UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E TRANSPORTES

TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO

PAIRS TRADING: UTILIZAÇÃO DE MACHINE LEARNING PARA SELEÇÃO DE PARES DE ATIVOS FINANCEIROS

MAIKE RONALD MOTA

Orientador: Michel J. Anzanello

PORTO ALEGRE
ABRIL/2023

PAIRS TRADING: Utilização de Machine Learning para seleção de pares de ativos financeiros

MAIKE R. MOTA[1]

MICHEL J. ANZANELLO[2]

Resumo

Este artigo tem como objetivo avaliar a eficácia do modelo proposto por Sarmento e Horta (2020) para solucionar o problema de alta dimensionalidade dos dados e a clusterização de pares de ativos financeiros, e desenvolver um código para implementar e testar a estratégia. Os resultados mostram que o modelo proposto apresenta um bom desempenho em termos de rentabilidade e é uma alternativa promissora para investidores em busca de estratégias de Pairs Trading. No entanto, é necessário avaliar constantemente os riscos e desafios envolvidos na implementação desse modelo e buscar constantemente melhorias e adaptações para garantir sua eficácia em longo prazo.

Palavras-chave: Pairs Trading; Machine Learning; PCA; Cluster.

1 INTRODUÇÃO

Algoritmos e sistemas de Machine Learning (ML) vêm sendo consistentemente utilizados nos mais diversos setores e cenários com vistas a dar suporte à decisão humana. Neste contexto, Pontes (2011) levanta a possibilidade de sistemas de ML superarem humanos em decisões de investimentos. Neste segmento, especialistas examinam detalhadamente os valores de diversas variáveis antes de tomar a decisão de compra ou venda de uma ação. Tal tarefa consome muito tempo, e por vezes, a decisão pode ser errônea ou vir com atraso, levando à perda de uma boa oportunidade de compra ou venda de uma ação.

Segundo Salles (1991), os preços das ações são influenciados por vários tipos de informação (preços passados, lucros futuros, volatilidade, índices econômico-financeiros da análise fundamentalista, variáveis econômicas, fatores políticos etc.), que provocam alterações

[1] Departamento de Engenharia de Produção e Transportes (DEPROT), Universidade Federal do Rio Grande do Sul (UFRGS)
[2]Departamento de Engenharia de Produção e Transportes (DEPROT), Universidade Federal do Rio Grande do Sul (UFRGS)

variadas dependendo do contexto do mercado, da relevância da informação e do momento que essa informação leva para ser incorporada pelo mercado. No mesmo ano, Fama (1991) assume que tais flutuações dos preços são justas, uma vez que a informação deve estar igualmente distribuída entre todos os agentes do mercado. Também considera que o movimento futuro dos preços não será afetado pela informação passada, fenômeno denominado como Hipótese do Mercado Eficiente (HME). A HME tem sido testada extensivamente em vários mercados e, como demonstrado por Taylor (2008), se a HME fosse verdadeira, qualquer tentativa de prever o mercado seria frustrada.

A percepção de que o mercado não é eficiente resultou em três escolas de pensamento sobre a previsão dos preços do mercado de ações: análise fundamentalista, análise técnica e análise quantitativa. A análise fundamentalista baseia-se na análise dos resultados setoriais e específicos de cada empresa, dentro do contexto da economia nacional e internacional (FORTUNA, 2015). A escola técnica, também definida por Fortuna (2015), tem como premissa o fato de que não há necessidade em fazer uma pesquisa dos fundamentos da empresa, pois o gráfico é a soma de todos os conhecimentos, esperanças e expectativas sobre uma determinada ação. Ele reflete o preço que o mercado está, naquele momento, disposto a pagar pela ação e, através de suas técnicas, indica a tendência futura. Já a análise quantitativa utiliza-se de modelos matemáticos e a implementação de modelos envolvendo inteligência artificial. As estratégias podem ser divididas em três grandes tipos: as operações seguidoras de tendência, arbitragem estatística e as operações de alta frequência (PONTES, 2011).

Dentre as divisões da análise quantitativa, o foco deste artigo será o *Pairs trading* que, segundo Avellaneda e Lee (2008), é o antecessor da arbitragem estatística. *Pairs trading* com ações buscam identificar os pares com base em correlação e outras regras de decisão não-paramétricas (CALDEIRA 2010). Algumas questões típicas devem analisadas no desenvolvimento de estratégias *pairs trading*, as quais incluem: (i) como identificar os pares, (ii) quando o portfólio combinado se distancia suficientemente da relação de equilíbrio para abrir uma posição de *pairs trading*, e (iii) quando a posição deve ser encerrada. Neste artigo, a seleção dos pares de ações para *pairs trading* é feita com base na presença de relação de cointegração entre duas séries de retornos de ações. A seleção de pares compreende duas etapas: (i) encontrar os pares de ativos candidatos apropriados e (ii) selecionar os mais promissores. A partir de (i), o investidor deve selecionar os ativos de interesse (por exemplo, ações, futuros, ETFs etc.) e a partir daí começar a procurar possíveis combinações de pares. Na literatura, normalmente são sugeridas duas abordagens para esta etapa: realizar uma busca exaustiva por

todas as combinações possíveis dos títulos selecionados, ou organizá-los em grupos, geralmente por setor, e esticar as combinações para pares formados por títulos dentro do mesmo grupo, fazendo uso de técnicas de ML.

Nos últimos anos, mais pesquisas se tornaram disponíveis sobre a aplicação de ML na área de finanças (CAVALCANTE et al., 2016). Com o aumento do poder computacional, ficou mais fácil treinar Redes Neurais Profundas complexas capazes de gerar resultados promissores. No entanto, a aplicação de Machine Learning em *Pairs Trading* ainda é escassa. Dunis, Laws e Evans (2006), Dunis, Laws e Evans (2009) e Dunis et al. (2015) são os principais autores que aplicam ML ao *País Trading*. Sarmento e Horta (2020) sugerem que, à medida que a popularidade do *Pairs Trading* cresce, torna-se cada vez mais difícil encontrar pares que consigam gerar bons resultados, corroborando as definições de HME descritas por Fama (1991). A escassez desses pares promissores força a expansão da busca para grupos mais amplos de títulos, na expectativa de que, ao considerar um grupo maior, a probabilidade de encontrar um bom par aumente.

A alta dimensionalidade dos dados envolvida na análise, no entanto, conduz a dois problemas: (i) na presença de mais ativos, a probabilidade de encontrar características irrelevantes aumenta, (ii) o problema da maldição da dimensionalidade, termo introduzido por Bellman (1966). A capacidade de análise humana acaba limitando a tomada de decisão com a expansão do grupo de ativos dado o enorme volume de dados e informações a serem processadas.

Este artigo tem como objetivo principal avaliar a eficácia do modelo de Pairs Trading baseado em Aprendizado de Máquina (ML) descrito por Sarmento e Horta (2020) para solucionar o problema de alta dimensionalidade dos dados e a clusterização de pares. A metodologia proposta contempla as seguintes etapas: (i) redução de dimensionalidade para encontrar uma representação compacta para cada ação; (ii) aprendizado não-supervisionado para definir potenciais clusters; (iii) seleção e definição de um conjunto de regras para selecionar pares para negociação. Além disso, será desenvolvido um código para implementar e testar o modelo proposto. A motivação deste estudo está na necessidade de se obter uma representação mais precisa dos dados, sem a necessidade de definir manualmente os grupos aos quais cada ação deve pertencer, através de técnicas de agrupamento automatizadas. A avaliação será feita comparando-o com o índice S&P 500 através de backtesting da estratégia, verificando se o modelo se mostra consistente ao longo do tempo, além de fornecer uma análise crítica do

modelo de Sarmento e Horta (2020) e contribuir para o campo de estudos de Pairs Trading baseado em Aprendizado de Máquina.

2 REFERENCIAL TEÓRICO

Ao longo da revisão da literatura acadêmica, a ênfase foi direcionada a temas relevantes ao escopo do trabalho, tais como *pairs trading* conhecida como uma estratégia quantitativa no mercado de capitais, assim como as recentes abordagens utilizando ML.

2.1.1 HIPÓTESE DE MERCADO EFICIENTE

A hipótese de mercado eficiente é uma teoria financeira que foi desenvolvida pela primeira vez por Eugene Fama (1970). Ela afirma que os preços dos ativos refletem toda a informação disponível no mercado, o que significa que é impossível obter lucros consistentes e acima da média através da análise fundamental ou técnica.

No entanto, alguns estudos, como o de Lo e MacKinlay (1990) e Lo e Wang (1993) apresentam evidências de ineficiência no mercado, devido às fontes de ineficiência como informação assimétrica, comportamento irracional dos investidores, barreiras de entrada e fragmentação do mercado, intervenção governamental e fatores externos. Um exemplo é o artigo de Gatev, Goetzmann e Rouwenhorst (2006) intitulado "Pairs trading: Performance of a relative value arbitrage rule" que mostra resultados de backtesting de uma estratégia de pairs trading baseada em diferenças de preços entre ativos co-integrados, e mostra que é possível obter lucros acima da média através da análise de dados e técnicas estatísticas. Outro exemplo é o artigo de Aït-Sahalia and Lo (2002) intitulado "Nonparametric Pricing of Contingent Claims" que também mostra resultados positivos na aplicação de Pairs Trading.

A hipótese de mercado eficiente é um tema de discussão contínua e tem sido objeto de muitos estudos e pesquisas, e apesar de ser uma teoria bastante aceita, os estudos citados e outros similares invalidam a hipótese, pois mostram que é possível obter lucros acima da média através da análise de dados e técnicas estatísticas, e não apenas com base na informação pública.

2.2 ARBITRAGEM ESTATÍSTICA

A arbitragem estatística baseia-se na suposição de que os padrões observados no passado serão repetidos no futuro. Isso se opõe à estratégia de investimento fundamental que explora e tenta prever o comportamento das forças econômicas que influenciam os preços das ações. Assim, a arbitragem estatística é uma abordagem puramente estatística projetada para explorar as ineficiências do mercado de ações definidas como o desvio do equilíbrio de longo prazo entre os preços das ações observados no passado (CALDEIRA; MOURA, 2013).

2.2.1 Pairs Trading

O *pairs trading* foi iniciado pelo grupo quant Nunzio Tartaglias no banco americano Morgan Stanley na década de 1980, e continua sendo uma importante técnica de arbitragem estatística usada por fundos de hedge. O grupo de Tartaglias descobriu que certos títulos estavam correlacionados em seus movimentos diários de preços (Vidyamurthy, 2004). Com base nessas investigações empíricas, estratégias de negociação podem ser formadas para explorar as ineficiências dos mercados de ações.

Vidyamurthy (2004) define *pairs trading* como uma estratégia neutra para o mercado, pois é uma operação que pode ser feita com qualquer condição de mercado. É considerado um modelo de arbitragem estatística feita com dois ativos com uma alta correlação ou cointegração no mercado. Após um momento em que os preços desses ativos não seguem a mesma tendência no curto prazo, é aberta uma posição de long-short, que consiste na compra de um ativo do par, e a venda a descoberto do outro ativo referente a esse par, com o objetivo de esperar uma reversão em longo prazo. Do mesmo modo, Caldeira e Moura (2013) definem *pairs trading* com uma estratégia de arbitragem estatística projetada para explorar desvios de curto prazo de um equilíbrio de longo prazo entre duas ações.

A estratégia de *pairs trading* está embasada no efeito de reversão à média, efeito que afirma que os desvios temporários do mercado tendem a se corrigir e a convergir ao padrão histórico. Em outras palavras, o efeito reversão está relacionado à mudança de desempenho das ações, onde é observado piora no desempenho das ações que tiveram melhor desempenho no passado e melhora do desempenho das ações que tiveram pior desempenho no passado. Como apresentado por Flori e Regoli (2021), esse efeito pode ser explicado pelo aumento da volatilidade dos papéis e pela reação exagerada dos agentes às novas informações distorcendo temporariamente os preços dos ativos.

Somado a isso, Jacobs e Weber (2015), que buscaram estudar os determinantes da lucratividade das estratégias *pairs trading* e observaram que os resultados dessas estratégias são consistentes em diferentes países, mostraram que a atenção e a reação ao noticiário dos investidores e o limite à arbitragem são fatores com alto poder explicativo na lucratividade dos pares selecionados. Sendo assim, pode-se classificar os trabalhos que desenvolvem estratégias *pairs trading* como pertencentes à literatura sobre anomalias de mercado. Essa classificação é explicada pelo fato dessas estratégias violarem a hipótese dos mercados eficientes desenvolvida por Fama (1971).

Krauss (2017), analisando a crescente literatura sobre *pairs trading*, mostra a variedade de abordagens que são possíveis na elaboração dessa estratégia. O autor destaca as cincos principais abordagens: (i) abordagem da distância; (ii) abordagem da cointegração; (iii) abordagem por séries temporais; (iv) abordagem de controle estocástico; e (v) outras abordagens.

Este trabalho foca-se em duas abordagens: a abordagem por cointegração e a abordagem por ML (enquadrada como outras abordagens de acordo com Kraus 2017).

2.2.2 Abordagem por cointegração

Esta abordagem requer o estudo dos padrões entre os pares de ativos. Este estudo envolve técnicas econométricas que, com o auxílio de ferramentas estatísticas, são aplicadas às séries temporais para verificar se o comportamento do par é estacionário e, portanto, apresenta tendência de reversão à média (CAVALCANTI et al.,2020). Por conta disso, uma das ferramentas adequadas para a modelagem da estratégia pairs trading é a cointegração.

Segundo Gujarati e Porter (2008), a cointegração é uma regressão de duas séries temporais com raiz unitária uma contra a outra, permitindo analisar a relação entre duas séries temporais não estacionárias sem produzir uma regressão espúria. Além disso, o uso isolado da correlação dos retornos das ações não seria adequado porque a correlação não garante a reversão à média, pois esse é um conceito que está relacionado ao conceito de séries estacionárias, tendo em vista que em um processo estacionário a variância é finita e, portanto, não é possível que a séria se afaste muito de sua média (SANTOS; PESSOA, 2017). Vidyamurthy (2004), autor do trabalho mais citado para *pairs tranding* com abordagem em cointegração, define que dadas duas (ou um conjunto) de séries temporais não estacionárias y(t) e x(t), se para certo valor γ a série $y(t) - \beta x(t)$ é estacionária, então as duas séries são ditas cointegradas.

Nas últimas décadas, o conceito de cointegração foi cada vez mais aplicado na econometria financeira e esta é uma técnica que permite a modelagem dinâmica de séries temporais não estacionárias (Alexander & Dimitriu 2002). A observação fundamental que justifica a aplicação do conceito de cointegração na análise de preços de ações é que um sistema envolvendo preços de ações não estacionários em níveis pode ter uma tendência estocástica comum (Stock & Watson 1988). Quando comparada ao conceito de correlação, a principal vantagem da cointegração é que ela possibilita a utilização das informações contidas nos níveis das variáveis financeiras. Alexander & Dimitriu (2005); Gatev et al. (2006); Caldeira & Portugal (2010), sugerem que a metodologia de cointegração oferece uma estrutura mais adequada para estratégias de arbitragem financeira.

Com a abordagem por cointegração, Caldeira (2013) utilizou dados do mercado de ações brasileiro no período de janeiro 2005 e dezembro de 2009 para elaborar uma carteira formada pelos 20 pares que apresentaram os melhores indicadores dentro da amostra utilizada. Como critério para iniciar a estratégia, o autor estimou o spread entre as séries cointegradas e, dessa maneira, abriu-se uma posição quando o spread apresentava dois desvios-padrão para cima ou para baixo. Em relação ao encerramento da posição, quando o spread estivesse a menos de 0,5 desvio da sua média ou já tivesse decorrido o período de meia vida do spread, a posição era encerrada. A estratégia elaborada por Caldeira (2013), levando em consideração os custos de transação, alcançou índice de Sharpe de 1,29 e rentabilidade média anual de 17,34%. Os resultados são robustos e reforçam o uso da cointegração como uma ferramenta importante para a gestão de fundos que utilizam estratégias pairs trading.

2.2.3 Teste de raiz unitária

No âmbito do estudo em questão, é importante destacar o uso do teste de raiz unitária e sua relação com a cointegração. Segundo (Hamilton, 1994), o teste de raiz unitária é uma técnica estatística amplamente utilizada para avaliar se uma série temporal possui uma raiz unitária, ou seja, se ela é estacionária. Já a cointegração é uma técnica estatística que é utilizada para verificar se duas ou mais séries temporais estão cointegradas, ou seja, se existe uma relação estatisticamente significativa entre elas (Enders, 2014). É importante notar que para se realizar testes de cointegração é necessário que as séries temporais sejam estacionárias, e é ai que entra o teste de raiz unitária, pois é ele que permite verificar se as séries são estacionárias, e assim poder realizar testes de cointegração. Portanto, antes de realizar testes de cointegração, é necessário realizar testes de raiz unitária para garantir que as séries são estacionárias, e assim.

2.2.4 **Abordagem por Machine Learning**

Com o avanço dos recursos computacionais nos últimos anos em conjunto com a maior facilidade de acesso a esses recursos, as técnicas de machine learning (ML) estão sendo cada vez mais utilizadas dentro do campo das finanças através de diferentes abordagens que são propostas para analisar dados financeiros com o intuito de alavancar as oportunidades de arbitragem estatística nos mercados financeiros (FLORI; REGOLI; 2021). Tais abordagens estão usualmente relacionadas com a previsão de séries temporais e identificação de parâmetros para estratégias de investimento. Além disso, essas ferramentas são particularmente interessantes ao campo das finanças pela sua capacidade de explorar relações não lineares (TROIANO et al., 2018).

O estudo de Alves (2015) aponta para o crescente interesse e desenvolvimento na área de inteligência artificial e machine learning (ML) aplicado ao mercado de ações. Isso se deve ao avanço tecnológico que permite a análise de grandes quantidades de dados e aplicação de técnicas avançadas de aprendizado de máquina. O uso do ML tem sido amplamente investigado para previsão de instrumentos financeiros e análise sentimental, como mencionado por Pontes (2011). A substituição de humanos no processo de tomada de decisão financeira tem sido um dos setores onde o uso de técnicas de ML tem mostrado resultados positivos.

Segundo o estudo de Silva e Barros (2018), o uso de técnicas de aprendizado de máquina tem permitido uma melhoria significativa na precisão das previsões de preços de ações. Eles mostraram como a utilização de técnicas de aprendizado de máquina, tais como redes neurais e árvores de decisão, pode superar as técnicas de previsão tradicionais.

O trabalho de Ferreira e Santos (2019) destaca como o uso de técnicas de aprendizado de máquina tem permitido a identificação de padrões e tendências no mercado de ações que são difíceis de serem identificadas manualmente. Eles mostraram como o uso de técnicas de aprendizado de máquina, como a análise de componentes principais e o agrupamento, pode ser usado para identificar relações entre diferentes ativos e melhorar a eficiência dos investimentos.

O uso de técnicas de machine learning tem permitido uma maior precisão e eficiência no processo de análise de dados e identificação de oportunidades de investimento. Como mencionado por Sarmento e Horta (2020), o uso de técnicas de aprendizado de máquina, como a redução de dimensionalidade e a clusterização, tem permitido a automatização do processo de seleção de pares de ações para negociação. Isso tem possibilitado uma maior eficiência e

rentabilidade no mercado de ações. Além disso, o uso de técnicas de aprendizado de máquina tem permitido a implementação de estratégias de investimento que são difíceis ou impossíveis de serem implementadas manualmente.

2.3 REDUÇÃO DE DIMENSIONALIDADE

Na busca por pares rentáveis, se faz necessário encontrar títulos com a mesma sistemática de exposição ao risco. De acordo com a Arbitrage Pricing Theory[3], esses títulos geram o mesmo retorno esperado de longo prazo. Quaisquer desvios do retorno teórico esperado podem, portanto, ser vistos como erros de precificação e servem como orientação para fazer negócios. Para extrair os fatores de risco subjacentes comuns para cada título, é proposta a aplicação do PCA na série de retornos, conforme descrito em Jolliffe (2011).

PCA é um procedimento estatístico que usa uma transformação ortogonal para converter um conjunto de observações de variáveis possivelmente correlacionadas em um conjunto de variáveis linearmente não correlacionadas, os componentes principais. A transformação é definida de tal forma que o primeiro componente principal seja responsável pelo maior número possível de variabilidade nos dados. Cada componente seguinte, por sua vez, tem a maior variância possível sob a restrição de que deve ser ortogonal aos componentes anteriores. É especialmente interessante notar que cada componente pode ser visto como representando um fator de risco (AVELLANEDA; LEE, 2010).

Sarmento e Horta (2020) descreve a aplicação do PCA da seguinte forma. Inicialmente, a série de retorno para um título *i* no tempo *t*, $R_{i,t}$, é obtida da série de preço do título P_i,

$$R_{i,t} = \frac{\left(P_{i,t} - P_{i,t-1}\right)}{P_{i,t-1}}$$

Em seguida, a série de retorno deve ser normalizada, pois o PCA é sensível ao dimensionamento relativo das variáveis originais. Isso é feito subtraindo a média, R_i, e dividindo pelo desvio padrão σ_i, como

[3]Arbitrage pricing theory (APT) é uma teoria geral de precificação de ativos que sustenta que os retornos esperados de um ativo financeiro podem ser modelados como uma função linear de vários fatores ou índices teóricos de mercado Ross SA (2013).

$$Y_i = \frac{R_i - \overline{R}_i}{\sigma_i}$$

A partir da série de retornos normalizados de todos os ativos, é calculada a matriz de correlação ρ, onde cada entrada é determinada por

$$\rho_{ij} = \frac{1}{T-1}\sum_{t=T}^{M} Y_{i,t} Y_{j,t}$$

De acordo com Sarmento e Horta (2020), a motivação para a aplicação do PCA em séries de retornos está no fato de que uma matriz de correlação de retornos é mais informativa para avaliar os co-movimentos de preços. O uso da série de preços pode resultar na detecção de correlações espúrias como resultado de tendências temporais subjacentes.

Para definir o número de características (features), k, Avellaneda e Lee (2010) utilizaram um procedimento que consiste em analisar a proporção da variância total explicada por cada componente principal e, em seguida, utilizar o número de componentes que explicam uma porcentagem fixa. Neste trabalho, no entanto, uma abordagem diferente é adotada. Como um algoritmo de Aprendizado Não Supervisionado é aplicado usando esses recursos de dados, deve haver uma consideração para a dimensionalidade dos dados. A alta dimensionalidade dos dados apresenta um problema duplo. O primeiro é que na presença de mais atributos, a probabilidade de encontrar características irrelevantes aumenta. O segundo é o problema da maldição da dimensionalidade, termo é introduzido por Bellman (1966) para descrever o problema causado pelo aumento exponencial do volume associado à adição de dimensões extras ao espaço euclidiano. Isso tem um grande impacto ao medir a distância entre pontos de dados aparentemente semelhantes que de repente se tornam muito distantes uns dos outros. Consequentemente, o procedimento de agrupamento se torna muito ineficaz. Segundo Berkhin (2006), o efeito começa a ser severo para dimensões maiores que 15. Levando isso em consideração, o número de dimensões do PCA é limitado superiormente neste valor e é escolhido empiricamente.

2.4 CLUSTERING

A tarefa de agrupar dados em subgrupos ou clusters é conhecida como clusterização (Han, Pei, & Kamber, 2011). Segundo Han, Pei e Kamber (2011), a clusterização é uma técnica estatística amplamente utilizada para identificar padrões ou relações entre elementos de um

grupo heterogêneo de dados. A clusterização permite organizar os dados em grupos com características similares, o que pode ser útil em várias áreas, como análise de mercado, análise de dados de saúde e análise de dados financeiros. Além disso, a clusterização pode ser combinada com outras técnicas estatísticas, como redução de dimensionalidade (Jolliffe, 2002), para melhorar a eficiência e precisão da análise dos dados. Um cluster pode ser entendido como uma coleção de registros que são similares entre si e dissimilares de objetos em outros grupos, onde objetos pertencentes a um dado cluster devem compartilhar um conjunto de propriedades comuns, sendo que essas propriedades não são compartilhadas com objetos de outros clusters Fayyad, Piatetsky-shapiro e Smyth (1996). Além de permitir a estruturação e fornecer uma melhor compreensão do conjunto de dados original, os resultados da tarefa de clustering também podem ser utilizados por outras técnicas de data mining, que realizariam seu trabalho nos clusters encontrados

Existem diversos métodos de clustering, dentre os quais podem ser destacados os métodos hierárquicos, de particionamento, baseados em grade, em modelos e em densidade (HAN; PEI; KAMBER, 2011). Métodos tradicionais de clustering, como os de particionamento, geralmente enfrentam dificuldades para encontrar agrupamentos com formatos arbitrários e não retornam bons resultados quando a base de dados em questão está contaminada com outliers (dados que destoam do padrão geral da distribuição). Outro ponto a ser considerado em alguns destes métodos, é que o usuário tem a necessidade de informar previamente o número de clusters que serão gerados, o que na maioria das vezes não é uma tarefa simples (ESTER et al., 1996; HAN; PEI; KAMBER, 2011).

De acordo com Kaufman e Rousseeuw (1990), a clusterização é uma técnica importante para a descoberta de estruturas ocultas em dados e é amplamente utilizada em diversas áreas, incluindo análise de mercado, análise de dados de saúde e análise de dados financeiros. Eles também mencionam que a clusterização pode ser combinada com outras técnicas estatísticas para melhorar a precisão e eficiência da análise de dados.

Jolliffe (2002) também destaca a importância da redução de dimensionalidade para a clusterização de dados, pois isso pode ajudar a remover a redundância e a melhorar a interpretabilidade dos resultados. Ele apresenta várias técnicas de redução de dimensionalidade, incluindo análise de componentes principais e análise de discriminante linear, e discute como elas podem ser aplicadas à clusterização de dados.

De acordo com Sarmento e Horta (2020), a escolha do método de clustering para utilização no *pairs trading* deve considera os seguintes pontos: (i) nenhuma suposição deve ser feita sobre o número apropriado de clusters nos dados, uma vez que não há informações prévias a esse respeito; (ii) ao tornar o número de clusters orientado a dados, deve-se introduzir o mínimo de viés possível; (iii) não se deve impor que cada título seja atribuído a um grupo, pois esperar-se encontrar títulos com séries de preços muito distintas e, em última análise, decomposições de PCA divergentes. Do ponto de vista de agrupamento, esses títulos serão identificados como outliers. Portanto, é necessário garantir que eles não interfiram no procedimento de agrupamento selecionando um algoritmo de agrupamento robusto para lidar com outliers; (iv) a atribuição deve ser rigorosa, caso contrário o número de combinações de pares possíveis aumenta, o que é conflitante com o objetivo inicial; (v) por fim, como não há informações prévias indicando que os clusters devem ser modelados regularmente, o algoritmo selecionado não deve adotar essa suposição. Para satisfazer tais requisitos, Sarmento e Horta (2020) sugerem a utilização de um método baseado em densidade. Na sequência são detalhadas duas técnicas de clusterização testadas neste artigo.

2.4.1 DBSCAN

O artigo acadêmico intitulado "A Density Based Spatial Clustering of Applications With Noise" (Ester et al., 1996) apresenta uma técnica de clusterização baseada na densidade. Essa técnica, conhecida como DBSCAN (Density-Based Spatial Clustering of Applications with Noise) é uma das técnicas mais populares de clusterização baseada em densidade e é amplamente utilizada em diversas áreas, incluindo mineração de dados, aprendizado de máquina e análise de redes. Nesse artigo foi apresentado o algoritmo DBSCAN, que pode ser aplicado a grandes conjuntos de dados que possuem outliers, ao mesmo tempo em que encontra clusters com diversos formatos com eficiência aceitável. O DBSCAN encontra agrupamentos baseando-se na vizinhança dos objetos, onde a densidade associada a um ponto é obtida por meio da contagem do número de pontos vizinhos em uma determinada região ao redor desse ponto (ERTÖZ; STEINBACH; KUMAR, 2003). Esse algoritmo possui a capacidade de encontrar clusters considerando as propriedades dos dados, pois não requer que seja informado antecipadamente o número de clusters, permitindo a formação de grupos com formatos arbitrários. Em contrapartida são necessários outros dois parâmetros de entrada para o algoritmo. Outras características importantes do algoritmo são a capacidade de identificar

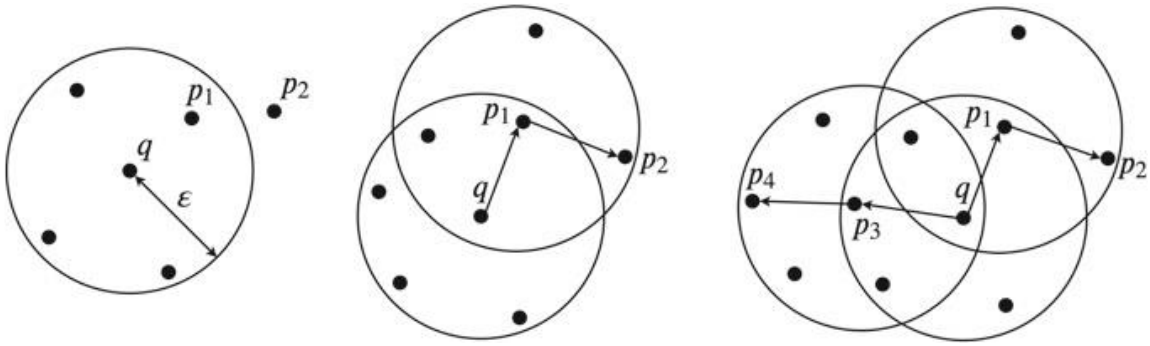
outliers, e a possibilidade de poder trabalhar com diversas funções de distância (ANKERST et al., 1999; ESTER et al., 1996).

Os dois parâmetros de entrada que o DBSCAN necessita são:

a. **raio de ε-vizinhança de um ponto:** determina o raio de vizinhança ε para cada ponto da base de dados. Dado o parâmetro ε, o algoritmo DBSCAN verifica a quantidade de pontos contidos no raio ε para cada ponto da base de dados, e se essa quantidade exceder certo número, um cluster é formado;

b. **número mínimo de pontos (η):** parâmetro que especifica o número mínimo de pontos, no dado raio de ε-vizinhança, que um ponto precisa possuir para ser considerado um ponto central e consequentemente, de acordo com as definições de cluster baseado em densidade, iniciar a formação de um cluster.

A Figura 1 ilustra os conceitos descritos anteriormente, quando o parâmetro minPts assume o valor de 5. A Figura 1a descreve um ponto central. O ponto q é um ponto central, dado que contém minPts dentro do círculo de raio ε, sua vizinhança ε. O ponto p1 pertence à vizinhança ε de q. Como p1, um ponto que não é um ponto central, mas está incluído em uma vizinhança ε, é chamado de ponto de fronteira. Um ponto que não pertence a nenhuma vizinhança, como p2, é considerado um outlier. A Figura 1b ilustra o conceito de pontos atingíveis diretamente por densidade e atingíveis por densidade. O ponto p1 é diretamente alcançável pela densidade de q. Além disso, p2 é a densidade alcançável a partir de q, pois q e p1 são pontos centrais. No entanto, q não é densidade alcançável a partir de p2, porque p2 não é um ponto central. Como acabamos de demonstrar, a densidade-acessibilidade não é simétrica em geral. Somente objetos centrais podem ser mutuamente alcançáveis por densidade. Finalmente, a Fig. 1c ilustra o conceito de conexão de densidade. Os pontos p2 e p4 são conexos por densidade, uma vez que ambos os pontos são densidade alcançáveis a partir de q.



(a) Pontos principais e outliers (b) Pontos alcançáveis por densidade (c) Pontos conectados por densidade

Fig. 1 - Ilustração DBSCAN de conceitos básicos, com minPts = 5
(FONTE: Sarmento e Horta (2020))

2.4.2 OPTICS

A técnica de clusterização escolhida por Sarmento e Horta (2020) é a OPTICS (Ordering Points To Identify the Clustering Structure). Esta escolha é justificada pela capacidade única do algoritmo de lidar com conjuntos de dados de alta dimensionalidade e heterogeneidade, como mencionado por Ankerst, Breunig, Kriegel e Sander (1999) e pelo seu desempenho eficiente em encontrar clusters de diferentes tamanhos e densidades, conforme descrito por Moulavi, Jain e Sander (2014). Além disso, a análise hierárquica dos clusters também é importante, pois permite explorar a estrutura do cluster de forma mais profunda, como mencionado por Campello *et al.* (2013).

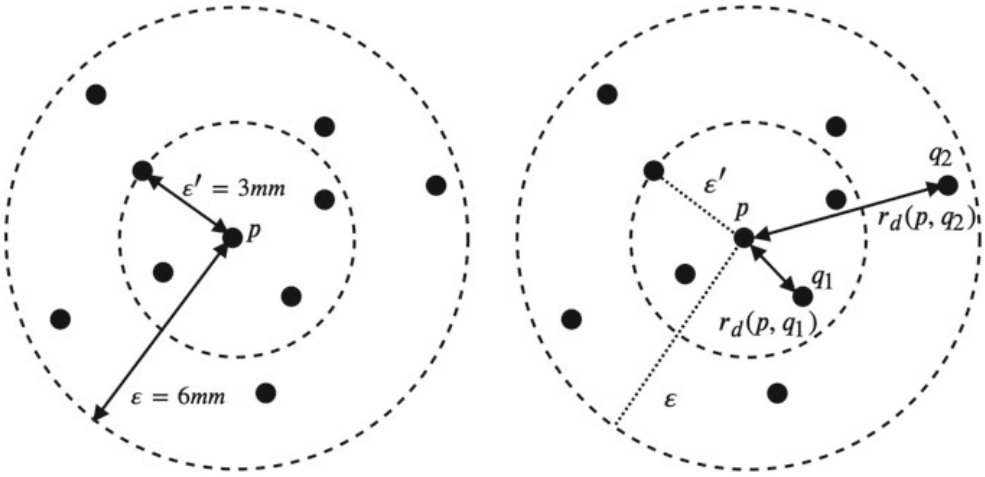
De acordo com o trabalho de Sarmento e Horta (2020), a tarefa principal é encontrar pares co-integrados, o que é possível através da utilização de técnicas de clusterização. O OPTICS se mostrou ser a escolha mais adequada para esta tarefa, pois é capaz de lidar com as complexidades dos dados de mercado e encontrar clusters de diferentes tamanhos e densidades, permitindo uma análise mais precisa e profunda dos pares co-integrados.

OPTICS, proposto por Ankerst *et al.* (1999), aborda o problema de detectar clusters significativos em dados de densidade variável. O algoritmo é inspirado no DBSCAN, com a adição de dois novos conceitos. Para as definições de OPTICS, será mantido a nomenclatura utilizada na definição do DBSCAN.

A distância central de um ponto p é, simplesmente, a menor distância ε' entre p e um ponto em sua vizinhança ε tal que p é um ponto central em relação a ε'. Se p não é um ponto central em primeiro lugar, a distância central não é definida. A distância de alcance de um ponto

p em relação a um ponto o pode ser interpretada como a menor distância tal que p é diretamente alcançável por densidade de o. Isso requer que o seja um ponto central, o que significa que a distância de alcance não pode ser menor que a distância central. Se fosse esse o caso, o não seria um ponto central.XXX corrigir a colagem da figura abaixo.

A Figura 2 representa os conceitos de distância do núcleo e distância de acessibilidade descrita acima. O cenário representado define minPts = 5 e ε = 6mm. Na imagem da esquerda, a distância-núcleo, representada como ε', retrata a distância mínima que faz de p um ponto-núcleo. Para um raio de ε' < ε, já existem cinco pontos (minPts) dentro do círculo. O conceito de alcance-distância está representado na imagem à direita. A distância de alcance entre q1 e p deve garantir que p seja um ponto central, embora a distância entre os dois pontos seja realmente menor. A distância de alcance entre q2 e p, corresponde à distância real entre os dois pontos, pois qualquer distância maior que ε' já garante que p é um ponto central.



(a) Distância do núcleo
$c_d(p) = \varepsilon'$

(b) Distância alcançável
$r_d(p, q_1) = \varepsilon \;\; e \;\; r_d(p, q_2) = d(p, q_2)$

Fig. 2 - Ilustração de distância do núcleo e distância de acessibilidade
(FONTE: Sarmento e Horta (2020))

3. METODOLOGIA

A seguir, apresentam-se os procedimentos metodológicos utilizados para testar a aplicabilidade da hipótese central de Sarmento e Horta (2020) para a seleção dos pares de ativos. Optou-se por utilizar um grupo de 700 ativos divididos entre ETFs, e principais ações

dos índices S&P500, Dow Jones e Nasdaq. Segundo os autores, o método utilizando a redução de dimensionalidade e o clustering deveria apresentar um conjunto de pares cointegrados e com maior probabilidade de retornos positivos - desde que preservadas certas condições de formulação matemática. Esta é uma asserção importante, ainda que não suficiente, para validar ou descartar a método descrito por Sarmento e Horta (2020), visto que existe mais de uma forma de testar a cointegração entre duas séries temporais e uma séries de parâmetros podem ser alterados dado o período e o *timeframe* avaliado.

O artigo classifica-se como pesquisa aplicada, dado que o método é aplicado a ativos reais e, as simulações obtidas têm aplicação real. A abordagem classifica-se como quantitativa, pois a tomada de decisão é totalmente baseada no resultado de cálculos dos algoritmos. Por se tratar de uma aplicação do modelo descrito por Sarmento e Horta (2020), trata-se de pesquisa explicativa, com a finalidade de justificar o potencial das tecnologias aplicadas para superar o retorno dos principais índices de ações no mercado. Quanto ao procedimento, classifica-se como experimental, pois há possibilidade de alterar diversos parâmetros dos modelos de machine learning e testes matemáticos e obter diferentes respostas para o problema.

3.1. REGRAS DE SELEÇÃO DE PARES

A proposta de Sarmento e Horta (2020) pode ser ilustrada na Figura 3, onde:

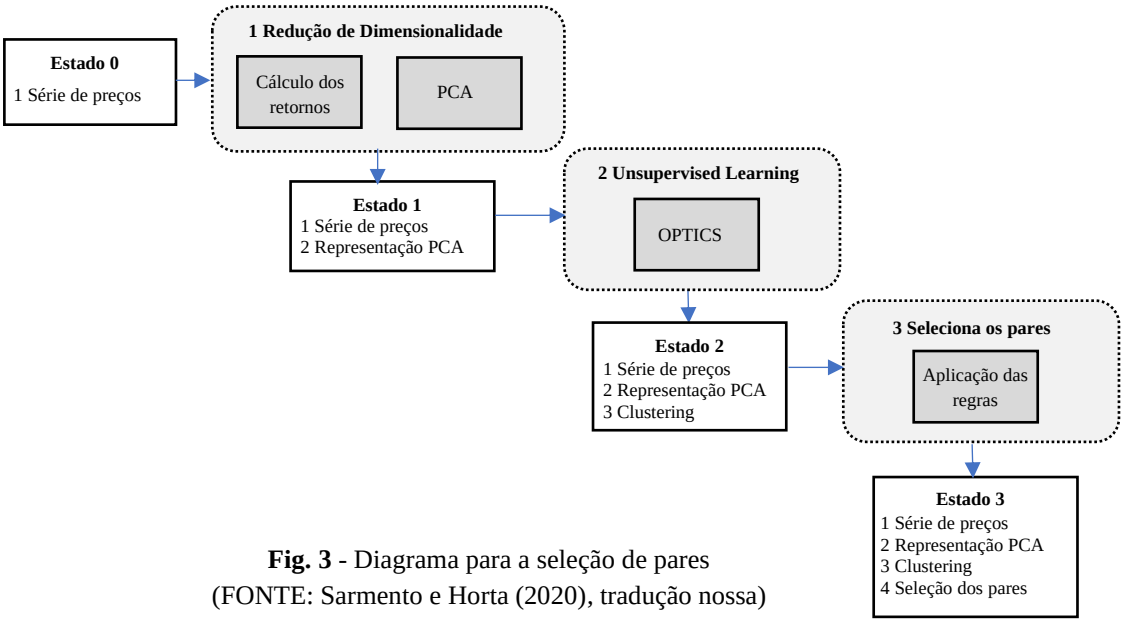


Fig. 3 - Diagrama para a seleção de pares
(FONTE: Sarmento e Horta (2020), tradução nossa)

- **Estado 0:** O estado inicial abrange a série de preços para todos os possíveis constituintes dos pares. Pode-se considerar que esta informação está disponível para o

investidor. Os principais ativos utilizados neste estudo são denominados como ETFs (Exchange-Traded Funds) e descritos por Sarmento e Horta (2020) como um tipo interessante de títulos a serem explorados para este modelo. Para eles, um ETF é um título que rastreia um índice, uma commodity ou uma cesta de ativos como um fundo de índice, mas é negociado como uma ação. De acordo com Chan (2013), a única vantagem de negociar pares de ETFs em vez de pares de ações é que, uma vez cointegrados, os pares de ETFs são menos propensos a sair da base amostral amostra. Isso ocorre porque a economia fundamental de uma cesta de ações muda mais lentamente do que a de uma única ação. Como adicional, para complementar a analise, será incluso ações à cesta de ativos financeiros;

- **Estado 1:** Então, reduzindo a dimensionalidade dos dados, cada título pode ser descrito não apenas por sua série de preços, mas também pela representação compacta que é oriunda da aplicação do PCA na série de retornos;

- **Estado 2:** Usando esta representação simplificada, o algoritmo OPTICS é capaz de organizar os ativos em clusters;

- **Estado 3:** Finalmente, é possível procurar combinações de pares dentro dos clusters e selecionar aqueles que verificam as regras descritas na Figura 4.

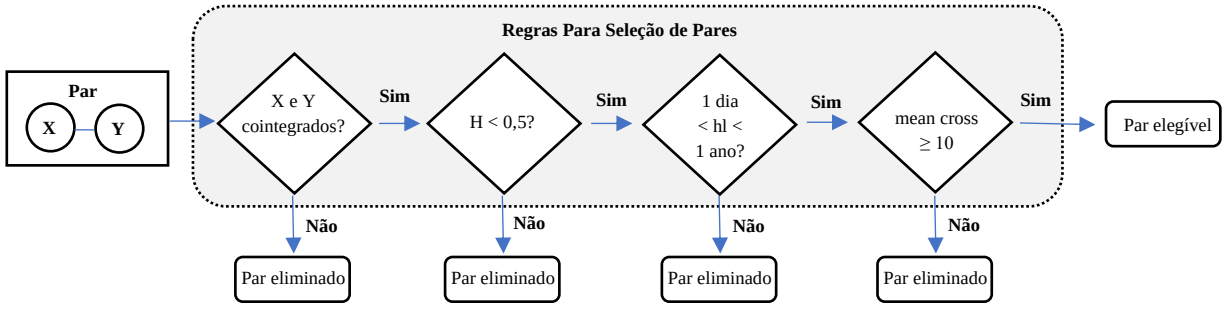


Fig. 4 - Regras para a seleção de pares
(FONTE: Sarmento e Horta (2020), tradução nossa)

De acordo com os critérios propostos, um par é selecionado se cumprir as quatro condições descritas a seguir:

1. Os constituintes do par são cointegrados: para testar esta condição, os autores propõem a aplicação do teste de Engle-Granger, devido à sua simplicidade.

2. O expoente de Hurst (H) do spread do par revela um caráter de reversão à média: uma etapa de validação adicional é sugerida para fornecer mais confiança no caráter de reversão à média do spread dos pares. Tem como objetivo restringir falsos positivos, possivelmente surgindo como efeito do problema de comparações múltiplas. A condição imposta é que o H associado ao spread de um determinado par seja menor que 0,5, garantindo que o processo se incline para a reversão à média.

3. O spread entre pares diverge e converge em períodos específicos: um spread de reversão à média isoladamente não garante lucros. É crucial estabelecer uma relação adequada entre a duração média da reveão e o período de negociação. A meia-vida (hl) pode ser interpretada como uma estimativa do tempo esperado para a reversão à média do spread. Portanto, os autores sugerem filtrar os pares cuja meia-vida não esteja alinhada com o período de negociação.

4. O spread do par reverte para a média com frequência suficiente: é necessário que cada spread ultrapasse sua média pelo menos uma vez por um período definido, para fornecer liquidez suficiente.

3.2. SIMULAÇÃO DE NEGOCIAÇÃO

A seguir, os detalhes sobre a simulação de negociação proposta por Sarmento e Horta (2020). Os aspectos considerados na construção da carteira são os custos de transação e as condições de liquidação das posições.

3.2.1 Custos de Transações

Todos os resultados apresentados neste trabalho contabilizam os custos de transação. Os custos de transação são baseados em estimativas de Do e Faff (2012). Os autores realizaram um estudo aprofundado sobre o impacto dos custos de transação no Pairs Trading. Os custos de comissão e impacto de mercado foram adaptados para contabilizar ambos os ativos do par. Os custos associados a uma transação são calculados conforme representado na Figura 5.

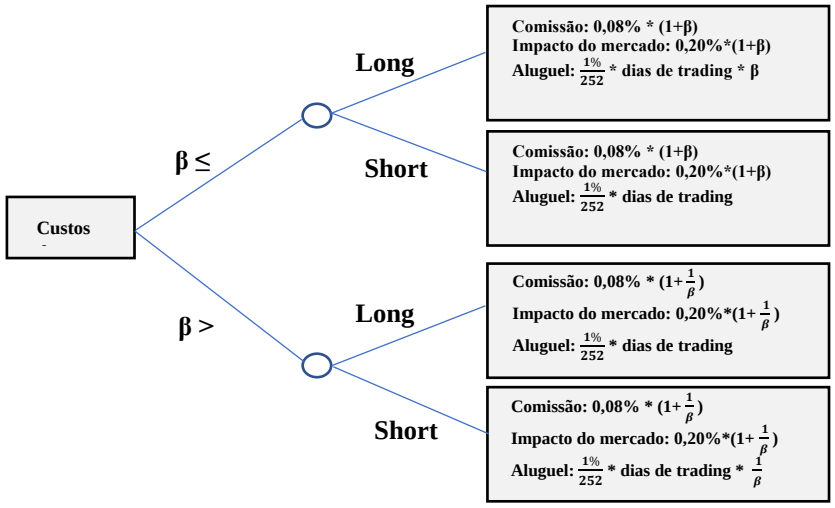


Fig. 5 - Cálculos para os custos de transações
(FONTE: Sarmento e Horta (2020), tradução nossa)

3.2.2. Modelo de Negociação Baseado em Limites

No modelo proposto por Gatev, Goetzmann e Rouwenhorst (2006), o critério para abertura de uma negociação é baseado na divergência de spread. Se o spread entre duas séries de preços que compõem um par divergir em mais de dois desvios padrão históricos, uma negociação é aberta. O negócio é fechado na convergência para a média, no final do período de negociação ou quando ocorre o fechamento. Este trabalho utilizará o mesmo modelo, pois corresponde à implementação introduzida por Sarmento e Horta (2020). Este modelo pode ser descrito mais formalmente da seguinte forma: (i) calcula-se a média ($S_t = Y_t - \beta X_t$) do spread, μ_s e o desvio padrão, σ_s durante o período de formação do par; (ii) definem-se os limites do modelo: o limite que aciona uma posição comprada, α_L, o limite que aciona uma posição vendida, α_S, e o limite de saída $\alpha_{saída}$ que define o nível em que uma posição deve ser encerrada; (iii) monitora-se a evolução do spread, S_t e controlar se algum limiar for ultrapassado; (iv) no caso de α_L ser cruzado, compre o spread comprando 1Y e vendendo βX. Se α_S for acionado, venda o spread vendendo 1Y e comprando βX. Saia da posição quando $\alpha_{saída}$ for acionado e uma posição estava sendo mantida.

3.3. MÉTRICAS DE AVALIAÇÃO

A metodologia utilizada neste estudo para avaliar o desempenho do modelo proposto por Sarmento e Horta (2020) inclui a análise de métricas relacionadas ao desempenho de negociação. Tais métricas incluem o lucro bruto, lucro líquido, retorno sobre o patrimônio

líquido (ROE) e o risco-ajustado retorno (Sharpe ratio). Estas métricas fornecerão uma visão geral do desempenho financeiro do modelo, permitindo uma comparação com outros modelos e com o índice S&P 500.

Além disso, será realizado backtesting da estratégia proposta, avaliando a consistência do modelo ao longo do tempo. Este processo permitirá avaliar se o modelo mostra-se consistente em diferentes condições de mercado e se é capaz de produzir resultados financeiros positivos.

Por fim, será realizada a avaliação de risco-retorno, que medirá o desempenho do modelo levando em consideração o risco envolvido. Isso permitirá uma análise mais completa do desempenho do modelo, levando em conta não apenas os lucros gerados, mas também o risco envolvido no processo.

3.4. AMBIENTE DE IMPLEMENTAÇÃO

O código desenvolvido neste trabalho é implementado em Python. A motivação para tal é a vasta quantidade de recursos disponíveis que facilitam a implementação dos algoritmos escolhidos, nomeadamente procedimentos de mineração de dados e análise de dados. Além disso, o Python tem sido, no momento da escrita, a linguagem de escolha para projetos relacionados ao Machine Learning, o que o torna mais adequado também do ponto de vista da colaboração. Algumas bibliotecas são particularmente úteis neste trabalho: o sci-kit learn é útil na implementação do PCA e do algoritmo OPTICS, enquanto o statsmodels fornece uma versão já implementada do teste ADF, útil para testar a cointegração.

A simulação é executada em uma VM EC2 na plataforma AWS (AMD EPYC 7R13 32 threads e 64 GB de memória RAM). Esses modelos envolvem um grande volume de multiplicações de matrizes que resultam em longos tempos de processamento ao usar a CPU.

3.5. BASE DE DADOS

Neste estudo, os dados empregados na análise e implementação do modelo de Pairs Trading baseado em Aprendizado de Máquina foram obtidos a partir do Yahoo Finance, uma plataforma amplamente reconhecida e confiável no campo das finanças. Os dados coletados referem-se aos preços de fechamento ajustados das ações, os quais são apropriados para análises de séries temporais, pois levam em consideração eventos como dividendos e desdobramentos de ações, garantindo assim a consistência e a comparabilidade dos dados ao longo do tempo.

A escolha dos preços de fechamento ajustados para o estudo se deve à sua capacidade de fornecer informações precisas e atualizadas sobre o comportamento das ações no mercado, permitindo uma análise mais profunda das relações de co-integração entre os pares. Esses dados são essenciais para a implementação e avaliação do modelo proposto, uma vez que possibilitam a identificação de padrões temporais e a comparação do desempenho do modelo com o índice S&P 500 através de backtesting da estratégia.

A diversidade da base permite uma compreensão mais ampla das tendências e padrões no mercado financeiro, que foi explorada neste estudo. A base de dados analisada contém um total de 729 ativos, incluindo empresas e ETFs. Esses ativos estão distribuídos entre diversos setores da economia, garantindo uma representatividade abrangente do mercado. Os setores incluem Tecnologia da Informação, Saúde, Financeiro, Consumo Discricionário, entre outros.

Ao analisar a distribuição por indústria, observa-se uma diversidade ainda maior, com ativos em áreas como Biotechnology, Software & Services, Pharmaceuticals, Banks, e muitas outras. Essa variedade permite explorar tendências e oportunidades de investimento em diferentes segmentos do mercado.

No que diz respeito às exchanges onde os ativos são negociados, a maioria está listada em grandes bolsas de valores, como a NYSE e a NASDAQ para equities e para ETFs a PCX e BTS.Em relação aos mercados, os ativos estão distribuídos entre diferentes classificações, como Large Cap, Mid Cap e Small Cap. Essa diversidade permite uma melhor compreensão do comportamento dos ativos com base no tamanho de mercado das empresas. Por fim, os ativos financeiros incluídos no estudo estão localizados em diversos países, com uma predominância de empresas dos Estados Unidos. A presença de ativos de diferentes países enriquece a análise, permitindo a identificação de oportunidades globais de investimento.

4. RESULTADOS

Com base na amostra composta por 729 ativos, incluindo empresas e ETFs, e utilizando os dados de fechamento ajustados coletados do Yahoo Finance, foi aplicado o modelo de Pairs Trading baseado em Aprendizado de Máquina descrito por Sarmento e Horta (2020). Inicialmente, buscou-se testar o funcionamento do modelo a fim de avaliar se a abordagem geral seria suficiente para identificar pares negociáveis corretamente.

Uma vez o modelo desenhado, foram realizados os seguintes testes:

- Filtro por PCA (Análise de Componentes Principais) para reduzir a dimensionalidade dos dados e encontrar uma representação compacta para cada ação. A figura 6 representa o peso percentual que cada ativo financeiro recebe em cada componente principal.

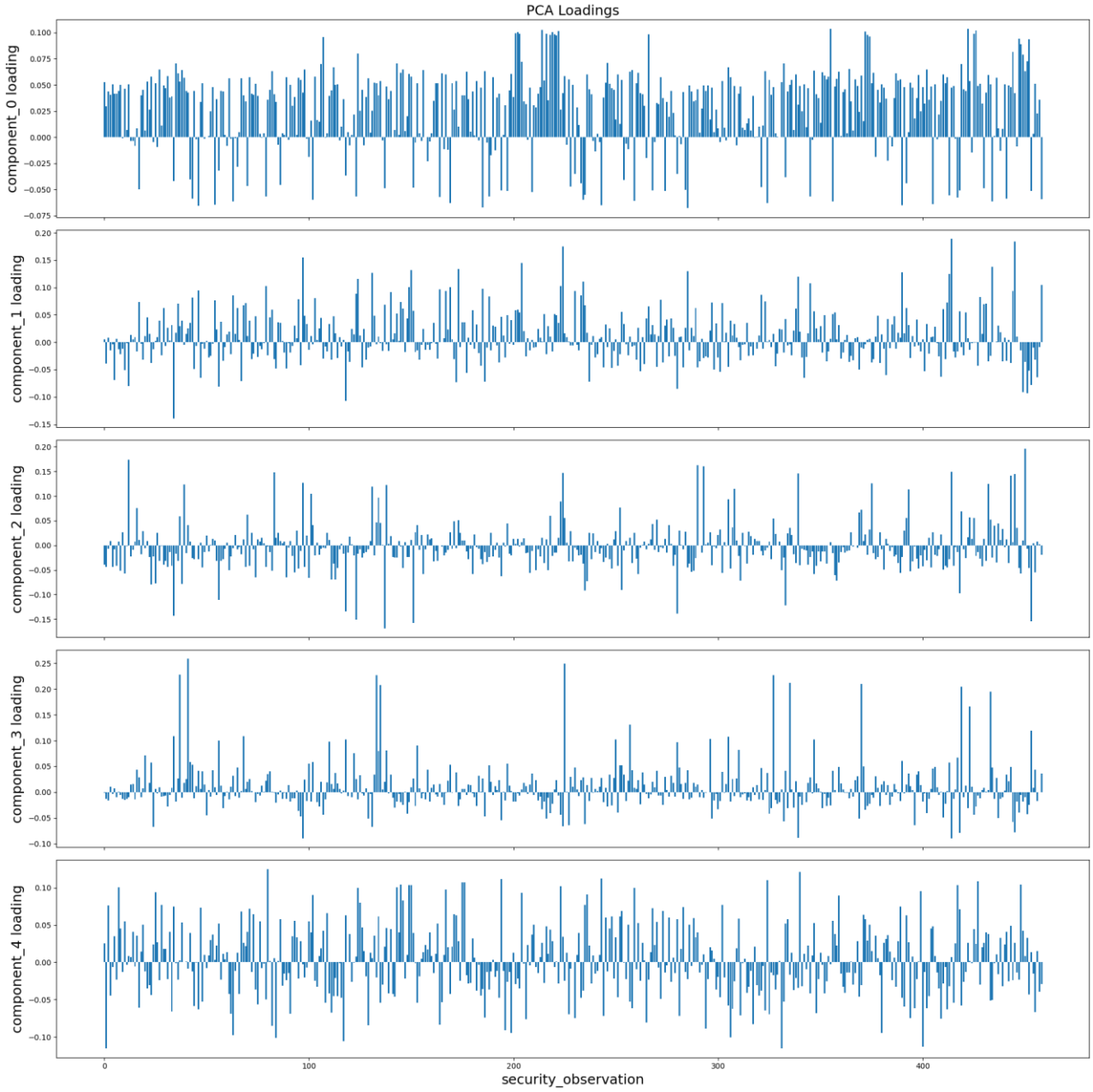


Fig. 6 – PCA
(FONTE: Elaborado pelo autor)

- Geração de clusters utilizando aprendizado não supervisionado, com o objetivo de agrupar os ativos em potenciais pares. A figura 7 ilustra em 3 dimensões os cluster gerados pelos 3 primeiros componentes principais e as cores destacam a separação dos clusters.

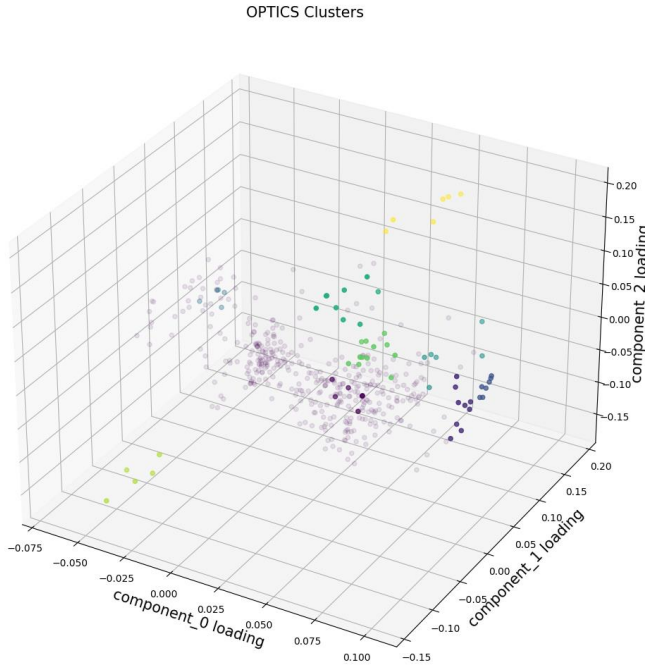


Fig. 7 – Clusters
(FONTE: Elaborado pelo autor)

- Aplicação de critérios de seleção para identificar os pares negociáveis dentro dos clusters gerados. A tabela 1 apresenta o resultado do PCA, da clusterizacao e a aplicação das regras de seleção, gerando os dados como ilustrado abaixo:

	pair	pvalue	hurst_exp	half_life	avg_cross_cou	alpha	beta	dependent	independent
67	(IWM, IWR)	0.025168	0.356335	10.270.227	23.688	0.259859	1.240.436	security_1	security_0
127	(SRE, WEC)	0.005387	0.330191	10.677.561	26.208	-0.158134	1.783.392	security_1	security_0
131	(WEC, XLU)	0.018504	0.315463	13.074.967	21.168	0.045734	0.512779	security_1	security_0
144	(AVB, EQR)	0.027569	0.247495	13.779.205	26.712	0.119522	1.248.484	security_1	security_0
154	(EQR, ESS)	0.005650	0.206672	14.814.869	32.256	-0.134638	0.857497	security_0	security_1
156	(EQR, MAA)	0.009082	0.197320	10.023.052	30.240	0.246261	1.476.035	security_1	security_0
158	(EQR, PSA)	0.023357	0.166810	18.592.668	23.688	0.006356	1.674.039	security_1	security_0
173	(IYR, PSA)	0.032522	0.310207	16.379.935	16.128	0.025065	2.221.128	security_1	security_0
190	(PSA, VNQ)	0.014717	0.263850	13.854.995	15.624	-0.442073	2.209.966	security_0	security_1

Tabela 1 – Pares de ativos financeiros selecionados.
(FONTE: Elaborado pelo autor)

Já a figura 8 ilustra o movimento dos preços entre o par de ativos financeiros bem como o spread gerado pelo movimento normalizado dos preços.

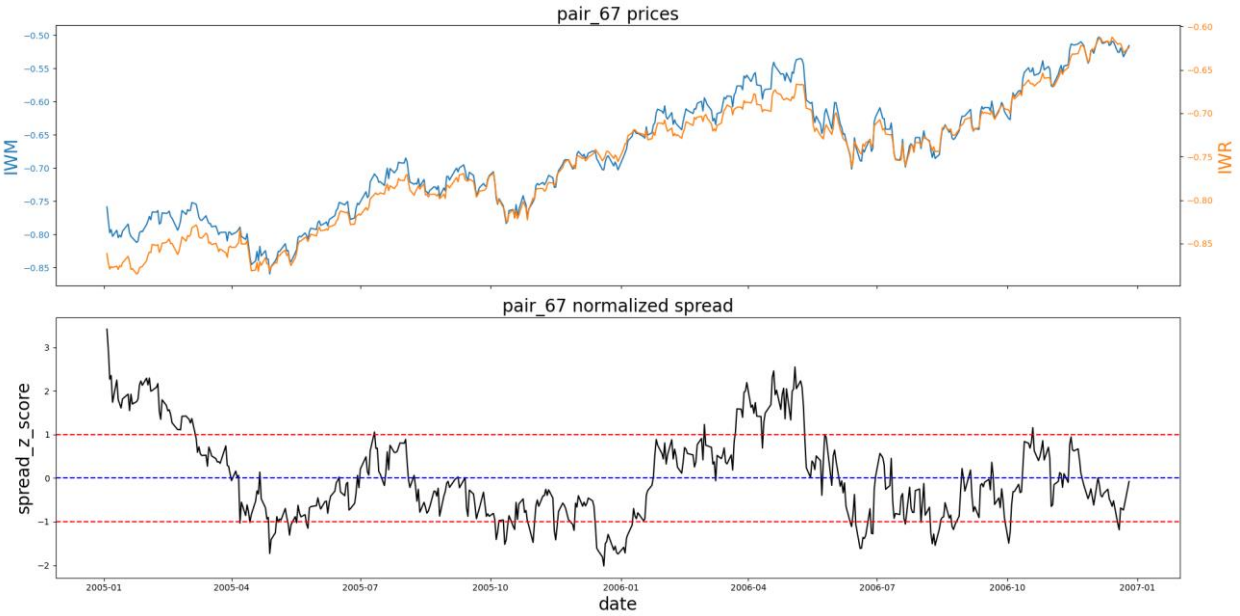


Fig. 8 – Movimento dos preços e z-spread de 1 par selecionado.
(FONTE: Elaborado pelo autor)

Essa abordagem resultou em uma lista de pares negociáveis, demonstrando que o modelo foi capaz de realizar as etapas necessárias para a identificação de oportunidades de negociação no mercado financeiro. Esses resultados preliminares forneceram uma base sólida para a implementação e avaliação da estratégia de Pairs Trading com base no modelo proposto.

4.1. RESULTADOS DO MODELO

Com base na execução diário realizada entre 2005 e final de 2022, foi possível avaliar a eficácia da estratégia de Pairs Trading proposta por Sarmento e Horta (2020). O modelo gerou diariamente uma lista de pares negociáveis com base em critérios de seleção como o filtro por PCA e a geração de clusters, e utilizou os parâmetros de Beta e desvio padrão como pontos de entrada e saída.

Inicialmente foi realizado a avaliação do número de trades por dia e a exposição de capital em cada operação, que são aspectos cruciais a serem considerados na implementação de uma estratégia de negociação. Segundo um estudo de Abdi et al. (2020), a seleção adequada do número de trades por dia pode influenciar significativamente o desempenho da estratégia de negociação. Por um lado, um número excessivo de trades pode levar a custos de transação mais elevados, enquanto que, por outro lado, um número insuficiente de trades pode limitar o potencial de lucro. Dessa forma, é necessário encontrar um equilíbrio entre o número de trades e o custo de transação, de modo a maximizar o retorno do investimento.

Além disso, a exposição de capital em cada operação também é um fator crítico a ser avaliado, uma vez que pode influenciar significativamente o risco e a rentabilidade da estratégia de negociação. De acordo com um estudo de Albergaria et al. (2020), a exposição de capital em cada operação deve ser cuidadosamente gerenciada, de modo a evitar perdas excessivas e maximizar o retorno do investimento. Para isso, é importante definir limites claros de perda máxima e exposição máxima de capital em cada operação, com base em uma análise cuidadosa dos riscos envolvidos e dos objetivos de investimento.

Em resumo, a avaliação do número de trades por dia e da exposição de capital em cada operação é fundamental para a implementação de uma estratégia de negociação eficaz. Como afirmam Abdi et al. (2020), "a seleção cuidadosa do número de trades por dia é essencial para maximizar o retorno do investimento, enquanto a gestão adequada da exposição de capital em cada operação pode ajudar a minimizar o risco e maximizar a rentabilidade". Dessa forma, é importante considerar esses aspectos durante a implementação da estratégia de negociação, a fim de alcançar um desempenho consistente e satisfatório.

4.1.1. Volume de Operações

Para a visualização e análise dos pares negociados, gerou-se a média móvel (21 dias) de pares negociados ao longo do tempo, permitindo uma visualização mais clara das tendências e padrões observados. A utilização da média móvel suaviza flutuações diárias e destaca padrões mais claros, o que facilita a análise dos dados coletados.

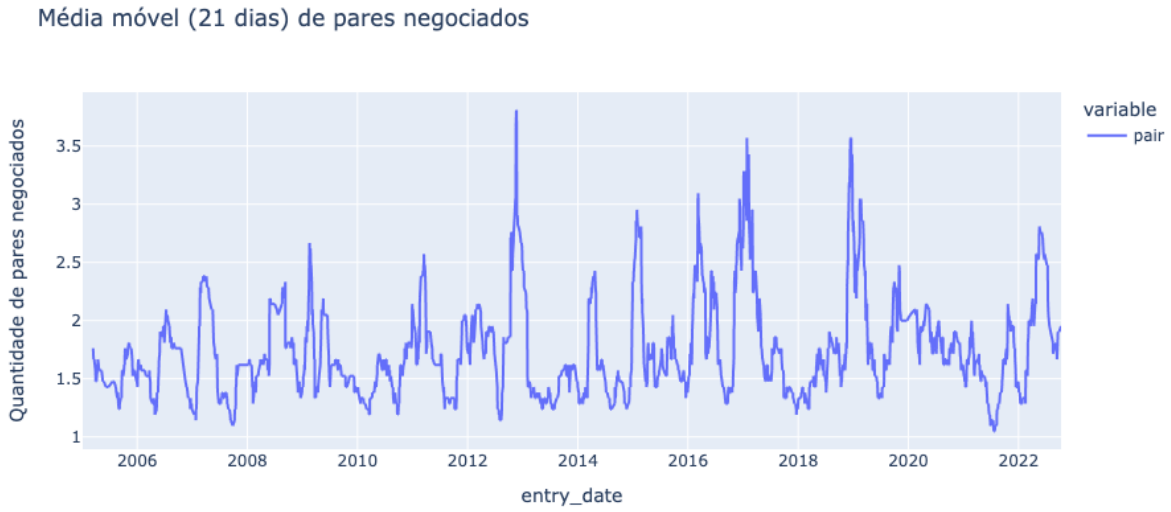


Fig. 9 – Média móvel (21 dias) de pares negociados por dia.
(FONTE: Elaborado pelo autor)

A figura 10 apresenta um histograma de pares negociados ao longo do tempo, fornecendo uma visão mais detalhada da distribuição dos dados. Juntos, esses gráficos fornecem uma visão geral do volume de negócios gerados pelo modelo e permitem uma análise mais aprofundada dos dados coletados. Esses dados expressam que o modelo não gera excesso de operações por dia, evitando a perda de rentabilidade devido aos custos atrelados às operações de compra e venda dos ativos.

Fig. 10 –Histograma de pares negociados.
(FONTE: Elaborado pelo autor)

4.1.2. Exposição de Capital

Para avaliar o desempenho das operações de Pairs Trading, é importante analisar a exposição financeira de cada operação, que representa o percentual do patrimônio total investido em cada par de ativos. A figura 8 apresenta a exposição financeira diária das operações de Pairs Trading, mostrando a variação do percentual de exposição ao longo do tempo.

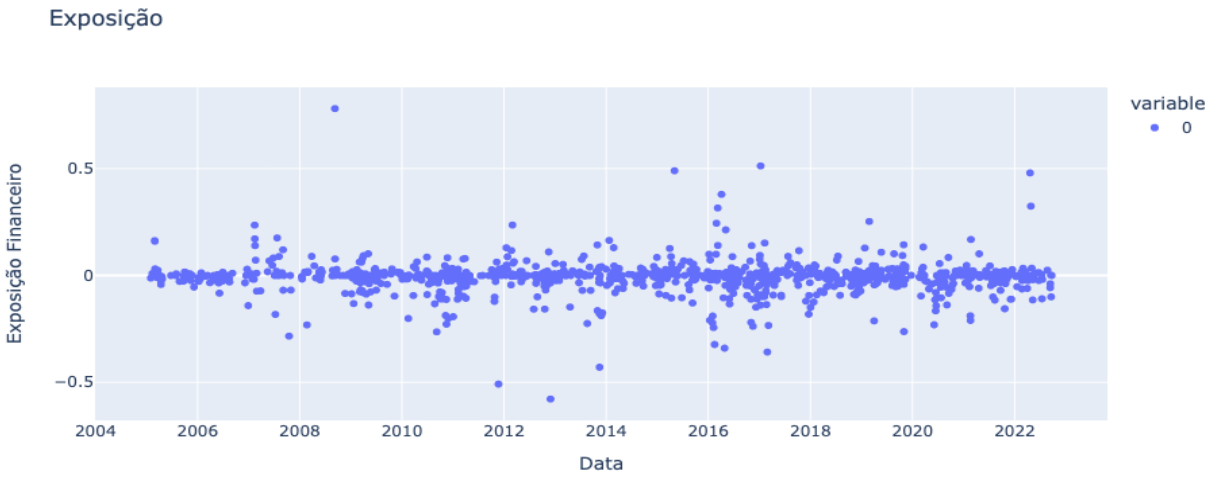


Fig. 11 – Exposição financeira por dia de trade.
(FONTE: Elaborado pelo autor)

Através dessa figura, é possível verificar a exposição de cada operação e avaliar a capacidade de gestão de risco do modelo de Pairs Trading. Além disso, a figura 12 apresenta um gráfico de caixa que possibilita visualizar a mediana e os desvios da exposição financeira de cada operação, permitindo uma análise mais detalhada da distribuição dos dados e identificando possíveis outliers que podem impactar o desempenho da estratégia. A análise da exposição financeira de cada operação é essencial para a gestão de risco e para a tomada de decisão em relação às operações de Pairs Trading.

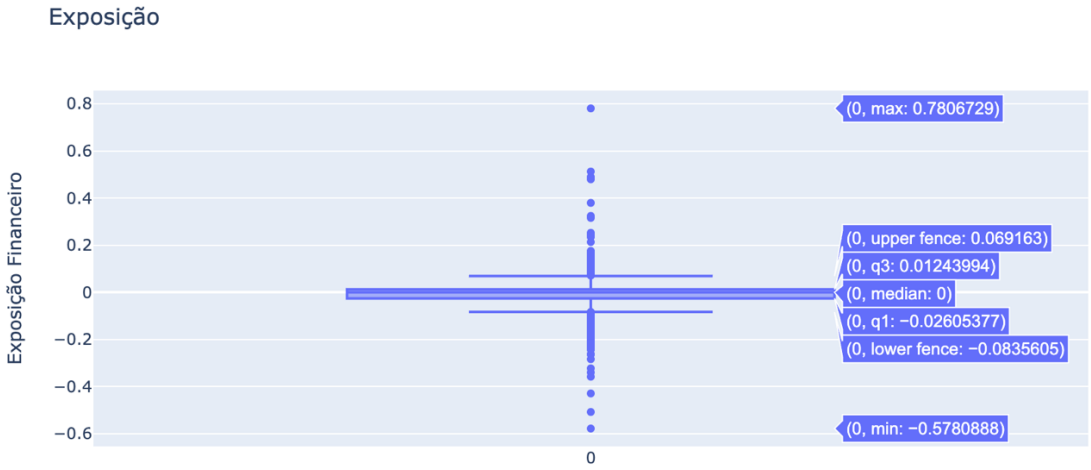


Fig. 12 – Distribuição da exposição.
(FONTE: Elaborado pelo autor)

Na figura 12, é possível verificar que a mediana da exposição financeira das operações de Pairs Trading é zero, indicando que as operações são relativamente balanceadas em termos

de exposição. No entanto, a lower e a upper fence do box mostram que a partir de +7% e -8% de exposição são encontrados os outliers. Quando outliers ocorrem é importante ressaltar uma preocupação significativa quanto ao risco na operação devido à alta exposição financeira. Essas exposições anormais podem ser explicadas pelo tamanho do beta de cada ativo. Ativos com betas muito altos podem gerar exposições anormais em relação ao par de ativo em questão.

4.2. RENTABILIDADE

Para avaliar o desempenho da estratégia de Pairs Trading, é importante considerar diversos fatores que afetam a rentabilidade da estratégia na vida real. Um desses fatores é a existência de taxas e comissões cobradas pelas corretoras e exchanges, que dificultam a avaliação precisa da rentabilidade da estratégia. Para contornar essa questão, é comum realizar uma aproximação dos valores cobrados, a fim de avaliar o impacto desses custos na rentabilidade da estratégia. Chan et al. (1996) destacam a importância da gestão adequada dos custos na avaliação de estratégias de negociação, ressaltando a necessidade de considerar esses custos de forma realista. Além disso, há também a incerteza quanto à execução da ordem de compra ou venda no valor determinado pelo modelo. Para avaliar o impacto dessa incerteza na rentabilidade da estratégia, é importante realizar simulações inserindo um valor aleatório nos preços dos ativos, tanto aumentando quanto diminuindo os preços artificialmente. Dessa forma, é possível avaliar a robustez da estratégia diante das variações de preço. Chan et al. (2013) discutem a importância da simulação de Monte Carlo na avaliação da robustez de estratégias de negociação, destacando a necessidade de avaliar o desempenho da estratégia em diferentes cenários de mercado.

Com base nessas considerações, a avaliação realista da rentabilidade da estratégia de Pairs Trading envolve a consideração cuidadosa dos custos adicionais e a simulação de preços aleatórios. Ao considerar esses fatores, é possível obter uma avaliação mais precisa da rentabilidade da estratégia, bem como identificar potenciais pontos de melhoria.

A figura 13 apresenta a curva de capital da estratégia de Pairs Trading no período de backtest. A linha em destaque representa a curva de capital considerando os preços reais dos ativos, sem a inclusão de fatores aleatórios nos preços. As demais linhas representam a curva de capital considerando preços com fatores aleatórios inseridos, aumentando ou diminuindo artificialmente os preços dos ativos.

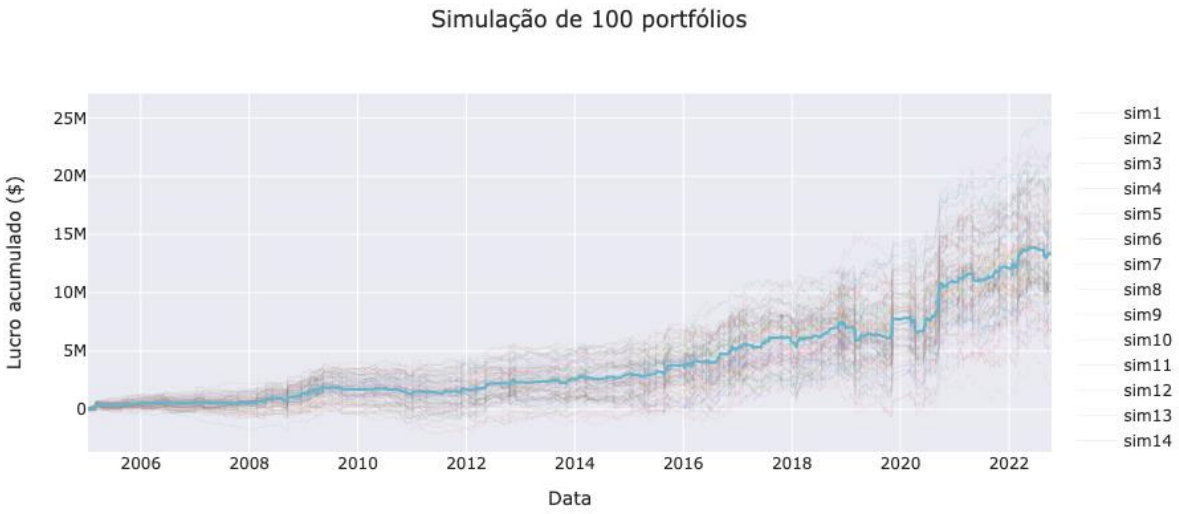


Fig. 13 – Curva de Capital considerando preço real e com fatores aleatórios.
(FONTE: Elaborado pelo autor)

Como pode ser observado na figura 10, a curva de capital considerando os preços reais apresenta um crescimento consistente ao longo do período de backtest. No entanto, a inclusão de fatores aleatórios nos preços pode levar a variações significativas na curva de capital, com períodos de lucro seguidos de períodos de prejuízo.

Essas variações na curva de capital ressaltam a importância da simulação de preços aleatórios na avaliação da robustez da estratégia de Pairs Trading. Ao considerar diferentes cenários de mercado, é possível identificar possíveis pontos de fragilidade na estratégia e buscar aprimoramentos que aumentem sua rentabilidade e robustez.

Embora os preços com fatores aleatórios tenham sido inseridos nas simulações, o modelo de Pairs Trading apresentou um crescimento no capital em todas as 100 simulações realizadas. Esse resultado indica a robustez da estratégia diante de variações de preço e sugere que a estratégia pode ser eficaz mesmo em cenários de mercado desafiadores.

4.3. BENCHMARK

Os resultados apresentados na figura 14 e na tabela 2 são oriundos de cálculos realizados pela biblioteca Pyfolio de Python, que é uma ferramenta comumente utilizada para avaliação de estratégias de negociação no mercado financeiro. A estratégia de Pairs Trading baseada em Aprendizado de Máquina descrita por Sarmento e Horta (2020) foi testada utilizando essa ferramenta em um período de backtest de 2005 a 2022, totalizando 98 meses.

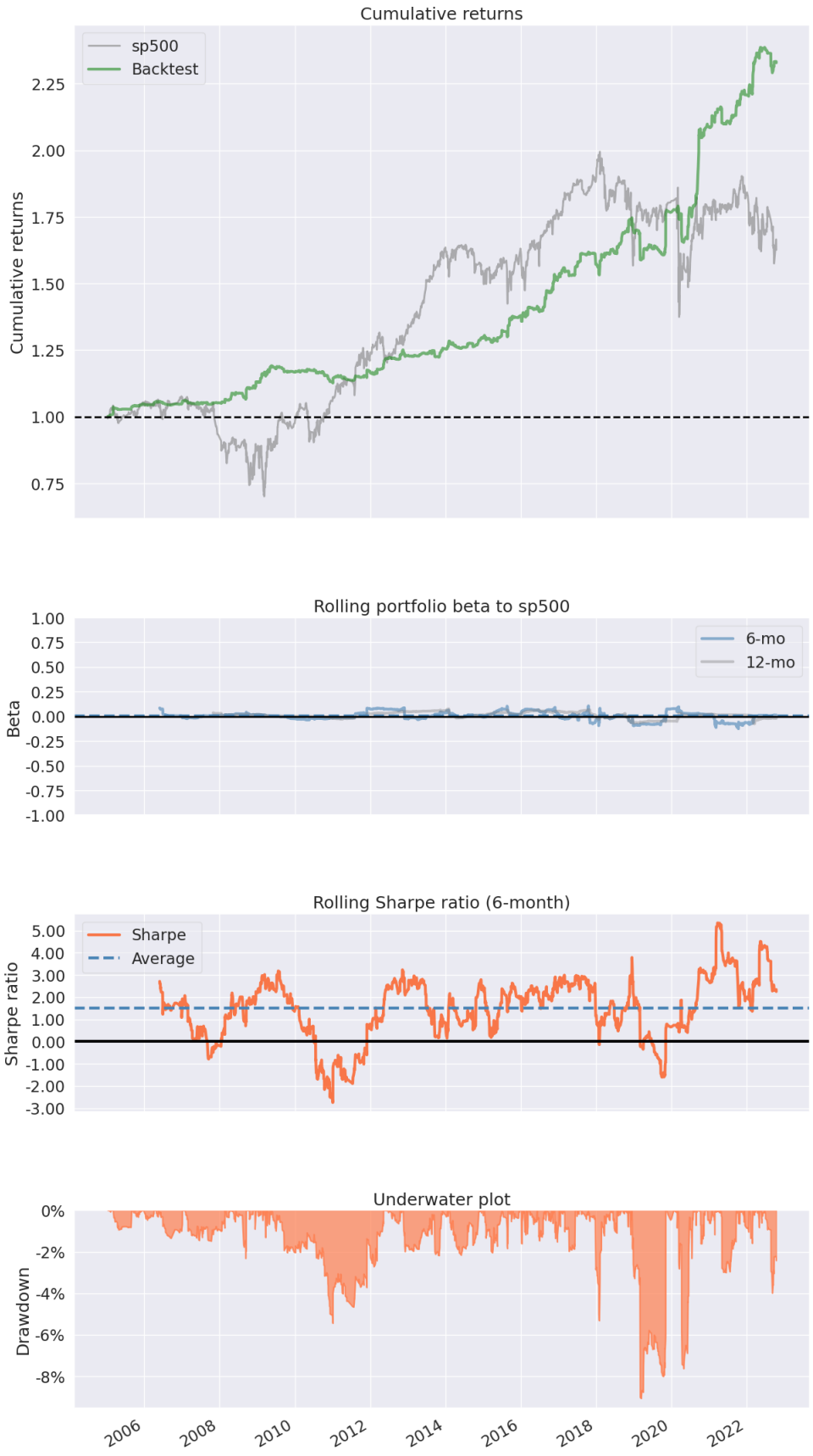


Fig. 14 – Análise utilizando pyfolio.
(FONTE: Elaborado pelo autor)

Start date	11/01/2005
End date	14/10/2022
Total months	98
Backtest	
Annual return	10.8%
Cumulative returns	132.8%
Annual volatility	6.9%
Sharpe ratio	1.53
Calmar ratio	1.20
Stability	0.92
Max drawdown	-9.1%
Omega ratio	1.60
Sortino ratio	2.70
Skew	-
Kurtosis	-
Tail ratio	1.92
Daily value at risk	-0.8%
Alpha	0.11
Beta	0.01

Tabela 2 – Análise utilizando pyfolio.
(FONTE: Elaborado pelo autor)

Os resultados indicam que a estratégia de Pairs Trading apresentou um desempenho superior ao índice de referência S&P500. O retorno anual foi de 10,8% e o retorno cumulativo foi de 132,8%, o que sugere uma performance acima da média do mercado. Além disso, o índice de Sharpe de 1,53 indica que a estratégia é capaz de gerar retornos ajustados ao risco atraentes, considerando a volatilidade anual de 6,9%. O Calmar ratio de 1,20 sugere que a estratégia é capaz de gerar retornos atraentes sem sofrer grandes drawdowns, enquanto o Omega ratio de 1,60 indica que a estratégia foi capaz de gerar retornos ajustados ao risco superiores ao índice de referência.

4.4. DISCUSSÃO

Os resultados obtidos são promissores e sugerem que a estratégia de Pairs Trading baseada em Aprendizado de Máquina pode ser uma opção interessante para investidores que buscam retornos acima da média do mercado. No entanto, é importante ressaltar que os resultados foram obtidos em um período de backtest e que a estratégia pode não ter o mesmo desempenho em um ambiente de negociação real.

Na implementação prática da estratégia é importante considerar alguns riscos operacionais que podem afetar a rentabilidade da estratégia. Um dos principais riscos é a incerteza em relação à execução das ordens de compra e venda, que pode afetar a efetividade

da estratégia. Além disso, a seleção de pares de ativos inadequados ou a utilização de parâmetros incorretos pode levar a resultados insatisfatórios.

Outro aspecto importante a ser considerado é o tempo necessário para a implementação da estratégia. A estratégia de Pairs Trading requer uma monitoração constante dos pares de ativos selecionados, a fim de identificar oportunidades de negociação e ajustar os parâmetros da estratégia. Isso pode requerer um investimento significativo de tempo e recursos, o que pode afetar a viabilidade da estratégia para alguns investidores.

Além disso, é importante destacar que os custos associados à negociação, como as taxas de transação e os custos de corretagem, podem afetar significativamente a rentabilidade da estratégia. É importante levar em consideração que os custos podem variar de acordo com a exchange utilizada, o que pode afetar a viabilidade da estratégia para alguns investidores.

Por fim, é importante destacar que o valor excedente ao que estava em exposição na operação poderia ter sido alocado em um ativo livre de risco, que poderia ter aumentado a rentabilidade da estratégia.

5. CONCLUSÃO

O objetivo deste artigo foi avaliar a eficácia do modelo proposto por Sarmento e Horta (2020) para solucionar o problema de alta dimensionalidade dos dados e a clusterização de pares. Este estudo apontou para a necessidade de uma representação mais precisa dos dados, sem a necessidade de definir manualmente os grupos aos quais cada ação deve pertencer, através de técnicas de agrupamento automatizadas. Para alcançar esse objetivo, a metodologia proposta neste artigo incluiu a redução de dimensionalidade para encontrar uma representação compacta para cada ação, aprendizado não supervisionado para definir potenciais clusters e seleção e definição de um conjunto de regras para selecionar pares para negociação. Além disso, este artigo também comparou o modelo proposto com o índice S&P 500 através de backtesting da estratégia, verificando se o modelo mostra-se consistente ao longo do tempo.

Os resultados mostraram que o modelo proposto apresentou um crescimento constante no capital em todas as 100 simulações realizadas durante 17 anos, mesmo com a inserção de valores aleatórios nos preços das ações.

É importante destacar que há riscos operacionais e o tempo necessário despendido para a implementação de tais modelos. Além disso, as fees e brokerage podem mudar de acordo com

a exchange e isso deve ser levado em consideração. Outro ponto a ser considerado é que o valor excedente ao que estava em exposição na operação poderia ter sido alocado em um ativo livre de risco e ter aumentado a rentabilidade. Portanto, é necessário avaliar cuidadosamente os riscos e benefícios envolvidos antes de implementar uma estratégia de Pairs Trading baseada em Machine Learning.

Em conclusão, o modelo proposto por Sarmento e Horta (2020) apresenta uma alternativa promissora para investidores interessados em Pairs Trading baseado em Machine Learning. Como sugestão para trabalhos futuros seria a incorporação da mensuração diária da variação dos preços, chamado mark to market (marcação a mercado). Isso permitiria uma leitura mais precisa sobre a real volatilidade das operações, bem como forneceria uma melhor gestão de risco e um monitoramento mais próximo do desempenho da estratégia. Além disso, essa análise diária poderia ser utilizada para realizar ajustes nos parâmetros do modelo e otimizar ainda mais a rentabilidade da estratégia.

REFERÊNCIAS

ABDI, H., WILLIAMS, L. J., & VALENTIN, D. (2020). Multiple factor analysis: An introduction and overview. Statistical Analysis and Data Mining: The ASA Data Science Journal

AÏT-SAHALIA, Y.; LO, A. Nonparametric Pricing of Contingent Claims. Journal of Finance, v. 47, n. 4, 2002.

ALBERGARIA, V., GIRALDI, J. M. E., & COSTA, R. M. (2020). A Machine Learning Algorithm for Stock Market Prediction: The Brazilian Case. Journal of Risk and Financial Management

ALEXANDER, S. S. Price movements in speculative markets: Trends or random walks. Industrial Management Review, v. 2, 1961.

ANKERST, M *et al*. OPTICS: ordering points to identify the clustering structure. Association for Computing Machinery, 1999.

AVELLANEDA, M.; LEE, J. Statistical arbitrage in the U.S. equities markets. Working papers, SSRN, 2008.

AVELLANEDA, M; LEE, JH. Statistical arbitrage in the US equities market. SSRN, 2010.

BAESSO, R *et al*. Teste da Hipótese de Eficiência do Mercado no Brasil: uma aplicação de filtros ótimos. XXXII ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PÓS-GRADUAÇÃO E PESQUISA EM ADMINISTRAÇÃO – ENANPAD., Rio de Janeiro, 2008.

BELLMAN, R. Dynamic programming. Science, v. 153, 1966.

BERKHIN, P. A Survey of Clustering Data Mining Techniques. Springer, 2006.

CALDEIRA, J. F. Arbitragem Estatística e Estratégia Long-Short Pairs Trading, Abordagem da Cointegração Aplicada a Dados do Mercado Brasileiro. UFRGS, 2010.

CALDEIRA, J; MOURA, G. Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy. SSRN. 28 p, 5 jan 2013.

CAMPBELL, M; HOANE JR., A; HSU, F. Deep Blue. Artificial Intelligence, v. 134, 2002.

CARVALHO, F *et al*. Economia monetária e financeira: teoria e política, f. 193. 2000.

CAVALCANTE, RC *et al*. Computational intelligence and financial markets: a survey and future directions. Expert Syst, 2016.

CHAN, Ernie. Algorithmic Trading: Winning Strategies and Their Rationale. John Wiley & Sons, v. 1, f. 112, 2013. 224 p.

CHAN, N., Goh, J., & Koh, S. (1996). Profitability of momentum strategies in the Asian stock markets. Pacific-Basin Finance Journal, 4(1), 53-68

CHAN, N., GAO, J., & WONG, W. (2013). Algorithmic trading: a review of current research. Pacific-Basin Finance Journal, 25.

DO, B; FAFF, R. Are pairs trading profits robust to trading costs?. J Financ Res, v. 35, 2012.

DUNIS, CL *et al*. Trading and hedging the corn/ethanol crush spread using time-varying leverage and nonlinear models. Eur J Financ, 2015.

DUNIS, CL; LAWS, J; EVANS, B. Modelling and trading the gasoline crack spread: a non-linear story. Deriv Use Trading, 2006.

DUNIS, CL; LAWS, J; EVANS, B. Modelling and trading the soybean-oil crush spread with recurrent and higher order networks: a comparative analysis. Artificial higher order neural networks for economics and business, IGI Global, 2009.

ERTÖZ, L; STEINBACH, M; KUMAR, V. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. Conference: Proceedings of the Third SIAM International Conference on Data Mining, 2003.

ESTER, M *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI: American Association for Artificial Intelligence, 1996.

FAMA, E. Efficient capital markets: a review of theory and empirical work. The Journal of Finance, Cambridge, v. 25, 1970.

FAMA, E. Efficient capital markets: II. The Journal of Finance, Cambridge, v. 46, 1991.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, v. 17, 1996. https://doi.org/10.1609/aimag.v17i3.1230.

FORTUNA, E. Mercado financeiro: produtos e serviços. 20 ed. 2015. 1100 p.

GATEV, E; GOETZMANN, WN; ROUWENHORST, KG. Pairs trading: performance of a relativevalue arbitrage rule. Rev Financ Stud, v. 19, 2006.

GRANTER, S; BECK, A; PAPKE, D. AlphaGo, Deep Learning, and the Future of the Human Microscopist. College of American Pathologists, 2017.

HAN, J; PEI, J; KAMBER, M. Data Mining: Concepts and Techniques. 2 ed. Elsevier, v. 3, f. 372, 2011. 744 p.

JENSEN, M. C. Some anomalous evidence regarding market efficiency. Journal of financial economics, v. 6, 1978.

JOLLIFFE, I. Principal component analysis. Springer, Berlin, 2011.

KOCH, C. How the Computer Beat the Go Master. Scientific American. 2016. Disponível em: https://www.scientificamerican.com/article/how-the-computer-beat-the-go-master/. Acesso em: 5 jul. 2022.

LO, A. W., et al. Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test. Review of Financial Studies, vol. 3, n. 1, 1990.

LONGERSTAEY, R., SOLNIK, B. International value and growth stock returns. Journal of finance, vol. 49, n. 5, p. 1639-1682, 1994.

MALKIEL, B. G. A Random Walk Down Wall Street. Norton, 2003.

MIAO, K; CHEN, F; ZHAO, Z.-g. Stock price forecast based on bacterial colony 546 RBF neural network. Journal of Qingdao University (Natural Science Edition), v. 2, 2007.

NETO, A. Mercado financeiro. 14 ed. 2018. 513 p.

PETERS, E. E. Chaos and order in the capital markets: a new view of cycles, prices, and market volatility. John Wiley & Sons, v. 1, 1996.

PONTES, R. Inteligência Artificial Nos Investimentos. Clube de Autores, v. 1, f. 65, 2011. 130 p.

ROBERTS, H. Statistical versus clinical prediction of the stock market. Unspublished Work presented in the Conference of Securities Price Analysis, Chicago, 1967.

SARMENTO, S; HORTA, N. A Machine Learning based Pairs Trading Investment Strategy. Springer Nature, v. 3, f. 52, 2020. 104 p.

SILVER, D *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, v. 529, 2016.

TAYLOR, S. J. Modelling financial time series. World Scientific, 2008.