

RAFFAELLO COUTO CASER

Introdução à arbitragem estatística com implementação de modelo em Python

São Paulo

2018

RAFFAELLO COUTO CASER

Introdução à arbitragem estatística com implementação de modelo em Python

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Especialista em
Engenharia Financeira

Orientador: Prof. Dr. Oswaldo Luiz do
Valle Costa

São Paulo

2018

RAFFAELLO COUTO CASER

Introdução à arbitragem estatística com implementação de modelo em Python

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Especialista em
Engenharia Financeira

Área de Concentração: Engenharia
Financeira

Orientador: Prof. Dr. Oswaldo Luiz do
Valle Costa

São Paulo

2018

RESUMO

Esta dissertação desenvolveu-se com objetivo de introduzir o conceito de arbitragem estatística e testar o desempenho no mercado de ações brasileiro de um modelo de investimento fundamentado nessa teoria. Para tanto assume-se que o leitor possui noções de teoria das probabilidades e linguagem de programação para que se estabeleça os conceitos de séries temporais e cointegração que servirão de base para o desenvolvimento do tema deste trabalho. Ainda, discute-se os principais tipos de operações existentes no mercado financeiro, a saber, cobertura, especulação e arbitragem, com ênfase na última uma vez que por extensão desta é possível definir arbitragem estatística. Outrossim detalha-se os conceitos acerca do tema, além de uma estratégia sistemática de investimento sustentada por esta teoria. Por fim simula-se qual seria o desempenho de tal estratégia no mercado de ações brasileiro entre os anos de 2015 e 2017, discute-se os resultados, além de sugerir abordagens complementares ao modelo.

Palavras-Chave: Engenharia financeira. Econometria. Séries temporais. Arbitragem estatística. Estratégias de investimento.

ABSTRACT

This essay has the goal of introducing the concept of statistical arbitrage and testing the performance in the Brazilian stock market of an investment model based on this theory. For this it is assumed that the reader has notions of probability theory and programming language to establish the concepts of time series and cointegration that will serve as a basis for the development of the subject of this work. Moreover the main types of operations in the financial market, namely hedging, speculation and arbitrage, are discussed, with emphasis on the latter, since by extension of this one it is possible to define statistical arbitrage. In addition it is detailed the concepts about the subject, as well as a systematic investment strategy supported by this theory. Finally it is simulated the performance of such a strategy in the Brazilian stock market between 2015 and 2017, the results are discussed and it is suggested complementary approaches to the model.

Keywords: Financial engineering. Econometrics. Time series. Statistical arbitrage. Investment strategies.

SUMÁRIO

1	INTRODUÇÃO	6
2	FUNDAMENTAÇÃO TEÓRICA	9
2.1	SÉRIES TEMPORAIS	9
2.1.1	Estacionariedade de séries temporais	10
2.1.2	Função de autocorrelação	10
2.1.3	Exemplos de séries temporais	11
2.1.3.1	Ruído Branco	11
2.1.3.2	Média móvel (MA)	14
2.1.3.3	Séries autorregressivas (AR)	17
2.1.3.4	Passeio aleatório	20
2.1.4	Cointegração	24
2.2	OPERAÇÕES DO MERCADO FINANCEIRO	26
2.2.1	Cobertura	26
2.2.2	Especulação	28
2.2.3	Arbitragem	28
2.3	ARBITRAGEM ESTATÍSTICA	31
2.3.1	Construção da relação estatística de preços	32
2.3.2	Identificação dos desvios de preço	34
2.3.3	Posicionamento no mercado	36
3	IMPLEMENTAÇÃO DO MODELO	38
3.1	COLETA DE DADOS	38
3.2	ALGORITMO EM PYTHON	40
3.3	RESULTADOS	46
4	CONCLUSÕES	55
5	BIBLIOGRAFIA	57

1 INTRODUÇÃO

O intenso desenvolvimento tecnológico do século passado influenciou, se não todas, a maior parte das áreas do saber humano devido, entre outros motivos, ao surgimento dos computadores somado a sua rápida evolução que implicou em um ganho formidável de velocidade de processamento de dados. Não foi diferente com finanças e o mercado financeiro. Tal evolução oportunizou aos financistas criar e aplicar estratégias de investimento e gestão de riscos que antes não seriam possíveis em razão justamente do requisito de trabalhar de forma rápida com grandes bases de dados de séries históricas de preços de ativos financeiros. Isso levou ao surgimento de campos de estudo inteiramente novos, tais como, engenharia financeira, finanças quantitativas, econometria financeira, entre outros, que combinam métodos de estatística e computação aplicados às finanças.

O trabalho de Bachelier (1900) publicado ainda décadas antes da criação e popularização dos computadores é considerado a fundação do estudo de teoria de probabilidades aplicada a finanças sendo precursor das áreas supracitadas. Neste trabalho o autor explicou o comportamento da bolsa de valores¹ francesa assumindo que os preços das ações² evoluíam no tempo conforme um modelo conhecido como passeio aleatório³. Infelizmente ele faleceu sem grande reconhecimento do seu trabalho provavelmente por não ter sido possível testá-lo empiricamente, justamente por não haver a tecnologia necessária à época. No entanto nas décadas de 1960 e 1970 Fama (1965) e Black e Scholes (1973) trouxeram o trabalho de Bachelier (1900) ao centro das atenções dos estudiosos de finanças: o primeiro, já com o poder de processamento dos computadores a sua disposição, elaborou testes estatísticos para verificar a hipótese de Bachelier (1900) de que os preços das ações seguem o modelo de passeio aleatório e constatou que existiam fortes indícios empíricos de que a tese do matemático francês estava correta, fato que o

¹ Bolsa de valores são mercados organizados autorregulados para negociação de ativos financeiros.

² Ação é a menor fração do capital social de uma sociedade anônima.

³ Daremos maiores detalhes sobre passeio aleatório em 2.1.3.4.

levou a elaborar a conhecida Hipótese dos Mercados Eficientes⁴, a qual diz basicamente que quanto mais eficiente é um mercado, mais rapidamente os agentes econômicos reagem às novas informações de forma que os preços atuais dos ativos financeiros incorporam toda a informação disponível e, por conseguinte, os preços passados possuem pouca ou nenhuma utilidade para o fim de prever os preços futuros; Black e Scholes (1973), por sua vez, elaborou o notável modelo de apreamento de opções⁵ de ações que tem como premissa que a ação que subscreve o contrato de opção segue o modelo proposto por Bachelier (1900).

O sucesso do passeio aleatório para descrever o comportamento de ações foi uma conquista muito importante para o campo da engenharia financeira, entretanto deixou os operadores do mercado financeiro em uma situação complicada, afinal o resultado de que os valores passados de uma série de preços de ação não possui nenhuma influência nos valores futuros não é nem um pouco animador para quem trabalha tentando prevêê-los. Mas, felizmente, novas teorias e ideias surgiram na esteira dos trabalhos acima, entre elas, a arbitragem estatística, tema desta dissertação. Segundo Vidyamurthy (2004), o primeiro grupo de pessoas que se sabe ter utilizado essa estratégia foi a equipe liderada por Nunzio Tartaglia no banco de investimento americano Morgan Stanley onde teriam obtido excelente performance no final da década de 1980. Posteriormente a equipe foi se desfazendo com seus membros passando a trabalhar em outras instituições de investimento de forma que a estratégia foi naturalmente se popularizando sendo objeto de estudo de acadêmicos e gestores de ativos.

A técnica consiste basicamente de identificar dois ativos financeiros, ou dois grupos de ativos financeiros, que possuem uma relação estatística tal que, construindo-se uma carteira de investimento a partir desses dois grupos com as devidas proporções de alocação de recursos em cada um deles, é possível obter uma série financeira que, ao contrário do passeio aleatório, possui um comportamento passível de predição, ou seja, o conhecimento do histórico da série é útil para antecipar movimentos de preço e, assim, auferir lucro. A grosso modo, a

⁴ Em 2.1.3.4 veremos o porquê dessa hipótese ser coerente com o passeio aleatório.

⁵ Veremos em 2.2.1 a definição de um contrato de opção de ativo financeiro.

metodologia busca ativos financeiros cujos preços tendem a variar de forma similar, por estarem expostos aos mesmo fatores de risco⁶, e os monitora para, quando houver uma incoerência entre a variação de preço dos mesmos, se posicionar comprando o ativo que estiver relativamente depreciado e vendendo o que estiver relativamente apreciado, pois uma vez que estão expostos ao mesmos fatores de risco, tais ativos deverão em algum momento do futuro retornar a relação de equilíbrio que mantinham.

Posto isso, para definirmos arbitragem estatística de maneira mais precisa e coerente organizamos a dissertação da seguinte forma: no segundo capítulo exporemos a teoria que fundamenta a estratégia iniciando pela conceituação de séries temporais e definição de duas importantes propriedades destas, estacionariedade e função de autocorrelação; em seguida discorreremos sobre os exemplos fundamentais de séries temporais, inclusive passeio aleatório, e finalizaremos com o conceito de cointegração, imprescindível à estratégia de investimento; ainda no segundo capítulo, trataremos das operações básicas do mercado financeiro, em especial a operação de arbitragem, pela qual, por extensão, é possível definir arbitragem estatística, último tópico desse capítulo; no terceiro capítulo, por seu turno, através do histórico de preços de ações negociadas na bolsa de valores de São Paulo no período de 2010 a 2017⁷, testaremos o desempenho de um modelo de investimento baseado na teoria exposta no capítulo anterior e implementado em Python⁸; por fim, discutiremos os resultados alcançados e sugeriremos direções a seguir.

⁶ Podemos citar como exemplos de fatores de risco, cotação do câmbio entre duas moedas distintas, taxa de juros base de uma economia (no Brasil, a taxa SELIC), cotação de insumos primários, contexto regulatório de um país, entre outros.

⁷ Ver: BM&FBovespa – **Cotações históricas**.

⁸ Python é uma linguagem de programação de alto nível cujo modelo de desenvolvimento é colaborativo, aberto e gerenciado pela organização sem fins lucrativos *Python Software Foundation*. Disponível em: <www.python.org>. Acesso em: 12 jan. 2017.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo exporemos a fundamentação teórica necessária para o entendimento e aplicação do modelo de investimento de arbitragem estatística. Este divide-se em três partes principais: séries temporais, operações do mercado financeiro e arbitragem estatística. As duas primeiras formam a base de estatística e finanças para o entendimento da última que, por sua vez, será aplicada no modelo do terceiro capítulo.

2.1 SÉRIES TEMPORAIS

Uma série temporal $\{x_t \mid t = 1, \dots, T\}$ é uma sequência de valores medidos ao longo do tempo que podem ser determinísticos ou aleatórios, isto é, obtidos a partir de uma fórmula determinística ou de uma distribuição de probabilidade. Apresentaremos a seguir de forma objetiva os principais conceitos desse tema necessários para o entendimento do modelo de arbitragem estatística, quais sejam, estacionariedade, autocorrelação e cointegração, além de alguns exemplos com o fim de obter maior clareza.

Naturalmente a bibliografia sobre esse tópico é extensa devido sua aplicabilidade nas mais diversas áreas, tais como, processamento de sinais, ciências sociais, epidemiologia, meteorologia, economia e, certamente, engenharia financeira, área de emprego desta dissertação. Posto isso, consultamos Tsay (2005) e Vidyamurthy (2004) para este fim, por serem referências que possuem rigor teórico além de apresentarem exemplos interessantes em finanças, inclusive arbitragem estatística.

Ainda, assumimos que o leitor já é familiarizado com alguns conceitos introdutórios de teoria das probabilidades, tais como, variável aleatória, valor esperado, variância, desvio padrão (comumente denominado volatilidade no mercado financeiro), covariância, correlação e distribuição normal. Caso contrário, recomendamos Ross (2010) e Magalhães e Lima (2004).

2.1.1 Estacionariedade de séries temporais

Uma série temporal é dita estritamente estacionária quando a distribuição conjunta de $(x_{t_1}, \dots, x_{t_k})$ e $(x_{t_1+\ell}, \dots, x_{t_k+\ell})$ são idênticas para qualquer inteiro k , ou seja, a distribuição conjunta de $(x_{t_1}, \dots, x_{t_k})$ é invariante sob deslocamento no tempo. Conforme é ressaltado em Tsay (2005), tal característica é muito difícil de se observar empiricamente, de forma que, em termos práticos, o tipo de estacionariedade com que se trabalha é a chamada estacionariedade fraca. Esta exige que a série temporal possua valor esperado constante e finito e autocovariância que dependa apenas do tamanho do passo ℓ da série, isto é:

- $E(x_t) = \mu$, onde μ é uma constante finita; e
- $Cov(x_t, x_{t-\ell}) = \gamma_\ell$, que depende apenas de ℓ .

Como consequência direta dessa definição, temos que:

- $Var(x_t) = \gamma_0$; e
- $\gamma_{-\ell} = \gamma_\ell$.

Uma propriedade importante apresentada por séries temporais estacionárias, uma vez que possui variância finita, é a de reversão à média, ou seja, a série tende ao longo do tempo a retornar ao seu valor médio. Essa característica será de extrema relevância na estratégia de investimento que desenvolveremos no terceiro capítulo devido ao seu poder de previsibilidade.

2.1.2 Função de autocorrelação

A correlação entre duas variáveis aleatórias é uma medida de dependência linear entre estas duas variáveis. Autocorrelação, por sua vez, é uma generalização da noção de correlação, que mede a dependência linear entre os valores observados em um instante t com os valores observados em períodos passados. Em outras palavras, sob a hipótese de estacionariedade fraca, a autocorrelação de passo ℓ (ρ_ℓ) é a correlação entre x_t e $x_{t-\ell}$, ou seja:

$$\rho_\ell = Corr(x_t, x_{t-\ell}) = \frac{Cov(x_t, x_{t-\ell})}{\sqrt{Var(x_t)Var(x_{t-\ell})}} = \frac{Cov(x_t, x_{t-\ell})}{Var(x_t)} = \frac{\gamma_\ell}{\gamma_0}. \quad (1)$$

Como resultado do estabelecido acima, temos que $\rho_0 = 1$, $\rho_{-\ell} = \rho_\ell$ e $-1 \leq \rho_\ell \leq 1$. Portanto a função de autocorrelação é a relação entre a autocorrelação da série temporal com o tamanho do passo ℓ , isto é, $\rho_\ell = f(\ell)$.

A relevância dessa função advém de sua relação com a propriedade de reversão à média apresentada pelas séries temporais estacionárias: quanto menor a autocorrelação de uma série temporal estacionária, maior sua velocidade de reversão à média. Em outras palavras, quanto menor a autocorrelação, menor é a memória da série temporal.

2.1.3 Exemplos de séries temporais

Apresentaremos em seguida alguns exemplos de séries temporais com intuito de solidificar os conceitos apresentados acima, além de introduzir novos que serão valiosos no desenvolvimento do modelo de arbitragem estatística. Iniciaremos pelo ruído branco e por extensões obteremos as séries de média móvel, autorregressiva e passeio aleatório. Estes e outros exemplos de séries temporais podem ser encontrados nos já citados Tsay (2005) e Vidyamurthy (2004) e, com maiores detalhes, em Morettin e Tolo (2004).

2.1.3.1 Ruído branco

Iniciemos com o caso mais simples de série temporal probabilística, porém extremamente útil aos modelos estatísticos, denominada ruído branco. Esta série é obtida tomando-se a cada instante de tempo, de forma independente, valores de uma específica distribuição de probabilidade com parâmetros constantes. Quando a distribuição de probabilidade que origina a série é a distribuição normal (ou gaussiana), cenário mais usual, esta série chama-se ruído branco gaussiano. A notação que se convencionou para o ruído branco é $y_t = \varepsilon_t$, onde ε_t é o valor obtido da distribuição de probabilidade no instante t . Daqui em diante utilizaremos essa notação para designar ruído branco com média nula e variância finita σ^2 , isto é, $\varepsilon_t \sim iid(0, \sigma^2)$. Dessa forma podemos calcular as principais estatísticas dessa série:

$$E(y_t) = E(\varepsilon_t) = 0, \quad (2)$$

$$\gamma_0 = \text{Var}(y_t) = \text{Var}(\varepsilon_t) = \sigma^2, \quad (3)$$

$$\gamma_\ell = \text{Cov}(y_t, y_{t-\ell}) = \text{Cov}(\varepsilon_t, \varepsilon_{t-\ell}) = E[(\varepsilon_t - E(\varepsilon_t))(\varepsilon_{t-\ell} - E(\varepsilon_{t-\ell}))] = E(\varepsilon_t \varepsilon_{t-\ell}) = 0, \quad (4)$$

$$\rho_\ell = \frac{\gamma_\ell}{\gamma_0} = 0. \quad (5)$$

Note que a série atende os requisitos de estacionariedade fraca (vide item 2.1.1) e sua autocorrelação é nula para qualquer $\ell \neq 0$, ou seja, os valores observados a cada instante de tempo são descorrelacionados com os valores passados e a série apresenta forte propriedade de reversão à média.

Na tabela 1 vê-se código em Python para geração de ruído branco gaussiano com média nula e desvio padrão unitário, além dos gráficos da série e da função de autocorrelação nas figuras 1 e 2, respectivamente. Na primeira nota-se a característica de reversão à média da série, enquanto que na segunda figura verifica-se que, de fato, a autocorrelação é insignificante para qualquer passo $\ell \neq 0$.

Tabela 1 – Código em Python para geração do ruído branco

```
# importação das bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf

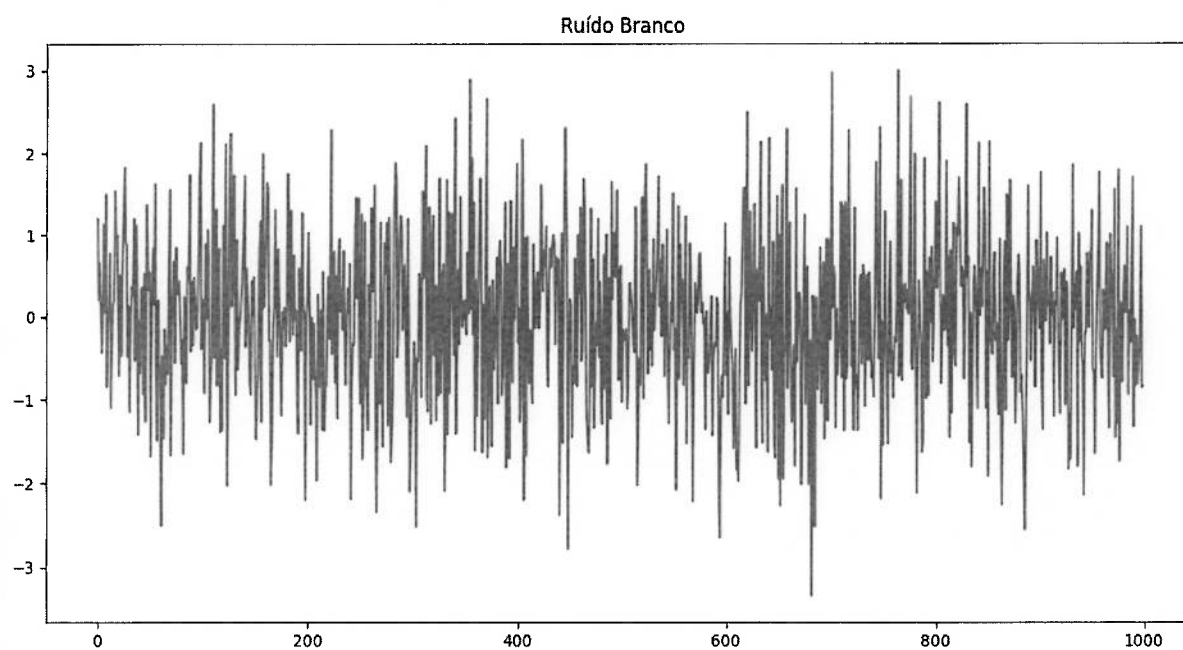
# definição dos parâmetros e tamanho da amostra
media = 0
desvio_padrao = 1
amostra = 1000

# geração da série
ruído_branco = np.random.normal(media, desvio_padrao, size=amostra)

# geração dos gráficos
plt.plot(ruído_branco)
plt.title('Ruído Branco')
plot_acf(ruído_branco, lags=30, title='Função de Autocorrelação do Ruído Branco')
plt.xlabel('passo')
plt.ylabel('autocorrelação')
plt.show()
```

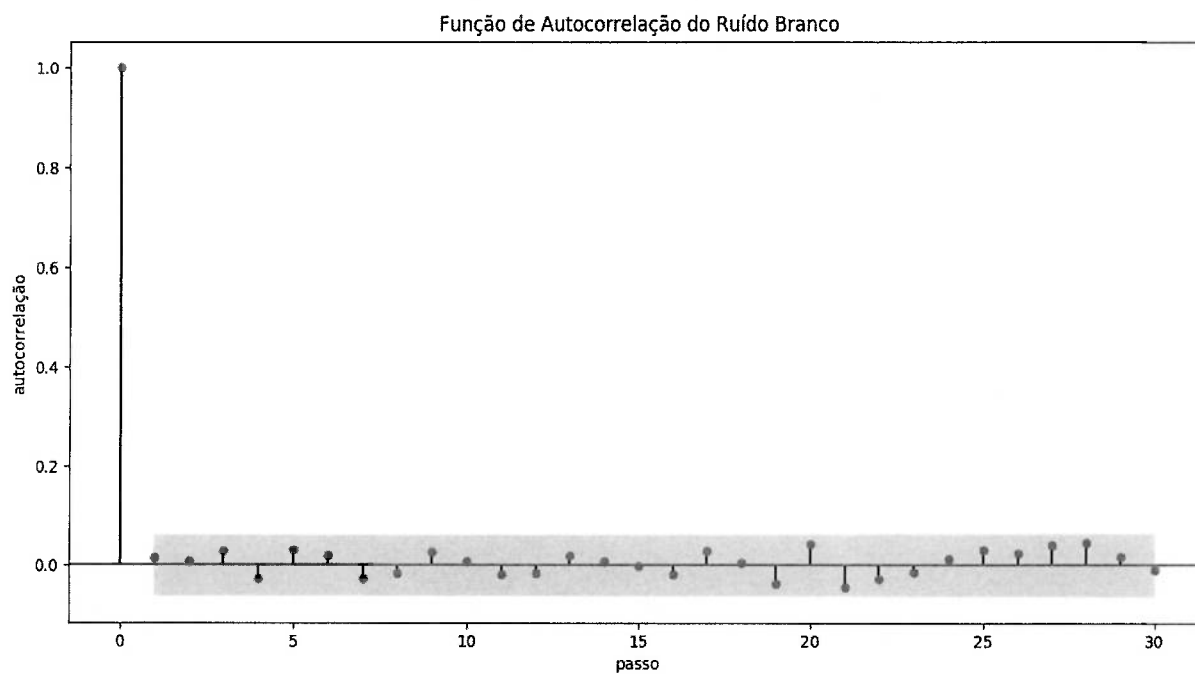
Fonte: Elaboração própria

Figura 1 – Ruído branco gaussiano com média nula e desvio padrão unitário



Fonte: Elaboração própria

Figura 2 – Função de autocorrelação do ruído branco



Fonte: Elaboração própria

2.1.3.2 Média móvel (MA)⁹

A média móvel de primeira ordem, em síntese MA(1), é uma generalização do ruído branco obtida pela seguinte relação:

$$y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1}, \quad (6)$$

onde β_1 é uma constante.

O primeiro termo da eq.(6) é interpretado como a informação nova do período t , enquanto que $\beta_1 \varepsilon_{t-1}$ é a informação do período anterior ponderada pelo fator β_1 . Por isso, para aplicações práticas, faz sentido utilizar $|\beta_1| < 1$, uma vez que espera-se que a informação mais recente tenha peso maior que informações passadas. Por esse motivo, também é usual denominar ε_t por choque ou inovação da série, termos que poderemos utilizar mais adiante. Posto isso, as estatísticas da série MA(1) são:

$$E(y_t) = E(\varepsilon_t + \beta_1 \varepsilon_{t-1}) = E(\varepsilon_t) + \beta_1 E(\varepsilon_{t-1}) = 0, \quad (7)$$

$$\gamma_0 = \text{Var}(y_t) = \text{Var}(\varepsilon_t + \beta_1 \varepsilon_{t-1}) = \text{Var}(\varepsilon_t) + \beta_1^2 \text{Var}(\varepsilon_{t-1}) = \sigma^2(1 + \beta_1^2) \quad (8)$$

$$\gamma_1 = \text{Cov}(y_t, y_{t-1}) = E(y_t y_{t-1}) = E[(\varepsilon_t + \beta_1 \varepsilon_{t-1})(\varepsilon_{t-1} + \beta_1 \varepsilon_{t-2})] = \beta_1 E(\varepsilon_{t-1}^2) = \beta_1 \sigma^2, \quad (9)$$

$$\ell > 1 : \gamma_\ell = \text{Cov}(y_t, y_{t-\ell}) = E(y_t y_{t-\ell}) = E[(\varepsilon_t + \beta_1 \varepsilon_{t-1})(\varepsilon_{t-\ell} + \beta_1 \varepsilon_{t-\ell-1})] = 0, \quad (10)$$

$$\rho_1 = \frac{\gamma_1}{\gamma_0} = \frac{\beta_1 \sigma^2}{\sigma^2(1 + \beta_1^2)} = \frac{\beta_1}{(1 + \beta_1^2)}, \quad (11)$$

$$\ell > 1 : \rho_\ell = \frac{\gamma_\ell}{\gamma_0} = 0. \quad (12)$$

Como era de se esperar, a série possui autocorrelação de passo unitário, porém não possui autocorrelação em passos maiores, uma vez que cada y_t possui termo em comum (ε_{t-1}) apenas com o passo imediatamente anterior y_{t-1} . Além disso,

⁹ A abreviatura MA deve-se à nomenclatura em inglês *moving average*.

¹⁰ Na terceira igualdade utilizamos o resultado de que a variância da soma de duas variáveis aleatórias independentes é igual a soma de suas variâncias. Daqui em diante o utilizaremos de forma direta sem menção.

observa-se que a série atende os requisitos de estacionariedade do item 2.1.1. Note ainda que:

$$E(y_t|y_{t-1}) = E(y_t|\varepsilon_{t-1}) = E(\varepsilon_t + \beta_1 \varepsilon_{t-1}|\varepsilon_{t-1}) = E(\varepsilon_t) + \beta_1 E(\varepsilon_{t-1}|\varepsilon_{t-1}) = \beta_1 \varepsilon_{t-1}. \quad (13)$$

Em palavras, o valor esperado da série em um instante t , dado que sabemos o valor imediatamente anterior, é a inovação ocorrida no passo anterior amortecida pelo fator β_1 . Ou seja, enquanto para o ruído branco o valor obtido a cada passo t advinha de uma distribuição de probabilidade com média nula e variância σ^2 , para a MA(1) o valor origina-se de uma distribuição com média $\beta_1 \varepsilon_{t-1}$ e variância σ^2 . Deste fato surge o termo *média móvel* para designar esta série temporal.

Na tabela 2 e figuras 3 e 4 adiante, vê-se, respectivamente, o código para a reprodução da MA(1) com fator beta igual a 0,5, o gráfico do mesmo, além da plotagem da função de autocorrelação da série. Nestes é possível observar a propriedade de reversão à média, além da existência de autocorrelação para o passo unitário.

Tabela 2 – Código em Python para geração da MA(1)

```
# importação das bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf

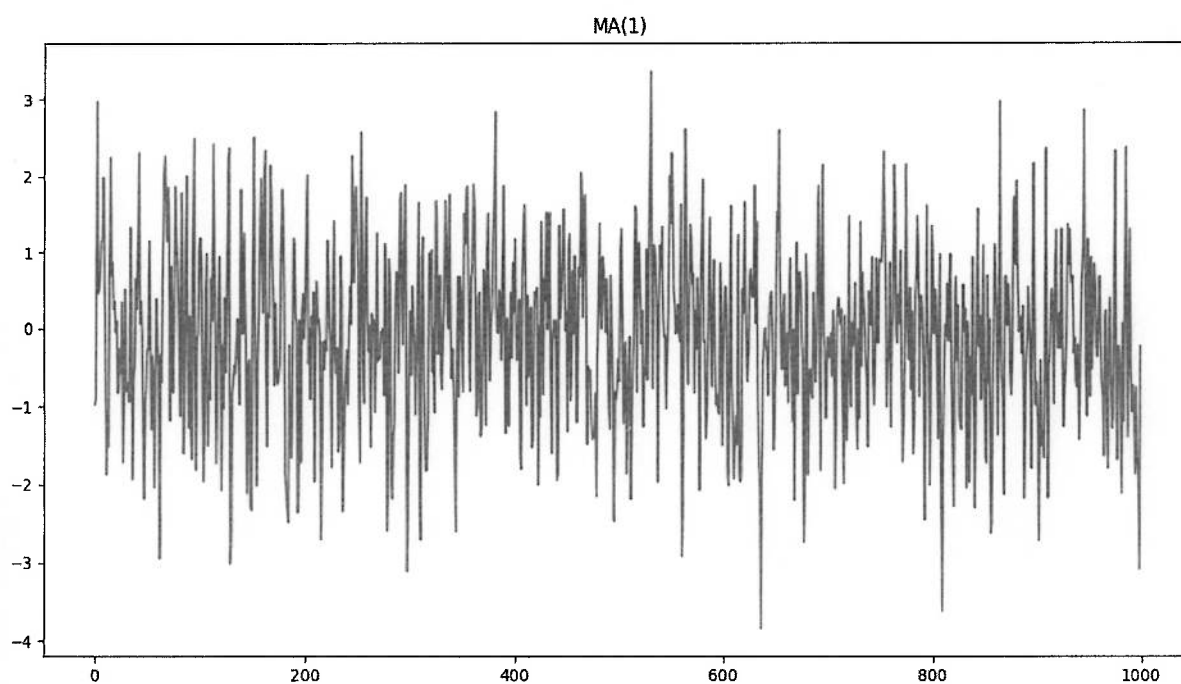
# definição dos parâmetros e tamanho da amostra
media = 0
desvio_padrao = 1
amostra = 1000
beta = 0.5

# geração da série MA(1)
inovacao = np.random.normal(media, desvio_padrao, size=amostra)
inovacao_beta = np.insert(np.delete(inovacao, amostra - 1), 0, 0) * beta
MA_1 = inovacao + inovacao_beta

# geração dos gráficos
plt.plot(MA_1)
plt.title('MA(1)')
plot_acf(MA_1, lags=30, title='Função de Autocorrelação da MA(1)')
plt.xlabel('passo')
plt.ylabel('autocorrelação')
plt.show()
```

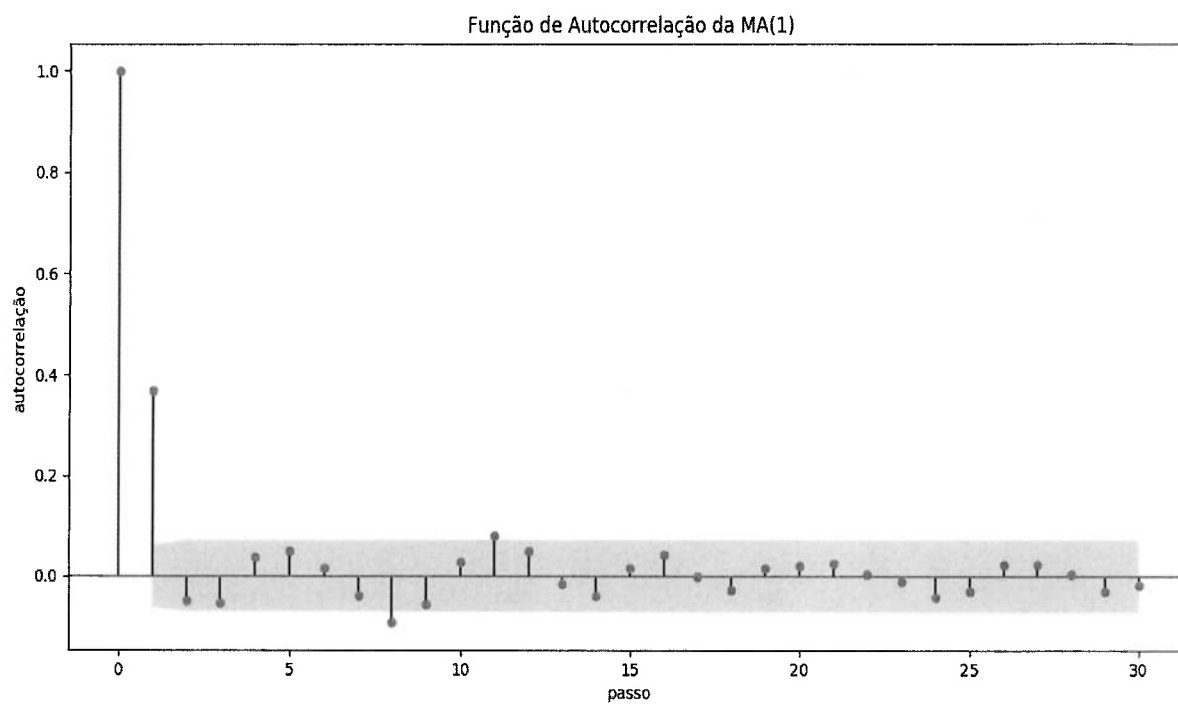
Fonte: Elaboração própria

Figura 3 – Média móvel de primeira ordem



Fonte: Elaboração própria

Figura 4 – Função de autocorrelação da média móvel de primeira ordem



Fonte: Elaboração própria

Por fim, é possível generalizar a média móvel de primeira ordem para uma ordem $q > 1$, MA(q), através da adição à série de uma combinação linear das inovações dos q passos anteriores do seguinte modo:

$$y_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \dots + \beta_q \varepsilon_{t-q} . \quad (14)$$

As estatísticas e características da MA(q) são análogas as da MA(1), de maneira que, enquanto esta apresenta autocorrelação de passo unitário, aquela apresenta autocorrelação até o q -ésimo passo.

2.1.3.3 Séries autorregressivas (AR)

Uma série autorregressiva de primeira ordem, AR(1), é obtida tomando-se o limite de MA(q) com $q \rightarrow \infty$ e $\beta_i = \alpha^i$:

$$y_t = \lim_{q \rightarrow \infty} MA(q) = \varepsilon_t + \alpha \varepsilon_{t-1} + \alpha^2 \varepsilon_{t-2} + \dots \quad (15)$$

Note que, da definição acima, tem-se:

$$y_{t-1} = \varepsilon_{t-1} + \alpha \varepsilon_{t-2} + \alpha^2 \varepsilon_{t-3} + \dots \quad (16)$$

Portanto, podemos reescrever AR(1):

$$y_t = \varepsilon_t + \alpha y_{t-1} . \quad (17)$$

Ou seja, enquanto MA(q) considerava apenas os choques até q passos anteriores, esta contempla, além da inovação do período t , toda a informação passada da série.

Vejamos as estatísticas da série:

$$E(y_t) = E(\varepsilon_t + \alpha \varepsilon_{t-1} + \alpha^2 \varepsilon_{t-2} + \dots) = E(\varepsilon_t) + \alpha E(\varepsilon_{t-1}) + \alpha^2 E(\varepsilon_{t-2}) + \dots = 0 , \quad (18)$$

$$\gamma_0 = Var(y_t) = Var(\varepsilon_t + \alpha \varepsilon_{t-1} + \alpha^2 \varepsilon_{t-2} + \dots) = Var(\varepsilon_t) + \alpha^2 Var(\varepsilon_{t-1}) + \alpha^4 Var(\varepsilon_{t-2}) + \dots$$

$$= (1 + \alpha^2 + \alpha^4 + \dots) \sigma^2 = \frac{\sigma^2}{(1-\alpha^2)} , \text{ se } |\alpha| < 1 , \quad (19)$$

$$\gamma_\ell = Cov(y_t, y_{t-\ell}) = E(y_t y_{t-\ell}) = E[(\varepsilon_t + \alpha \varepsilon_{t-1} + \alpha^2 \varepsilon_{t-2} + \dots)(\varepsilon_{t-\ell} + \alpha \varepsilon_{t-\ell-1} + \alpha^2 \varepsilon_{t-\ell-2} + \dots)]$$

$$= \alpha^\ell E(\varepsilon_{t-\ell}^2) + \alpha^{\ell+2} E(\varepsilon_{t-\ell-1}^2) + \alpha^{\ell+4} E(\varepsilon_{t-\ell-2}^2) + \dots = \alpha^\ell (1 + \alpha^2 + \alpha^4 + \dots) \sigma^2 = \alpha^\ell \frac{\sigma^2}{(1-\alpha^2)} , \quad (20)$$

$$|\alpha| < 1 ,$$

$$\rho_\ell = \frac{y_\ell}{y_0} = \frac{\alpha^\ell (1 + \alpha^2 + \alpha^4 + \dots) \sigma^2}{(1 + \alpha^2 + \alpha^4 + \dots) \sigma^2} = \alpha^\ell, \text{ se } |\alpha| < 1. \quad (21)$$

Perceba que a soma $(1 + \alpha^2 + \alpha^4 + \dots)$ converge quando $|\alpha| < 1$, e nesse caso temos uma série estacionária. Por outro lado, para $|\alpha| \geq 1$, a soma $(1 + \alpha^2 + \alpha^4 + \dots)$ diverge e, conseqüentemente, a variância da série também diverge. Portanto a série com $|\alpha| \geq 1$ não é estacionária e não apresenta a propriedade de reversão à média. Outro ponto a se ressaltar é que para $|\alpha| < 1$, quanto mais recente a inovação maior é sua relevância para o modelo. De fato, pela eq.(21), vemos que a magnitude do fator da ℓ -ésima inovação decai exponencialmente.

Pode-se ver na tabela 3 o código em Python para geração de uma série AR(1) com $\alpha = 0,7$, além das figuras 5 e 6 onde há, respectivamente, o gráfico da série e da função de autocorrelação da série. Nestes nota-se claramente a propriedade de reversão à média, além do decaimento exponencial da autocorrelação.

Tabela 3 – Código em Python para geração da série AR(1)

```
# importação das bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf

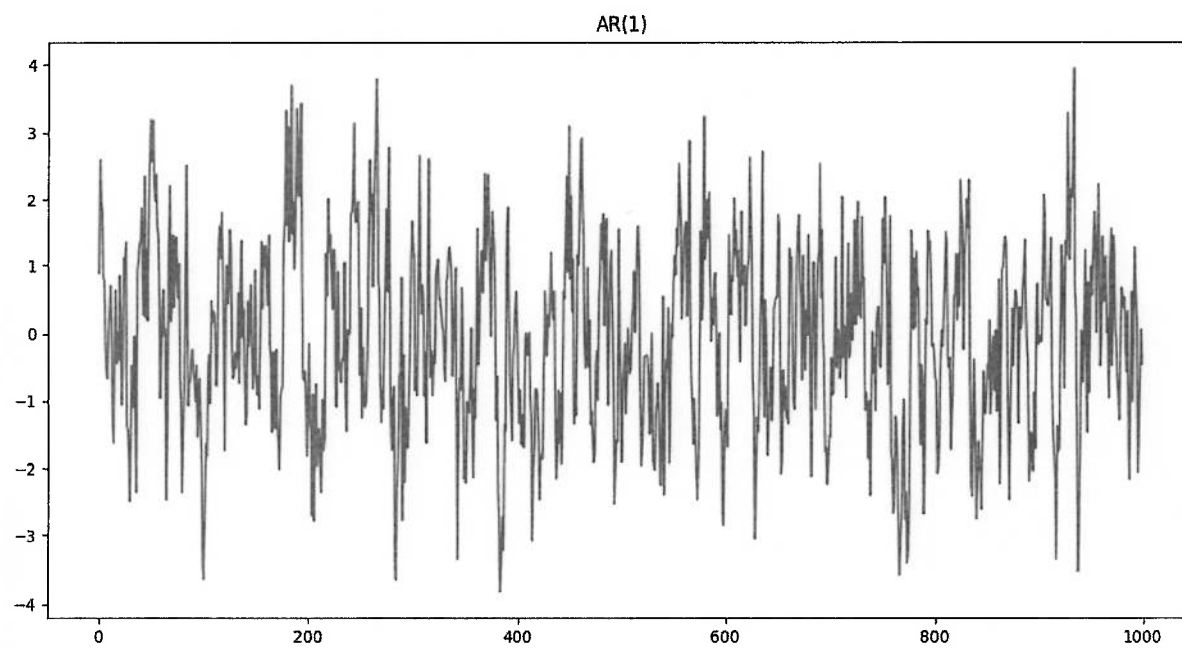
# definição dos parâmetros e tamanho da amostra
media = 0
desvio_padrao = 1
amostra = 10000          # tamanho da amostra deve ser suficientemente grande para aproximar de
                          # infinito
alfa = 0.7

# geração da série AR(1)
inovacao = np.random.normal(media, desvio_padrao, size=2 * amostra)
alfa_i = np.array([alfa ** i for i in range(amostra, -1, -1)])
AR_1 = np.empty((amostra,), float)
for i in range(amostra):
    AR_1[i] = np.sum(inovacao[i:amostra+1+i] * alfa_i)

# geração dos gráficos
plt.plot(AR_1)
plt.title('AR(1)')
plot_acf(AR_1, lags=30, title='Função de Autocorrelação da AR(1)')
plt.xlabel('passo')
plt.ylabel('autocorrelação')
plt.show()
```

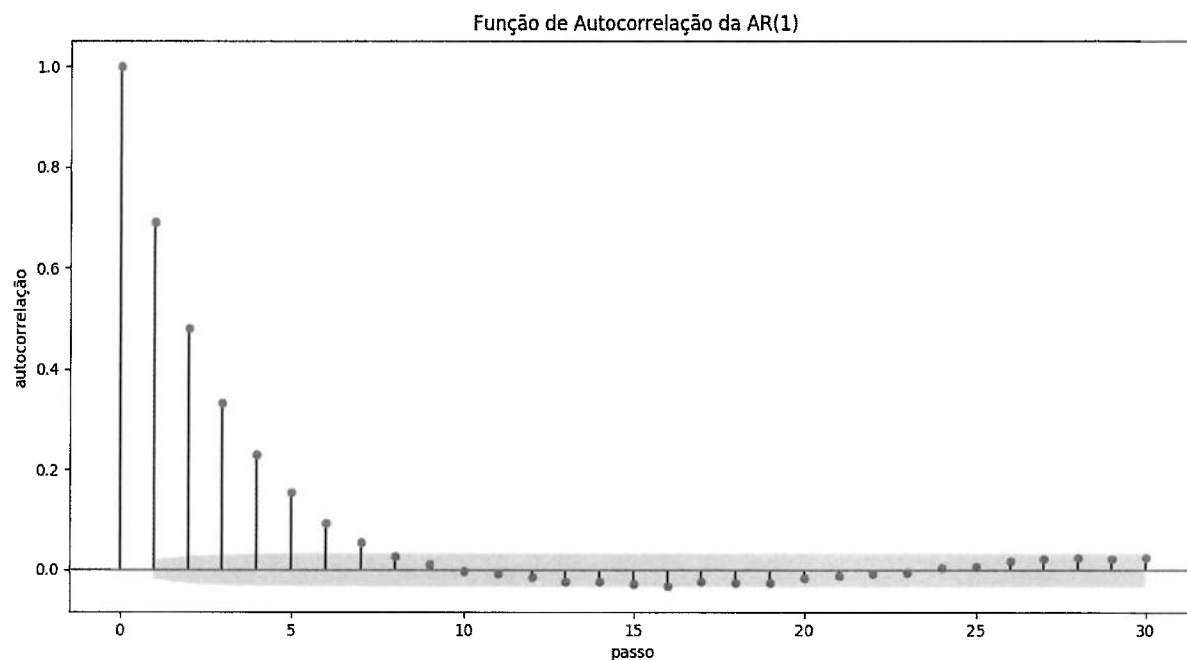
Fonte: Elaboração própria

Figura 5 – Série autorregressiva de primeira ordem



Fonte: Elaboração própria

Figura 6 – Função de autocorrelação da série autorregressiva de primeira ordem



Fonte: Elaboração própria

Finalmente, assim como foi possível em 2.1.3.2 estender o conceito de MA(1) para MA(q), é possível generalizar a série AR(1) para uma ordem $p > 1$, AR(p), através da adição à série de uma combinação linear dos valores da série nos p passos anteriores, de forma que obtemos a seguinte relação:

$$y_t = \varepsilon_t + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} . \quad (22)$$

2.1.3.4 Passeio aleatório

Passeio aleatório é um modelo de séries temporais importante por ser aplicável nas mais variadas áreas, desde biologia e física até finanças. Esta série é um caso especial da AR(1) onde $\alpha = 1$. Portanto tomando eq.(17) com $\alpha = 1$, obtemos sua relação:

$$y_t = \varepsilon_t + y_{t-1} . \quad (23)$$

Neste caso, considerando uma série finita $\{y_t | t = 1 \dots T\}$, a variância do passeio aleatório em um instante t é dada por:

$$\begin{aligned} Var(y_t) &= Var(\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \dots + \varepsilon_1) = Var(\varepsilon_t) + Var(\varepsilon_{t-1}) + Var(\varepsilon_{t-2}) + \\ &+ \dots + Var(\varepsilon_1) = t \sigma^2 . \end{aligned} \quad (24)$$

Ou seja, a variância do passeio aleatório cresce linearmente em função do tempo, de forma que a série pode assumir valores cada vez mais distantes da sua média e de seu valor inicial, ou seja, tal série não é estacionária e não apresenta a propriedade de reversão à média. Investiguemos agora seu valor esperado quando sabemos o valor imediatamente anterior:

$$E(y_t | y_{t-1}) = E(\varepsilon_t + y_{t-1} | y_{t-1}) = E(\varepsilon_t | y_{t-1}) + E(y_{t-1} | y_{t-1}) = y_{t-1} . \quad (25)$$

Isso quer dizer que, em um passeio aleatório, o valor atual da série é a melhor previsão que se pode fazer para o período posterior. Calculando sua autocorrelação a partir da eq.(21) e considerando uma série finita, obtemos:

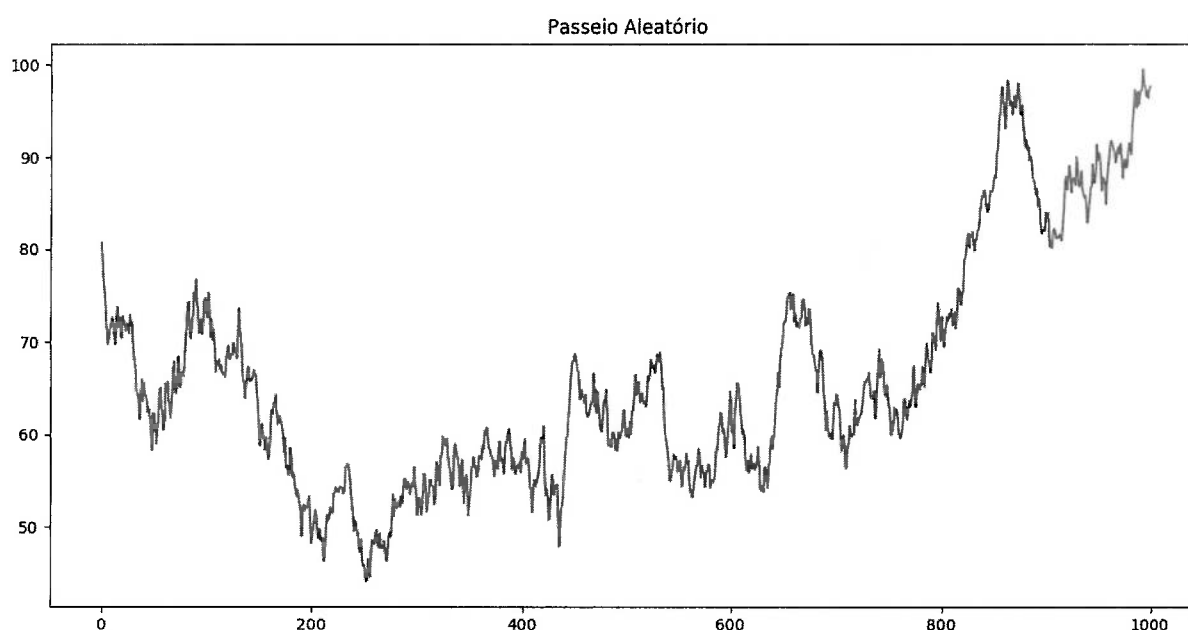
$$\rho_\ell = \frac{y_\ell}{y_0} = \frac{\alpha^\ell (1 + \alpha^2 + \alpha^4 + \dots) \sigma^2}{(1 + \alpha^2 + \alpha^4 + \dots) \sigma^2} = \alpha^\ell = 1, \text{ pois } \alpha = 1 . \quad (26)$$

Isso significa que todos os valores passados da série influenciam o valor atual do passeio aleatório. Os resultados da eq.(25) e eq.(26) podem ser observados, respectivamente, nas figuras 7 e 8 geradas a partir do código presente na tabela 3 com o parâmetro *alfa* igual a 1. Um vislumbre na figura 7 nos lembra de imediato um gráfico da evolução no tempo do preço de uma ação qualquer negociada nos mercados de bolsa de valores. De fato, é comum adotar o passeio aleatório para modelar a variação no tempo do preço (P) de ações do seguinte modo¹¹:

$$\ln P_t = \varepsilon_t + \ln P_{t-1}, \quad (27)$$

onde P_t é o preço da ação no instante t . A escolha de modelar o logaritmo do preço ao invés do preço diretamente se dá pelo princípio de *responsabilidade limitada*, que diz que a responsabilidade dos acionistas de sociedades anônimas é limitada ao preço de aquisição ou subscrição das ações destas¹².

Figura 7 – Passeio aleatório

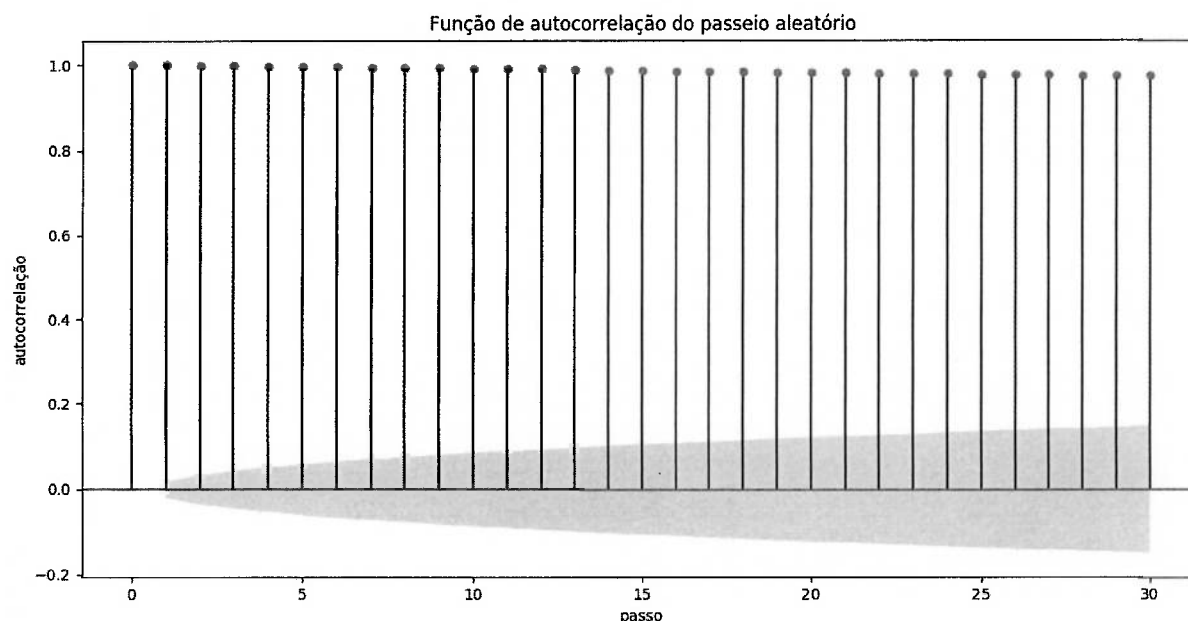


Fonte: Elaboração própria

¹¹ Um exemplo clássico é o modelo de apreçamento de opções de Black e Scholes (1973), útil no mercado financeiro até os dias atuais, que assume como premissa que o logaritmo do preço da ação que subscreve a opção objeto do modelo se comporta como um passeio aleatório.

¹² BRASIL, Código Civil, Lei No 6.404, 15 de Dezembro de 1976, Capítulo I, Art. 1º. **Características e Natureza da Companhia ou Sociedade Anônima.** Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L6404compilada.htm>. Acesso em: 01 fev. 2018.

Figura 8 – Função de autocorrelação do passeio aleatório



Fonte: Elaboração própria

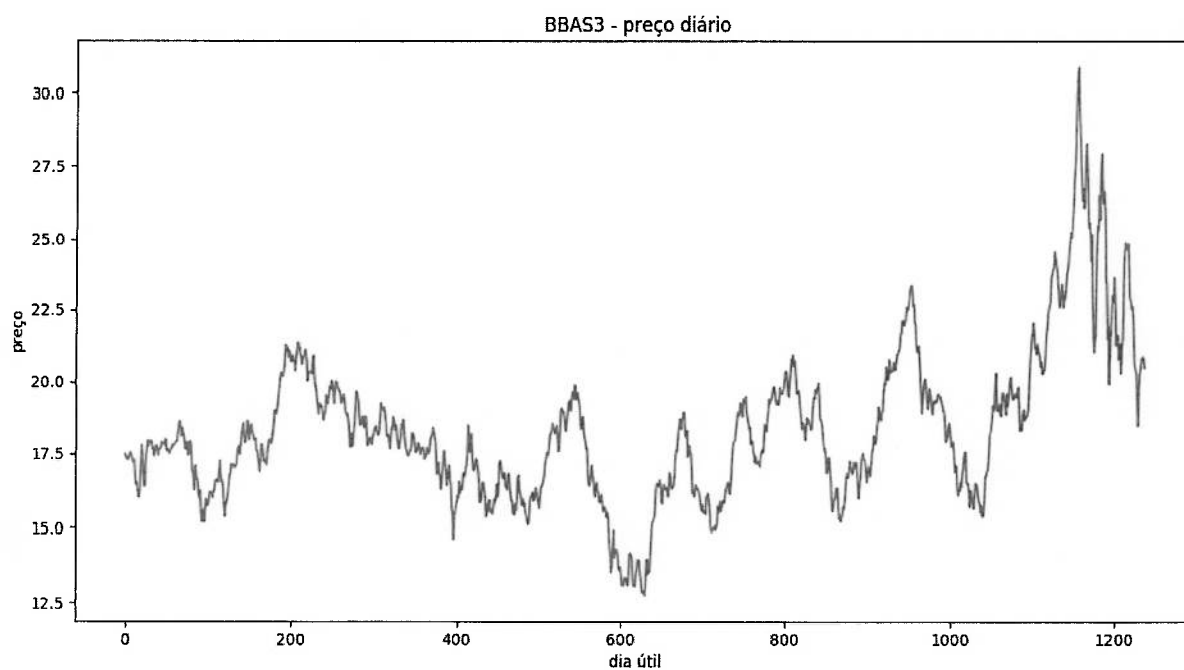
Isso quer dizer que o preço mínimo das ações é zero e, portanto, o modelo de passeio aleatório para o log-preço, também conhecido como modelo log-normal, faz mais sentido, uma vez que garante que o preço não assumirá em nenhum momento valores negativos. Além disso, ao tomarmos a aproximação $R_t \sim \ln(1 + R_t)$, onde R_t é o retorno da ação no instante t , junto à eq.(27), temos que:

$$R_t \sim \ln(1 + R_t) = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln(P_t) - \ln(P_{t-1}) = \varepsilon_t. \quad (28)$$

Ou seja, a primeira diferença da série de log-preços é uma série estacionária que equivale à série de retornos da ação, fato que vai ao encontro do que é observado empiricamente¹³ vide figuras 9 e 10, respectivamente, histórico de preços e retornos diários das ações do *Banco do Brasil S.A.* negociada na bolsa de valores de São Paulo sob o código *BBAS3*, entre os anos de 2010 a 2014. Quando a n -ésima diferença de uma série temporal não-estacionária é estacionária dizemos que esta série é integrada de ordem n , em síntese $I(n)$, portanto a série de log-preços acima é uma série integrada de primeira ordem, $I(1)$. Outro ponto a se destacar é que a relação presente na eq.(25), ou seja, de que a melhor previsão para um passeio

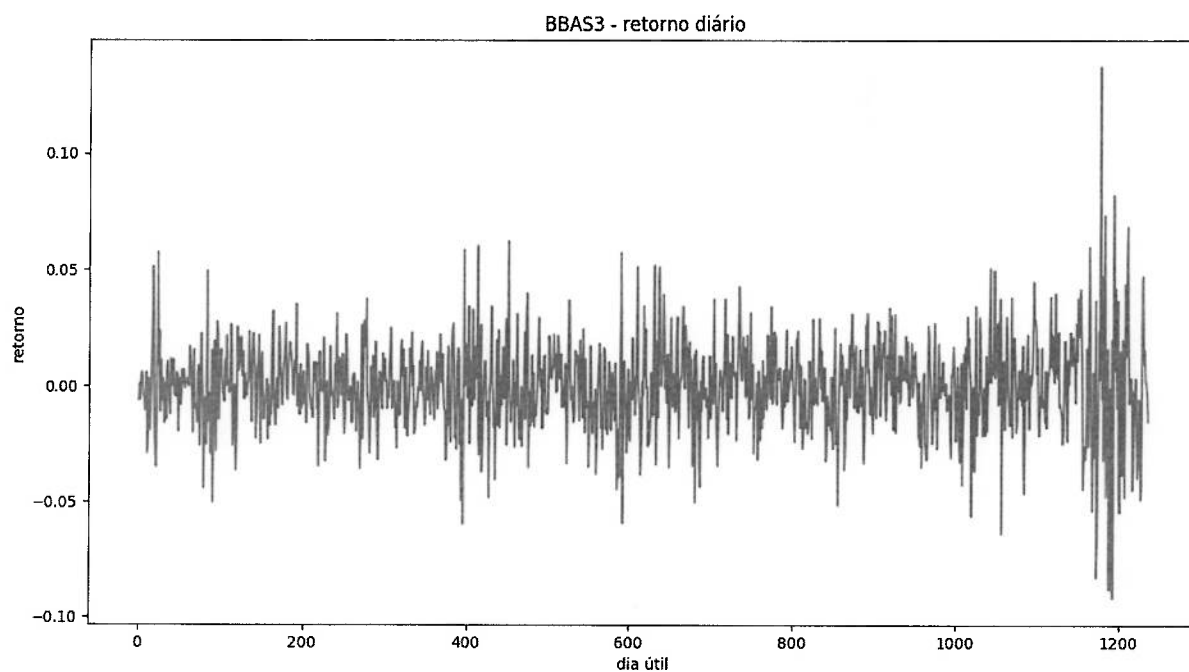
¹³ Detalhes de teste estatísticos de séries financeiras para esse fim podem ser encontrados em Zivot (2006) fora o trabalho pioneiro de Fama (1965).

Figura 9 – Histórico de preços diários de BBAS3



Fonte: Elaboração própria a partir de dados da BM&FBovespa – Cotações Históricas.

Figura 10 – Histórico de retornos diários de BBAS3



Fonte: Elaboração própria a partir de dados da BM&FBovespa – Cotações Históricas.

aleatório é seu valor atual, vai ao encontro da Hipótese de Mercados Eficientes elaborada por Fama (1965) conforme destacamos na introdução deste trabalho.

Apesar de descrever satisfatoriamente a evolução no tempo do preço e do retorno de uma ação, a característica de que a melhor previsão de preço para o período seguinte é o preço atual, isto é, retorno zero, torna o passeio aleatório inadequado em termos práticos para previsão de retorno de uma ação. Afinal, como destaca Vidyamurthy (2004), o objetivo ao modelarmos séries de preços de ações é auferir lucros com a previsão da mesma, de forma que um modelo em que o valor esperado do retorno é zero é inútil para esse fim. Entretanto veremos no próximo tópico que é possível a partir de uma combinação linear de séries não-estacionárias, tais como eq.(27), obter uma série estacionária. Esta, como vimos em 2.1.1, ao contrário do passeio aleatório, possui alta previsibilidade devido sua propriedade de reversão à média, tornando-se bastante útil no contexto de séries financeiras.

2.1.4 Cointegração

Segundo Zivot (2006), um conjunto de séries temporais não-estacionárias $I(1)$ é cointegrada quando existe uma combinação linear destas que é estacionária. Em notação vetorial, temos que um vetor $Y_t = (y_{1t}, \dots, y_{nt})'$, onde y_{it} , $i = 1, \dots, n$ é uma série temporal $I(1)$, é dito cointegrado quando existe um vetor $\beta = (\beta_1, \dots, \beta_n)'$ tal que:

$$\beta'Y_t = \beta_1 y_{1t} + \dots + \beta_n y_{nt} \sim I(0). \quad (30)$$

Este vetor não é único uma vez que para um escalar c qualquer temos que:

$$(c\beta)'Y_t = c(\beta_1 y_{1t} + \dots + \beta_n y_{nt}) \sim I(0). \quad (31)$$

Logo é necessário uma normalização para tornar a solução única. A normalização mais comum, e que também utilizaremos no modelo de investimento descrito no terceiro capítulo, é a seguinte:

$$\beta = (1, -\beta_2, \dots, -\beta_n)', \quad (32)$$

portanto,

$$\beta'Y_t = y_{1t} - \beta_2 y_{2t} - \dots - \beta_n y_{nt} \sim I(0), \quad (33)$$

que pode ser reescrito como

$$y_{1t} = \beta_2 y_{2t} + \dots + \beta_n y_{nt} + \varepsilon_t, \quad (34)$$

onde $\varepsilon_t \sim I(0)$ é o ruído branco conforme definido em 2.1.3.1. A interpretação econômica da eq.(34) é a de equilíbrio de longo prazo, a menos de ε_t , que seria o erro de desequilíbrio ou resíduo de cointegração. A ideia é que, no longo prazo, a relação entre as variáveis y_{it} , $i = 1, \dots, n$ acima é mantida pois as forças econômicas agem após os choques ε_t restaurando o equilíbrio. Dessa forma, neste trabalho, buscaremos séries de preços de ações y_{it} , $i = 1, \dots, n$, que mantêm uma relação de equilíbrio conforme eq.(34) e adotaremos uma estratégia de investimento que visa obter lucros a partir dos eventuais desequilíbrios desta relação.

A fim de solidificar conceitos finalizaremos esse tópico com um exemplo simples presente em Parreiras (2007) de duas séries temporais cointegradas. Seja $Y_t = (y_{1t}, y_{2t})$ um vetor de séries temporais não-estacionárias tais que:

$$y_{1t} = w_t + \varepsilon_{1t}, \quad (35)$$

$$y_{2t} = w_t + \varepsilon_{2t}, \quad (36)$$

onde,

$$w_t = w_{t-1} + \varepsilon_t, \quad (37)$$

$$\varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_t \sim iid. \quad (38)$$

Calculando-se a primeira diferença de cada uma tem-se,

$$\begin{aligned} \Delta y_1 &= y_t^{(1)} - y_{t-1}^{(1)} = (w_t + \varepsilon_t^{(1)}) - (w_{t-1} + \varepsilon_{t-1}^{(1)}) = (w_{t-1} + \varepsilon_t + \varepsilon_t^{(1)}) - (w_{t-1} + \varepsilon_{t-1}^{(1)}) = \\ &= \varepsilon_t + \varepsilon_t^{(1)} - \varepsilon_{t-1}^{(1)} \Rightarrow \Delta y_1 \sim I(0) \Rightarrow y_1 \sim I(1), \end{aligned} \quad (37)$$

$$\begin{aligned} \Delta y_2 &= y_t^{(2)} - y_{t-1}^{(2)} = (w_t + \varepsilon_t^{(2)}) - (w_{t-1} + \varepsilon_{t-1}^{(2)}) = (w_{t-1} + \varepsilon_t + \varepsilon_t^{(2)}) - (w_{t-1} + \varepsilon_{t-1}^{(2)}) = \\ &= \varepsilon_t + \varepsilon_t^{(2)} - \varepsilon_{t-1}^{(2)} \Rightarrow \Delta y_2 \sim I(0) \Rightarrow y_2 \sim I(1). \end{aligned} \quad (38)$$

Por fim, calculando a diferença entre as duas séries, vê-se,

$$y_1 - y_2 = (w_t + \varepsilon_{1t}) - (w_t + \varepsilon_{2t}) = \varepsilon_{1t} - \varepsilon_{2t} \Rightarrow y_1 - y_2 \sim I(0). \quad (39)$$

Portanto o vetor $Y_t = (y_{1t}, y_{2t})$ é cointegrado com vetor de cointegração $\beta = (1, -1)$.

2.2 OPERAÇÕES DO MERCADO FINANCEIRO

Nesta seção revisaremos, conforme Hull (2009), as três operações fundamentais que há no mercado financeiro, quais sejam, cobertura, especulação e arbitragem, com especial atenção à última, uma vez que as operações de arbitragem estatística, tema deste trabalho, podem ser estudadas como uma generalização das operações de arbitragem, segundo veremos nos próximos tópicos.

2.2.1 Cobertura

O primeiro tipo de operação, cobertura¹⁴, é a principal motivação para a criação do mercado derivativos, isto é, mercado de instrumentos financeiros que derivam de outros ativos financeiros que os subscrevem. Neste tipo de operação o negociador objetiva reduzir o risco relacionado a alguma variável de mercado a qual esteja exposto.

Os principais tipos de derivativos são quatro: contratos futuros, contratos a termo, opções e *swaps*. Os dois primeiros tratam-se de um acordo entre duas partes para negociar um ativo a um preço pré-determinado em uma data futura pré-determinada. A diferença entre contratos futuros e a termo deve-se a maior padronização do primeiro que é negociado exclusivamente em mercados organizados, onde há maior liquidez e garantia de liquidação por parte do administrador do mercado. As opções, por sua vez, são contratos que conferem ao titular a opção de comprar (no caso de opção de compra) ou vender (no caso de opção de venda) determinado ativo a um preço específico (preço de exercício) em uma data futura específica (se opção do tipo europeia) ou em qualquer momento até

¹⁴ Usualmente denominada no mercado brasileiro em sua versão na língua inglesa: *hedge*.

uma data específica (se opção do tipo americana), data esta de vencimento da opção. Na hipótese do titular exercer a opção, o lançador desta tem a obrigação de vender ao titular (no caso de opção de compra) ou comprar do titular (no caso de opção de venda) o ativo que subscreve a opção pelo preço de exercício. Em contrapartida, o lançador da opção no momento do lançamento da mesma recebe um valor monetário do titular (chamado de prêmio da opção). Por último, os *swaps* são acordos entre duas partes de troca de fluxo de caixa em uma ou mais datas futuras. O cálculo do fluxo de caixa tem como base índices de mercado, tais como taxa de juros, cotação de câmbio de moedas, taxa pré-fixada, entre outros.

Não discorreremos sobre maiores detalhes de cada um dos derivativos por fugir do escopo deste trabalho¹⁵. No entanto, ilustraremos a seguir como esses contratos podem ser utilizados para operações de cobertura através de um exemplo com contrato futuro de dólar.

Imaginemos uma empresa brasileira cuja principal fonte de receita seja a venda de produtos no exterior em dólares de forma que sua receita é fortemente correlacionada com a cotação do dólar em reais. Suponhamos ainda que por motivos macroeconômicos a direção da companhia acredite que exista uma probabilidade significativa de que ao longo do ano vigente a cotação do dólar em relação ao real descenderá impactando negativamente o resultado anual da empresa. Uma forma da companhia se proteger da exposição ao risco de mercado da cotação do dólar em reais é vender uma quantidade de contratos futuros de taxa de câmbio de reais por dólar comercial, negociados na Bolsa de Mercadorias e Futuros (BM&F)¹⁶, que represente um volume financeiro mais próximo possível da receita anual esperada. Dessa forma a empresa passa a estar neutra em relação à oscilação do dólar, pois caso o cenário esperado pela direção se concretize e a cotação do dólar em reais caia, a empresa terá uma redução da receita em reais, porém obterá um lucro de mesma magnitude devido à operação de venda de contratos futuros. Por outro lado, na hipótese da cotação do dólar subir, o prejuízo

¹⁵ Ao leitor que tiver maior curiosidade sobre o assunto recomendamos Hull (2009).

¹⁶ Para maiores detalhes deste contrato ver: BM&FBovespa – **Futuro de taxa de câmbio de reais por dólar comercial**. Disponível em: http://www.bmfbovespa.com.br/pt_br/produtos/listados-a-vista-e-derivativos/moedas/futuro-de-taxa-de-cambio-de-reais-por-dolar-comercial.htm. Acesso em: 22 de dez. 2017.

que a empresa obter na operação de venda de contratos futuros será neutralizado pelo aumento não esperado da receita anual em reais.

2.2.2 Especulação

A operação de especulação é realizada com objetivo oposto à operação de cobertura: enquanto que nesta o agente busca se proteger de algum risco de mercado ao qual esteja exposto, naquela o negociador se expõe aos riscos de um determinado ativo por acreditar que o preço do mesmo aumentará (posição comprada) ou decrescerá (posição vendida).

Existem inúmeras formas de especular no mercado de bolsa de valores, umas mais conservadoras, como aquisições de ações de companhias sólidas com boa liquidez¹⁷ no mercado à vista, outras mais agressivas, como, por exemplo, operações alavancadas¹⁸ no mercado a termo, futuro ou de opções. Em geral, podemos dizer que uma carteira de ativos bem administrada combina operações de especulação com operações de cobertura.

2.2.3 Arbitragem

Por fim a operação de arbitragem se constitui de transações simultâneas em dois ou mais mercados de forma que o agente consegue arbitrar um lucro sem risco devido a eventuais desvios de preços teóricos de um ativo ou grupo de ativos negociados em mercados diferentes. Com objetivo de solidificar o conceito mostraremos em seguida dois exemplos clássicos de operação de arbitragem.

O primeiro consta em Hull (2009) e trata-se de uma ação que é negociada em duas bolsas de valores distintas, quais sejam, *New York Stock Exchange*¹⁹ e *London*

¹⁷ Liquidez, em finanças, trata-se da facilidade com que um agente econômico consegue monetizar um determinado ativo, isto é, encontrar alguém disposto a comprá-lo.

¹⁸ Nas operações alavancadas o agente toma uma posição cujo volume financeiro é maior que o capital que o mesmo dispõe.

¹⁹ **New York Stock Exchange**. Disponível em: <www.nyse.com>. Acesso em: 22 dez. 2017.

*Stock Exchange*²⁰. Suponhamos que em um dado momento o preço dessa ação seja US\$200,00 (duzentos dólares) em Nova York e £100,00 (cem libras) em Londres quando a cotação do dólar por libra é US\$2,0300/£. Dessa forma um arbitrador poderia simultaneamente comprar um lote de 100 ações em Nova York e vender um lote de 100 ações em Londres obtendo, portanto, um lucro sem risco de $100 \times [(US\$2,03/\text{£} \times \text{£}100,00) - US\$200,00]$, ou seja, US\$300,00 (trezentos dólares) por lote de ação, sem considerar os custos de transação.

O segundo exemplo encontra-se em Parreiras (2007) e apresenta uma arbitragem de índice, ou seja, uma operação simultânea de compra (ou venda) no mercado à vista das ações que constituem um determinado índice de mercado²¹ e venda (ou compra) no mercado futuro do contrato futuro do referido índice²², com objetivo de lucrar com eventual desvio de preço teórico entre os dois. A relação teórica de preço entre o contrato futuro de índice e o grupo de ações que o constitui é dada por²³:

$$F_t = \sum_i \omega_i S_t^i e^{(r-q_i)(T-t)}, \quad (40)$$

onde F_t é o preço do contrato futuro no dia t ; S_t^i é o preço da i -ésima ação que compõe o índice no dia t ; ω_i é o peso da i -ésima ação na composição do índice; r é a taxa de juros livre de risco²⁴; q_i é a taxa contínua de dividendos da i -ésima ação²⁵;

²⁰ **London Stock Exchange**. Disponível em: <www.londonstockexchange.com>. Acesso em: 22 dez. 2017.

²¹ Índice de mercado é um portfólio teórico de ações com objetivo de indicar o desempenho das empresas de algum setor da economia ou da economia de um país como um todo, como, por exemplo, o Índice Bovespa, principal índice de ações do Brasil, composto por ações listadas na BM&FBovespa que atendem critérios de liquidez.

²² Um exemplo de contrato futuro de índice: BM&FBovespa – **Futuro de Ibovespa**. Disponível em: <http://www.bmfbovespa.com.br/pt_br/produtos/listados-a-vista-e-derivativos/renda-variavel/futuro-de-ibovespa.htm>. Acesso em: 22 dez. 2017.

²³ Maiores detalhes da teoria por trás da formação de preços dos contratos futuros podem ser encontrados em Hull (2009).

²⁴ Na economia brasileira considera-se taxa livre de risco a taxa SELIC, taxa base do financiamento público federal, cuja meta é determinada em reuniões periódicas pelo Comitê de Política Monetária (COPOM) composto pelo Presidente e Diretores do Banco Central do Brasil.

e T é o dia de vencimento do contrato futuro. Dessa forma, se ocorrer um desvio de preço entre os dois termos da eq.(40) maior que os custos de transação, um arbitrador pode aproveitar a oportunidade para obter lucros sem risco comprando o termo que estiver depreciado e vendendo o termo que estiver apreciado.

Note que oportunidades de arbitragem como as expostas acima quando surgem tendem a desaparecer rapidamente uma vez que os agentes do mercado, em especial grandes bancos de investimentos e gestoras de recursos que possuem menores custos de transação, as percebem e se posicionam para obterem lucros através dos desvios de preço apresentados. Por esse motivo, muitos modelos matemáticos em finanças, como o conhecido modelo de apreçamento de opções de Black e Scholes (1973), assumem como pressuposto a ausência de oportunidades de arbitragem.

É importante ressaltar que a expressão “sem risco” utilizada para descrever esse tipo de operação significa que não há risco teórico pois trata-se de uma operação que obtém vantagem de eventuais desvios de preços teóricos entre os ativos. No entanto, existem riscos nesse tipo de operação, em especial risco operacional, pois muitas vezes não é possível se posicionar da maneira exata requerida pela relação teórica de preços entre os ativos. Na eq.(40) por exemplo, é exigido que o arbitrador se posicione com peso ω_i para cada ação i que compõe o índice, valor que não é necessariamente um múltiplo de 100, fator complicador considerando que as ações são negociadas no mercado brasileiro em lotes de 100 unidades. Ainda na eq.(40) tem-se como fator dificultador a taxa de dividendos q_i que é uma estimativa. Outro ponto a se considerar é a dificuldade que pode surgir, por questões de liquidez dos ativos, para atender com precisão suficiente o requisito de simultaneidade entre as transações de compra do ativo ou grupo de ativos depreciados e venda do ativo ou grupo de ativos apreciados.

Uma vez solidificados os conceitos com os dois exemplos anteriores, podemos definir de maneira geral uma operação de arbitragem sem risco teórico.

²⁵ Dividendos são parcelas do lucro de uma companhia distribuídas aos acionistas, em geral com frequência anual, semestral ou trimestral. Taxa contínua de dividendos é uma aproximação do percentual de dividendos que uma ação remunera seu titular de forma contínua no tempo.

Conforme exposto em Parreiras (2007), tem-se que uma operação de arbitragem sem risco é possível quando a relação abaixo não é atendida:

$$|X_t - AS(X_t)| < CT, \quad (41)$$

onde X_t é preço do ativo ou grupo de ativos base da operação; $AS(X_t)$ é o preço do ativo sintético que replica o ativo base da operação; e CT são os custos das transações.

Portanto a implementação de uma operação de arbitragem sem risco é composta de três passos: construção de relações teóricas de preços entre os ativos, isto é, $X_t = AS(X_t)$; identificação de desvios de preços acima dos custos de transação, ou seja, quando a relação na eq.(41) não é atendida; e, por último, se posicionar no mercado de acordo com o desvio observado, quer dizer, comprando o ativo depreciado e vendendo o ativo apreciado.

2.3 ARBITRAGEM ESTATÍSTICA

Como adiantamos em 2.2 arbitragem estatística nada mais é do que uma extensão do conceito de arbitragem teórica. Enquanto nesta o primeiro passo para implementar a estratégia de investimento é construir uma **relação teórica** de preços entre os ativos financeiros, naquela o primeiro passo consiste em construir uma **relação estatística** de preços entre os ativos financeiros. A partir daí, o segundo e terceiro passos são os mesmo para ambos, ou seja, identificar os desvios de preços conforme a relação de preços construída e, por fim, se posicionar no mercado de acordo com o desvio observado.

Dessa forma tem-se que uma operação de arbitragem estatística é possível quando a seguinte relação não é obedecida:

$$|E(X_t) - AS(X_t)| < CT, \quad (42)$$

onde $E(X_t)$ é o valor esperado do preço do ativo ou grupo de ativos base da operação; $AS(X_t)$ é o preço do ativo sintético que replica o ativo base da operação; e CT são os custos de transação. Note que a diferença entre a eq.(41) e a eq.(42) é sutil porém fundamental: na primeira consideramos diretamente o preço do ativo base X_t ao passo que na última leva-se em conta o valor esperado do preço do ativo

base $E(X_t)$, consequência da construção estatística do modelo. Por conseguinte vejamos a seguir as três etapas supracitadas para executar a estratégia de arbitragem estatística: construção da relação estatística de preços conforme eq.(42); identificação dos desvios de preço através de um teste de estacionariedade; e posicionamento no mercado consoante ao desvios identificados.

2.3.1 Construção da relação estatística de preços

A fim de construir relações estatísticas de preços $E(X_t) = AS(X_t)$, conforme eq.(42), de forma a obter desvios de preços $D_t = X_t - AS(X_t)$ estacionários, empregaremos o procedimento desenvolvido por Burgess (1999), também utilizado por Parreiras (2007), que consiste em, dado um conjunto de ativos, para cada um deles, realizar uma regressão linear da série de preços do mesmo contra as séries de preços dos ativos remanescentes do conjunto com o objetivo de obter ao final do processo uma relação análoga à eq.(34) para cada um dos ativos do conjunto. No entanto tal regressão não é realizada de forma direta considerando toda a base de ativos, ao contrário, ela é efetuada passo a passo²⁶ da seguinte forma: inicia-se o procedimento executando uma regressão linear da série de preços do ativo objeto contra a série do ativo do conjunto que apresentar maior correlação com o ativo objeto; em seguida escolhe-se dentre os ativos remanescentes do conjunto, aquele cuja série de preços apresentar maior correlação com os resíduos da regressão anterior, isto é, com os desvios de preço resultantes da regressão anterior; ato contínuo, realiza-se uma nova regressão linear da série de preços do ativo objeto contra essas novas séries dos dois ativos selecionados anteriormente; então, escolhesse mais um ativo de maneira análoga ao anterior, repete-se a regressão considerando-se esse novo ativo, e assim sucessivamente até um critério de parada que pode ser um número máximo de ativos definido para explicar o ativo objeto ou uma significância estatística alcançada.

Uma interpretação econômica destacada por Burgess (1999) para o método acima é a de que no primeiro passo escolhe-se o ativo do conjunto que possui maior

²⁶ Burgess (1999) alcunhou por *stepwise regression* esse tipo de regressão.

semelhança ao ativo objeto no que tange às exposições de risco de mercado dos mesmos. Posteriormente, uma vez que a tendência estocástica foi eliminada no primeiro passo, escolhe-se os ativos que possuem maior correlação com as exposições de riscos residuais que permaneceram dos passos anteriores.

Em resumo, dado um conjunto A de ativos financeiros e um conjunto X de séries de preços tal que cada $x \in A$ possui uma série de preços associada $x_t \in X$, $t = 1, 2, \dots, T$, a fim de obter uma relação de preços conforme eq.(34), adotaremos o seguinte algoritmo:

- 1) $C^{(0)} = \emptyset$, $d_t^{(0)} = x_t$, $k = 0$;
- 2) $C^{(k+1)} = C^{(k)} \cup \{\operatorname{argmax}_{c_t \in X - \{x_t\} - C^{(k)}} \operatorname{Corr}(d_t^{(k)}, c_t)\}$;
- 3) $d_t^{(k+1)} = x_t - \sum_{c_{i,t} \in C^{(k+1)}} \beta_i^{(k+1)} c_{i,t}$, onde $\beta^{(k+1)} = \operatorname{argmin}_{\beta_i^{(k+1)}} \sum_{t=1}^T d_t^{(k+1)^2}$;
- 4) $k = k + 1$;
- 5) Enquanto $k \leq L$, voltar ao segundo passo; onde L é o limite de ativos arbitrado para formar o ativo sintético $AS(x_t)$ de x_t .

Os ativos $c \in A$ cujas séries de preços associadas c_t são tais que $c_t \in C^{(L)}$ serão os escolhidos para formar o ativo sintético $AS(x_t)$ e, por conseguinte, a série de desvios de preço do ativo objeto x_t :

$$d_t = x_t - AS(x_t) = x_t - \sum_{i=1}^L \beta_i^{(L)} c_{i,t}. \quad (43)$$

Note que o vetor $(1, -\beta_1^{(L)}, -\beta_2^{(L)}, \dots, -\beta_L^{(L)})$ obtido no procedimento acima é o potencial vetor de cointegração normalizado (ver 2.1.4, eq.(32)) da série $(x_t, c_{1,t}, c_{2,t}, \dots, c_{L,t})$. No tópico imediatamente posterior veremos como testar se cada série de desvios de preço formada conforme eq.(43) é estacionária com um nível de certeza arbitrado, ou seja, como verificar se cada vetor $(1, -\beta_1^{(L)}, -\beta_2^{(L)}, \dots, -\beta_L^{(L)})$ é de fato um vetor de cointegração (para um certo nível de certeza). Os desvios que passam no teste são escolhidos para o posicionamento de mercado, enquanto os que falham são desconsiderados.

2.3.2 Identificação dos desvios de preço

Uma vez construída as relações de preço $E(x_t) = AS(x_t)$ devemos identificar quais desvios de preço $d_t = x_t - AS(x_t)$ são estacionários e, por consequência, quais vetores $(X_t, AS(X_t))$ são cointegrados, pois estes serão utilizados na estratégia de investimento a definir no próximo tópico. Por outras palavras, testaremos a hipótese nula $H_0 : d_t \sim I(1)$ contra a hipótese alternativa $H_1 : d_t \sim I(0)$. O teste de estacionariedade mais comum, conforme ressaltado por Burgess (1999), é o teste de Dickey-Fuller (1976) que é realizado por meio de uma regressão da série de interesse contra ela própria deslocada por um período, isto é,

$$d_t = \alpha + \beta d_{t-1} + \varepsilon_t, \quad (44)$$

caso β não seja significativamente menor do que um, a hipótese nula não é rejeitada e deve-se fazer a regressão tomando-se a primeira diferença da série,

$$d_t - d_{t-1} = \Delta d_t = \alpha + (\beta - 1)d_{t-1} + \varepsilon_t = \alpha + \bar{\beta}d_{t-1} + \varepsilon_t, \quad (45)$$

e testa-se então se $\bar{\beta}$ é significativamente menor do que zero, caso em que d_t seria uma série estacionária. Há ainda uma extensão do teste acima denominada por *Dickey-Fuller aumentado* (ADF)²⁷ onde se considera deslocamento maiores da série a fim de eliminar a autocorrelação entre os erros:

$$\Delta d_t = \alpha + \beta d_{t-1} + \beta_1 \Delta d_{t-1} + \beta_2 \Delta d_{t-2} + \dots + \beta_n \Delta d_{t-n} + \varepsilon_t, \quad (46)$$

onde testa-se também se β é significativamente menor que zero para rejeitar a hipótese de não-estacionariedade.

Vale destacar que existem outros testes de estacionariedade como, por exemplo, Philips (1987), Philips e Perron (1988) e Hall (1989), no entanto, como destaca Burgess (1999), os indícios são de que tais testes não apresentaram melhora em relação ao ADF, de forma que para nossos propósitos seguiremos com o teste ADF e não discutiremos sobre os demais. Em Python, no módulo

²⁷ Abreviatura do termo em inglês *Augmented Dickey-Fuller*, tradução sugerida por Parreiras (2007).

statsmodels, há a função *adfuller*²⁸ para a realização do teste. Seus principais argumentos são: *x*, a série a ser testada (único argumento obrigatório da função); *maxlag*, argumento opcional, define o número máximo de deslocamentos (o padrão é $\sqrt[4]{12 \times \text{obs}/100}$, onde *obs* é o número de observações da amostra)²⁹; *regression*, também opcional, determina o tipo de regressão a ser realizada, se é apenas com uma constante ('c'), ou com constante somada a tendência linear ('ct'), ou com constante, tendência linear e tendência quadrática ('ctt'), ou se é sem constante e sem tendência ('nc'), sendo que o padrão da função é a opção 'c'; e *autolag*, argumento que possibilita definir o número de deslocamentos da série através de teste t ('t-stat') ou critério de informação bayesiano ('BIC') ou ainda critério de informação de Akaike ('AIC') (se esse argumento for deixado vazio, a função *adfuller* utiliza a quantidade de deslocamentos definida em *maxlag*)³⁰. Em contrapartida, a função retorna a estatística do teste ('adf')³¹, seu valor-p ('pvalue')³², o número de deslocamentos e observações utilizados ('usedlag' e 'nobs'), além dos valores críticos para os níveis de significância de 1%, 5% e 10% conforme MacKinnon (2010)³³.

Para melhor entendimento, finalizaremos o tópico aplicando o teste ADF, a título de exemplo, em uma série temporal de desvios de preços: observa-se nas tabelas 4 e 5, respectivamente, o código em Python para realização do teste ADF com todos os argumentos opcionais na opção padrão; e a saída dos retornos da função *adfuller*. Note que a estatística do teste é de -4,61, valor menor que o valor

²⁸ *Statsmodels. Adfuller*. Disponível em:

<<http://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html>>. Acesso em: 22 dez. 2017.

²⁹ Maiores detalhes sobre a fórmula para determinar o número máximo de deslocamentos ideal para o teste ADF podem ser encontrados em Greene (2012).

³⁰ Idem nota 29.

³¹ A estatística do teste ADF é o resultado do quociente entre o parâmetro a estimar β da eq.(46) e seu erro padrão, que, por sua vez, é o resultado do quociente entre o desvio padrão da estimativa de β e a raiz quadrada do número de observações da amostra.

³² Assumindo que a hipótese nula é verdadeira, o valor-p fornece a probabilidade de que a estatística do teste assuma os valores observados, ou seja, quanto menor o valor-p, maior é o nível de certeza para que se rejeite a hipótese nula.

³³ O valor crítico para um nível de significância p é o qual a estatística do teste de hipótese deve superar para que se possa rejeitar a hipótese nula com nível de certeza de $1-p$.

crítico para significância de 1% (-3,43), ou seja, podemos concluir que a série temporal exemplificada é estacionária com nível de certeza de pelo menos 99%. Com efeito, ainda na tabela 5, vê-se que o valor-p da série é de 0,012%, quer dizer, a série temporal é estacionária com nível de certeza de 99,988%. Por fim podemos verificar também que havia 1729 observações na amostra e que foi três o número de deslocamentos utilizado para o teste.

Tabela 4 – Código em Python para teste ADF

```
# importação das bibliotecas necessárias

import numpy as np
from statsmodels.tsa.stattools import adfuller

# abertura do arquivo com os dados da série e definição do respectivo vetor

arq1 = 'serie.csv'
serie = np.asarray(np.loadtxt(arq1, delimiter=',')/100)

# realização do teste ADF com todas os argumentos opcionais na opção padrão

adf = adfuller(serie)

# saída dos resultados do teste
print('estatística: %s\nvalor-p: %s\nnúmero de deslocamentos: %s\nnúmero de observações: '
      '%s\nvalores críticos: %s' % (adf[0], adf[1], adf[2], adf[3], adf[4]))
```

Fonte: Elaboração própria

Tabela 5 – Saída dos retornos do teste ADF

```
estatística: -4.61033604569
valor-p: 0.000123620438235
número de deslocamentos: 3
número de observações: 1729
valores críticos: {'1%': -3.4341377588624709, '5%': -2.8632130839838936, '10%':
-2.567660702509265}
```

Fonte: Elaboração própria

2.3.3 Posicionamento no mercado

Cumprida a primeira e segunda etapas, respectivamente, construir as relações de preço $E(x_t) = AS(x_t)$ e identificar quais desvios $d_t = x_t - AS(x_t)$ são estacionários, devemos estabelecer uma estratégia de investimento para nos posicionarmos de forma adequada em relação aos desvios de preços selecionados.

A estratégia que descreveremos a seguir foi adotada por Burgess (1999) e Parreiras (2007) e consiste, em primeiro lugar, se posicionar em cada período para cada desvio selecionado, de forma oposta ao desvio observado e com peso proporcional ao tamanho do desvio elevado a um fator $k > 0$, tal que: se $k > 1$, dar-se-á maior peso aos maiores desvios; se $k = 1$, o peso será linearmente proporcional ao tamanho dos desvios; se $k < 1$, os menores desvios ganharão peso relativamente aos maiores. Ou seja, a posição p em cada período de tempo t , para cada desvio d , é obtida por:

$$p_t = -\text{sin}(\text{al}(d_{t-1})) |d_{t-1}|^k, \quad (47)$$

onde $\text{sin}(x)$ é uma função que retorna -1 se $x < 0$, 0 se $x = 0$, e 1 se $x > 0$. Dessa forma, o volume total no período t , isto é, a soma das posições absolutas de cada ativo, considerando a possibilidade de alavancagem³⁴, é o seguinte:

$$v_t = \frac{x_t + AS(x_t)}{\text{alavancagem}} = \frac{2x_t - d_t}{\text{alavancagem}}, \quad (48)$$

portanto, o retorno em cada período t é dado por:

$$r_t = p_t \times \frac{(d_t - d_{t-1})}{v_{t-1}} - c \times |p_t - p_{t-1}|, \quad (49)$$

onde c é uma aproximação dos custos de transação. Por conseguinte tem-se o retorno acumulado ao fim de um período $t = 1 \dots T$ ³⁵:

$$R_t = \sum_{t=1}^T r_t, \quad (50)$$

e, por fim, tem-se que o retorno do portfólio em um período t é dado pela média dos retornos de cada desvio em cada período. Além da alavancagem, outra estratégia que poderemos adotar é a de interrupção de perda, amplamente conhecida no mercado financeiro pela alcunha em inglês *stop loss*. Nela, o investidor define um limite de retorno negativo que o mesmo aceita em uma determinada operação e uma vez que tal operação eventualmente alcance esse retorno negativo, ela é finalizada, de forma a evitar maiores perdas.

³⁴ Idem nota 18.

³⁵ Note que não estamos compondo o retorno pois considera-se capital fixo, ou seja, os lucros dos períodos anteriores não são reinvestidos.

3 IMPLEMENTAÇÃO DO MODELO

Com base na fundamentação teórica exposta no capítulo anterior, em especial tópico 2.3 que é a base do algoritmo desenvolvido em Python no item 3.2, simularemos um modelo de investimento de arbitragem estatística aplicado ao mercado de ações brasileiro no período entre 2015 e 2017, por meio dos dados históricos de preços a partir de 2010. Ao final discutiremos os resultados obtidos.

3.1 COLETA DE DADOS

Felizmente, ao contrário de outras áreas de aplicação de análise estatística, a coleta de dados de preços diários de ações de companhias listadas em bolsas de valores em geral não é trabalhosa. No Brasil há apenas uma bolsa de valores, a *BM&FBovespa*, e esta fornece as séries de preços diários das ações de forma estruturada em seu endereço eletrônico³⁶, de onde obtivemos os dados que utilizaremos no algoritmo descrito no próximo tópico.

A simulação da estratégia de investimento se dará no período de 2015 a 2017 utilizando dados a partir de 2010 para formar os desvios de preços estacionários. Ainda, escolhemos trabalhar com as ações que compunham o Índice Brasil 100 (IBrX 100)³⁷ na data de 22 de dezembro de 2017 e que já estavam sendo negociadas na bolsa de valores de São Paulo em janeiro de 2010, ou seja, ativos que possuem histórico de preços suficiente para que possamos inferir se são cointegrados com outras séries. A escolha desse índice se deu por ele ser indicador do desempenho médio das 100 ações listadas na BM&FBovespa de maior negociabilidade e representatividade, isto é, para que uma ação faça parte dele ela deve atender critérios de liquidez, característica essencial para implementação da metodologia de arbitragem estatística. Dessa forma, chegamos a uma lista de 66 ativos que formarão o conjunto A definido em 2.3.1, cujos códigos de negociação na

³⁶ Ver BM&FBovespa – **Cotações históricas**.

³⁷ BM&FBovespa – **Índice Brasil 100**. Disponível em:
<http://www.bmfbovespa.com.br/pt_br/produtos/indices/indices-amplos/indice-brasil-100-ibrx-100.htm>. Acesso em: 22 dez. 2017.

BM&FBovespa são³⁸: BBAS3, BBDC3, BBDC4, BEEF3, BRAP4, BRFS3, BRML3, BRSR6, BTOW3, BVMF3, CCRO3, CESP6, CIEL3, CMIG4, CPFE3, CPLE6, CSAN3, CSMG3, CSNA3, CYRE3, DTEX3, ELET3, ELET6, EMBR3, ENBR3, EQTL3, ESTC3, EZTC3, FIBR3, FLRY3, GFSA3, GGBR4, GOAU4, GOLL4, GRND3, HGTX3, HYPE3, IGTA3, ITSA4, ITUB4, LAME3, LAME4, LIGT3, LREN3, MDIA3, MRFG3, MRVE3, MULT3, MYPK3, NATU3, ODPV3, PETR3, PETR4, POMO4, PSSA3, RAPT4, RENT3, SANB11, SBSP3, SMT03, SULA11, TOTS3, TRPL4, USIM5, VALE3, WEGE3.

Por fim, a base de dados fornecida pela *BM&FBovespa*, não é ajustada por eventos de pagamento de proventos³⁹, restituição, cisão, bonificação, desdobramento⁴⁰ ou agrupamento⁴¹ de ações. Portanto, devemos fazer esses ajustes nas séries de preços dos 66 ativos selecionados acima. No caso de proventos (com exceção de bonificação), cisão e restituição, o ajuste é feito da seguinte forma:

$$x'_t = x_t \times \left(1 - \frac{prov}{x_{T'}}\right), \quad t = 1 \dots T', \quad (51)$$

onde x_t é o preço da ação no período t antes do ajuste; x'_t é o preço da ação no período t após o ajuste; $prov$ é o valor recebido pelo acionista devido ao benefício; e T' é o último dia antes do titular da ação perder o direito ao recebimento do benefício

³⁸ Empresas listadas na *BM&FBovespa* com seus respectivos códigos de negociação. Disponível em: <http://www.bmfbovespa.com.br/pt_br/produtos/listados-a-vista-e-derivativos/renda-variavel/empresas-listadas.htm>. Acesso em: 20 jan. 2018

³⁹ Proventos são benefícios que uma companhia pode distribuir aos seus acionistas tais como, dividendos, juros sobre capital próprio e bonificação.

⁴⁰ Dado que as ações listadas na *BM&FBovespa* são negociadas em lotes de 100 unidades, eventualmente o preço das ações de uma determinada companhia pode alcançar um patamar que inviabilize a aquisição por pequenos investidores. Quando isso ocorre, a administração pode optar pelo desdobramento das ações, ou seja, aumenta-se o número de ações sob titularidade de cada investidor por um multiplicador n , ao mesmo tempo em que se reduz o preços destas pelo mesmo fator n .

⁴¹ Agrupamento é a situação inversa ao desdobramento, ou seja, a quantidade de ações por investidor é reduzida na mesma proporção do aumento do preço da ação.

daquele evento específico⁴². O ajuste devido a bonificação, por sua vez, é realizado do modo abaixo:

$$x'_t = x_t \times \frac{1}{1+b}, \quad t = 1 \dots T', \quad (52)$$

onde x_t , x'_t e T' são definidos conforme eq.(51) e b é a bonificação que é definida em pontos percentuais. Por último, temos a fórmula para o ajuste de eventos de desdobramento e agrupamento:

$$x'_t = x_t \times \frac{1}{r}, \quad t = 1 \dots T', \quad (53)$$

onde r é razão entre a nova quantidade de ações em circulação e a anterior; x_t , x'_t e T' são definidos conforme acima.

3.2 ALGORITMO EM PYTHON

Neste tópico detalharemos o algoritmo em Python para simular a estratégia de arbitragem estatística conforme as três etapas descritas em 2.3.1, 2.3.2 e 2.3.3, respectivamente, construção da relação estatística de preço, identificação dos desvios de preço e posicionamento no mercado.

Na tabela 6 pode-se ver o código completo cuja estrutura está dividida em nove partes. Na primeira importamos os módulos cujas funções utilizaremos no decorrer do código: *numpy*⁴³, pacote útil para lidar com números, vetores, matrizes e dados multi-dimensionais; e *statsmodels*⁴⁴, biblioteca para modelagem estatística cujas funções *OLS*⁴⁵ e *adfuller* serão empregadas, respectivamente, na primeira etapa do algoritmo para realização das regressões lineares e na realização do teste ADF de estacionariedade das séries de desvios de preços conforme explicado em 2.3.2.

⁴² O dia do pagamento do benefício não é necessariamente o mesmo dia em que se cessa o direito a recebê-lo. Em geral aquele ocorre depois deste. O primeiro dia em que a ação é negociada com encerramento do direito ao provento é denominado *data ex-dividendos*. Para que um investidor tenha direito ao benefício de um evento de pagamento de proventos de uma ação ele deve iniciar o pregão da data *ex-dividendos* com essa ação em sua carteira.

⁴³ **Numpy**. Disponível em: <<http://www.numpy.org/>>. Acesso em: 12 jan. 2018.

⁴⁴ **StatsModels**. Disponível em: <<http://www.statsmodels.org>>. Acesso em: 12 jan. 2018.

⁴⁵ Abreviatura em inglês para método dos mínimos quadrados.

A segunda parte do código centraliza todas as entradas do modelo. A entrada *arq1* é o arquivo com a base de dados das séries de preços dentro-da-amostra⁴⁶ de cada um dos 66 ativos do conjunto *A*, organizados portanto em 66 colunas – uma para cada ativo – e em um número de linhas equivalente ao número total de observações de cada ativo da amostra, ou seja, total de dias úteis do período. Esses são os dados que serão utilizados para definirmos um ativo sintético $AS(x_t)$ para cada um dos 66 ativos do conjunto *A*. O *arq2*, por seu turno, é o arquivo com a base de dados fora-da-amostra, organizado de maneira análoga ao *arq1*. A estratégia de investimento será simulada para o período de 2015 a 2017, porém ela será implementada em três fases, uma para cada ano do período: na primeira fase, o *arq1* contemplará os dados de 2010 a 2014, e o *arq2* os de 2015; na fase seguinte, os dados de 2015 passam a integrar o *arq1*, enquanto que os dados de 2016 formam o *arq2*; e na última fase, tem-se os preços de 2010 a 2016 formando a base dentro-da-amostra (*arq1*) e os dados de 2017 formando a base fora-da-amostra (*arq2*). Com essa forma de implementar a estratégia estamos simulando a passagem do tempo. Na primeira fase, por exemplo, 2015 é um ano por vir, enquanto que na segunda ele é um ano que já passou, por isso, na transição entre as fases 1 e 2 ele deixa de fazer parte dos dados fora-da-amostra e passa a fazer parte dos dados dentro-da-amostra. Ainda na segunda parte do código, tem-se as seguintes entradas: *q*, quantidade de ativos estabelecidos para formar o ativo sintético $AS(x_t)$, critério de parada da regressão passo-a-passo conforme definida em 2.3.1; *p*, valor crítico determinado para o teste ADF, isto é, valor-p mínimo que as séries de desvio de preço deverão obter para que sejam consideradas estacionárias e portanto selecionadas para a carteira de investimento (ver 2.3.2); *fat*, é o fator *k* da eq.(47); *c*, é uma aproximação percentual dos custos de transação⁴⁷; *stop_loss*,

⁴⁶ A expressão *dentro-da-amostra* em modelos de previsão designa os dados que serão utilizados para ajustar o modelo, no nosso caso, construir as relações estatísticas de preço. A expressão *fora-da-amostra*, por sua vez, refere-se aos dados que serão utilizados para avaliar o desempenho do modelo, isto é, servirão como um comparativo para os valores obtidos pelo modelo a partir dos dados *dentro-da-amostra*.

⁴⁷ Tanto Burgess (1999) quanto Parreiras (2007) utilizam a aproximação $c = 0,25\%$, valor que também adotaremos. Para detalhes sobre o porquê desse valor, ver uma das duas referências.

variável booleana⁴⁸ para definir se será adotada estratégia com *stop loss* (*stop_loss* = *True*) ou não (*stop_loss* = *False*); *p_stop_loss*, define o percentual para interrupção de perda na estratégia com *stop loss*; e *alav* que define multiplicador da alavancagem (estratégia sem alavancagem equivale a *alav* = 1).

A terceira parte do código é apenas um passo preliminar onde criamos as matrizes da base de dados dentro-da-amostra e fora-da-amostra, além das matrizes que guardarão os índices dos ativos selecionados para cada desvio, os pesos e desvios de preço do portfólio. Nessa etapa também tomamos o total de ativos do conjunto *A*, e número de observação das duas bases de dados.

A quarta parte, por seu turno, equivale à primeira etapa da estratégia de investimento, ou seja, iremos construir as relações de preços através das regressões passo-a-passo descritas em 2.3.1 para então, na quinta parte do código equivalente à segunda etapa da metodologia, identificar quais desvios de preço são estacionários através do teste ADF discutido em 2.3.2. Note que no código aplicamos a função *OLS* dentro de um laço⁴⁹ *for* que itera a função de 1 até a quantidade *q* de ativos arbitrado para compor o ativo sintético, ou seja, fazemos uma regressão para cada novo ativo adicionado à composição do ativo sintético⁵⁰.

A sexta parte do código é um passo intermediário entre a segunda e terceira etapas do método de arbitragem estatística. Aqui calculamos os desvios de preço da base de dados fora-da-amostra (oriunda do *arq2*) conforme as relações de preços determinadas na primeira etapa (quarta parte do código).

A terceira e última etapa da estratégia de investimento definida em 2.3.3 foi implementada na sétima parte do código. Na oitava parte do código calculamos o retorno diário e retorno acumulado de cada desvio de preço e também da carteira como um todo, além de calcular a média e volatilidade dos retornos, e o índice de

⁴⁸ Variáveis booleanas, em referência a George Boole, matemático britânico criador dessa álgebra, podem assumir apenas dois valores, 1 (verdadeiro) ou 0 (falso). Elas são fundamentais para operações lógicas em computação.

⁴⁹ Usualmente denominado em inglês *loop*.

⁵⁰ Python começa a contagem de índices de listas, vetores, matrizes, entre outros objetos, a partir do zero, portanto, de forma precisa a iteração do laço *for* citado é de zero a *q-1*.

Sharpe⁵¹ da carteira. Por fim, na última parte do código produzimos as saídas do modelo em extensão csv, em especial, as estatísticas do portfólio: retorno acumulado, média e volatilidade.

Tabela 6 – Código em Python para simulação da estratégia de investimento

```
#### importação dos módulos necessários

import numpy as np
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller

#####

### entradas do algoritmo

# cada coluna de ambos arquivos representam um ativo (logo os dois arquivos devem ter o mesmo
# número de colunas)
# cada linha representa um dia útil

arq1 = 'BD1.csv'      # arquivo com a base de dados do histórico de preços dos ativos dentro
                      # da amostra
arq2 = 'BD2.csv'      # arquivo com a base de dados do histórico de preços dos ativos fora da
                      # amostra
q = 4                # quantidade de ativos desejados para replicar o ativo base (critério de
                      # parada da regressão)
p = '1%'             # nível crítico do teste ADF (valor-p mínimo)
fat = 1              # fator de posicionamento em relação ao desvio de preço
c = 0.0025           # custo percentual de implementação das operações
stop_loss = True     # booleano para definir modelo com ou sem stop loss
p_stop_loss = 0.2    # porcentagem para opção com stop loss
alav = 1             # alavancagem

#####

### passo preliminar

# criação das matrizes com o histórico de preços dentro e fora da amostra

BD1 = np.asmatrix(np.loadtxt(arq1, delimiter=',')/100)      # abertura do arquivo 1
BD2 = np.asmatrix(np.loadtxt(arq2, delimiter=',')/100)      # abertura do arquivo 2
n = BD1.shape[1]      # total de ativos da base de dados
du = BD1.shape[0]     # total de dias úteis na base de dados dentro da amostra
du2 = BD2.shape[0]    # total de dias úteis na base de dados fora da amostra

# criação das matrizes que guardarão respectivamente os índices, pesos e desvios do portfólio

I = np.empty((n, q+1), int)
P = np.empty((n, q+1), float)
D = np.empty((du,0), float)
```

⁵¹ Índice de Sharpe é uma medida de risco de carteira calculada pelo quociente do retorno com a volatilidade, ou seja, quanto maior o índice melhor a relação risco-retorno da carteira.

```
#####

### 1ª ETAPA da estratégia de investimento (tópico 2.3.1 da dissertação)

# formação dos desvios dos n ativos da base de dados para período dentro da amostra

for i in range(n):
    l[i][0] = i                # formação da primeira coluna com os índices dos n ativos base
    ativo_base_i = BD1[:,i]    # vetor com histórico de preços do i-ésimo ativo base
    desvio_i = ativo_base_i

    # criação da matriz dos preços dos ativos replicantes, variáveis exógenas do modelo de
    # regressão
    X = np.empty((du,0), float)

    # loop para escolher os q ativos que melhor replicam o i-ésimo ativo base
    for j in range(q):
        corr_i = np.corrcoef(desvio_i, BD1, rowvar=0)[0][1:] # vetor com as correlações dos desvios
                                                             # do i-ésimo ativo
        ind_ord = np.argsort(-corr_i)                        # índices em ordem decrescente de correlações

        # loop para certificar que não há repetição na escolha do j-ésimo ativo
        vetor_i = l[i].tolist()                             # selecionando a linha do i-ésimo ativo base
        for k in range(q+1):
            try:
                vetor_i.index(ind_ord[k])                    # verificando se ativo já consta na i-ésima linha de l
            except ValueError:
                l[i][j+1] = ind_ord[k]                       # caso não encontre, é adicionado à i-ésima linha de l
                X = np.append(X, BD1[:,ind_ord[k]], axis=1)  # e adicionado à matrix X das variáveis
                                                             # exógenas do modelo
            break

        RLN = sm.OLS(ativo_base_i, X).fit()                  # regressão linear do i-ésimo ativo base contra os
                                                             # ativos replicantes
        desvio_i = RLN.resid                                # atualização do desvio_i conforme regressão linear
        pesos = np.asmatrix(RLN.params)                     # tomando os pesos finais da regressão
        P[i] = np.append([1.0], -pesos)                     # guardando os pesos finais na matriz de pesos
        desvio_i = np.reshape(desvio_i, (len(desvio_i), 1))
        D = np.append(D, desvio_i, axis=1)                  # guardando os desvios de preço na matriz correspondente

#####

### 2ª ETAPA da estratégia de investimento (tópico 2.3.2 da dissertação)

# realização do teste ADF para verificar quais desvios são estacionários

ind_desvios_estac = np.empty((0), int)                     # guardará os índices dos desvios estacionários
for i in range(n):
    adf = adfuller(D[:,i])
    if adf[0] < adf[4][p]: ind_desvios_estac = np.append(ind_desvios_estac, i)

#####

### passo intermediário entre a 2ª e 3ª etapas

# formação dos desvios dos n ativos da base de dados para o período fora da amostra
```

```

D2 = np.empty((du2, 0)) # criação da matriz que guardará os desvios do portfólio fora da amostra
for i in range(n):
    v_col = np.zeros((du2, 1), float)
    for j in range(q+1):
        v_col = v_col + BD2[:, I[i][j]] * P[i][j]
    D2 = np.append(D2, v_col, axis=1)

```

```

#####

```

```

### 3ª ETAPA da estratégia de investimento (tópico 2.3.3 da dissertação)

```

```

# simulação da estratégia de arbitragem estatística no período fora da amostra

```

```

m = len(ind_desvios_estac) # quantidade de desvios estacionários
POS = np.zeros((du2, m), float) # matriz que guardará o tamanho da posição para cada
# desvio em cada d.u.
R = np.zeros((du2, m), float) # matriz que guardará os retornos diários de cada desvio
# em cada d.u.
R_acum = np.zeros((du2, m), float) # matriz que guardará o retorno acumulado de cada desvio
# em cada d.u.

```

```

for j in range(m):
    ind = ind_desvios_estac[j]
    for i in range(1, du2):
        POS[i][j] = -np.sign(D2[i-1, ind]) * abs(D2[i-1, ind]) ** fat # eq.(47): tamanho da posição
        delta_pos = POS[i][j] - POS[i-1][j] # variação da posição
        delta_desvio = D2[i, ind] - D2[i-1, ind] # variação do desvio
        vol_tot = (2 * BD2[i-1, ind] - D2[i-1, ind]) / alav # eq.(48): volume total
        R[i][j] = POS[i][j] * delta_desvio / vol_tot - c * abs(delta_pos) # eq.(49): retorno no período
        R_acum[i][j] = R_acum[i-1][j] + R[i][j] # eq.(50): retorno acumulado
        if stop_loss and R_acum[i][j] < -p_stop_loss: # loop para modelo com stop loss
            R_acum[i][j] = -p_stop_loss
            R[i][j] = -p_stop_loss - R_acum[i-1][j]
            for k in range(i+1, du2):
                R[k][j] = 0
                R_acum[k][j] = -p_stop_loss
            break

```

```

#####

```

```

### resultados

```

```

# estatísticas de cada desvio de preço

```

```

media = np.mean(R, axis=0) # média para cada desvio de preço
vol = np.std(R, axis=0) # volatilidade para cada desvio de preço

```

```

# estatísticas do portfolio

```

```

ret_port = R.mean(1) # retorno diário do portfolio
ret_port_acum = R_acum.mean(1) # retorno acumulado do portfolio
media_port = np.mean(ret_port, axis=0) # média do portfolio
vol_port = np.std(ret_port, axis=0) # volatilidade do portfolio
sharpe_port = media_port / vol_port # índice de sharpe do portfolio
est_port = [ret_port_acum[du2-1], vol_port, sharpe_port] # estatísticas do portfolio

```

```

#####

```

saídas do algoritmo em .csv

```
np.savetxt('indices.csv', I, fmt='%i', delimiter=',')
np.savetxt('pesos.csv', P, fmt='%f', delimiter=',')
np.savetxt('desvios_dentro_da_amostra.csv', D, fmt='%f', delimiter=',')
np.savetxt('ind_desvios_estac.csv', ind_desvios_estac, fmt='%i', delimiter=',')
np.savetxt('desvios_fora_da_amostra.csv', D2, fmt='%f', delimiter=',')
np.savetxt('retorno_diario.csv', R, fmt='%f', delimiter=',')
np.savetxt('retorno_acumulado.csv', R_acum, fmt='%f', delimiter=',')
np.savetxt('retorno_portfolio.csv', ret_port, fmt='%f', delimiter=',')
np.savetxt('retorno_portfolio_acumulado.csv', ret_port_acum, fmt='%f', delimiter=',')
np.savetxt('media.csv', media, fmt='%f', delimiter=',')
np.savetxt('vol.csv', vol, fmt='%f', delimiter=',')
np.savetxt('est_portfolio.csv', est_port, fmt='%f', delimiter=',')
```

Fonte: Elaboração própria

3.3 RESULTADOS

Conforme afirmamos em 3.2 implementamos o algoritmo em três fases, uma para cada ano do período escolhido para a simulação, ou seja, 2015, 2016 e 2017. Nas tabelas 7, 8 e 9 abaixo há as relações de preços $E(X_t) = AS(X_t)$ obtidas para os desvios que passaram no teste ADF de estacionariedade em cada uma das fases, 2015, 2016 e 2017 respectivamente.

Tabela 7 – Desvios de preço estacionários obtidos para o ano de 2015

Ativo Objeto	Peso X Ativo Sintético			
BBAS3	-1,8201 X ITSA4	-0,1816 X RAPT4	-0,5111 X SMT03	-0,1389 X PETR4
BBDC3	-0,3366 X MDIA3	0,1886 X ELET3	-0,2451 X BRSR6	-1,8314 X ITSA4
BBDC4	-0,2075 X BBDC3	0,0655 X ELET3	-0,1986 X BRML3	-2,1621 X ITSA4
BRML3	-0,2416 X CSMG3	-0,1678 X BRSR6	-0,6221 X ODPV3	-0,0549 X NATU3
BVMF3	-0,1847 X CSAN3	-0,0716 X ELET3	-0,1476 X HYPE3	-0,1415 X BRSR6
CCRO3	-0,0160 X MULT3	-0,2770 X RENT3	-0,3748 X SBSP3	-0,0854 X BRFS3
CPFE3	-1,3390 X ENBR3	-0,0164 X ELET6	-0,0890 X TRPL4	-0,1681 X LIGT3
CPL6	-0,9056 X CESP6	-0,6374 X DTEX3	-0,2417 X ELET3	-0,3232 X BRSR6
CSAN3	-1,4980 X CCRO3	0,0565 X USIM5	-0,1539 X RAPT4	-0,9520 X DTEX3

CSMG3	-0,9648 X SBSP3	-0,4040 X ELET6	-0,5835 X BRML3	-0,0376 X CMIG4
CYRE3	-0,3833 X GGBR4	0,0254 X LREN3	-0,6610 X RAPT4	-0,3158 X SANB11
DTEX3	-0,6201 X RAPT4	-0,0924 X FIBR3	-0,1784 X BEEF3	-0,0721 X CYRE3
ELET6	-0,6938 X ELET3	0,0245 X CCRO3	-0,2090 X LIGT3	-0,0423 X MYPK3
EMBR3	-1,3885 X WEGE3	-0,0448 X TRPL4	-0,4403 X POMO4	-0,0119 X FIBR3
ENBR3	-0,3356 X ODPV3	-0,0556 X PETR3	-0,7058 X POMO4	-0,1398 X CESP6
FLRY3	-0,1272 X HGTX3	-0,0161 X BTOW3	-0,1699 X ELET3	-0,1086 X EMBR3
GGBR4	-0,7210 X GOAU4	-0,0529 X GOLL4	-0,0112 X FIBR3	-0,0381 X USIM5
GOAU4	-1,2015 X GGBR4	0,1369 X ENBR3	-0,0457 X MYPK3	-0,0386 X CSMG3
GOLL4	-0,1362 X GFSA3	0,6236 X RENT3	-0,3158 X ITSA4	-0,4572 X ITUB4
IGTA3	-0,3397 X MULT3	-0,0096 X GOLL4	-0,1828 X BRSR6	-0,3373 X SULA11
ITSA4	-0,2334 X ITUB4	-0,0137 X BRAP4	-0,0005 X HGTX3	-0,0494 X BRSR6
ITUB4	-3,2436 X ITSA4	-0,1032 X ESTC3	-0,0399 X BTOW3	-0,0294 X CESP6
MDIA3	-0,0278 X EQTL3	-0,8861 X GRND3	-0,6254 X CIEL3	-0,0387 X GOAU4
MRVE3	-0,2589 X BRAP4	0,0238 X ITSA4	-0,1453 X CYRE3	-0,1188 X BRSR6
MULT3	-1,7127 X CCRO3	-0,1041 X MRFG3	-0,3139 X NATU3	-0,7208 X LAME4
NATU3	-2,5279 X POMO4	-0,2570 X ELET3	-0,5750 X HYPE3	-0,9043 X BRML3
ODPV3	-0,0711 X TOTS3	-0,0216 X CPFE3	-0,1326 X HGTX3	-0,1119 X CIEL3
PETR3	-0,7874 X PETR4	-0,3878 X ELET3	0,0309 X GOAU4	-0,1656 X USIM5
PETR4	-0,7657 X PETR3	-0,1063 X MDIA3	-0,0792 X CMIG4	-0,0561 X BVMF3
PSSA3	-0,5675 X ESTC3	0,0427 X BRAP4	-1,1220 X SULA11	-0,1361 X GOLL4
RAPT4	-0,4188 X DTEX3	-0,1007 X BRSR6	-0,0565 X BRAP4	-0,0743 X MYPK3
RENT3	-0,5580 X LREN3	-0,0025 X CSNA3	-0,0527 X CESP6	-0,0893 X PSSA3
SANB11	-0,3000 X GOLL4	-0,2125 X CCRO3	-0,1611 X CESP6	-0,0797 X GOAU4
SBSP3	-0,7573 X CCRO3	-0,2056 X CSMG3	-1,1034 X POMO4	0,2676 X CMIG4
SMT03	-0,7379 X WEGE3	-0,0737 X LIGT3	-0,0075 X BBAS3	0,0111 X SULA11
SULA11	-0,2623 X PSSA3	-0,0911 X BRAP4	-0,0579 X HGTX3	-0,2081 X DTEX3
TOTS3	-1,4662 X CCRO3	-0,1014 X ELET3	-1,0340 X FLRY3	-0,1789 X LIGT3

WEGE3	-0,3927 X ESTC3	0,0192 X BRAP4	-0,2770 X BRSR6	-0,0334 X FIBR3
-------	-----------------	----------------	-----------------	-----------------

Fonte: Elaboração própria

Tabela 8 – Desvios de preço estacionários obtidos para o ano de 2016

Ativo Objeto	Peso X Ativo Sintético			
BBAS3	-0,6229 X BBDC4	-0,0714 X PETR4	-0,1112 X BTOW3	-0,2818 X BRSR6
BBDC3	-1,5441 X RENT3	-0,0993 X FIBR3	-0,2485 X RAPT4	-0,3007 X GRND3
BRFS3	-1,4248 X CIEL3	-0,1511 X BRAP4	-2,0443 X POMO4	-0,2675 X FIBR3
BRML3	-0,3897 X HGTX3	0,0251 X FIBR3	-0,2065 X BBDC4	-0,3175 X BEEF3
BRSR6	-0,6766 X RAPT4	-0,0469 X TRPL4	-0,0566 X BBDC4	-0,1797 X LIGT3
BTOW3	-1,6662 X GOLL4	-0,7745 X CIEL3	-0,0370 X FIBR3	0,8376 X BBAS3
BVMF3	-0,1787 X EQTL3	0,0159 X BRAP4	-0,4561 X DTEX3	-0,3428 X MRVE3
CCRO3	-0,0728 X MULT3	-0,2729 X EZTC3	-0,0608 X BRFS3	-0,1834 X SBSP3
CPFE3	-1,0239 X ENBR3	-0,0214 X BRAP4	-0,2626 X CMIG4	-0,5405 X BRSR6
CPLE6	-1,0295 X CESP6	-0,1527 X HYPE3	-0,8202 X FLRY3	-0,2093 X MRFG3
CSNA3	-0,0843 X GFSA3	0,0770 X CPFE3	-0,3412 X ELET6	-0,3138 X GGBR4
CYRE3	-0,4716 X GGBR4	-0,1641 X LAME4	-0,4309 X RAPT4	-0,2173 X MRVE3
DTEX3	-0,8209 X RAPT4	-0,0455 X EQTL3	-0,0151 X BTOW3	-0,1230 X HYPE3
ELET6	-0,7208 X ELET3	0,1202 X GRND3	-0,4011 X ENBR3	-0,1862 X BVMF3
ENBR3	-0,6698 X ODPV3	-0,0242 X MRFG3	-0,2120 X SULA11	-0,0408 X TRPL4
EZTC3	-0,1071 X CSAN3	0,0112 X TOTS3	-2,2140 X POMO4	-0,1576 X CIEL3
GGBR4	-0,6349 X GOAU4	-0,0214 X EQTL3	-0,1149 X ELET3	-0,1102 X HYPE3
IGTA3	-0,3118 X MULT3	0,0333 X GOLL4	-0,9135 X ITSA4	-0,0690 X TOTS3
ITSA4	-0,2109 X ITUB4	-0,0285 X BRAP4	-0,0338 X EMBR3	-0,0148 X CESP6
ITUB4	-2,9660 X ITSA4	-0,1440 X EQTL3	-0,0572 X BTOW3	-0,0623 X ELET6
LAME4	-0,5844 X LREN3	-0,0959 X GOAU4	-0,2550 X EMBR3	-0,0486 X HGTX3
LREN3	-0,3927 X EQTL3	0,0587 X BRML3	-0,0208 X PSSA3	-0,2411 X LAME3
MRFG3	-0,1034 X GFSA3	0,0319 X SBSP3	-0,0750 X TRPL4	-0,1324 X GOAU4

MRVE3	-0,2165 X ELET3	0,1060 X BEEF3	-0,1277 X HGTX3	-0,2773 X HYPE3
MULT3	-1,3334 X IGTA3	-0,6667 X SBSP3	-0,1562 X CYRE3	-0,1240 X ESTC3
MYPK3	-0,5002 X HGTX3	0,0916 X HYPE3	-0,2757 X GGBR4	-0,5192 X SULA11
NATU3	-5,2809 X POMO4	-0,1393 X GFSA3	-0,2803 X LREN3	-0,3197 X CESP6
ODPV3	-0,5099 X ENBR3	-0,3643 X RENT3	0,0221 X TOTS3	-0,1119 X LAME3
PETR3	-0,7618 X PETR4	-0,1639 X ELET3	-0,2726 X MRFG3	-0,1411 X CSNA3
PSSA3	-1,3986 X WEGE3	-0,0557 X VALE3	-0,4516 X SULA11	-0,0832 X BTOW3
RENT3	-0,2363 X MDIA3	0,0071 X BRAP4	-0,0228 X TRPL4	-0,1117 X CESP6
SANB11	-0,3066 X HYPE3	-0,1034 X CPFE3	-0,2707 X ELET6	-0,1956 X LREN3
SULA11	-0,4587 X SMT03	-0,0354 X VALE3	-0,1186 X FIBR3	-0,1709 X MYPK3
TOTS3	-1,5296 X CCRO3	-0,0284 X GFSA3	-0,4140 X CESP6	-0,0891 X FIBR3
USIM5	0,4109 X GGBR4	-0,1009 X GFSA3	-0,5509 X GOAU4	-0,2443 X CSNA3
VALE3	-1,2448 X BRAP4	-0,0858 X CIEL3	-0,0626 X FIBR3	-0,2424 X DTEX3
WEGE3	-0,3717 X CIEL3	0,0022 X GFSA3	-0,2541 X FLRY3	-0,0619 X FIBR3

Fonte: Elaboração própria

Tabela 9 – Desvios de preço estacionários obtidos para o ano de 2017

Ativo Objeto	Peso X Ativo Sintético			
BBDC3	-1,4003 X RENT3	-0,0633 X FIBR3	-0,2938 X RAPT4	-0,1236 X BRFS3
BRFS3	-1,4288 X CIEL3	-0,1526 X BRAP4	-2,0308 X POMO4	-0,2668 X FIBR3
BRML3	-0,4900 X HGTX3	-0,0484 X FIBR3	-0,1550 X ESTC3	-0,0466 X BTOW3
BRSR6	-0,6644 X RAPT4	-0,0511 X TRPL4	-0,0597 X BBDC4	-0,1740 X LIGT3
BVMF3	-0,1522 X EQTL3	-0,0185 X GOAU4	-0,4540 X MRVE3	-0,1104 X CSAN3
CCRO3	-0,1492 X MULT3	-0,3541 X EZTC3	-0,0484 X BRFS3	0,0187 X TOTS3
CPL6	-1,0656 X CESP6	-0,1777 X HYPE3	-0,6715 X FLRY3	-0,2098 X MRFG3
CSAN3	-0,8110 X EZTC3	0,1246 X ELET6	-1,1246 X BVMF3	-0,2204 X CSMG3
CSNA3	-0,0875 X GFSA3	0,0672 X CPFE3	-0,3176 X ELET6	-0,3093 X GGBR4
CYRE3	-0,3436 X GGBR4	-0,2474 X LAME4	-0,5903 X RAPT4	-0,0326 X GFSA3

DTEX3	-0,8248 X RAPT4	-0,0385 X EQTL3	-0,0178 X BTOW3	-0,1242 X HYPE3
EMBR3	-0,2229 X BRFS3	-0,1419 X FIBR3	0,1631 X BRML3	-0,2063 X TOTS3
ENBR3	-0,7653 X ODPV3	-0,0017 X MRFG3	-0,0887 X ELET6	-0,2002 X SULA11
EZTC3	-2,4281 X POMO4	-0,1155 X CIEL3	-0,0056 X CSNA3	-0,3289 X RENT3
GGBR4	-0,6135 X GOAU4	-0,1419 X SANB11	-0,1061 X ELET3	-0,0631 X HYPE3
IGTA3	-0,3640 X MULT3	-0,0563 X GOLL4	-0,1505 X PSSA3	-0,0288 X TRPL4
ITSA4	-0,2125 X ITUB4	-0,0144 X GOAU4	-0,0425 X CESP6	-0,0138 X FIBR3
ITUB4	-2,9797 X ITSA4	-0,1418 X EQTL3	-0,0571 X BTOW3	-0,0586 X ELET6
LAME3	-0,5980 X LAME4	-0,0360 X RAPT4	-0,1120 X CMIG4	-0,4249 X POMO4
LAME4	-0,5749 X LREN3	-0,0834 X GOAU4	-0,2325 X EMBR3	-0,1173 X BRML3
LREN3	-0,3643 X EQTL3	0,1280 X BRML3	-0,0496 X TOTS3	-0,2939 X LAME3
MRFG3	-0,0768 X GFSA3	0,0158 X SBSP3	-0,2706 X PETR3	-0,0309 X FIBR3
MRVE3	-0,2505 X ELET3	-0,0199 X TOTS3	-0,1812 X HYPE3	-0,1302 X MYPK3
MULT3	-1,3352 X IGTA3	-0,6638 X SBSP3	-0,1570 X CYRE3	-0,1241 X ESTC3
MYPK3	-0,5007 X HGTX3	-0,0782 X HYPE3	-0,2186 X USIM5	-0,3317 X EZTC3
NATU3	-5,2902 X POMO4	-0,1385 X GFSA3	-0,2584 X LREN3	-0,3302 X CESP6
ODPV3	-0,5000 X ENBR3	0,0011 X TOTS3	-0,3650 X RENT3	-0,0375 X EMBR3
PETR3	-0,7608 X PETR4	-0,1309 X ELET3	-0,2929 X MRFG3	-0,1552 X CSNA3
PSSA3	-1,3207 X WEGE3	-0,1078 X VALE3	-0,2779 X EMBR3	-0,0701 X BTOW3
RENT3	-0,2224 X MDIA3	0,0488 X BRAP4	-0,1386 X CESP6	-0,1600 X MRVE3
SANB11	-0,9435 X FLRY3	-0,0556 X BTOW3	-0,0480 X FIBR3	-0,1858 X HYPE3
SMT03	-0,2939 X EQTL3	-0,0390 X GOLL4	-0,1131 X EMBR3	-0,0151 X CSMG3
SULA11	-0,4687 X SMT03	-0,0464 X BRAP4	-0,1214 X FIBR3	-0,1703 X MYPK3
TOTS3	-1,4832 X CCRO3	-0,0286 X GFSA3	-0,4499 X CESP6	-0,0805 X FIBR3
USIM5	-0,0977 X GFSA3	0,5926 X POMO4	-0,4636 X GOAU4	-0,0276 X MDIA3
VALE3	-1,2579 X BRAP4	-0,0442 X CIEL3	-0,0931 X FIBR3	-0,0717 X CSAN3
WEGE3	-0,3528 X CIEL3	0,0501 X ELET3	-0,0654 X FIBR3	-0,3475 X FLRY3

Fonte: Elaboração própria

Dos desvios obtidos acima podemos ver alguns padrões na formação dos ativos sintéticos. O primeiro deles, e talvez o mais natural de todos, é a relação entre ações preferenciais⁵² e ações ordinárias⁵³ da mesma companhia, como, por exemplo, LAME3 x LAME4 e PETR3 x PETR4. Outro caso é o de relações de preço formadas entre companhias que possuem vínculo de propriedade entre si ou que pertencem ao mesmo grupo, casos de GGBR4 x GOAU4, ITUB4 x ITSA4 e VALE3 x BRAP4⁵⁴. Também há o cenário de relações de preço formadas por companhias que atuam no mesmo setor: IGTA3 x MULT3, ambos atuam no mercado imobiliário; BRML3 x HGTX3 e LAME4 x LREN3, todas empresas de varejo; e BBAS3 x BBDC4, ambos bancos. Há ainda outras relações menos óbvias entre as companhias que formam uma determinada relação de preços, como sensibilidade a flutuações cambiais, modificações na taxa de juros básica da economia ou expectativa de inflação por exemplo. No entanto, não buscaremos justificar economicamente cada uma das relações obtidas, uma vez que estamos construindo estatisticamente tais correspondências. Com os padrões supracitados podemos ficar confortáveis que, de uma forma geral, os resultados estatísticos estão fazendo sentido econômico e isso deve ser suficiente para prosseguirmos com a metodologia. Na figuras 11 e 12, a título de exemplo, pode-se ver o gráfico do desvio de preço para o ativo objeto BBAS3 em 2016 e VALE3 em 2017.

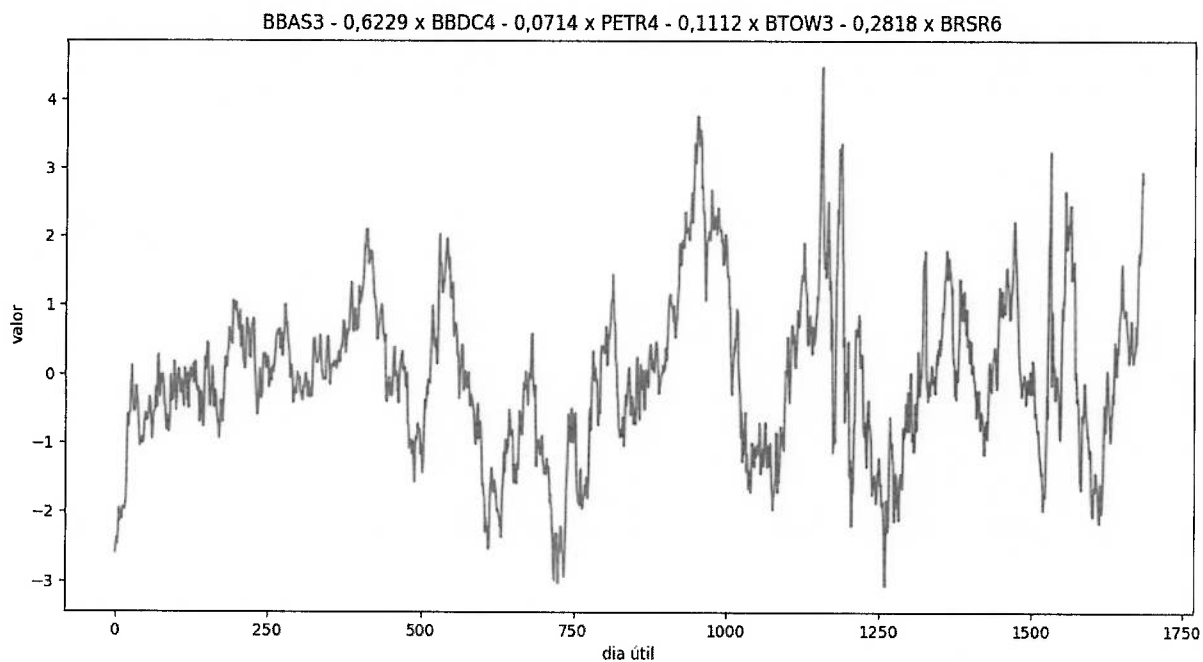
Posteriormente aplicamos a estratégia de investimento definida em 2.3. para diferentes valores dos parâmetros k (vide eq.(47)), *alavancagem* e *stop loss*. A tabela 10 sumariza o retorno, volatilidade diária e índice de Sharpe da carteira em cada um dos cenários para cada ano do período estabelecido.

⁵² Negociadas sob código com dígito final 4, são as ações com preferência no recebimento de dividendos, porém não dão direito a voto em assembleia ao acionista titular da mesma.

⁵³ Negociadas sob código com dígito final 3, são as ações com que dão ao acionista direito a voto em assembleia.

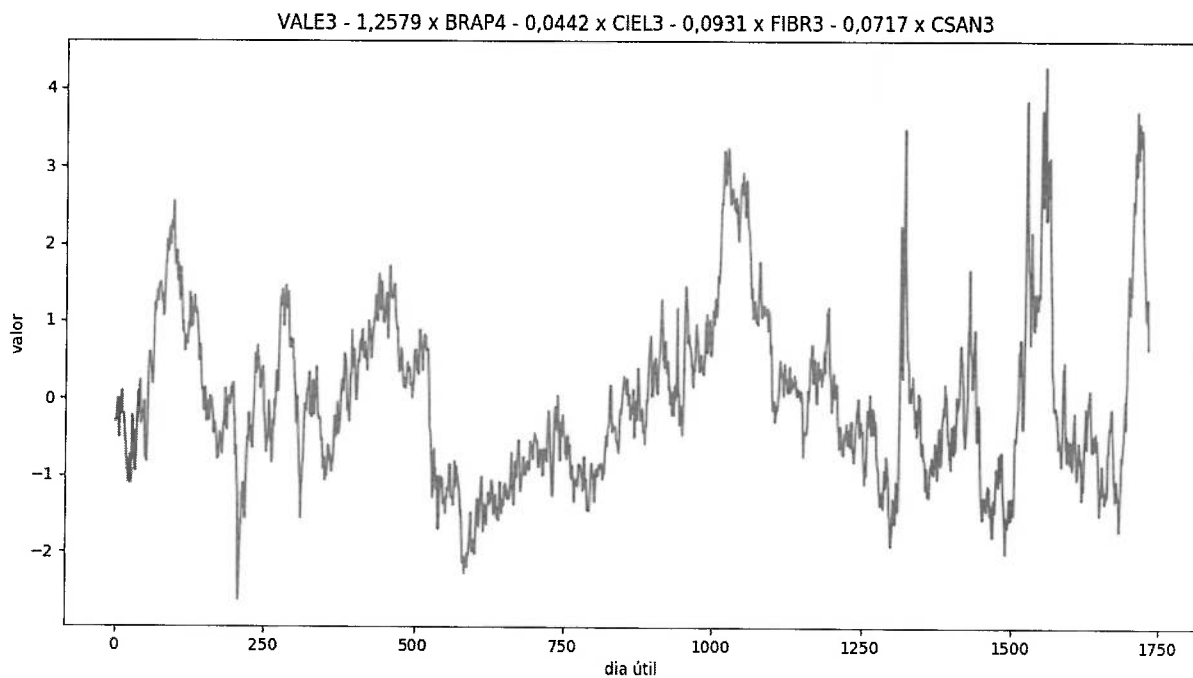
⁵⁴ Informações sobre a posição acionária de companhias listadas na *BM&FBovespa*. Disponível em: <http://www.bmfbovespa.com.br/pt_br/produtos/listados-a-vista-e-derivativos/renda-variavel/empresas-listadas.htm>. Acesso em: 20 jan. 2018.

Figura 11 – Desvio de preço para o ativo objeto BBAS3 em 2016



Fonte: Elaboração própria

Figura 12 – Desvio de preço para o ativo objeto VALE3 em 2017



Fonte: Elaboração própria

Tabela 10 – Retorno, volatilidade diária e índice de Sharpe para diferentes valores de *stop loss*, *alavancagem* e fator *k*.

sem stop loss										
alavancagem		1			2			3		
fator k	dado / ano	2015	2016	2017	2015	2016	2017	2015	2016	2017
1	retorno;	-0,211 ;	0,153 ;	-0,183 ;	-0,293 ;	0,455 ;	-0,212 ;	-0,374 ;	0,757 ;	-0,243 ;
	volatilidade;	0,010 ;	0,006 ;	0,008 ;	0,021 ;	0,011 ;	0,017 ;	0,033 ;	0,017 ;	0,025 ;
	sharpe	-0,078	0,107	-0,089	-0,054	0,16	-0,051	-0,0459	0,177	-0,040
2	retorno;	-0,497 ;	0,682 ;	-1,179 ;	-0,016 ;	1,875 ;	-1,086 ;	0,465 ;	3,067 ;	-0,991 ;
	volatilidade;	0,096 ;	0,026 ;	0,080 ;	0,193 ;	0,052 ;	0,161 ;	0,288 ;	0,078 ;	0,240 ;
	sharpe	-0,020	0,105	-0,059	-0,0003	0,144	-0,028	0,007	0,157	-0,017
3	retorno;	3,133 ;	3,850 ;	-11,720 ;	16,398 ;	10,130 ;	-8,217 ;	29,66 ;	16,408 ;	-4,713 ;
	volatilidade;	1,061 ;	0,154 ;	1,077 ;	2,124 ;	0,309 ;	2,149 ;	3,186 ;	0,464 ;	3,223 ;
	sharpe	0,012	0,100	-0,044	0,031	0,132	-0,016	0,037	0,142	-0,006
stop loss = 10%										
alavancagem		1			2			3		
fator k	dado / ano	2015	2016	2017	2015	2016	2017	2015	2016	2017
1	retorno;	-0,059 ;	0,014 ;	-0,058 ;	-0,017 ;	0,036 ;	-0,040 ;	0,027 ;	0,082 ;	-0,053 ;
	volatilidade;	0,002 ;	0,002 ;	0,002 ;	0,003 ;	0,003 ;	0,003 ;	0,004 ;	0,004 ;	0,003 ;
	sharpe	-0,144	0,024	-0,137	-0,024	0,053	-0,050	0,027	0,090	-0,067
2	retorno;	-0,014 ;	-0,022 ;	-0,067 ;	0,114 ;	0,019 ;	-0,069 ;	0,241 ;	0,081 ;	-0,053 ;
	volatilidade;	0,004 ;	0,003 ;	0,003 ;	0,01 ;	0,005 ;	0,005 ;	0,014 ;	0,007 ;	0,008 ;
	sharpe	-0,012	-0,033	-0,107	0,047	0,016	-0,058	0,065	0,047	-0,029
3	retorno;	-0,071 ;	-0,050 ;	-0,090 ;	0,003 ;	-0,001 ;	-0,085 ;	1,403 ;	0,053 ;	-0,0810
	volatilidade;	0,006 ;	0,005 ;	0,007 ;	0,012 ;	0,008 ;	0,014 ;	0,071 ;	0,014 ;	0,025 ;
	sharpe	-0,050	-0,041	-0,053	0,001	-0,0	-0,023	0,08	0,014	-0,012
stop loss = 20%										
alavancagem		1			2			3		
fator k	dado / ano	2015	2016	2017	2015	2016	2017	2015	2016	2017
1	retorno;	-0,12 ;	0,039 ;	-0,09 ;	-0,098 ;	0,1409 ;	-0,088 ;	-0,055 ;	0,211 ;	-0,049 ;
	volatilidade;	0,002 ;	0,003 ;	0,002 ;	0,004 ;	0,006 ;	0,004 ;	0,005 ;	0,008 ;	0,005 ;
	sharpe	-0,221	0,052	-0,161	-0,109	0,1010	-0,090	-0,043	0,112	-0,036
2	retorno;	-0,091 ;	-0,043 ;	-0,100 ;	0,042 ;	0,066 ;	-0,106 ;	0,175 ;	0,090 ;	-0,063 ;
	volatilidade;	0,005 ;	0,006 ;	0,004 ;	0,01 ;	0,007 ;	0,006 ;	0,0149 ;	0,008 ;	0,008 ;
	sharpe	-0,069	-0,032	-0,099	0,017	0,04	-0,071	0,047	0,0429	-0,029
3	retorno;	-0,143 ;	-0,042 ;	-0,125 ;	0,768 ;	-0,039 ;	-0,136 ;	1,366 ;	-0,019 ;	-0,078 ;
	volatilidade;	0,014 ;	0,006 ;	0,008 ;	0,047 ;	0,01 ;	0,016 ;	0,071 ;	0,0179 ;	0,025 ;
	sharpe	-0,040	-0,025	-0,068	0,066	-0,014	-0,034	0,076	-0,004	-0,012

Fonte: Elaboração própria

Note que em linhas gerais a adoção de *stop loss* melhorou o desempenho da carteira. Ainda, maiores valores de k , ou seja, um peso maior nos desvios de preço de maior magnitude, boa parte das vezes também ajudou a melhorar o retorno da carteira, assim como a adoção de alavancagem. No entanto, o aumento do retorno veio acompanhado de um significativo aumento na volatilidade (veja por exemplo opção sem *stop loss*, *alavancagem* = 3 e $k=3$) fato que não ajudou o índice de Sharpe da carteira mudar de patamar. Em termos práticos isso quer dizer que apesar de terem ocorrido alguns retornos muito acima da média, o oposto disso, retornos negativos de grande magnitude também podem ocorrer (ver por exemplo opção sem *stop loss*, *alavancagem* = 2 $k = 3$), e a carteira possui um risco alto. Outro ponto a se destacar é que em nenhuma das estratégias houve um triênio com retornos positivos em todos os anos, o que mostra que a carteira não ficou com retornos consistentes.

4 CONCLUSÕES

Desenvolvemos este trabalho com objetivo de introduzir os principais conceitos necessários para o entendimento de operações de arbitragem estatística, além de testar um modelo de investimento no mercado brasileiro com base nessa teoria, com expectativa de obter retornos consistentes, com volatilidade sob controle e consequentemente um bom índice de Sharpe. Entretanto, verificamos que para diferentes valores dos parâmetros k , *alavancagem* e *stop loss*, teve-se uma variedade de retornos sem que se apresentasse algum padrão que pudéssemos adotar.

Diante disso, nos cabe agora refletir sobre eventuais melhorias aplicáveis ao modelo, além de abordagens complementares. Entendemos que a primeira e segunda etapa da metodologia, isto é, construção das relações de preços e identificação dos desvios de preços estacionários, foram bem executadas pois conseguimos chegar em uma lista de 37 desvios de preços estacionários com nível de certeza de 99% para cada ano do período de 2015 a 2017. As figuras 11 e 12 ilustram bem isso. A terceira etapa, posicionamento no mercado, por sua vez, é um passo interessante para testar outras fórmulas de investimento no lugar da eq.(47). Uma alternativa simples, por exemplo, seria ao invés de se posicionar diariamente, tomar uma posição apenas quando o desvio de preço assumir valores maiores que uma banda que pode ser fixa ou determinada por múltiplos de desvio padrão da série. Dessa maneira reduziria-se os custo de transação, além aumentar a probabilidade de auferir lucro em uma dada operação. Outra abordagem interessante seria determinar pesos para posicionamento nos desvios de preços estacionários por critérios de maximização de retorno e minimização de variância⁵⁵. Na mesma linha, é possível determinar um modelo de otimização numérica para determinar de forma mais rigorosa os melhores intervalos de valores para os parâmetros de entrada do modelo.

Uma abordagem complementar adotada tanto por Burgess (1999) e Parreiras (2007) foi de inserir variáveis exógenas, tais como cotações cambial, taxa de juros

⁵⁵ Markowitz (1952) foi pioneiro no desenvolvimento de modelos para seleção de carteira com critérios de maximização de retorno e minimização de variância.

livre de risco e índices de bolsa, entre outros, junto com técnicas de redes neurais a fim de aprimorar a capacidade de predição do modelo. De fato, ambos obtiveram retornos melhores após adotar tal estratégia, enquanto que quando seguiram uma linha de arbitragem estatística pura, isto é, sem a combinação com redes neurais, alcançaram resultados irregulares. Por fim, vale destacar um trabalho recente de Moura (2014) onde descreve uma metodologia de arbitragem estatística baseada em filtro de Kalman.

5 BIBLIOGRAFIA

BACHELIER, L. **Théorie de la spéculation**. Annales scientifiques de l'É.N.S., v.3, n. 17, p. 21-86, 1900.

BLACK, F. S.; SCHOLES, M. S. **The Pricing of Options and Corporate Liabilities**. Journal of Political Economy, v. 81, n. 3, p. 637-654, 1973.

BM&FBovespa – **Cotações históricas**. Disponível em:

http://www.bmfbovespa.com.br/pt_br/servicos/market-data/historico/mercado-a-vista/cotacoes-historicas/>. Acesso em: 22 dez. 2017.

BURGESS, A. N. **A Computational Methodology for Modelling the Dynamics of Statistical Arbitrage**. Londres: Tese de Doutorado, London Business School, 1999.

FAMA, E. F. **The Behavior of Stock-Market Prices**. Journal of Business, v. 38, n. 1, p. 34-105, 1965.

GREENE, W. H. **Econometric Analysis**. Harlow: Pearson Education, 2012.

HALL, A. **Testing for a Unit Root in the Presence of Moving Average Errors**. Biometrika, v. 76, n. 1, p. 49-56, 1989.

HULL, J. **Options, Futures and Other Derivatives**. Toronto: Prentice Hall, 2009.

MACKINNON, J. G. **Critical Values for Cointegration Tests**. Kingston: Working Papers, Queen's University, 2010.

MAGALHÃES, M. N.; de LIMA, A. C. P. **Noções de Probabilidade e Estatística**. São Paulo: Edusp, 2004.

MARKOWITZ, H. **Portfolio Selection**. The Journal of Finance, v. 7, n. 1, p. 77-91, 1952.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de Séries Temporais**. São Paulo: Edgard Blücher, 2004.

MOURA, C. E. S. **A new pairs trading strategy based on linear state space models and the Kalman filter**. Rio de Janeiro: Dissertação de Mestrado, Instituto de Matemática Pura e Aplicada, 2014.

PARREIRAS, L. P. **Arbitragem Estatística e Inteligência Artificial**. São Paulo: Dissertação de Mestrado, Universidade de São Paulo, 2007.

PHILIPS, P. C. B. **Time Series Regression with a Unit Root**. Econometrica, v. 55, n.2, p. 277-301, 1987.

PHILIPS, P. C. B; PERRON, P. **Testing for a Unit Root in Time Series Regression**. Biometrika, v. 75, n. 2, p. 335-346, 1988.

ROSS, S. **Probabilidade**: Um curso moderno com aplicações. Porto Alegre: Grupo A, 2010.

TSAY, R. S. **Analysis of Financial Time Series**. Hoboken: John Wiley & Sons, 2005.

VIDYAMURTHY, G. **Pairs Trading**: Quantitative Methods and Analysis. Hoboken: John Wiley & Sons, 2004.

ZIVOT, E.; WANG, J. **Modelling Financial Time Series with S-PLUS**. New York: Springer Science+Business Media, 2006.