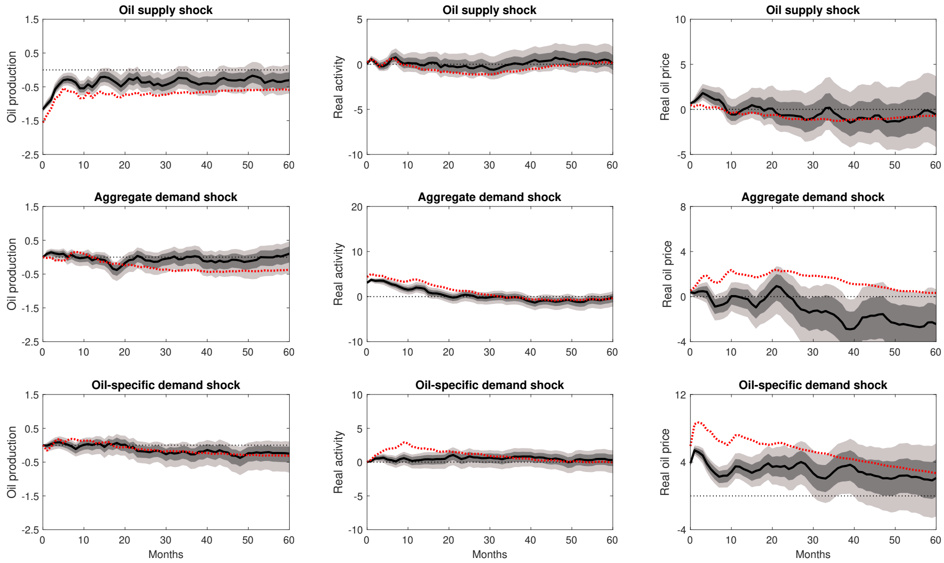


Figure C.7: Model 4: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through short-run restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

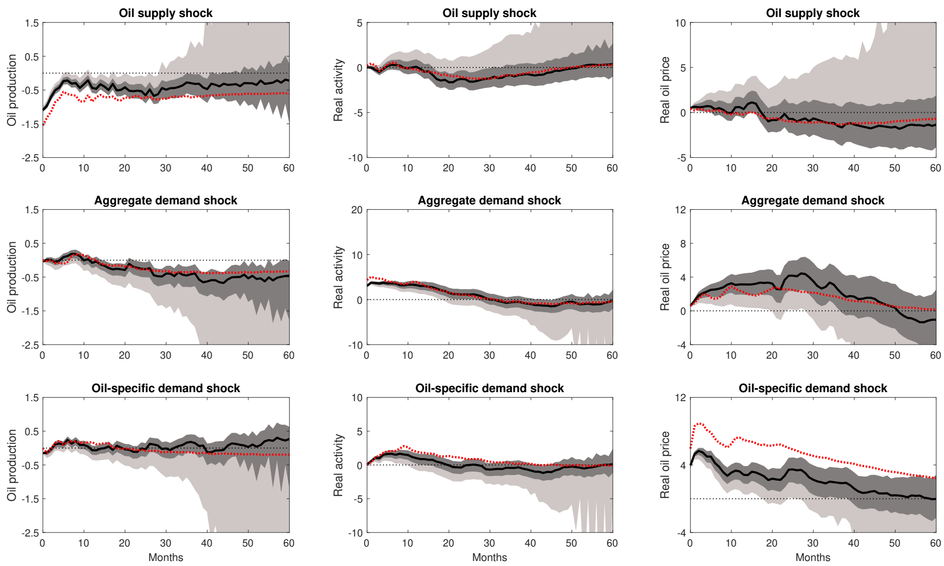


Figure C.8: Model 4: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through sign restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

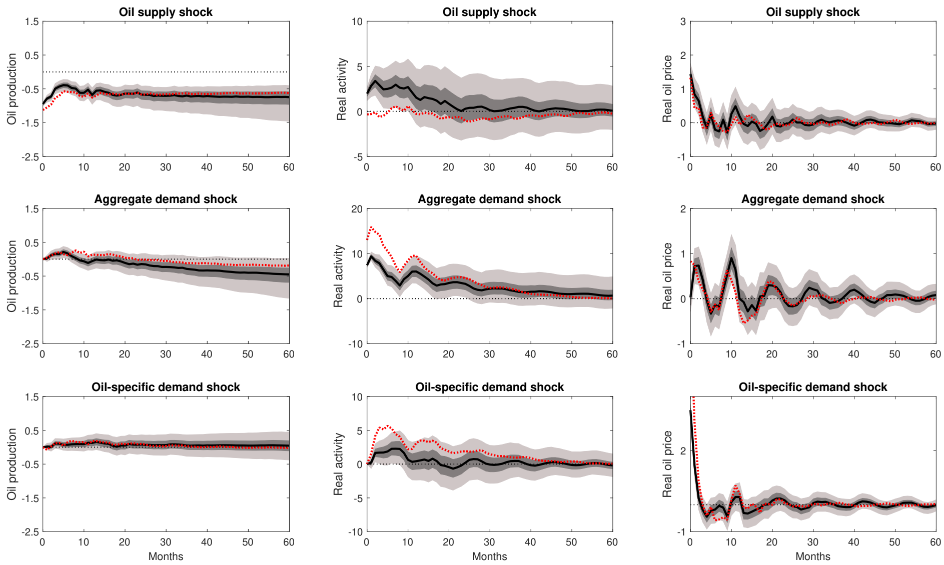


Figure C.9: Model 4: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified via heteroscedasticity. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

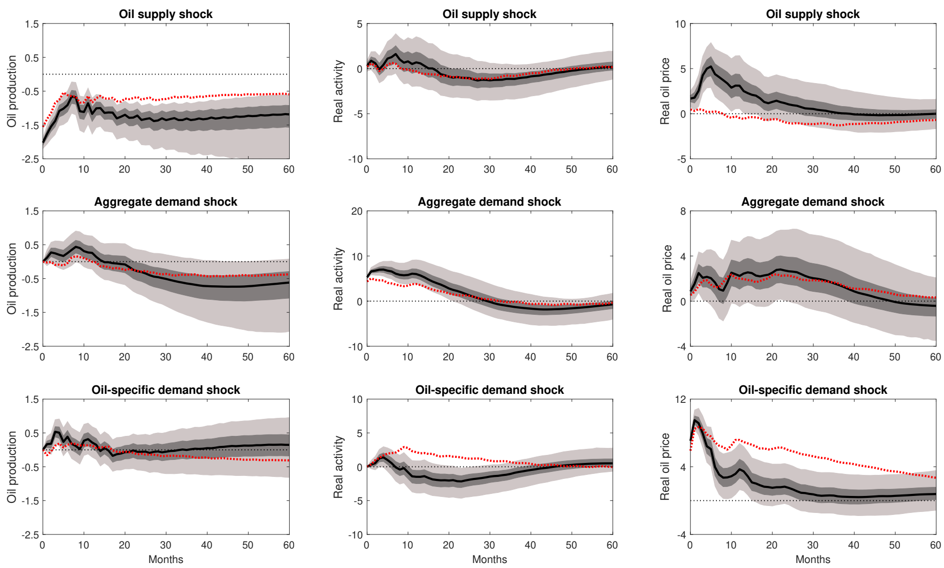


Figure C.10: Model 5: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through short-run restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

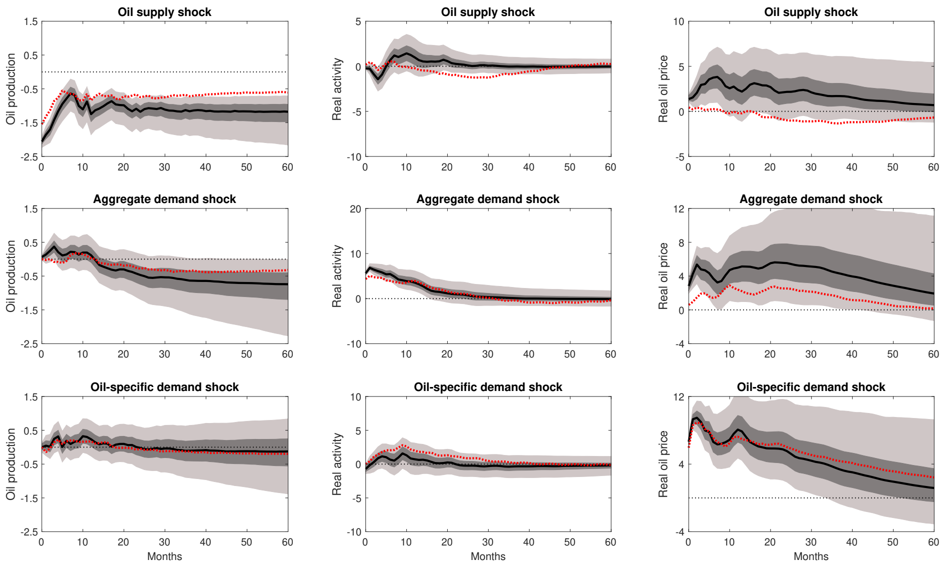


Figure C.11: Model 5: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified through sign restrictions. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

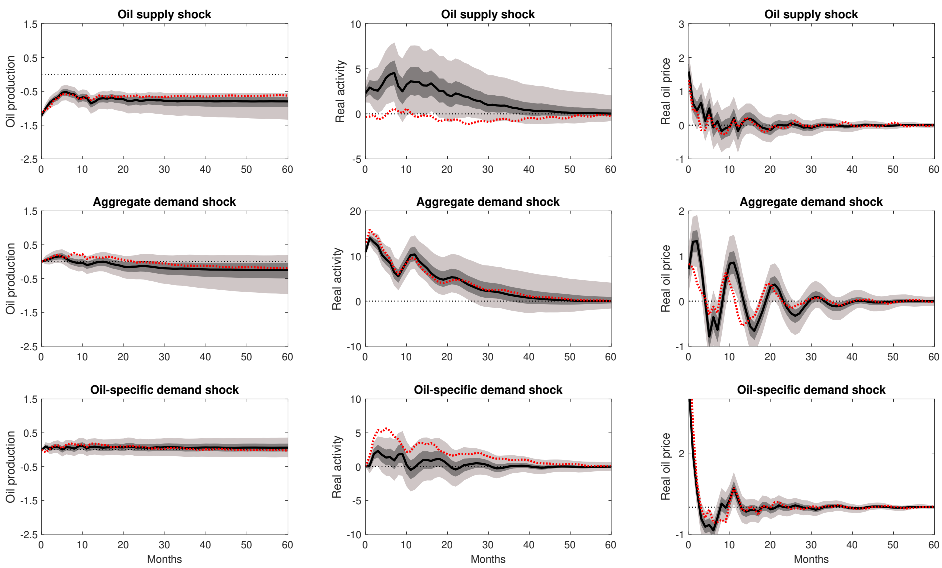


Figure C.12: Model 5: IRFs of the fixed 3×3 matrix of artificial generated data, where SVARs are identified via heteroscedasticity. Black solid lines show the median responses of the correctly specified model and the shaded dark and light regions describe the relative 90% and 68% posterior credible set. Red dotted lines show the response of the true model.

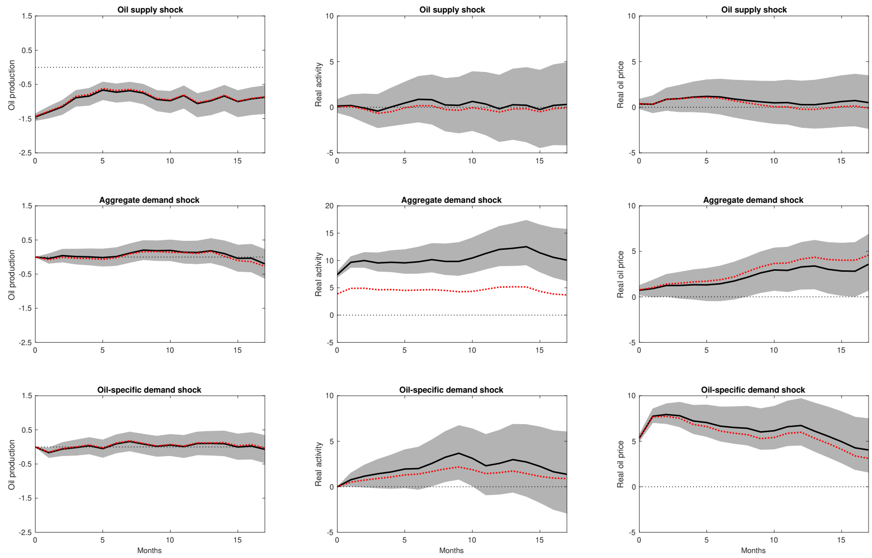


Figure C.13: Black solid lines show the median responses of Kilian (2009) with the updated index, whereas red dotted lines show the median response with previous measure of global real economy. Shaded regions describe the relative 95% posterior credible set.

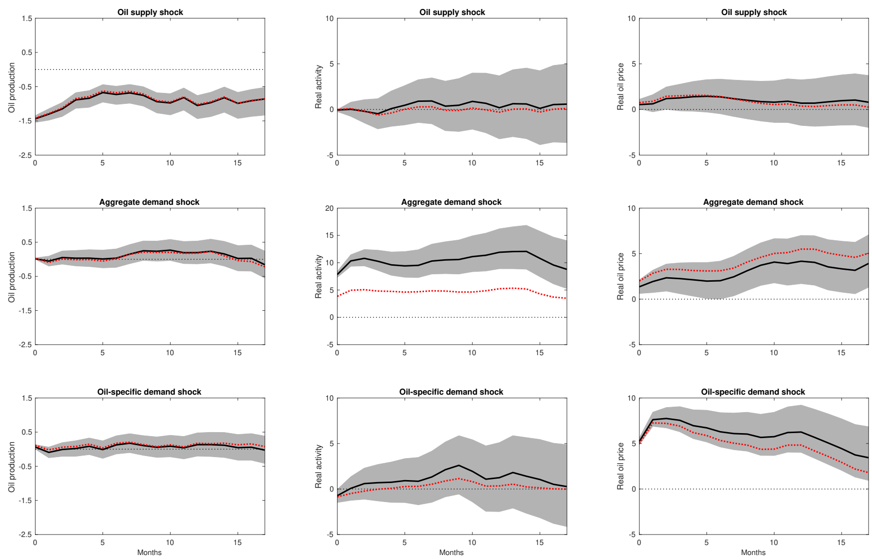


Figure C.14: Black solid lines show the median responses of Kilian and Murphy (2012) with the updated index, whereas red dotted lines show the median response with previous measure of global real economy. Shaded regions describe the relative 95% posterior credible set.

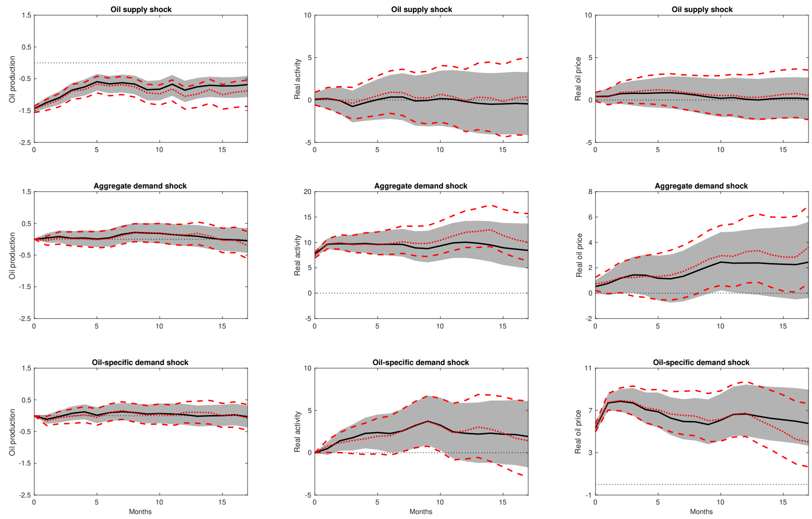


Figure C.15: IRFs of the fixed 3×3 matrix, where SVARs are identified through short-run restrictions and Kilian's updated index proxy for real world economy. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 3-variables misspecified VAR.

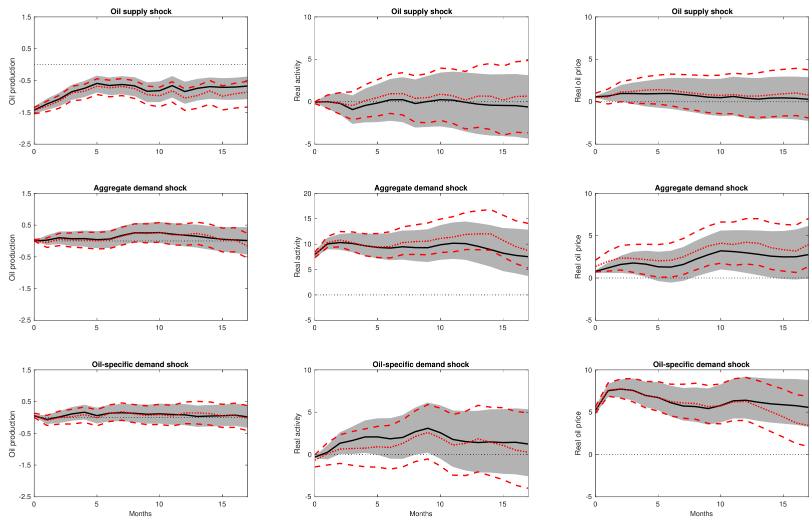


Figure C.16: IRFs of the fixed 3×3 matrix, where SVARs are identified through sign restrictions and Kilian's updated index proxy for real world economy. Black solid lines show the median responses of all model combinations and the shaded regions describe the relative 95% posterior credible set. Red dotted lines show the response of the 3-variables misspecified VAR.

Bibliography

Aastveit, K. A., Cross, J. L. and van Dijk, H. K. [2022], 'Quantifying time-varying forecast uncertainty and risk for the real price of oil', *Journal of Business & Economic Statistics* pp. 1–15.

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. and Jiao, J. [2015], 'An integrated text analytic framework for product defect discovery', *Production and Operations Management* **24**(6), 975–990.

Acosta, J. [2014], 'Fomc responses to calls for transparency: Evidence from the minutes and transcripts using latent semantic analysis', *Mimeograph, University of Stanford* .

Akram, Q. F. [2004], 'Oil prices and exchange rates: Norwegian evidence', *The Econometrics Journal* **7**(2), 476–504.

Alexopoulos, M., Cohen, J. et al. [2009], 'Uncertain times, uncertain measures', *University of Toronto Department of Economics Working Paper* **352**(7), 8.

Almon, S. [1965], 'The distributed lag between capital appropriations and expenditures', *Econometrica: Journal of the Econometric Society* pp. 178–196.

Aloui, C. and Jammazi, R. [2009], 'The effects of crude oil shocks on stock market shifts behaviour: A regime switching approach', *Energy economics* **31**(5), 789–799.

Alquist, R. and Kilian, L. [2010], 'What do we learn from the price of crude oil futures?', *Journal of Applied econometrics* **25**(4), 539–573.

Alquist, R., Kilian, L. and Vigfusson, R. J. [2013], Forecasting the price of oil, *in* 'Handbook of economic forecasting', Vol. 2, Elsevier, pp. 427–507.

Andersen, T. G. and Bollerslev, T. [1998], 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International economic review* pp. 885–905.

Ando, T. and Li, K.-C. [2014], 'A model-averaging approach for high-dimensional regression', *Journal of the American Statistical Association* **109**(505), 254–265.

Antweiler, W. and Frank, M. Z. [2004], 'Is all that talk just noise? the information content of internet stock message boards', *The Journal of finance* **59**(3), 1259–1294.

Apergis, N. and Miller, S. M. [2009], 'Do structural oil-market shocks affect stock prices?', *Energy economics* **31**(4), 569–575.

Ashenfelter, O. and Pencavel, J. H. [1969], 'American trade union growth: 1900-1960', *The Quarterly Journal of Economics* pp. 434–448.

Bachmeier, L. J. and Griffin, J. M. [2003], 'New evidence on asymmetric gasoline price responses', *Review of Economics and Statistics* **85**(3), 772–776.

Bai, Y., Li, X., Yu, H. and Jia, S. [2022], 'Crude oil price forecasting incorporating news text', *International Journal of Forecasting* **38**(1), 367–383.

Baker, S. R., Bloom, N. and Davis, S. J. [2016], 'Measuring economic policy uncertainty', *The quarterly journal of economics* **131**(4), 1593–1636.

Bakshi, G., Panayotov, G. and Skoulakis, G. [2010], 'The baltic dry index as a predictor of global stock returns, commodity returns, and global economic activity', *Commodity Returns, and Global Economic Activity (October 1, 2010)* .

Banchs, R. E. [2012], *Text mining with MATLAB®*, Springer Science & Business Media.

Banerjee, A., Marcellino, M. and Masten, I. [2005], 'Leading indicators for euro-area inflation and gdp growth', *Oxford Bulletin of Economics and Statistics* **67**, 785–813.

Barro, R. J. [1991], 'World interest rates and investment'.

Barro, R. J. and Sala-i Martin, X. [1990], 'World real interest rates', *NBER macroeconomics annual* **5**, 15–61.

Bates, J. M. and Granger, C. W. [1969], 'The combination of forecasts', *Journal of the Operational Research Society* **20**(4), 451–468.

Baumeister, C., Guérin, P. and Kilian, L. [2015], 'Do high-frequency financial data help forecast oil prices? the midas touch at work', *International Journal of Forecasting* **31**(2), 238–252.

Baumeister, C. and Hamilton, J. D. [2015], 'Sign restrictions, structural vector autoregressions, and useful prior information', *Econometrica* **83**(5), 1963–1999.

Baumeister, C. and Hamilton, J. D. [2019], 'Structural interpretation of vector autoregressions with incomplete identification: Revisiting the role of oil supply and demand shocks', *American Economic Review* **109**(5), 1873–1910.

Baumeister, C. and Kilian, L. [2012], 'Real-time forecasts of the real price of oil', *Journal of business & economic statistics* **30**(2), 326–336.

Baumeister, C. and Kilian, L. [2014], 'What central bankers need to know about forecasting oil prices', *International economic review* **55**(3), 869–889.

Baumeister, C. and Kilian, L. [2015], 'Forecasting the real price of oil in a changing world: A forecast combination approach', *Journal of Business & Economic Statistics* **33**(3), 338–351.

Baumeister, C., Korobilis, D. and Lee, T. K. [2020], 'Energy markets and global economic conditions', *The Review of Economics and Statistics* pp. 1–45.

Baumeister, C. and Peersman, G. [2013], 'Time-varying effects of oil supply shocks on the us economy', *American Economic Journal: Macroeconomics* **5**(4), 1–28.

Baumer, E., Elovic, E., Qin, Y., Polletta, F. and Gay, G. [2015], Testing and comparing computational approaches for identifying the language of framing in political news, *in* 'Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', pp. 1472–1482.

Baylis, P. [2020], 'Temperature and temperament: Evidence from twitter', *Journal of Public Economics* **184**, 104161.

Bernanke, B. S. and Boivin, J. [2003], 'Monetary policy in a data-rich environment', *Journal of Monetary Economics* **50**(3), 525–546.

Bernanke, B. S., Gertler, M., Watson, M., Sims, C. A. and Friedman, B. M. [1997], 'Systematic monetary policy and the effects of oil price shocks', *Brookings papers on economic activity* **1997**(1), 91–157.

Bholat, D., Hansen, S., Santos, P. and Schonhardt-Bailey, C. [2015], 'Text mining for central banks', *Available at SSRN 2624811* .

Blanchard, O. J. and Gali, J. [2007], The macroeconomic effects of oil shocks: Why are the 2000s so different from the 1970s?, Technical report, National Bureau of Economic Research.

Blinder, A. S. [1981], 'Temporary income taxes and consumer spending', *Journal of Political Economy* **89**(1), 26–53.

Borenstein, S., Cameron, A. C. and Gilbert, R. [1997], 'Do gasoline prices respond asymmetrically to crude oil price changes?', *The Quarterly journal of economics* **112**(1), 305–339.

Brunnermeier, M., Palia, D., Sastry, K. A., Sims, C. A. et al. [2017], 'Feedbacks: financial markets and economic activity', *Manuscript, Princeton University* .

Buckland, S. T., Burnham, K. P. and Augustin, N. H. [1997], 'Model selection: an integral part of inference', *Biometrics* pp. 603–618.

Bybee, L., Kelly, B. T., Manela, A. and Xiu, D. [2021], Business news and business cycles, Technical report, National Bureau of Economic Research.

Caldara, D., Cavallo, M. and Iacoviello, M. [2019], 'Oil price elasticities and oil price fluctuations', *Journal of Monetary Economics* **103**, 1–20.

Caldara, D. and Iacoviello, M. [2022], 'Measuring geopolitical risk', *American Economic Review* **112**(4), 1194–1225.

Carriero, A., Clark, T. E. and Marcellino, M. [2015], 'Bayesian vars: specification choices and forecast accuracy', *Journal of Applied Econometrics* **30**(1), 46–73.

Carriero, A., Clark, T. E. and Marcellino, M. [2019], 'Large bayesian vector autoregressions with stochastic volatility and non-conjugate priors', *Journal of Econometrics* **212**(1), 137–154.

Carriero, A., Clark, T. E. and Marcellino, M. [2022], 'Nowcasting tail risk to economic activity at a weekly frequency', *Journal of Applied Econometrics* **37**(5), 843–866.

Carriero, A., Clark, T. E. and Marcellino, M. G. [2016], 'Large vector autoregressions with stochastic volatility and flexible priors'.

Carruth, A. A., Hooker, M. A. and Oswald, A. J. [1998], 'Unemployment equilibria and input prices: Theory and evidence from the united states', *Review of economics and Statistics* **80**(4), 621–628.

Chan, J. C. [2022], 'Large hybrid time-varying parameter vars', *Journal of Business & Economic Statistics* pp. 1–16.

Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y. and Dhillon, I. [2019], 'X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers', *arXiv preprint arXiv:1905.02331* .

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. [2018], 'Double/debiased machine learning for treatment and structural parameters'.

Christiano, L. J., Motto, R. and Rostagno, M. [2014], 'Risk shocks', *American Economic Review* **104**(1), 27–65.

Claeskens, G., Hjort, N. L. et al. [2008], 'Model selection and model averaging', *Cambridge Books* .

Clarida, R. H., Sarno, L., Taylor, M. P. and Valente, G. [2003], 'The out-of-sample success of term structure models as exchange rate predictors: a step beyond', *Journal of International Economics* **60**(1), 61–83.

Clark, T. E., Huber, F., Koop, G., Marcellino, M. and Pfarrhofer, M. [2022], 'Tail forecasting with multivariate bayesian additive regression trees', *International Economic Review* .

Clements, M. P. and Galvão, A. B. [2008], 'Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states', *Journal of Business & Economic Statistics* **26**(4), 546–554.

Clements, M. P. and Galvão, A. B. [2009], 'Forecasting us output growth using leading indicators: An appraisal using midas models', *Journal of Applied Econometrics* **24**(7), 1187–1206.

Cochrane, J. H. and Piazzesi, M. [2002], 'The fed and interest rates-a high-frequency identification', *American economic review* **92**(2), 90–95.

Cogley, T. and Sargent, T. J. [2005], 'Drifts and volatilities: monetary policies and outcomes in the post wwii us', *Review of Economic dynamics* **8**(2), 262–302.

Colladon, A. F. [2020], 'Forecasting election results by studying brand importance in online news', *International Journal of Forecasting* **36**(2), 414–427.

Colón-Ruiz, C. and Segura-Bedmar, I. [2020], 'Comparing deep learning architectures for sentiment analysis on drug reviews', *Journal of Biomedical Informatics* **110**, 103539.

Correa, R., Garud, K., Londono, J. M., Mislang, N. et al. [2017], 'Constructing a dictionary for financial stability', *IFDP notes. Board of Governors of the Federal Reserve System, Washington, DC* .

Davis, S. J. and Haltiwanger, J. [2001], 'Sectoral job creation and destruction responses to oil price changes', *Journal of monetary economics* **48**(3), 465–512.

De Smedt, T. and Daelemans, W. [2012], 'Pattern for python', *The Journal of Machine Learning Research* **13**(1), 2063–2067.

Degiannakis, S. and Filis, G. [2018], 'Forecasting oil prices: High-frequency financial data are indeed useful', *Energy Economics* **76**, 388–402.

Degiannakis, S. and Filis, G. [2022], 'Oil price volatility forecasts: What do investors need to know?', *Journal of International Money and Finance* **123**, 102594.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. [2018], 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805* .

d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q. and Vidal, M. [2020], 'Fquad: French question answering dataset', *arXiv preprint arXiv:2002.06071* .

Diederich, J., Kindermann, J., Leopold, E. and Paass, G. [2003], 'Authorship attribution with support vector machines', *Applied intelligence* **19**(1), 109–123.

Doan, T. [2013], 'Rats user's guide, version 8.2'.

Doan, T., Litterman, R. and Sims, C. [1984], 'Forecasting and conditional projection using realistic prior distributions', *Econometric reviews* **3**(1), 1–100.

Draper, D. [1995], 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society: Series B (Methodological)* **57**(1), 45–70.

Durbin, J. and Koopman, S. J. [2012], *Time series analysis by state space methods*, Vol. 38, OUP Oxford.

Edelstein, P. and Kilian, L. [2009], 'How sensitive are consumer expenditures to retail energy prices?', *Journal of Monetary Economics* **56**(6), 766–779.

Elder, J. and Serletis, A. [2010], 'Oil price uncertainty', *Journal of Money, Credit and Banking* **42**(6), 1137–1159.

Engelberg, J. E. and Parsons, C. A. [2011], 'The causal impact of media in financial markets', *The Journal of Finance* **66**(1), 67–97.

Evans, M. D. and Lewis, K. K. [1995], 'Do expected shifts in inflation affect estimates of the long-run fisher relation?', *The Journal of Finance* **50**(1), 225–253.

Fan, J., Ke, Y. and Wang, K. [2020], 'Factor-adjusted regularized model selection', *Journal of Econometrics* **216**(1), 71–85.

Fang, L. and Peress, J. [2009], 'Media coverage and the cross-section of stock returns', *The Journal of Finance* **64**(5), 2023–2052.

Favero, C. A., Hashem Pesaran, M. and Sharma, S. [1994], 'A duration model of irreversible oil investment: Theory and empirical evidence', *Journal of Applied Econometrics* **9**(S1), S95–S112.

Feldstein, M. and Eckstein, O. [1970], 'The fundamental determinants of the interest rates', *The Review of Economics and Statistics* pp. 363–375.

Fernandez, C., Ley, E. and Steel, M. F. [2001], 'Model uncertainty in cross-country growth regressions', *Journal of applied Econometrics* **16**(5), 563–576.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., Sargent, T. J. and Watson, M. W. [2007], 'Abcs (and ds) of understanding vars', *American Economic Review* **97**(3), 1021–1026.

Ferraro, D., Rogoff, K. and Rossi, B. [2015], 'Can oil prices forecast exchange rates? an empirical analysis of the relationship between commodity prices and exchange rates', *Journal of International Money and Finance* **54**, 116–141.

Forni, M. and Gambetti, L. [2014], 'Sufficient information in structural vars', *Journal of Monetary Economics* **66**, 124–136.

Foroni, C., Ghysels, E. and Marcellino, M. [2013], Mixed-frequency vector autoregressive models., *in* 'VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims', Vol. 32, Emerald Group Publishing Limited, pp. 247–272.

Foroni, C. and Marcellino, M. G. [2013], 'A survey of econometric methods for mixed-frequency data', *Available at SSRN 2268912* .

Foroni, C., Marcellino, M. and Schumacher, C. [2015], 'Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* pp. 57–82.

Frankel, J. A. [1986], 'Expectations and commodity price dynamics: The overshooting model', *American Journal of Agricultural Economics* **68**(2), 344–348.

Fung, G. P. C., Yu, J. X. and Lam, W. [2003], Stock prediction: Integrating text mining approach using real-time news, *in* '2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.', IEEE, pp. 395–402.

Garratt, A., Vahey, S. P. and Zhang, Y. [2019], 'Real-time forecast combinations for the oil price', *Journal of Applied Econometrics* **34**(3), 456–462.

Gentzkow, M. and Shapiro, J. M. [2010], 'What drives media slant? evidence from us daily newspapers', *Econometrica* **78**(1), 35–71.

Gentzkow, M., Shapiro, J. M. and Sinkinson, M. [2014], 'Competition and ideological diversity: Historical evidence from us newspapers', *American Economic Review* **104**(10), 3073–3114.

Ghysels, E. [2016], 'Macroeconomics and the reality of mixed frequency data', *Journal of Econometrics* **193**(2), 294–314.

Ghysels, E., Santa-Clara, P. and Valkanov, R. [2004], 'The midas touch: Mixed data sampling regression models'.

Ghysels, E., Santa-Clara, P. and Valkanov, R. [2005], 'There is a risk-return trade-off after all', *Journal of financial economics* **76**(3), 509–548.

Ghysels, E., Santa-Clara, P. and Valkanov, R. [2006], 'Predicting volatility: getting the most out of return data sampled at different frequencies', *Journal of Econometrics* **131**(1-2), 59–95.

Ghysels, E., Sinko, A. and Valkanov, R. [2007], 'Midas regressions: Further results and new directions', *Econometric reviews* **26**(1), 53–90.

Ghysels, E. and Wright, J. H. [2009], 'Forecasting professional forecasters', *Journal of Business & Economic Statistics* **27**(4), 504–516.

Giannone, D., Lenza, M. and Primiceri, G. E. [2015], 'Prior selection for vector autoregressions', *Review of Economics and Statistics* **97**(2), 436–451.

Giannone, D., Reichlin, L. and Small, D. [2008], 'Nowcasting: The real-time informational content of macroeconomic data', *Journal of Monetary Economics* **55**(4), 665–676.

Gifuni, L. [2021], 'Real oil price forecasting: Gains and pitfalls of text data', *Available at SSRN 3957549* .

Gneiting, T., Raftery, A. E., Westveld III, A. H. and Goldman, T. [2005], 'Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation', *Monthly Weather Review* **133**(5), 1098–1118.

Golub, S. S. [1983], 'Oil prices and exchange rates', *The Economic Journal* **93**(371), 576–593.

Gong, Y., Bu, R. and Chen, Q. [2022], 'What affects the relationship between oil prices and the us stock market? a mixed-data sampling copula approach', *Journal of Financial Econometrics* **20**(2), 253–277.

Granger, C. W. and Ramanathan, R. [1984], 'Improved methods of combining forecasts', *Journal of forecasting* **3**(2), 197–204.

Griliches, Z. [1967], 'Distributed lags: A survey', *Econometrica: journal of the Econometric Society* pp. 16–49.

Groseclose, T. and Milyo, J. [2005], 'A measure of media bias', *The quarterly journal of economics* **120**(4), 1191–1237.

Groth, S. S. and Muntermann, J. [2011], 'An intraday market risk management approach based on textual analysis', *Decision Support Systems* **50**(4), 680–691.

Guérin, P. and Marcellino, M. [2013], 'Markov-switching midas models', *Journal of Business & Economic Statistics* **31**(1), 45–56.

Hagenau, M., Liebmann, M. and Neumann, D. [2013], 'Automated news reading: Stock price prediction based on financial news using context-capturing features', *Decision Support Systems* **55**(3), 685–697.

Hamilton, J. D. [1983], 'Oil and the macroeconomy since world war ii', *Journal of political economy* **91**(2), 228–248.

Hamilton, J. D. [2019], 'Measuring global economic activity', *Journal of Applied Econometrics* .

Hansen, S., McMahon, M. and Prat, A. [2018], 'Transparency and deliberation within the fomc: a computational linguistics approach', *The Quarterly Journal of Economics* **133**(2), 801–870.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. [1999], 'Bayesian model averaging: a tutorial', *Statistical science* pp. 382–401.

Hotelling, H. [1931], 'The economics of exhaustible resources', *Journal of political Economy* **39**(2), 137–175.

Hutto, C. and Gilbert, E. [2014], Vader: A parsimonious rule-based model for sentiment analysis of social media text, *in* 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 8.

Inoue, A. and Kilian, L. [2013], 'Inference on impulse response functions in structural var models', *Journal of Econometrics* **177**(1), 1–13.

Joachims, T. [1996], A probabilistic analysis of the rocchio algorithm with tfidf for text categorization., Technical report, Carnegie-mellon univ pittsburgh pa dept of computer science.

Joachims, T. [1998], Text categorization with support vector machines: Learning with many relevant features, *in* 'European conference on machine learning', Springer, pp. 137–142.

Johnson, K., Jin, D. and Goldwasser, D. [2017], Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter, *in* 'Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', pp. 741–752.

Jones, C. M. and Kaul, G. [1996], 'Oil and the stock markets', *The journal of Finance* **51**(2), 463–491.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G. and Kapadia, S. [2020], 'Making text count: economic forecasting using newspaper text', *Journal of Applied Econometrics* .

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.-t. [2020], 'Dense passage retrieval for open-domain question answering', *arXiv preprint arXiv:2004.04906* .

Karrenbrock, J. D. et al. [1991], 'The behavior of retail gasoline prices: symmetric or not?', *Federal Reserve Bank of St. Louis Review* **73**(4), 19–29.

Kass, R. E. and Raftery, A. E. [1995], 'Bayes factors', *Journal of the american statistical association* **90**(430), 773–795.

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), 1053–69.

Kilian, L. and Murphy, D. P. [2012], 'Why agnostic sign restrictions are not enough: understanding the dynamics of oil market var models', *Journal of the European Economic Association* **10**(5), 1166–1188.

Kilian, L. and Murphy, D. P. [2014], 'The role of inventories and speculative trading in the global market for crude oil', *Journal of Applied econometrics* **29**(3), 454–478.

Kilian, L. and Taylor, M. P. [2003], 'Why is it so difficult to beat the random walk forecast of exchange rates?', *Journal of International Economics* **60**(1), 85–107.

Knetsch, T. A. [2007], 'Forecasting the price of crude oil via convenience yield predictions', *Journal of Forecasting* **26**(7), 527–549.

Kollias, C., Kyrtsou, C. and Papadamou, S. [2013], 'The effects of terrorism and war on the oil price–stock index relationship', *Energy Economics* **40**, 743–752.

Koop, G. and Korobilis, D. [2010], *Bayesian multivariate time series methods for empirical macroeconomics*, Now Publishers Inc.

Koop, G. and Korobilis, D. [2012], 'Forecasting inflation using dynamic model averaging', *International Economic Review* **53**(3), 867–886.

Koop, G. and Korobilis, D. [2013], 'Large time-varying parameter vars', *Journal of Econometrics* **177**(2), 185–198.

Laffont, J.-J. and Garcia, R. [1977], 'Disequilibrium econometrics for business loans', *Econometrica: Journal of the Econometric Society* pp. 1187–1204.

Lamperti, F., Roventini, A. and Sani, A. [2018], 'Agent-based model calibration using machine learning surrogates', *Journal of Economic Dynamics and Control* **90**, 366–389.

Lanne, M. and Lütkepohl, H. [2008], 'Identifying monetary policy shocks via changes in volatility', *Journal of Money, Credit and Banking* **40**(6), 1131–1149.

Lanne, M., Lütkepohl, H. and Maciejowska, K. [2010], 'Structural vector autoregressions with markov switching', *Journal of Economic Dynamics and Control* **34**(2), 121–131.

Larsen, V. H. and Thorsrud, L. A. [2019], 'The value of news for economic developments', *Journal of Econometrics* **210**(1), 203–218.

Leamer, E. E. [1978], *Specification searches: Ad hoc inference with nonexperimental data*, Vol. 53, Wiley New York.

Lenza, M. and Primiceri, G. E. [2020], How to estimate a var after march 2020, Technical report, National Bureau of Economic Research.

Li, W., Shao, W., Ji, S. and Cambria, E. [2022], 'Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis', *Neurocomputing* **467**, 73–82.

Li, X., Shang, W. and Wang, S. [2019], 'Text-based crude oil price forecasting: A deep learning approach', *International Journal of Forecasting* **35**(4), 1548–1560.

Li, Y., Jiang, S., Li, X. and Wang, S. [2021], 'The role of news sentiment in oil futures returns and volatility forecasting: Data-decomposition based deep learning approach', *Energy Economics* **95**, 105140.

Lippi, F. and Nobili, A. [2012], 'Oil and the macroeconomy: a quantitative structural analysis', *Journal of the European Economic Association* **10**(5), 1059–1083.

Litterman, R. B. [1986], 'Forecasting with bayesian vector autoregressions—five years of experience', *Journal of Business & Economic Statistics* **4**(1), 25–38.

Liu, Y. [2019], 'Fine-tune bert for extractive summarization', *arXiv preprint arXiv:1903.10318* .

Liu, Y. and Lapata, M. [2019], 'Text summarization with pretrained encoders', *arXiv preprint arXiv:1908.08345* .

Loughran, T. and McDonald, B. [2011], 'When is a liability not a liability? textual analysis, dictionaries, and 10-ks', *The Journal of finance* **66**(1), 35–65.

Ma, Y.-r., Ji, Q. and Pan, J. [2019], 'Oil financialization and volatility forecast: Evidence from multidimensional predictors', *Journal of Forecasting* **38**(6), 564–581.

MacDonald, R. and Taylor, M. P. [1994], 'The monetary model of the exchange rate: long-run relationships, short-run dynamics and how to beat a random walk', *Journal of international money and finance* **13**(3), 276–290.

Madigan, D. and Raftery, A. E. [1994], 'Model selection and accounting for model uncertainty in graphical models using occam's window', *Journal of the American Statistical Association* **89**(428), 1535–1546.

Martens, M. [2001], 'Forecasting daily exchange rate volatility using intraday returns', *Journal of International Money and Finance* **20**(1), 1–23.

McAllester, D. A. [1999], Pac-bayesian model averaging, *in* 'Proceedings of the twelfth annual conference on Computational learning theory', pp. 164–170.

McCracken, M. W. and Ng, S. [2016], 'Fred-md: A monthly database for macroeconomic research', *Journal of Business & Economic Statistics* **34**(4), 574–589.

McGurk, Z., Nowak, A. and Hall, J. C. [2020], 'Stock returns and investor sentiment: textual analysis and social media', *Journal of Economics and Finance* **44**(3), 458–485.

Meese, R. A. and Rogoff, K. [1983], 'Empirical exchange rate models of the seventies: Do they fit out of sample?', *Journal of international economics* **14**(1-2), 3–24.

Miao, H., Ramchander, S., Wang, T. and Yang, D. [2017], 'Influential factors in crude oil price forecasting', *Energy Economics* **68**, 77–88.

Min, C.-k. and Zellner, A. [1993], 'Bayesian and non-bayesian methods for combining models and forecasts with applications to forecasting international growth rates', *Journal of Econometrics* **56**(1-2), 89–118.

Ming, F., Wong, F., Liu, Z. and Chiang, M. [2014], Stock market prediction from wsj: text mining via sparse matrix factorization, *in* '2014 IEEE International Conference on Data Mining', IEEE, pp. 430–439.

Miranda-Agrippino, S. and Ricco, G. [2021], 'The transmission of monetary policy shocks', *American Economic Journal: Macroeconomics* **13**(3), 74–107.

Mittermayer, M.-A. [2004], Forecasting intraday stock price trends with text mining techniques, *in* '37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the', IEEE, pp. 10–pp.

Modigliani, F. and Sutch, R. [1966], 'Innovations in interest rate policy', *The American Economic Review* **56**(1/2), 178–197.

Mork, K. A. [1989], 'Oil and the macroeconomy when prices go up and down: an extension of hamilton's results', *Journal of political Economy* **97**(3), 740–744.

Mosteller, F. and Wallace, D. L. [1963], 'Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers', *Journal of the American Statistical Association* **58**(302), 275–309.

Mugaloglu, E., Polat, A. Y., Tekin, H. and Dogan, A. [2021], 'Oil price shocks during the covid-19 pandemic: evidence from united kingdom energy stocks', *Energy Research Letters* **2**(1).

Mumtaz, H. and Zanetti, F. [2013], 'The impact of the volatility of monetary policy shocks', *Journal of Money, Credit and Banking* **45**(4), 535–558.

Nguyen, D. K. and Walther, T. [2020], 'Modeling and forecasting commodity market volatility with long-term economic and financial variables', *Journal of Forecasting* **39**(2), 126–142.

Nielsen, F. Å. [2011], 'A new anew: Evaluation of a word list for sentiment analysis in microblogs', *arXiv preprint arXiv:1103.2903* .

Nyman, R., Kapadia, S. and Tuckett, D. [2021], 'News and narratives in financial systems: exploiting big data for systemic risk assessment', *Journal of Economic Dynamics and Control* **127**, 104119.

Pang, B., Lee, L. and Vaithyanathan, S. [2002], 'Thumbs up? sentiment classification using machine learning techniques', *arXiv preprint cs/0205070* .

Peersman, G. [2005], 'What caused the early millennium slowdown? evidence based on vector autoregressions', *Journal of Applied Econometrics* **20**(2), 185–207.

Peress, J. [2014], 'The media and the diffusion of information in financial markets: Evidence from newspaper strikes', *the Journal of Finance* **69**(5), 2007–2043.

Poria, S., Cambria, E., Howard, N., Huang, G.-B. and Hussain, A. [2016], 'Fusing audio, visual and textual clues for sentiment analysis from multimodal content', *Neurocomputing* **174**, 50–59.

Porter, M. F. [1980], 'An algorithm for suffix stripping', *Program* .

Pota, M., Ventura, M., Catelli, R. and Esposito, M. [2021], 'An effective bert-based pipeline for twitter sentiment analysis: a case study in italian', *Sensors* **21**(1), 133.

Price, S. M., Doran, J. S., Peterson, D. R. and Bliss, B. A. [2012], 'Earnings conference calls and stock returns: The incremental informativeness of textual tone', *Journal of Banking & Finance* **36**(4), 992–1011.

Primiceri, G. E. [2005], 'Time varying structural vector autoregressions and monetary policy', *The Review of Economic Studies* **72**(3), 821–852.

Rachlin, G., Last, M., Alberg, D. and Kandel, A. [2007], Admiral: A data mining based financial trading system, *in* '2007 ieee symposium on computational intelligence and data mining', IEEE, pp. 720–725.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. [2005], 'Using bayesian model averaging to calibrate forecast ensembles', *Monthly weather review* **133**(5), 1155–1174.

Raftery, A. E., Madigan, D. and Hoeting, J. A. [1997], 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**(437), 179–191.

Renault, T. [2017], 'Intraday online investor sentiment and return patterns in the us stock market', *Journal of Banking & Finance* **84**, 25–40.

Rigobon, R. [2003], 'Identification through heteroskedasticity', *Review of Economics and Statistics* **85**(4), 777–792.

Rigobon, R. and Sack, B. [2003], 'Measuring the reaction of monetary policy to the stock market', *The quarterly journal of Economics* **118**(2), 639–669.

Rodriguez, A. and Puggioni, G. [2010], 'Mixed frequency models: Bayesian approaches to estimation and prediction', *International Journal of Forecasting* **26**(2), 293–311.

Rossi, B. [2014], *Density forecasts in economics and policymaking*, CREI, Centre de Recerca en Economia Internacional.

Rubio-Ramirez, J. F., Waggoner, D. F. and Zha, T. [2010], 'Structural vector autoregressions: Theory of identification and algorithms for inference', *The Review of Economic Studies* **77**(2), 665–696.

Sala-i Martin, X. X. [1997], I just ran four million regressions, Technical report, National Bureau of Economic Research.

Salton, G. and Buckley, C. [1988], 'Term-weighting approaches in automatic text retrieval', *Information processing & management* **24**(5), 513–523.

Sari, R., Soytas, U. and Hacihasanoglu, E. [2011], 'Do global risk perceptions influence world oil prices?', *Energy Economics* **33**(3), 515–524.

Schorfheide, F. and Song, D. [2015], 'Real-time forecasting with a mixed-frequency var', *Journal of Business & Economic Statistics* **33**(3), 366–380.

Schorfheide, F., Song, D. and Yaron, A. [2018], 'Identifying long-run risks: A bayesian mixed-frequency approach', *Econometrica* **86**(2), 617–654.

Schumaker, R. P. and Chen, H. [2009], 'Textual analysis of stock market prediction using breaking financial news: The azfin text system', *ACM Transactions on Information Systems (TOIS)* **27**(2), 1–19.

Schumaker, R. P., Zhang, Y., Huang, C.-N. and Chen, H. [2012], 'Evaluating sentiment in financial news articles', *Decision Support Systems* **53**(3), 458–464.

Sévi, B. [2014], 'Forecasting the volatility of crude oil futures using intraday data', *European Journal of Operational Research* **235**(3), 643–659.

Shapiro, A. H., Sudhof, M. and Wilson, D. J. [2020], 'Measuring news sentiment', *Journal of econometrics* .

Shapiro, A. H., Sudhof, M. and Wilson, D. J. [2022], 'Measuring news sentiment', *Journal of econometrics* **228**(2), 221–243.

Shapiro, A. H. and Wilson, D. J. [2022], 'Taking the fed at its word: A new approach to estimating central bank objectives using text analysis', *The Review of Economic Studies* **89**(5), 2768–2805.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D. and Liu, H. [2020], 'Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media', *Big data* **8**(3), 171–188.

Sims, C. A. and Zha, T. [2006], 'Were there regime switches in us monetary policy?', *American Economic Review* **96**(1), 54–81.

Singh, M., Jakhar, A. K. and Pandey, S. [2021], 'Sentiment analysis on the impact of coronavirus in social life using the bert model', *Social Network Analysis and Mining* **11**(1), 1–11.

Smith, J. E. and McCardle, K. F. [1998], 'Valuing oil properties: integrating option pricing and decision analysis approaches', *Operations Research* **46**(2), 198–217.

Stock, J. H. and Watson, M. [2003], 'Forecasting output and inflation: The role of asset prices', *Journal of Economic Literature* **41**(3), 788–829.

Strachan, R. W. and Van Dijk, H. K. [2013], 'Evidence on features of a dsge business cycle model from bayesian model averaging', *International Economic Review* **54**(1), 385–402.

Suardi, S., Rasel, A. R. and Liu, B. [2022], 'On the predictive power of tweet sentiments and attention on bitcoin', *International Review of Economics & Finance* **79**, 289–301.

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W. and Jiang, P. [2019], Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, *in* 'Proceedings of the 28th ACM international conference on information and knowledge management', pp. 1441–1450.

Taylor, M. P. [1995], 'The economics of exchange rates', *Journal of Economic literature* **33**(1), 13–47.

Tetlock, P. C. [2007], 'Giving content to investor sentiment: The role of media in the stock market', *The Journal of finance* **62**(3), 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M. and Macskassy, S. [2008], 'More than words: Quantifying language to measure firms' fundamentals', *The journal of finance* **63**(3), 1437–1467.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. [2010], 'Sentiment strength detection in short informal text', *Journal of the American society for information science and technology* **61**(12), 2544–2558.

Townsend, R. M. [1983], 'Forecasting the forecasts of others', *Journal of Political Economy* **91**(4), 546–588.

Vafa, K., Naidu, S. and Blei, D. M. [2020], 'Text-based ideal points', *arXiv preprint arXiv:2005.04232* .

Viallefont, V., Raftery, A. E. and Richardson, S. [2001], 'Variable selection and bayesian model averaging in case-control studies', *Statistics in medicine* **20**(21), 3215–3230.

Vrugt, J. A., Diks, C. G. and Clark, M. P. [2008], 'Ensemble bayesian model averaging using markov chain monte carlo sampling', *Environmental fluid mechanics* **8**(5-6), 579–595.

Vrugt, J. A. and Robinson, B. A. [2007], 'Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and bayesian model averaging', *Water Resources Research* **43**(1).

Weale, M. and Wieladek, T. [2016], 'What are the macroeconomic effects of asset purchases?', *Journal of monetary Economics* **79**, 81–93.

Wright, J. H. [2008], 'Bayesian model averaging and exchange rate forecasts', *Journal of Econometrics* **146**(2), 329–341.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K. and Zhang, J. [1998], Daily stock market forecast from textual web data, *in* 'SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)', Vol. 3, IEEE, pp. 2720–2725.

Xing, F. Z., Cambria, E. and Welsch, R. E. [2018], 'Natural language based financial forecasting: a survey', *Artificial Intelligence Review* **50**(1), 49–73.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M. and Lin, J. [2019], 'End-to-end open-domain question answering with bertserini', *arXiv preprint arXiv:1902.01718* .

Yin, G. and Yuan, Y. [2009], 'Bayesian model averaging continual reassessment method in phase i clinical trials', *Journal of the American Statistical Association* **104**(487), 954–968.

Zhai, Y., Hsu, A. and Halgamuge, S. K. [2007], Combining news and technical indicators in daily stock price trends prediction, *in* 'International symposium on neural networks', Springer, pp. 1087–1096.

Zhang, X., Wei, F. and Zhou, M. [2019], 'Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization', *arXiv preprint arXiv:1905.06566* .

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. and Gao, J. [2020], Unified vision-language pre-training for image captioning and vqa, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 13041–13049.