

PortFlow Invest: Uma ferramenta para análise e suporte de decisões para o mercado financeiro

Lucas A. S. Gonçalves, Michele A. Brandão, Ozório J. S. Camargos e Adriano C. M. Pereira

Departamento de Ciência da Computação

Universidade Federal de Minas Gerais

Belo Horizonte, Brasil

Resumo—Este trabalho apresenta o desenvolvimento do PortFlow Invest, um sistema de suporte à decisão para a diversificação de carteiras na B3 que integra análise de sensibilidade macroeconômica e agrupamento temporal de ativos. Orientado pelo processo de Extração de Conhecimento em Bases de Dados (KDD), o sistema compreende desde a engenharia e consolidação de dados até uma modelagem dinâmica capaz de capturar quebras estruturais na relação entre ativos financeiros e indicadores macroeconômicos brasileiros (IMecs). A metodologia proposta engloba: (i) modelagem de correlações móveis por janelas deslizantes; (ii) aplicação de algoritmos de clusterização temporal shape-based (Dynamic Time Warping - DTW e k-Shape) sobre trajetórias de retornos e assinaturas de sensibilidade; (iii) estruturação de um framework de avaliação baseado na métrica *Silhouette Score*; e (iv) implementação de um dashboard interativo em Streamlit para auxílio ao investidor.

Index Terms—Mercado financeiro, B3, séries temporais financeiras, clusterização de séries temporais, Dynamic Time Warping, k-Shape, correlações móveis, sensibilidade macroeconômica, diversificação de portfólio, KDD, Streamlit.

I. INTRODUÇÃO

O mercado de capitais brasileiro, operado pela B3, tem passado por uma transformação estrutural expressiva, caracterizada pela democratização do acesso a investimentos de renda variável. Dados da própria instituição indicam que a participação de investidores individuais atingiu cerca de 25% do volume total negociado em ações. Esse influxo massivo de capital de pessoas físicas ocorre em um ecossistema macroeconômico historicamente volátil, marcado por sensibilidades fiscais, oscilações na taxa básica de juros (Selic) e acentuada instabilidade cambial. Nesse panorama, a diversificação de portfólio — princípio basilar da Teoria Moderna de Portfólio — consolida-se como mecanismo indispensável para a mitigação de riscos sistêmicos.

Contudo, investidores não institucionais carecem de instrumental analítico acessível para identificar interdependências complexas entre ativos financeiros e a conjuntura macroeconômica. A diversificação ingênua, fundamentada exclusivamente na classificação do setor nominal declarado das empresas, mostra-se insuficiente em momentos de estresse financeiro, período em que as correlações tendem a convergir e a dissipar o efeito de proteção esperado da carteira.

Para mitigar essa vulnerabilidade, o sistema *PortFlow Invest* propõe um modelo de diversificação orientado pela sensibilidade macroeconômica dinâmica dos ativos. A plataforma mensura como os retornos de cada ativo reagem a quatro

Indicadores Macroeconômicos (IMecs) de referência: inflação (IPCA), câmbio (Dólar), atividade econômica (IBC-Br) e política monetária (Selic). Esse paradigma baseia-se no fato de que ativos de setores distintos podem manifestar perfis de resposta idênticos a choques macroeconômicos. Tornar essas assinaturas de sensibilidade explícitas e visualizáveis constitui o núcleo desta solução.

A. A limitação da modelagem estática

Embora análises de correlação tradicionais ofereçam um panorama geral do mercado, abordagens puramente estáticas falham ao assumir que as relações financeiras são imutáveis no tempo. Calcular coeficientes fixos para horizontes extensos desconsidera fenômenos críticos como quebras estruturais e mudanças de regime econômico. A sensibilidade de uma empresa exportadora em relação à moeda estrangeira, por exemplo, altera-se dinamicamente sob diferentes condições de mercado. Consequentemente, superar o reducionismo estático por meio de agrupamentos baseados na evolução temporal dessas relações configura o desafio científico central deste estudo.

B. Objetivos e contribuições

O presente trabalho tem como objetivo desenvolver o *PortFlow Invest* como uma ferramenta interativa e exploratória de suporte à decisão e apoio à diversificação de portfólios. As contribuições técnico-científicas específicas desta pesquisa compreendem:

- 1) **Modelagem dinâmica de dependência Estatística:** Implementação de um ecossistema analítico baseado em janelas deslizantes para avaliar o comportamento temporal de ativos frente à economia.
- 2) **Agrupamento temporal baseado em forma (*Shape-Based*):** Aplicação de algoritmos focados na morfologia de séries temporais para agrupar ativos pela similaridade de suas trajetórias históricas e assinaturas de sensibilidade.
- 3) **Framework de validação comparável:** Estruturação de uma esteira de avaliação baseada no *Silhouette Score*, viabilizando o contraste quantitativo direto entre diferentes algoritmos de distâncias morfológicas.
- 4) **Plataforma de visualização analítica:** Desenvolvimento de um *dashboard* interativo focado na usabili-

dade, traduzindo topologias matemáticas complexas em recursos visuais interpretáveis.

O restante deste artigo está estruturado da seguinte forma: a Seção II revisita o referencial teórico de finanças e séries temporais; a Seção III detalha a metodologia de mineração de dados proposta; a Seção IV apresenta e discute os resultados empíricos obtidos; e a Seção V encerra o trabalho com as conclusões.

II. TRABALHOS CORRELATOS E REFERENCIAL TEÓRICO

A fundamentação teórica que sustenta o desenvolvimento deste trabalho estrutura-se em três eixos principais: os fundamentos econômicos clássicos de seleção de carteiras; a interseção entre finanças quantitativas e métodos computacionais; e os desenvolvimentos em mineração de dados voltados à clusterização de séries temporais.

A. Teoria de Portfólio e Precificação de Ativos

Os alicerces matemáticos para a diversificação moderna foram estabelecidos por Markowitz [1], que formalizou a seleção de portfólios demonstrando que o risco agregado de uma carteira é condicionado pelas covariâncias mútuas entre os ativos, e não apenas pela soma de suas variâncias. Esse princípio justifica o uso de matrizes de dependência estatística e agrupamentos matemáticos para mitigar o risco não sistemático.

Expandindo a dinâmica de alocação, Sharpe [2] introduziu o *Capital Asset Pricing Model* (CAPM), postulando que o retorno esperado de um ativo guarda uma relação proporcional com o seu risco sistemático (β). A abordagem arquitetada no presente estudo estende essa formulação: em vez de estimar um único parâmetro contra um índice de mercado agregado, mapeia-se um vetor multidimensional de sensibilidade dinâmica face a múltiplos indicadores econômicos.

A convergência original entre a teoria financeira e os métodos de agrupamento foi proposta por Mantegna [3], que introduziu a clusterização baseada em correlação (*correlation-based clustering*) para decodificar a taxonomia latente em mercados de capitais, validando a premissa de que a similaridade matemática é capaz de isolar comportamentos de risco homólogos.

B. Métodos computacionais aplicados a finanças

A literatura contemporânea evidencia o protagonismo de arquiteturas computacionais no tratamento de séries financeiras complexas. Kim e Won [5] desenvolveram estruturas híbridas que integram modelos econométricos tradicionais (GARCH) a Redes Neurais Recorrentes (LSTM) para a modelagem de volatilidade. Os autores demonstraram que mecanismos não lineares possuem capacidade superior para assimilar nuances, quebras estruturais e caudas pesadas que violam as premissas estatísticas clássicas. Tal arcabouço legitima o uso de algoritmos de mineração de dados para o processamento escalável e a extração de padrões robustos em ambientes voláteis.

C. Clusterização de Séries Temporais

A decisão metodológica central desta pesquisa consiste em modelar o agrupamento de ativos rejeitando o reducionismo estático de período único. Adota-se a taxonomia formal de Paparrizos et al. [6], que categoriza os métodos de agrupamento sequencial. Esta pesquisa posiciona-se na família dos métodos baseados em distância (*distance-based*), mobilizando algoritmos morfológicos (*shape-based*) complementares:

- **Dynamic Time Warping (DTW) e Soft-DTW:** O algoritmo DTW mensura a similaridade entre duas sequências realizando um alinhamento elástico no eixo do tempo. Essa mecânica permite que séries com morfologias semelhantes, porém defasadas cronologicamente, sejam identificadas como pertencentes a um mesmo domínio. Tal propriedade é essencial para o mercado, onde repasses econômicos ocorrem com atrasos distintos. Adicionalmente, explora-se a variante *Soft-DTW*, que substitui a métrica rígida por uma formulação matemática suavizada e diferenciável. Essa abordagem confere maior estabilidade à formação dos centroides frente aos ruídos estocásticos típicos de séries financeiras de alta volatilidade.
- **k-Shape:** Desenvolvido por Paparrizos e Gravano [4], o *k-Shape* é um método que se baseia em uma distância computada a partir da correlação cruzada normalizada. O algoritmo apresenta como propriedade fundamental a invariância a deslocamentos de fase e variações de amplitude. Contudo, devido ao uso interno de Transformadas Rápidas de Fourier (FFT), o método requer sequências temporais de dimensões idênticas, exigindo estratégias rigorosas de janelamento de dados históricos.

A coexistência dessas técnicas viabiliza um contraste empírico relevante: enquanto o DTW tolera distorções temporais locais, o *k-Shape* foca estritamente na similaridade geométrica global das curvas.

D. Avaliação de agrupamentos temporais

A validação das estruturas topológicas geradas apoia-se no índice de silhueta (*Silhouette Score*), que equaciona os conceitos de coesão intra-cluster e separação inter-cluster. No contexto deste trabalho, o cálculo matemático foi adaptado para ser processado diretamente sobre as matrizes intrínsecas de cada método: a matriz de distâncias acumuladas do DTW e a *Shape-Based Distance* (SBD) do *k-Shape*. A normalização de ambas para uma escala homogênea viabiliza o julgamento quantitativo comparável entre os modelos.

III. METODOLOGIA E DESENVOLVIMENTO

A metodologia desta pesquisa foi estruturada com base no processo de Extração de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* — KDD). O fluxo de trabalho inicia-se com uma caracterização estatística da base histórica, que justifica as principais escolhas de modelagem do estudo: o uso de correlações dinâmicas e o conceito de *assinatura macroeconômica*. Em seguida, o desenvolvimento avança pelas etapas tradicionais do KDD: Seleção,

Pré-processamento, Transformação, Mineração, Avaliação e Apresentação.

Para suportar esse processo, o sistema foi projetado sob uma arquitetura modular, separando a extração de dados, o cálculo matemático de correlações, os algoritmos de agrupamento e a interface de visualização. O diagrama completo dessa estrutura lógica está sintetizado na Fig. 1.

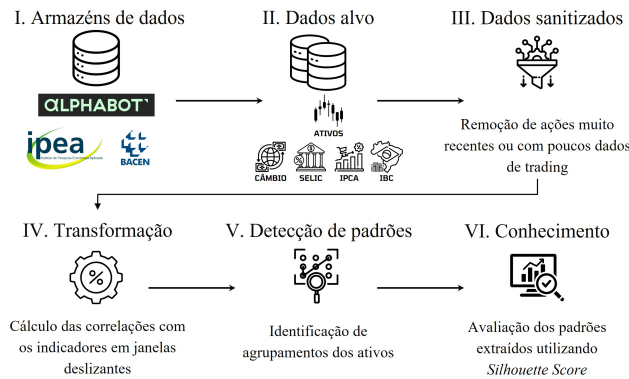


Figura 1. Pipeline metodológico do PortFlow Invest estruturado segundo as etapas do processo KDD e as camadas da biblioteca modular.

A. Fundamentação empírica: caracterização da base

A análise prévia dos dados forneceu as evidências necessárias para guiar a modelagem. Destacam-se as seguintes observações de mercado que fundamentaram as decisões arquiteturais:

Não-normalidade e caudas pesadas: Os retornos dos ativos analisados mostraram desvio em relação à distribuição Normal. Observou-se a presença de eventos extremos, com quedas abruptas de até -79% e valorizações súbitas de $+96\%$ em ativos específicos, que ocorrem com frequência superior à prevista por modelos tradicionais. Esse comportamento justifica o uso do coeficiente de correlação de Spearman, que é não-paramétrico e mais robusto a valores extremos, em detrimento do coeficiente de Pearson. Adicionalmente, validou-se o uso da correlação de distância (*distance correlation*) para capturar relações não-lineares.

Volatilidade macroeconômica: O período estudado foi marcado por forte instabilidade econômica. A taxa Selic variou de 2% a 15% ao ano, enquanto o Dólar apresentou alta volatilidade. Essa exposição a choques sistêmicos reforça a necessidade de transformar todas as séries para formatos estacionários (retornos e variações percentuais), permitindo uma análise mais fiel da realidade.

Paradigma da assinatura macroeconômica: A análise revelou que o Dólar atua como um fator de precificação dominante e polarizado. Enquanto o mercado em geral apresenta forte aversão a altas cambiais (correlação negativa), empresas exportadoras descolam-se desse padrão, beneficiando-se da desvalorização do real. Movimentos semelhantes ocorrem no setor varejista frente a choques na Selic e no IPCA. O perfil único de resposta de cada ativo a esses indicadores forma a

sua assinatura macroeconômica, exemplificada pela Fig. 3 e formalizada na Seção III-D.

Limitação da abordagem estática: Modelos de correlação calculados sobre um período único mascaram mudanças de regime. O ativo VALE3, por exemplo, apresentou fortes inversões em sua correlação com o Dólar entre 2015 e 2024 (Fig. 2). Essa limitação torna essencial a transição de uma análise estática para uma modelagem dinâmica baseada em janelas móveis (*rolling windows*).

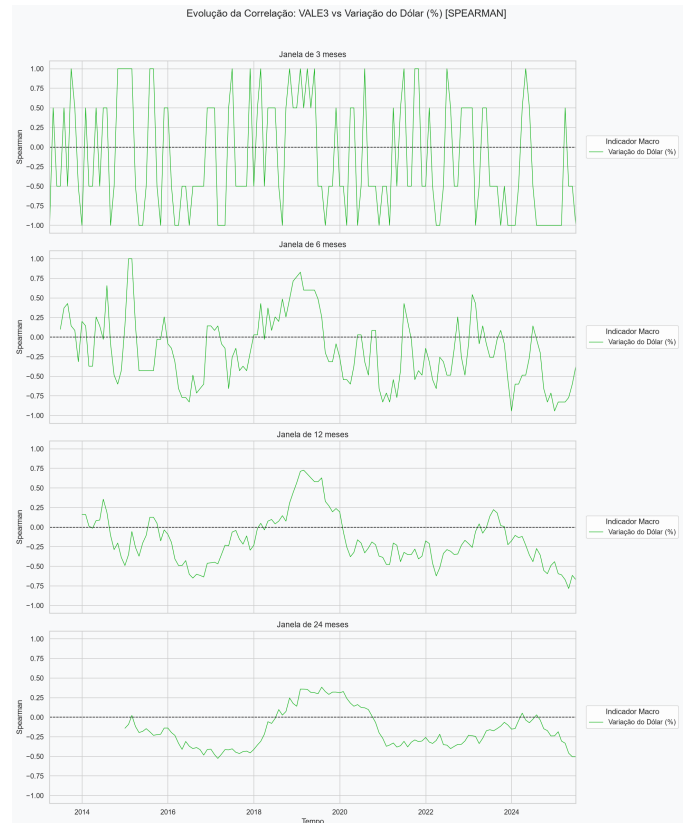


Figura 2. Evolução da correlação móvel (VALE3 $\times$ Dólar) sob coeficientes de Spearman segmentados em janelas de 3, 6, 12 e 24 meses. O contraste ilustra o *trade-off* estatístico entre responsividade temporal e ruído estocástico comparado à observação de transições estruturais e de mudança de regimes através de janelas longas e suavizadas.

B. Seleção de dados

Esta etapa deriva os *Dados alvo* (II) a partir dos *Armazéns de dados* (I) do pipeline da Fig. 1, combinando preços históricos de ações da B3 (amostrados a cada 15 minutos) com indicadores econômicos emitidos por órgãos oficiais, incluindo o IPCA, a Selic, a cotação do Dólar e o índice IBC-Br.

Para evitar o viés de sobrevivência e distorções causadas por ativos de baixa liquidez, aplicou-se um filtro estrito: foram selecionados apenas ativos com histórico de negociação de pelo menos quatro anos consistentes ou um mínimo de 10.000 registros intradiários. O resultado foi um painel consolidado com 85 ativos, cobrindo 150 meses entre janeiro de 2013 e junho de 2025.

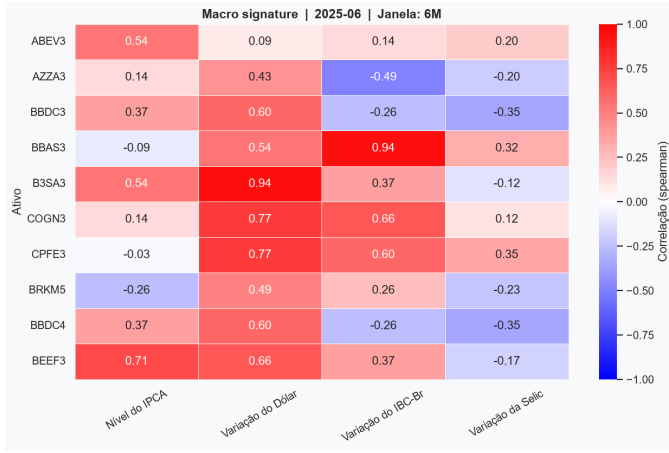


Figura 3. Exemplificação em Heatmap de matriz (Ativo $\times$ IMec) da assinatura macroeconômica para uma data de referência fixa. O mapa identifica similaridades, mensuradas por Spearman, entre papéis isolados da B3 em relação aos indicadores selecionados num período de 6 meses.

C. Pré-processamento e transformação

Os dados intradiários foram agregados para uma periodicidade mensal, adotando-se o fechamento do último dia útil de cada mês. Os preços das ações foram convertidos em retornos geométricos, e os índices econômicos foram transformados em variações percentuais para garantir a estacionaridade das séries.

O foco central desta etapa é o cálculo das matrizes de correlação móveis, utilizando janelas de 3, 6, 12 e 24 meses. Para lidar com meses em que taxas (como a Selic) se mantêm inalteradas, adotou-se o preenchimento inercial dos valores anteriores, garantindo a continuidade da série sem vaziar informações futuras. Além disso, introduziu-se um parâmetro de defasagem temporal (*lag*) para capturar possíveis atrasos entre um choque econômico e o reflexo nos preços das ações.

Dado o elevado custo computacional de calcular correlações dinâmicas cruzadas ao longo do tempo, implementou-se um sistema robusto de *cache*, que salva resultados intermediários em disco e otimiza significativamente o tempo de resposta da aplicação final.

D. Formalização da assinatura macroeconômica

A assinatura macroeconômica é o objeto matemático central deste trabalho e atua como o principal descritor levado à etapa de mineração. Os algoritmos de agrupamento operam sobre dois objetos complementares: a própria assinatura macroeconômica, definida a seguir, e a trajetória de valor de cada ativo, reconstruída a partir de seus retornos, que será formalizada na seção seguinte. A assinatura capta a sensibilidade aos indicadores macroeconômicos, enquanto a trajetória de valor serve como objeto de contraste baseado na forma para validar os agrupamentos de maneira independente. Em termos formais, a assinatura traduz em um vetor de baixa dimensão a forma como os retornos de um ativo covariam com cada indicador em um dado horizonte temporal. Essa abordagem substitui a noção de um vetor de sensibilidade estático, no qual cada ativo

é descrito por coeficientes de uma regressão linear múltipla, por um descritor dinâmico baseado em correlação móvel não paramétrica.

Sejam $P_{i,t}$ o preço de fechamento mensal do ativo i no mês t e $R_{i,t}$ o seu retorno mensal simples, calculado para normalizar os dados em uma base percentual comparável com os indicadores macroeconômicos:

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

Denota-se por $X_{m,t}$ o valor do indicador macroeconômico m . As quatro séries macroeconômicas são analisadas em variação, conforme justificado na caracterização da base: a Selic como variação mensal em pontos percentuais, o Dólar e o IBC-Br como variação percentual mensal, e o IPCA como a taxa mensal de inflação.

Define-se a correlação móvel entre o ativo i e o indicador m , sob a janela de w meses encerrada em t , como o coeficiente de correlação calculado sobre as w observações mais recentes:

$$\rho_{i,m}^{(w)}(t) = \text{Corr}\left(\{R_{i,\tau}\}_{\tau \in \mathcal{W}(t,w)}, \{X_{m,\tau}\}_{\tau \in \mathcal{W}(t,w)}\right), \quad (2)$$

$$\mathcal{W}(t,w) = \{t-w+1, \dots, t\}.$$

A medida de correlação pode ser escolhida entre três alternativas: Pearson, Spearman e correlação de distância (distance correlation). O estudo utiliza as duas últimas devido às características empíricas dos dados de mercado. O coeficiente de Spearman confere imunidade contra o efeito de caudas pesadas e desvios de normalidade típicos dos retornos financeiros, enquanto a correlação de distância diferencia-se por mensurar associações não lineares arbitrárias entre os ativos e o ambiente macroeconômico, o que a permite revelar interdependências estruturais que seriam ignoradas pelas abordagens lineares puras.

A assinatura macroeconômica do ativo i no instante t , considerando a janela w , é o vetor que compila as quatro correlações móveis, sendo uma para cada indicador:

$$\mathbf{S}_i^{(w)}(t) = \left(\rho_{i,\text{IPCA}}^{(w)}(t), \rho_{i,\text{USD}}^{(w)}(t), \rho_{i,\text{IBC-Br}}^{(w)}(t), \rho_{i,\text{Selic}}^{(w)}(t)\right),$$

$$\mathbf{S}_i^{(w)}(t) \in [-1, 1]^4 \quad (3)$$

Cada coordenada situa-se no intervalo de -1 a 1, ou de 0 a 1 no caso da correlação de distância. A assinatura representa um ponto no espaço quadridimensional que varia com o tempo. Ao fixar o ativo i e variar t , obtém-se a trajetória da sensibilidade do ativo. A mineração opera sobre essa trajetória em relação a um indicador guia, que descreve a evolução da sensibilidade a um choque específico.

Por fim, nos casos em que um indicador se mantém inalterado dentro de uma janela, a variância nula inviabiliza o cálculo da correlação. Nessas situações, preserva-se a última sensibilidade válida por meio de um preenchimento inercial por ativo ao longo do tempo, evitando a interferência de dados entre diferentes papéis.

E. Mineração de dados

A fase de mineração corresponde à *Deteção de padrões* (V) da Fig. 1 e busca identificar padrões ao agrupar os ativos de acordo com seu comportamento. Foram aplicados dois algoritmos de clusterização projetados especificamente para lidar com séries temporais, focando na forma geométrica das curvas (*shape-based*):

- 1) **Dynamic Time Warping (DTW) e Soft-DTW:** O algoritmo DTW alinha sequências de forma elástica, sendo capaz de ignorar pequenas defasagens ou assimetrias temporais entre os ativos. O agrupamento é feito iterativamente e o método acomoda perfeitamente painéis desbalanceados (ativos com diferentes tempos de vida na bolsa). O algoritmo foi utilizado para agrupar tanto as trajetórias brutas de retornos quanto a evolução das assinaturas de sensibilidade ao longo do tempo. Como refinamento à modelagem, o sistema incorpora o uso do *Soft-DTW*, que atua como um mecanismo de suavização das distâncias elásticas, otimizando a estabilidade e a convergência dos clusters em ambientes de alta volatilidade estrutural.
- 2) **k-Shape:** Este algoritmo agrupa os ativos utilizando a correlação cruzada normalizada, tornando os grupos invariantes a oscilações de fase e amplitude. Por exigir séries de tamanho idêntico, a aplicação do *k-Shape* operou sobre um recorte padronizado dos dados recentes de cada ativo, permitindo uma comparação estruturada com os resultados obtidos pelo DTW.

Em paralelo à análise da assinatura, a fase de mineração também agrupa a trajetória de valor de cada ativo com base em sua forma. Como a base consolidada armazena exclusivamente retornos mensais, essa trajetória deve ser reconstruída por meio de capitalização composta. Considerando $r_{i,t} = R_{i,t}/100$ o retorno simples em formato decimal, define-se o índice de valor de base 100 no início do período t_0 como o produto acumulado dos fatores de retorno bruto:

$$\tilde{P}_{i,t} = 100 \prod_{\tau=t_0+1}^t (1 + r_{i,\tau}), \quad \tilde{P}_{i,t_0} = 100. \quad (4)$$

Essa reconstrução preserva o formato da série, o que é essencial para o agrupamento baseado na geometria das curvas. O índice reconstruído corresponde ao preço real a menos de uma constante multiplicativa, que representa o preço inicial normalizado. Essa constante é eliminada na normalização prévia ao alinhamento elástico, etapa na qual as séries são padronizadas para se tornarem invariantes à amplitude. Desse modo, o procedimento recupera o formato exato da trajetória de preço, e o nível absoluto é descartado intencionalmente no espaço em que a clusterização opera.

F. Avaliação do conhecimento

A qualidade de separação e coesão dos grupos formados foi medida matematicamente por meio do Coeficiente de Silhueta (*Silhouette Score*). Para que a métrica fosse justa e condizente

com a natureza temporal dos algoritmos, o cálculo não utilizou distâncias puramente lineares, mas foi processado diretamente sobre as matrizes geradas pelos próprios algoritmos: a matriz de distâncias acumuladas para o DTW e a matriz de distâncias baseadas em forma (*Shape-Based Distance*) para o *k-Shape*. Ambas foram normalizadas para o intervalo de -1 a 1 , viabilizando a comparação direta de desempenho entre os modelos. O número de grupos k foi selecionado por varredura do parâmetro, combinando o ponto de inflexão da curva de silhueta com um critério de não degeneração da partição; o detalhamento empírico dessa seleção encontra-se na Seção IV-B.

G. Apresentação do conhecimento

Os modelos matemáticos desenvolvidos foram integrados a um *dashboard* web interativo utilizando a biblioteca Streamlit. O objetivo é traduzir cálculos complexos em uma ferramenta de usabilidade intuitiva para o gerenciamento de portfólios. A interface está dividida em cinco módulos analíticos:

- **Visão geral:** Consolidação e panorama estatístico do conjunto de dados.
- **Correlações móveis:** Exploração visual do comportamento das dependências entre ativos e indicadores ao longo do tempo, com suporte ao teste de defasagens (*lags*).
- **Assinatura Macro:** Representação da sensibilidade econômica dos ativos por meio de mapas de calor tridimensionais.
- **Similaridade entre ativos:** Ferramenta ágil para identificar pares de ativos com comportamento análogo.
- **Clusterização Temporal:** Ambiente interativo final em que o usuário pode alternar algoritmos (DTW e *k-Shape*) e observar os agrupamentos topológicos sendo formados com base no perfil comportamental dinâmico dos papéis da B3.

IV. RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos a partir das clusterizações temporais baseadas em forma (*shape-based*) aplicadas ao universo de 85 ativos da B3, englobando o período de janeiro de 2013 a junho de 2025. Para garantir a robustez da análise e evitar o viés de uma configuração única, conduziu-se uma varredura sistemática composta por 510 execuções. Esse procedimento cruzou diferentes algoritmos (DTW e *k-Shape*), objetos de estudo (a trajetória bruta de preços versus a evolução da assinatura de sensibilidade ao Dólar), recortes temporais (épocas recentes e antigas), janelas de observação e número de agrupamentos (k).

A interpretação empírica exige duas ressalvas metodológicas fundamentais. Primeiramente, embora o Coeficiente de Silhueta (*Silhouette Score*) seja normalizado para a escala de $[-1, 1]$, ele é computado sobre espaços topológicos distintos (matrizes acumuladas para o DTW, divergência para o *Soft-DTW* e distância baseada em correlação cruzada para o *k-Shape*). Logo, comparações diretas de magnitude são válidas

principalmente intra-algoritmo. Em segundo lugar, um coeficiente elevado não assegura uma estrutura informacional útil. Configurações com janelas curtas e poucos grupos tendem a degenerar em um único *cluster* majoritário, insuflando a métrica de forma artificial. Além disso, recortes temporais mais antigos (2016–2020) favorecem partições mais nítidas devido ao viés de sobrevivência, já que o universo amostral era menor e majoritariamente composto por grandes companhias. Portanto, a discussão a seguir pondera a qualidade matemática pela distribuição e coesão econômica dos agrupamentos.

A. Visão geral e melhores topologias

A Tabela I consolida as execuções de maior destaque para cada método e período, evidenciando o desempenho superior do agrupamento focado na assinatura macroeconômica.

Tabela I
MELHORES RESULTADOS GERAIS DA VARREDURA SISTEMÁTICA

Método	Data Ref.	Janela	k	Ativos	Distribuição	Silhueta
Soft-DTW (Assinatura)	2024-06	24	4	85	47/20/11/7	0,553
DTW (Preço Bruto)	2018-12	6	3	74	10/18/46	0,490
k-Shape (Preço Bruto)	2019-12	12	3	74	23/7/44	0,463
DTW (Assinatura)	2017-12	12	4	68	31/5/26/6	0,424
DTW (Preço Bruto)	2025-06	6	3	85	56/17/12	0,422
k-Shape (Preço Bruto)	2023-12	12	4	85	6/12/12/55	0,401
DTW (Assinatura)	2023-12	12	4	85	25/35/9/16	0,343

Analisada de forma isolada, a tabela poderia sugerir que as configurações focadas no traçado bruto de preços com janelas curtas são excelentes. Contudo, a coluna de distribuição de ativos denuncia o processo de degeneração algorítmica: concentrações como 56/17/12 (recente) e 10/18/46 (antiga) agrupam a quase totalidade do mercado em um único bloco. Quando se impõe a restrição de equilíbrio econômico na separação, os valores estabilizam-se na faixa de 0,30 a 0,37, que representa a zona de utilidade real. A exceção a essa regra é o campeão baseado em assinatura macroeconômica, cujo índice de 0,553 decorre de uma coesão genuína distribuída por múltiplos subgrupos bem definidos.

B. Seleção do número de grupos (k)

A escolha do número de grupos k não foi arbitrária nem fixada a priori: o parâmetro foi tratado como uma das dimensões da varredura sistemática, sendo avaliado por dois critérios conjuntos. O primeiro é o coeficiente de silhueta, que mensura a coesão intragrupo em contraposição à separação intergrupos. O segundo é a não degeneração da partição, um critério qualitativo que rejeita configurações nas quais um único grupo absorve a quase totalidade do mercado.

Para as famílias de trajetória de preço, fez-se variar o número de grupos em $k \in \{3, 4, 5, 6\}$, inspecionando o ponto de inflexão da curva de silhueta, de forma análoga ao método do cotovelo. A família de assinatura, por sua vez, foi avaliada diretamente em $k = 4$. A varredura revela um viés sistemático da métrica em favor de valores pequenos de k , uma vez que os maiores índices de silhueta surgem em $k = 3$ sobre as trajetórias brutas de preço. Esse ganho, porém, é analiticamente enganoso, pois a distribuição dos ativos denuncia a degeneração do agrupamento, com um único

grupo concentrando a maioria absoluta dos papéis, resultando em partições assimétricas do tipo 56, 17 e 12 ativos, o alto valor da métrica reflete uma separação quase trivial em vez de uma estrutura de mercado informativa.

Aumentar k além de quatro grupos traz ganhos apenas marginais de silhueta e logo conduz à sobre-segmentação, enquanto valores menores recaem na partição degenerada já descrita. O número de grupos $k = 4$ representa, portanto, o ponto de equilíbrio empírico. É a partir desse limite que a partição deixa de ser dominada por um único bloco e passa a exibir subgrupos economicamente coerentes, conciliando boa coesão com uma distribuição mais homogênea dos ativos. É também nessa configuração que emerge o agrupamento mais coeso de todo o estudo, a assinatura de sensibilidade ao Dólar, o que reforça $k = 4$ como a escolha natural.

A definição final de $k = 4$ é corroborada por dois argumentos de interpretação técnica. Primeiro, observa-se a paridade com o número de fatores explorados: o modelo descreve cada ativo por sua reação a quatro indicadores macroeconômicos centrais, e quatro arquétipos de regime de sensibilidade constituem uma partição logicamente interpretável desse espaço vetorial. Segundo, nota-se uma continuidade metodológica, visto que a clusterização estática por K-Means com $k = 4$, conduzida na fase descritiva inicial do trabalho, já havia produzido quatro grupos economicamente legíveis, correspondentes a financeiras e varejo doméstico, utilidades e consumo básico, exportadoras de commodities, e materiais básicos e energia. Mantém-se, assim, $k = 4$ como configuração de referência para os resultados estáticos, sem prejuízo de a ferramenta interativa desenvolvida permitir que o usuário final explore dinamicamente outros valores para o parâmetro.

C. Assinatura cambial como eixo discriminante

O agrupamento mais expressivo de todo o estudo foca na trajetória de sensibilidade do mercado em relação à variação do Dólar americano (janela de 24 meses, coeficiente de Spearman, $k = 4$), processado pelo algoritmo *Soft-DTW* com normalização por média e variância (Fig. 4). A varredura demonstra que ampliar a janela de observação de 6 para 24 meses associada à suavização elástica (*Soft-DTW*) filtra os ruídos estocásticos do mercado de câmbio e eleva o poder de separação.

Como a normalização neutraliza a amplitude escalar para isolar estritamente a forma das curvas, os grupos refletem o comportamento evolutivo da sensibilidade cambial das empresas, superando as fronteiras setoriais clássicas. Observou-se a formação de um núcleo expressivo de commodities e mineração (VALE3, BRAP4, BEEF3, JBSS3) ao lado de blocos menores de transição e um núcleo voltado ao varejo doméstico. O achado metodológico central comprova que guiar a clusterização pela morfologia da sensibilidade a um choque exógeno específico entrega arranjos muito mais nítidos do que agrupar preços diretamente.

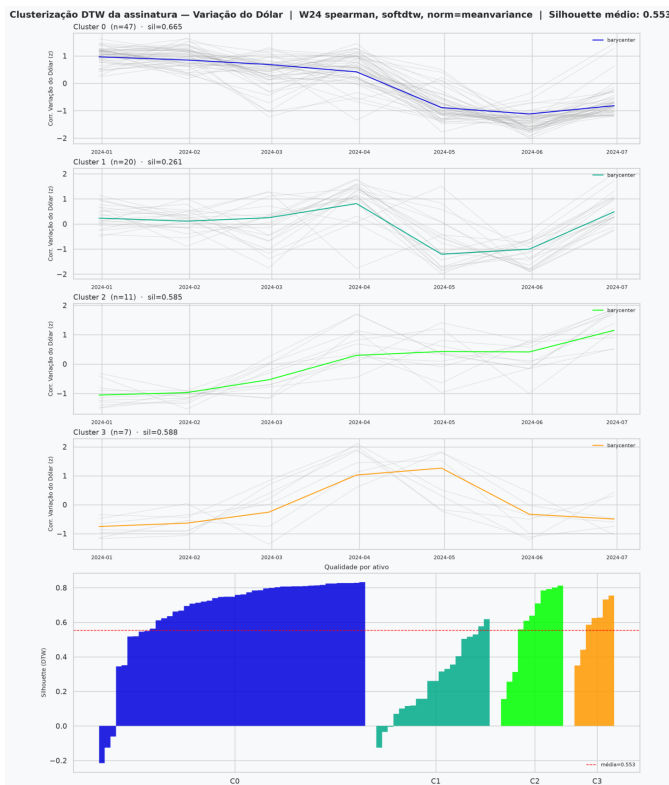


Figura 4. Clusterização *Soft-DTW* da trajetória de sensibilidade ao Dólar ($k = 4$). A análise baseada no comportamento cambial formou subgrupos altamente coesos sem degenerar a partição.

D. Trajetória bruta de preços: invariância econômica

Ao operar sobre os índices de preços históricos, as abordagens *shape-based* convergiram para a mesma estrutura fundamental. Ambas as técnicas (DTW elástico e *k-Shape* por correlação cruzada) isolaram repetidamente um núcleo pesado de mineração, siderurgia e petróleo, composto predominantemente pelo bloco VALE3, BRAP4, USIM5, CSNA3, GGBR4 e PETR4. O fato de um algoritmo independente de alinhamento temporal capturar exatamente as mesmas topologias do agrupador *k-Shape* valida que o núcleo de commodities representa um invariante econômico real da bolsa brasileira, e não um artefato matemático residual de uma única rotina de software.

A decisão prática entre os modelos impõe uma troca tecnológica: o DTW processa naturalmente o painel desbalanceado de empresas através do preenchimento adaptativo de dados ausentes, acomodando ativos novos e antigos. Em contrapartida, o *k-Shape* força a homogeneização das séries, podando o histórico de todos os ativos pela extensão do integrante mais recente, o que eleva a coesão interna mas descarta informações passadas valiosas.

E. Similaridade dinâmica e mudanças de regime

Além dos agrupamentos topológicos densos, a arquitetura validou a extração direta de distâncias ponto a ponto para suporte rápido à decisão (“*quais ativos comportam-se como*

este?”). Utilizando a PETR4 como alvo em dezembro de 2022, o modelo cruzou três lentes de observação distintas (Fig. 5, 6 e 7).

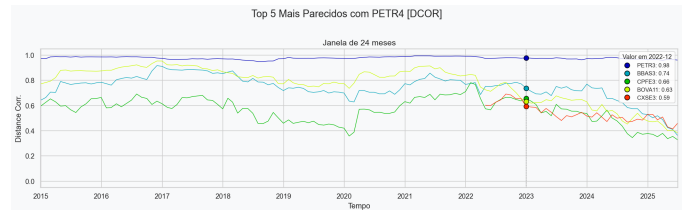


Figura 5. Ativos de maior similaridade com a PETR4 identificados via *distance correlation* (retornos brutos, Dez/2022).

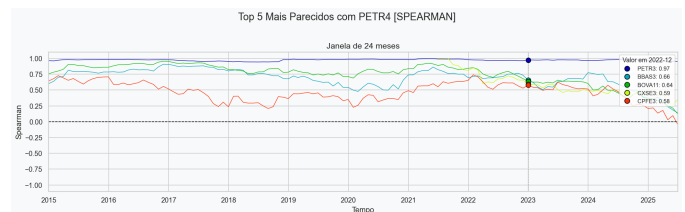


Figura 6. Identificação de similaridade via coeficiente não linear de Spearman (retornos brutos, Dez/2022). O alinhamento das duas métricas corrobora a estabilidade dos movimentos de preço no período.

Para as trajetórias de retornos puras, tanto a correlação de distância multivariada quanto o coeficiente de Spearman ranquearam perfis quase idênticos (PETR3, BBAS3, CPFE3, BOVA11 e CXSE3), provando forte concordância matemática em recortes estritos. No entanto, ao se avaliar a similaridade através da assinatura macroeconômica no mesmo período, a vizinhança da petrolífera alterou-se.

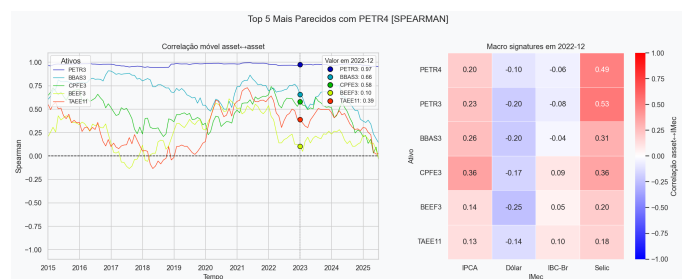


Figura 7. Similaridade avaliada pelo perfil da assinatura macroeconômica (Dez/2022). A matriz de sensibilidade evidencia um alinhamento governado pela correlação positiva com a curva de juros (Selic) e negativa com a de câmbio (Dólar).

O traço de aproximação não obedeceu ao setor industrial, mas à conjuntura monetária. Em 2022, a PETR4 demonstrou elevada sensibilidade aos juros, movendo-se de forma coordenada com instituições financeiras e seguradoras, apresentando neutralidade ao câmbio no período. Esse dinamismo demonstra o valor central do sistema: a correlação entre ativos não é fixa e depende da lente analítica e do regime macroeconômico vigente.

F. Discussão analítica

A consolidação dos testes empíricos permite extrair constatações determinantes para o processo de suporte à diversificação de carteiras no cenário brasileiro:

- 1) **O Câmbio como vetor primário:** O acompanhamento isolado da sensibilidade ao Dólar gera os agrupamentos mais coesos e economicamente fundamentados da B3. Janelas mais extensas aliadas à suavização geométrica superam abordagens de curto prazo baseadas puramente em retorno.
- 2) **A invariância do setor de Base:** As empresas exportadoras e extrativistas brasileiras respondem a matrizes exógenas de modo uníssono, emergindo repetidamente como o *cluster* de maior densidade e obviedade matemática, independentemente do algoritmo subjacente.
- 3) **Dinamicidade de proximidade:** Prova-se a premissa fundamental de que a vizinhança comportamental de um ativo depende do contexto. Uma empresa pode ser pareada ao setor elétrico durante um ciclo de estresse monetário, mas reagir de forma análoga a mineradoras sob choques cambiais.

Em perspectiva tecnológica, embora o *Soft-DTW* entregue fronteiras notavelmente superiores, seu alto custo computacional restringe o uso massivo. As adaptações otimizadas do DTW clássico e do *k-Shape* fornecem o balanço adequado entre performance algorítmica e recuperação da estrutura econômica real. Em suma, as explorações validam a tese de que mapear matrizes dinâmicas de sensibilidade macroeconômica fomenta arranjos estratégicos indisponíveis nas leituras estáticas.

V. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho cumpriu o objetivo central de desenvolver o *PortFlow Invest* enquanto ferramenta dinâmica de suporte à diversificação de carteiras no mercado de capitais (B3). A partir de uma infraestrutura de dados consolidada, a investigação entregou: (i) a modelação de interdependências através de janelas deslizantes otimizadas, suportando métricas lineares e não-lineares; (ii) a aplicação prática da taxonomia de agrupamento de séries temporais focada em métricas morfológicas (*shape-based*), nomeadamente o *Dynamic Time Warping* (e a variante *Soft-DTW*) e o *k-Shape*; (iii) um ecossistema de avaliação parametrizado pelo Coeficiente de Silhueta adaptado a matrizes de distâncias sequenciais; e (iv) uma plataforma analítica interativa projetada para investidores não especialistas.

Sob a perspectiva científica, o estudo validou que o agrupamento mais coeso e estruturalmente informativo não deriva da trajetória nominal de preços absolutos, mas da evolução temporal da sensibilidade dos ativos face ao Dólar. Esse resultado atesta empiricamente a premissa de que o risco macroeconômico se altera por regimes ao longo do tempo, comprovando que orientar a clusterização pelas reações a fatores exógenos revela agrupamentos mais nítidos do que a análise estática tradicional.

Para consolidar o sistema como uma plataforma de recomendação exaustiva e otimizada, delineiam-se as seguintes frentes para trabalhos futuros:

- 1) **Integração de Otimização Clássica:** Incorporação da biblioteca *Riskfolio-Lib* como motor matemático de referência (*baseline*) para o cálculo de fronteiras eficientes e para a simulação de alocações de portfólio baseadas nos clusters gerados.
- 2) **Avaliação retrospectiva (*Backtesting*):** Construção de um ambiente de testes históricos para validar se as carteiras desenhadas por regimes de sensibilidade macroeconômica exibem maior proteção a quedas abruptas (*drawdowns*) quando contrastadas com a diversificação setorial ingênua.
- 3) **Generalização por indicadores isolados:** Uma vez que o câmbio revelou ser o eixo discriminante mais acentuado, propõe-se estender o modelo analítico aos demais vetores e subíndices (como recortes setoriais do IBC-Br e o impacto quantificado de surpresas inflacionárias sobre a curva do IPCA).
- 4) **Otimização computacional do *Soft-DTW*:** Investigar rotinas de aceleração e aproximações algorítmicas que mitiguem a elevada exigência de processamento desta métrica elástica, viabilizando a sua implementação fluida em ambientes de produção e *dashboards* dinâmicos.
- 5) **Expansão da taxonomia de agrupamento:** Evoluir das atuais métricas baseadas em distância para a exploração de métodos suportados pela aprendizagem de representações subjacentes (*representation-learning-based*), integrando arquiteturas de *Deep Learning* aplicadas a sequências temporais.

Em suma, esta pesquisa estabeleceu o núcleo de engenharia e mineração do *PortFlow Invest*, demonstrando que o rastreo dinâmico da sensibilidade econômica pode revelar padrões latentes de mercado e complementar análises de coesão e diversificação de portfólios.

APÊNDICE: DISPONIBILIDADE DE CÓDIGO-FONTE

Com o intuito de garantir a transparência, fomentar a reprodutibilidade científica desta pesquisa e contribuir para o desenvolvimento de ferramentas analíticas de código aberto voltadas ao mercado financeiro brasileiro, todo o ecossistema computacional desenvolvido para o sistema *PortFlow Invest* foi disponibilizado publicamente.

O repositório do projeto inclui a arquitetura modular completa descrita neste documento e sintetizada na Fig. 1, cobrindo todo o fluxo dos *Armazéns de dados* (I) ao *Conhecimento* (VI), e abrangendo:

- Os *pipelines* de engenharia e pré-processamento de dados da B3 e dos indicadores macroeconômicos;
- O motor de cálculo otimizado via *cache* para correlações dinâmicas e extração da assinatura macroeconômica;
- As implementações dos algoritmos de clusterização de séries temporais (*Dynamic Time Warping*, *Soft-DTW* e *k-Shape*);

- O código-fonte integral da interface web interativa (*dashboard*) desenvolvida em Streamlit.

O acesso ao repositório pode ser realizado por meio da plataforma GitHub, através do seguinte endereço eletrônico: <https://github.com/lucasaugs/b3-pattern-recon>.

REFERÊNCIAS

- [1] Harry Markowitz. "Portfolio selection". *The Journal of Finance*, 7(1): 77–91, 1952.
- [2] William F. Sharpe. "Capital asset prices: A theory of market equilibrium under conditions of risk". *The Journal of Finance*, 19(3):425–442, 1964.
- [3] Mantegna, R.N. Hierarchical structure in financial markets. *Eur. Phys. J. B* 11, 193–197, 1999. <https://doi.org/10.1007/s100510050929>
- [4] Paparrizos, John, and Luis Gravano. "k-shape: Efficient and accurate clustering of time series." *Proceedings of the 2015 ACM SIGMOD international conference on management of data*. 2015.
- [5] Hyeon Yong Kim and Chang Hee Won. "Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple garch-type models". *Expert Systems with Applications*, 103:25–37, 2017
- [6] Paparrizos, John, Fan Yang, and Haojun Li. "Bridging the gap: A decade review of time-series clustering methods." *arXiv preprint arXiv:2412.20582*, 2024.
- [7] McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). Disponível em: <https://pandas.pydata.org>
- [8] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- [9] Streamlit Inc. (2024). Streamlit Documentation. Disponível em: <https://docs.streamlit.io>
- [10] Tavenard, Romain, et al. "Tsllearn, a machine learning toolkit for time series data." *Journal of machine learning research* 21.118 (2020): 1-6. Disponível em: <https://tslearn.readthedocs.io/en/stable/index.html>