Universidade Federal Fluminense
Alex Sandro Oliveira

Uma Análise do Mercado do Petróleo Utilizando Aprendizado de Máquina

Niterói
2016

Alex Sandro Oliveira

Uma Análise do Mercado do Petróleo Utilizando Aprendizado de Máquina

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense Trabalho de Conclusão de Curso como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Universidade Federal Fluminense

Orientador: Mariana Tasca Fontenelle Lôbo

Niterói

2016

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

O48 Oliveira, Alex Sandro
Uma análise do mercado do petróleo utilizando aprendizado de máquina / Alex Sandro Oliveira. – Niterói, RJ : [s.n.], 2016.
35 f.

Projeto Final (Tecnólogo em Sistemas de Computação) – Universidade Federal Fluminense, 2016.
Orientador: Mariana Tasca Fontenelle Lôbo.

1. Aprendizado de máquina. 2. Commodity. 3. Série temporal. 4. Petróleo. I. Título.

CDD 006.31

Alex Sandro Oliveira

Uma Análise do Mercado do Petróleo Utilizando Aprendizado de Máquina

Trabalho de Conclusão de Curso submetido ao Curso de Tecnologia em Sistemas de Computação da Universidade Federal Fluminense Trabalho de Conclusão de Curso como requisito parcial para obtenção do título de Tecnólogo em Sistemas de Computação.

Niterói, 24 de maio de 2016:

Mariana Tasca Fontenelle Lôbo, D.Sc.
Orientadora
UFF, Universidade Federal Fluminese

Julliany Sales Brandão, Dsc.
Avaliadora
CEFET-RJ, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca

Niterói
2016

Dedico este trabalho aos meus filhos,
Ana Laura e Lucas Matheus.

AGRADECIMENTOS

Agradeço a minha orientadora, Professora Mariana, pelo suporte, incentivo e orientações.
Aos meus pais, minha irmã, minha esposa e meus filhos Lucas Matheus e Ana Laura pelos esforços dedicados para que eu realizasse este trabalho.
A Deus pela oportunidade do convívio com estas pessoas.

RESUMO

O mercado financeiro é o local onde são negociados ativos financeiros. Para o investidor, o mercado financeiro apresenta problemas complexos relacionados à decisão de comprar ou vender um ativo. Isto se deve, entre outros fatores, à quantidade de variáveis que precisam ser analisadas e pela velocidade que é necessária para tomar cada decisão.

Pelo fato da velocidade de processamento dos computadores ser alta, algoritmos para análise do comportamento de ativos são interessantes. O ramo da estudo de aprendizagem de máquina possui algoritmos que são promissores no sentido de auxiliar o investidor na tomada de decisão na hora de comprar ou vender no mercado financeiro.

Neste trabalho, foi realizada uma análise pelo método de Box e Jenkins [1] da série temporal dos preços da cesta da OPEC, utilizando basicamente o software R. Posteriormente, foi realizada uma análise utilizando o modelo de mínimos quadrados do pacote de aprendizagem de máquina Sklearn. O objetivo é confrontar os resultados destas duas metodologias. O algoritmo de mínimos quadrados do sklearn mostrou resultados não satisfatórios tanto utilizando somente a série temporal com atrasos como utilizando variáveis ligadas à oferta e demanda. O método de Box e Jenkins mostrou resultados satisfatórios de curto prazo.

Palavras-chave: *commodity*, aprendizado de máquina, séries temporais

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

BM&FBovespa	Bolsa de Valores de São Paulo
CVM	Comissão de Valores Mobiliários
EIA	*U.S. Energy Information Administration*
OPEC	Organização dos Países Exportadores de Petróleo
OECD	Organização para Cooperação e Desenvolvimento Econômico
IEA	*International Energy Agency*

SUMÁRIO

1 INTRODUÇÃO

O Mercado do petróleo está inserido dentro da dinâmica do mercado financeiro internacional. O Mercado financeiro é um nome genérico para locais físicos ou virtuais em que são negociados ativos financeiros, participações em empresas, contratos agrícolas ou que envolvem áreas como energia, minerais, transporte, moedas, etc. No Brasil existe uma instituição reguladora, a CVM (Comissão de Valores Mobiliários), e um único centro de negociação virtual, a BM&FBovespa. Em países com um mercado financeiro desenvolvido é comum existir mais de um centro de negociação.

O mercado financeiro possui vários agentes, que podem ser classificados como[2]:

Hedgers: são aqueles que realizam uma ou mais operações para reduzir sua exposição ao risco.

Bons especuladores: são aqueles que aceitam o risco para obter lucro em operações com um volume financeiro pequeno.

Maus especuladores: são aqueles que aceitam assumir um risco maior do que o apropriado para seu patrimônio. Estas posições são grandes o suficiente para gerar um efeito em cadeia quebrando outros agentes.

Arbitradores: realizam mais de uma operação sem risco, por vezes em diferentes mercados, para auferir lucro.

Tesouraria: operadores de tesouraria que consistem em captar recursos a uma taxa e aplicá-los de maneira a obter lucro, compatibilizando o prazo de captação e aplicação.

Market makers: instituições altamente especializadas em determinadas ações que se comprometem a manter ofertas de compra e venda do ativo no mercado.

Manipuladores: alteram artificialmente o valor de um ativo com o objetivo de auferir lucro.

Inside Traders: participantes com informação privilegiada.

Fabricantes de resultados: são aqueles que utilizam o mercado para gerar lucro ou prejuízo artificiais.

Bancos Centrais atuam no mercado com o objetivo de por em prática políticas de Estado.

Existem dois tipos de análises amplamente conhecidas para avaliar o valor de um ativo financeiro. Análise **fundamentalista**, onde o objetivo é avaliar o valor intrínseco de cada ativo utilizando fatores econômicos, entre outros. Análise **técnica**, que faz uso de

indicadores que utilizam como principal variável de entrada o preço e o volume financeiro histórico negociado. A partir de séries temporais destas duas variáveis, busca-se padrões gráficos e indicadores que possam apontar o melhor momento para compra ou venda de um ativo.

Cada agente possui um objetivo e sua ação tem uma influência diferente no comportamento dos preços dos ativos. Alguns tomam a decisão de comprar ou vender baseada em indicadores fundamentalistas, outros em indicadores técnicos e alguns não são motivados por análises[1].

O mercado de *commodity* é um local físico ou virtual onde são negociados contratos que envolvem áreas como energia, produtos agrícolas, metais e alguns produtos mais sofisticados como contratos de afretamento de navios. Neste mercado são comercializados produtos como ouro, trigo, gás e petróleo. Existem mais de 50 mercados de *commodities* ao redor do mundo.

Os fatores que influenciam o preço do petróleo são derivados, em sua grande maioria, da oferta e demanda no mercado.

No lado da oferta, existe uma divisão básica entre os países que são exportadores do petróleo. Há países pertencentes e não pertencentes à OPEC (Organização de Países Exportadores de Petróleo). Segundo a EIA [3] (*U.S. Energy Information Administration*), os países pertencentes à OPEC respondem por mais de 30% da produção mundial do petróleo. Entres os membros da OPEC, a Arábia Saudita ocupa uma posição de destaque, sendo o maior país produtor. A produção dos países membros da OPEC sofre uma influência grande das decisões nas assembléias da OPEC. Nos países não membros da OPEC, as decisões que influenciam a oferta decorrem de fatores econômicos e financeiros.

No lado da demanda, existe uma divisão entre países pertencentes à OECD (Organização para Cooperação e Desenvolvimento Econômico), na sua guande maioria países desenvolvidos, e países não pertencentes à OECD. O consumo do petróleo nos países pertencentes à OECD corresponde a quase 53% do consumo mundial. Segundo IEA[4] (*International Energy Agency*) o consumo de petróleo nos países pertencentes à OECD é maior que nos países não pertencentes, porém o crescimento do consumo ano a ano é menor em realação aos países não pertencentes à OECD.

Variações no consumo e na demanda afetam diretamente o mercado do petróleo orientando, inclusive, outras variáveis mais subjetivas, como a expectativa do preço futuro.

É razoável utilizar como premissa que o objetivo do investidor seja prever o valor futuro de um ativo utilizando esta informação para obter o maior lucro com o menor risco. Mas na formulação de um modelo matemático que possua um grau de incerteza aceitável, talvez a primeira pergunta a ser respondida esteja relacionada ao tipo de comportamento

[1] Fabricantes de Resultados podem ser motivados somente por questões fiscais.

matemático que os preços dos ativos possuem. Eles são determinísticos ou estocásticos? A resposta a este questionamento fornece um guia para realizar uma abordagem mais adequada ao problema de previsão apresentado.

Os autores de [5, 6] estudaram a Teoria dos Mercados Eficientes, a qual considera que os preços refletem todas as informações disponíveis e os agentes compreendem toda a informação tomando a melhor decisão a partir delas. Foram estudadas as hipóteses de mercados eficientes considerando três grupos de informações: *strong-form*, informações relevantes para formação de preço de ativos; *semi-strong-form*, informações óbvias que são públicas e *weak-form*, que consideram somente o histórico dos preços. A abordagem utilizando a Teoria dos Mercados Eficientes faz uso modelos estocásticos.

Mais recentemente, o trabalho apresentado em [7] analisa a possibilidade do índice *Dow Jones Industrial Avarage* ter um comportamento caótico ou de passeio aleatório. Foram encontradas pequenas evidências contra a teoria que a série possui um sistema gerador caótico de pequena dimensão. Sistemas caóticos são determinísticos e um passeio aleatórios é um processo estocástico. Este resultado não direciona de maneira rígida a abordagem, considerando séries temporais similares à estudada, para modelos estocásticos. Porém, com visão flexível e pragmática o modelo estocástico é o mais adequado.

Aprendizado de máquina é um subcampo da inteligência artificial que emergiu do estudo teoria do aprendizado de máquina e reconhecimento de padrões. Pela sua característica de resolver problemas complexos, como a previsão dos preços de ativos financeiros, suas técnicas vem sendo estudadas para este propósito.

Foi realizada, uma análise da série temporal dos preços da cesta do petróleo da OPEC pelo método de Box e Jenkins [1] e uma análise utilizando o modelo de mínimos do pacote Sklearn. O objetivo central é confrotar estas metodologias.

Este trabalho está organizado da seguinte maneira. No primeiro cápitulo é apresentada a motivação e o trabalho em linhas gerais. No segundo capítulo é apresentada uma breve descrição do mercado de *commodities*. No terceiro capítulo é apresentada a teoria de aprendizado de máquina. No quarto capítulo é apresentada a metodologia de Box e Jenkins. No quinto capítulo é descrito o processo de coleta e tratamento dos dados. No sexto capítulo são apresentados os resultados. O sétimo capítulo é reservado para as conclusões.

2 APRENDIZADO DE MÁQUINA

Mas a definição mais citada para aprendizado de máquina vem de Tom M. Mitchell: "Um programa de computador é dito aprender de uma experiência E com respeito a alguma classe de problema T e performance P se sua performance em T, mensurada por P, melhora com a experiência E"[8, p. 1].

O aprendizado de máquina pode ser separado em três tipos [9]

1. Aprendizado supervisionado – O sistema aprende utilizando um conjunto de dados. O objetivo é adquirir uma habilidade de mapear variáveis de entradas e variáveis de saídas ou alvo. Após a etapa de aprendizagem, o sistema deve ser capaz de fornecer valores de saída para dados de entradas que não estavam na base de dados utilizada durante o processo de treinamento.

2. Aprendizado não supervisionado – O sistema busca uma representação dos dados de entrada de uma maneira que reflita sua estrutura. Ao contrário do aprendizado supervisionado e do aprendizado por reforço, não existe nenhum tipo de avaliação dos dados de saída com o alvo ou o retorno fornecido pela base de dados para que o sistema possa comparar com a saída gerada do aprendizado já adquirido.

3. Aprendizado por reforço – Da mesma maneira que o aprendizado supervisionado, o objetivo é aprender utilizando um conjunto de dados. A diferença é no processo de treinamento, nele o sistema não sabe exatamente se uma saída está correta ou não – como acontece durante o processo de aprendizagem supervisionado – o que ele recebe é uma resposta do nível de exatidão da saída.

Entre as abordagens em problemas de aprendizagem de máquinas existe [9, 10]:

1. Classificação – Inclusa na categoria de aprendizado supervisionado. O objetivo é classificar os dados de entrada em subconjuntos conhecidos. Neste caso os valores alvo, considerados os valores das variáveis de saída do sistema real, pertencem a um domínio finito e não ordenado.

2. Ranking – Inclusa na categoria de aprendizado supervisionado, quando os valores alvo pertencem a um domínio finito ordenado.

3. Regressão – Inclusa na categoria de aprendizado supervisionado, quando os valores alvo pertencem a um domínio infinito ordenado.

4. Clustering – Este é um tipo de aprendizagem não supervisionada. Parecido com a classificação, exceto pelo fato que as saídas pertencem a um domínio desconhecido, porém finito. O objetivo é classificar os dados de entrada em subconjuntos (chamados *clusters*) de acordo com uma medida de similaridade.

A seguir, são apresentados alguns algoritmos de regressão.

2.1 Problemas de Regressão

Dado um conjunto de dados para treinamento contendo N amostras da variável aleatória X como entrada e um conjunto de mesmo tamanho de variáveis alvo Y correspondente aos valores de X. Utilizando a notação X_n para referenciar o conjunto de variáveis de entrada e Y_n para referenciar o conjunto de variáveis objetivo, o problema de regressão constitui-se em prever o valor de Y para um novo valor de X.

A seguir são apresentados alguns métodos para realizar esta tarefa e o método de regularização de Tikhonov, utilizado para evitar ploblemas de *overfitting*.

2.1.1 Técnica de Regularização

Um problema comum em aprendizado de máquina é *overfitting*. Ele acontece quando o modelo tenta capturar uma quantidade de comportamentos da variável alvo que resulta em uma piora da performance do modelo.

Uma das técnicas utilizadas como solução para *overfitting* é a regularização. Como exemplo, será aplicada a regularização a regressão de mínimos quadrados.

Na expressão 2.1, x é uma matriz com amostras de variáveis de entrada, A um vetor linha para expressar uma interação linear das variáveis em x com y, sendo y um vetor de dados de saída ou variável alvo. O objetivo é encontrar A para no passo posterior poder realizar previsões de y dado um novo x. Portanto, considerando a equação 2.2 como solução para o problema de encontrar A, é necessário que x^{-1} exista. Assim, x não dever ser singular. Esta é a origem da necessidade de x não ter colienariedade.

$$xA = y \tag{2.1}$$

$$A = x^{-1}y \tag{2.2}$$

Conhecendo o valor de A, o passo seguinte é realizar previsões dado um novo x'. Nesta etapa é comum surgir uma diferença, chamada de ϵ, entre o alvo observado y' e o previsto $\hat{y}$.

$$x'A = \hat{y} \tag{2.3}$$

$$\hat{y} - y' = \epsilon \tag{2.4}$$

$$||x'A - y'||^2 = ||\epsilon||^2 \tag{2.5}$$

O método de regularização de Tikhonov consiste em achar funções f que minimizem a equação 2.6

$$V(f(x'), y') - \lambda||f||^2, \tag{2.6}$$

onde $\lambda > 0$ é um parametro dado e V uma função dada.

Este método é utilizado em situações que o problema não é bem-posto. Um problema bem-posto, pela definição de Jacques Hadamard [11], tem as seguintes características:

- Existe uma solução
- A solução é única
- O comportamento da solução varia continuamente com as condições iniciais.

Aplicando 2.6 em 2.3, tem-se 2.7. Neste caso, a função f é o vetor A e V é $||x'A - y'||^2$.

$$||x'A - y'||^2 - \lambda||A||^2 \tag{2.7}$$

2.1.2 Mínimos quadrados

A técnica de mínimos quadrados consiste em resolver a equação 2.8.

$$min_A \left\|XA - y\right\|_2^2 \tag{2.8}$$

onde, A é o vetor de parâmetros, X é a matriz de dados de entrada e y o vetor das variáveis alvo.

O modelo 2.8 não é regularizado. Assim, é um pressuposto que X não tenha colunariedade colinear.

2.1.3 Ridge

O método de regressão do ridge é uma derivação do método de mínimos quadrados. O método de ridge faz uso do coeficiente de ridge que penaliza o uso de parêmtros A com valores grandes.

$$min_A \left\| XA - y \right\|_2^2 + \alpha ||A||_2^2 \tag{2.9}$$

O parâmetro α é dado. O termo $\alpha ||A||_2^2$ é considerado um regularizador.

3 MÉTODO BOX E JENKINS

A metodologia de Box e Jenkins [1] para análise de séries temporais está baseda em 4 passos:

1. Identificação do modelo apropriado – MA, AR ou ARMA.
2. Estimação dos parâmetros do modelo.
3. Verificação do modelo segundo sua adequação através de análise de resíduos.
4. Previsão de valores futuros.

O objetivo desta metodologia é definir o processo gerador dos dados. É considerada uma premissa que a série temporal é oriunda de um processo estocástico.

Definição 1. Processo estocástico: Um processo estocástico Y é uma coleção de variáveis aleatórias $(Y_t)_{t\in T}$ definidas em um espaço Ω e indexadas por um subconjunto T de $\mathcal{R}$. Sendo o indexador T o conjunto $[0, \infty)$, para auxílio no tratamento de séries temporais.

Definição 2. Ruído Branco: Seja $\{\epsilon_t\}$ uma coleção de variáveis aleatórias indenticamente distribuídas e independentes (iid) com $\mu = 0$ e σ^2_ϵ. $\{\epsilon\}$ é considerado um ruído branco e possui as seguintes caracteríticas.

1. $\epsilon_t \sim iid$
2. $E[\epsilon_t] = \mu$
3. $Var[\epsilon_t] = \sigma^2_\epsilon$
4. $Cov[\epsilon_t, \epsilon_{t+k}] = 0 \quad \forall \quad k \neq 0$

Definição 3. Estacionariedade: Um processo estocástico é dito estacionário quando é invariante no tempo. Neste caso, a função distribuição de y_t é idêntica independente da defasagem no tempo.

Um processo estocástico possui uma estacionariedade de segunda ordem quando

1. $E[y_t] = \mu$ é constante.
2. $Cov[y_t, y_{t+k}] = \gamma_k$ é função somente de k.

Definição 4. Série temporal: Uma série temporal Y de tamanho t, chamada Y^t, é uma coleção de variáveis y_t indexadas por um conjunto $T \in \mathcal{R}_+^*$ tal que

$$Y^t = \{y_1, y_2, \ldots, y_t\} \tag{3.1}$$

O modelo proposto é da forma

$$y_t = \mu + \sum_{k=0}^{\infty} \Psi_k u_{t-k} = \Psi(L) u_t \tag{3.2}$$

Considerando $y_t - \mu = \tilde{y}_t$,

$$\tilde{y}_t = \Psi(L) u_t. \tag{3.3}$$

Ψ é um filtro linear definido como

$$\Psi(L) = \frac{\Theta(L)}{\Phi(L)} \tag{3.4}$$

u_t é considerado um ruído branco. Um ruído branco entra em um filtro linear que gera a série temporal.

$$\Phi_p(L)\tilde{y}_t = \Theta_q(L)\epsilon_t \tag{3.5}$$

L é considerado o operador *Lag* que introduz um atraso nos dados. Assim, discriminando 3.5, tem-se.

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \tag{3.6}$$

O modelo matemático 3.6 é chamado ARMA(p, q).

4 DADOS

4.1 Coleta de Dados

O preço da mistura teórica de petróleo da OPEC e em qual sentido será a sua variação são as variáveis que representam o papel de alvo durante o trabalho. A série temporal, utilizada neste trabalho, dos preços da mistura teórica da OPEC foi retirada do *site* da OPEC em 05 de novembro de 2016 (ver Gráfico 1).

A série do sentido da variação temporal do preço e o percentual de variação é apresentada no Gráfico 2. A variação é tratada como uma série de diferenças definida na equação 4.1.

$$Dif[D+1] = \frac{Preço[D+1] - Preço[D]}{Preço[D]} \tag{4.1}$$

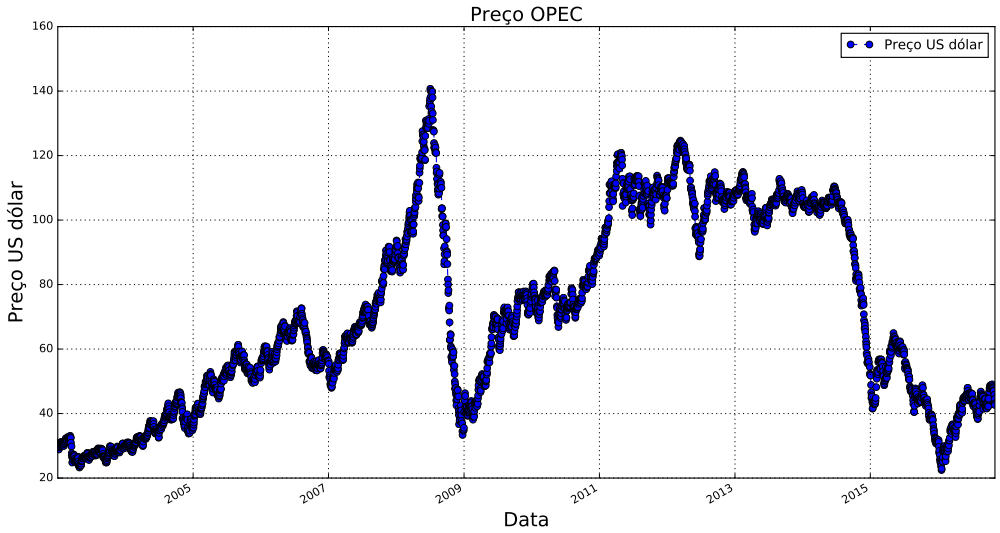
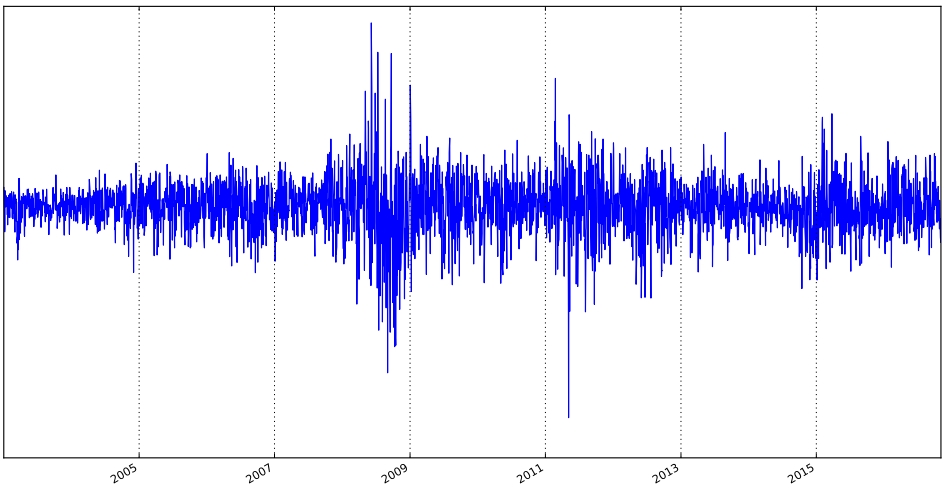


Gráfico 1 – Preço da cesta do petróleo da OPEC. Cotação diária de fechamento. Início em 02/01/2003 término 05/01/2016.

Em alguns momentos da análise dos resultados apresentados pelos algoritmos foram utilizadas como variáveis independentes dados da produção e demanda (ver Gráfico 3). A extração destes dados foi realizada no dia 27 de novembro de 2016 do *site* da EIA. Os dados são uma variação percentual dos variáveis de ano para ano.

Entre os dados de produção, tem-se a **capacidade produtiva** e a **produção da Arábia Saudita**. A capacidade produtiva dos países membros da OPEC é definida pela



2015

Gráfico 2 – Série temporal da primeira diferença – definida na equação 4.1 – aplicada a série do preço cesta OPEC – Gráfico 1. Início em 03/01/2003 teŕmino 05/01/2016.

EIA como o volume produtivo que pode ser entregue em 30 dias e mantido por no mínimo 60 dias.

Entre os dados de consumo, foi utilizado o consumo de países pertencentes (**Consumo OECD**) e não pertencentes (**Consumo não OECD**) à OECD.

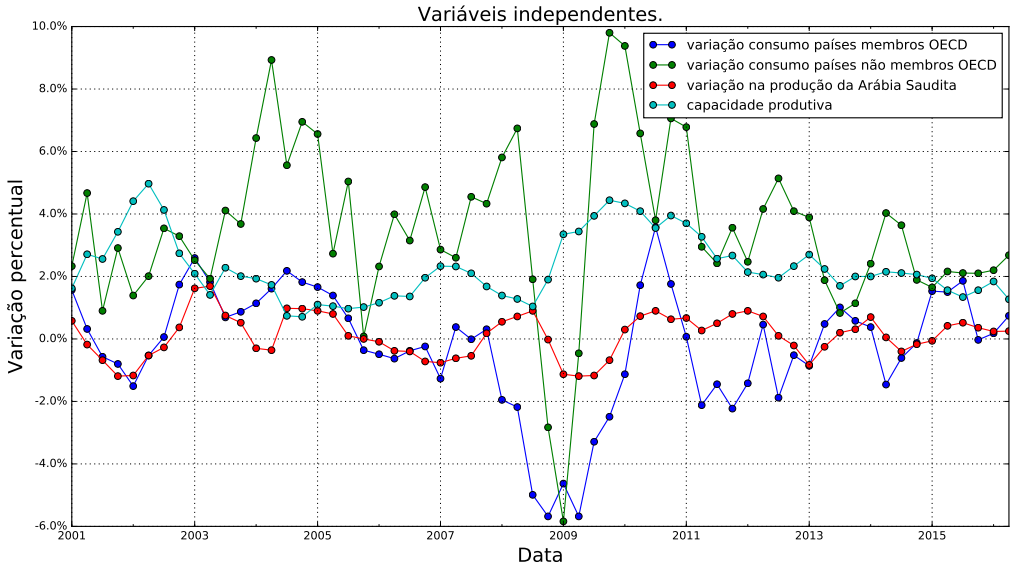


Gráfico 3 – Variação percentual ano a ano das variáveis de demanda e produção.

4.2 Descrição dos Dados

Na Tabela 1 é apresentada a correlação das variáveis independentes e o preço do petróleo OPEC.

Tabela 1 – Tabela de correlação.

	Preço OPEC	Consumo OECD	Consumo não OECD	Produção na Arábia Saudita	Capacidade produtiva
Preço OPEC	1,000000	-0,330185	0,070935	-0,005672	0,247008
Consumo OECD	-0,330185	1,000000	0,308926	0,466600	-0,241375
Consumo não OECD	0,070935	0,308926	1,000000	0,130094	0,209274
Produção na Arábia Saudita	-0,005672	0,466600	0,130094	1,000000	-0,279325
Capacidade produtiva	0,247008	-0,241375	0,209274	-0,279325	1,000000

É importante ressaltar o fato que a existência de correlação alta entre duas variáveis não implica em causalidade, porém as variáveis escolhidas são apontadas em estudos de EIA como fatores importantes na formação do preço do petróleo. Existe também o fato de a correlação não ser capaz de capturar a depêndencia entre duas variáveis, em alguns casos. Percebe-se que existe uma baixa correlação dos preços do petróleo com os valores de **Produção na Arábia Saudita** e o **Consumo de países não pertencentes** à OECD. Porém, estas duas variáveis possuem uma grande correlação com o **Consumo de países da OECD** e com a **Capacidade produtiva**. O **Consumo de países da OECD** e a **Capacidade produtiva** possuem grande correlação com os preços da OPEC.

As variáveis independentes de consumo e produção serão tratadas no decorrer do texto como uma matriz conforme o formato dado pela equação 4.2.

$$X^T = \begin{bmatrix} \text{Consumo OECD} \\ \text{Consumo não OECD} \\ \text{Produção na Arábia Saudita} \\ \text{Capacidade produtiva} \end{bmatrix} \tag{4.2}$$

5 RESULTADOS

5.1 Análise Linear Box e Jenkins

Neste capítulo é utilizado o método de Box e Jenkins para uma análise dos dados.

O *primeiro passo* na utilização do método é a identificação da necessidade de utilização de diferenças sucessivas na série para torná-la normal. A diferença de 1, por exemplo, deve ser entendida como uma série $\{y_1 - y_2, y_2 - y_3, \ldots, y_{t-1} - y_t\}$.

O número de diferenças corresponde ao parâmetro d no modelo ARIMA(p, d, q). O modelo ARIMA(p, d, q) é similar ao modelo ARMA(p, q) com a diferença que é aplicado d diferenças nos dados de entrada antes de um tratamento idêntico ao realizado no modelo ARMA(p, q)

Para uma verificação da estacionariedade da série, são verificados os Gráficos de variância 4 e média 5 no tempo. O cálculo foi realizado com amostras temporais de tamanho 50.

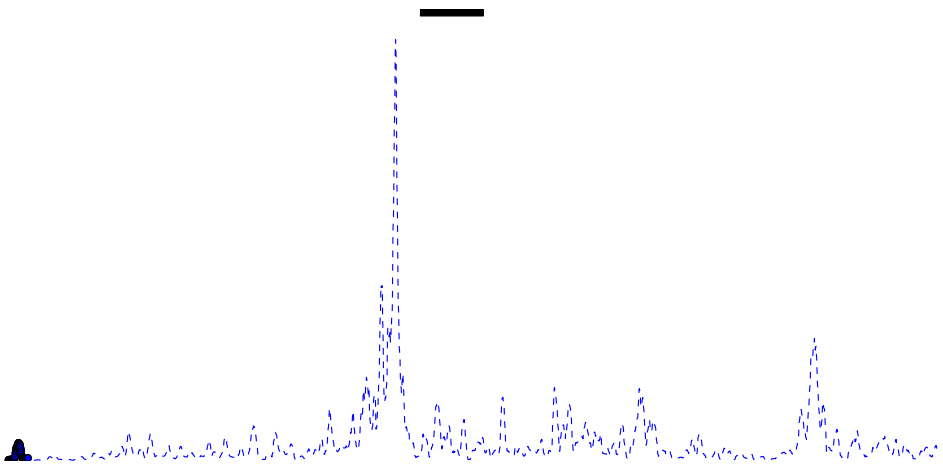


Gráfico 4 – Variância móvel dos preços da cesta do petróleo da OPEC calculada com amostras com período de 30 dias úteis.

De acordo com os Gráficos 4 e 5 é visível que a série não é estacionária.

Para reforçar a percepção de não estacionariedade da série é verificada a distribuição dos resíduos de um modelo ARIMA$(0, 0, 0)$. Assim, é verificada somente a distribuição de desvios-padrão em relação à média do resíduos.

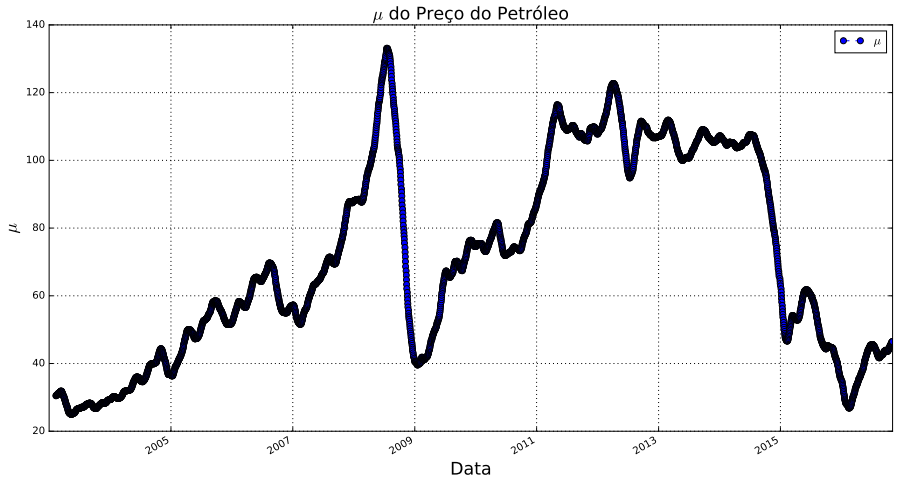


Gráfico 5 – Média dos preços da cesta do petróleo da OPEC calculada com amostras com período de 30 dias úteis..

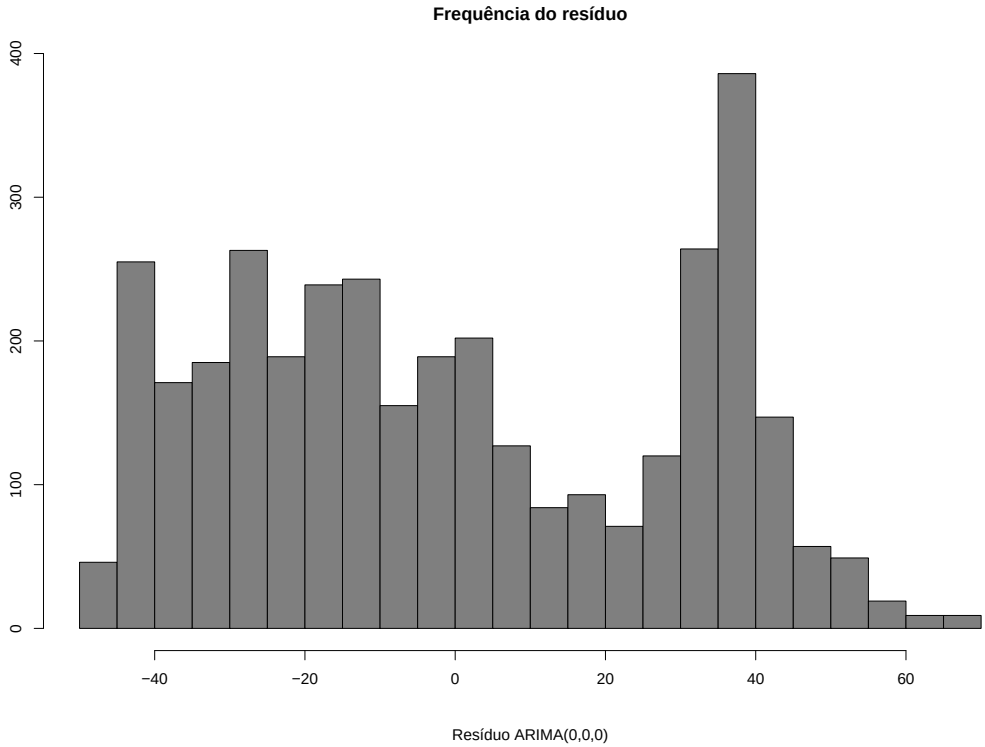
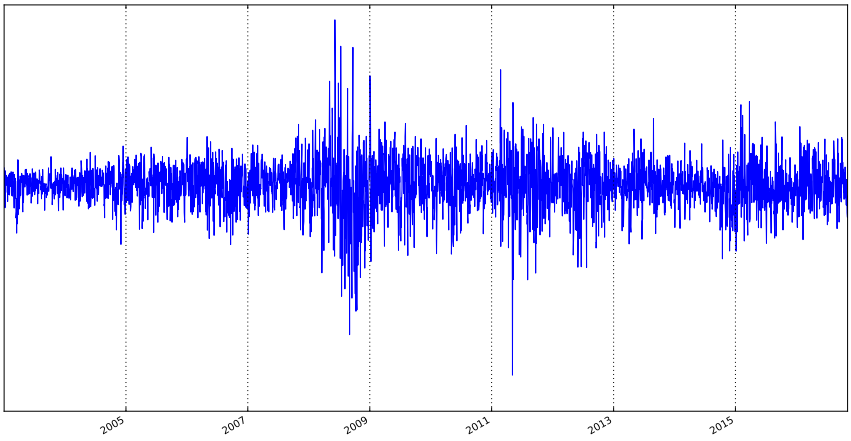


Gráfico 6 – Frequência dos resíduos no modelo ARIMA$(0, 0, 0)$. $\mu = 70,7481$ e $\sigma^2 = 858,137$.

Pelo Gráfico 6 percebe-se que a distribuição não é normal, possuindo $\mu \neq 0$.

Na metodologia de análise utilizada, é importante que os resíduos apresentem uma distribuição normal. Com o objetivo de conseguir resíduos com esta caracterítica, é realizada uma primeira diferença no modelo para análise da média, variância e distribuição de frequência.



2015

Gráfico 7 – Primeira diferença dos preços OPEC. $\mu = 0$ e $\sigma^2 = 1,1492$.

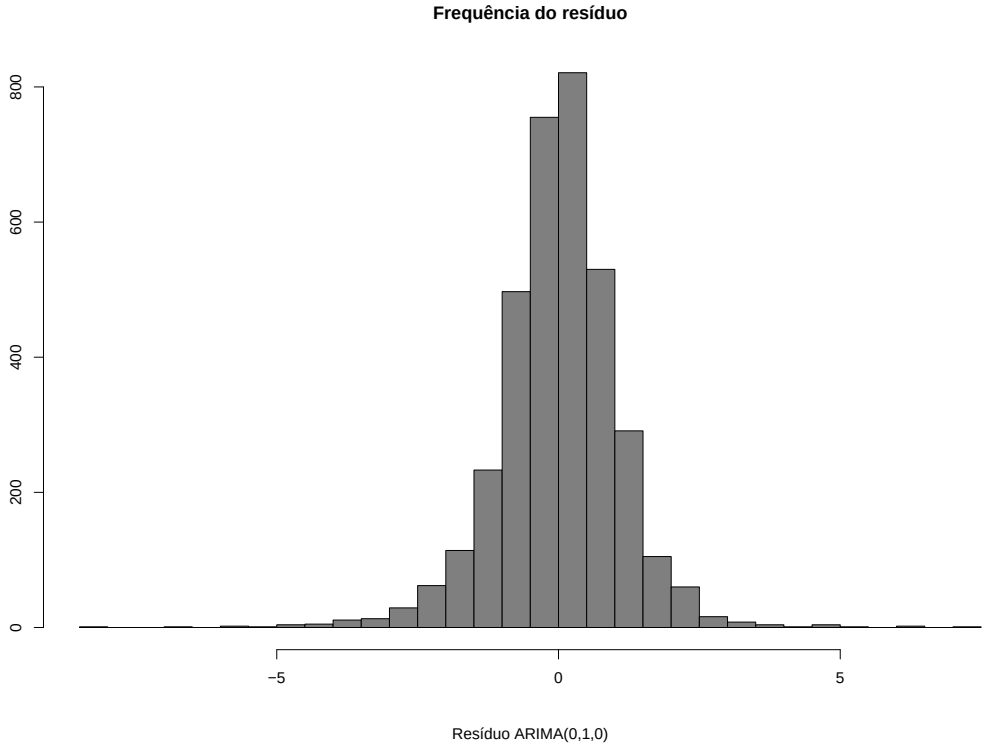


Gráfico 8 – Frequência dos resíduos no modelo ARIMA$(0, 1, 0)$. $\mu = 0$ e $\sigma^2 = 1,1492$.

Pelos Gráficos 7 e 8, percebe-se que a série temporal possui uma distribuição de frequências próxima da normal com média zero.

A partir deste ponto, é considerado razoável realizar a análise da série utilizando somente o modelo ARIMA(p, d, q) com $d = 1$.

O *segundo passo* é a estimativa dos parâmetros p e q no modelo ARIMA$(p, 1, q)$. Estes parâmetros são os mesmos da equação 3.5.

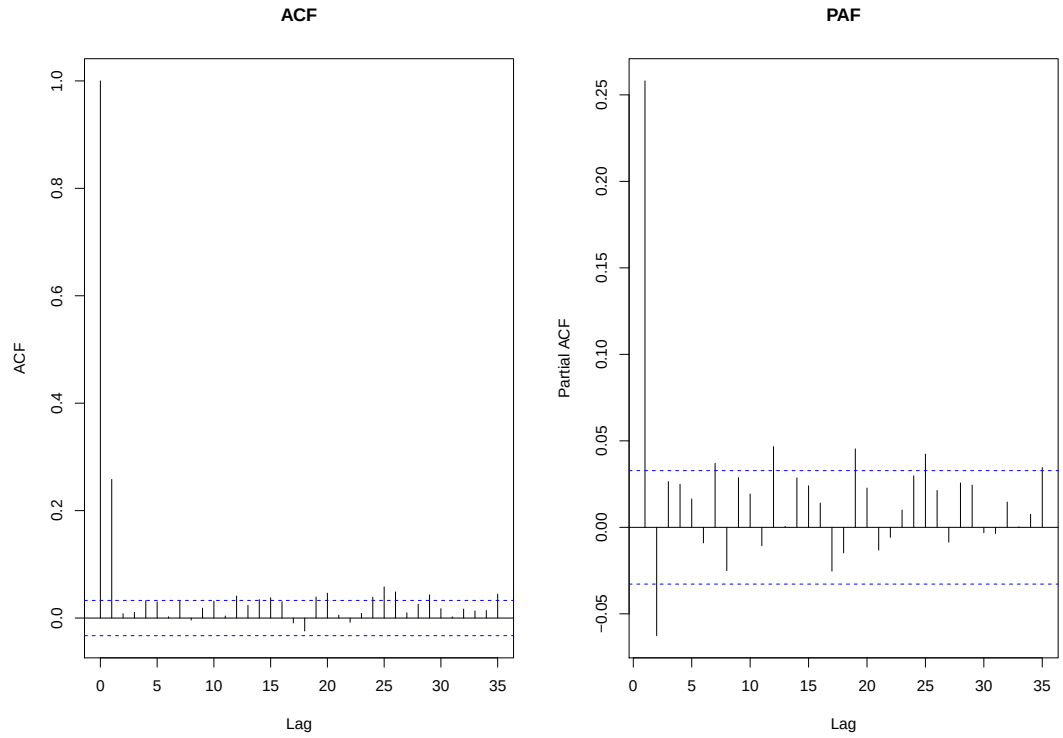


Gráfico 9 – FAC – Função de Auto Correlação, e PFAC – Função Parcial de Auto Correlação. Todas estas variáveis foram calculadas em relação a *lags* na série temporal.

O parâmetro p possui uma relação com *lags* que têm valores de PACF altos. O parâmetro q possui uma relação com *lags* que têm valores de ACF altos. Assim, foram escolhidos como parâmetros $p = 1$ e $q = 1$ e verificada a adequação modelo ARIMA$(1, 1, 1)$.

Como verificado no Gráfico 10, o resíduo possui uma distribuição próxima da normal. Na Tabela 2 é apresentado um resumo dos coeficientes do modelo. Na Tabela 3 são apresentadas a média e a variância.

Tabela 2 – Coeficientes modelo ARIMA$(1, 1, 1)$.

	ϕ_1	θ_1	μ
Valor	0,0263	0,2488	0,0053
σ^2	0,0610	0,590	0,0222

Tabela 3 – Média e variância do resíduo modelo ARIMA$(1, 1, 1)$.

	μ	σ^2
Valor	1,07	2,74

O *terceiro passo* é a realização de previsões. Para tanto, utiliza-se o modelo para a previsão dos preços de 03/11/2016 até 08/12/2016.

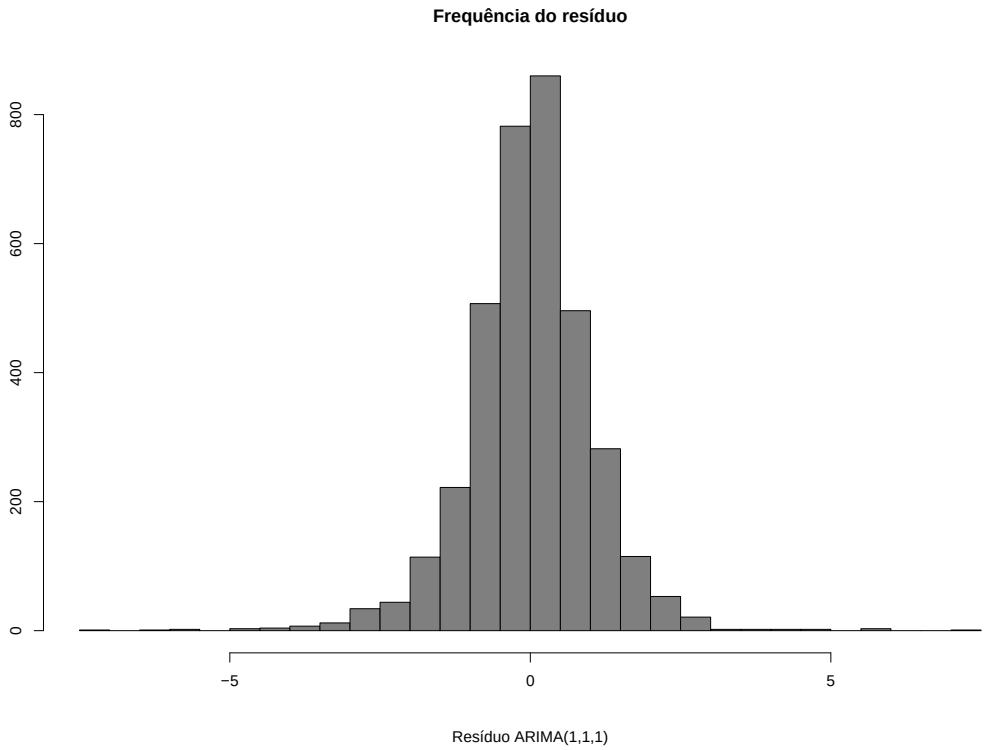


Gráfico 10 – Distribuição de frequências do resíduo do modelo ARIMA$(1, 1, 1)$.

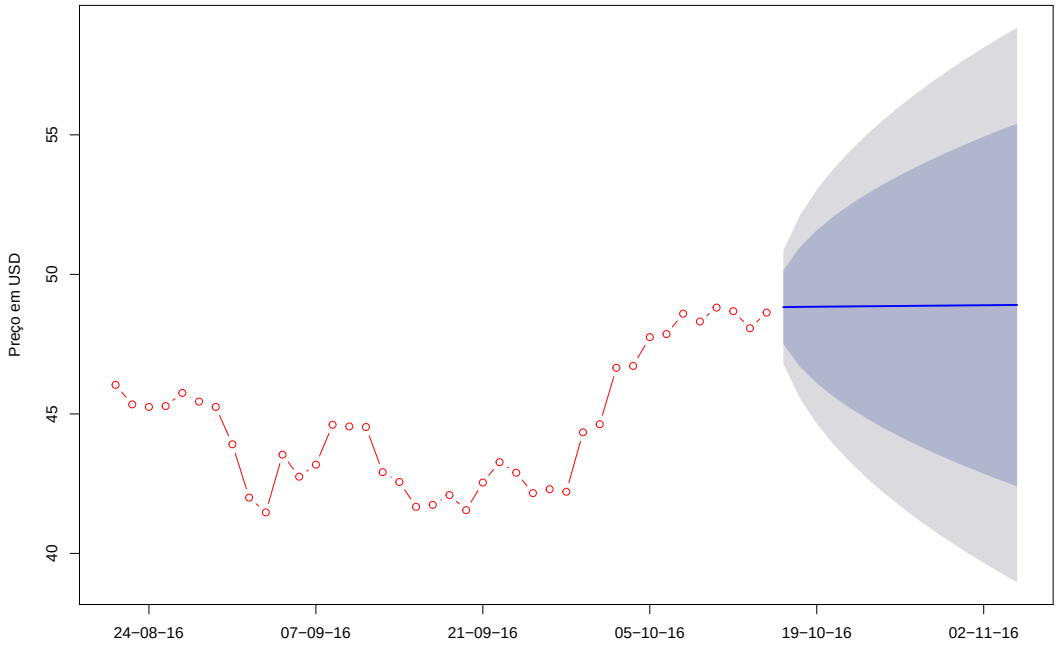


Gráfico 11 – Previsão dos preços de 03/11/2016 até 08/12/2016 realizada com o modelo ARIMA$(1, 1, 1)$. Em cinza escuro é a área do preços com um intervalo de confiaça de 95%. Em cinza claro é o intervalo de confiança com 80%. Os pontos em verde é o valor do preço realizado.

5.2 Análise com Sklearn

Na análise realizada na Seção 5.1 foi utilizado essencialmente o software R com os seus pacotes. Nesta seção é utilizado o pacote Sklearn [12, 13] em Python, com o intuito de confrontar os resultados da Seção 5.1.

Os modelos utilizados são os modelos gerais de aprendizado de máquina. Estes modelos são variações do modelo ARIMA.

Um primeiro cuidado tomado é no teste no modelo. Testar o algoritmo de aprendizado e treinar no mesmo conjunto de dados pode induzir a conclusões erradas. Os parâmetros do modelo matemático serão tais que apresentarão um erro otimizado, mas quando utilizados fora daquele conjunto de dados o algoritmo pode sofrer de *overfitting*. Para solucionar este problema, utiliza-se a técnica de validação cruzada, onde uma parte da amostra é separada somente para teste.

Entre as técnicas de validação existe o método *k-fold* onde uma amostra de tamanho n é separada em k subconjuntos. O algoritmo é treinado com os $k-1$ subconjuntos e treinado no subconjunto restante. O algoritmo repete a iteração até todos os k subconjuntos terem sido utilizados como teste. Este método é similar ao método da classe *TimeSeriesSplit* utilizado neste trabalho.

Em séries temporais, é necessária uma estratégia um pouco diferente no momento de realizar o teste e o treinamento. Deve-se evitar a utilização de dados com um passo de tempo a frente para treinamento. No *k-fold* os subconjuntos são escolhidos de maneira aleatória. No *timeseriesSplit* os dados são dividos retornando os k primeiros valores para uso durante o treinamento sendo o valor $k+1$ para teste. Este processo se repete aumentando o valor de k até $k+1=n$ sendo n o número de amostras. Para maiores detalhes consultar a documentação do pacote [14]. A Figura 1 apresenta um exemplo para uma amostra de tamanho 6.

Figura 1 – Exemplo de funcionamento do método de validação cruzada de tamanho 6. Considerar que as amostras estão em ordem temporal da esquerda para a direita. De cima para baixo está a maneira como ela é divida. Em azul temos o conjunto de treinamento e em verde o conjunto de teste.

Utilizando a classe *TimeSeriesSplit* do pacote Sklearn, é realizada uma verificação do número razoável de divisões entre conjuntos de treinamento e teste da amostra.

No Gráfico 12, é apresentado um parâmetro para decisão do número de divisões entre subconjuntos de treinamento e teste. Quanto mais próximo de

R^2, melhor o resultado. Assim, percebe-se que quanto mais divisões na amostra, pior o resultado do algoritmo. Devido a tal fato foi utilizado o número de 2 subconjuntos.

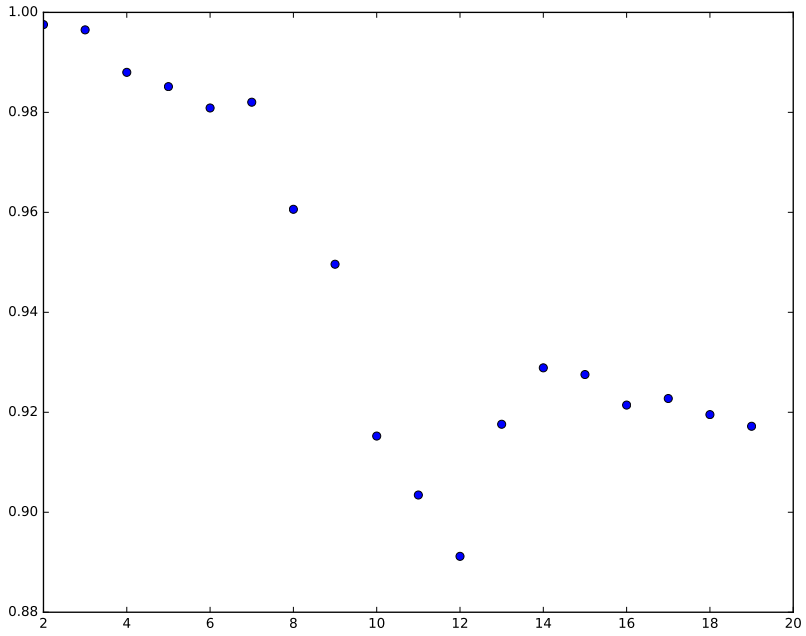


Gráfico 12 – No eixo y estão os valores do coeficiente de determinação R^2 e no eixo x o número de divisões entre conjuntos de treinamento e teste. Valores utilizando o preditor mínimos quadrados do pacote Sklearn.

No Gráfico 13, é apresentada a previsão do modelo entre os dias 07/11/2016 e 17/11/2016. O gráfico superior apresenta a evolução da previsão e o preço realizado. O gráfico inferior apresenta a evolução da diferença entre esses valores.

A Tabela 4 apresenta os valores dos parâmetros do modelo.

	w_0	w_1	μ	σ^2
Valor	0,1437	0,9986	42,06	0,2986

Tabela 4 – coeficientes do modelo linear sklearn.linearmodel.

O algoritmo utilizado foi o sklearn.linearmodel que utiliza o modelo matemático dado na equação 5.1

$$y_t = w_0 + w_1 y_{t-1} \tag{5.1}$$

O Gráfico 14 apresenta a evolução dos preços previstos, utilizando o algoritmo sklearn.linearmodel, e realizados por trimestre. O modelo utilizado é o de mínimos quadrados, porém as variáveis consideradas agora são as mesmas apresentadas pelo vetor (4.2) X^T.

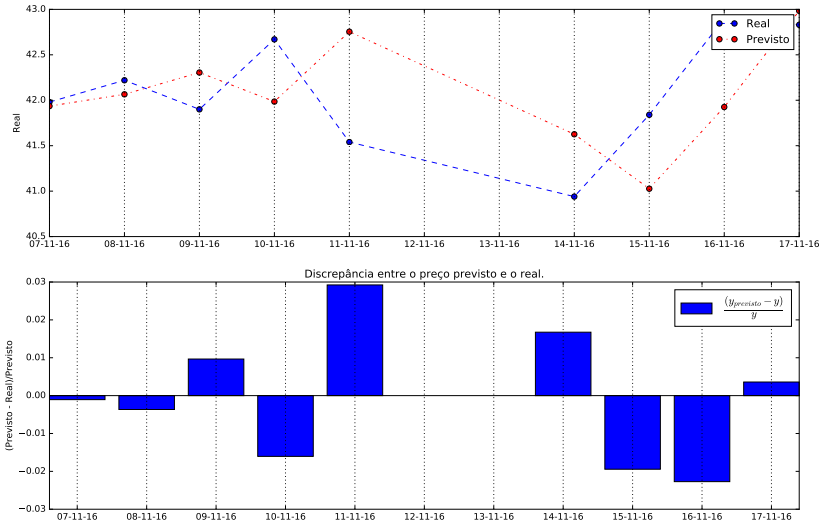


Gráfico 13 – Previsão dos preços OPEC por mínimos quadrados e utilizando somente a série dos preços como dado de entrada. Gráfico superior – evolução do preço previsto da cesta do petróleo da OPEC e o preço realizado. Gráfico inferior – diferença entre o previsto e o preço realizado (Dados diários).

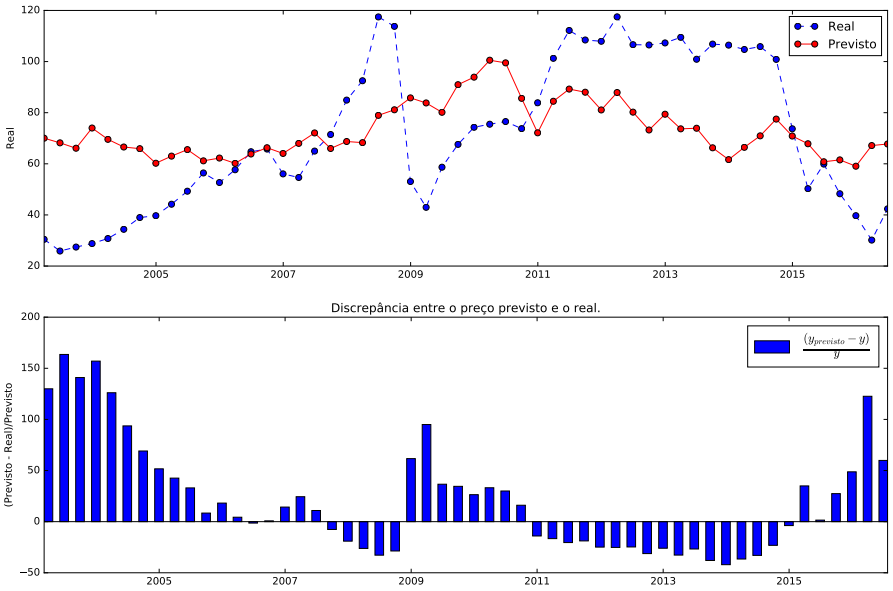


Gráfico 14 – Previsão dos preços OPEC por mínimos quadrados e utilizando dados de oferta e demanda como dado de entrada. Gráfico superior – evolução do preço previsto da cesta de petróleo OPEC e o previsto pelo modelo. Gráfico inferior – diferença entre o previsto e o realizado (Dados trimestrais).

A matriz de coeficientes gerada é

$$w = \begin{bmatrix} -4,71 \\ 2,62 \\ 4,74 \\ 6,54 \end{bmatrix} \tag{5.2}$$

Com o modelo (5.1), (4.2) e (5.2), tem-se (5.3).

$$Y = w_0 + wX^T \tag{5.3}$$

Onde $w_0 = 37,04$.

Uma medida de qualidade do modelo é o coeficiente R^2 que na simulação realizada utilizando como dado de entradas o vetor (4.2) é $R^2 = 0.72$.

6 CONCLUSÃO

Este trabalho apresenta um estudo da capacidade de previsão, em séries temporais do mercado financeiro, de algoritmos de apredizagem de máquina comparados com métodos já tradicionais em econometria. O caso específico abordado compara o método de Box e Jenkins com o método linear não regularizado do pacote Sklearn.

Foi realizada uma análise pelo método de Box e Jenkins [1] da série temporal dos preços da cesta do petróleo da OPEC, utilizando basicamente o software R. Posteriormente, foi realizada uma análise utilizando o modelo de mínimos do pacote Sklearn. O objetivo central é confrotar estas metodologias verificando os prós e contras para um investidor.

Nos casos específicos análisados pelo método de Box e Jenkins, as previsões de curto prazo foram satisfatórias com o nível de confiança de 95%. O que pode fornecer um indicativo para operações financeiras estruturadas com opções ou somente com perspectiva de risco financeiro. Mesmo com grandes oscilações, o modelo baseado em passeio aletório desmonstrou-se útil de maneira semelhante ao citado em [7].

Os algoritmos de mínimos quadrados do pacote Sklearn estudados introduzem uma diferença maior entre os valores observados e previstos, oferecendo uma visão conservadora da série.

O mesmo algoritmo do pacote Sklearn, quando consideradas como variáveis de entrada dados de oferta e demanda, fornece uma previsão conservadora e proibitiva para utilização real entre os valores observados e previstos.

Possivelmente, a correlação existente nos dados de oferta e demanda do preço do pretróleo podem oferecer uma grande informação para algoritmos de aprendizagem de máquina. Uma análise com algoritmos de classificação é um trabalho interessante, pois tem o potencial de fornecer sinais de movimentos positivos ou negativos do mercado o que é uma informação interessante para especuladores e *hedgers*, por exemplo.

A utilização de dados diferentes do mercado financeiro como volume financeiro de negociações, série temporal de negociações é interessante para uma análise futura.

REFERÊNCIAS

1 BOX, G. E.; JENKINS, G. M. **Time series analysis: forecasting and control, revised ed**. [S.l.: s.n.]. 6, 14, 19, 33

2 MARINS, A. C. **Mercado de Derivativos e Análise de Risco**. [S.l.]: AMS, 2004. v. 1. 12

3 OPEC. **OPEC Annual Statistical Bulletin**. 2016. 128 p. Disponível em: <http://asb.opec.org>. 13

4 IEA. **Oil Medium-Term Market Report 2015**. 2015. 140 p. Disponível em: <http://www.iea.org/publications/freepublications/publication/MTOMR_2015_Final.pdf>. 13

5 FAMA, E. F. The Behavior of Stock-Market Prices. **The Journal of Business**, University of Chicago Press, v. 38, n. 1, p. 34–105, 1965. ISSN 00219398, 15375374. Disponível em: <http://www.jstor.org/stable/2350752>. 14

6 FAMA, E. F. Efficient Capital Markets: A Review of Theory and Empirical Work. **The Journal of Finance**, [American Finance Association, Wiley], v. 25, n. 2, p. 383–417, 1970. ISSN 00221082, 15406261. Disponível em: <http://www.jstor.org/stable/2325486>. 14

7 SERLETIS, A.; SHINTANI, M. No evidence of chaos but some evidence of dependence in the {US} stock market. **Chaos, Solitons & Fractals**, v. 17, n. 2–3, p. 449–454, 2003. ISSN 0960-0779. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960077902003879>. 14, 33

8 MITCHELL, T. M. **Machine Learning**. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISSN 9780070428072. ISBN 0070428077. 15

9 SUGIYAMA, M. Chapter 1 - Statistical Machine Learning. In: SUGIYAMA, M. (Ed.). **Introduction to Statistical Machine Learning**. Boston: Morgan Kaufmann, 2016. p. 3–8. ISBN 978-0-12-802121-7. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780128021217000121>. 15

10 BISHOP, C. About. **Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn**. 1. ed. Springer, New York, 2006. 738 p. (Information Science and Statistics). ISSN 16139011. ISBN 9780387310732. Disponível em: <http://www.springer.com/us/book/9780387310732>. 15

11 HADAMARD, J. **Lectures on Cauchy's Problem in Linear Partial Differential Equations**. [S.l.]: JSTOR, 1925. 17

12 PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. 29

13 BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**. [S.l.: s.n.], 2013. p. 108–122. 29

14 SKLEARN. **Cross-validation**. Disponível em: <http://scikit-learn.org/stable/modules/cross_validation.html>. 29