UNIVERSIDADE D
COIMBRA

Gonçalo Wilson da Silva Punza

COMPARAÇÃO DE MODELOS DE PREVISÃO DO PREÇO DO PETRÓLEO

Dissertação no âmbito do Mestrado em Economia, com especialização em Economia Financeira orientada pelo Professor Doutor Pedro Miguel Avelino Bação e apresentado à Faculdade de Economia da Universidade de Coimbra

Junho de 2019

FEUC FACULDADE DE ECONOMIA
UNIVERSIDADE DE COIMBRA

Gonçalo Wilson da Silva Punza

Comparação de modelos de previsão do preço do petróleo

Trabalho de Projeto no âmbito do Mestrado em Economia, com especialização em Economia Financeira, orientado pelo Professor Doutor Pedro Miguel Avelino Bação e apresentado à Faculdade de Economia da Universidade de Coimbra

Junho de 2019

AGRADECIMENTOS

Gostaria de agradecer a minha família e amigos que me apoiaram durante todo o meu percurso escolar. À minha mãe, obrigado por estares sempre ao meu lado e pelo apoio incondicional!

Um agradecimento especial ao Professor Doutor Pedro Miguel Avelino Bação pela ajuda e disponibilidade demonstrada ao longo do processo de escrita do documento.

"Was Du erlebst, kann keine Macht der Welt Dir rauben"

RESUMO

Dado a importância que o petróleo tem para o funcionamento de muitas atividades económicas que têm estado na base do crescimento económico das nações, o petróleo é ainda a matéria-prima mais importante transacionada a nível mundial. A sua importância é tal que continua a ter um impacto muito grande na definição de políticas económicas por parte dos governos, para além da influência que tem sobre o comportamento das empresas e dos consumidores privados. O que é que determina a evolução do preço do petróleo? Este trabalho tem como objetivo avaliar as previsões obtidas com dois grupos de modelos de previsão. O primeiro grupo engloba os modelos mais tradicionais: ARIMA, VAR e regressão múltipla. O modelo VAR e o modelo de regressão múltipla incorporam informação sobre possíveis determinantes do preço do petróleo. Já o segundo grupo é composto por duas classes de modelos que se baseiam no *machine learning*: *support vector machine* (SVM) e rede neural artificial (ANN). Estes modelos são considerados "caixas negras", pois é difícil perceber de que forma as variáveis explicativas estão relacionadas com a variável dependente. Entre os modelos testados, apenas o modelo SVM com uma função kernel linear teve um desempenho semelhante ao do modelo de previsão *naive*, tendo os outros modelos apresentado desempenhos inferiores. Portanto, prever o preço do petróleo parece continuar a ser muito difícil, mesmo fazendo uso de modelos de *machine learning*.

PALAVRAS-CHAVE: ARIMA, modelo Vetor Autorregressivo (VAR), preço do petróleo, rede neural artificial (ANN), Support Vector Machine (SVM).

***Classificação JEL*:** C320, C450

ABSTRACT

Given the importance that oil has to the functioning of many economic activities that are the basis for the economic growth of nations, oil is still the most important raw material traded worldwide. Its importance is such that it continues to have a very big impact on the definition of economic policies by governments, besides its influence on the behaviour of companies and private consumers. What determines the price of oil? This work project evaluates the forecasts obtained with two groups of forecast models. The first group includes the more traditional models: ARIMA, VAR and multiple regression. The VAR model and the multiple regression model incorporate information about possible determinants of the price of oil. The second group is composed by machine learning models: support vector machine (SVM) and artificial neural network (ANN). These models are considered "black boxes", because it is difficult to understand the relation between the explanatory variables and the dependent variable. Among the models that we tested, only the SVM model with a linear kernel function performed at the level of the naïve forecast model, with all the other models performing worse. Therefore, to forecast the oil price is still very hard, even if one uses machine learning models.

***KEYWORDS**:* ARIMA, VAR model, oil price, Support Vector Machine (SVM), Artificial Neural Network (ANN).

JEL Classification: C320, C450

ÍNDICE

1. INTRODUÇÃO

O crude continua a ser a matéria prima mais importante, sendo a mais comercializada a nível mundial Yu et. al. (2008). Isto verifica-se apesar da tendência para a utilização de outras fontes de energia nos países desenvolvidos, tendência que tem vindo a contribuir para a redução do peso do petróleo na energia consumida. No entanto, o petróleo ainda continua a estar presente com grande peso na indústria – cf. Godarzi et. al. (2014). Dada a importância do petróleo, as empresas, os governos e os cidadãos são constantemente afetados pelas flutuações que o preço do petróleo sofre, de que foram exemplos maiores as crises petrolíferas nos anos 1970-1980 do século passado. Sendo uma matéria-prima tão essencial para o normal funcionamento das economias mundiais, o petróleo é comercializado internacionalmente por diferentes agentes: as nações produtoras de petróleo, empresas petrolíferas, refinarias independentes e as nações importadoras de petróleo – cf. Hamilton (2002). Os fatores que acabam por determinar o preço do petróleo são, por isso, variados. Entre esses fatores deverão estar o crescimento da produção (que usa o petróleo como matéria-prima), aspetos políticos (que podem alterar os impostos que oneram o consumo de petróleo, ou que podem fazer aumentar ou diminuir a produção de petróleo) ou mesmo as expectativas dos agentes económicos. O que é facto é que se têm observado flutuações extremas do preço do petróleo ao longo das últimas décadas. O estudo do comportamento do preço do petróleo poderá dar pistas sobre os fatores que mais influenciam hoje em dia a evolução do preço do petróleo, e ajudar os agentes económicos a formular previsões melhores.

O objetivo deste trabalho é efetuar comparações das previsões da evolução do preço do petróleo obtidas por diferentes métodos, incluindo os modelos econométricos mais tradicionais e também métodos de *machine learning*. O trabalho está organizado da seguinte forma. Na secção 2 é feita a revisão bibliográfica. Na secção 3 apresento os modelos a utilizar no corrente trabalho. Na secção 4 é feita a discussão dos resultados empíricos. A secção 5 conclui o trabalho.

2. REVISÃO BIBLIOGRÁFICA

O peso dos países produtores de petróleo, mais concretamente dos que fazem parte da Organização dos Países Exportadores de Petróleo (OPEP), no impacto do preço do petróleo tem vindo a diminuir com o passar das décadas – cf. Almoguera et. al. (2010). No ano de fundação desta organização, em 1960, os seus membros detinham um poder bastante significativo na determinação do preço do crude.

No entanto, as condições geopolíticas, crises económicas, as demais instabilidades políticas e sociais, os conflitos armados que ocorreram em alguns membros, e mais recentemente a crise política e económica na Venezuela, constituem fatores que contribuíram para a redução desse peso – cf. os relatórios da Agência Internacional da Energia (International Energy Agency, 2018). Brown e Huntington (2017) confirmam a ideia de que os países da OPEP têm contribuído para a instabilidade do mercado petrolífero. No seu estudo mostram que o desvio-padrão da taxa de crescimento da produção nos países da OPEP é mais elevado do que nos outros países. Esta redução do poder dos países da OPEP permitirá o aumento da influência dos países que não fazem parte desta organização, especificamente os Estados Unidos da América, que têm vindo a aumentar a sua produção nos últimos anos.

A tarefa de prever o preço futuro do crude é muito difícil, quer pela instabilidade dos países produtores, quer pelas demais variáveis que em determinado momento podem vir a influenciar esse preço. No entanto, a previsão do preço do petróleo é uma atividade importante, tendo passado a ter uma importância fulcral após as crises petrolíferas da década de setenta do século passado. Alguns setores da economia são extremamente dependentes das previsões do preço do petróleo para a tomada de decisões. Por exemplo, as companhias aéreas apoiam-se nessas previsões para definir as suas tarifas futuras. As previsões do preço do petróleo também influenciam decisões de investimento no setor da energia, com consequências para as emissões de gases de efeito de estufa, especificamente as emissões de dióxido de carbono.

Ao longo do tempo têm sido utilizados diversos métodos para tentar prever a evolução do preço do petróleo.

Baumeister e Kilian (2011) efetuaram um estudo de previsão para o preço real do petróleo WTI, em que recorreram a dados mensais para o período compreendido entre o início do ano de 1991 e finais de 2010. Neste estudo, foram utilizados modelos VAR e ARMA nas previsões dos preços reais do petróleo, tendo as previsões sido comparadas através do *mean square prediction error* (MSPR). Os modelos VAR apresentaram melhores resultados em comparação com os modelos autorregressivos puros (AR) e com os modelos (ARMA).

Hamilton (2008) examinou os fatores responsáveis pela mudança que o preço do petróleo sofreu no ano de 1997. Este estudo analisou o comportamento do preço tendo em consideração as influências que a procura e a oferta exercem. Ao longo do texto é feita uma revisão das várias teorias que contribuíram para o aumento do preço no ano de 2008, como por exemplo a procura mundial da matéria-prima (mais concretamente por parte da China) e o peso da OPEP.

A aplicação de métodos econométricos para estudar o comportamento de séries temporais é muito comum na previsão. O uso de modelos que se baseiam no *Machine Learning* é algo relativamente mais recente. Por exemplo, Yu et al. (2008) utilizaram o método EMD (*Empirical Mode Decomposition*) combinado com uma rede artificial neural. No estudo também foi utilizado um modelo de rede neural artificial (*ANN*) composto por três camadas de redes neurais. Os autores reconheceram que a tarefa de efetuar previsões para o preço do crude é uma tarefa complexa devido aos factores internos e externos, e pela volatilidade que essas variáveis impõem a qualquer que seja o preço do petróleo que seja alvo de análise.

Movagharnejad et al. (2011) investigaram as diferenças nos preços comerciais do petróleo para sete regiões diferentes do Golfo Pérsico desde o início de 2000 até abril de 2010. No estudo em questão foi utilizado um modelo ANN para calcular as previsões para as sete regiões. Este método permitiu aumentar a qualidade das previsões.

No estudo de Godarzi et al. (2014) foi efetuada uma comparação entre os modelos NARX (*Nonlinear AutoRegressive model with eXogenous variables*), um modelo ANN estático e modelos de séries temporais (ARIMA). O estudo utilizou dados de países da OCDE para o período de 1974-2005 e os autores recorreram aos valores do MAE (*mean absolute error*,

erro absoluto médio) como medida da qualidade das previsões dos modelos. O modelo NARX apresentou melhores resultados face ao modelo econométrico clássico, ARIMA.

Em Ramyar e Kianfar (2017) foram comparadas as previsões para o preço do crude para o mercado norte americano produzidas por um modelo baseado no *machine learning* (o modelo ANN) e por um modelo de regressão múltipla (com um desfasamento do preço do crude e outras variáveis explicativas). O critério utilizado para decidir qual dos modelos era o melhor foi o MSE, tendo a conclusão sido favorável ao modelo ANN.

3. METODOLOGIA

Nesta secção serão apresentados os quatro modelos que serão utilizados no processo de previsão. Em primeiro lugar, serão expostos os métodos de previsão econométricos "clássicos": o modelo ARIMA e o modelo VAR (para mais detalhes sobre estes modelos, para lá dos que serão apresentados em seguida, ver Hamilton, 1994). Embora seja possivelmente controverso chamar "clássicos" a estes modelos, optei por denominá-los assim para melhor expressar a separação que existe entre os dois grupos de modelos que serão aplicados. O segundo grupo que será apresentado contém os modelos que têm como base o *machine learning*. Os dois novos modelos aqui considerados são o modelo ANN e o modelo SVM (*Support Vector Machine*) – para mais detalhes ver Lantz (2015). Também será usado um modelo econométrico de regressão múltipla, que será apresentado no final desta secção.

3.1. MEDIDA DE QUALIDADE DAS PREVISÕES

O indicador que será usado como base de comparação entre as previsões produzidas pelos modelos será o *root mean square error* (RMSE), ou seja, raiz quadrada do erro quadrático médio, que pode ser calculado da seguinte forma:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (1)$$

Na equação 1, *n* é o número de previsões, y_i é o valor efectivo e $\hat{y}_i$ é o valor previsto.

3.2. Modelos ARIMA

O elemento fundamental dos modelos ARIMA é o ruído branco (que será representado por ε), o qual apresenta as características seguintes:

$$E\ (\varepsilon_t) = 0 \tag{2}$$

$$\mathrm{V}(\varepsilon_t^2)\ =\ \sigma_\varepsilon^2 \tag{3}$$

$$Cov(\varepsilon_t, \varepsilon_s) = 0, \qquad s \neq t \tag{4}$$

Por palavras, a esperança é zero, a variância é constante e as autocorrelações são sempre nulas.

3.2.1 Média móvel

Um processo de média móvel de ordem *q* pode ser escrito da seguinte forma:

$$Y_t = \ \mu + \ \varepsilon_t + \ \theta_1\varepsilon_{t-1} + \ \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \tag{5}$$

Consequentemente,

$$E\ (Y_t) = \ \mu \tag{6}$$

$$V(Y_t) = E\ [(Y_t - \ \mu)^2] = (1 + \theta_1^2\ + \theta_2^2 + \cdots + \theta_q^2)\sigma_\varepsilon^2 \tag{7}$$

3.2.2. Processo autorregressivo

Um modelo autorregressivo expressa o comportamento de uma variável como função do seu comportamento passado. Um processo autorregressivo de ordem *p* pode ser escrito da seguinte forma:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \quad (8)$$

A esperança de Y_t será

$$E(Y_t) = \frac{c}{1 - \phi_1 - \phi_2 - \cdots - \phi_p} \quad (9)$$

A variância será uma função mais complexa dos parâmetros.

3.2.3 Processo ARMA

O processo ARMA (*p*, *q*) agrega os dois processos anteriores, os processos AR (*p*) e MA (*q*). Temos então a seguinte formulação:

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (10)$$

A esperança continua a ser dada pela equação 9.

O processo ARMA está na base do processo ARIMA (*p*, *d*, *q*), o qual não é mais do que um processo ARMA (*p*, *q*) para a *d*-ésima diferença da variável Y_t.

3.3. Modelo VAR

Os modelos VAR são bastante populares e têm uma grande tradição em estudos econométricos, concretamente em previsões económicas e estudos macroeconómicos. A

popularidade deste tipo de modelos advém do facto de ser um modelo linear e de fácil utilização, que permite incorporar informação proveniente de um conjunto de variáveis na modelação da dinâmica temporal. A ideia por detrás de um modelo VAR é que o valor corrente de cada variável incluída no VAR pode ser explicado pelos valores passados de todas variáveis que estão a ser consideradas no estudo e não apenas da própria variável. Um modelo VAR pode ser apresentado da seguinte forma:

$$\mathbf{y}_{\mathrm{t}} = \boldsymbol{\nu} + \boldsymbol{\gamma}_1 \mathbf{y}_{\mathrm{t}-1} + \boldsymbol{\gamma}_2 \mathbf{y}_{\mathrm{t}-2} + \cdots + \boldsymbol{\gamma}_{\mathrm{p}} \mathbf{y}_{\mathrm{t}-\mathrm{p}} + \boldsymbol{\varepsilon}_{\mathrm{t}} \quad (11)$$

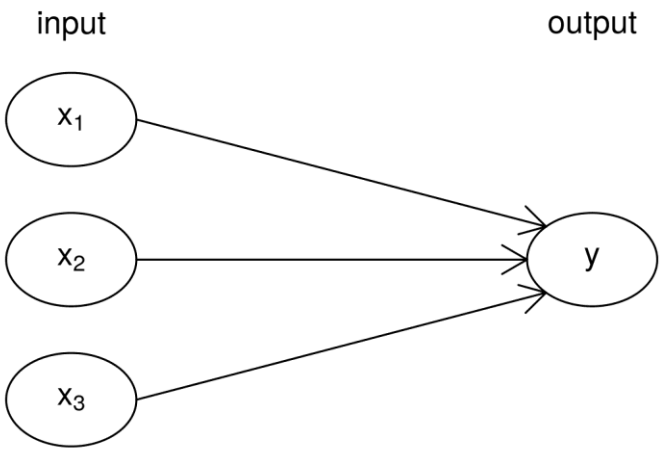
Na equação 11, $\boldsymbol{y_t}$ representa uma matriz coluna de dimensão K X 1 da variável dependente com desfasamentos de ordem *p*, o símbolo $\boldsymbol{\nu}$ representa uma matriz coluna com as constantes de dimensão K X 1, os elementos $\boldsymbol{\gamma_i}$ representam matrizes de dimensão K X K com os coeficientes associados a cada desfasamento, e finalmente o termo $\boldsymbol{\varepsilon_t}$ é uma matriz de dimensão K X 1 com os termos de erro (por hipótese, um vetor com as propriedades do ruído branco).

3.4. Rede Neural Artificial

Os modelos de Redes Neurais Artificiais (ANN) são modelos matemáticos, dentro do paradigma do *machine learning*, que procuram ter semelhanças com as redes neurais que existem no cérebro humano. A versão mais simples encontra-se representada na Figura 1. No modelo representado nessa figura, há uma camada (*layer*) com três *inputs* (x_1, x_2, x_3) que são combinados para dar origem a um *output* (*y*, mas note-se que a camada de *output* também pode ter mais do que um elemento). Cada linha representa uma sinapse entre os neurónios no modelo. Quem estiver a construir o modelo terá de escolher os *inputs* (a informação observável usada no modelo) e a função (de "ativação") que transforma a combinação linear desses *inputs* no *output*. O algoritmo de otimização procurará os valores dos parâmetros dessa combinação linear para os quais o ajustamento do modelo aos dados é o melhor. Algumas das funções de ativação mais comuns são a função degrau (*threshold*), a função linear, a função sigmóide, a função densidade da distribuição normal

estandardizada e a função ReLU (*Rectified Linear Unit*, que toma o valor do argumento quando este é positivo e o valor zero no caso contrário). Em face da natureza dos dados que irei analisar (variáveis contínuas) e do tipo de modelo ANN que será utilizado (discutido a seguir), usarei a função sigmóide e a função ReLU.

FIGURA 1: Rede neural artificial com uma camada de *input* e outra de *output*

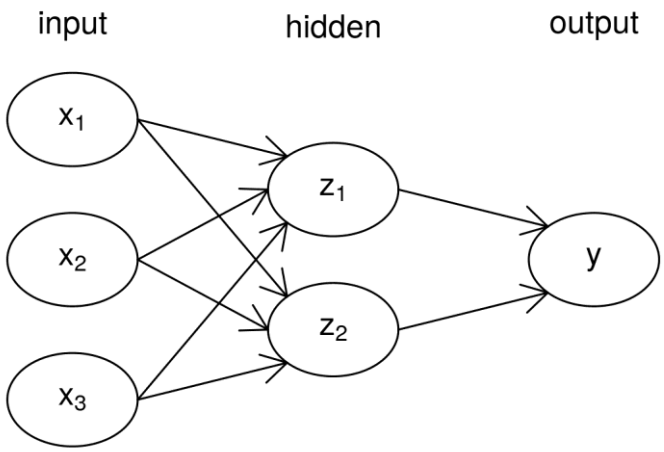


Na Figura 2 está representado um modelo que também inclui uma camada oculta (*hidden*) com dois nós, ou "neurónios", (z_1, z_2). Este é um exemplo de um *Multilayer Perceptron* (MLP), que é o modelo mais comum dentro desta classe e que será o modelo usado neste trabalho. No entanto, há modelos mais complexos, nomeadamente modelos que envolvem a transmissão de informação na direção oposta à representada na Figura 2, isto é, de camadas mais próximas dos *outputs* para camadas mais próximas dos *inputs* (modelos com *feedback* em vez de meramente *feedforward*).

Um resultado muito importante no contexto dos modelos ANN é que uma rede neural com uma camada oculta com um número suficiente de neurónios pode ser usada para construir uma função arbitrariamente próxima de uma função contínua num certo intervalo. Por um lado, este resultado incentiva a construção de modelos mais complexos, pois serão melhores aproximações à função que originou os dados. Por outro lado, os modelos mais

complexos estão mais sujeitos ao problema do *overfitting*, isto é, são otimizados (na fase de "treino do modelo") para reproduzir muito bem os dados históricos, mas com isso perdem validade fora da amostra usada para essa otimização (na fase de "teste do modelo").

FIGURA 2: Rede neural artificial com uma camada oculta

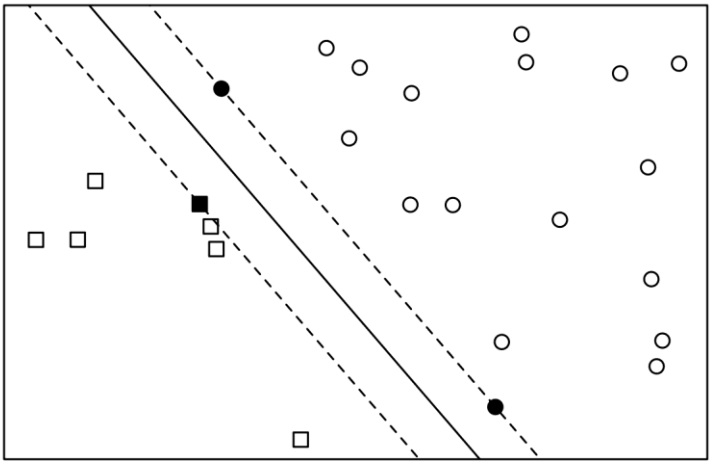


Os modelos com base em redes neurais têm sido utilizados em vários campos de investigação, tais como engenharia, ciências informáticas, ciências económicas, etc. Um exemplo é o estudo efetuado por Ekonomou (2010) relativo ao consumo de energia no território grego. Outro exemplo, para a previsão do preço do petróleo em vários países, é o estudo efetuado por Movagharnejad et al. (2011).

3.5. MODELOS SVM

Quando a variável a prever é discreta, ou seja, quando o problema é de classificação, a ideia fundamental do modelo SVM é a de encontrar uma linha (ou de forma mais geral, um hiperplano) que separe as regiões correspondentes às várias classificações. Esta ideia está representada na Figura 3.

FIGURA 3: SVM num problema de classificação



Na Figura 3 está identificada a linha que separa os círculos dos quadrados. Essa linha maximiza a distância entre as linhas paralelas que passam pelos pontos preenchidos, os quais correspondem aos "vetores de suporte". Encontrar essa linha é o objetivo do SVM quando aplicado a um problema de classificação (neste exemplo, classificar como círculo ou como quadrado). Nem sempre os círculos podem ser separados perfeitamente dos quadrados através de funções lineares. Nesse caso, o SVM pode dar uma "tolerância" a infrações, permitindo que alguns pontos fiquem mal classificados, ou pode ser possível aplicar uma transformação (não linear) aos dados que produza um conjunto de valores que possa ser classificado de forma mais correta. É aqui que entra a escolha da função *kernel*, que será a função que, em certo sentido, corresponderá a essa transformação dos dados.

O Gretl permite usar como função *kernel* uma função linear, uma função polinomial, a *radial basis function* (RBF) e a função sigmóide. A função *kernel* do tipo linear corresponde ao produto interno:

$$K(u; v) = u'v \tag{12}$$

A função *kernel* polinomial acrescenta um termo constante e um expoente:

$$K(u; v) = (\gamma u' v + c)^d \quad (13)$$

A função RBF é:

$$RBF(u; v) = e^{-\gamma \|u-v\|^2}, \text{ com } \gamma > 0 \quad (14)$$

A função sigmóide é:

$$K(u; v) = \tanh(\gamma\, u' v + c) \quad (15)$$

Na equação 15, tanh é a tangente hiperbólica.

No caso deste trabalho, o problema não é de classificação, pois a variável a prever é contínua. Assim, será utilizado o modelo *support vector regression* (SVR) do género ε-SVR. Isto significa que os coeficientes do modelo de regressão serão escolhidos de forma a minimizar a norma desses coeficientes, impondo-se a restrição de que os erros estejam, em valor absoluto, abaixo de um certo limiar (ε). Novamente, o algoritmo dá uma tolerância a infrações dessa restrição, penalizando-as de forma linear.

3.6. Apresentação do modelo de regressão múltipla e dos dados

O modelo de regressão múltipla a ser estimado é o seguinte:

$$P_t = \alpha + \beta_1 P_{t-1} + \beta_2 PIB_t + \beta_3 ER_t + \beta_4 CI_t + \beta_5 IR_t + \beta_6 OR_t + \beta_7 OS_t + \beta_8 OP_t + \beta_9 OC_t + \beta_{10} GN_t + u_t \quad (16)$$

A notação na equação 16 é a seguinte: α é a constante; P_t representa o preço do barril de petróleo da *West Texas Intermediate* e P_{t-1} é o primeiro desfasamento; PIB é o produto interno bruto total dos países da OCDE (em volume); ER é uma taxa de câmbio efetiva do dólar americano; CI é o índice de preços no consumidor nos EUA; IR é a taxa de juro da Reserva Federal dos EUA; OR são as reservas de crude nos EUA; OS são as existências de crude nos EUA; OP é a produção de crude nos EUA; OC é o consumo de petróleo nos EUA e GN representa o preço do gás natural nos EUA. Este modelo foi escolhido por ser aquele que Ramyar e Kianfar (2017) usaram como termo de comparação no seu estudo sobre a utilização de ANN na previsão do preço do petróleo.

Enquanto Ramyar e Kianfar (2017) usaram dados anuais, neste trabalho de projeto foram utilizados dados mensais. Em relação ao preço do petróleo, foi selecionado o preço do barril do petróleo West Texas Intermediate (WTI), visto que este é considerado um preço de referência para compradores e vendedores no mercado norte-americano. Esta série foi obtida na página da internet da Federal Reserve Economic Data (FRED) do Banco da Reserva Federal de St. Louis. Nesta mesma página recolhemos os dados das seguinte variáveis: a taxa de câmbio efetiva do dólar, o índice de preços do consumidor nos EUA e a taxa de juro da Reserva Federal norte-americana. O produto interno bruto total dos membros da OCDE foi obtido na plataforma Thomson Reuters. Os dados desta variável naquela plataforma são trimestrais, pelo que foi necessário transformá-los em dados mensais através de interpolação linear. Esta variável reflete a atividade económica dos membros desta organização mundial e como tal deverá ser um indicador da procura mundial de petróleo, visto que os membros da OCDE são responsáveis por boa parte dessa procura de petróleo.

Da agência de U.S. Energy Information Administration foram retirados os dados relativos à produção de crude, ao consumo de petróleo, às existências de crude e ao preço do gás natural. Obtive na mesma fonte dados quanto às reservas de crude dos EUA. Porém estes dados são anuais, pelo que foi necessário aplicar o processo de interpolação linear de modo a transforma-los em dados mensais.

O período de análise das observações estendeu-se desde o início de janeiro de 1981 até dezembro de 2018, contabilizando um total de 457 observações para todas as variáveis com a exceção das reservas de petróleo, relativamente às quais só foi possível encontrar dados até final de 2017, com isto contabilizando um total de 446 observações. O período

utilizado como *pseudo-in-sample* (para a estimação dos parâmetros dos modelos) foi o período compreendido entre 1981:12 e 2009:12. O período *pseudo-out-of-sample* para o teste dos modelos (avaliação das previsões) foi definido como o intervalo de 2010:1 até ao final da amostra. A Tabela 1 resume a informação sobre as variáveis usadas nos modelos.

TABELA 1: VARIÁVEIS USADAS NA PREVISÃO

SIGLA	VARIÁVEL	MIN	MAX	UNIDADE
P	preço do barril de petróleo WTI	11,28	133,93	Dólares por barril
PIB	PIB total dos países da OCDE	20837	52064	Mil milhões de dólares
ER	Taxa de câmbio efetiva do USD	69,061	143,91	-
CI	Índice preço do consumidor	18,445	264,61	-
IR	Taxa de juro da Reserva Federal	0,07	19,10	-
OR	Reservas de crude	19121	39160	Milhões de barris
OS	Existências de crude	486,45	1230,1	Milhões de barris
OP	Produção de crude	119,21	370,79	Milhões de barris
OC	Consumo de crude	414,17	671,65	Milhões de barris
GN	Preço do gás natural	3,940	20,770	$\$/1000ft^2$

Fonte: cálculos do autor

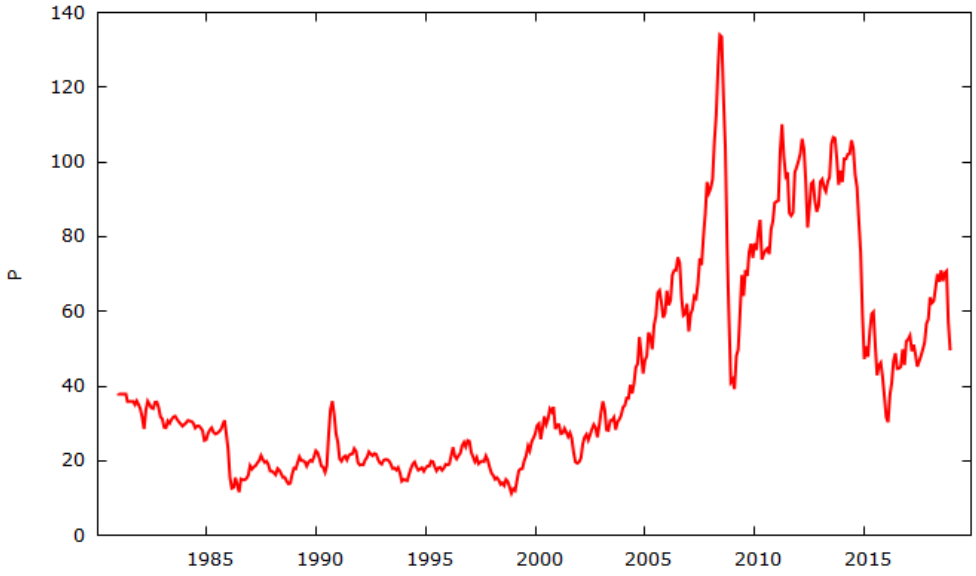
4. RESULTADOS EMPÍRICOS

Nesta secção serão apresentados os resultados das estimações empíricas dos modelos acima mencionados. Primeiramente é exibido o gráfico 1, onde está representado a evolução do preço do barril de petróleo ao longo da amostra, de janeiro de 1981 até dezembro de 2018.

Em janeiro de 1981, o barril de petróleo WTI foi transacionado a 38 dólares e este foi o valor mais alto observado nesse ano. Só a meio do ano de 1986 é que encontramos o

primeiro “choque” significativo no preço, quando o barril foi transacionado a uns meros 11,575 dólares; este foi também o valor mais baixo para o período em análise.

GRÁFICO 1: EVOLUÇÃO DO PREÇO DO BARRIL DA WTI



Fonte: FRED.

Como foi mencionado no início do texto, as flutuações do preço do petróleo são consequência das forças de mercado e de outros acontecimentos, como por exemplo conflitos armados ou eleições que tragam uma mudança de política, contribuindo de forma direta ou indireta para as variações do preço.

Assim sendo, as oscilações que o preço do barril de petróleo apresentou desde o início de 1981 até ao final do ano de 2018 podem ser associadas a vários acontecimentos que contribuíram para tais variações. Tendo em conta isto, torna-se relevante referir um evento que teve bastante relevância nos últimos anos: a crise que teve origem no mercado imobiliário norte-americano, que originou a última grande crise económica e financeira mundial, por vezes designada por crise das hipotecas norte-americanas (*subprime*), que levou a muitas falências e a um período de recessão económica não só nos EUA, mas em todo o globo. Até a crise entrar na sua fase mais aguda, na segunda metade de 2008, o preço do barril de petróleo foi subindo, atingindo o seu valor mais elevado em junho de 2008, a 133,93 dólares. Porém, dadas as vicissitudes que o mercado mundial atravessava, tal marco não se manteve e o preço registou a maior queda, chegando a 39,160 dólares em fevereiro de 2009. No entanto, após esta descida, o preço voltou a crescer de modo

sustentado. Este crescimento apenas foi contrariado por pequenas oscilações, que não impediram que se viesse a verificar, em junho de 2014, o valor mais alto nos últimos cinco anos, com o barril a ser transacionado a 105,790 dólares. Estas subidas e descidas acentuadas dificultam a tarefa dos modelos de previsão.

A Tabela 2 apresenta o valor-p dos testes ADF aplicados às variáveis na base de dados. A hipótese nula de raiz unitária nunca é rejeitada (ao nível de significância de 10%) para o nível das variáveis, exceto no caso da taxa de juro (IR). Para a primeira diferença, a hipótese nula de raiz unitária é sempre rejeitada ao nível de significância de 10%, exceto no caso da produção de crude (OP). Portanto, o teste ADF sugere que as variáveis são integradas de ordem um, exceto possivelmente a taxa de juro, que poderá ser estacionária em nível.

TABELA 2: VALOR-P DO TESTE ADF

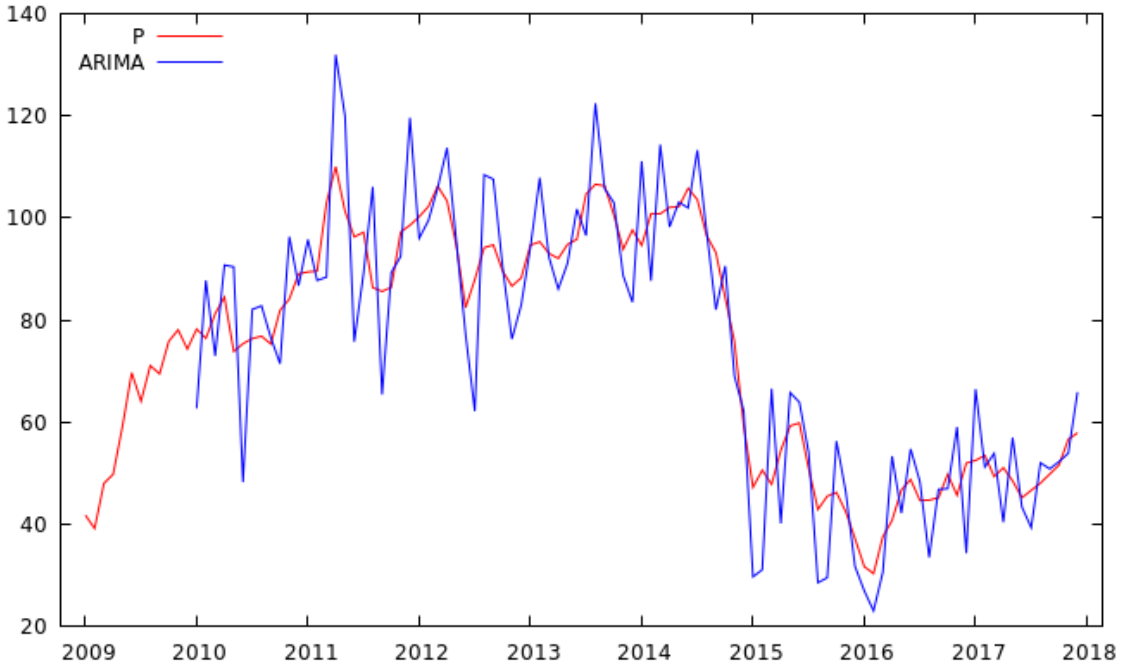
Variável	Nível		Diferença	
	CT	C	C	NC
P	0,327	0,424	0,000	0,000
PIB	0,243	0,965	0,000	0,059
ER	0,588	0,412	0,000	0,000
CI	0,334	0,576	0,000	0,000
IR	0,001	0,073	0,000	0,000
OR	0,997	0,955	0,090	0,011
OS	0,116	0,240	0,000	0,000
OP	1,000	0,976	0,172	0,026
OC	0,485	0,442	0,000	0,000
GNP	0,640	0,733	0,000	0,000

Notas: "CT" – teste com constante e tendência. "C" – teste com constante. "NC" – teste sem constante.

Fonte: cálculos do autor

Deste modo, estimei um modelo ARIMA (com o software Gretl) impondo uma raiz unitária no preço do petróleo. O modelo selecionado pelo critério de informação de Akaike foi o ARIMA (3,1,1). O Gráfico 2 mostra as previsões recursivas (com um horizonte temporal de um período) para o período de janeiro de 2010 até dezembro de 2017. As previsões acompanham a evolução geral do preço do petróleo, mas com oscilações acentuadas em torno da tendência.

GRÁFICO 2: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO ARIMA



Fonte: cálculos do autor

Para o modelo VAR também recorremos ao critério de Akaike para determinar o número de desfasamentos. Uma vez que o número da variável no modelo é grande, limitei o número de desfasamentos a três (um trimestre). O critério de Akaike selecionou um modelo VAR com três desfasamentos, enquanto os critérios de Schwarz e de Hannan-Quinn selecionaram o modelo VAR com dois desfasamentos – ver Tabela 3.

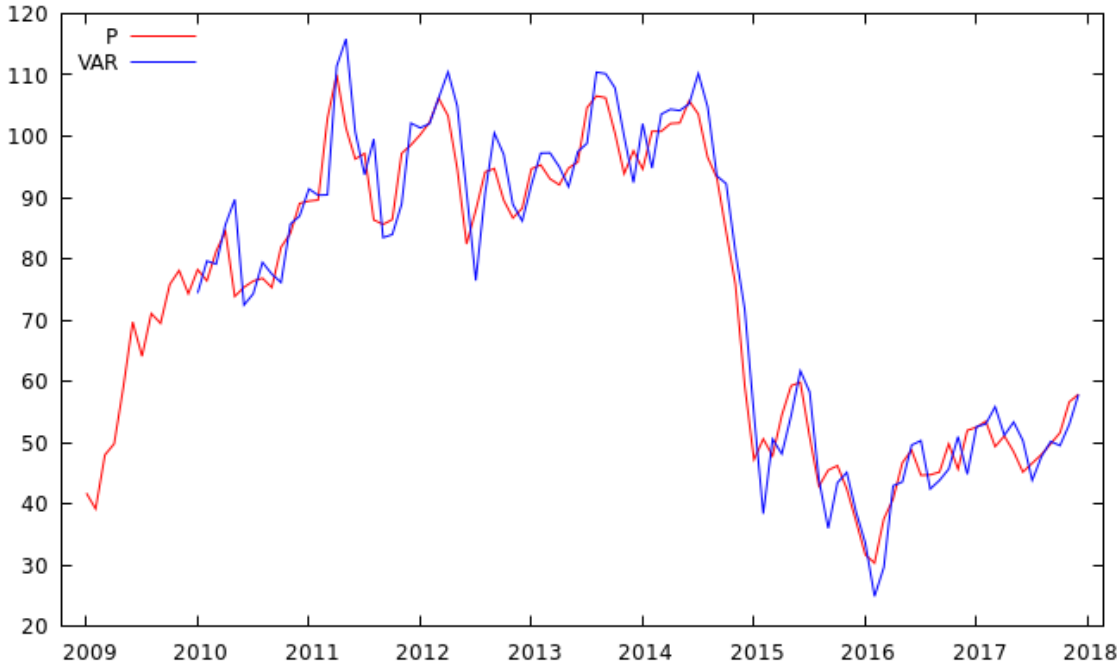
TABELA 3: SELEÇÃO DO NÚMERO DE DESFASAMENTOS NO MODELO VAR

Ordem	Akaike	Schwarz	Hannan-Quinn
1	119,528917	120,54886	119,931248
2	113,784357	115.731522*	114.552444*
3	113.591585*	116,465971	114,725428

Fonte: cálculos do autor

O Gráfico 3 mostra como se comportaram as previsões recursivas, também com um horizonte temporal de um período, produzidas pelo modelo VAR para o período *pseudo-out-sample*. Tal como as previsões do ARIMA, as previsões do VAR conseguem acompanhar a tendência geral, mas aparentam desviar-se menos dessa tendência do que as previsões do ARIMA.

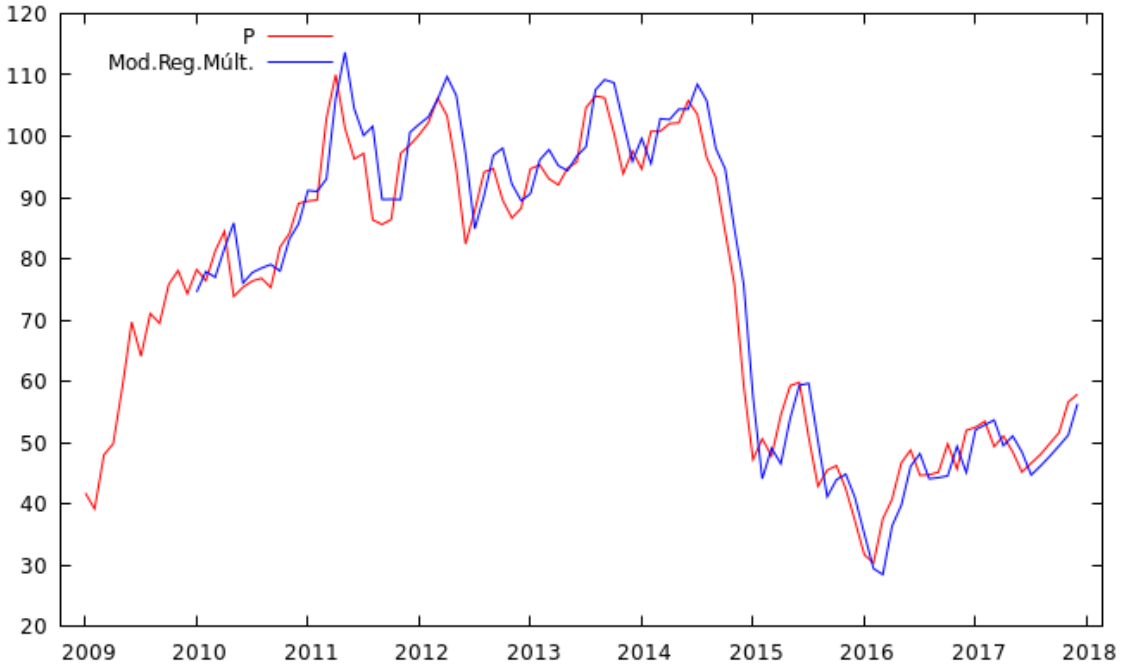
GRÁFICO 3: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO VAR



Fonte: cálculos do autor

O modelo de regressão múltipla sugerido por Ramyar e Kianfar (2017) tem a vantagem de usar informação corrente para prever o valor do preço do petróleo. Os parâmetros estimados estão no Anexo, juntamente com todas as outras estatísticas produzidas pelo Gretl na estimação. O Gráfico 4 mostra as previsões recursivas com um horizonte temporal de um mês obtidas com este modelo. Curiosamente, neste caso as previsões parecem demorar mais tempo a seguir a tendência do preço do petróleo, mas os desvios em relação à tendência também parecem ser mais pequenos.

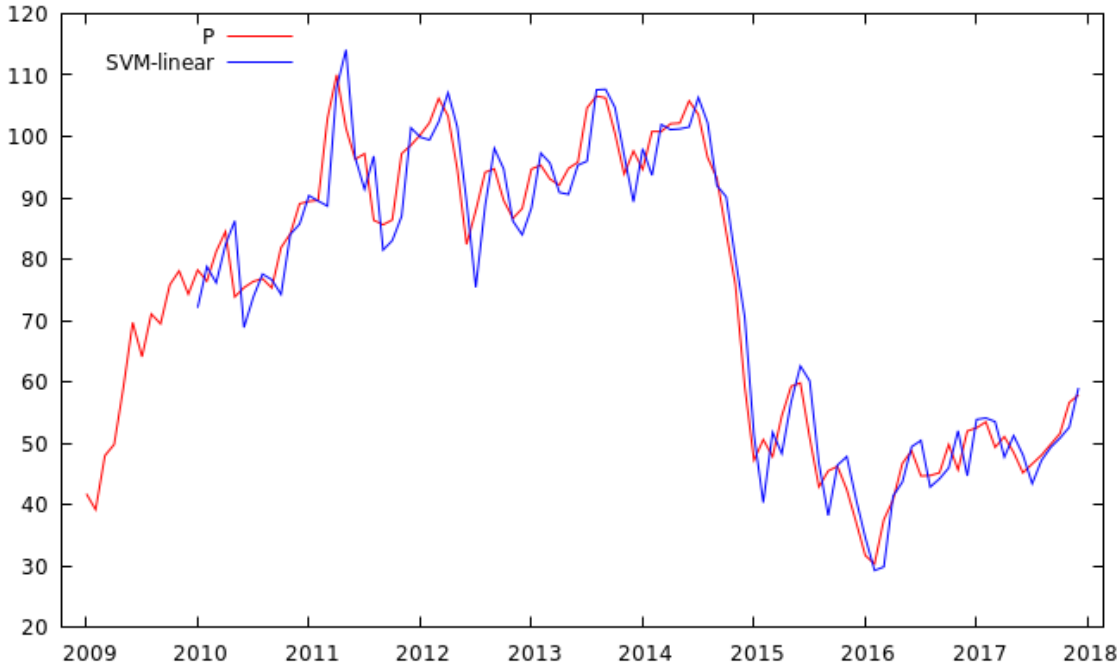
GRÁFICO 4: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO DE REGRESSÃO MÚLTIPLA



Fonte: cálculos do autor

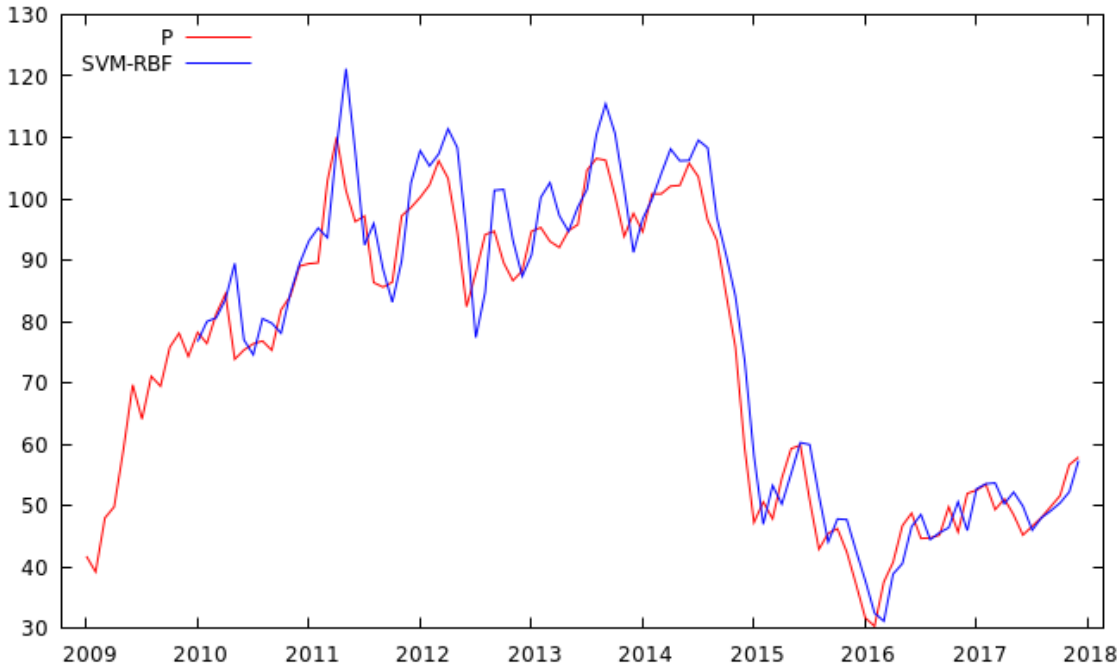
Na classe de modelos SVM, utilizei três funções kernel: linear, RBF e sigmóide. Também experimentei a função kernel polinomial, mas, tanto com grau dois como com grau três, os resultados foram insatisfatórios, não merecendo ser apresentados neste trabalho. Os Gráficos 5, 6 e 7 mostram as previsões obtidas com aquelas funções kernel. Como características a serem usadas na previsão usei os primeiros quatro desfasamentos do preço do petróleo, pois também é isso que o modelo ARIMA (3,1,1) faz. Algo inesperadamente, a função kernel linear parece ter resultado em previsões mais próximas dos valores efetivos do preço do petróleo. Ou seja, as funções não lineares originaram previsões de qualidade inferior.

GRÁFICO 5: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO SVM-LINEAR



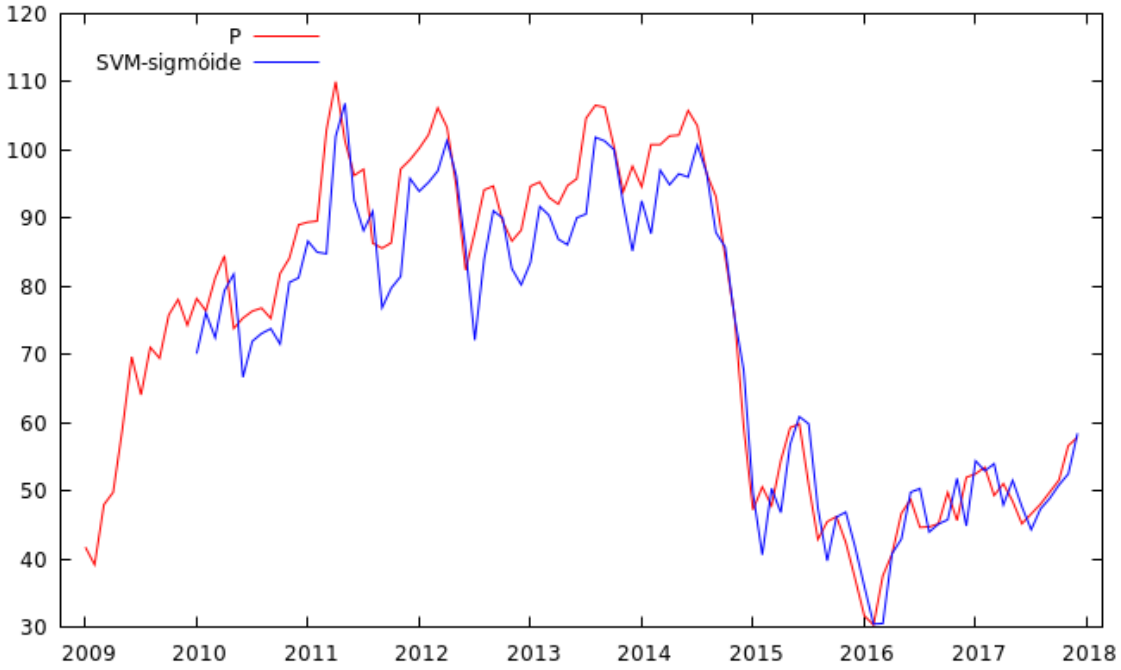
Fonte: cálculos do autor

GRÁFICO 6: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO SVM-RBF



Fonte: cálculos do autor

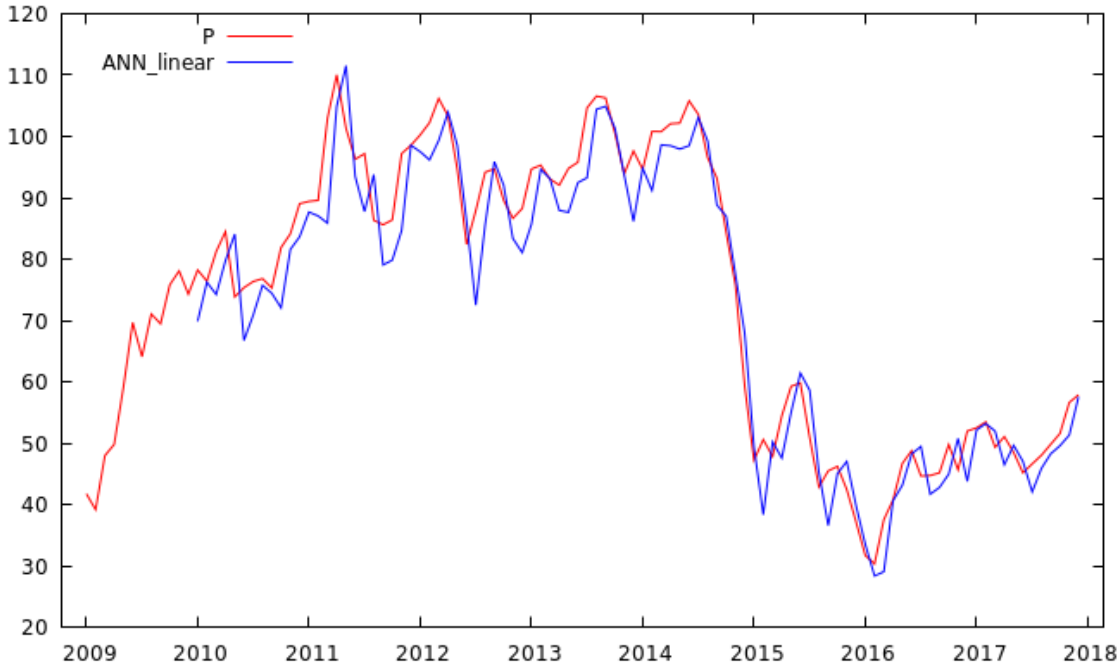
GRÁFICO 7: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO SVM-SIGMÓIDE



Fonte: cálculos do autor

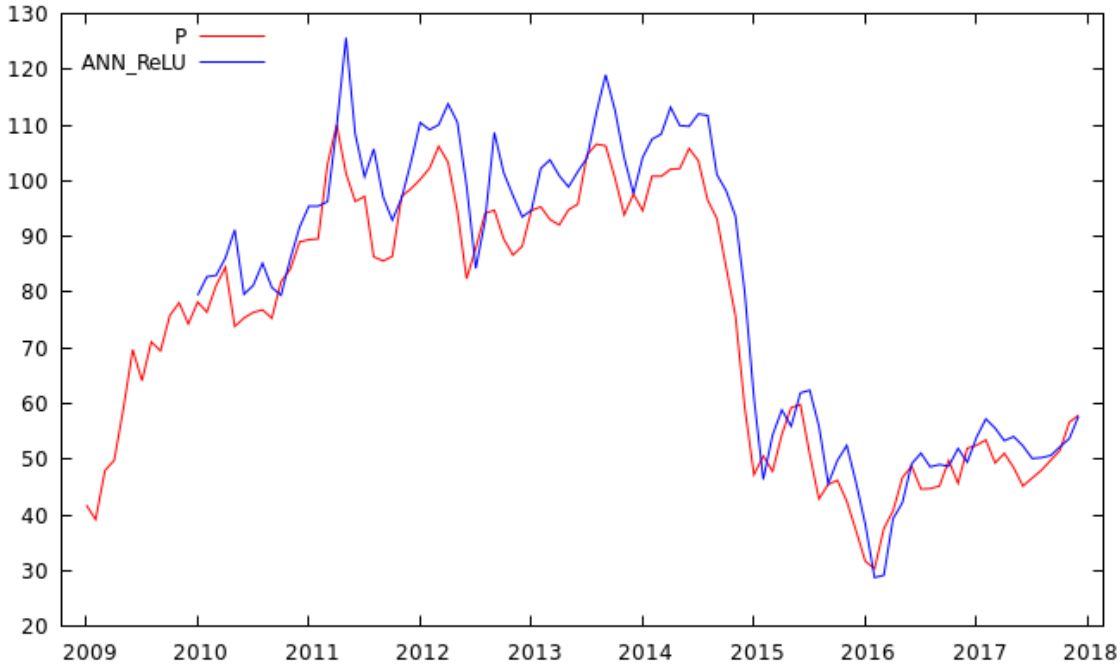
Finalmente, na classe de modelos ANN usei um modelo com funções de ativação linear (que dispensa a camada oculta), um modelo com uma camada oculta com funções de ativação ReLU e outro com funções de ativação sigmóide. As previsões obtidas com estes três modelos estão nos Gráficos 8, 9 e 10. Para estes cálculos fiz uso do pacote de instruções Keras (com Tensorflow) no software estatístico R. Tal como nos modelos SVM, as variáveis explicativas foram os primeiros quatro desfasamentos do preço do petróleo. Para obter bons resultados, foi necessário normalizar as variáveis, subtraindo a média do preço do petróleo na amostra e dividindo pelo seu desvio padrão. As previsões nos Gráficos 8, 9 e 10 resultam da aplicação da transformação inversa às previsões que o programa estatístico calculou. Novamente como no caso dos modelos SVM, a versão que usa uma função linear parece dar resultados melhores.

GRÁFICO 8: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO ANN-LINEAR



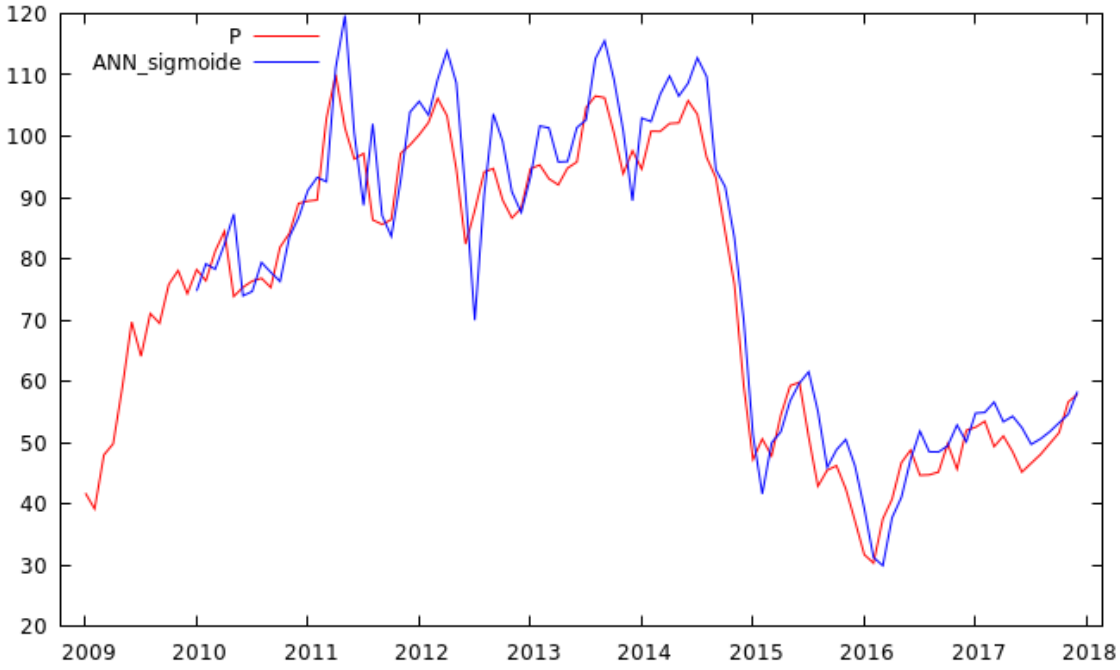
Fonte: cálculos do autor

GRÁFICO 9: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO ANN-RELU



Fonte: cálculos do autor

GRÁFICO 10: PREÇO DO PETRÓLEO (P) E PREVISÃO COM UM MODELO ANN-SIGMÓIDE



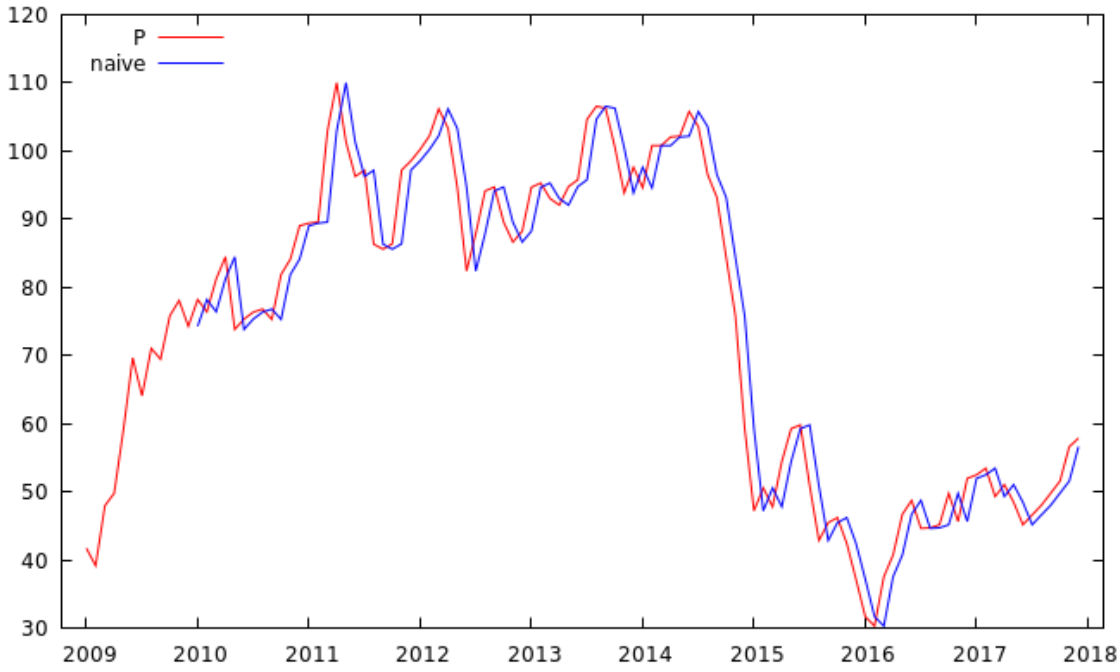
Fonte: cálculos do autor

Os modelos apresentados até aqui implicaram a escolha de formas funcionais e a estimação de parâmetros. Seria de esperar que tal esforço desse melhores resultados do que abordagens mais simples. Comparemos então as previsões daqueles modelos com as previsões *naive*, ou seja, previsões que consistem no último valor observado da série que estamos a prever. Estas previsões são adequadas quando a série segue um passeio aleatório, hipótese que os resultados dos testes ADF (Tabela 2) não invalidam. As previsões *naive* para o preço do petróleo estão no Gráfico 11.

Até agora a comparação entre as previsões foi baseada apenas na análise dos gráficos respectivos. Devemos, no entanto, analisar também algumas medidas habituais de avaliação da qualidade das previsões. Estas estatísticas estão na Tabela 4. Os modelos ANN tendem a ser bastante enviesados (o erro médio afasta-se de zero), tal como os modelos SVM com excepção do SVM-linear. A previsão *naive* é a menos enviesada. Contudo, o modelo com o menor RMSE é o SVM-linear, embora apenas bata a previsão naive por uma

pequena margem. O modelo ARIMA apresenta claramente o pior desempenho de acordo com o RMSE (e com os critérios que se seguem). O MAE penaliza menos os erros de maior magnitude. Neste caso a previsão *naive* é ligeiramente melhor do que o modelo SVM-linear. Se olharmos para os erros relativos (MAPE e U de Theil), o SVM-linear volta a ser (por pouco) o modelo com as melhores previsões.

GRÁFICO 11: PREÇO DO PETRÓLEO (P) E PREVISÕES NAIVE



Fonte: cálculos do autor

Tabela 4: Avaliação das previsões

Fonte: cálculos do autor

	ME	RMSE	MAE	MAPE	U
ARIMA	0,33	10,8	8,67	12,73	2,08
Reg.Múlt.	-1,32	5,70	4,42	6,46	1,07
VAR	-0,93	5,59	4,36	6,35	1,07
SVM-linear	0,31	5,22	4,09	5,95	0,99
SVM-RBF	-2,47	6,32	4,93	6,91	1,08
SVN-sigmóide	3,15	6,39	5,07	6,84	1,11
ANN-linear	2,51	5,84	4,55	6,40	1,10
ANN-ReLU	-5,77	8,47	6,71	9,42	1,43
ANN-sigmóide	-2,85	6,64	5,28	7,59	1,20
Naive	-0,17	5,29	4,05	6,01	1,00

Notas: ME- erro médio. RMSE- raiz quadrada do erro quadrático médio. MAE- erro absoluto médio. MAPE- erro percentual absoluto médio. U- U de Theil.

5. CONCLUSÃO

Este trabalho teve como objetivo efetuar uma comparação entre modelos de previsão do preço do petróleo WTI, que é o *benchmark* no mercado norte-americano. Usei modelos baseados em informação sobre os determinantes fundamentais do preço do petróleo (incluindo medidas da produção, da procura e das existências de petróleo), nomeadamente modelos VAR e de regressão múltipla, e modelos que apenas usam a informação contida nos valores passados do preço do petróleo. Entre estes últimos modelos estão o modelo ARIMA tradicional e modelos que se estão a tornar populares mais recentemente: SVM e ANN. No entanto, nenhum dos modelos utilizados demonstrou ser claramente preferível ao modelo mais simples: o modelo de previsão *naive*, que não exige qualquer trabalho de construção e estimação do modelo. Apenas o modelo SVM com uma função kernel linear apresentou um desempenho semelhante ao do modelo *naive*. Estas conclusões aplicam-se quer se avaliem as previsões usando o RMSE, o MAE, o MAPE ou o U de Theil.

Assim sendo, a principal conclusão é que a previsão do preço do petróleo é extremamente difícil. Tal resultará do facto de o preço do petróleo ser influenciado por muitos fatores, por vezes relacionados com comportamentos mais de natureza política (interna e externa) do que económica. Por outro lado, sendo o preço de um ativo, não é de estranhar que o seu comportamento se aproxime de um passeio aleatório, tal como acontece com os preços de outros ativos, como os preços das ações cotadas na bolsa ou a taxa de câmbio.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Almoguera, P. A., Douglas, C. C., & Herrera, A. M. (2011). Testing for the cartel in OPEC: non-cooperative collusion or just non-cooperative?, Oxford Review of Economic Policy, 27(1), 144-168.

Baumeister, C., & Kilian, L. (2012). Real-time forecasts of the real price of oil. *Journal of Business & Economic Statistics*, 30(2), 326–336.

Berwick, R. (2003). An Idiot's guide to Support vector machines (SVMs). Consultado em 24/06/2019, https://web.cs.dal.ca/~tt/CSCI415009/Berwick03.pdf.

Brown, S. P., & Huntington, H. G. (2017). OPEC and world oil security. Energy Policy, 108, 512-523.

Ekonomou, L. (2010). Greek long-term energy consumption prediction using artificial neural networks. Energy, 35(2), 512-517.

Godarzi, A. A., Amiri, R. M., Talaei, A., & Jamasb, T. (2014). Predicting oil price movements: A dynamic Artificial Neural Network approach. Energy Policy, 68, 371-382.

Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *Journal of political economy*, 91(2), 228-248.

Hamilton, J. D. (1994). Time series analysis. Princeton University Press.

Hamilton, J. D. (2008). Understanding crude oil prices (No. w14492). National Bureau of Economic Research.

International Energy Agency, 2018, Market Report Series: Oil 2018: Analysis and Forecasts to 2023. Consultado em 24/06/2019, https://webstore.iea.org/market-report-series-oil-2018-pdf

Lantz, B. (2015). *Machine Learning with R*, 2ª ed. Birmingham: Packt Publishing Ltd.

Movagharnejad, K., Mehdizadeh, B., Banihashemi, M., & Kordkheili, M. S. (2011). Forecasting the differences between various commercial oil prices in the Persian Gulf region by neural network. Energy, 36(7), 3979-3984.

Ramyar, S., & Kianfar, F. (2017). Forecasting crude oil prices: A comparison between artificial neural networks and vector autoregressive models. Computational Economics, 1-19.

Yu, L., Wang, S., & Lai, K. K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, 30(5), 2623-2635.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International journal of forecasting*, 14(1), 35-62.

ANEXO: **Output da estimação do modelo de regressão múltipla**

```
Model 1: OLS, using observations 1981:02-2017:12 (T = 443)
Dependent variable: P

             coefficient    std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       1,11984       5,86142        0,1911    0,8486
  PIB         8,72298e-09   1,05058e-07    0,08303   0,9339
  ER         -0,0226778     0,0229603     -0,9877    0,3239
  CI          0,0623282     0,0354801      1,757     0,0797    *
  IR          0,224844      0,114455       1,964     0,0501    *
  OR          3,16363e-06   0,000194150    0,01629   0,9870
  OS          2,67867e-06   4,07998e-06    0,6565    0,5118
  OP         -1,11303e-05   1,98503e-05   -0,5607    0,5753
  OC          4,82269e-07   8,46997e-06    0,05694   0,9546
  GNP         0,0131652     0,110957       0,1187    0,9056
  P_1         0,846511      0,0724406     11,69      1,33e-027 ***

Mean dependent var   41.58673   S.D. dependent var   27.97333
Sum squared resid    6829.140   S.E. of regression   3.975952
R-squared            0.980255   Adjusted R-squared   0.979798
F(10, 432)           2144.704   P-value(F)           0.000000
Log-likelihood      -1234.477   Akaike criterion     2490.955
Schwarz criterion    2535.984   Hannan-Quinn         2508.714
rho                  0.392476   Durbin-Watson        1.214967
```