FM

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS SOCIAIS E HUMANAS
CURSO DE GRADUAÇÃO EM CIÊNCIAS ECONÔMICAS

Mateus Machado de Pereira

EFICIÊNCIA DE MERCADO: EVIDÊNCIAS A PARTIR DO GOOGLE TRENDS E ALGORITMOS DE TRADING

Santa Maria, RS
2021

Mateus Machado de Pereira

EFICIÊNCIA DE MERCADO: EVIDÊNCIAS A PARTIR DO GOOGLE TRENDS E ALGORITMOS DE TRADING

Monografia apresentada ao Curso de Graduação em Ciências Econômicas, Área de Concentração em Economia, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Barachel em Ciências Econômicas**.

ORIENTADOR: Prof. Reisoli Bender Filho

Santa Maria, RS
2021

Mateus Machado de Pereira

EFICIÊNCIA DE MERCADO: EVIDÊNCIAS A PARTIR DO GOOGLE TRENDS E ALGORITMOS DE TRADING

Monografia apresentada ao Curso de Graduação em Ciências Econômicas, Área de Concentração em Economia, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Barachel em Ciências Econômicas**.

Aprovado em 29 de janeiro de 2021:

Reisoli Bender Filho, Dr. (UFSM)
(Presidente/Orientador)

Clailton Ataídes de Freitas, Dr. (UFSM)

Kalinca Leia Becker, Dra. (UFSM)

Santa Maria, RS
2021

RESUMO

EFICIÊNCIA DE MERCADO: EVIDÊNCIAS A PARTIR DO GOOGLE TRENDS E ALGORITMOS DE TRADING

AUTOR: Mateus Machado de Pereira
ORIENTADOR: Reisoli Bender Filho

A Informação tem sido um dos temas mais estudados em economia, com destaque para o mercado financeiro. A partir disso, o trabalho analisou a relação entre o valor do Ibovespa, o volume de negociações e uma *proxy* da atenção do investidor, no período que compreende 2016 a 2020. Para a elaboração da *proxy* demanda de informação ou atenção do investidor foi aplicada a Análise de Componentes Principais (ACP) no Volume Histórico de Pesquisa (VHP) disponibilizado pelo *Google Trends*. A partir disso, foi analisada a relação dessa *proxy*, com o valor do Ibovespa e volume de negociações do Ibovespa, a partir da metodologia VEC-VAR e pela elaboração de um algoritmo de *Trading*. Os principais resultados apontam que a atenção do investidor aumenta quando o mercado cai, quando há maior incerteza, por outro lado, quando o mercado está em alta, o viés de confirmação provoca uma sensação de segurança no investidor, reduzindo a demanda de informação, que passa a não acompanhar com a mesma frequência o comportamento das cotações. Ainda, o mercado apresenta um comportamento eficiente na maior parte do período analisado, assim algoritmo de trading tem um desempenho inferior ao retorno do mercado nesse período, entretanto para o período que se refere à pandemia, o mercado não se mostra eficiente, nesse período o algoritmo superou o retorno o mercado, indicando que estratégias desenvolvidas a partir de métodos quantitativos são indicadas em períodos de incerteza e instabilidades.

Palavras-chave: VEC-VAR. Ibovespa. Google Trends. Componentes Principais

ABSTRACT

MARKET EFFICIENCY: EVIDENCES FROM GOOGLE TRENDS AND TRADING ALGORITHMS

AUTHOR: Mateus Machado de Pereira
ADVISOR: Reisoli Bender Filho

Information is one of the most studied topics in the economy and finance, therefore, this work analysis the relationship between Ibovespa, amount traded and a proxy about investor attention in the period between 2016 and 2020. The present work innovates when using a new methodology to elaborate the proxy demand for information or investor attention by using Principal Component Analysis (ACP) in several tickers researched in Google Trends. From that, this work analyzes the relationship of this proxy and the value of the Ibovespa and the amount traded of Ibovespa by an VEC-VAR methodology and the elaboration of an trading strategy. Still, the market shows an efficient behavior in most of the analyzed period, so the trading algorithm has a lower performance than the market return in that period, however for the period that refers to the pandemic, the behavior shows a different behavior from what is characterized efficient, in this period the algorithm surpassed the return to the market, indicating that strategies developed from quantitative methods are indicated in periods of uncertainty and panic.

Keywords: Ibovespa. Google Trends. Principal Component Analysis. VEC-VAR. Trading

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao estudar a teoria tradicional de finanças um dos temas comumente discutido é a Hipótese do Mercado Eficiente (HME), desenvolvida por (FAMA, 1970) e (FAMA, 1991). De forma contextual, essa hipótese postula que o preço reflete todas as informações possíveis para a correta precificação de um determinado ativo, servindo assim de indicador para o investidor. Para tanto, a hipótese apresenta três formas: i) fraca; ii) semi-forte; iii) forte.

A forma fraca (i) postula que os preços das ações refletem as informações dos preços passados, dessa forma é possível prever os preços a partir de séries temporais. A versão semi-forte (ii) considera a hipótese anterior e ainda aponta que os preços refletem as todas as informações disponíveis ou públicas da empresa, de tal modo que os preços tendem a refletir todos fatos relevantes da empresa. Enquanto a terceira forma (iii) e mais restritiva, considera que preços refletem toda informação pública e privada, assim nenhuma informação, pública ou privada permite ao investidor obter lucros extraordinários no longo prazo, pois inclusive informações privilegiadas são refletidas nos preços dos ativos.

Apesar de consolidada, há muitos trabalhos que a criticam a HME, devido as suas premissas, ver (CAMPBELL; SHILLER, 1988), (HAUGEN; JORION, 1996) e; (WOUTERS, 2006) entre outros), os autores consideram as premissas da HME distante da realidade, esses autores ainda apontam situações que o mercado não se comporta de forma racional ou esperada. Sobre a irracionalidade, há um campo da ciência econômica em expansão dedada à crítica da hipótese da racionalidade dos agentes, ver (KAHNEMAN; TVERSKY, 1972),(TVERSKY; KAHNEMAN, 1974), (KAHNEMAN; TVERSKY, 2013) (KAHNEMAN, 2003), (ARIELY; JONES, 2008), (BOWER, 2010), entre outros, em que se busca explicar os motivos e desvios de raciocínios que levam ao comportamento irracional.

Essa discussão é encontrada no trabalho de (NETO, 2006), em que é discutido que as anomalias são resultado da interpretação dos investidores, que faz parte da eficiência de mercado, (NETO, 2006) ainda argumenta que há confusão em pesquisas que buscam rejeitar a HME apresentando evidências de irracionalidade, pois (FAMA, 1970) aponta que o mercado preserva a eficiência na presença de investidores irracionais.

Neste contexto de divergência, há estudos que buscam incorporar novos instrumentos a fim de melhorar previsões de preços e retornos, nessa linha (VARIAN, 2014) aponta que o *Big Data* é uma nova fonte de dados que pode ser explorada por econometristas, possibilitando o uso de novas *proxies* que anteriormente não estavam disponíveis. Já em (CHOI; VARIAN, 2009), encontra-se uma ferramenta específica, o *Google Trends*, que têm sido usado em diferentes estudos aplicados em economia, finanças, saúde e em outras áreas. Situações que reforçam a argumentação de (NASEER; TARIQ et al., 2015), quando enfatizam que os modelos de precificação dos ativos ao incorporar esses novos instrumentos, caso do *Big Data*, tem se mostrado eficazes para analisar o comportamento dos agentes e realizar previsões.

Esses instrumentos possuem vantagens operacionais pela elevada capacidade de armazenar uma enorme quantidade de informação, caso do buscador da *Google*. Corroborando (DZIELINSKI, 2012), que afirma que 70% do tráfego de pesquisas é realizado a partir do buscador da *Google* que armazena as pesquisas e transforma em um indicador de Volume Histórico de Pesquisas (VHP) e disponibiliza em uma ferramenta chamada *Google Trends*. Especificamente, esse indicador pode apresentar também outras propriedade interessantes, como tendências do que está chamando a atenção de um público específico, além da facilidade de trabalhar com diferentes frequências de dados bem como a extensa disponibilidade temporal de informações.

A informação disponibiliza pelo *Google Trends* consiste no Volume Histórico de Pesquisa (VHP) de todos usuários da ferramenta, como dúvidas simples, previsão do tempo, pesquisas a artigos, busca por nomes de sites e empresas. Segundo (HU et al., 2018), o VHP tem sido usado como *proxy* da demanda de informação ou atenção do investidor. (CHALLET; AYED, 2014) complemetam, expondo que essa *proxy* é usada para fazer previsão e inferência de variáveis mercadológicas, já que buscas pelo *ticker* [1]. de uma empresa listada em uma bolsa de valores, pode indicar que há investidores acompanhando os fatos relevantes dessa empresa, preços e notícias. Logo pesquisas *Google Trends* podem ser usadas para prever preço de abertura ações, preço de fechamento, retorno, volatilidade e volume de negociações.

Os trabalhos com focos nas finanças quantitativas buscam desenvolver modelos de previsão capazes de prever preço e/ou retorno, com essa previsão os pesquisadores desenvolvem estratégias de *trading* em que a previsão é usada para decidir simulações de compra ou venda de ativos financeiros. Se a previsão é de alta dos preços ou retornos positivos, simula-se uma compra do ativo, enquanto se a previsão é de queda do preço ou retorno negativo, simula-se uma venda do ativo. São diversos os estudos internacionais ver (DZIELINSKI, 2012), (LOUGHLIN; HARNISCH, 2013), (PERLIN et al., 2017), (CHALLET; AYED, 2014)), já trabalhos nacionais são mais escassos, mas (RAMOS; RIBEIRO; PERLIN, 2017) desenvolveram uma estratégia de trading para o Índice Ibovespa, para realizar essa previsão os autores usam o VHP de *tickers* do que compõem o Índice.

Além disso, o VHP já foi usado em estudos para previsão de indicadores econômicos (ver (CHOI; VARIAN, 2009)), previsão de índices de bolsa de valores (ver (HEIBERGER, 2015), (HU et al., 2018)), analisar a relação com retornos (ver (DZIELINSKI, 2012), (LOUGHLIN; HARNISCH, 2013), (PERLIN et al., 2017), (BIJL et al., 2016), (CHALLET; AYED, 2014), (RAMOS; RIBEIRO; PERLIN, 2017)), volatilidade ((PERLIN et al., 2017) (RAMOS; RIBEIRO; PERLIN, 2017)). Nessa linha analítica, há também trabalhos que usam *tweets* para compreender o humor do mercado (ver (BOLLEN; MAO; ZENG, 2011)), ou que combinam *tweets* com VHP (LOUGHLIN; HARNISCH, 2013).

Estruturado nessa discussão e fundamentado na intersecção da teoria tradicional de finanças, das finanças comportamentais e finanças quantitativas, o presente trabalho propõe-se a analisar a relação entre o valor do Ibovespa (preço de fechamento), volume de negociações

[1] *ticker* é um identificador único de uma ação

do Ibovespa e uma *proxy* da atenção do investidor gerada a partir da Análise de Componentes Principais (ACP) do Volume Histórico de Pesquisa (VHP) disponibilizado pelo *Google Trends*. A análise é feita pela metodologia VEC-VAR. Para tanto busca: i)Elaborar a *proxy* da demanda atenção do investidor por meio de Componentes principais; ii) Elaborar um algotimo de *trading* e testar a consequência da HME; iii) Analisar o processo de pesquisa, consumo de informação e tomada de decisão; iv) Sintetizar a relação entre o *Google Trends* e o impacto no Ibovespa.

Sobre esse último aspecto, conquanto se encontre um volume das pesquisas de viés quantitativo que buscam testar a HME, ainda há reduzidos trabalhos que avaliaram as técnicas quantitativas no mercado brasileiro, sendo o único que aplicou uma estratégia de *trading* foi desenvolvido por (RAMOS; RIBEIRO; PERLIN, 2017); dessa forma encontra-se espaço para novas pesquisas em testar novas técnicas e metodologias que busquem retornos maiores que o mercado, pois uma estratégia que supere o mercado oferece *insights* sobre o processo de investimento e tomada de decisão.

Estruturalmente, o trabalho conta com o capítulo de introdução e outros seis. O segundo apresenta a revisão bibliográfica com os principais conceitos sobre a HME, a escola de economia comportamental, as finanças quantitativas, e ainda, apresenta os trabalhos que buscaram no *Google Trends* uma variável de previsão, que buscaram no *Google Trends*, uma variável para a previsão. O terceiro capítulo apresenta as instituições do mercado brasileiro e o processo de investimento. O quarto capítulo sobre material e métodos discorre sobre a análise de componentes principais e a análise de séries temporais, bem como apresenta as características específicas dos dados analisados. E no quinto capítulo é discutido os resultados descritivos e empíricos dos métodos estatísticos e o resultado estratégia de *trading* baseado na predição realizada. E, por fim, o último capítulo apresenta a conclusão do trabalho.

2 FINANÇAS: ABORDAGENS TEÓRICAS

Este capítulo tem como objetivo apresentar a revisão bibliográfica para então realizar a intersecção das abordagens teóricas. Assim, inicilamente, são apresentadas a teoria tradicional de finanças, a escola da economia comportamental, as novas abordagens de finanças quantitativas, para então relacionar as três abordagens. Em seguida,são apresentado os principais estudos empíricos.

2.1 TEORIA TRADICIONAL DE FINANÇAS

A teoria tradicional de finanças busca explicar os retornos e os preços dos ativos, principalmente das ações. Um dos primeiros trabalhos a analisar essa relação foi o de (BACHELIER, 1900) que concluiu que o comportamento dos preços das ações é o mesmo de um passeio aleatório (*random walk*), o que implica que as mudanças nesses preços são independentes. O preço como sinalizador do mercado é um conceito definido a décadas, que pode ser visto no trabalho de (H, 1959), cuja ideia principal é de que o preço reflete todas as informações possíveis do ativo, assim sendo, ele é precificado corretamente e serve de indicador para o investidor, esse conceito serve de base para a pressuposição da Hipótese do Mercado Eficiente (HME), conforme pode ser encontrado em (FAMA, 1970).

Anos mais tarde, (FAMA, 1991) aperfeiçoa a HME, em que postula ser impossível obter lucros anormais no longo prazo, desenvolvendo as já citadas três versões: i) fraca, ii) semi-forte, e, iii) forte. A forma fraca (i) postula que os preços das ações refletem as informações nos preços passados, dessa forma, os preços passados podem ser capazes de prever os preços futuros, ao passo que nenhum investidor obterá lucros anormais analisando preços passados, pois eles já foram incorporados ao novo equilíbrio. As pesquisas realizadas a partir dessa hipótese têm carácter de previsão.

Na versão semi-forte (ii), a hipótese postula que os preços refletem seu comportamento passado e ainda refletem as informações disponíveis da empresa, de forma que novas informações levam a um novo patamar de preços. A partir desta estrutura informacional o investidor não conseguirá lucro anormal com as informações públicas. Estudos a partir dessa hipótese estudam o quão rápido o mercado se ajusta a novas informações.

Já a terceira hipótese (iii) postula os preços refletem toda informação pública e privilegiada, de tal modo que os preços partem para um novo equilíbrio que reflete toda e qualquer informação. Assim sendo, nem as informações públicas e nem as informações privilegiadas permitem o investidor obter lucros extraordinários, pois essa informação também já está refletida no preço. Estudos considerando essa hipótese buscam a evidências de agentes com informações privilegiadas.

A base teórica da Hipótese do Mercado Eficiente segundo (SHILLER et al., 1981) não pressupõe que os investidores são racionais, assim as ações são precificadas corretamente; caso os investidores sejam irracionais, as negociações são aleatórias e não afetam o preço das ações. Todavia, mesmo na presença de investidores irracionais, há arbitradores racionais cuja atuação leva a precificação correta das ações. Assim, o pressuposto da racionalidade não se torna rígido para determinação da eficiência de mercado na versão mais fraca.

Complementando, (FAMA, 1970) postula as condições para a eficiência do mercado: i) inexistência dos custos de transação; ii) informações estão disponíveis para todos; iii) expectativa homogênea dos futuros retornos, ou seja, as implicações das informações são as mesmas para todos. Entretanto (MUNIZ, 1980) considera todas as proposições necessárias para um mercado eficiente pouco realistas.

Assim sendo, a Hipótese do Mercado Eficiente implica que não é possível superar o retorno do mercado a partir de uma estratégia, caso aconteça, foi um caso esporádico e atribuído a outros fatores. Logo, análises técnicas, em que um agente analisa informações passadas em gráficos ou estatísticas específicas em cada estratégia para obter retornos maiores que o mercado também não são eficazes, mas resultado de sorte, conforme discute (SAFFI, 2003).

A HME também implica que outras estratégias como a análise fundamentalista, não são capazes de superar retornos de mercado. Historicamente a análise fundamentalista tem origem com Benjamin Graham após o colapso do mercado amercaido em 1929, a estratégia busca determinar o valor intrínseco da ação e comparar com o valor justo, assim como indicadores e fundamentos que são informações públicas, conforme visto em (COSTA; VARGAS, 2020).

Entretanto, apesar de consolidada, essa teoria também é largamente criticada ver (CAMPBELL; SHILLER, 1988), (HAUGEN; JORION, 1996) e; (WOUTERS, 2006), sobremaneira por utilizar anomalias que contrapõem-se a eficiência do mercado. Dentre as críticas feitas à HME, destaca-se (HAUGEN; JORION, 1996) que afirma que as alterações nos preços das ações não se ajustam instantaneamente, mas de forma atrasada, como também de (HAUGEN; HAUGEN, 2001) que enfatiza que a hipótese semiforte aponta que profissionais do mercado não tem valor.

Sob uma perspectiva comportamental comportamental, (TVERSKY; KAHNEMAN, 1992) colocam em dúvida a racionalidade na Teoria da Perspectiva, em que a irracionalidade afeta os retornos de investimentos, a teoria define os agentes como avesso ao risco para ganhos e propenso ao risco para perdas, sendo essa é uma das principais teorias em que se contrapõem a Hipótese do Mercado Eficiente.

Esta escola de economia é um dos principais expoentes em contraposição a teoria tradicional de finanças, pois o comportamento irracional dos investidores pode ser uma fonte de anomalias de mercado. Logo, as críticas mais comuns à eficiência de mercado partem da irracionalidade dos agentes ou das anomalias de mercado, pois conforme (LIMA, 2003) a eficiência de mercado na versão forte é sensível a racionalidade dos investidores.

Os aspectos descritos são relevantes para compreender como a teoria tradicional prevê

o funcionamento de mercado e suas condições, bem como consequências de um mercado eficiente. É importante destacar ainda a sensibilidade da HME à simetria de informação, bem como sobre a racionalidade do investidor e suas consequências como anomalias. Pois, a HME impossibilita ganhos (retornos) acima do mercado, de forma que uma estratégia de trading pode não ser eficaz nesse cenário. Também as críticas não apontam sobre a irracionalidade de algoritmos, assim as finanças quantitativas podem se mostrar um caminho alternativo para a eficiência de mercado.

2.2 ECONOMIA COMPORTAMENTAL

A economia comportamental está centrada em dois conceitos que servem de base para suas análises: i) vieses e ii) heurística. Esses fatores são apontados como geradores de irracionalidade no mercado. Porém, (CVM, 2020a) destaca sete vieses como os principais que afetam os investidores: ancoragem, aversão a perda, falácia do jogador, viés de confirmação, lacuna de empatia, autoconfiança excessiva e efeito de enquadramento.

Para compreensão dos conceitos de viés cognitivo, conforme discute (ARIELY; JONES, 2008) faz referência a um erro sistemático de lógica ou pensamento do investidor, quando o julgamento desvia do desejável, ou do que seria caso usado lógica racional. Já heurísticas são comumente definidas como atalhos cognitivos ou regras práticas para simplificar decisões, dado que representam um processo de substituir uma questão difícil por outra mais fácil, de acordo com (KAHNEMAN, 2003). Ressalta-se que vieses estão geralmente relacionados a heurísticas, mas também podem estar relacionados a outros fatores como emoções, pressões sociais e motivações pessoais.

As heurísticas formalizadas por Tversky e Kahneman foram: disponibilidade, representatividade e ancoragem. A primeira induz indivíduos a fazer julgamento a partir da facilidade de imaginar um exemplo, conforme ressaltado por (TVERSKY; KAHNEMAN, 1974) quando expôs a situação de avaliar a qualidade de um investimento a partir de notícias recentes, sendo que essa notícia se sobrepõe a uma análise profunda do investidor. Ainda, essa heurística indica que o indivíduo estima a probabilidade de um evento acontecer pela velocidade de imaginar ou lembrar um exemplo desse evento, logo essa estimativa não é baseada em cálculo algum. (KAHNEMAN; TVERSKY, 1972) apontam que essa heurística tem um alicerce em estereótipos.

Já a representatividade, apresentada em (TVERSKY; KAHNEMAN, 1981) é a estimação da probabilidade de um evento A ocorrer, dado que ocorreu um evento B. Como os eventos são parecidos, o investidor considera essa aproximação boa. Um exemplo é inferir a qualidade e probabilidade de retorno do investimento A, a partir do conhecimento do investimento B.

A terceira, ancoragem, é descrita como uma apresentação a um valor inicial, que influencia todos julgamentos de valores subsequentes e que acontece de forma inconsciente. A

(CVM, 2020a) considera a ancoragem um viés no qual a exposição prévia a uma informação influencia essa mesma informação na tomada de decisão ou na formulação de alguma estimativa, independente da correlação ou relevância dessa informação.

Aversão a perda, também considerado um viés pela (CVM, 2020a) indica que o indivíduo atribui maior importância à perda que à ganhos. Kahneman e Tversky (1979) afirmam que a dor da perda é maior que o prazer do ganho, contradizendo algumas curvas de utilidades mais comuns.

O viés da falácia do jogador, segundo (BOWER, 2010) se origina da falha em compreender a independência de eventos dado a quantidade de vezes que já ocorreu. (RONEY; TRICK, 2009) afirmam que a falácia do jogador aparece sempre que um indivíduo tenta fazer previsões sobre eventos aleatórios.

Já o viés de confirmação descreve como indivíduos tendem a usar informações para ratificar as próprias convicções. (NICKERSON, 1998) associa esse viés a processos não motivados, como já discutida no viés da ancoragem. Isso fica mais claro quando o indivíduo se baseia em informações que já encontrou em um momento anterior, ou seja, reforça as hipóteses de uma ação influencia em um resultado, quando não há uma relação causal.

A lacuna de empatia segundo a (CVM, 2020a) aponta que a capacidade de interpretar acontecimentos dependente do estado emocional do investidor. Esse é o caso em que se tem uma relação próxima com uma heurística, a do afeto, em que o sentimento do investidor se sobrepõe a análise, ou até uma reação emocional a um estímulo. Segundo (FINUCANE et al., 2002), essas reações são rápidas e praticamente automáticas.

O excesso de confiança é definido por (MOORE; HEALY, 2008) quando a confiança do investidor é superior ao seu desempenho real. Os autores apontam que o excesso de confiança pode ser um fator que explica altas taxas de empreendedores que entram em mercados com baixas probabilidades de sucesso. Complementam (BUEHLER; GRIFFIN; ROSS, 1994) quando afirmam que a falácia do planejamento, em que pessoas subestimam o tempo de um projeto, é resultado do excesso de confiança. Assim, no processo do excesso de confiança o investidor confia mais na sua experiência que em dados reais, deixando de lado informações importantes.

O último viés destacado pela (CVM, 2020a) é o efeito de enquadramento, que descreve como a tomada de decisão pode ser afetada pela maneira como o problema é formulado, ou como as informações são apresentadas. (KAHNEMAN; TVERSKY, 1979) desenvolveram testes em que apresentam a mesma informação de formas diferentes e constataram que a forma com que a informação é apresentada pode induzir a respostas diferentes. Logo, uma mesma informação pode levar a decisões diferentes.

A partir dos pressupostos da escola comportamental é possível compreender o que a teoria prevê sobre como e por que os indivíduos podem agir irracionalmente. Muitos dos vieses e heurísticas apresentados, dependem da percepção e capacidade dos agentes lidar com informação e emoção.

É quase impossível imaginar um mercado eficiente com investidores irracionais, como

destacado por (LIMA, 2003) em que uma das versões da HME é sensível a racionalidade dos investidores, pois a implica que a interpretação de informações podem ser afetadas pelas emoções dos agentes. Situação que abre o debate sobre o que pode influenciar os retornos dos investidores, tendo como um papel crucial não apenas a simetria de informação, mas também como ela chega até o investidor e como se essa informação pode afetar o mercado.

Essa discussão, para (NETO, 2006) pode ser falha, pois é possível conciliar a teoria tradicional e a teoria comportamental a partir da denominada Hipótese de Eficiência Comportamental do Mercado (HECM), a qual considera as falhas de processamento de informações dos investidores. Nesse caso, ao acontecer um fato relevante, o investidor ao consumir essa informação interpreta o acontecimento enquanto sofre com vieses e heurísticas, resultando em uma análise viesada, mas a ação do investidor ainda gera o preço compatível com a informação. (NETO, 2006) aponta que ainda pode haver ineficiências, mas elas são fontes de vieses de preferências como aversão ao risco ou a Teoria do Prospecto, mas o preço é considerado justo de acordo com a informação naquele momento. O processo de decisão considerando a HECM ocorre conforme fluxo exposto na Figura 2.1.

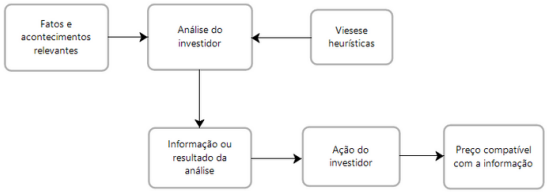

Figura 2.1 – Processo de decisão segundo a HECM

Fonte: Adaptado de (NETO, 2006)

Além disso, (NETO, 2006) argumenta que frequentemente há confusão em pesquisas que buscam rejeitar a HME a partir da apresentação de evidências de irracionalidade, pois para (FAMA, 1970) o mercado preserva a eficiência na presença de investidores irracionais, dado que as seguintes condições tendem a ser mantidas: i) não há custos de transação, ii) informações estão disponíveis e iii) todos concordam com as implicações das informações para os preços dos ativos. Entretanto, isso aponta uma nova sensibilidade, a forma como o investidor interpreta as informações, já que alguns vieses ainda podem ocorrer mesmo com HECM.

Outrossim, a discussão que a economia comportamental trouxe aprofunda a arcabouço analítico sobre as decisões do investidor, contrapondo o consenso da HME, pois considera o processo de tomada de decisão a um nível individual. Também aponta a sensibilidade do investidor às informações, pois a forma, o conteúdo e até as emoções dos investidores podem afetar sua decisão de investimento.

A partir disso pode-se questionar se algoritmos que não sejam sensíveis a vieses e heurísticas podem prever preços. Uma vez que esses algoritmos podem ser capazes de compreender um grande conjunto de informações e encontrar o preço correto a partir dessas informações sem sofrer algum tipo de irracionalidade. Essa é a premissa de estratégias de trading baseadas em

dados, pois a decisão de compra ou venda é inteiramente tomada pelo algoritmo que aprende com os dados.

2.3 FINANÇAS QUANTITATIVAS

O ramo das finanças quantitativas busca aplicar técnicas estatísticas e aprendizado de máquina (*machine learning*), em investimentos. A principal vantagem é afastar a decisão de investimento dos possíveis vieses, conforme discutem (BEUNZA; STARK, 2012b) e (BEUNZA; STARK, 2012a). Historicamente o primeiro trabalho de finanças quantitativas é desenvolvido por (MARKOWITZ, 1952), que atribuiu com uma abordagem matemática para mensurar risco e uma abordagem para decisão entre risco e retorno, formando a base para a teoria moderna do portfólio, discutido em (BECKER; REINGANUM, 2018) e (KAHN, 2018).

O modelo *Capital Asset Pricing Model*, conhecido como CAPM é outro avanço importante no campo das finanças quantitativas, desenvolvido por (SHARPE, 1964), discutido em (MOSSIN, 1966), (FRENCH, 2003) e citejensen1972capital. O avanço possibilita interpretar o mercado como decisão de investimento a partir do *trade-off* entre risco e retorno.

Por fim, o último trabalho que deu a base das finanças quantitativas é o trabalho de (BLACK; SCHOLES, 1973), os autores desenvolveram o mais famoso modelo de precificação de opções que levam seu nome *Black e Scholes*. Esses trabalhos formam a base do que hoje é considerado finanças quantitativa, assim finanças quantitativa é a intersecção da matemática, finanças, estatística e ciências da computação, conforme visto em (LI; WU; BU, 2016).

Com a Hipótese do Mercado Eficiente desenvolvida por (FAMA, 1970), a academia manteve em um consenso de que investimentos ativos (que buscam retorno acima do mercado) é um comportamento inútil, de que superar o retorno do mercado é praticamente impossível, mesmo explorando as ineficiências do mercado. Entretanto (EMERSON et al., 2019) aponta que os investidores têm adotado novas técnicas das finanças quantitativas, especialmente técnicas de aprendizado de máquina que buscam relações não lineares nos investimentos que podem auxiliar servir de instrumental matemático para superar o retorno do mercado.

Para tanto, são diversas os instrumentos que as finanças quantitativas dispõem, há trabalhos empíricos da área que buscam otimização de portfólio como Prado (2016) e Heaton, Polson e Wittie (2017), Previsão de risco de títulos, que pode ser visto em Bianchini Bucher e Tamoni (2018), Previsão de preço de derivativos e uso de hedge visto em Spiegeleer, Madan e Schoutens (2018), ou ainda *back-testing* Prado e Lewis (2018).

Os algoritmos de *trading* usam as técnias de finanças quantitativas como insumo para tomada de decisão, os trabalhos geralmente comparam os resultados com a estratégia de *Buy and Hold*, a estratégia consiste na compra de ativos e manter eles em carteira como em (CHAN, 2009). A partir disso é possível descobrir se o algoritmo é capaz de superar os retornos do mercado, o que a HME implica que não é possível. Todavia, se o algoritmo superar o retorno

do mercado, a estratégia é considerada vencedora.

Nesse sentido técnicas de aprendizado de máquina, mineração de dados e econometria financeira ganham destaque na tomada de decisão de investimento. Essas estruturas e a intersecção da teoria tradicional de finanças, da economia comportamental e da finanças quantitativas que formam o arcabouço teórico-aplicado utilizado no presente estudo, pois busca em algoritmos de *trading* superar vieses e heurísticas.

Esse algoritmo é desenvolvido por técnicas de aprendizado de máquina e econometria, com objetivo de superar a consequência da HME, que é a impossibilidade de retornos acima do mercado. O uso de um buscador de pesquisa como do *Google*, pode também revelar o comportamento do investidor, especialmente sobre como o mesmo atende sua demanda de informação, ou ainda a tendência do mercado.

No entanto é interessante conhecer os trabalhos aplicados, dessa forma a próxima seção, aprofunda as evidência empíricas que usam o *Google Trends* para prever variáveis de mercado.

2.4 EVIDÊNCIAS EMPÍRICAS

Cronologicamente, um dos primeiros trabalhos relacionando o VHP fornecido pelo Google Trends, com indicadores econômicos foi o de (CHOI; VARIAN, 2009), o qual tem carácter informativo sobre como usar a ferramenta, demonstrando a eficácia da ferramenta para melhorar previsões desde a venda de veículos, desemprego e até a confiança do consumidor, a partir de diferentes metodologias. Assim o trabalho também destaca a flexibilidade do banco de dados e aplicabilidade nos diversos setores.

Dentre os trabalhos aplicados, (BOLLEN; MAO; ZENG, 2011) encontrou correlação positiva entre os tweets e o indicador da atividade industrial do *Dow Jones* a partir da análise de sentimentos e do teste de causalidade de Granger. Logo, os resultados apontam que os *tweets* melhoram a acurácia de modelos preditivos do fechamento do *Dow Jones Industral Average*. Nesta mesma linha, (LOUGHLIN; HARNISCH, 2013) ao evoluir a análise, usaram os *tweets* e o volume de buscas a partir do *Google Trends* para prever o retorno de empresas específicas. Os resultados apontaram que o VHP não é um bom instrumento, mas os *tweets* se mostraram significantes.

Sobre os distintos resultados, (DZIELINSKI, 2012) afirma que a frequência de pesquisas em buscadores têm influência no retorno agregado das ações e na volatilidade, citando como vantagens do uso do *Google Trends* alta frequência dos dados e o fato de que eles são gerados de forma espontânea. No estudo de (DZIELINSKI, 2012) foi utilizado o modelo de regressão linear, do retorno contra a diferença do índice de buscas e a diferença de outros índices de mercado e o retorno. Os resultados levaram a conclusão de que a maior incerteza na economia eleva a demanda de informação que, por sua vez, eleva o índice de pesquisas do *Google Trends*.

Em outro estudo, (PREIS; MOAT; STANLEY, 2013) relacionaram 98 termos de pes-

quisa as operações compradas e vendidas a partir de uma estratégia de *trading* em que se simula uma compra do ativo se o valor do VHP aumenta, e se simula uma venda se o VHP diminui. O resultado indicou que é possível montar uma estratégia de investimentos baseadas em pesquisas do buscador. De outra forma, o *Google Trends* pode ser capaz de indicar as tendências de curto prazo no mercado acionário.

Para tanto,(CHALLET; AYED, 2014) alertam sobre alguns detalhes que podem comprometer a pesquisa, como viés de escolha, pois as pesquisas são sensíveis aos termos escolhidos, assim o grupo de palavras pode levar a conclusões espúrias. Para evitar esse problema, os autores usaram o nome da empresa e o *ticker*, ambas variáveis expressas em dados semanais, a partir da modelagem *Suport Vector Machines* (SVM), por sugerirem que modelos lineares não são adequados para previsão do mercado. Como resultados, concluíram que o os preços são fracamente dependentes das buscas no *Google Trends*, justificando que as próprias buscas são sensíveis a eventos externos, como más notícias.

Já (HEIBERGER, 2015), ao analisar a influência do volume de pesquisas, encontra resultados diferentes, concluindo que o *Google Trends* funciona adequadamente como indicador de más notícias, o que implica que os indivíduos usam o buscador em situações negativas economicamente. Assim a atenção coletiva segue a tendência dos jornais, contribuído para reações exageradas em situações de más notícias.

Focados na demanda de informação, (VLASTAKIS; MARKELLOS, 2012) apontam que a relação com as pesquisas como em buscadores tem relação com a volatilidade dos ativos, assim, ativos mais pesquisados são os mais voláteis. (MOUSSA; DELHOUMI; OUDA, 2017) ainda reforçam que essa volatilidade depende do retorno do ativo, onde ativos com maior rentabilidade são muito pesquisados, enquanto ativos menos rentáveis são pouco pesquisados.

Na mesma linha,(BIJL et al., 2016) buscaram no *Google Trends* uma forma de prever o retorno do mercado a partir da aplicação de modelos lineares. Os resultados possibilitaram concluir que o volume de pesquisas no buscador é significante, mas seu impacto é pequeno nos retornos.Além disso, os pesquisadores apontam a dificuldade em generalizar a pesquisa, uma vez que a relação entre o volume de pesquisas e retorno mudam ao longo do tempo.

Ainda nessa linha, há estudos puramente quantitativos que exploram a aplicação de aprendizado de máquina, os estudos foram desenvolvidos por (HU et al., 2018), (AHMED et al., 2017) e (PYO et al., 2017). Os trabalhos utilizaram Redes Neurais Artificiais em diferentes variáveis, sendo elas: preço de abertura do *Dow Jones* e S & P; preço de abertura, fechamento e volume da Bolsa do Paquistão, Shangai; Índice da Coreia do Sul respectivamente, os trabalhos de (AHMED et al., 2017) e de (HU et al., 2018) concluem que os dados do *Google Trends* melhoram significativamente os modelos de previsão. Entretanto (PYO et al., 2017) encontram resultados diferentes para a bolsa sul-coreana, em que o índice do *Google Trends* não melhora a qualidade da previsão.

Para o mercado brasileiro os estudos são recentes e em quantidade reduzida. Entre eles está o estudo de (PERLIN et al., 2017), que seleciona um grupo específico de palavras pes-

quisadas no *Google Trends*, baseado em quatro livros de investimento. O trabalho desenvolve três modelos que buscam avaliar o impacto das pesquisas no retorno e na volatilidade de quatro índices a partir de Vetores Autorregressivos (VAR). Os resultados apontam que buscas por *stock* tem impacto significativo no retorno e na volatilidade e indicam que os investidores pesquisam sobre o mercado antes de investir.

Outro trabalho aplicado é o de (RAMOS; RIBEIRO; PERLIN, 2017), o qual estudou a influência do volume de pesquisas no buscador da *Google Trends* no mercado acionário brasileiro e americano fazendo uso das modelagens VAR e causalidade de Granger para as variáveis volume de negociação, retorno e volatilidade. Os resultados apontaram que há relação causal entre as pesquisas e o mercado acionário.

O trabalho mais recente, desenvolvido por (PEREIRA; ROSA; FILHO, 2020) também específico para o mercado brasileiro foi desenvolvido a partir Vetores Autorregressivos para dados em painel (PVAR), os autores também encontraram Granger Causalidade entre VHP e retorno das ações, volume de negociações e volatilidade do mercado americano, mas apontam que empresas com maior retorno atraem mais a atenção dos investidores.

Decorrente destas aplicações e conhecendo o uso aplicado do VHP é possível compreender que a intersecção entre finanças quantitativas, teoria tradicional de finanças e economia comportamental se dá na informação e em como ela é usada. Dependendo do impacto e de suas características, a informação pode influenciar a eficiência do mercado já que esse processo inclui os possíveis vieses e heurísticas do investidor.

Como constatado, a HME considera um sistema eficiente de informação, tanto que desconsidera qualquer assimetria informacional (FAMA, 1970). Logo, buscadores como *Google Trends* aproximam o mercado da eficiência de informação, reduzindo a chance de assimetrias que permitem retornos acima do mercado. Todavia, os mesmos buscadores também podem servir de insumos para modelos de aprendizado de máquinas e algoritmos de *trading* que desafiam a Hipótese da Eficiência do Mercado.

3 ESTRUTURA DO MERCADO BRASILEIRO E PROCESSO DE INVESTIMENTO

Este capítulo busca apresentar a estrutura e principais características institucionais do mercado acionário brasileiro, bem como explicar as principais terminologias usadas no trabalho e discutir o processo de investimento.

3.1 ESTRUTURA DO MERCADO

O principal órgão regulador do mercado brasileiro é a (CVM, 2020b) criada na forma de autarquia vinculada ao Ministério da Economia em 07/12/1976 pela Lei 6.385/76 com o objetivo de normatizar, fiscalizar e desenvolver o mercado de valores mobiliários no Brasil. É a CVM quem determina as regras que as bolsas de valores devem seguir.

No Brasil existe apenas uma bolsa de valores, chamada de Brasil Bolsa Balcão (B3), a qual tem atuação na criação e administração de sistemas de negociação, compensação, liquidação, depósito e registro para todas as principais classes de ativos, desde ações e títulos de renda fixa corporativa até derivativos de moedas, operações estruturadas e taxas de juro e de *commodities* ver (B3, 2020a). Na bolsa, os investimentos são realizados por sociedades corretoras de valores mobiliários, mais conhecidas como corretoras, a partir de ordens de compra e venda de investidores.

As corretoras são agentes intermediários entre investidores e bolsa de valores para operar nas bolsas de valores, mercados e futuros para terceiros (os investidores), bem como encarregam-se da administração das carteiras e custódia dos títulos e valores mobiliários. As corretoras também estão sujeitas as regras e fiscalizações da CVM (CVM, 2020b).

Outra instituição importante é a *clearing houses* que é responsável por realizar o registro, compensação, liquidação e gerenciamento de de operações com derivativos e *comoddities*. A B3 possui sua própria clearing house, chamada *clearing* B3 que também atua como depositária central de títulos de ações.

A partir do desenvolvimento do mercado, (TERRA; LIMA, 2006) aponta que foram criados mecanismos pela CVM para melhorar a divulgação de informação e governança corporativa, dentre eles é os segmentos de mercados, os mais atuais são quatro: Mercado tradicional, Nível 1, Nível 2 e Novo Mercado. O nível Tradicional é o mais básico e o Novo Mercado é o mais moderno, assim é o mais rígido sobre como as informações devem ser divulgadas.

Para tanto, toda informação importante para o mercado, deve ser informada ao mercado como fato relevante, que é um documento público com informações que podem alterar a cotação da ação. (SILVA; FELIPE, 2010) afirma que há uma tendência no decréscimo no retorno após a divulgação de fatos pessimistas, enquanto a divulgaçao de fatos otimistas tem pouco impacto.

3.2 PROCESSO DE INVESTIMENTO

Para acompanhar as variações da bolsa de valores brasileira há diversos indicadores, o principal deles é o índice Ibovespa ou só Ibovespa foi criado em 1968 (B3, 2020a). A composição teórica é reavaliada a cada quatro meses geralmente em janeiro, maio e setembro. O Ibovespa busca destacar as empresas mais líquidas da bolsa de valores, ainda segundo a (B3, 2020a) o Ibovespa corresponde à 80% das negociações na bolsa brasileira, na composição atual o índioce contempla 77 empresas. Assim, se o Ibovespa está aumentando, há um consenso de que o mercado está em alta, portanto o Ibovespa é um indicador importante para o mercado brasileiro.

Para comprar ações, os investidores precisam usar o *home broker* da sua corretora, nesses sistema o investidor pesquisa a ação desejada a partir do *ticker* que é um identificador único da ação, informar o preço de compra ou venda desejado no sistema, a quantidade desejada e a data de validade da ordem. Se o valor de ordem equivaler ao valor de mercado do ativo, a operação será liquidada e o investidor passa a ser dono dessa ação. O indicador único ticker tem relação com o nome da empresa e contém um número que indica o tipo de ação, por exemplo: "ITSA4".

Logo, para operar, comumente os investidores fazem uso de algum tipo de estratégia para realizar investimentos. (COSTA; VARGAS, 2020) apontam que há dois grandes grupos de análises: análise técnica ou análise fundamentalista. A análise técnica tem como objetivo ganhos no curto prazo com ganhos na valorização ou desvalorização de ativos, escolhendo ativos a partir de gráficos e estatísticas, enquanto a análise fundamentalista busca escolher ativos para manter no longo prazo a partir dos fundamentos dos ativos, buscando ganhos de dividendos e valorização patrimonial. Uma variante da estratégia fundamentalista é *Buy and Hold* (comprar e segurar),

Já para obter lucro com a queda do mercado, é necessário realizar operações vendido ou a descoberto, essa operação é descrita por (JAEGER, 2003) como uma operação arriscada, em que se acredita que o preço do ativo está acima do preço ideal, assim usa-se mecanismos para lucrar com a queda do preço do ativo.

Todavia, independente da estratégia ou análise escolhida, o investidor precisa buscar algumas informações sobre os ativos, ou ainda para enviar suas ordens para a corretora é necessário conhecer o *ticker* da ação e informar o preço que ele deseja pagar. Então, em algum momento o investidor precisa coletar informaçoes sobre preço de mercado e *ticker* da ação. Para reforçar a hipótese, segundo (B3, 2020b) em novembro haviam 2.348.612 Cadastros de Pessoas Físicas (CPF) registrados na bolsa de valores. O número de investidores pessoa física cresceu nos últimos anos, já que em em 2010 haviam aproximadamente 600.000 cadastros. Esse aumento pode indicar mudanças no comportamento do mercado, como tornar a relação entre VHP e investimento intrínseca.

4 MATERIAL E MÉTODOS

O capítulo de material e métodos é dividido em cinco seções: A primeira discute a fonte dos dados e característica das variáveis usadas. A segunda seção apresenta o modelo empírico estimado e detalhes sobre as estimações. A terceira seção apresenta o modelo teórico da Análise de Componentes Principais (ACP). A quarta seção discute detalhes sobre modelos de séries temporais e a metodologia VEC-VAR. A última seção discute a elaboração dos algoritmos de *trading*.

4.1 DESCRIÇÃO DOS DADOS E ESTRATÉGIA METODOLÓGICA

Para entender a demanda de informação dos investidores são coletadas 79 Volumes Históricos de Pesquisas (VHP) da ferramenta *Google Trends* referentes a 77 tickers que compõem o Ibovespa e aos termos Ibovespa e IBOV. Essas variáveis servem como proxy da demanda de informação, pois além da vantagem da simplicidade, são resultados da ação voluntária dos investidores, sem sofrer intervenção alguma. Após a coleta é passado um filtro, em que apenas tickers com mais de 200 pontos na ferramenta são usadas, isso elimina os *tickers*: "GNDI3", "HAPV3", "HYPE3", "IGTA3", "MULT3", "NTCO3", "PCAR3", "PRIO3", "QUAL3", "TIMS3", "UGPA3", "VIVT3", "WEGE3", "YDUQ3". O filtro busca impedir vetores com muitos zeros, que prejudica a distribuição multivariada.

Os dados fornecidos pela ferramenta *Google Trends* não são brutos, pois o buscador de pesquisas normaliza o número de pesquisas realizadas na semana dividindo o número de pesquisas pelo valor máximo no intervalo, passando para um intervalo entre 0 e 100. A frequência semanal para o VHP do *Google Trends* tem início no domingo e termina no sábado. No *software* **R** os dados podem ser importados a partir do pacote **gtrendsR** (MASSICOTTE; EDDELBUETTEL, 2020).

Os dados referentes ao volume de negociações e valor do Ibovespa, são coletados do *Yahoo Finance*, mas diferente dos dados anteriores, as ações são negociadas apenas em dias úteis, assim o período começa na segunda-feira e terminando na sexta-feira. No *software* **R**, os dados sobre as variáveis de mercado podem ser coletados a partir do pacote **BatchGetSymbols** (PERLIN, 2020).

Este conjunto de informações foi coletado para o período entre entre 03 de abril 2016 a 03 de abril de 2020, com frequência semanal, totalizando 209 observações, esse conjunto é chamado de dados de treino. A partir do conjunto de treino, é desenvolvido o modelo e realizado testes econométricos. Ainda há o conjunto de dados de teste, que contém mais 37 observações, que compreendem as semanas entre 06/04/2020 a 08/12/2020. Os dados de teste são usados para avaliar a qualidade de previsão, e nesse caso, o algoritmo de *trading* numa

situação mais próxima da realidade, uma vez que os modelos de previsão podem ser muito bons dentro da amostra, mas em situações reais podem ter uma qualidade de previsão ruim, então é uma boa prática a separação entre dados de treino e teste, então o conjunto total de dados são 246 observações. Como a escala entre os valores é diferente, é interessante normalizar ou logaritmizar os dados. Dessa forma o aplica-se o logaritmo natural no Volume e o Ibovespa.

Após a coleta do VHP dos *tickers* é usado a Análise de Componentes Principais (ACP) para se reduzir a dimensão da quantidade de variáveis, de onde é extraída a *proxy* da atenção do investidor, usada para prever o Ibovespa. Para compreender a relação entre as variáveis é feito a cointegração de Johansen, por meio do método do autovalor, a partir disso é realizado o Vetor de Correção de Erros (VEC) que posteriormente é transformado um Vetor Auto Regressivo (VAR), processo conhecido como modelo VEC-VAR, para realizar a previsões e desenvolver uma estratégia de *trading* para o Ibovespa. De forma prática todas etapas desse estudo é realizado no *software* **R**, desde coletada de dados até tratamento e a estimação.

Entretanto para compreender a proposta de trabalho é preciso partir do princípio de que investidores demandam informações e para atender essa demanda o investidor pesquisa no *Google* o *ticker* da empresa para acompanhar notícias, encontrar sites com análises ou apenas se informar o preço desse *ticker* para então tomar uma decisão, seja de venda ou compra de um ativo ou papel, esse comportamento também pode ser considerado a atenção do investidor. Comportamento que também pode ser considerado a atenção do investidor. Decisão que também pode ser tomada no dia quando o investidor vai aportar (comprar um ativo), sendo que ele pode simplesmente verificar na internet os preços passados do ativo para decidir o preço de compra que irá enviar para sua corretora.

Logo, no caso de muitos investidores estarem pesquisando sobre um ativo ou *ticker*, então pode ocorrer um impacto no preço ou no volume de negociações desse ativo. Entretanto, isso pode não ser significativo, já que algumas empresas precisam de milhares de negociações para uma pequena alteração no preço. Mas um conjunto grande de pesquisas sobre as mais diferentes empresas impactam o Ibovespa, pois espera-se que essa massa de dados contenha a expectativa dos investidores.

Entretanto é difícil para o investidor compreender todo esse conjunto de informação, como qual é o impacto da pesquisa no *Google* sobre mais de 70 *tickers* no Ibovespa. Dessa forma é necessário simplificar toda essa informação, por isso a Análise de Componentes Principais é um passo importante. Essa abordagem tem o objetivo de simplificar esse conjunto de informações em *scores* que servem de *proxy* da atenção do investidor quando usada para previsão. Essa redução de dimensionalidade é interessante por manter em um mesmo índice *blue chips* (empresas grandes e consolidadas) e *small caps*(empresas novas e menores na bolsa).

Um mercado operado apenas por computadores atende com mais facilidade os pressupostos da HME já que conseguem reagir de forma eficiente à informação quantitativa, nesse sentido, os algoritmos são capazes de compreender o que significa uma alteração do número de pesquisas em buscadores, também são os algoritmos os melhores candidatos a prever preços, já

que não reproduzem o comportamento irracional dos investidores.

4.2 ANÁLISE DE COMPONENTES PRINCIPAIS

Estatística multivariada conceitualmente é um conjunto de técnicas que visam simplificar a interpretação de um grande conjunto de dados. As técnicas são geralmente associadas à redução de dimensão sem perder informação, além disso, as técnicas são úteis para predição e analisar a relação entre variáveis. Uma das principais técnicas de redução de dimensionalidade é a análise de componentes principais (ACP). Aplicações práticas são feitas a partir do pacote *psych*, ver (REVELLE, 2017). O desenvolvimento da ACP pode ser vista em (HAIR et al., 2009) e (CORRAR; FILHO; PAULO, 2009).

As técnicas de componentes principais transformam um grupo de variáveis em outro conjunto de variáveis chamado de Componentes Principais (CP's) que são basicamente combinaçoes lineares estimadas de forma a captar o máximo de variação do conjunto original de dados. Desta forma em p variáveis, cria-se k componentes principais de forma que $k < p$. Dentre as vantagens da técnica é a capacidade de lidar com multicolinearidade, enquanto a desvantagem é a sensibilidade à *outliers*, zeros e valores faltantes, como apontam (CORRAR; FILHO; PAULO, 2009).

Para compreender, considere as variáveis $X_1, X_2, \ldots, X_p$, para cada variável há n observações, como as variáveis não são dependentes considere a matriz de covariância como $\Sigma_{p\times p}$. Esse conjunto de dados é na prática uma matriz $X(n \times p)$. As médias e variâncias podem ser expressas na forma vetorial, conforme exposto em 4.1.

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{4.1}$$

Para obter os componentes principais, é necessário conhecer a matriz de covariância Σ, vista em 4.2:

$$Var - Cov(x) = \Sigma_{p\times p} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{np}^2 \end{bmatrix} \tag{4.2}$$

Assim associado a matriz Σ há autovalores $\lambda_1, \lambda_2, \ldots, \lambda_p$ e autovetores $e_1, e_2, \ldots, e_p$, então é possível escrever o sistema linear 4.3:

$$Z_i = e_i' X = e_{i1} X_1 + e_{i2} X_2 + \ldots + e_{ip} X_p \tag{4.3}$$

Assim, é estimado o i-ésimo componente principal. Além disso, usando a decomposição espectral $\Sigma = P\Lambda P$ em que P é a matriz de autovetores de Σ e Λ a matriz diagonal de autovalores toma-se o traço em 4.4:

$$tr(\Sigma) = tr(P\Lambda P) = tr(\Lambda P' P) = tr(\Lambda I) = tr(\Lambda) = \sum_{i=1}^{p} \lambda_i \tag{4.4}$$

Em que:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \tag{4.5}$$

Enquanto o $tr(\Sigma)$ é dado pela soma dos elementos da diagonal.

$$tr(\Sigma) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i \tag{4.6}$$

A contribuição de cada componente Z_i é expresso em percentual e é calculado a partir de:

$$C_k = \frac{Var(Z_1)}{\sum_{i=1}^{p} Var(Z_1)} \times 100 = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100 \tag{4.7}$$

O primeiro componente é aquele que explica a maior parte da variância. Assim como na Análise Fatorial na ACP é possível rotacionar as matrizes para facilitar a interpretação, sendo a rotação mais comum é a varimax. Esse método "forma um novo sistema de eixos ortogonais com o mesmo número de fatores e permite que o grupo de variáveis apareça com maior nitidez, facilitando a interpretação e a análise"(ZAMBRANO; LIMA, 2004). Matematicamente o processo é dado por:

$$V = \frac{1}{P} \sum_{j=1}^{r} \left[\sum_{i=1}^{p} \tilde{\alpha}_{ij}^4 - \frac{1}{p} \left(\sum_{i=1}^{p} \tilde{\alpha}_{ij}^2 \right) \right] \tag{4.8}$$

Em que $\tilde{\alpha}_{ij} = \dfrac{\hat{\alpha}_{ij}}{\hat{h}_i}$ é a carga do componente principal escalonada pela raiz quadrada da comunalidade da variável X_i. Com a variável rotacionada, obtém-se os escores (*scores*) que são a estimativa de cada componente por observação, análogo a predição de séries temporais. Como poder ser visto em 4.9.

$$X_i = A_i C_i + \varepsilon_i \tag{4.9}$$

Assim, para obter o resíduo basta isolar ε_i. A forma mais simples de estimar os coeficientes é a partir dos métodos de regressão.

$$\hat{C}_i = A'(AA' + \Psi)^{-1} X_i \tag{4.10}$$

Para tanto é necessário avaliar a qualidade e o ajustamento do modelo além da consistência interna dos componentes gerados. Para isso, utilizam-se o teste de esfericidade de Bartlett, o Kaiser-Meyer-Olkin (KMO) e o Alfa de Cronbach. Também é interessante avaliar se a amostra possui distribuição normal multivariada.

O teste de esfericidade de Bartlett verifica se a matriz de correlações é estatísticamente igual a matriz identidade, assim a $H_0 : P_{p\times p} = I_{p\times p}$. Para os dados ser adequados para a análise de Componentes Principais é necessário rejeitar a hipótese nula. A estatística do teste é definida por:

$$T = -\left[n - \frac{1}{6}(2p - 11)\right]\left[\sum_{i=1}^{p} ln(\lambda_i)\right] \sim \chi_{\frac{1}{2}p(p-1)} \tag{4.11}$$

O critério de adequabilidade mais usado é o KMO(Kaiser-Meyer-Olkin) que compara a correlação simples e parcial a partir de:

$$KMO = \frac{\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} r_{ij}^2}{\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} r_{ij}^2 + \Sigma_{i=1}^{p}\Sigma_{j=1}^{p} \alpha_{ij}^2} \tag{4.12}$$

Em que r_{ij}^2 é o coeficiente da correlação simples e α_{ij}^2 é o coeficiente de correlação parcial entre X_i e X_j, segundo (HAIR et al., 2009) uma estatística KMO acima de 0,6 é considerado o mínimo para ACP, um valor acima de 0,9 indica que os dados são adequados para análise multivariada.

Para análise de consistência interna, o Alfa de Cronbach é uma estatística de confiabilidade e consistência interna dos dados baseado na correlação entre as variáveis. Quanto mais próximo de um o alpha de Cronbach, mais adequado são os dados, um valor mínimo é conside-

rado 0,6 como visto em (HAIR et al., 2009). Formalmente a estatística é dada por:

$$\alpha = \frac{(cov/var)}{1+(p-1)(cov/var)} \tag{4.13}$$

Complementando, para a análise de componentes principais e análise fatorial algumas hipóteses são importantes (HAIR et al., 2009), afirma que as hipóteses que sustentam a análise fatorial são mais conceituais do que estatisticamente comprovadas, caso da normalidade multivariada, da multicolinearidade e da linearidade, assim, o mais importante são as características as quais impactam nas correlações entre as variáveis. Entretanto, a hipótese da normalidade multivariada é requerida quando se utiliza o método de máxima verossimilhança para a extração dos fatores, já para a análise dos componentes principais esta hipótese não é muito importante. A multicolinearidade é importante para ACP, uma vez que esta ferramenta identifica relacionamentos entre as variáveis.

Dessa forma, para avaliar a normalidade multivariada se usa o teste de Mardia. O teste analisa assimetria e curtose, se definirmos $m_r = [\sum(X - mx)^r]/n$ então é possível encontrar a assimetria e curtose multivariada a partir:

$$g_1 = m_3/(m_2)^{3/2} \; e \; g_2 = m_4/(m_2)^2 - 3 \tag{4.14}$$

Em que g_1 é estimativa da assimetria e g_2 é a estimativa da curtose.

O conjunto de dados que será aplicado a ACP será o VHP do *tickers* das ações, isso comporta 70 termos relacionados ao mercado brasileiro, para gerar a *proxy* da atenção do investidor ou demanda de informação, será obtido o *score* de cada observação. Nesse novo vetor será aplicado técnicas econométricas de séries temporais que são discutidas na próxima seção. O script da análise está disponível no Apêndice B.

4.3 INTRODUÇÃO A SÉRIES TEMPORAIS

Séries temporais são uma sequência de variáveis aleatórias estocásticas que são indexadas a um identificador de tempo. Dessa forma há um processo gerador de dados estocásticos por trás de séries temporais. Antes de apresentar o modelo de séries temporais usado é necessário entender os processos estacionários, testes de raiz unitária e integração. Todos detalhes de séries temporais são vistos em (BUENO, 2008), (MORETTIN, 2017) e (PFAFF et al., 2008).

O primeiro conceito é o de estacionariedade, basicamente uma série estacionária se desenvolve em torno de uma média, o conceito de estacionariedade pode ser dividido em duas categorias: (i) Estacionariedade estrita em que o processo estocástico $X_t; \quad t = 1,2\ldots \quad se \; \forall \; t$

tem a mesma distribuição conjunta, ou seja a Função Distribuição de Probabilidade é invariante no tempo, essa condição é difícil analisar em dados reais, por isso geralmente trabalha-se com estacionariedade Fraca(ii).

A estacionariedade Fraca (ii) é o suficiente para a maioria dos modelos de séries temporais, pois permite o uso da Lei dos Grande Números (LGN) e o Teorema do Limite Central (TLC), além disso as condições de estaconariedade Fraca são mais simples. Para o processo $X_t, \quad t = 1,2 \ldots N$ ser considerado fracamente estacionário, é necessário que:

- $E(X_t) = c$: isso significa que a média é constante para todo período t.
- $E|X_t|^2 \leq \infty$ isso significa que a Variância é finita.
- $\gamma(t,t-h) = \gamma(0,h) = \gamma(h)$ para todo tempo t e passo h. Isso significa que a autocovariância é finita e depende apenas da defasagem h.

As séries temporais que atendem esses 3 requisitos são chamados estacionárias, ou de ordem $I(0)$. Um processo estacionário comum é o Ruído Branco, denotado por $\varepsilon_t \sim RB(0,\sigma^2)$, além de ser estacionário o Ruído branco tem mais características importantes:

- $E(\varepsilon_t) = 0 \; \forall \; t \in Z$
- $E(\varepsilon_t^2) = \sigma^2 \; \forall \; t \in Z$
- $E(\varepsilon_t \varepsilon_s) = 0 \; \forall \; t \neq s \in Z$

Mas algumas séries temporais que não são estacionárias, podem se tornar ao aplicar o operador de diferença ∇ um número suficiente de vezes. O operador de diferença é basicamente a diferença do valor atual pelo anterior $\nabla X_t = (1-B)X_t$, onde $B = X_{t-1}$. Assim ao aplicar a diferença em um processo X_t que não é estacionário, cria-se um novo vetor, caso esse vetor seja estacionário após aplicar uma diferença, o processo é chamado de Processo estacionário de primeira ordem $I(1)$.

É importante não confundir o operador de diferença ∇ com o operador de retardo (*lag*) B. Ao operador de retardo temos a mesma série temporal defasada um período, também se perde uma observação, mas os valores não são alterados, assim o operador de retardo é: $BX_t = X_{t-1}$. Esses dois operadores, retardo e diferença são muito usados em análise de séries temporais.

São muitas as razões de uma série não ser estacionária, os casos mais comuns são processos com memória forte, tendência ou sazonalidade. Um processo que também comum é o passeio aleatório ou *random walk*, esse processo é um processo não estacionário, além disso, o processo tem diversas formas de se especificar com diferentes características como tendência e deslocamento, para fins de exemplos considere um passeio aleatório mais generalista que é o passeio aleatório com deslocamento δ.

$$X_t = X_{t-1} + \sum_{i=1}^{t} \varepsilon_t + \delta t \tag{4.15}$$

A média $E(X_t) = t\delta$ e variância $Var(X_t) = \sigma^2 t$, desse processo passam a ser influênciadas por t. Portanto, o processo não é estacionário, pois a série é uma recorrências de choques externos e uma tendência determinística, por isso o nome passeio aleatório com deslocamento. Porém a primeira diferença desse processo é estacionário. Pois $X_t - X_{t-1} = \delta + \varepsilon_t$. *Random walk* são processos I(1) muito comuns na literatura.

4.3.1 Vetores Autorregressivos (VAR)

O modelo Vetores Autorregressivos (VAR) é uma técnica multivariada para análise de séries temporais de curto prazo, a técnica é geralmente usada quando as séries quando todas variáveis são estacionárias. A forma básica do VAR pode ser vista em (BUENO, 2008) e consiste no uso de K variáveis endónegas do processo $X_t = (X_{1t}, \ldots, X_{kt}, \ldots, X_{Kt}, \ldots)$, para $k = 1, \ldots, K$. O processo VAR é definido por:

$$X_t = A_1 X_{t-1} + \ldots + A_p X_{t-p} + u_t \tag{4.16}$$

Em que A_i é a matriz de coeficientes com dimensão $K \times K$ para $i = 1$, enquanto u_t é o processo com dimensão K em que $E(u_t) = 0$, por fim $E(u_t u_t') = \sum_u$ que é um processo ruído branco. Uma característica importante dos processos VAR é estabilidade. A estabilidade implica em média, variância e covariância constantes no tempo. Uma forma de checar a estabilidade é a partir do polinômio característico de forma que a partir da equação 4.17.

$$det(I_K - A1z - \ldots - A_p z^p) \neq 0, para |z| \leq 1 \tag{4.17}$$

Se a solução tem raiz para $z = 1$ então as variáveis no processo VAR(p) é integrado de ordem I(1). Esse é o caso em que o Vetor de Correção de Erros (VEC) é mais adequado. Na prática os testes de estabilidade avaliam os autovalores do polinômio característico. Caso não haja raiz, o modelo VAR é estável. O processo VAR(p) pode ser reescrito como VAR(1).

$$\xi_t = A\xi_{t-1} + \upsilon_t \tag{4.18}$$

Em que.

$$\xi_t = \begin{bmatrix} X_t \\ \vdots \\ X_{t-p+1} \end{bmatrix}, A = \begin{bmatrix} A_1 & A_2 & \cdots & Ap-1 & Ap \\ I & 0 & \cdots & 0 & 0 \\ 0 & I & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I & 0 \end{bmatrix}, \upsilon_t = \begin{bmatrix} u_t \\ \vdots \\ 0 \end{bmatrix} \tag{4.19}$$

A dimensão de ξ_t e υ_t é $KP \times 1$, enquanto a dimensão de A é $K_p \times K_p$. Se o módulo dos autovalores de A está dentro do círculo unitário, o modelo é considerado estável. Para os vetores de variáveis endógenas $X_t, \ldots, X_N$, com suficientes amostras passadas como $X_{-p+1}, \ldots, X_0$, os coeficientes do processo VAR(p) pode ser estimado por mínimos quadrados, apliacados separadamente para cada equação.

O modelo VAR(p), além dos coeficientes estimados, gera informações sobre a decomposição dos erros da previsão (FEVD), útil para avaliar a contribuição de cada variável para cada passo de previsão e as funções impulso resposta (IRF).

A FEVD pode ser estimada a partir da decomposição de Wold, exposta pela equação 4.20:

$$X_t = \Phi_0 u_t + \Phi_1 u_{t-1} + \Phi_2 u_{t-2} + \ldots, \tag{4.20}$$

Em que $\Phi_0 = I_K$ e Φ_s é calculado recursivamente a partir de.

$$\Phi_s = \sum_{j=1}^{s} \Phi_{s-j} A_j, s = 1, 2, \ldots, \tag{4.21}$$

Note que $A_j = 0$ para $j > p$. Para prever horizontes $h \geq 1$ do VAR(p) empírico pode-se usar.

$$X_{N+h|N} = A_1 X_{N+h-1|N} + \ldots + A_p X_{N+h-p|N}; \quad X_{N+j|N} = X_{N+j} \forall j \leq 0 \tag{4.22}$$

O erro de previsão da matriz de covariância Σ_u merece destaque merece, afinal, segundo (BUENO, 2008) o determinante da matriz é usado para testar hipóteses nos casos multivariados, análogo o quadrado do resíduo em casos univariados.:

$$Cov = \begin{bmatrix} X_{N+1} - X_{N+1|N} \\ \vdots \\ X_{N+h} - X_{N+h|N} \end{bmatrix} = \begin{bmatrix} I & 0 & \cdots & 0 \\ \Phi_1 & I & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \Phi_{h-1} & \Phi_{h-2} & \cdots & I \end{bmatrix} (\sum u \otimes I_h) \begin{bmatrix} I & 0 & \cdots & 0 \\ \Phi_1 & I & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \Phi_{h-1} & \Phi_{h-2} & \cdots & I \end{bmatrix}^{\mathsf{T}} \tag{4.23}$$

Dessa forma as matrizes Φ_i contém os coeficientes empíricos da média móvel de Wold. O símbolo $\otimes$ é o produto Kronecker.

Por sua vez, a função resposta ao impulso (IRF) simula um choque em uma das variáveis do modelo VAR(p) baseado na versão de Wold, é uma técnica útil para entender a dinâmica entre variáveis endógenas. Os (i, j) são os coeficientes das matrizes Φ_s são interpretados como a resposta esperada da variável $X_{r,t+s}$ dado um mudança na variável X_{jt}. Os efeitos são acumulados no tempo $s = 1, 2, \ldots,$, assim se tem o impacto simulado na variável i da mudança na variável j.

A IRF deriva da decomposição de Choleski ($\Sigma_u = PP^T$), onde P é uma matriz triangular. A representação de médias móveis pode ser transformada em:

$$X_t = \Psi_0 \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \ldots, \tag{4.24}$$

Em que $\varepsilon_t = P^{-1} u_t$ e $\Psi_i = \Phi_i P$ e $\Psi_o = P$. Segundo (BUENO, 2008) a ordem das variáveis no modelo podem gerar diferentes Funções Resposta ao Impulso.

O intervalo de confiança é feito a partir técnicas de *bootstrap*, que é estimado por geradores de números pseudo-aleatórios. O desvio padrão do percentil é definido por:

$$CI_s = [s^*_{\alpha/2}, s^*_{1-\alpha/2}] \quad , \tag{4.25}$$

Em que $s^*_{\alpha/2}$ e $s^*_{1-\alpha/2}$ formam o intervalo superior e inferior e $\alpha/2$ e $1-\alpha/2$ os quantis da estimativa *bootstrap*. Mais detalhes são vistos em (LUTKEPOHL, 2006).

Uma forma alternativa de visualizar o processo VAR é a forma apresentada por (BUENO, 2008) o modelo VAR é expresso na forma reduzida por:

$$X_t = A^{-1} B_0 + \sum_{i=1}^{p} A^{-1} B_1 X_{t-i} + A^{-1} B \varepsilon_t = \Phi_0 + \sum_{i=1}^{p} \Phi_i X_{t-i} + e_t \tag{4.26}$$

Sendo $\Phi_i = A^{-1} B_i; i = -, \ldots, p; B\varepsilon_t = Ae_t$. Essa segunda forma apresentada condensa muita informação, mas deixa claro quanta informação a matriz Φ agrega. Uma forma bivariada

do modelo VAR com variáveis y_t e _t é

$$y_t = b_{10} - a_{10}z_t + b_{11}y_{t-1} + b_{12}z_{t-1} + \sigma_y\varepsilon_{yt}. \tag{4.27}$$

$$z_t = b_{20} - a_{20}y_t + b_{21}y_{t-1} + b_{22}z_{t-1} + \sigma_z\varepsilon_{zt}. \tag{4.28}$$

Nessa notação a condição de estabilidade desse modelo VAR é ter autovalores $(I - \Phi L)$ dentro do círculo unitário.

4.3.2 Cointegração e Vetor de Correção de Erros (VEC)

O processo de cointegração de Johansen ver (JOHANSEN, 1995a) e (JOHANSEN, 1995b) é uma técnica multivariada que analisa a relação de longo prazo entre séries temporais integradas de primeira ordem. Considerando que X_t e Z_t como processos I(1) e μ_t o resíduo dessa regresão seja um processo I_0, então X_t e Y_t são processos cointegrados. A vantagem dessa técnica é que segundo (MORETTIN, 2017), ativos financeiros geralmente apresentam uma relação comum de longo prazo.

Formalmente, os elementos do vetor X_t $n \times 1$ são ditos cointegrados de ordem(db) denotados por $X_t \sim CI(db)$ se existe pelo menos um vetor β não nulo tal que $u_t = X_t'\beta - \mu_1 t - \mu_0 \sim I(d-b), b > 0$, essa é a definição de cointegração de Campbell e Perron estendida para incluir tendência e constante:

Dessa forma, em um sistema com n variáveis endógenas, podem haver $n-1$ processos de cointegração. Segundo (BUENO, 2008) a vantagem dessa definição é que as variáveis não precisam ter a ordem ordem de integração.

De forma prática, o processo VAR é um caso especial do processo VEC, em que as variáveis são estacionárias. Entretanto a maneira mais simples de se chegar ao VEC é a partir do VAR, portanto reconsidere o processo VAR.

$$X_t = A_1X_{t-1} + \ldots + A_pX_{t-p} + u_t \tag{4.29}$$

O processo VAR pode ser transformado em um Vetor de Correção de Erros (VEC) a partir de operações de diferença, conforme 4.30.

$$\Delta X_t - \alpha\beta' X_{t-p} + \Gamma_1\Delta X_{t-1} + \ldots + \Gamma_{p-1}X_{t-p+1} + u_t \tag{4.30}$$

Em que.

$$\Gamma_i = -(I - A_1 - \ldots - A_i), \quad i = 1, \ldots, p-1 \tag{4.31}$$

Ainda a matriz $\prod$ carrega a informação de quantos vetores de cointegração existem.

$$\prod = \alpha\beta' = -(I - A_1 - \ldots - A_p) \tag{4.32}$$

A matriz Γ_i capta a relação de longo prazo, uma vez que o VEC foi especificado na forma de longo prazo. A forma transitória é:

$$\Delta X_t - \alpha\beta' X_{t-1} + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_{p-1} X_{t-p+1} + u_t \tag{4.33}$$

Em que.

$$\Gamma_i = -(A_{i+1} - \ldots - A_p), \quad i = 1, \ldots, p-1 \tag{4.34}$$

A forma reestruturada e condensada do VEC apresentada em (BUENO, 2008) é melhor para compreender o processo de Johansen, por isso considere essa especificação de VEC.

$$\Delta X_t = \Phi X_{t-1} + \sum_{i=1}^{p-1} \Lambda_i \Delta X_{t-i} + e_t \tag{4.35}$$

Ao maximimizar essa equação com restrições na matriz de covariância, é possível obter os autovalores da matriz Φ, após obter os autovalores, ordena-se em ordem decrescente: $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. Os autovalores são testados iterativamente se esse é o número de processos cointegrados reais a partir da estatística do autovalor de Johansen.

$$LR(r) = -Tln(1 - \hat{\lambda}_{r+1}) \tag{4.36}$$

A hipótese nula do teste é $H_0 : r = r^*0$, ou seja, o valor testado r é o número real de cointegrações. Como a hipótese alternativa é $H_1 : r = r^* + 1$ o teste é um processo iterativo, o primeiro teste considera $r = 0$, se rejeitar essa primeira hipótese, há pelo menos um $r = 0 + 1$ vetores de cointegração.

Porém, (BUENO, 2008) apresenta cinco formas de se especificar o processo de cointegração, em nível, com intercepto em ou um dois vetores, tendência em um ou dois vetores.

Dessa forma, a forma mais completa do processo de cointegração é:

$$\Delta X_t = \alpha[\beta'[X_{t-1} + \mu_0 + \mu_1(t-1)]] + (\delta_0 + \delta_{1t}) + \sum_{i=1}^{p-1} \Lambda_i \Delta X_{t-1} + e_t \tag{4.37}$$

As cinco formas de cointegração derivam desse modelo generalista. Todo processo de transformação de séries temporais pode se chegar no caminho inverso, então é possível se estimar um VEC e transformar em um VAR, a partir da metodologia VEC-VAR proposta por (LUTKEPOHL, 2006), por meio do rank da matriz $\prod$ que tem a informação de quantos vetores de cointegração existem.

4.3.3 Teste de raiz unitária

Os testes de raiz unitária são úteis para se descobrir se a série é estacionária, ou até a ordem de integração, que é o número de diferenciações necessárias para a série se tornar estacionária. Um teste adequado é o teste desenvolvido por Denis Kwiatkowski , Peter C. B. Phillips, Peter Schmidt e Yongcheol Shin chamado de KPSS, ver (KWIATKOWSKI et al., 1992).

Para entender o teste KPSS suponha que é possível decompor a série temporal $X_t, \quad t = 1,2\ldots N$ em uma tendência $\delta_t, \quad t = 1,2\ldots N$, um passeio aleatório $\alpha_t = \alpha_{t-1} + \upsilon t, \quad \upsilon \sim i.i.d.(0.\sigma^2), \quad t = 1,2\ldots N$ e um erro $\varepsilon_t, \quad t = 1,2\ldots N$ e uma constante μ. Definindo $e_t = \alpha_t + u_t$ tem-se.

$$X_t = \mu + \delta_t + e_t \tag{4.38}$$

Calculando o resíduo dessa regressão 4.38 em todo t, obtêm-se:

$$\hat{e_t} = X_t - \hat{\mu} - \hat{\delta}_t \tag{4.39}$$

Como a soma parcial dos resíduos é definida por $S_t = \sum_{j=1}^{t} \hat{e_j}$ A estatística do teste KPSS é uma estatística LM definida por:

$$LM = \sum_{t=1}^{N} \frac{S_t^2}{N^2 \hat{\sigma_{lp}^2}} \tag{4.40}$$

Em que $\hat{\sigma}_{lp}^2$ é a variância de longo prazo estimada por.

$$\hat{\sigma}_{lp}^2 = \frac{\sum_{t=1}^{N} \hat{e_t^2}}{N} + \frac{2}{N} \sum_{j=1}^{M} \omega\left(\frac{j}{M+1}\right) \sum_{t=j+1}^{N} \hat{e}_t \hat{e_{t-j}} \tag{4.41}$$

A Hipótese nula do teste é $H_0 : X \sim (0)$ ou $H_0 : \sigma^2 = 0$, ou seja, estacionário. Então a hipótese alternativa se torna, $H_1 : X \sim (1)$ ou $H_1 : \sigma^2 > 0$, que indica um processo não estacionário. Se X_t é $I(1)$ o numerador vai explodir o que torna a estatística grande o suficiente para rejeitar a hipótese nula, mas se aplicar o mesmo teste na primeira diferença dessa série, o teste indicará que a série é estacionária. A ordem de integração das variáveis influencia bastante no modelo a ser utilizado, o teste de raiz unitária guia a escolha entre Vetores Autorregressivos (VAR) ou Vetores de Correção de Erros (VEC).

4.3.4 Número ótimo de defasagens

Para estimar corretamento os modelos VAR ou VEC é importante usar o número certo de defasagens no modelo VAR, por isso, usa-se critérios de informação que mostram o melhor número de defasagens. São quatro critérios de informação usados para obter essa informação, Akaike (AIC), Hannan Quinn (HQ), Schwarz (SC) e Erro de Previsão Final (FPE). A estimativa dos critérios de informação são feitas a partir de um VAR, como em:

$$y_t = A_1 y_{t-1} + + A_p y_{t-p} + CD_t + u_t \tag{4.42}$$

Então cada critério de seleção aplica uma função com o resíduo e matriz de variância. Com isso, o modelo aponta o número de defasagens (lags) que leva ao melhor resíduo. A formalização de cada um dos quatro critérios está apresenta nas equações de 4.43 a 4.46.

$$AIC(n) = \ln\det(\tilde{\Sigma}_u(n)) + \frac{2}{T} nK^2 \quad , \tag{4.43}$$

$$HQ(n) = \ln\det(\tilde{\Sigma}_u(n)) + \frac{2\ln(\ln(T))}{T} nK^2 \quad , \tag{4.44}$$

$$SC(n) = \ln\det(\tilde{\Sigma}_u(n)) + \frac{\ln(T)}{T} nK^2 \quad , \tag{4.45}$$

$$FPE(n) = ft(\frac{T+n^*}{T-n^*})^K \det(\tilde{\Sigma}_u(n)) \quad , \tag{4.46}$$

Em que $\tilde{\Sigma}_u(n) = T^{-1}\Sigma_{t=1}^{T}\hat{u}_t\hat{u}_t'$ é análogo a soma quadrática dos resíduos e n^* é o número total de parâmetros em cada equação e n indica o a ordem de defasagens. Assim cada critério de seleção apresenta um número ótimo de retardos, geralmente eles convergem para um mesmo valor, que é o mais adequado dentre todos testados.

4.3.5 Correlação serial

E para avaliar o ajustamento do modelo são aplicados testes de diagnóstico para auto-correlação, heterocedasticidade e normalidade. Os testes de autoccorelação mais comuns nos modelos VAR são de Portementau, Breusch-Godfrey e Edgerton-Shukur. O ideal para amostras pequenas é usar as versões com correção para amostras pequenas ver (EDGERTON; SHUKUR, 1999). O teste de autocorrelaçao é importante pois segundo (LUTKEPOHL, 2006) é ele quem garante um bom modelo VEC-VAR. O teste de Portmanteau, avalia a ausência de autocorrelação de ordem h, tendo a estatística do teste definida como:

$$Q_h = T\sum_{j=1}^{h} tr(\hat{C}_j'\hat{C}_0^{-1}\hat{C}_j\hat{C}_0^{-1}) \quad , \tag{4.47}$$

Em que:

$$\hat{C}_i = \frac{1}{T}\sum t = i+1^T \hat{u}_t\hat{u}_{t-i}' \tag{4.48}$$

A estatística do teste é aproximadamente distribuída como $\chi^2(K^2(h-p))$. A estatística pode ser gerada na versão assintótica ou na versão para pequenas amostras. Para pequenas amostras a estatística pode ser vista na equação 4.49:

$$Q_h^* = T^2\sum j = 1^h \frac{1}{T-j} tr(\hat{C}_j'\hat{C}_0^{-1}\hat{C}_j\hat{C}_0^{-1}) \quad , \tag{4.49}$$

A versão usada é para pequenas amostras. A H_0 indica ausência de autocorrelação, enquanto a H_1 indica a presença da correlação serial.

4.3.6 Normalidade

Para avaliar a normalidade do resíduo, o teste de normalidade Jarque-Bera pode ser aplicado nas versões univariadas e multivariadas nos resíduos do modelo VAR. O teste multivariado usa os resíduos que são normalizados pela decomposição de Choleski da variância e covariância. Uma discussão sobre o teste pode ser vista em (BERA; JARQUE, 1981), (JARQUE; BERA, 1980) (JARQUE; BERA, 1987) e (BRUGGEMANN; LUTKEPOHL; SAIKKONEN, 2006).

A estatística do teste é:

$$JB_{mv} = s_3^2 + s_4^2 \tag{4.50}$$

Em que s_3^2 e s_4^2 é definido por:

$$s_3^2 = T b_1^T b_1 \frac{1}{6} \tag{4.51}$$

E.

$$s_4^2 = T(b_2 - 3_K)^T \frac{(b_2 - 3_k)}{24} \tag{4.52}$$

Onde b_1 e b_2 são os momentos dos vetores normalizados do resíduo $\hat{u}_t^s = \tilde{P}^{-(\hat{u}_t - \bar{\hat{u}}_t)}$ e $\tilde{P}$ é a matriz triangular de baixo da matriz com diagonal positiva em que $\tilde{P}\tilde{P}^T = \tilde{\Sigma}$, i.e., a decomposição de Choleski do resíduo da matriz de covariância. A estatística do teste JB_{mv} tem distribuição $\chi^2(2K)$, a assimetria multivariada é s_3^2 e a curtose multivariada é s_4^2 tem distribuições $\chi^2(K)$

4.3.7 Efeitos ARCH

Para avaliar a heterocedasticidade, utiliza-se o teste multivariado ARCH-LM (heteroscedasticidade condicional autorregressiva). Esse teste pode ser aplicado em modelos univariados e multivariados, um aprofundamento do teste é visto em (ENGLE, 1982) e (HAMILTON; SUSMEL, 1994). O teste é baseado na regressão 4.53:

$$\overrightarrow{h}(\hat{u}_t \hat{u}_t') = \beta_0 + B_1 \overrightarrow{h}(\hat{u}_{t-1}\hat{u}_{t-1}') + + B_q \overrightarrow{(}\hat{u}_{t-q}\hat{u}_{t-q}' + v_t) \tag{4.53}$$

Em que v_t representa o erro do processo e $\overrightarrow{h}$ é o operador que empilha as colunas. A dimensão de β_0 é $\frac{1}{2}K(K+1)$ e a dimensão da matriz de coeficientes B_i é $i = 1,,q, \frac{1}{2}K(K+1) \times \frac{1}{2}K(K+1)$. A hipótese nula é: $H_0 := B_1 = B_2 = = B_q = 0$ enquanto a alternativa: $H_1 : B_1 \neq 0 or B_2 \neq 0 ou B_q \neq 0$. A estatística do teste é:

$$VARCH_{LM}(q) = \frac{1}{2}TK(K+1)R_m^2 \quad , \tag{4.54}$$

Em que.

$$R_m^2 = 1 - \frac{2}{K(K+1)}tr(\tilde{}_0^{-1}) \quad , \tag{4.55}$$

E $\hat{\Omega}$ indica a matriz de covariância obtida no modelo de regressão O teste tem distribuição $\chi^2(qK^2(K+1)^2/4)$. O teste pode ser aplicado na versão multivarada ou versão univariada.

4.3.8 Metodologia VEC-VAR

A transformação de um modelo VEC em um modelo VAR é uma proposta feita por (LUTKEPOHL, 2006). Para manter a notação dos autores, considere o processo y_t integrado de primeira ordem. Esse processo é aplicado um procsso VEC na forma:

$$\Delta y_t - \alpha\beta' y_{t-1} + \Gamma_1 \Delta y_{t-1} + \ldots + \Gamma_{p-1} y_{t-p+1} + u_t, \quad t-1,2,\ldots, \tag{4.56}$$

Onde X_t é o vetor K dimensional das variáveis, e α e β são matrizes $K \times r$ de rank r. Precisamente β é a matriz de cointegraçãoe e r é o rank de cointegração. Vale notar que $\alpha\beta' X_{t-1}$ é o termo de correção de erro. O vetor Γ tem dimensão $K \times K$ e representa as matrizes de coeficientes de curto prazo, o erro u_t tem comportamento ruído branco, e matriz de covariância $\Sigma_u, u_t \sim (0, \Sigma_u)$. Ainda $X_{-p+1}, \ldots, X_0$ são as condições iniciais do sistema multivariado. É possível reescrever o sistema em uma forma VAR(p):

$$y_t = A_1 y_{t-1} + \ldots + A_p y_{t-p} + u_t \tag{4.57}$$

Onde o vetor $A_1 = \alpha\beta' + I_K + \Gamma_1$, $A_i = \Gamma_i - \Gamma_{i-1}$, ainda $A_p = -\Gamma_{p-1}$. Então o modelo VAR inclui p retardos, quando há p-1 diferenças no modelo VEC.

Agora considere os parâmetros estimados a partir de Johansen (1995), então considere:

$$X_{t-1} \begin{bmatrix} \Delta y_{t-1} \\ \vdots \\ \Delta y_{t-p+1} \end{bmatrix} \tag{4.58}$$

De forma compacta essa mesma matriz pode ser reescrita na forma: $\Delta Y = \alpha\beta' Y_{t-1} + \Gamma X + U$.

$$\hat{\Gamma} = (\Delta Y - \alpha\beta' Y_{t-1} M + \hat{U}) \tag{4.59}$$

Onde $M = I = X'(XX')^{-1}X$. As estimativas de α e β são feitas pela correlação canônica, ver (ANDERSON, 1962) ou redução de rank como em johansen (1995) (JOHANSEN, 1995a). A solução de Johansen é vista em:

$$S_{00} = T^{-1}\Delta Y M \Delta Y', \quad S_{01} = T^{-1}\Delta Y M Y'_{-1}, \quad S_{11} = T^{-1}\Delta Y_{-1} M Y'_{-1} \tag{4.60}$$

Resolvendo o sistema e obtendo o autovalor, temos:

$$det(\lambda S_{11} - S'_{01} S_{00}^{-1} S_{01}) \tag{4.61}$$

Ordenando o autovalor do maior para o menor, com os associados autovatores $V = [b_1, , b_k]$ pode-se verificar a condição $\lambda_i S_{11} b_i = S'_{01} S_{00}^{-1} S_{01} b_i$, e assume-se que os valores são normalizados por $V' S_{11} V = I_K$. A partir disso, obtêm os estimadores α e β através:

$$\hat{\alpha} = \Delta Y M Y'_{-1} \hat{\beta} (\hat{\beta}' Y_{-1} M Y'_{-1} \hat{\beta})^{-1} \tag{4.62}$$

$$\hat{\beta} = [b_1, \ldots, b_r] \tag{4.63}$$

O estimador $\hat{\alpha}$ pode ser considerado o estimador de mínimos quadrados de

$$\Delta Y M = \alpha \hat{\beta}' Y_{-1} M + \tilde{U} \tag{4.64}$$

Se considerar $\hat{\Gamma} = (\Delta Y - \alpha\beta' Y_{t-1}) X'(XX')^{-1}$ um estimador factível de Γ, todos os estimadores sob condições gaussianas são consistentes assimptóticamente, além disso, é possível estimar os parâmetros estruturais a partir de uma da logartimização natural, isso pode ser visto abaixo:

$$lnl_c(B) = C_0 - \frac{T}{2} ln|B|^2 - \frac{T}{2} tr(B^{'-1} B^{-1} \tilde{\Sigma}_u) \tag{4.65}$$

A maximização dessa função em B pode ser feita por meio de métodos numéricos, já que não há forma fechada de estimação. Esse estimador é assimptóticamente consistente e assimptóticamente normal. (LUTKEPOHL, 2006) Afirma que para a análise de resíduo basta garantir que não há presença de autocorrelação, ou ainda garantir a estabilidade. A previsão realizada pelo modelo VEC-VAR é usada no algoritmo de *trading*. O script da análise de séries temporais está disponível no Apêndice C.

4.4 ALGORITMO DE *TRADING*

Elaboração de algoritmos de *trading* encontrada em (CONLAN, 2016), (CHAN, 2009) e (DUNIS; LAWS; NAIM, 2004). Esse processo consiste na aplicação de métodos quantitativos, estatísticos e econométricos para prever o retorno ou preço, e apartir disso desenvolver um sistema de regras para decisão de compra ou venda de um ativo. Com o modelo VEC-VAR é possível fazer previsões de curto prazo e desenvolver um algoritmo para tomar decisões de quando comprar ou vender ativos. As pesquisas sobre algoritmos de *trading* ignoram todas taxas do processo de investimento, para facilitar a pesquisa, uma vez que diferentes corretoras cobram diferentes taxas de corretagem.

São dois algoritmos de *trading* desenvolvidos, o primeiro simula ganhos apenas na compra e venda de ativos, aproveitando todas altas do mercado. O segundo simula ganhos com compra e além da venda, simula posições a descoberto internacionalmente conhecido como *short*, dessa forma é um pouco mais complexo, arriscado e por isso, espera-se que mais volátil.

As regras que compõem o primeiro algoritmo desenvolvido são simples: se o valor previsto para o Ibovespa para a próxima semana é maior que o valor previsto para a semana atual (o que indica uma alta do mercado), o algoritmo simula uma compra, caso contrário é simulado uma venda. O retorno acumulado da estratégia é comparado com o retorno da estratégia *Buy and Hold*, que é basicamente o retorno acumulado do Ibovespa para o mesmo período. O algoritmo simples é denotado por $Algoritmo\ S$ e pode ser visto em 4.66.

$$Regra^{Simples}(\widehat{Ibov_{t+1}}, \widehat{Ibov_t}) = \begin{cases} simula\ \ compra\ se\ \widehat{Ibov_{t+1}} > \widehat{Ibov_t}; \\ simula\ \ venda\ se\ cc. \end{cases} \tag{4.66}$$

O conjunto de regras do segundo algoritmo é ligeiramente mais complexo: se o valor previsto para o Ibovespa para a próxima semana é maior que o valor previsto para a semana atual (o que indica uma alta do mercado), o algoritmo simula uma compra, caso contrário é simulado uma venda e operação a descoberto ou vendido. Então o algoritmo busca simular ganhos sob duas perspectivas, a alta do mercado e baixa do mercado, por isso o modelo é denotado por $AlgoritmoD$ e pode ser visto na equação 4.67.

$$Regra^{Duplo}(\widehat{Ibov_{t+1}}, \widehat{Ibov_t}) = \begin{cases} simula \;\; compra \; se \; \widehat{Ibov_{t+1}} > \widehat{Ibov_t}; \\ simula \;\; venda \; e \; posio \; vendido \; se \; cc. \end{cases} \tag{4.67}$$

Ainda são produzaidas algumas estatísticas descritivas com o retorno de cada algoritmo a fins de comparação. Se o retorno da estratégia de *trading* é maior que a estratégia *Buy and Hold* a consequência da HME é violada, caso contrário a HME é preservada. O script referente ao algoritmo de trading está disponível no Apêndice D.

4.5 MODELO EMPÍRICO

O modelo empírico parte da Análise de Componentes Principais que reduz todos vetores baseados no VHP dos *tickers* em que compõem o Ibovespa em um fator: $PC1$. Esse único fator representa a *proxy* da demanda de informação, apesar da literatura indicar que se deve elaborar uma quantidade de componentes que expliquem um mínimo de 60% da variação dos dados, ou usar testes informais como do "cotovelo", a presente pesquisa limita-se a elaborar um componente principal, para manter o significado econômico da *proxy* da demanda de informação.

A etapa posterior consiste em verificar o comportamento das séries, isso é, a ordem de integração das séries. Caso sejam da mesma ordem de integração é preciso escolher entre a metodologia mais adequada: VAR se todas variáveis estacionárias ou VEC para se todas variáveis apresentarem a mesma ordem de integração. Após isso é necessário usar um critério de informação para saber o número ótimo de defasagens.

Objetivamente, é analisado a relação da *proxy* da demanda de informação com o Volume de negociações do Ibovespa e Valor ou Preço do Ibovespa. Para isso, se usa a metodologia de Johansen para verificar quantos vetores de cointegração existem. Caso exista pelo menos um vetor de cointegração é estimado o vetor de correção de erros (VEC), esse vetor é posteriormente transformado em um VAR, pelo método VEC-VAR proposto por (LUTKEPOHL, 2006). A previsão do modelo VAR é usada como entrada das regras do algoritmo de *trading*.

Portanto, considere o modelo VEC que relaciona Ibovespa (IBOV) e a *proxy* da atenção do investidor (PC1), com dois retardos, denominado modelo 1. Uma versão reduzida dessa equação é apresentada na seção de Cointegração e VEC.

$$\begin{aligned} \Delta P_t^{Ibov} = \Phi_t^{ibov} + \alpha^{Ibov}(P_{t-1}^{Ibov} - \beta_1 \Delta PC1t - 1) + \Gamma_1 \Delta P_{t-1}^{Ibov} + \Gamma_2 \Delta P_{t-2}^{Ibov} + \\ \delta_1 \Delta PC1_{t-1} + \delta_2 \Delta PC1_{t-2} + u_t \end{aligned} \tag{4.68}$$

Ou ainda o mesmo modelo VEC onde se busca entender a relação entre Volume de negociações (Vol) e a *proxy* da demanda (PC1), com quatro retardos, denominado modelo 2.

Para mais defasagens como 4, basta adicionar os vetores onde $t = 1, \ldots, 4$.

$$\Delta Vol_t = \Phi_t^{Vol} + \alpha^{Vol}(P_{t-1}^{Vol} - \beta_1 \Delta PC1t - 1) + \Gamma_1 \Delta P_{t-1}^{Vol} + \Gamma_2 \Delta P_{t-2}^{Vol} + \delta_1 \Delta PC1_{t-1} + \delta_2 \Delta PC1_{t-2} + u_t \tag{4.69}$$

Esse modelo é transformado em um modelo VAR, assim, o resíduo u_t é analisado a por meio do modelo VAR, a partir dos testes de correlação serial, normalidade e efeitos ARCH. Para entender melhor a relação entre as variáveis ainda é realizado a decomposição da variância (FEVD) e Função Resposta ao Impulso (FRI). Por fim, para o modelo do Ibovespa, ainda é feito a previsão desse modelo passos a frente, que é a entrada do algoritmo de *trading*.

5 RESULTADOS

O capítulo de resultados é dividido por análise, iniciando com os resultados da análise de componentes principais e dos *scores* estimados, pois são os *scores* as *proxies* da atenção do investidor. A segunda seção apresenta os resultados descritivos a partir da análise exploratória de dados. A terceira seção apresenta os resulados econométricos, com teste de raiz unitária, Cointegração de Johansen, Vetor de Correção de Erros e VAR. Por último, é apresentada comparação entre as estratégias de *trading* e a estratégia *Buy and Hold*.

5.1 COMPONENTES PRINCIPAIS

Os dados em que são aplicados a Análise de Componentes Principais (ACP) são as pesquisas pelos *tickers* das empresas que compõem do Ibovespa e as pesquisas sobre o Ibovespa. Para aplicar a ACP é necessário uma correlação muito forte entre as variáveis, a correlação pode ser vista no mapa de calor abaixo, tonalidades de azul forte indicam a correlação mais próximo de um (1), tonalidade branco indicam valores mais próximo de zero (0), enquanto tonalidades de vermelho indicam correlação próxima de menos um (-1). De forma reduzida existe correlação positiva entre a maioria das variáveis que pode ser vista na Figura 5.1, o que indica um sinal positivo para a ACP.

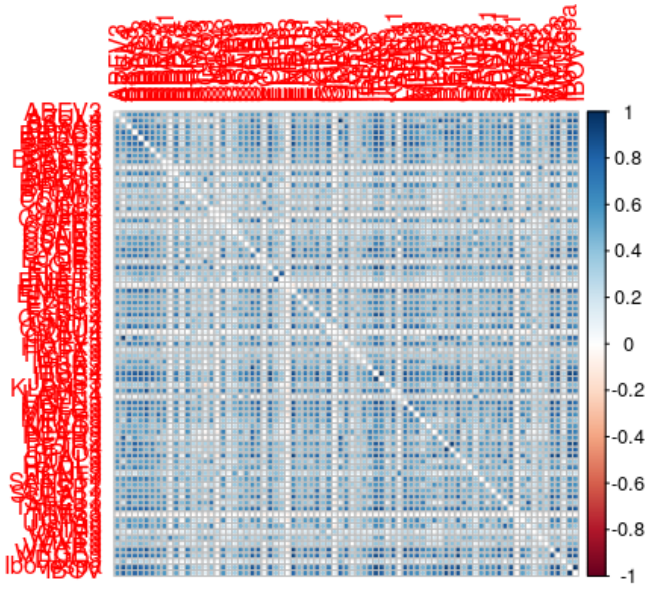

Figura 5.1 – Mapa de correlação do VHP sobre *tickers*.

Fonte: Resultados da pesquisa (2021)

Porém, a correlação é preciso ser confirmada, para isso aplicaram-se testes estatísticos nos dados para verificar a adequabilidade. Na Tabela 5.1 é apresentado o teste de distribuição

Tabela 5.1 – Testes de adequabilidade para análise multivariada

Teste	Estatística	P-valor
Barlett	26089.956	0,000
Assimetria	158220.811	0,000
Curtose	88.967	0,000
KMO	0.943	-
$\bar{\alpha}$ Cronbach	0.965	-

Fonte: Resultado da pesquisa (2021)

Tabela 5.2 – Resultado dos modelo de componentes principais estimado

Estatística	PC1
Variância explicada	0.458

Fonte: Resultado da pesquisa (2021)

normal multivariada, teste de Barlett as estatísticas MO e a média de todos Alpha de Cronbach. Os testes de distribuição normal multivariada foram rejeitados, enquanto a estatística de Barlett não foi rejeitada, além disso a estatística KMO e o Alpha de Cronbach indicam que os dados são adequados para a ACP.

Atendida a adequabilidade dos dados, segue-se com a estimação do modelo. Apesar de a literatura indicar criar componentes explicando no mínimo de 60% da variação dos dados, ou ainda usar testes informais como o do cotovelo, buscando manter o significado econômico da proxy da demanda de informação foi estimado apenas um componente principal. Dessa forma é apresentada a variância explicada pelo único componente estimado, conforme resultado exposto na Tabela .

O único componente tem o poder explicativo de 45% da variância da massa de dados originais dos *tickers*. Conquanto tenha ficado baixo se comparado às recomendações da literatura, demonstra informação relevante e constitui-se na *proxy* da demanda de informação ou *proxy* da atenção do investidor denominado $PC1$. A importância dessa variável é fornecer sentido econômico para os modelos estimados.

5.2 RESULTADOS DESCRITIVOS

Os resultados das estatísticas descritivas são apresentadas na Tabela 5.3. Para tanto, verifica-se que as variáveis apresentam características peculiares, o *score* estimado possui média zero e desvio padrão próximo de 1. Já a média semanal de negociações do Ibovespa é quase 19 milhões, enquanto a média semanal do Ibovespa é de quase 80 mil pontos. A assimetria e curtose das séries indicam que nenhuma das séries segue a distribuição normal, o que é esperado se tratando de séries temporais onde não há estacionaridade em nível.

Tabela 5.3 – Estatísticas descritivas

Variável	Média	Desvio Padrão	Min	Max	Assimetria	Curtose
PC1	0	1	-1	5.17800e+00	2.0	6.1
Volume	19543141	8181915	5454900	7.58423e+07	3.3	17.5
Ibovespa	79765	18116	49051	1.18478e+05	0.3	-0.9

Fonte: Resultado da pesquisa (2021)

Ao analisar o comportamento das séries temporais em nível nas figuras 5.2 a 5.4, observa-se que a demanda de informação tem um comportamento muito parecido com o volume negociado e ligeiramente parecido com o Ibovespa, indicando que é possível relacionar as variáveis entre si. Sobre o comportamento das séries, parece que o Ibovespa apresenta uma tendência linear, enquanto as outras séries o comportamento é um pouco mais estocástico. Também é visível o choque da pandemia em todas as variáveis, correspondente as alterações após a semana 199.

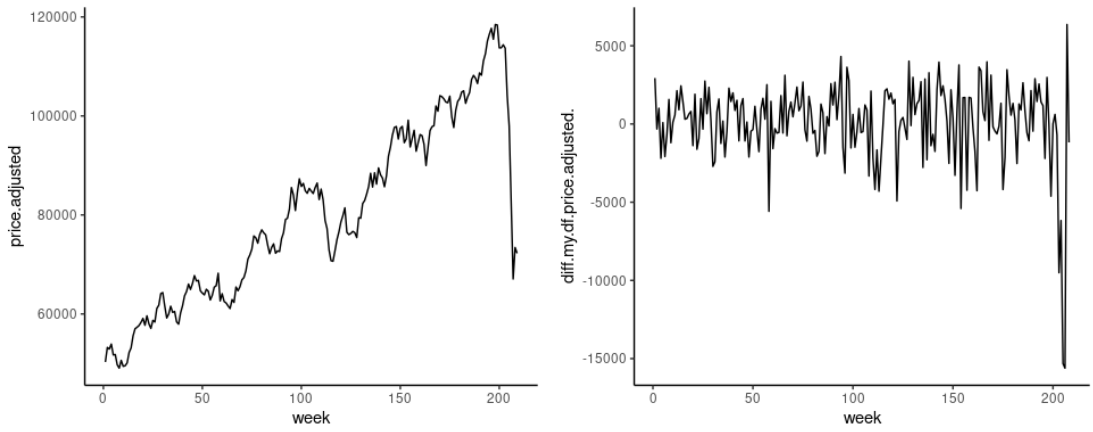

Figura 5.2 – Gráfico do Ibovespa em nível e primeira diferença

Fonte: Resultado da pesquisa (2021)

Quando analisada a primeira diferença nas figuras 5.2 a 5.4. Verifica-se que primeira diferença apresenta o comportamento de ruído branco, mas também é visível o choque do efeito pandemia. Algo que deverá ser controlado por meio da inclusão de *dummies* no modelo aplicado. Isoladamente as séries parecem ter um comportamento parecido, mas para confirmar ainda é preciso avaliar a relação entre as variáveis.

Isoladamente as séries parecem ter um comportamento parecido, mas quando visualizadas em conjunto demonstram outras informações. A relação entre as séries pode ser observada na Figura 5.5 apresentadas em gráficos de dispersão, bem como apresentada a reta de melhor ajuste, estimada por MQO. Novamente a relação entre a demanda de informação e volume de negociações parece apontar para a mesma direção, bem como indica um comportamento linear. Já a relação entre a demanda de informação e Ibovespa sofre com *outliers* e parece não ajustar tão adequadamente como o modelo do volume.

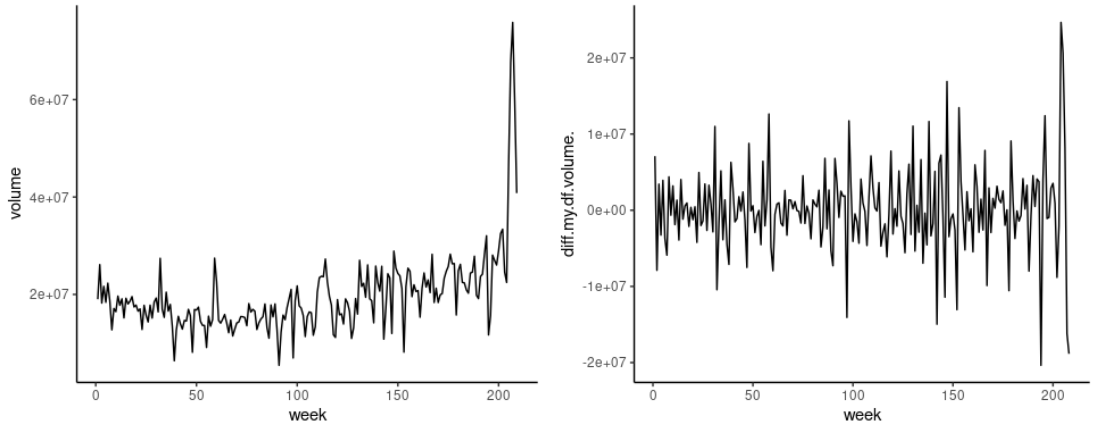

Figura 5.3 – Gráfico do Volume de negociações em nível e primeira diferença

Fonte: Resultado da pesquisa (2021)

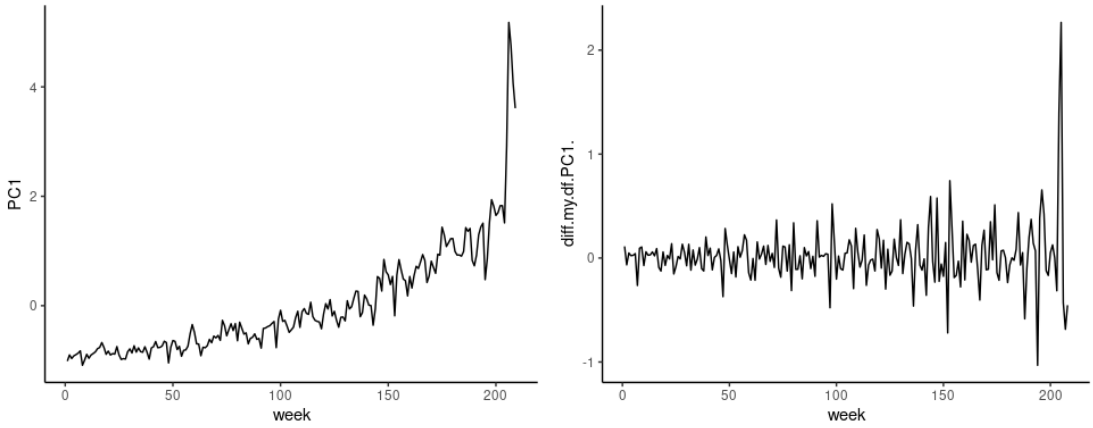

Figura 5.4 – Gráfico da *proxy* atenção do investidor em nível e primeira diferença

Fonte: Resultado da pesquisa (2021)

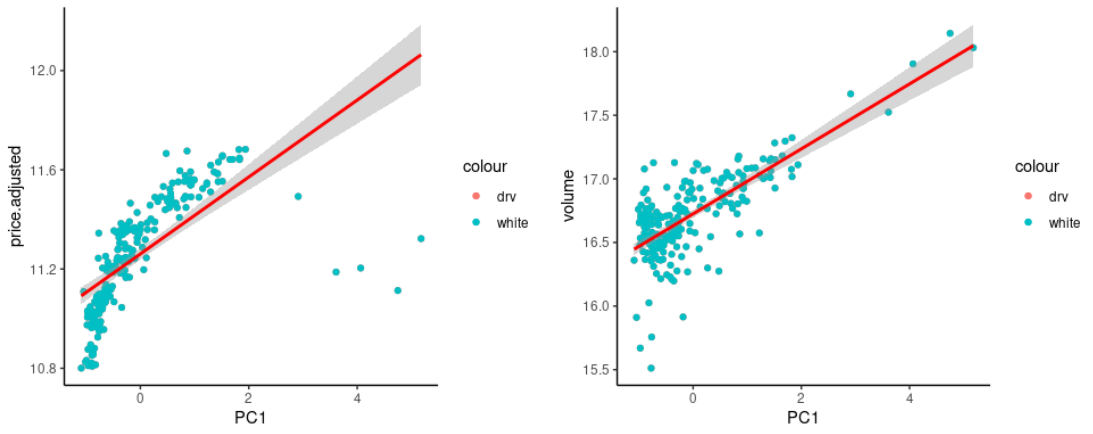

Figura 5.5 – Gráfico de dispersão e reta de melhor ajuste.

Fonte: Resultado da pesquisa (2021)

Por fim, as análises descritiva e gráfica das séries indicam que há uma relação econô-

Tabela 5.4 – Resultado dos testes de raiz unitária KPSS

Variável	Estatística	Valor crítico
PC1	3.373***	0.463
Ibovespa	1.756***	0.463
Volume	3.759***	0.463
PC1 - diferença	0.220	0.463
Ibovespa - diferença	0.057	0.463
Volume - diferença	0.220	0.463

Fonte: Resultado da pesquisa (2021). ***, **, * Indicam respectivamente a rejeição da hipótese nula a 1%, 5% e 10% de significância.

Tabela 5.5 – Critério de informação para número ótimo de retardos

Modelo	AIC(n)	HQ(n)	SC(n)	FPE(n)
Volume	4	1	1	4
Ibovespa	9	2	2	9

Fonte: Resultado da pesquisa (2021)

mica entre a *proxy* da atenção do investidor, volume de negociações e o valor do Ibovespa. A hipótese levantada é que investidores pesquisam no *Google* sobre os *tickers* antes de fazer qualquer negociação, o que justifica a *proxy* possuir um comportamento mais próximo do volume de negociações se comparado com o valor do Ibovespa. Essa relação pode ser modelada, as medidas de qualidade do ajuste são apresentadas a seguir.

5.3 RESULTADOS ECONOMÉTRICOS

5.3.1 Resultados do teste de estacionariedade, de definição das defasagens e de cointegração

Os resultados da estimação do modelo autorregressivo iniciam com a apresentação dos testes de estacionariedade das séries. Como já era esperado, o teste de raiz unitária KPSS aponta que as séries são integradas de primeira ordem I(1), o que favorece estimação por VEC, conforme 5.1.

Na sequência, para se estimar o VEC e realizar o teste de cointegração é preciso saber quantas defagens são necessárias para chegar ao ruído branco. Os resultados, conforme a Tabela 5.5, demonstram que para o modelo que relaciona volume e VHP é necessário apenas 1 retardo pelas estatísticas HQ e SC, mas o software exige no mínimo dois retardos, então opta-se por quatro retardos. Enquanto para o modelo que relaciona o Ibovespa e VHP, são necessários 2 retardos pelas estatísticas Hannan-Quinn (HQ) e Schwarz (SC).

Tabela 5.6 – Teste de cointegração de Johansen

Teste	Ibovespa	Volume	10pct	5pct	1pct
r <= 1	3.671	1.549	7.52	9.24	12.97
r = 0	39.967***	37.834***	13.75	15.67	20.20

Fonte: Resultado da pesquisa (2021). ***, **, * Indicam respectivamente a rejeição da hipótese nula a 1%, 5% e 10% de significância.

O resultado do teste de cointegração pelo autovalor é apresesentado na Tabela 5.6. São dois modelos estimados, o modelo que relaciona volume e VHP com quatro defasagens, e o modelo que relaciona o valor do Ibovespa e o VHP com duas. Para ambos os modelos é rejeitado a hipótese nula de que não há vetor de cointegração, assim há pelo menos um vetor de cointegração para ambos os modelos. Dessa forma é confirmado a hipótese de há uma relação econômica entre a *proxy* da atenção do investidor e volume de negociações e valor do Ibovespa.

A cointegração entre as variáveis estão de acordo com os trabalhos de (HEIBERGER, 2015), (HU et al., 2018) e (AHMED et al., 2017), em que há uma relação entre o preços e o *Google trends*. A partir disso, (CHALLET; AYED, 2014), sugere que o *Google trends* pode ser usado como *proxy* da demanda de informação por ativos. Dessa forma, o VHP sobre ativos ou na forma de *proxy* da demanda de informação pode melhorar modelos de previsão, por captar um efeito específico que geralmente não é considerado pela maioria das pesquisas.

O uso da ACP para gerar a *proxy* da atenção do investidor se mostrou uma abordagem adequada, os resultados reforçam a hipótese de que investidores pesquisam no *Google* sobre os ativos antes de investir, ainda reforça os achado de (VLASTAKIS; MARKELLOS, 2012) em que as pesquisas em buscadores tem relação com a volatilidade uma vez que há relação com o volume de negocições, ou ainda (MOUSSA; DELHOUMI; OUDA, 2017) em que essa relação de volatilidade pode aumentar ou diminuir dependendo do diferente ativo.

Além disso, os resultados evidenciam a relação para o mercado brasileiro, já apontada por (RAMOS; RIBEIRO; PERLIN, 2017) e (PEREIRA; ROSA; FILHO, 2020) os autores encontraram uma relação de Granger Causalidade entre volume, retorno e VHP. Nesse sentido, pode-se considerar que o mercado brasileiro apresenta relação entre o Ibovespa, o volume e a demanda de informação de longo prazo. Logo, entende-se que os investidores brasileiros fazem uso do buscador *Google* antes de tomar uma decisão, ao passo que a ferramenta pode mostrar tendências do mercado financeiro brasileiro.

5.3.2 Resultados das funções impulso resposta e da decomposição da variância

Após a definição da estrutura do modelo, foram estimadas as funções impulso resposta e a decomposição da variância. Para tanto, procedeu-se complementarmente com os testes de diagnóstico dos modelos, conforme Tabela 5.7. Os resultados indicam ausência de autocorre-

Tabela 5.7 – Análise de resíduo modelo VAR

Modelo	Correlação Serial	Normalidade	Efeito ARCH
Ibovespa	64.750	1586.748***	206.381***
Volume	38.797	4332.446***	49.676*

Fonte: Resultado da pesquisa (2021). ***, **, * Indicam respectivamente a rejeição da hipótese nula a 1%, 5% e 10% de significância.

lação, conforme resultado do teste de autocorrelação de Portmentau com ajuste para pequenas amostras, pois não rejeitam a hipótese nula. Em relação a variância constante, evidenciou-se que o teste de ARCH-LM indicou a rejeição da hipótese nula de homocedasticidade para o modelo do preço, considerando 5% de significância, enquanto que não rejeita a pressuposição de igual variância para o modelo do volume ao nível de 5%. E quanto ao teste de normalidade, os resultados indicam que para os dois modelos apresentam distribuição não normal para os resíduos a 1% de significância.

A partir desses resultados, as dinâmicas das Funções de Resposta ao Impulso, analisadas a partir da medida de um desvio padrão, são apresentadas na Figura 5.6. Um observa-se que um choque não antecipado no Ibovespa (*price.adjusted*), reduz a demanda de informação, o que pode indicar que investidores sabendo que as carteiras estão perfomando bem, reduzem a necessidade de acompanhar as informações. Um choque negativo do Ibovespa, desperta a curiosidade dos investidores pelo motivo do Ibovespa estar caindo, assim quando o Ibovespa cai, investidores estudam mais seus ativos.

Já o choque da demanda de informação (PC1) leva a uma pequena alteração no Ibovespa, negativa muito próxima de zero. Enquanto o choque na próxima informação é positiva e se dissipa em poucos passos. Dessa forma, sugere-se que, quando um investidor começa a acompanhar um ativo, o investidor mantém esse comportamento por um período curto de tempo até se sentir seguro, após adquirir essa sensação de segurança a demanda de informação reduz (passa a acompanhar menos) em relação ao preço do ativo. Ou ainda, essa é uma evidência da estratégia *Buy and Hold*, em que algumas vertentes sugerem que preço não é importante, dado que o investidor acompanha o ativo até comprar e depois para de acompanhar a cotação, mas como um choque negativo do Ibovespa elevaria a demanda de informação, essa hipótese pode ser descartada.

Esse resultado ainda parece divergir de (VLASTAKIS; MARKELLOS, 2012), uma vez que choques positivos no retorno aumentam a volatilidade dos ativos, mas reduzem a atenção do investidor. Entretanto, isso pode ser causado por uma segurança irracional, como vista no viés de confirmação, nesse sentido se o mercado está em alta, essa situação reforça a tese de investimento do indivíduo que se sente seguro o suficiente para não acompanhar mais as cotações.

Entretanto, o principal resultado da FRI está de acordo com (DZIELINSKI, 2012) quando enfatizam que, em que tempos de maior de incerteza, maior tende a ser a demanda de informa-

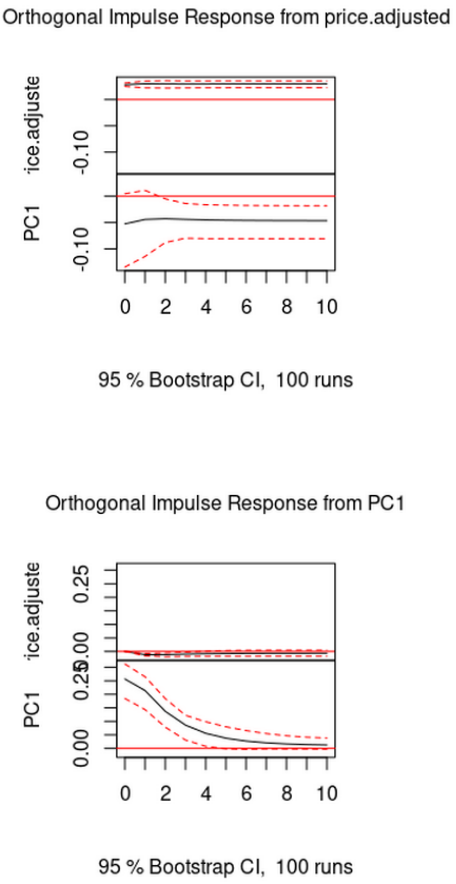

Figura 5.6 – Função Resposta ao Impulso (IRF).

Fonte: Resultado da pesquisa (2021)

ção. Na mesma linha, (HEIBERGER, 2015) indicam que o VHP é um bom indicador de más notícias, sendo que esse efeito pode ser justificado pela economia comportamental, pois segundo (KAHNEMAN, 2011), a dor da perda é maior que o prazer do ganho, como também reforça a hipótese do viés de confirmação. Ou ainda os achados evidenciados por (MOORE; HEALY, 2008) em que o investidor apresenta excesso de confiança, ao atender a demanda de informação o investidor superestima seu conhecimento e deixa e acompanhar novas informações.

Apesar de parecer uma anomalida do mercado, os resultado dirvegem de (CAMPBELL; SHILLER, 1988); (HAUGEN; JORION, 1996) (WOUTERS, 2006), já que não parece ser uma anomalia do mercado, a irracionalidade no comportamento dos investidores não impacta no resultado do mercado, o que pode fazer parte pelo processo evidenciado por (NETO, 2006). Por outro lado, poderia-se pensar em um anomalia do mercado se o choque da atenção do investidor tivesse um impacto maior no valor do Ibovespa, nesse caso a irracionalidade não tem impacto significativo no mercado, conforme as hipóteses de (NETO, 2006).

Complementando, apresentam-se os resultados à decomposição da variância dos dois

Tabela 5.8 – Resultados da decomposição da variância

Períodos	Volume	PC1 relacionado ao volume	Ibovespa	PC1 relacionado ao Ibovespa
1	1.000	0.000	1.000	0.000
2	0.948	0.052	0.919	0.081
3	0.923	0.077	0.897	0.103
4	0.914	0.086	0.897	0.103
5	0.912	0.088	0.901	0.099
6	0.906	0.094	0.907	0.093
7	0.894	0.106	0.911	0.089
8	0.882	0.118	0.915	0.085
9	0.875	0.125	0.919	0.081
10	0.869	0.131	0.922	0.078

Fonte: Resultado da pesquisa (2021)

modelos, os quais se encontram na Tabela 5.8. A decomposição da variância aponta que as variações do volume são relativamente pequenas, sendo que passados 10 períodos (semanas) o PC1 explica aproximadamente 13% da variância do volume de negociações. Enquanto que para o modelo do Ibovespa a demanda de informação explica menos, cerca de 7% decorridos 10 períodos. Ainda que pequeno, o efeito é importante reconhecer que o volume também tem uma grande variância dada a atenção do investidor, o efeito reduzido no volume pode ser um indicativo ainda há poucos investidores, pessoa física, no mercado brasileiro, mas esse investidor tem um potencial impacto no mercado acionário.

Os resultados apontam que a demanda de informação tem um impacto pequeno na variância, mas não pode ser ignorado. Esse resultado está de a cordo com (BIJL et al., 2016) que apontam a dificuldade de generalização de pesquisas com o VHP pois a relação entre o VHP e preços e volume de negociações mudam ao longo do tempo. Também é importante ressaltar que para (PERLIN et al., 2017) o VHP têm um impacto positivo em retornos e volume de negociações, mas o impacto é muito pequeno. Nesse sentido a *proxy* da demanda de informação deve ser usada apenas para melhorar modelos de previsão e tornar ruído branco o resíduo dos modelos que analisem ativos financeiros, pois é um efeito significativo ignorado.

5.4 ALGORITMO DE *TRADING*

O algoritmo de *trading* possibilita avaliar a capacidade de previsão do modelo analítico, isso é se o modelo é capaz de prever com boa acurácia o comportamento do Ibovespa. Os resultados da dinâmica da previsão do Ibovespa e dos valores reais do Ibovespa apontam que o VAR prevê bem o comportamento do Ibovespa conforme a Figura 5.7. Porém para realizar *trading* é importante acertar também quando o ativo vai subir e quando vai cair, nesse sentido, em alguns momentos a previsão parece se atrasar ou adiantar-se. A Figura 5.7 apresenta os

resultados de treino e de teste, as quais são separados pela linha pontilhada. À esquerda da linha pontilhada contempla os dados de treino, enquanto à direita os dados de teste.

Sobre o retorno, o algoritmo em média performa pior que o Ibovespa, superando apenas após a pandemia. Dessa forma a consequência da HME situação a qual não é possível ter retornos maiores que o mercado é mantida para a maior parte do período analisado. Entretanto, para o período após a semana 199 que contempla um período que houve uma reação exagerada do mercado, o algoritmo passou a superar o retorno do mercado, indicando que nesse período a HME não se sustenta. Assim os algoritmos se destacam no período da pandemia, onde há maior incerteza no mercado, período o qual os dados de teste contemplam. Uma vez que a incerteza causada pelo Covid-19 pode ter gerado assimetrias na expectativa dos investidores, o mercado rompeu com a HME justificando o desempenho superior do algoritmo com os dados de teste.

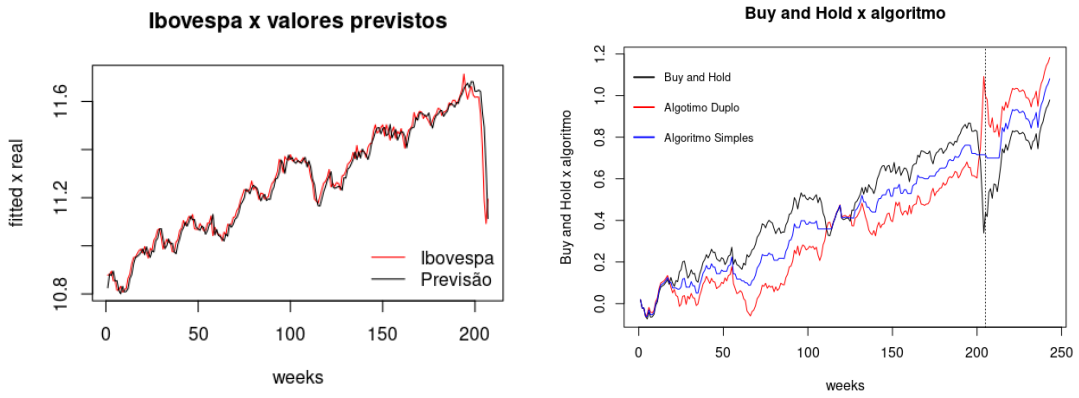

Figura 5.7 – Resultados da previsão e retorno acumulado das estratégias.

Fonte: Resultado da pesquisa (2021)

Com efeito, observando o período de treino em que a HME se sustenta, os resultados diferem de (RAMOS; RIBEIRO; PERLIN, 2017), (PREIS; MOAT; STANLEY, 2013), (PERLIN et al., 2017) e (HEIBERGER, 2015), que desenvolveram algoritmos de *trading* a partir do VHP que superam o retorno do mercado. (PREIS; MOAT; STANLEY, 2013) ainda apontam que os investidores pesquisam mais pelo buscador *Google* informações das empresas quando desejam vender essas ações, se comparado quando desejam comprar ativos, por isso é mais indicado para estratégias com posição a descoberto. Ainda (RAMOS; RIBEIRO; PERLIN, 2017) desenvolveram um algoritmo de *trading* que supera o retorno o Ibovespa incluindo outras variáveis como Selic, taxa CDI e Tesouro direto.

Assim sendo, o resultado obtido indicou que apenas a demanda de informação dos agentes não é o suficiente para superar o retorno de mercado em períodos que a HME se sustenta. Todavia, por conseguir evitar a queda brusca que a pandemia causou, a demanda de informação dos agentes pode ser útil para descobrir situações de pânico e incerteza no mercado.

Por outro lado, o resultado está de acordo com (DZIELINSKI, 2012), (CHALLET; AYED, 2013), (CHALLET; AYED, 2014), (LOUGHLIN; HARNISCH, 2013). Esses traba-

lhos apontam que apenas o VHP ou *proxy* da atenção do investidor não é o suficiente para superar o retorno de mercado. Enquanto, a *proxy* adequada quando há períodos de incerteza, o que alinha-se ao exposto por (HEIBERGER, 2015) quando expôs que o VHP é um bom indicador de más notícias, os resultados apontam que o VHP também é um indicador de incerteza. Por sua vez, como afirma (DZIELINSKI, 2012), o *twitter* é uma melhor *proxy* da atenção do investidor.

(CHALLET; AYED, 2013) ainda apontam que algoritmos de trading baseados no *Google Trends* são muito sensíveis à escolha de termos, tornando essa etapa um processo importante para determinar se o algoritmo irá superar o mercado. Nesse sentido a redução de dimensão também é sensível a esses termos, uma vez que muitas empresas que compõem Ibovespa hoje, não eram tão muito pesquisadas alguns anos atrás, além disso, por optou-se um componente principal, mas uma outra proposta que capte mais a variância dos dados originais pode mostrar resultados diferentes. Ainda (CHALLET; AYED, 2014) afirmam que o *Google Trends* tem um comportamento parecido com o próprio índice com a mesma dificuldade de previsão que o próprio ativo, assim, também é sensível a eventos extremos.

Assim sendo, é possível que o algoritmo tenha apresentado melhor desempenho ao final da amostra por não sofrer os mesmos vieses e heurísticas que investidores podem sofrer, conforme (KAHNEMAN, 2011), (ARIELY; JONES, 2008) e outros. Como já foi discutido, a pandemia pode ser a causa de algumas assimetrias no mercado que afasta o mercado dos pressupostos da HME, apresentando oportunidades para ganhos acima da média por estratégias baseadas em *machine learning*.

Embora seja difícil saber exatamente os motivos de o algoritmo que considera posições a descoberto ter apresentado retornos maiores no período de teste, alguns motivos podem ser expostos. Entre eles está o risco maior da operação vendido, como também a hipótese em que investidores pesquisam mais quando desejam vender seus ativos (PREIS; MOAT; STANLEY, 2013). Além disso, pode resultar da da racionalidade do algoritmo, o qual não sofre os vieses já discutidos, ou ainda um caso de loteria. Entretanto, todas as evidências apontam que algoritmos de *trading* merecem ser estudados, especialmente em períodos de incerteza no mercado, quando os mercado rompe com os pressupostos da HME, surgindo assim oportunidades de retornos acima da média.

6 CONCLUSÃO

O trabalho buscou nas pesquisas de investidores no *Google* o desenvolvimento de uma *proxy* da atenção do investidor (ou da demanda do investidor por informações). A partir disso foi analisada a relação entre a atenção do investidor, o volume de negociações do Ibovespa e o valor do Ibovespa, por meio da aplicação da metodologia VEC-VAR. Além disso, foi desenvolvido um algoritmo de *Trading* com objetivo de testar os retornos do mercado, verificando se ele é capaz de superar ou não o retorno do Ibovespa.

A pesquisa no *Google* faz parte do processo individual de investimento, em que há cointegração entre o VHP e volume de negociações. Essa relação deixa implícito que investidores pesquisam antes de comprar ou vender ativos. Com isso, o VHP do buscador *Google*, conhecido como *Google Trends* é uma boa ferramenta para gerar essa *proxy*. Além disso, uma forma de se gerar essa *proxy* é pela Análise de Componentes Principais.

O principal resultado sobre o comportamento do investidor, situação em que demanda mais informação quando o mercado cai, ou seja, quando há maior incerteza. Nesse sentido, os investidores estudam mais sobre suas carteiras quando estão com retornos negativos, por uma situação de pânico, dado o choque simulado da Função Resposta ao Impulso. Por outro lado, quando o mercado está em alta, o viés de confirmação passa uma sensação de segurança, reduzindo a demanda de informação do investidor, que passa a não acompanhar com a mesma frequência as cotações, característica do excesso de confiança.

Por fim, evidenciou-se que a Hipótese de Mercado Eficiente é preservada até o início dos impactos da pandemia relacionada ao Covid-19, uma vez que o retorno do algoritmo não supera o retorno do mercado para a maior parte do período. Isso implica que o mercado brasileiro é eficiente? Não, necessariamente, uma vez que para um período de próximo de março de 2020 o algoritmo passou a ter um retorno superior ao do mercado. Assim, o algoritmo de *trading* se mostrou adequado no período de elevada incerteza, quando o mercado passou a não atender os pressupostos da HME, mesmo período em que, invariavelmente, a demanda de informação aumenta. Outro achado, é que os investidores acompanham cotação quando estão vendendo seus ativos, o que sugestiona que algoritmos que consideram posições vendido se mostram mais rentáveis, mas também mais arriscados.

Conquanto os resultados tenham demonstrado evidências relevantes, também faz-se importante lembrar que se sacrificou um número maior de componentes para preservar uma interpretação econômica da relação entre a *proxy* da demanda de informação, o volume de negociações e o Ibovespa, o que se constitui em uma limitação da análise proposta. Desta forma, futuras pesquisas abordar formaas que captem mais a variação dos dados e desenvolver previsões e algoritmos de *Trading* que superem o retorno do mercado, também é interessante avaliar em horizontes de tempo, como em diferentes frequências temporais.

REFERÊNCIAS BIBLIOGRÁFICAS

AHMED, F. et al. Financial market prediction using google trends. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 7, p. 388–391, 2017.

ANDERSON, T. W. **An introduction to multivariate statistical analysis**. [S.l.], 1962.

ARIELY, D.; JONES, S. **Predictably irrational**. [S.l.]: Harper Audio New York, NY, 2008.

B3. 2020. <http://www.b3.com.br/pt_br/>. Accessed: 2020-10-30.

B3. 2020. <http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/historico-pessoas-fisicas/>. Accessed: 2020-10-30.

BACHELIER, L. Théorie de la spéculation. **Annales scientifiques de l'École Normale Supérieure**, Elsevier, v. 3e série, 17, p. 21–86, 1900. Disponível em: <www.numdam.org/item/ASENS_1900_3_17__21_0/>.

BECKER, Y. L.; REINGANUM, M. R. **The Current State of Quantitative Equity Investing**. [S.l.]: CFA Institute Research Foundation, 2018.

BERA, A. K.; JARQUE, C. M. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte carlo evidence. **Economics letters**, Elsevier, v. 7, n. 4, p. 313–318, 1981.

BEUNZA, D.; STARK, D. From dissonance to resonance: Cognitive interdependence in quantitative finance. **Economy and Society**, Taylor & Francis, v. 41, n. 3, p. 383–417, 2012.

____. Seeing through the eyes of others: dissonance within and across trading rooms. **C. Knorr, Karin, & A. Preda (Eds.), Oxford handbook of the sociology of finance. Oxford handbooks in business and management**, p. 203–222, 2012.

BIJL, L. et al. Google searches and stock returns. **International Review of Financial Analysis**, Elsevier, v. 45, p. 150–156, 2016.

BLACK, F.; SCHOLES, M. The pricing of options and corporate liabilities. **Journal of political economy**, The University of Chicago Press, v. 81, n. 3, p. 637–654, 1973.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of computational science**, Elsevier, v. 2, n. 1, p. 1–8, 2011.

BOWER, B. Gambling on experience: Perceptions of risk can get pulled in opposite directions. **Science News**, Wiley Online Library, v. 177, n. 9, p. 26–29, 2010.

BRUGGEMANN, R.; LUTKEPOHL, H.; SAIKKONEN, P. Residual autocorrelation testing for vector error correction models. **Journal of Econometrics**, Elsevier, v. 134, n. 2, p. 579–604, 2006.

BUEHLER, R.; GRIFFIN, D.; ROSS, M. Exploring the"planning fallacy": Why people underestimate their task completion times. **Journal of personality and social psychology**, American Psychological Association, v. 67, n. 3, p. 366, 1994.

BUENO, R. d. L. da S. **Econometria de séries temporais**. [S.l.]: Cengage Learning, 2008.

CAMPBELL, J. Y.; SHILLER, R. J. Stock prices, earnings, and expected dividends. **The Journal of Finance**, Wiley Online Library, v. 43, n. 3, p. 661–676, 1988.

CHALLET, D.; AYED, A. B. H. Predicting financial markets with google trends and not so random keywords. **arXiv preprint arXiv:1307.4643**, 2013.

____. Do google trend data contain more predictability than price returns? **arXiv preprint arXiv:1403.1715**, 2014.

CHAN, E. **Quantitative trading: how to build your own algorithmic trading business**. [S.l.]: John Wiley & Sons, 2009. v. 430.

CHOI, H.; VARIAN, H. Predicting initial claims for unemployment benefits. **Google Inc**, Citeseer, v. 1, p. 1–5, 2009.

CONLAN, C. **Automated Trading with R**. [S.l.]: Springer, 2016.

CORRAR, L. J.; FILHO, J. M. D.; PAULO, E. **Análise multivariada para os cursos de administração, ciências contábeis e economia**. [S.l.]: Editora Atlas, 2009.

COSTA, I. J.; VARGAS, J. Análise fundamentalista e análise técnica: agregando valor a uma carteira de ações. **Destarte**, v. 1, n. 1, p. 9–25, 2020.

CVM. 2020. <https://www.investidor.gov.br/publicacao/ListaCVMComportamental.html>. Accessed: 2020-10-30.

CVM. 2020. <https://www.gov.br/cvm/pt-br>. Accessed: 2020-10-30.

DUNIS, C. L.; LAWS, J.; NAIM, P. **Applied quantitative methods for trading and investment**. [S.l.]: John Wiley & Sons, 2004.

DZIELINSKI, M. Measuring economic uncertainty and its impact on the stock market. **Finance Research Letters**, Elsevier, v. 9, n. 3, p. 167–175, 2012.

EDGERTON, D.; SHUKUR, G. Testing autocorrelation in a system perspective testing autocorrelation. **Econometric Reviews**, Taylor & Francis, v. 18, n. 4, p. 343–386, 1999.

EMERSON, S. et al. Trends and applications of machine learning in quantitative finance. In: **8th International Conference on Economics and Finance Research (ICEFR 2019)**. [S.l.: s.n.], 2019.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 987–1007, 1982.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **The journal of Finance**, JSTOR, v. 25, n. 2, p. 383–417, 1970.

____. Efficient capital markets: Ii. **The journal of finance**, Wiley Online Library, v. 46, n. 5, p. 1575–1617, 1991.

FINUCANE, M. L. et al. Aging and decision-making competence: An analysis of comprehension and consistency skills in older versus younger adults considering health-plan options. **Journal of Behavioral Decision Making**, Wiley Online Library, v. 15, n. 2, p. 141–164, 2002.

FRENCH, C. W. The treynor capital asset pricing model. **Journal of Investment Management**, California, v. 1, n. 2, p. 60–72, 2003.

H, R. Statistical versus clinical predictionof the stock market. **American Finance Associaton**, mar. 1959.

HAIR, J. F. et al. **Análise multivariada de dados**. [S.l.]: Bookman editora, 2009.

HAMILTON, J. D.; SUSMEL, R. Autoregressive conditional heteroskedasticity and changes in regime. **Journal of econometrics**, North-Holland, v. 64, n. 1-2, p. 307–333, 1994.

HAUGEN, R. A.; HAUGEN, R. A. **Modern investment theory**. [S.l.]: Prentice Hall Upper Saddle River, NJ, 2001. v. 5.

HAUGEN, R. A.; JORION, P. The january effect: Still there after all these years. **Financial Analysts Journal**, Taylor & Francis, v. 52, n. 1, p. 27–31, 1996.

HEIBERGER, R. H. Collective attention and stock prices: evidence from google trends data on standard and poor's 100. **PloS one**, Public Library of Science, v. 10, n. 8, p. e0135311, 2015.

HU, H. et al. Predicting the direction of stock markets using optimized neural networks with google trends. **Neurocomputing**, Elsevier, v. 285, p. 188–195, 2018.

JAEGER, R. A. **All about hedge funds**. [S.l.]: McGraw-Hill New York, 2003.

JARQUE, C. M.; BERA, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. **Economics letters**, North-Holland, v. 6, n. 3, p. 255–259, 1980.

____. A test for normality of observations and regression residuals. **International Statistical Review/Revue Internationale de Statistique**, JSTOR, p. 163–172, 1987.

JOHANSEN, S. Identifying restrictions of linear equations with applications to simultaneous equations and cointegration. **Journal of econometrics**, Elsevier, v. 69, n. 1, p. 111–132, 1995.

____. A statistical analysis of cointegration for i (2) variables. **Econometric Theory**, JSTOR, p. 25–59, 1995.

KAHN, R. N. **The Future of Investment Management**. [S.l.]: CFA Institute Research Foundation, 2018.

KAHNEMAN, D. Maps of bounded rationality: Psychology for behavioral economics. **American economic review**, v. 93, n. 5, p. 1449–1475, 2003.

____. **Thinking, fast and slow**. [S.l.]: Macmillan, 2011.

KAHNEMAN, D.; TVERSKY, A. Subjective probability: A judgment of representativeness. **Cognitive psychology**, Elsevier, v. 3, n. 3, p. 430–454, 1972.

____. On the interpretation of intuitive probability: A reply to jonathan cohen. Elsevier Science, 1979.

____. Prospect theory: An analysis of decision under risk. In: **Handbook of the fundamentals of financial decision making: Part I**. [S.l.]: World Scientific, 2013. p. 99–127.

KWIATKOWSKI, D. et al. Testing the null hypothesis of stationarity against the alternative of a unit root. **Journal of econometrics**, v. 54, n. 1-3, p. 159–178, 1992.

LI, Y.; WU, J.; BU, H. When quantitative trading meets machine learning: A pilot survey. In: IEEE. **2016 13th International Conference on Service Systems and Service Management (ICSSSM)**. [S.l.], 2016. p. 1–6.

LIMA, L. A. D. O. Auge e declínio da hipótese dos mercados eficientes. **Brazilian Journal of Political Economy**, SciELO Brasil, v. 23, n. 4, p. 531–546, 2003.

LOUGHLIN, C.; HARNISCH, E. The viability of stocktwits and google trends to predict the stock market. **Unpublished Research**, 2013.

LUTKEPOHL, H. Structural vector autoregressive analysis for cointegrated variables. **Allgemeines Statistisches Archiv**, Springer, v. 90, n. 1, p. 75–88, 2006.

MARKOWITZ, H. The utility of wealth. **Journal of political Economy**, The University of Chicago Press, v. 60, n. 2, p. 151–158, 1952.

MASSICOTTE, P.; EDDELBUETTEL, D. **gtrendsR: Perform and Display Google Trends Queries**. [S.l.], 2020. R package version 1.4.7. Disponível em: <https://CRAN.R-project.org/package=gtrendsR>.

MOORE, D. A.; HEALY, P. J. The trouble with overconfidence. **Psychological review**, American Psychological Association, v. 115, n. 2, p. 502, 2008.

MORETTIN, P. A. **Econometria financeira: um curso em séries temporais financeiras**. [S.l.]: Editora Blucher, 2017.

MOSSIN, J. Equilibrium in a capital asset market. **Econometrica: Journal of the econometric society**, JSTOR, p. 768–783, 1966.

MOUSSA, F.; DELHOUMI, E.; OUDA, O. B. Stock return and volatility reactions to information demand and supply. **Research in International Business and Finance**, Elsevier, v. 39, p. 54–67, 2017.

MUNIZ, C. J. Testes preliminares de eficiência do mercado de ações brasileiro. **Revista Brasileira do Mercado de Capitais**, v. 6, n. 16, 1980.

NASEER, M.; TARIQ, D. B. et al. The efficient market hypothesis: A critical review of the literature. **The IUP Journal of Financial Risk Management**, v. 12, n. 4, p. 48–63, 2015.

NETO, J. Wanderley da F. **A hipótese de eficiência de mercado e as finanças comportamentais: evidências empíricas no mercado acionário brasileiro e uma proposta teórica integrativa**. 2006.

NICKERSON, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. **Review of general psychology**, SAGE Publications Sage CA: Los Angeles, CA, v. 2, n. 2, p. 175–220, 1998.

PEREIRA, M. M. de; ROSA, T. G. da; FILHO, R. B. Influência do google trends em ações listadas na bolsa de valores brasileira: evidências a partir da modelagem pvar. **Revista Eletrônica de Administração**, v. 26, n. 3, p. 796–818, 2020.

PERLIN, M. **BatchGetSymbols: Downloads and Organizes Financial Data for Multiple Tickers**. [S.l.], 2020. R package version 2.6.1. Disponível em: <https://CRAN.R-project.org/package=BatchGetSymbols>.

PERLIN, M. S. et al. Can we predict the financial markets based on google's search queries? **Journal of Forecasting**, Wiley Online Library, v. 36, n. 4, p. 454–467, 2017.

PFAFF, B. et al. Var, svar and svec models: Implementation within r package vars. **Journal of Statistical Software**, v. 27, n. 4, p. 1–32, 2008.

PREIS, T.; MOAT, H. S.; STANLEY, H. E. Quantifying trading behavior in financial markets using google trends. **Scientific reports**, Nature Publishing Group, v. 3, p. 1684, 2013.

PYO, S. et al. Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the korean stock markets. **PloS one**, Public Library of Science San Francisco, CA USA, v. 12, n. 11, p. e0188107, 2017.

RAMOS, H. P.; RIBEIRO, K. K. M.; PERLIN, M. S. The forecasting power of internet search queries in the brazilian financial market. **RAM. Revista de Administração Mackenzie**, SciELO Brasil, v. 18, n. 2, p. 184–210, 2017.

REVELLE, W. R. psych: Procedures for personality and psychological research. 2017.

RONEY, C. J.; TRICK, L. M. Sympathetic magic and perceptions of randomness: The hot hand versus the gambler's fallacy. **Thinking & reasoning**, Taylor & Francis, v. 15, n. 2, p. 197–210, 2009.

SAFFI, P. A. Análise técnica: sorte ou realidade? **Revista Brasileira de Economia**, SciELO Brasil, v. 57, n. 4, p. 953–974, 2003.

SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. **The journal of finance**, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964.

SHILLER, R. J. et al. Alternative tests of rational expectations models: The case of the term structure. **Journal of Econometrics**, Elsevier, v. 16, n. 1, p. 71–87, 1981.

SILVA, C. A. T.; FELIPE, E. da S. Avaliação da influência de textos narrativos de fatos relevantes no preço das ações de empresas brasileiras. **Revista Contabilidade e Controladoria**, v. 2, n. 2, 2010.

TERRA, P. R. S.; LIMA, J. B. N. d. Governança corporativa e a reações do mercado de capitais à divulgação das informações contábeis. **Revista Contabilidade & Finanças**, SciELO Brasil, v. 17, n. 42, p. 35–49, 2006.

TVERSKY, A.; KAHNEMAN, D. Judgment under uncertainty: Heuristics and biases. **science**, American association for the advancement of science, v. 185, n. 4157, p. 1124–1131, 1974.

____. **Judgments of and by representativeness**. [S.l.], 1981.

____. Advances in prospect theory: Cumulative representation of uncertainty. **Journal of Risk and uncertainty**, Springer, v. 5, n. 4, p. 297–323, 1992.

VARIAN, H. R. Big data: New tricks for econometrics. **Journal of Economic Perspectives**, v. 28, n. 2, p. 3–28, 2014.

VLASTAKIS, N.; MARKELLOS, R. N. Information demand and stock market volatility. **Journal of Banking & Finance**, Elsevier, v. 36, n. 6, p. 1808–1821, 2012.

WOUTERS, T. **Style investing: behavioral explanations of stock market anomalies**. [S.l.]: University of Groningen, 2006.

ZAMBRANO, C.; LIMA, J. Análise estatística multivariada de dados socioeconômicos. **Métodos quantitativos em economia. Viçosa: UFV**, p. 556–577, 2004.

APÊNDICE A - PACOTES NECESSÁRIOS

```
## pacotes necessários ---------
library(lmtest)
library(urca)
library(dplyr)
library(forecast)
library(vars)
library(tsDyn)
library(ggplot2)
library(psych)
library (readr)
```

APÊNDICE B ANÁLISE DE COMPONENTES PRINCIPAIS

```
    #importar dados para PCA----------------
my.url <- "https://raw.githubusercontent.com/517127/Monografia/master/dados/dados_b
my.df <- read_csv(url(my.url)) #importar dados do github
det(cor(my.df)) > 0 # condição necessária
KMO <- KMO(my.df) # teste KMO
KMO$MSA # # teste KMO
bart <- bartlett.test(my.df) #teste bartlett
bart$statistic # estatística de barlett
bart$p.value # p valor associado
cortest <- cortest.bartlett(my.df) # teste de hiptoese
cortest$chisq # estatistica
cortest$p.value # p valor
alpha.c <- alpha(my.df,
check.keys = TRUE) #O Alfa de Cronbach
alpha.c$alpha.drop # alpha de cronbach único
mean(alpha.c$alpha.drop[,1]) # média da série
corrplot::corrplot(corr = cor(my.df),
xlab = "", ylab = "") #correlação
```

```
scree(my.df) # recomendação para fatores e componentes
fa.parallel(my.df) # sugestão de componentes e fatores
my.fac <- pca(my.df,
nfactors = 1, rotate = "varimax") # componentes principais
my.scores <- my.fac$scores # estimativas dos scores
```

APÊNDICE C ANÁLISE DE SÉRIES TEMPORAIS

```
## series temporais -----------
my.url <- "https://raw.githubusercontent.com/517127/Monografia/master/dados/dados_f
my.df <- read_csv(url(my.url)) #importar dados do github
my.model.volume <- my.df[,c(4, 1)]
my.model.volume$volume <- log(my.model.volume$volume)
my.model.price <- my.df[,c(9, 1)]
my.model.price$price.adjusted <- log(my.model.price$price.adjusted)
# seleção ótima de defasagens----------
VARselect(my.model.volume, type = "cons",
          exogen = my.df$my.dummy) #  4
VARselect(my.model.price, type = "trend",
          exogen = my.df$my.dummy) # 2
# cointegração---------
coint.volume <- ca.jo(my.model.volume,
type = "eigen",
ecdet = "cons",
K = 4,
dumvar = my.df$my.dummy
) # 2 vetores
summary(coint.volume)
coint.price  <- ca.jo(my.model.price,
type = "eigen",
ecdet = "trend",
K = 2,
dumvar = my.df$my.dummy
)
summary(coint.price)
#vec2var ------------
vecvar.volume <- vec2var(coint.volume, r = 1) # transformaçao de vec em var
vecvar.price <- vec2var(coint.price, r = 1) #t transformação de vec em var
```

```
# residuo --------------
serial.test(vecvar.volume, #autocorrelação
lags.pt = 15, type  = "PT.adjusted")
serial.test(vecvar.price, #autocorrelação
lags.pt = 15, type = "PT.adjusted")
normality.test(vecvar.volume) #jarquebera
normality.test(vecvar.price) #jarquebera
arch.test(vecvar.volume, #heterocedasticidade
lags.multi = 4)
arch.test(vecvar.price,#heterocedasticidade
lags.multi = 2)
# decomposição da variância ------------
fevd(vecvar.volume)  #decomposição volume
fevd(vecvar.price) #decomposição ibovespa
# irf ------------
impulse_price <- irf(vecvar.price,
impulse = "price.adjusted",
ortho = TRUE) # impulso resposta
impulse_PC1 <- irf(vecvar.price,
impulse = "PC1") # impulso resposta
plot(impulse_PC1)
plot(impulse_price)
```

APÊNDICE D ALGORITMO DE TRADING

```
#trading sistem -----------------
fit.var.price <- fitted(vecvar.price) # valores previstos dentro da amostra
var.fit <- length(fit.var.price[,1]) # tamanho
real.price <- lag(my.model.price$price.adjusted,2) %>%  na.omit # remover NA
plot(fit.var.price[,1], # grafico
     type = "l",
     col = "red",
     #lwd = 3, lty = 3,
     main = "Ibovespa x valores previstos",
     xlab = "weeks",
     ylab = "fitted x real")
lines(real.price)
legend(
```

```
  "bottomright",
  legend = c("Ibovespa", "Previsão"),
  col = c("red", "black"),
  lty = 1.1,
  lwd = 1.1,
  bty = "n"
)
price.prev <- fit.var.price[,1] # valor previsto
l.price.prev <- lag(price.prev,1) # valor previsto com 1 lag
# regra do algoritmo de trading  compra se o valor previsto do futuro é maior que o
regra <- ifelse(price.prev > l.price.prev, 1, 0)
ret.var <- regra[2:207] * my.df$ret.adjusted.prices[4:209] # calculo do retorno
ret.var <- na.omit(ret.var)  # remover NA
ret.real <- my.df$ret.adjusted.prices[4:209] # retorno real
ret.real <- na.omit(ret.real) #remover NA
plot(cumsum(ret.real), # grafico do retorno acumulado
     type = "l",
     col = "red",
     #lwd = 3, lty = 3,
     main = "Buy and Hold x algoritmo",
     xlab = "weeks",
     ylab = "Buy and Hold x algoritmo"
)
lines(cumsum(ret.var))
legend(
  "bottomright",
  legend = c("Buy and Hold", "Algoritmo"),
  col = c("red", "black"),
  lty = 1.1,
  lwd = 1.1,
  bty = "n"
)
## trading system dados treino e teste
library(urca)
library(vars)
library(BatchGetSymbols)
firstdate <- "2020-03-30"
lastdate <- "2020-12-18"
frq <- "weekly"
```

```
ibov <- BatchGetSymbols(tickers = "^BVSP",
                        first.date = firstdate,
                        last.date = lastdate,
                        freq.data = frq,
                        do.cache = FALSE)
saveRDS(ibov$df.tickers, file = "dados_teste.RDS")
n_linhas <- nrow(ibov$df.tickers)-1

coint.price  <- ca.jo(my.model.price,
                      type = "eigen",
                      ecdet = "trend",
                      K = 2,
                      dumvar = my.df$my.dummy
)
vecvar.price <- vec2var(coint.price, r = 1)


#DADOS DE TREINO ----------------

fit.var.price <- fitted(vecvar.price)
price.prev <- fit.var.price[,1]
ret.real <- my.df$ret.adjusted.prices[4:209]


# DADOS DE TESTE--------------------------------
exo1 <- data.frame(rep(0,n_linhas)) # não sei sobre o futuro da pandemia
names(exo1) <- "exo1"
pred.var <- predict(vecvar.price,
                    dumvar =  exo1,
                    n.ahead = n_linhas)
prev.ibov_t <- pred.var$fcst$price.adjusted[,1]
real.ret_2 <- ibov$df.tickers$ret.adjusted.prices[2:38]


prev_pred <- c(price.prev,prev.ibov_t)

ret_ibov <- c(ret.real, real.ret_2)

#algoritmo comprado
```

```
l.prev_pred <- lag(prev_pred,1)
regra <- ifelse(prev_pred > l.prev_pred, 1, 0)
ret.var <- regra[2:length(regra)] *  ret_ibov

## comprado e vendido ------------------------
regra <- ifelse(prev_pred > l.prev_pred, 1, -1)
ret.var_cv <- regra[2:length(regra)] *  ret_ibov

plot(cumsum(ret.var_cv),
     type = "l",
     col = "red",
     #lwd = 3, lty = 3,
     main = "Buy and Hold x algoritmo",
     xlab = "weeks",
     ylab = "Buy and Hold x algoritmo"
)
lines(cumsum(ret_ibov))
lines(cumsum(ret.var), col = "blue")
abline(v = 205, lty = 3)
legend(  "topleft",
  legend = c("Buy and Hold", "Algotimo Duplo",
             "Algoritmo Simples"),
  col = c("black", "red", "blue"),
  lty = 1,
  lwd = 2,
  cex = 0.8,
  bty = "n"
)
```