Forecasting oil prices: alternative approaches

Rennan Kertlly de Medeiros[*]
Cássio da Nóbrega Besarria [†]
Diego Pitta de Jesus [‡]
Vinícius Phillipe de Albuquerquemello [§]

Abstract

This paper proposes alternative methodologies for oil price forecasting using data sets of mixed frequencies and by constructing a sentiment indicator from oil market reports. We used the root mean square error (RMSE) to evaluate the forecasting accuracy of the econometric models. Our findings indicate that, compared to other econometric models, the mixed data sampling (MIDAS) model, with high-frequency financial indicators and the sentiment index as explanatory variables, performs better for forecasting crude oil prices.

Keywords: Forecasting oil prices; Time Series; MIDAS model; Sentiment indicator.
JEL Code: C53; C58; Q02.

Resumo

Este artigo propõe metodologias alternativas para a previsão de preços do petróleo usando um conjunto de dados de frequências mistas e a construção de um indicador de sentimento a partir de relatórios do mercado de petróleo. Utilizamos a raiz do erro quadrático médio (RMSE) para avaliar a precisão da previsão dos modelos econométricos. Nossas descobertas indicam que, comparado a outros modelos econométricos, o modelo de amostragem de dados mistos (MIDAS), com indicadores financeiros de alta frequência e o índice de sentimentos como variáveis explicativas, tem melhor desempenho na previsão de preços do petróleo.

Palavras-chave: Previsão de preços do petróleo; Séries temporais; Modelo MIDAS; Indicador de sentimento.

Área 8: Microeconomia, Métodos Quantitativos e Finanças

[*] PhD student of the Graduate Program in Economics, UFPB. E-mail:rennanmedeiros22@hotmail.com.
[†] Professor of the Graduate Program in Economics, UFPB. E-mail:cassiodanobrega@yahoo.com.br.
[‡] PhD student of the Graduate Program in Economics, UFPB. diegopitta13@hotmail.com.
[§] Professor of DCSA, UFPB - Campus III. PhD student PIMES, UFPE. vinicius.phillipe@hotmail.com.

1 Introduction

Crude oil has a high representation in the global economy. According to Eika and Magnussen (2000) and Kilian (2009), the oil price trajectory impacts an economy at different levels, from family budgets to corporate earnings. In this sense, Kilian and Park (2009) state that oil demand and oil supply shocks are responsible for 22% of long-term variation in the real returns of U.S. stocks.

Moreover, Salisu, Raheem and Ndako (2019) demonstrate its causality on stock market returns, and Hamilton (1996) points out its relation to macroeconomic variables. These effects occur by many channels: Baffes (2007) and Carpio (2019) mention the impact of oil prices on transportation costs; Casamassima, Fiorello and Martino (2009) analyze the impact of oil prices on inflation; and Cuñado and Pérez de Gracia (2003) investigate the relation of oil price volatility to exchange rates.

However, forecasting oil prices is a difficult task to accomplish once the occurrence of exogenous shocks, which reduce or increase its value and have no unanimous explanation by literature, are common. For example, in 2014, oil prices fell below US$50.00, and according to Baumeister and Kilian (2014a), this was due to the recovery of oil production in Libya and the resilience of production in Iraq; however, Kilian and Murphy (2014) pointed out that it was more related to the increase in shale oil production by the United States, a substitute.

The literature on oil price forecasting is thus in constant development, always attempting to explain the sources of fluctuations and to make more accurate predictions. The following are some examples of papers in this branch of the literature: Coppola (2008), who tested autoregressive vectors based on future prices to explain spot prices; Agnolucci (2009), who used GARCH models to forecast oil price implied volatility; Hamilton (2009), who included an instrumental variable to account for the effect of cartel practices on oil prices; Chai et al. (2018), who applied hybrid models which mix genetic algorithms with econometric methods, and relies on real-time data for oil price forecasting; Miao et al. (2017) who employed penalty methods (LASSO); Movagharnejad et al. (2011) and Chiroma, Abdulkareem and Herawan (2015), who implemented different methods of neural networks; Albuquerquemello et al. (2018), who examined numerous econometric models with multiple specifications for the explanatory variables; Li, Shang and Wang (2019), who used deep learning with the inclusion of time series generated from textual analysis of financial reports; Olofin et al. (2020), who forecasted oil prices based on shale oil production, following the econometric method of Westerlund and Narayan (2015); and Salisu, Swaray and Oloko (2019), who tested a multi-factor model for forecasting the channel of oil prices on the U.S. stock market.

Against of the econometric methods mentioned above, we expand the discussion and add relevant contributions to increase the accuracy of oil price forecasting. In the previously mentioned oil studies, most of the methods used were estimated with low-frequency data and based on time series made available by official bodies with some degree of delay. To fill this gap, we propose mixed-data sampling (MIDAS) to allow the inclusion of high-frequency variables to help predict the oil price. In addition, the key innovation of this study is the inclusion of the sentiment predictor variable in the MIDAS econometric model, built using text regression techniques.

Our results are as follows: (i) the series of oil prices are characterized by a non-linear quadratic pattern; (ii) for in-sample prediction, the model ADL-MIDAS$_{Umidas}(sent)$ presents the best accuracy among models; (iii) the SETAR model has better accuracy only when compared with univariate models (linear and nonlinear); (iv) for out-of-sample forecasting, we point out MIDAS$_{Umidas}(sent)$ as the best predictor; (v) the inclusion of a variable that captures the tone (sentiment) of oil market reports augments the forecast capacity of every model. (vi) the use of mixed-frequency data tends to augment the precision of oil price forecasts.

A possible explanation for the enhancements brought by the models suggested here is that financial data (high frequency) are accessible in real time and incorporate forward-looking information, while the conventional macro-economic variables used in low-frequency designs are prone to gaps, constantly requiring corrections. In addition, the sentiment index reflects the mood of public authorities and U.S. governmental analysts for the oil market.

Our findings are unique as they contribute to reducing asymmetric information in a) the financial market, thereby facilitating the formation of agents' expectations, and b) oil price forecasting. We specifically regard its importance in uncertain scenarios when governmental actors perceptions about the future must be taken into account.

Apart from this introduction, this paper is structured in four additional sections. The second section describes the empirical methodology and data used in the research. The results are presented in the third section, and the fourth section is dedicated to a robustness analysis. Finally, the fifth section contains concluding remarks.

2 Data and Methodology

Since our objective is to test the inclusion of high-frequency data and the sentiment index of the oil market on monthly oil price forecasts' accuracy, we tested models based on low-frequency data and those based on mixed frequencies. Then, we compared the results in-sample and out-of-sample. Subsection 2.1 describes the database used and its details, such as the frequency of each time series, the treatment of the data, and the variables used in each model. Thereafter, Subsection 2.2 presents the econometric models that we tested, which are based on low-frequency data, and Subsection 2.3 describes the MIDAS models that allow for the inclusion of high-frequency data. The construction of the sentiment index is then explained in Subsection 2.4, and Subsection 2.5 presents our methods for comparison of forecasting accuracy between models.

2.1 Database and models

The database is composed of low-frequency variables (monthly observations) and high-frequency variables (daily observations). The period of the observations is 1995–2017, which is the maximum compatibility. The variables' descriptions, frequencies and sources are presented in Table 1, and Appendix B presents the descriptive statistics of the variables.

Table 1 – Summary of the variables used

Variable	**Frequency**	**Source**
Oil spot price	Monthly	Index Mundi
Oil futures price	Monthly	Investing website
World oil production	Monthly	Energy Information Administration (EIA)
World industrial production index	Monthly	Central Bank of St. Louis (FRED)
OECD oil stock	Monthly	Organization for Economic Cooperation and Development (OECD)
OPEC oil production	Monthly	EIA
Non-OPEC oil production	Monthly	EIA
Oil market sentiment index	Monthly	Constructed based on EIA reports
Constant rate of the U.S. treasure bonds to maturity of 10 years (DGS10)	Daily	FRED
Volatility index of S&P500 (VIX)	Daily	FRED
Inter-bank interest rate (LIBOR 3-month)	Daily	FRED
Dollar index weighted by the trade of the major currencies (DTWEXM)	Daily	FRED
NYSE composition ratios (NYA)	Daily	Yahoo Finance
Oil and Gas NYSE ARCA (XOI)	Daily	Yahoo Finance
Gasoline	Daily	EIA
Oil futures price (RCLC)	Daily	EIA

Source: authors' elaboration.

2.1.1 Treatment of the variables

As indicated in Ratkowsky and Giles (1990) and Enders (2008), for models that rely on low-frequency observations, the analysis of time series begins with their treatment. For mixed-frequency models, there is a consensus in the literature that no previous treatment of the variables is required, subsection 2.3 presents further details. The treatment of low-frequency variables involves the following steps: (i) detecting of seasonality and removing it if necessary, (ii) detecting the unit root and removing the stochastic trend if necessary, (iii) detecting structural breaks and using regime switching models if necessary, (iv) detecting nonlinear relationships, (v) implementing optimal lag selection with information criteria and (vi) performing a co-integration analysis and performing an error correction for the long-term relationship if necessary. Only then can we proceed to model estimation, analyses of the residuals and comparison of the forecasting accuracy.

The first step is necessary for removing repetitive effects in certain periods of time that may affect the quality of prediction. The second step is important to ensure that the series are stationary and can be predicted. The third step is important to check whether the observations derive from the same data generator process. The fourth step is important to adjust models to capture the full dynamics of the time series. The fifth step is important to adjust models to incorporate the effects of exogenous shocks that are not diluted over time. Finally, the sixth step is important to check whether series co-move over time and carry a spurious relation.

The detection of seasonality was done through an X-13 Arima test. To investigate whether the series are stationary, we used the following unit root tests: the augmented Dickey-Fuller (ADF) test and the Phillips-Perron (PP) test. Moreover, nonlinearities were detected with a Keenan test, a McLeod-Li test and a Tsay test. To check for the presence of structural breaks, we applied the Zivot-Andrews test, and we made optimal lag choices with the Akaike information criterion for finite samples (AICc). Furthermore, the co-integration analysis was performed using the Johansen test, and after the estimation of the models, we turned to the diagnosis of models' residuals for the presence of autocorrelation and heteroscedasticity with a multivariate portmanteau test and Engle's ARCH test respectively.

As listed in Table 2, 10 models based on the treated low-frequency data were tested: random walk, AR (p), MA (q), ARIMA (p, d, q), three VARs, SETAR, LSTAR and ESTAR. With the exception of the non-linear models (SETAR, LSTAR and STAR), the rest were estimated taking into account the existence of the constant regression to capture fluctuations through the use of intercepts, as shown by Hotelling (1931). The random walk models, ARMA class and VARs are consolidated in the oil price forecasting literature, the first being the most fragile and generally used as a comparative basis for other more sophisticated models. We used the autoregressive vector model without error correction because no long-term relationship existed between the variables, according to the Johansen co-integration test. In addition, we justify the use of the SETAR, LSTAR and ESTAR models because they accommodate structural breaks, a strong characteristic of the oil price series, signaled by the Zivot-Andrews test.

The first four models—random walk, AR, MA and ARIMA—are univariate and simple, as they only consider the variable of interest as endogenous. This approach via the auto-regressive integrated model of moving averages is quite consolidated and disseminated in forecasting literature, either for the sake of comparison between models, singly or associated with the models of the ARCH family to analyze price volatility, as demonstrated in Mohammadi and Su (2010).

In turn, we have the VARs as multivariate models. First, VAR A was based on the idea proposed by Reichsfeld and Roache (2011), in which the prices of futures contracts can be used to predict the spot price.[1] Second, VAR B consists of variables linked to the oil market coupled with the world industrial production, as in Hamilton (2009). Third, VAR C consists of the VAR B variables with oil production broken down between OPEC members and non-members, based on the categorization used in Kilian and Murphy (2014).

The last three models' specifications are typically applied to time series as an extension of auto-regressive models. Albuquerquemello et al. (2018) point out that among the multivariate and univariate models tested, the SETAR had the lowest RMSE (i.e., the best accuracy). However, these authors tested only one type of regime transition threshold. In this study, we added two models of smooth transition, namely ESTAR and LSTAR.

[1] We verified a bicausality between spot and future prices of oil through the Granger test.

Table 2 – Specification of estimated models

Model	Variables
Random walk	Oil spot price
AR(p)	Oil spot price
MA(q)	Oil spot price
ARIMA(p,d,q)	Oil spot price
VAR A	Oil spot price and Oil futures price
VAR B	VAR A variables + World industrial production + World oil production + OCDE oil stock
VAR C	VAR B variables + OPEC oil production + Non-OPEC oil production
SETAR	Oil spot price
LSTAR	Oil spot price
ESTAR	Oil spot price
MIDAS	Oil spot prices + Oil market sentiment index + DGS10 + VIX + LIBOR 3-month + DTWEXM + NYA + XOI + Gasoline + RCLC

Source: authors' elaboration.

The studies by Ghysels and Wright (2009), Armesto et al. (2010), Andreou, Ghysels and Kourtellos (2013) and Pan et al. (2018) indicate the ability of MIDAS regression to improve the performance prediction of quarterly macroeconomic variables on the basis of monthly variables, when compared with traditional univariate models (benchmark). Another advantage of this method is improvement in the prediction accuracy of monthly financial variables through the use of data with mixed frequencies (daily or weekly basis).

2.2 VAR, threshold and smooth transition

According to Sims (1980), in the presence of concurrency among a group of variables, these variables are treated in a homogeneous way; that is, all the variables are considered endogenous to the model. A multivariate analysis of interrelated time series violates the principle of standard minimum hands square ordinaries (OLS); using algebraic techniques to transform equations in matrix notation, called a vector autoregressive (VAR) as shown in Equation 1, is thus essential.

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \epsilon_t \tag{1}$$

where B is the coefficient matrix over time; Γ_0 is the intercept matrix, Γ_1 is the parameter matrix of lagged variables; and ϵ_t represents the error terms, assumed to be white noise. Multiplying Equation 1 by the inverse of B, we obtain

$$B^{-1}Bx_t = B^{-1}\Gamma_0 + B^{-1}\Gamma_1 x_{t-1} + B^{-1}\epsilon_t \tag{2}$$

The Equation 2 results in VAR estimation:

$$X_t = A_0 + A_i X_{t-i} + \epsilon_t \tag{3}$$

where X_t is the matrix containing the set of defined variables, A_0 is the matrix of constants, A_i is the model parameter matrix, and ϵ_t is the vector of error terms.

For identification and estimation of VAR, we followed Enders (2008) and Lütkepohl (2005), defining the following steps: a) obtaining the order of integration of each variable by a test of unit root; b) determining the optimal number of lags from information criteria; c) identifying the co-integration relation between variables through a Johansen (1988) test; d) evaluating auto-correlation, heteroscedasticity and stability conditions; e) making predictions in and out of the sample and f) analyzing the performance measures of the forecasts.

Regarding smooth transition models, the STAR model can be described as a nonlinear autoregressive model in which the parameters are determined by an array of explanatory variables incorporated into a transition space. The STAR model was proposed by Granger and Teräsvirta (1993) and Teräsvirta (1994), and its variants are LSTAR (logistic STAR) and ESTAR (exponential STAR).

According to Enders (2008), the auto-regressive model enables smooth transition the transition parameter if switch slowly. The following is a general definition is

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_1 y_{t-1} f(y_{t-1}) + \epsilon_t \tag{4}$$

where $f()$ is a continuous function of smoothing, and the auto-regressive coefficient $(\alpha_1 + \beta_1)$ represents the change in the values y_{t-1}.

The logistics version of the STAR model generalizes the standard auto-regressive model, where the auto-regressive coefficient is a logistic function:

$$y_t = \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \theta[\beta_0 + \beta_1 y_{t-1} + ... + \beta_p y_{t-p}] + \epsilon_t \tag{5}$$

where $\theta = [1 + \exp\left(-\gamma\left(y_{t-1} - c\right)\right)]^{-1}$ and γ is the smoothing parameter

In turn, the exponential form of the smooth transition model can be demonstrated by

$$\theta = 1 - \exp\left[-\gamma\left(y_{t-1} - c\right)^2\right] \quad \gamma > 0 \tag{6}$$

where θ contains a quadratic term, so that the model coefficients are symmetrical around $y_{t-1} = c$.

Finally, following Tsay (1989), we estimated the threshold model in three stages. First, we determined the level of AR **(p)**; second, we selected the lag parameter, **d**; third, we established the AR level for each regime by setting the value of the threshold in the direction of the values of p and d obtained.[2]

2.3 MIDAS

In the search for models that consider heterogeneous frequencies in a single regression, Ghysels, Santa-Clara and Valkanov (2004) present a category of models designated MIDAS. The MIDAS method allows the dependent variable to be explained by variables of diverse frequency (mixed). Some studies, such as those by Ghysels and Wright (2009), Clements and Galvão (2009), Baumeister, Guérin and Kilian (2015) and Zhang and Wang (2019), indicate the importance of MIDAS use in predicting macroeconomic and financial variables.

The sampling procedure for different frequencies causes high parameterization in MIDAS regression. For this, the responses of the independent variables are modeled from lagged polynomials to correct problems related to the selection of lag order. In this study, we estimate a multivariate MIDAS model in its simple form and in an auto-regressive (ADL) version. The simple version is based in Foroni and Marcellino (2013) and is expressed by Equation 7:

$$y_{t_m} = \beta_0 + \sum_{i=1}^{K} \sum_{j=1}^{L} b_{ij}\left(L_{m_i}; \theta\right) x_{t_m + w - h_m}^{(m_i)} + \varepsilon_{t_m} \tag{7}$$

where β_0 is a constant, y_{t_m} is the low-frequency variable (monthly spot price of oil), and $x_{t_m + w - h_m}$ indicates the high-frequency lagged variables.

The second estimation using MIDAS regression follows Clements and Galvão (2008). In this study, we estimate a version of the MIDAS autoregressive model (ADL) to eight high-frequency explanatory variables. The form of generalized estimation is described in Equation 8:

$$y_t = \beta_0 + \lambda y_{t-h} + \sum_{i=1}^{K} \beta_i B\left(L^{\frac{1}{m}}; \theta_i\right)\left(1 - \lambda L^h\right) x_{i,t-h}^{(m)} + \varepsilon_t \tag{8}$$

The weights $b_{ij}\left(L_{m_i}; \theta\right)$ and $B\left(L^{\frac{1}{m}}; \theta_i\right)$ of Equations 7 and 8 were employed in polynomial, lagged and unrestricted Almon (UMIDAS)[3].

[2] For further details, see Tong and Lim (1980).

[3] See Ghysels, Sinko and Valkanov (2007) and Foroni, Marcellino and Schumacher (2015).

2.4 Analysis of textual sentiment and construction of the tone of the reports

Starting from monthly reports on crude oil between January 1995 and December 2017, issued by Energy Information Administration (EIA), and using the algorithm of Jockers (2017), we extracted the textual sentiment.

According to Jockers (2017), the computational routine performs the reading of texts and classifies words by their positive and negative cognitive aspects. These features are predefined by a dictionary; here we use Deeney et al. (2015). With the count of positive and negative words in the reports, we built an index that represents the tone or mood of the report, as described by Equation 9:

$$sent_t = \frac{\sum PositiveWords - \sum NegativeWords}{\sum PositiveWords + \sum NeutralWords + \sum NegativeWords} \tag{9}$$

The index $sent_t \in [-1, 1]$. Therefore, the oil report tone is considered: positive if it is greater than 0 and less than or equal to 1, neutral when it is equal to 0 and negative if it is less than 0.

2.5 Forecast performance evaluation

The evaluation of the forecasting performance took place in two stages: i) calculation of root mean square error (RMSE) for in-sample and out-of-sample forecasts and ii) pairing and comparison of RMSE among models, following the methodology proposed by Diebold and Mariano (1995).

The Diebold-Mariano test begins with the calculation of the differential loss between two prediction methods, uniformly weighting the loss functions of the error terms. The RMSE measure and Diebold-Mariano test are obtained from Equations 10 and 11 respectively:

$$RMSE_h = \sqrt{\frac{1}{T-h}\sum_{t=1}^{T-h}(\hat{y}_{t+h} - y_{t+h})^2} \tag{10}$$

$$DM_h = \frac{\bar{d}}{\sqrt{\frac{\hat{w}_d}{T-h}}} \tag{11}$$

where $\bar{d} = \frac{1}{T-h}\sum_{t=1}^{T-h}[\hat{u}^B_{t+h})^2 - (\hat{u}^A_t + h)^2]$, $\hat{u}^A_{t+h}, \hat{u}^B_{t+4}$ are the oil price forecast errors for h-periods forward, and $\hat{w}_d$ is the covariance matrix $\bar{d}$ considering serial auto correlation.

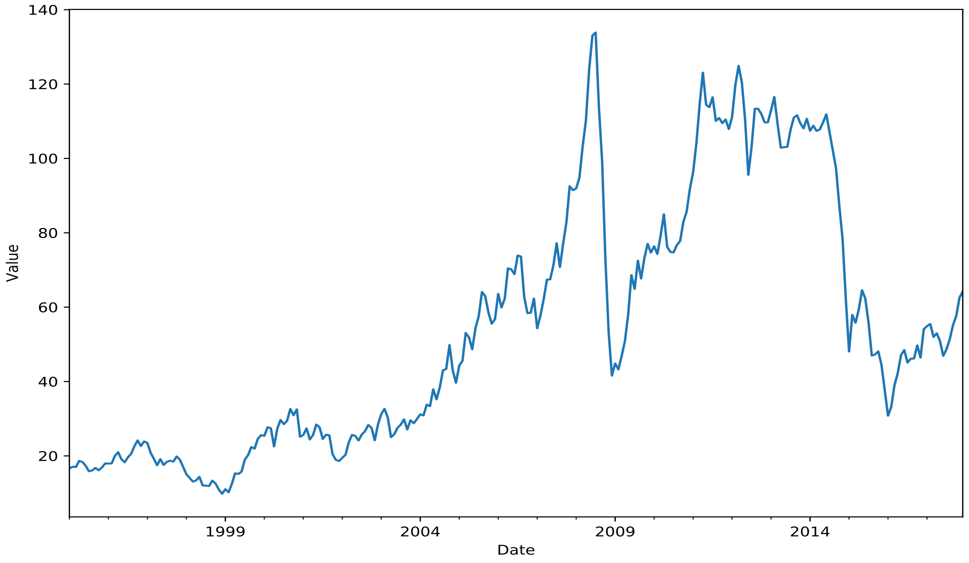
3 Results and discussion

In this section, we present the treatment and analysis of the time series used in the specification of the models. Section 3.1 lists the variables used in the study, and it presents statistical facts found by some authors, culminating in a brief research and contextualization of the world scenario on oil prices. Then, in Section 3.2 we describe the statistical tests applied to the time series and the treatment adopted in the presence of unit root, structural breaks, nonlinearity and co-integration, among others. Section 3.3 is designed to evaluates the forecasting performance of econometric models, while Section 3.4 explain the oil price prediction performance from real-time data plus a sentiment indicator. In general, the purpose of this section is to demonstrate that the MIDAS model, with daily financial indicators and an index of textual sentiments as explanatory variables, performs better at predicting oil prices compared to more traditional models in the literature.

3.1 Time-series analysis

The data-set for oil price is made up of 276 observations ranging from January 1995 to December 2017. Figure 1 illustrates the trajectory of our variable of interest over time.

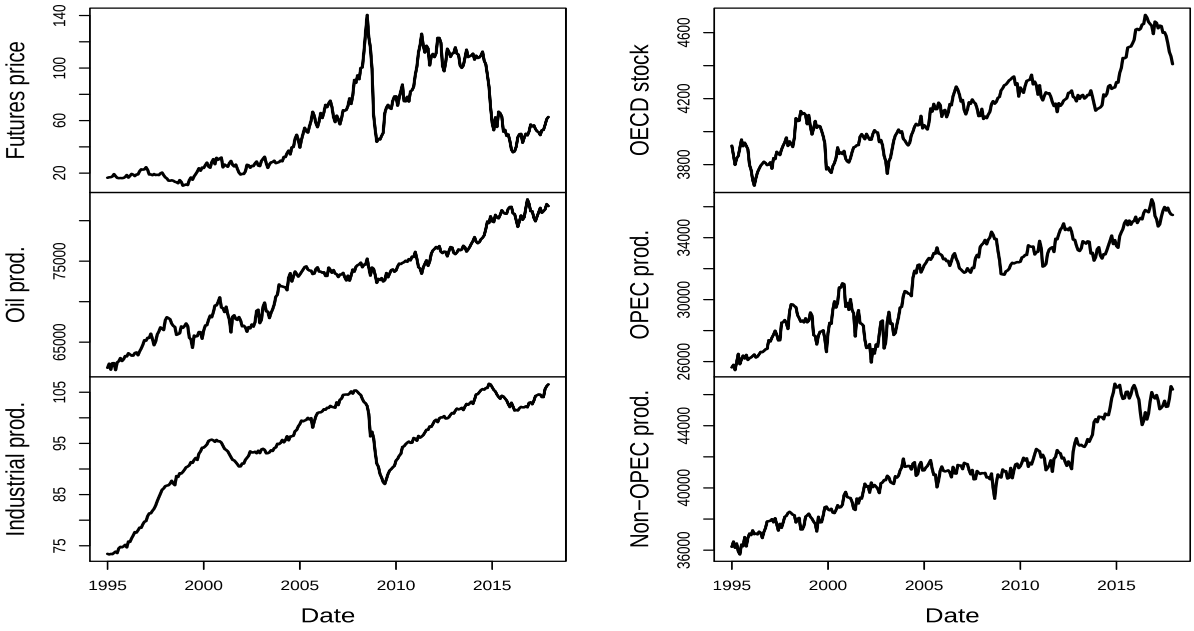
Figure 1 – Oil price trajectory



Source: authors' elaboration.

As illustrated in Figure 1, the trajectory of the oil price has a positive tendency, with high volatility and shifts during the period of analysis. A high fluctuation in the oil price over time can be observed, reaching its peak of US$ 133.87 in the financial crisis of 2008–2009. After the crisis, there was a drastic drop in price and subsequently a recovery with mild stability during 2010–2013. Another moment of sudden price drop can be seen during 2014; Arezki and Blanchard (2015) and Baumeister and Kilian (2016) attribute this occurrence to the increase in supply of shale oil and consequently an increase in U.S. oil supply.[4] The Figure 2 depicts the monthly time series treated and used in this study as oil price determinants in the VAR A, VAR B and VAR C models.

Figure 2 – Oil price determinants



Source: authors' elaboration.

[4] The distribution of oil prices data is shown in Appendix A.

According to Figure 2, all variables follow a positive trend throughout the study period. Another common feature of the variables is their decline during the U.S. subprime crisis in 2008, with an exception only of oil stock in Organization for Economic Cooperation and Development (OECD) countries.

3.2 Treatment of low-frequency time series

This section describes the treatment time series through econometric techniques. Table 3 lists the seasonal patterns. We checked for the possible presence, if any, of a seasonal component in the series, and the correction was made through the X-13 Arima Seats test, where the null hypothesis is the existence of no seasonal effects.

We noticed that the oil price, oil price futures and world industrial production index series do not have seasonality, based on the original series (qsori). The other variables presented some intrinsic seasonal component over time.

Table 3 – Seasonality test

Variable	**qsori**	**qsorievadj**	**qsrsd**	**qssadj**	**qssadjevadj**
Oil spot price	0.4039	0.4039	0.6685	0.4039	0.4039
Oil futures price	0.9870	0.0658	0.9768	0.9870	0.0658
World oil production	0.0000	0.0000	1.0000	1.0000	1.0000
World industrial production index	1.0000	1.0000	1.0000	1.0000	1.0000
OCDE oil stock	0.0000	0.0000	1.0000	1.0000	1.0000
OPEC oil production	0.0050	0.2496	0.1568	0.0050	0.2496
Non-OPEC oil production	0.0000	0.0000	1.0000	1.0000	1.0000

Source: authors' elaboration.

[1] Note: The output of R software does not have critical values and calculated values; p values were obtained by an approximation made by simulating a chi-square distribution with two degrees of freedom.

After testing for the presence of seasonality, the next step is to check the stationarity of the time series. We used two tests consolidated in the time series literature, namely the ADF and PP tests [5]. The null hypothesis of both tests is the presence of a unit root. The result is described in Table 4, and as can be seen, we found that all variables have a stochastic trend in both tests. To make these series stable, we applied the first difference.

Table 4 – Unit root tests

Variable	**ADF**		**Phillips-Perron**	
	Level	**Difference**	**Level**	**Difference**
Oil spot price	0.52	0.00	0.39	0.00
Oil futures price	0.41	0.00	0.34	0.00
World oil production	0.79	0.00	0.75	0.00
World industrial production index	0.20	0.00	0.20	0.00
OCDE oil stock	0.48	0.00	0.77	0.00
OPEC oil production	0.60	0.00	0.55	0.00
Non-OPEC oil production	0.84	0.00	0.86	0.00

Source: authors' elaboration.

[2] Note: We considered the significance level of 5%.

We then checked for the presence of structural breaks in the series through a Zivot-Andrews test, presented in Table 5. As Zivot and Andrews (2002) indicate, this test provides the detection of structural breaks in time series without its knowledge in advance. The main limitation of this test is the possibility of detecting only a structural deviation.

As displayed in Table 5, the Zivot-Andrews test indicates structural breaks in the following series during or after the U.S. international crisis of 2008: spot price, future price, world oil production and world industrial production index. Furthermore, the stock of oil in OECD member countries,

[5] For further details, see Dickey and Fuller (1979) and Phillips and Perron (1988).

oil production of OPEC members and non-OPEC members presented some structural breaks in the middle of the 2000s.

Table 5 – Structural break assessment by Zivot Andrews test

Variable	Potential break	t-cal	t-crit
Oil spot price	Sep/2014	-3.14	-5.08
Oil futures price	Sep/2010	-3.33	-5.08
World oil production	Jun/2008	-2.95	-5.08
World industrial production index	Mar/2008	-3.87	-5.08
OCDE oil stock	Jan/1999	-3.09	-5.08
OPEC oil production	Jul/2003	-3.78	-5.08
Non-OPEC oil production	May/2005	-2.906	-5.08

Source: authors' elaboration.

[3] Note: The significance level was 5%.

[4] Note: The test was performed considering intercept and trend in the estimation.

After verification of the presence of structural breaks, we investigated the existence of quadratic nonlinear behavior in time series, as in McLeod and Li (1983), Keenan (1985) and Tsay (1989). These three procedures' null hypothesis is that the model follows an autoregressive process, with linear behavior. According to Table 6, the oil price series has nonlinear behavior.

Table 6 – Nonlinearity tests

Test	p-value
Keenan	0.0120
McLeod-Li	0.0000
Tsay	0.0000

Source: authors' elaboration.

In Table 7 we present the optimal lag of econometric models based on AIC criteria. For the ARIMA model, the optimal lag was (2,1,3), and for VAR models, it was 10. Moreover, for threshold models and smooth transition models, the optimal lag was 2.

Table 7 – Optimal lag tests – AIC

Model	Optimal lag	AIC
Random walk	-	-543.19
AR(p)	2	-552.50
MA(q)	1	-551.89
ARIMA(p,d,q)	2,1,3	-552,61
VAR A	10	-11,02
VAR B	10	-38,41
VAR C	10	-49.24
SETAR(TAR)	2	-1367.91
SETAR(MTAR)	2	-1345.64
LSTAR	2	-1363.99
ESTAR	2	-1364.47

Source: authors' elaboration.

We then carried out the Johansen (1988) co-integration test, to verify the existence of a long-term relationship between the series used in the multivariate models (i.e., the VARs). As shown in Appendix C, the variables do not co-integrate. We thus proceeded with the estimation of a model of vector autoregression without error correction.

Finally, prior to the performance evaluation of the predictions, we assessed the residuals of the estimations for auto-correlation and heteroscedasticity, presented in Table 8. We observed that the residuals of most models are not auto-correlated and are homoscedastic. The only exception was the random walk, which rejected the null hypothesis of no auto-correlation, measured by Portmanteau test.

Table 8 – Test of estimated models' residuals

	Portmanteau (Auto-correlation)	Arch (Heteroscedasticity)
Model	**P-value**	**P-value**
Random walk	0.0008	0.3819
AR(p)	0.9647	0.6819
MA(q)	0.9004	0.5539
ARIMA(p,d,q)	0.8719	0.8266
VAR A	0.8140	0.3436
VAR B	0.6160	0.2819
VAR C	0.9418	0.2139
SETAR(TAR)	0.7951	0.6388
SETAR(MTAR)	0.9712	0.7224
LSTAR	0.8666	0.4661
ESTAR	0.9046	0.5686

Source: authors' elaboration.

[5] Note: For univariate models, we applied the tests of Box and Pierce (1970) and White et al. (1980) to detect auto-correlation and heteroscedasticity respectively.

3.3 Evaluation of predictions

After the treatment of the series, we proceeded with the estimation and evaluation of prediction accuracy.

Table 9 – Comparison of the models' accuracy

Model	**RMSE**
Random Walk	0.0893
AR(2)	0.0868
MA(1)	0.0876
ARIMA(2,1,3)	0.0858
VAR A	0.0640
VAR B	0.0381
VAR C	0.0326
SETAR(TAR)	0.0821
SETAR(MTAR)	0.0856
LSTAR	0.0823
ESTAR	0.0824

Source: authors' elaboration.

According to Table 9, VAR C has the lowest RMSE among all multivariate models tested. This result corroborates the studies by Mirmirani and Li (2004) and Murat and Tokat (2009). However, considering the nonlinear models, the SETAR has the best prediction accuracy, as also found in Albuquerquemello et al. (2018).

Table 10 – Root mean square error (10 periods out of the sample)

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Random walk	0.0647	0.0520	0.0536	0.0893	0.1432	0.1799	0.2071	0.2244	0.2612	0.2990
AR(2)	0.0489	0.0179	0.0014	0.0128	0.0449	0.0571	0.0576	0.0458	0.0549	0.0644
MA(1)	0.0634	0.0501	0.0513	0.0867	0.1402	0.1767	0.2036	0.2206	0.2572	0.2948
ARIMA(2,1,3)	0.0450	0.0337	0.0329	0.0643	0.1207	0.1536	0.1817	0.1991	0.2340	0.2742
VAR A	0.0705	0.0650	0.0729	0.1129	0.1788	0.2264	0.2642	0.2959	0.3443	0.3946
VAR B	0.0350	0.0031	0.0077	0.0408	0.1057	0.1417	0.1730	0.2086	0.2639	0.3253
VAR C	0.0191	0.0442	0.0752	0.0439	0.0088	0.0343	0.0611	0.0938	0.1485	0.2116
SETAR(TAR)	0.0548	0.0302	0.0177	0.0386	0.0770	0.0960	0.1035	0.0989	0.1145	0.1293
SETAR(MTAR)	0.0633	0.0492	0.0493	0.0835	0.1357	0.1707	0.1960	0.2113	0.2462	0.2820
LSTAR	0.0555	0.0316	0.0201	0.0420	0.0815	0.1018	0.1107	0.1078	0.1248	0.1412
ESTAR	0.0559	0.0323	0.0213	0.0437	0.0837	0.1046	0.1142	0.1118	0.1292	0.1458

Source: authors' elaboration.

Considering the out-of-sample forecasts for 10 periods ahead, the best performances were exhibited by VAR C and the autoregressive (AR) model, as indicated in Table 10. For a better understanding of the accuracy of prediction, we compared different models, taking into account the loss function of residuals with a Diebold-Mariano test.

Table 11 – Comparison of models by Diebold-Mariano test

Model	**Random Walk**	**AR(2)**	**MA(1)**	**ARIMA (2,1,3)**	**VAR C**	**SETAR (TAR)**	**SETAR (MTAR)**	**LSTAR**	**ESTAR**
Random Walk	-	0.1669	0.1721	0.0896	0	0.1116	0.0868	0.1177	0.1193
AR(2)	0.8331	-	0.8004	0.2333	0	0.1042	0.0273	0.1128	0.1149
MA(1)	0.8279	0.1996	-	0.0980	0	0.1095	0.0585	0.1169	0.1189
ARIMA (2,1,3)	0.9105	0.7667	0.9020	-	0	0.1846	0.3649	0.1979	0.2015
VAR C	1	1	1	1	-	1	1	1	1
SETAR (TAR)	0.8884	0.8958	0.8905	0.8154	0	-	0.8449	0.8913	0.9236
SETAR (MTAR)	0.9132	0.9726	0.9415	0.6351	0	0.1551	-	0.1682	0.1723
LSTAR	0.8823	0.8872	0.8831	0.8021	0	0.1087	0.8318	-	0.8689
ESTAR	0.8807	0.8851	0.8811	0.7985	0	0.0764	0.8277	0.1311	-

Source: authors' elaboration.

[6] Note: each number presents the p-value of the DM test when comparing two models. The null hypothesis that the two forecasts have no difference will be rejected if the p-value is higher than the significance level of 0.05.

As presented in Table 11, we found that the best model indicated by the test is VAR C, and the second best model. Finally, as expected, the weakest econometric model is the random walk.

3.4 Performance prediction for high-frequency data – MIDAS

The previous sections described the analysis of performance prediction considering data with single frequency. This section presents the models with high-frequency data, estimated by MIDAS regression.

In the estimation of MIDAS, we used a monthly indicator for the sentiment of authorities on the oil market, which is the tone of reports issued by the EIA. This latter variable was constructed with computational algorithms that extract the cognitive aspects of texts. In addition, we also used eight daily time series as predictors of the BRENT oil price.

The DGS10 variable is the constant rate of the treasure bonds to a maturity of 10 years, and VIX is the volatility index of the implied stock market in the options of the S&P 500. In addition, the inter-bank interest rate used is the LIBOR 3-month, established in London, DTWEXM is the index of the dollar weighted by trade for the major currencies. The remaining four variables are as follows: NYSE composition ratios (NYA), oil and gas NYSE ARCA (XOI), regular spot price of conventional gas from New York Harbor (gasoline) and daily future oil prices (RCLC).

Due to the mixed frequency of data, the number of observations was different. We solved this problem in two steps: (i) we excluded the missings of the high-frequency time series, and (ii) we selected common data among the variables. The number of observations, totaling 5,590, was consequently made compatible for all indicators.

Table 12 – Performance prediction in sample

Model	**RMSE**
MIDAS$_{Almon}$	0.0412
MIDAS$_{Umidas}$	0.0265
ADL-MIDAS$_{Almon}$	0.0413
ADL-MIDAS$_{Umidas}$	0.0259
MIDAS$_{Almon}(sent)$	0.0399
MIDAS$_{Umidas}(sent)$	0.0264
ADL-MIDAS$_{Almon}(sent)$	0.0411
ADL-MIDAS$_{Umidas}(sent)$	0.0257

Source: authors' elaboration.

After estimation, we performed forecasts in sample and out of the sample for the oil price time series. Table 12, present the accuracy of each model used, indicated by the RMSE. The ADL-MIDAS$_{Umidas}(sent)$ had higher accuracy when compared to the remaining MIDAS extensions. Moreover, we observed that the inclusion of the sentiment indicator improved all MIDAS models and that the ADL-MIDAS$_{Umidas}$, MIDAS$_{Umidas}$, MIDAS$_{Umidas}(sent)$ and ADL-MIDAS$_{Umidas}(sent)$ models had better accuracies compared to the estimated models with low-frequency data, from Section 3.3.

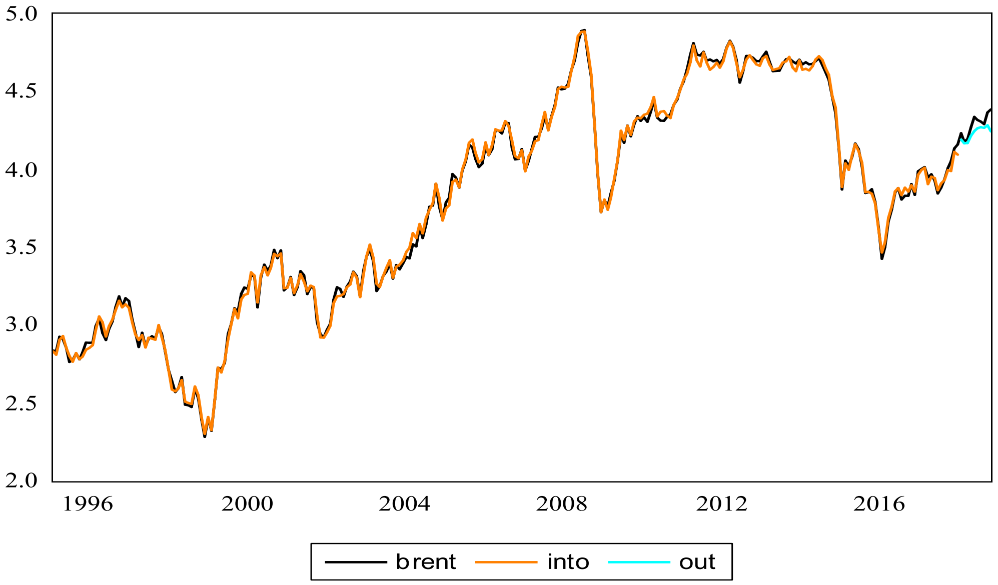
Table 13 – Performance of 10 out-of-sample periods – RMSE

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MIDAS$_{Almon}$	0.067	0.085	0.103	0.121	0.128	0.134	0.134	0.137	0.157	0.185
MIDAS$_{Umidas}$	0.009	0.018	0.049	0.117	0.127	0.169	0.162	0.147	0.166	0.175
ADL-MIDAS$_{Almon}$	0.006	0.009	0.022	0.043	0.068	0.095	0.102	0.108	0.128	0.154
ADL-MIDAS$_{Umidas}$	0.008	0.017	0.035	0.101	0.129	0.158	0.167	0.163	0.184	0.196
MIDAS$_{Almon}(sent)$	0.070	0.088	0.114	0.129	0.141	0.146	0.140	0.143	0.161	0.191
MIDAS$_{Umidas}(sent)$	0.002	0.023	0.060	0.129	0.137	0.181	0.175	0.160	0.179	0.186
ADL-MIDAS$_{Almon}(sent)$	0.011	0.017	0.033	0.057	0.081	0.106	0.115	0.119	0.139	0.165
ADL-MIDAS$_{Umidas}(sent)$	0.002	0.022	0.046	0.114	0.141	0.171	0.181	0.177	0.199	0.209

Source: authors elaboration.

For forecasts out of the sample, according to Table 13, for one period ahead, the ADL-MIDAS$_{Umidas}(sent)$ model has superior performance over any other model. However, considering a 10-period horizon, ADL-MIDAS$_{Almon}$ presented the best performance. As proposed by Ghysels (2016), the textual sentiment variable positively adds to the accuracy of financial data, such as oil prices. Figure 3 illustrates the comparison of oil price predictions to the original series.

Figure 3 – Prediction into and out of the sample using MIDAS



Source: authors' elaboration.

[7] Note: The crude oil price series for the analysis of out-of-sample forecasts consists of period from 2018.1 to 2018.10.

[8] Note: For the predictions in and out of the sample, we considered the ADL-MIDAS$_{Umidas}(sent)$ and ADL-MIDAS$_{Almon}$ models.

[9] Note: The predictions were performed with the time series values in a natural logarithm, except for the textual indicator.

As illustrated in Figure 3, we observed that the MIDAS model has a high capacity to predict oil prices, as it was possible to capture oscillations and the positive trend of the original series with high precision. Albuquerquemello et al. (2018) investigated traditional models in the literature; however, they did not consider high-frequency financial indicators in econometric models for forecasting oil prices. Therefore, our study contributes strongly to the literature, since our analysis involves high-frequency (daily) variables and a monthly indicator of sentiment, as explanatory variables, which improve the performance of the forecast of oil prices.

4 Robustness test

This section presents a robustness analysis to predict the price of oil. Here, we consider the period in-sample period, and we use the models in Table 12 to verify the results. The exercise of robustness used was the substitution of the variable spot price dependent on BRENT oil for the spot price of oil quoted on the New York Stock Exchange, the WTI.

The table in Appendix D contains the results obtained by MIDAS models with the price of WTI oil as a dependent variable. We found that the results are similar to those obtained by the price of BRENT oil. The model with the greatest potential to predict the price of oil remains the ADL-MIDAS$_{Umidas}(sent)$ model. Therefore, it is valid to use the spot price of WTI to capture the same effect as the estimated models. Appendix E presents a comparison of in- and out-of-sample forecasts for WTI oil prices.

5 Conclusions

In this article, we proposed alternative econometric approaches to improve the prediction of oil prices. We focused on the MIDAS model with the inclusion of high- and low-frequency variables and an indicator of textual sentiment as explanatory variables for the spot price of Brent oil.

Our findings suggest that the following: (i) the oil price series has a nonlinear quadratic pattern; (ii) within the sample prediction, the ADL-MIDAS$_{Umidas}(sent)$ model is the most accurate predictor; (iii) the SETAR model has better accuracy only in comparison to other univariate models (linear and nonlinear); (iv) for out-of-the sample forecasting (10 periods ahead), the MIDAS$_{Umidas}(sent)$ and MIDAS$_{Almon}$ models had the better performance; (v) the inclusion of the sentiment index augments the forecast accuracy for every model; and (vi) the temporal frequency and the manner of data collection and construction are essential for improvements in prediction performance.

Among the results of our study, the main contribution to the oil price forecast literature is the inclusion of a textual sentiment indicator in the mixed frequency models (MIDAS). Based on computational algorithms, this innovative indicator was extracted from texts in the oil market reports issued by the EIA. A robustness test with the spot price of WTI oil as a dependent variable confirms the model with the highest prediction performance, namely the econometric model ADL-MIDAS$_{Umidas}(sent)$.

With this study, we provide tools for policy makers and financial agents regarding the oil market. This contribution can occur in several ways; we highlight two: helping in the daily routine with respect to portfolio assets and reducing information asymmetries in the expectations of the oil market. Furthermore, our results motivate future research that considers two different models of mixed frequencies for forecasting oil prices: MF-VAR and MIDAS. The selection of explanatory variables can be established using machine learning techniques.

ACKNOWLEDGEMENT

This study was financed in party by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

AGNOLUCCI, P. Volatility in crude oil futures: a comparison of the predictive ability of garch and implied volatility models. *Energy Economics*, Elsevier, v. 31, n. 2, p. 316–321, 2009.

ALBUQUERQUEMELLO, V. P. de; MEDEIROS, R. K. de; BESARRIA, C. da N.; MAIA, S. F. Forecasting crude oil price: Does exist an optimal econometric model? *Energy*, Elsevier, v. 155, p. 578–591, 2018.

ANDREOU, E.; GHYSELS, E.; KOURTELLOS, A. Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, Taylor & Francis, v. 31, n. 2, p. 240–251, 2013.

AREZKI, R.; BLANCHARD, O. The 2014 oil price slump: Seven key questions. *VoxEU, January*, v. 13, 2015.

ARMESTO, M. T.; ENGEMANN, K. M.; OWYANG, M. T. et al. Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, Federal Reserve Bank of St. Louis, v. 92, n. 6, p. 521–36, 2010.

BAFFES, J. Oil spills on other commodities. *Resources Policy*, Elsevier, v. 32, n. 3, p. 126–134, 2007. ISSN 0301-4207.

BAUMEISTER, C.; GUÉRIN, P.; KILIAN, L. Do high-frequency financial data help forecast oil prices? the midas touch at work. *International Journal of Forecasting*, Elsevier, v. 31, n. 2, p. 238–252, 2015.

BAUMEISTER, C.; KILIAN, L. Do oil price increases cause higher food prices? *Economic Policy*, Oxford University Press, v. 29, n. 80, p. 691–747, 2014a.

BAUMEISTER, C.; KILIAN, L. Understanding the decline in the price of oil since june 2014. *Journal of the Association of Environmental and Resource Economists*, University of Chicago Press Chicago, IL, v. 3, n. 1, p. 131–158, 2016.

BOX, G. E.; PIERCE, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, Taylor & Francis Group, v. 65, n. 332, p. 1509–1526, 1970.

CARPIO, L. G. T. The effects of oil price volatility on ethanol, gasoline, and sugar price forecasts. *Energy*, Elsevier, 2019.

CASAMASSIMA, G.; FIORELLO, D.; MARTINO, A. The impact of oil prices fluctuactions on transport and its related sectors. *European Parliament, Directorate General Internal Policies of the Union, Policy Department Structural and Cohesion Policies, Transport and Tourism, PE*, v. 419, 2009.

CHAI, J.; XING, L.-M.; ZHOU, X.-Y.; ZHANG, Z. G.; LI, J.-X. Forecasting the wti crude oil price by a hybrid-refined method. *Energy Economics*, Elsevier, v. 71, p. 114–127, 2018.

CHIROMA, H.; ABDULKAREEM, S.; HERAWAN, T. Evolutionary neural network model for west texas intermediate crude oil price prediction. *Applied Energy*, Elsevier, v. 142, p. 266–273, 2015.

CLEMENTS, M. P.; GALVÃO, A. B. Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 26, n. 4, p. 546–554, 2008.

CLEMENTS, M. P.; GALVÃO, A. B. Forecasting us output growth using leading indicators: An appraisal using midas models. *Journal of Applied Econometrics*, Wiley Online Library, v. 24, n. 7, p. 1187–1206, 2009.

COPPOLA, A. Forecasting oil price movements: Exploiting the information in the futures market. *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, Wiley Online Library, v. 28, n. 1, p. 34–56, 2008.

CUÑADO, J.; Pérez de Gracia, F. Do oil price shocks matter? Evidence for some European countries. *Energy Economics*, v. 25, n. 2, p. 137–154, 2003. ISSN 01409883.

DEENEY, P.; CUMMINS, M.; DOWLING, M.; BERMINGHAM, A. Sentiment in oil markets. *International Review of Financial Analysis*, Elsevier, v. 39, p. 179–185, 2015.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, Taylor & Francis, v. 74, n. 366a, p. 427–431, 1979.

DIEBOLD, F. X.; MARIANO, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, v. 13, n. 3, 1995.

EIKA, T.; MAGNUSSEN, K. A. Did norway gain from the 1979–1985 oil price shock? *Economic Modelling*, Elsevier, v. 17, n. 1, p. 107–137, 2000.

ENDERS, W. *Applied econometric time series*. [S.l.]: John Wiley & Sons, 2008.

FORONI, C.; MARCELLINO, M.; SCHUMACHER, C. Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 178, n. 1, p. 57–82, 2015.

FORONI, C.; MARCELLINO, M. G. A survey of econometric methods for mixed-frequency data. *Available at SSRN 2268912*, 2013.

GHYSELS, E. Macroeconomics and the reality of mixed frequency data. *Journal of Econometrics*, Elsevier, v. 193, n. 2, p. 294–314, 2016.

GHYSELS, E.; SANTA-CLARA, P.; VALKANOV, R. The midas touch: Mixed data sampling regression models. *Working Paper*, 2004.

GHYSELS, E.; SINKO, A.; VALKANOV, R. Midas regressions: Further results and new directions. *Econometric Reviews*, Taylor & Francis, v. 26, n. 1, p. 53–90, 2007.

GHYSELS, E.; WRIGHT, J. H. Forecasting professional forecasters. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 27, n. 4, p. 504–516, 2009.

GRANGER, C.; TERäSVIRTA, T. *Modelling Non-Linear Economic Relationships*. [S.l.]: Oxford University Press, 1993.

HAMILTON, J. Understanding crude oil prices. *Energy Journal*, International Association for Energy Economics, Cleveland, USA, v. 30, n. 2, p. 179–206, 2009.

HAMILTON, J. D. This is what happened to the oil price-macroeconomy relationship. *Journal of Monetary Economics*, Elsevier, v. 38, n. 2, p. 215–220, 1996. ISSN 03043932.

HAMILTON, J. D. Understanding Crude Oil Prices. *NBER Working Paper Series*, n. 14492, p. 1–44, 2009. ISSN 01956574.

HOTELLING, H. The economics of exhaustible resources. *Journal of political Economy*, The University of Chicago Press, v. 39, n. 2, p. 137–175, 1931.

JOCKERS, M. Package ‘syuzhet’. *URL: https://cran. r-project. org/web/packages/syuzhet*, 2017.

JOHANSEN, S. Statistical analysis of cointegration vectors. *Journal of economic dynamics and control*, Elsevier, v. 12, n. 2-3, p. 231–254, 1988.

KEENAN, D. M. A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, Oxford University Press, v. 72, n. 1, p. 39–44, 1985.

KILIAN, L. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, v. 99, n. 3, p. 1053–69, 2009.

KILIAN, L.; MURPHY, D. P. The role of inventories and speculative trading in the global market for crude oil. *Journal of Applied Econometrics*, Wiley Online Library, v. 29, n. 3, p. 454–478, 2014.

KILIAN, L.; PARK, C. The impact of oil price shocks on the U.S. stock market. *International Economic Review*, v. 50, n. 4, p. 1267–1287, 2009. ISSN 00206598.

LI, X.; SHANG, W.; WANG, S. Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, Elsevier, v. 35, n. 4, p. 1548–1560, 2019.

LÜTKEPOHL, H. *New introduction to multiple time series analysis.* [S.l.]: Springer Science & Business Media, 2005.

MCLEOD, A. I.; LI, W. K. Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of time series analysis*, Wiley Online Library, v. 4, n. 4, p. 269–273, 1983.

MIAO, H.; RAMCHANDER, S.; WANG, T.; YANG, D. Influential factors in crude oil price forecasting. *Energy Economics*, Elsevier, v. 68, p. 77–88, 2017.

MIRMIRANI, S.; LI, H. C. A comparison of var and neural networks with genetic algorithm in forecasting price of oil. In: *Applications of Artificial Intelligence in Finance and Economics.* [S.l.]: Emerald Group Publishing Limited, 2004. p. 203–223.

MOHAMMADI, H.; SU, L. International evidence on crude oil price dynamics: Applications of arima-garch models. *Energy Economics*, Elsevier, v. 32, n. 5, p. 1001–1008, 2010.

MOVAGHARNEJAD, K.; MEHDIZADEH, B.; BANIHASHEMI, M.; KORDKHEILI, M. S. Forecasting the differences between various commercial oil prices in the persian gulf region by neural network. *Energy*, Elsevier, v. 36, n. 7, p. 3979–3984, 2011.

MURAT, A.; TOKAT, E. Forecasting oil price movements with crack spread futures. *Energy Economics*, Elsevier, v. 31, n. 1, p. 85–90, 2009.

OLOFIN, S. O.; OLOKO, T. F.; ISAH, K. O.; OGBONNA, A. E. Crude oil price–shale oil production nexus: a predictability analysis. *International Journal of Energy Sector Management*, Emerald Publishing Limited, 2020.

PAN, Z.; WANG, Q.; WANG, Y.; YANG, L. Forecasting us real gdp using oil prices: A time-varying parameter midas model. *Energy Economics*, Elsevier, v. 72, p. 177–187, 2018.

PHILLIPS, P. C.; PERRON, P. Testing for a unit root in time series regression. *Biometrika*, Oxford University Press, v. 75, n. 2, p. 335–346, 1988.

RATKOWSKY, D. A.; GILES, D. E. *Handbook of nonlinear regression models.* [S.l.]: M. Dekker New York, 1990.

REICHSFELD, D. A.; ROACHE, S. K. *Do Commodity Futures Help Forecast Spot Prices?* [S.l.], 2011.

SALISU, A. A.; RAHEEM, I. D.; NDAKO, U. B. A sectoral analysis of asymmetric nexus between oil price and stock returns. *International Review of Economics & Finance*, Elsevier, v. 61, p. 241–259, 2019.

SALISU, A. A.; SWARAY, R.; OLOKO, T. F. Improving the predictability of the oil–us stock nexus: The role of macroeconomic variables. *Economic Modelling*, Elsevier, v. 76, p. 153–171, 2019.

SIMS, C. A. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, JSTOR, p. 1–48, 1980.

TERÄSVIRTA, T. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the american Statistical association*, Taylor & Francis Group, v. 89, n. 425, p. 208–218, 1994.

TONG, H.; LIM, K. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 245–292, 1980.

TSAY, R. S. Testing and modeling threshold autoregressive processes. *Journal of the American statistical association*, Taylor & Francis Group, v. 84, n. 405, p. 231–240, 1989.

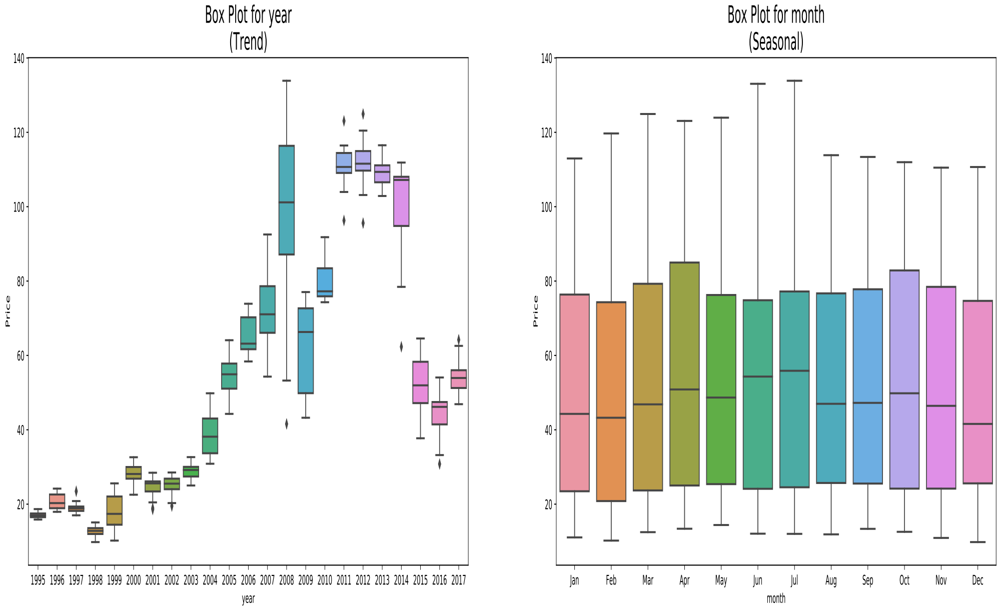
WESTERLUND, J.; NARAYAN, P. Testing for predictability in conditionally heteroskedastic stock returns. *Journal of Financial Econometrics*, Oxford University Press, v. 13, n. 2, p. 342–375, 2015.

WHITE, H. et al. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *econometrica*, Princeton, v. 48, n. 4, p. 817–838, 1980.

ZHANG, Y.-J.; WANG, J.-L. Do high-frequency stock market data help forecast crude oil prices? evidence from the midas models. *Energy Economics*, Elsevier, v. 78, p. 192–201, 2019.

ZIVOT, E.; ANDREWS, D. W. K. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of business & economic statistics*, Taylor & Francis, v. 20, n. 1, p. 25–44, 2002.

APPENDIX A – Box plot - Distribution of spot price series



Source: authors elaboration.

APPENDIX B – Time series summary statistics

Variable	Average	Median	Maximum	Minimum	Standard deviation	Asymmetry	Kurtosis	Observations
Oil spot price	54.14	47.19	137.87	9.80	34.01	0.59	2.09	276
Oil futures price	54.56	49.58	140.44	10.44	34.17	0.57	2.06	276
World oil production	72302.19	73398.23	82631.99	61555.18	5413.66	-0.06	2.07	276
World industrial production index	95.19	95.84	106.66	73.30	8.45	-0.94	3.26	276
OECD oil shock	4122.00	4126.84	4707.31	3672.46	230.97	0.54	2.90	276
OPEC oil production	31272.97	32222.47	36475.85	25451.92	2942.04	-0.31	1.84	276
Non-OPEC oil production	41029.21	41062.16	46699.53	35723.28	2689.05	0.31	2.47	276
Oil market sentiment index	0.10	0.08	0.67	-0.54	0.27	-0.01	2.38	276
DGS10	4.08	4.15	7.89	1.37	1.56	0.12	1.96	5590
VIX	19.95	18.36	80.86	9.14	8.28	1.99	9.92	5590
LIBOR 3-month	2.82	1.90	6.87	0.22	2.33	0.28	1.38	5590
DTWEXM	86.88	86.37	113.10	68.03	10.80	0.34	2.24	5590
NYA	7411.58	7066.81	12853.09	2651.15	2276.27	0.13	2.35	5590
XOI	880.61	957.83	1726.22	258.24	398.68	0.06	1.57	5590
Gasoline	1.51	1.41	3.67	0.29	0.85	0.44	1.95	5590
RCLC	52.63	47.40	145.29	10.72	30.22	0.53	2.14	5590

Source: authors elaboration.

APPENDIX C – Johansen co-integration test

	A	B	C
Statistic	**r=0**	**r=0**	**r=0**
$\lambda_{eigenvalue}$ (calculated at 1%)	11.65	25.75	32.14
$\lambda_{eigenvalue}$ (critical at 1%)	19.19	32.14	38.78
λ_{trace} (calculated at 1%)	11.65	37.22	55.43
λ_{trace} (critical at 1%)	23.52	55.43	78.87

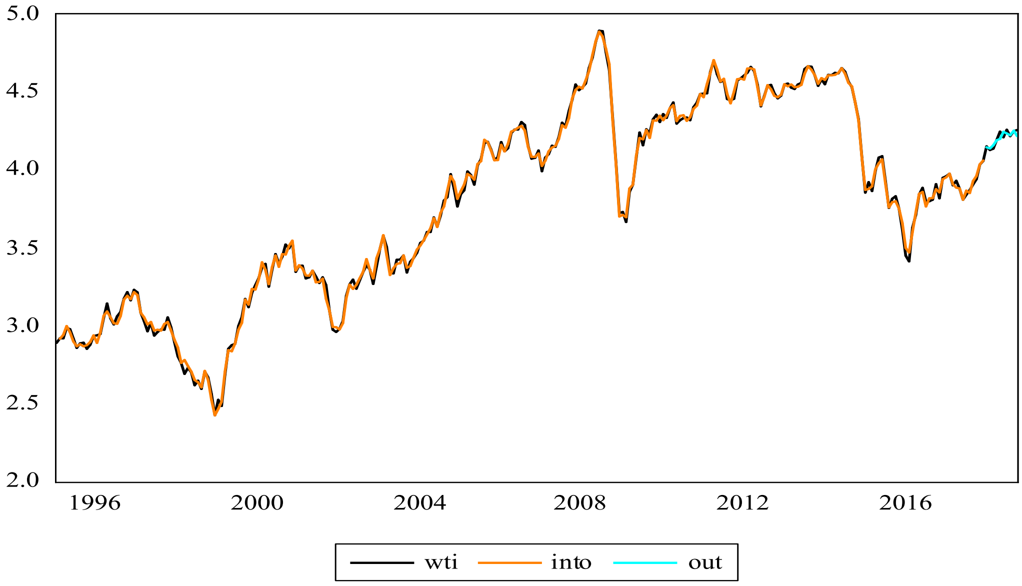
Source: authors elaboration.

APPENDIX D – Prediction accuracy in sample for WTI oil price

Model	**RMSE**
MIDAS$_{Almon}$	0.0362
MIDAS$_{Umidas}$	0.0359
ADL-MIDAS$_{Almon}$	0.0248
ADL-MIDAS$_{Umidas}$	0.0247
MIDAS$_{Almon}(sent)$	0.0350
MIDAS$_{Umidas}(sent)$	0.0346
ADL-MIDAS$_{Almon}(sent)$	0.0246
ADL-MIDAS$_{Umidas}(sent)$	0.0244

Source: authors elaboration.

APPENDIX E – Prediction for WTI crude oil



Source: authors elaboration.

[10] Note: The crude oil price series for the analysis of out of sample forecasts, consists in the period: 2018.1 to 2018.10.

[11] Note: For the predictions in and out of the sample we considered ADL-MIDAS$_{Umidas}(sent)$ and ADL-MIDAS$_{Almon}$ models.

[12] Note: The predictions were performed with the time series values in natural logarithm, with except for the textual indicator.