



UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA
MESTRADO ACADÊMICO EM FÍSICA

EWERTON DA SILVA COSTA

ANÁLISE DE SÉRIES TEMPORAIS DO MERCADO FINANCEIRO UTILIZANDO
TEORIA DE REDES COMPLEXAS

FORTALEZA

2023

EWERTON DA SILVA COSTA

ANÁLISE DE SÉRIES TEMPORAIS DO MERCADO FINANCEIRO UTILIZANDO TEORIA
DE REDES COMPLEXAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Física do Programa de Pós-Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Física. Área de Concentração: Física da Matéria Condensada.

Orientador: Prof. Dr. Saulo Davi Soares e Reis.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C871a Costa, Ewerton da Silva.

Análise de séries temporais do mercado financeiro utilizando teoria de redes complexas / Ewerton da Silva Costa. – 2023.
63 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Física, Fortaleza, 2023.

Orientação: Prof. Dr. Saulo Davi Soares e Reis.

1. Minimal spanning tree. 2. Estruturas de grupos financeiros. 3. Séries temporais. I. Título.

CDD 530

EWERTON DA SILVA COSTA

ANÁLISE DE SÉRIES TEMPORAIS DO MERCADO FINANCEIRO UTILIZANDO TEORIA
DE REDES COMPLEXAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Física do Programa de Pós-Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Física. Área de Concentração: Física da Matéria Condensada.

Aprovada em: 27/01/2023.

BANCA EXAMINADORA

Prof. Dr. Saulo Davi Soares e Reis (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Cesar Ivan Nunes Sampaio Filho
Universidade Federal do Ceará (UFC)

Prof. Dr. Rilder de Sousa Pires
Universidade de Fortaleza (UNIFOR)

AGRADECIMENTOS

Agradeço, em especial, a Débora Gomes e aos meus pais Everane e Deusimar, por todo apoio proporcionado.

Ao Prof. Dr. Saulo Davi Soares e Reis, pela orientação e formação durante o mestrado, pelos excelentes tópicos de pesquisas propostos e por diversos ensinamentos e debates construtivos, que resultaram neste trabalho. Aos membros do corpo docente do Departamento de Física, em especial, o Prof. Dr. Carlos Lenz César, por todos os ensinamentos que carrego desde a minha graduação e aos professores Wandemberg Paiva, José Alexandre Paschoal, Carlos William Paschoal e João Milton.

Também gostaria de agradecer aos colegas do Departamento: Higor Monteiro, por ter me incentivado a mudar de área; Lemuel Ferreira; Israel Lima e aos antigos colegas do GTMC, Davi Dantas, Nathanaell Sousa, Jorge Coelho e Jorge Araújo.

Ao Departamento de Física da Universidade Federal do Ceará pela infraestrutura e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro e por não ter me deixado desistir do mestrado.

"O único homem que está isento de erros é
aquele que não arrisca acertar." (EINSTEIN,
[19--])

RESUMO

Embora a área de Sistemas Econômicos seja tradicionalmente estudada por economistas e matemáticos, é crescente o número de estudos nessa área realizada por físicos. Consequentemente, ferramentas de Sistemas Complexos e Ciência de Redes tem se provado úteis na modelagem de mercados financeiros. Em particular, conceitos como distância de correlação e *minimal spanning tree* (MST) tem sido largamente utilizados na literatura para estudar a dinâmica de mercados financeiros, através da similaridade de medidas, análise de risco e comportamento em períodos de crise financeira. Neste trabalho, utilizaremos dados do mercado financeiro composto por índices de diversos países e regiões, moedas, *commodities*, títulos públicos, entre outros. Fazemos uma análise de como diferentes grupos financeiros se relacionam levando em conta sua correlação. Além disso, utilizamos a correlação entre ativos para construir a MST e estudar suas propriedades de rede.

Palavras-chave: minimal spanning tree; estruturas de grupos financeiros; séries temporais.

ABSTRACT

In spite of being an area commonly studied by economists and mathematicians, the analysis of economic systems has attracted a growing number of physicists. Consequently, complex system and network science tools have been proven to be extremely useful in the modeling of financial markets. In particular, concepts such as correlation distance and minimal spanning tree (MST) have been widely used in the literature to study the dynamics of financial markets, through similarities of measures, risk analysis and behaviour in periods of financial crises. In this work, we use financial market data composed of indices of several countries and regions, currencies, commodities, government bonds, among others. We analyse how different financial groups are related by taking their correlation into account. In addition to that, we use the correlations between assets to create the MST and study its network properties.

Keywords: minimal spanning tree; financial groups structures; time series.

LISTA DE FIGURAS

Figura 1 – Uma rede da ciência de redes.	14
Figura 2 – Representação de um grafo.	16
Figura 3 – Grafos direcionado e não-direcionado.	16
Figura 4 – Menor caminho entre os vértices v_1 e v_4	18
Figura 5 – Árvore e Floresta.	20
Figura 6 – Algoritmo de Kruskal.	22
Figura 7 – Algoritmo de Prim.	23
Figura 8 – Soma de vetores representando distâncias de correlação.	31
Figura 9 – Série Temporal dos preços de fechamento do índice <i>S&P500</i>	34
Figura 10 – MST do Índice <i>S&P500</i> para os anos de 2010 a 2020.	38
Figura 11 – Mapas de calor da matriz de distância de correlação e da matriz de distância de correlação ordenada pela MST.	39
Figura 12 – Ampliação do mapa de calor da matriz de distância de correlação ordenada pela MST.	40
Figura 13 – MSTs para o Índice <i>S&P500</i>	41
Figura 14 – Séries Temporais	46
Figura 15 – Log-retorno semanal	48
Figura 16 – Matriz de correlações	49
Figura 17 – Análise da conectividade dos grupos financeiros	51
Figura 18 – Dendrograma da Matriz de Correlação	52
Figura 19 – Matriz de Correlação de todos os ativos	54
Figura 20 – MST para todos os ativos no período de 10 anos	55
Figura 21 – Matriz de Correlação de todos os ativos reordenada	56
Figura 22 – Análise dos Valores Médios da Correlação e da Distância de Correlação	57

LISTA DE ABREVIATURAS E SIGLAS

APAC	Ásia e Pacífico, na sigla em inglês
CN	China
COMM	Abreviação para <i>commodities</i>
CUR	Abreviação para moedas (<i>currencies</i> em inglês)
EE	Europa Oriental, na sigla em inglês
FOREX	<i>Foreign Exchange Rates</i>
LATAM	América Latina e México, na sigla em inglês
MENA	Oriente Médio e África, na sigla em inglês
MST	Minimal Spanning Tree
NA	América do Norte, na sigla em inglês
WE	Europa Ocidental, na sigla em inglês

LISTA DE SÍMBOLOS

$\langle x \rangle = E[x]$	Valor esperado de x
$\bar{x}$	Valor médio de x
M_n	Momento de n -ésima ordem
σ^2	Variância
m_n	Momento centrado de n -ésima ordem
$cov(x, y)$	Covariância entre x e y
ρ_{ij}	Coeficiente de correlação
d_{ij}	Distância de correlação

SUMÁRIO

1	INTRODUÇÃO	11
2	CONCEITOS FUNDAMENTAIS	13
2.1	Teoria de Grafos	14
2.1.1	<i>Definições e Notações</i>	14
2.1.2	<i>Caminho</i>	17
2.1.3	<i>Conectividade</i>	18
2.1.4	<i>Árvores</i>	20
2.1.5	<i>Minimal Spanning Trees</i>	21
2.2	Teoria da Probabilidade	23
2.2.1	<i>Valor Esperado, Momentos e Covariância</i>	25
2.2.2	<i>Coefficiente e distância de correlação</i>	28
3	SÉRIES TEMPORAIS DO MERCADO FINANCEIRO	33
3.1	O que são ações?	33
3.2	Retorno e Log-Retorno	35
3.3	Análise de séries temporais	36
3.4	Análise do SP500	38
4	RESULTADOS E DISCUSSÃO	42
4.1	Dados	42
4.2	Eventos e Ciclos	44
4.3	Análise Exploratória dos Dados	45
4.4	Análise dos agrupamentos a partir da correlação	53
5	CONCLUSÃO	59
	REFERÊNCIAS	60

1 INTRODUÇÃO

Muitos dos sistemas complexos observados em Física, Biologia e Ciências sociais são organizados, hierarquicamente, em aglomerados e estruturas correlacionadas. A hierarquia das interações desses sistemas está diretamente ligada a dinâmica dos mesmos. Portanto, faz-se necessária uma descrição quantitativa desses sistemas para a extração de informações que caracterizem a dinâmica da interação. Neste contexto, é possível investigar propriedades de séries temporais provenientes do mercado financeiro (MANTEGNA, 1999; BONANNO *et al.*, 2001; BONANNO *et al.*, 2003; MARDIA *et al.*, 1979; MICCICHÈ *et al.*, 2003; CORONNELLO *et al.*, 2005).

O crescente interesse de físicos em sistemas econômicos tem propiciado o surgimento de novas técnicas e modelos para o entendimento da evolução temporal do preço de um determinado ativo financeiro, de um índice de uma bolsa de valores, ou do mercado de moedas (MANTEGNA; STANLEY, 2000; SINHA *et al.*, 2011; GABAIX *et al.*, 2003; STAUFFER *et al.*, 1999). Métodos de redes complexas como matriz de covariância, correlação, distância de correlação e técnicas de rede (CORMEN *et al.*, 2009; BARABÁSI, 2016) (como graus de interação, formação de árvores e aglomerados) são ferramentas importantes para a extração de informações do mercado.

No mercado financeiro, a técnica de Minimal Spanning Tree (Minimal Spanning Tree (MST)) foi primeiramente proposta por Mantegna (MANTEGNA, 1999) como um método para analisar similaridades entre preços de ações e encontrar otimizações de portfólios. Esse tipo de método é denominado análise de topologia de redes e tem se mostrado bastante útil na análise de dados de mercado financeiro (ONNELA *et al.*, 2003; SHIN *et al.*, 2020). Em particular, há estudos com aplicações de teoria de redes para o mercado financeiro (TANG *et al.*, 2018; BONANNO *et al.*, 2004), mercados de ações coreano (JUNG *et al.*, 2006), europeu (GILMORE *et al.*, 2008), americano (POZZI *et al.*, 2008), brasileiro (TABAK *et al.*, 2010), mercado de commodities (SIECZKA; HOLYST, 2009) e de moedas (MCDONALD *et al.*, 2005; BRIDA *et al.*, 2009).

Sabe-se que mercados em diferentes localizações geográficas reagem de maneira distinta à notícias externas (EDERINGTON; LEE, 1993; BALDUZZI *et al.*, 2001; ANDERSEN *et al.*, 2007), o que sugere que as informações contidas em séries temporais financeiras estão, de alguma forma, correlacionadas (NOBI *et al.*, 2014). Os índices financeiros compõem uma rede financeira global e se reorganizam devido à crises ou informações externas (KUMAR; DEO,

2012; SENSOY *et al.*, 2013). Além disso, sistemas financeiros compostos por tipos distintos de companhias tendem a ser sensíveis ao mesmo tipo de notícias externas (NOBI *et al.*, 2014).

Isso ocorre porque mercados financeiros são sistemas complexos em constante evolução, ocasionada por mudanças provenientes de novas tecnologias, alianças e parcerias econômicas, entre outros fatores. Por esse motivo, explorar a dinâmica das MSTs pode ampliar o entendimento sobre o mercado, pois a interpretação dos conceitos de redes complexas, já citados acima, pode ser utilizada para caracterizar os mercados e como os diferentes setores dentro destes se correlacionam.

Por ser uma área de grande interesse de estudo, o mercado financeiro possui, possui uma vasta literatura baseada em análises qualitativas. Esses estudos foram desenvolvidos a partir do comportamento temporal dos preços ao longo dos anos. Com isso em mente, este trabalho busca viabilizar algumas análises quantitativas, com base nas propriedades de redes complexas.

Com essa finalidade, constrói-se a matriz de distância de correlação de grupos financeiros, compostos por dados dos mercados financeiros das principais regiões econômicas do mundo, informações de outros ativos financeiros que compõem a cadeia comercial mundial – *Commodities*, Moedas e Metais – além da informação de títulos governamentais de dívida pública dessas regiões. A partir disto, pode-se construir a MST para os índices, extrair propriedades das vizinhanças de cada grupo financeiro, determinar o caminho mínimo médio desses grupos e reordenar a matriz de distância de correlação partindo da ordem crescente dos pesos obtidos a partir da MST.

Esse trabalho, está organizado da seguinte forma: o capítulo 2 aborda os conceitos fundamentais de redes e teoria da probabilidade. O capítulo 3 se dedica a discussão de séries temporais do mercado financeiro. O capítulo 4 contém os resultados e discussões e o capítulo 5 contém a conclusão e perspectivas.

2 CONCEITOS FUNDAMENTAIS

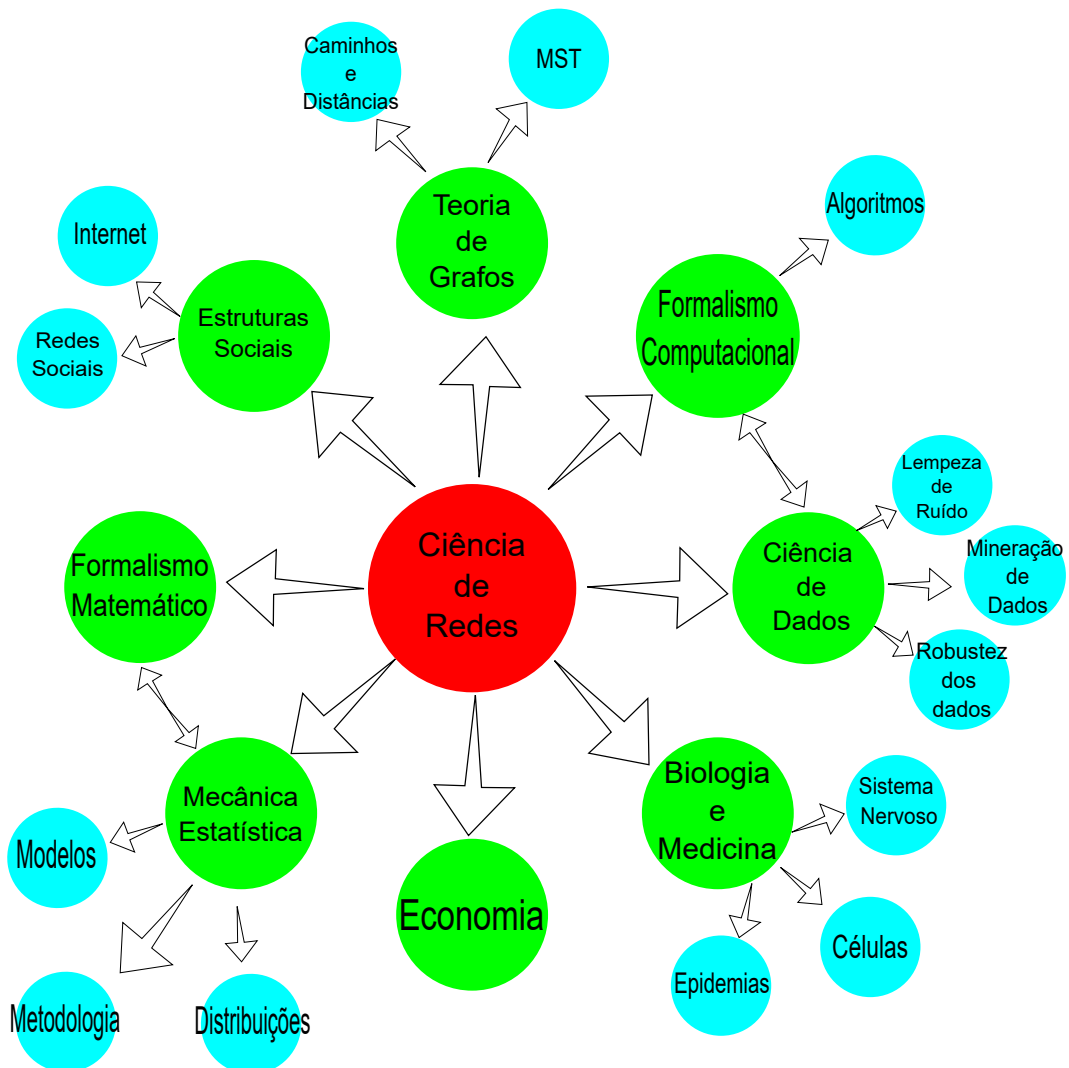
A ciência de redes se tornou uma ferramenta inovadora, largamente utilizada no estudo de sistemas complexos em diversas áreas científicas e tecnológicas (WATTS; STROGATZ, 1998; BARABÁSI; ALBERT, 1999; BARABÁSI, 2002). O estudo de redes, mais especificamente de teoria de grafos, surgiu a partir de 1736 quando foi proposto o problema teórico de percorrer toda a cidade de Königsberg, capital da Prússia Oriental, passando por cada uma de suas sete pontes apenas uma vez. Apesar de muitas tentativas, ninguém teve sucesso em encontrar tal caminho e, como provado por Euler utilizando o formalismo de grafos, esse caminho não existe (BARABÁSI, 2016).

Euler resolveu o problema das pontes de Königsberg utilizando a teoria dos grafos representando a cidade e suas pontes por um grafo, onde as ilhas e margens eram os vértices e as pontes eram as arestas. Ele observou que, para que fosse possível caminhar por todas as pontes exatamente uma vez, cada vértice do grafo teria que ter um número par de arestas incidentes (duas ilhas e duas margens do rio, por exemplo). No entanto, a cidade de Königsberg tinha quatro vértices com um número ímpar de arestas incidentes. Portanto, Euler concluiu que era impossível caminhar por todas as pontes exatamente uma vez, retornando ao ponto de partida (EULER, 1968).

A solução de Euler para o problema das pontes de Königsberg foi um marco importante na teoria dos grafos e levou ao desenvolvimento de novas técnicas para resolver problemas semelhantes. Atualmente, a teoria dos grafos é uma área importante da matemática que tem aplicações em diversas áreas, como ciência da computação, redes sociais e otimização.

A ciência de redes se desenvolveu formalmente e compreende diversas áreas científicas, entre elas economia, mecânica estatística, biologia (Figura 1). Apesar de permear áreas de conhecimento tão distintas, o que a ciência de redes tem em comum em todas elas são os princípios relacionados a topologia. Por exemplo, a topologia de redes sociais afeta como a difusão de informações ou doenças ocorre; a topologia de uma rede de abastecimento de energia elétrica afeta sua robustez e estabilidade (STROGATZ, 2001); a topologia de planos de voos e a localização de *hubs* determinam a concentração de tráfego aéreo em determinadas regiões (BONGIORNO *et al.*, 2015).

Figura 1 – Uma rede da ciência de redes.



Fonte: O autor. Áreas da ciência de redes representadas por um grafo.

2.1 Teoria de Grafos

Quando busca-se o entendimento de um sistema complexo, precisamos conhecer como ocorre a interação entre seus componentes. Assim, nesta seção, busca-se a definição de conceitos básicos necessários a nossa aplicação como, por exemplo: grafos, distância de correlação, teoria da probabilidade e obtenção da MST.

2.1.1 Definições e Notações

Muitas situações do mundo real podem ser descritas na forma de um diagrama, consistindo de um conjunto de pontos e linhas que, por sua vez, ligam pares de pontos convenientemente. Desta forma, os pontos podem representar: pessoas e seus pares de amigos (com a

interação dada por linhas); centros de comunicação e suas redes de distribuições (interligadas por linhas); ações de um mercado financeiro e suas distâncias de correlação (onde as linhas possuem diferentes pesos); distâncias entre bairros e qual o caminho mais rápido (com as linhas representando as ruas e o trânsito) e entre outros.

Buscando uma notação matemática e precisa, define-se um *grafo* G como $G = \{V(G), E(G)\}$, onde o conjunto finito e com N elementos $V(G) = \{v_1, v_2, \dots, v_N\}$ é a representação dos *vértices* ou *nós* da rede e o conjunto $E(G) = \{(v_i, v_j), v_i, v_j \in V\}$ é a representação das *arestas* entre dois vértices de V (BONDY; MURTY, 2008).

A vantagem de se trabalhar com um grafo consiste em sua representação gráfica que, por sua vez, simplifica o entendimento de propriedades provenientes de um determinado sistema analisado. Assim, indicando cada vértice como um ponto ou um círculo e cada aresta como uma linha que une dois nós, temos a representação de um grafo.

Partindo dessa abordagem, vemos que a Figura 2 representa um exemplo de um grafo. Com os nós representados por cada círculo, sendo $V = \{1, 2, \dots, 9\}$, totalizando 9 componentes e $E = \{(1, 2), (2, 3), (2, 8), (2, 6), (3, 8), (3, 4), (3, 6), (6, 7), (8, 9), (4, 5), (6, 4), (8, 4)\}$ são as arestas entre os pares de vértices, com um total de 12 possibilidades.

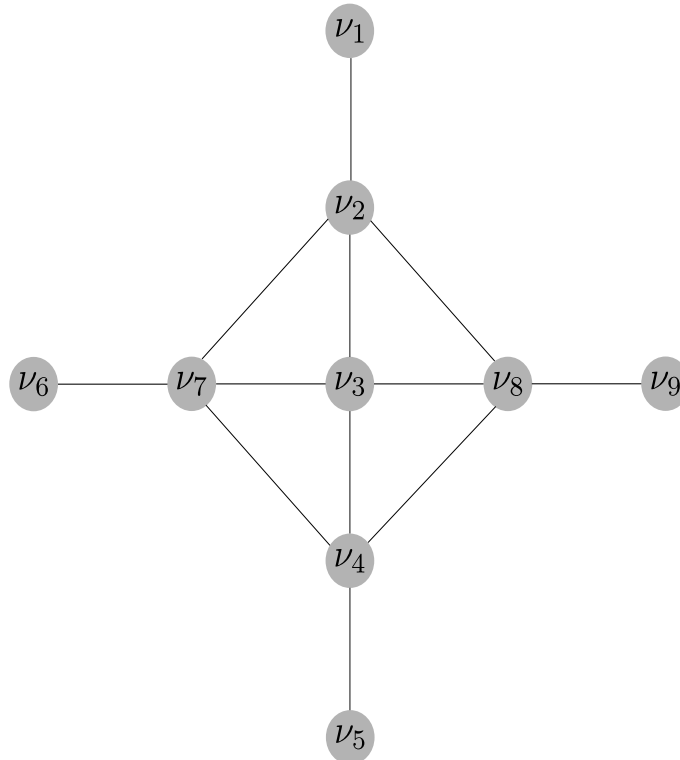
Então, pode-se afirmar que:

- A *ordem* do sistema, dada por $|G|$, representa o número de vértices, N , em $G(V, E)$. Ou seja, é a quantidade de entidades que o grafo possui;
- O número de arestas, ou número de ligações L , representa o número de conexões existentes em uma rede. Trazendo como informação o seu *tamanho*, denotado por $\|G\|$ (DIESTEL, 2005).

Além das propriedades já citadas, podemos observar que a posição relativa dos pontos que representam os nós ou do formato de cada linhas que os unem usualmente não tem significância. No entanto, vale salientar algumas características importantes na configuração de um grafo que são as seguintes:

- Um *grafo não-direcionado* ocorre quando suas arestas tem a função única de indicar a conexão entre cada vértice, o que pode ser representado por um conjunto não ordenado $(v_i, v_j) = (v_j, v_i)$, como ocorre na Figura 2;
- Um *grafo direcionado* ocorre quando um par (V, E) de G contém ligações direcionais, representadas por flechas, impondo uma característica unidirecional de interação entre dois vértices, além de possuir um ordenamento específico na representação das arestas

Figura 2 – Representação de um grafo.

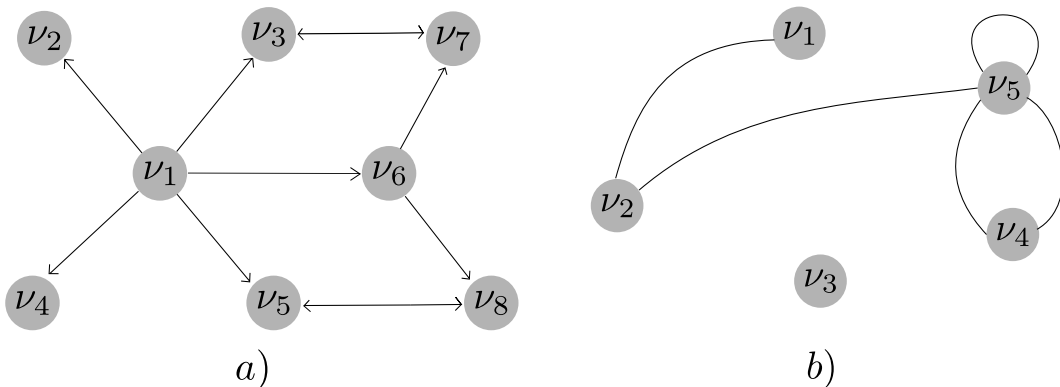


Fonte: O autor. Demonstração de um grafo $G = (V, E)$, com $V = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ e $E = \{(1, 2), (2, 3), (2, 8), (2, 6), (3, 8), (3, 4), (3, 6), (6, 7), (8, 9), (4, 5), (6, 4), (8, 4)\}$.

(CORMEN *et al.*, 2009), como na Figura 3-a;

- *loops* ou *auto-ciclo*, possuem a finalidade de indicar uma ligação que se inicia em um vértice e termina nele mesmo, representada por arestas circulares, como na Figura 3-b.

Figura 3 – Grafos direcionado e não-direcionado.



Fonte: O autor. a) Um grafo G direcionado, com $|G| = 8$ e $\|G\| = 11$, onde $V = \{1, 2, 3, 4, 5, 6, 7, 8\}$ e $E = \{(1, 2), (1, 4), (1, 5), (1, 3), (3, 7), (7, 3), (5, 8), (8, 5), (1, 6), (6, 7), (6, 8)\}$. b) Um grafo G não-direcionado e com auto-ciclo, possui $|G| = 5$ e $\|G\| = 5$, onde $V = \{1, 2, 3, 4, 5\}$ e $E = \{(1, 2), (2, 5), (5, 5), (5, 4), (4, 5)\}$, sendo $(5, 5)$ o auto-ciclo e o vértice v_3 isolado.

2.1.2 Caminho

Segundo (BARABÁSI, 2016), “Em redes, o conceito de distância é desafiador”, pois qual será a distância entre duas páginas de internet ou a distância de interação entre dois indivíduos desconhecidos? Ou seja, o conceito físico de distância, ao qual conhecemos, não se aplica no estudo de grafos, tornando-se necessário a utilização de *caminho*.

Desta forma, o caminho é um grafo que começa e termina em vértices diferentes. Ou seja, o comprimento ℓ de um vértice u até um vértice v , em um grafo $G = (V, E)$ é uma sequência dada por: $\{v_1, v_2, \dots, v_\ell\}$, sendo $u = v_1$ e $v = v_\ell$ e $(v_{i-1}, v_i) \in E$ para $i = 1, 2, \dots, \ell$. Assim, o *comprimento* do caminho, será o número de arestas percorrido de u a v .

Logo, seria intuitivo nos perguntarmos qual seria o menor caminho entre dois pontos de um grafo. Então, dado um vértice i qualquer, dentre uma rede com N nós, para acharmos o menor caminho entre o ponto i e o ponto j basta percorremos por todos os $N - 1$ caminhos (arestas) possíveis, em que os dois pontos estão contidos. Portanto, podemos definir o *mínimo caminho médio* $\langle \ell \rangle$ (BOCCALETTI *et al.*, 2006) de uma rede como:

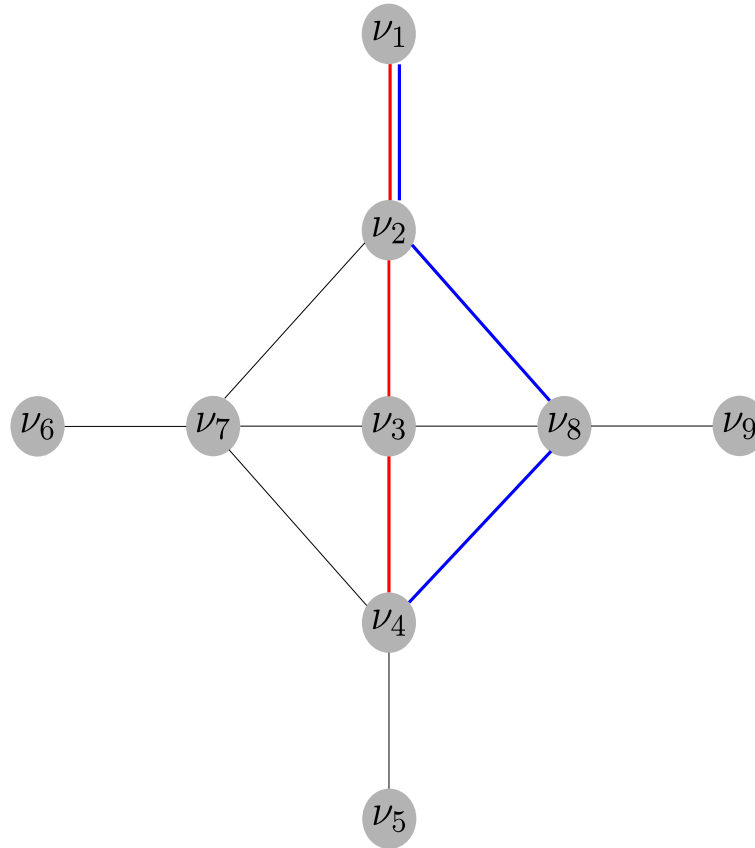
$$\langle \ell \rangle = \frac{1}{N(N-1)} \sum_i^N \sum_{j:j \neq i}^N \ell_{ij}. \quad (2.1)$$

Como precaução para a equação acima, devemos garantir que os vértices dessa rede estejam todos conectados, excluindo-se os vértices que não possuem um caminho ℓ que os conecte.

Outro ponto importante e que deve ser discutido é o caso de um grafo com arestas ponderadas por um peso. Até agora, só discutimos grafos com ligações idênticas e com o mesmo peso, ou seja, $A_{u,v} = 1$. Mas, em muitos casos, torna-se necessário o estudo de *grafos ponderados*, onde cada aresta (u, v) possui um peso dado por w_{uv} . Essa modificação altera os elementos da matriz de adjacência, equação 2.2, que, por sua vez, será: $A_{ij} = w_{ij}$.

Agora, vamos usar como exemplo a Figura 4, onde há um grafo com nove vértices e observa-se o menor caminho entre os dois vértices v_1 e v_4 . Para o caso de uma rede com adjacência $A_{ij} = 1$, o menor caminho é aquele onde há o menor número de ligações, ou seja, não haverá distinção entre o caminho azul ou o caminho vermelho, pois as ligações são de peso equivalente; para o caso de um rede ponderada onde o comprimento ℓ varia em cada ligação, tomando $w_{2,8} = 1,4$ e $w_{8,4} = 1,4$, temos que o menor caminho possível é o vermelho, com $\ell = 3$ e note que o caminho entre as arestas (v_8, v_3) , (v_3, v_4) é maior do que o caminho (v_8, v_4) .

Figura 4 – Menor caminho entre os vértices v_1 e v_4 .



Fonte: O autor. Um exemplo de uma rede com $N = 9$ vértices. Observa-se o menor caminho entre os dois vértices v_1 e v_4 para o caso de uma rede com adjacência $A_{ij} = 1$, onde o menor caminho é aquele onde há o menor número de ligações: caminho azul ou vermelho; e para o caso de um rede ponderada onde o comprimento ℓ varia em cada ligação.

2.1.3 Conectividade

Antes da definição de conectividade, vamos contextualizar o conceito de vizinhança pois, como visto anteriormente, as arestas tem a finalidade de conectividade entre dois vértices. Assim, dois vértices serão tidos como *vizinhos* se, e somente se, estiverem conectados por uma aresta que incide de ambos. Desta forma, um grafo *conectado* é aquele cujo os nós estão representados em um conjunto não vazio de E , onde dado os vértices v_i e v_j existe um $E = (v_i, v_j)$ e, caso contrário, o grafo é tido como *desconectado*.

Assim, a *conectividade* de um grafo, ou *grau* k , está intimamente relacionada com o número de ligações que determinado nó k_i apresenta. Sendo, assim, possível a determinação do conjunto de vértices *adjacentes* A_{ij} ao nó de referência (BARABÁSI, 2016). Sendo a matriz de

adjacência construída da seguinte forma:

$$A_{ij} = 1, \text{ se existir uma ligação entre o nó } j \text{ e o nó } i \quad (2.2)$$

$$A_{ij} = 0, \text{ caso não existe ligação entre os nós.}$$

Desta forma, é possível inferir que: um vértice *isolado* é aquele cujo o grau é nulo; um grafo não-direcionado é tido como conectado somente quando todos os vértices interagem entre si; um grafo direcionado *fortemente conectado* deve possuir uma conexão em ambas as direções de um par de nós. Outra característica da conectividade de grafos direcionados está presente na parcela de arestas de saem de determinado vértice i , k_i^{sada} , e da parcela de arestas que incidem sobre ele, $k_i^{entrada}$, de forma que o grau desse tipo de rede será dado por:

$$k_i = k_i^{entrada} + k_i^{sada}. \quad (2.3)$$

Como exemplo, podemos verificar o grafo não-direcionado da Figura 2, onde temos $k_1 = 1, k_2 = 4, k_3 = 4, k_4 = 4, k_5 = 1, k_6 = 1, k_7 = 4, k_8 = 4$ e $k_9 = 1$. Então, o número total de ligações L , será dado pela soma de todos os graus da rede: $L = 24$. No entanto, o grafo dessa figura possui $E = 12$ e percebemos que esse fator de 2 ocorre porque cada ligação foi contada referente ao nó de incidência, totalizando o dobro de ligações. Portanto, como generalização, temos:

$$L = \frac{1}{2} \sum_{i=1}^N k_i. \quad (2.4)$$

Sabemos que a média de uma medida é dada pela seguinte relação:

$$\langle x \rangle = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i, \quad (2.5)$$

daí, aplicando essa notação à equação 2.4 e substituindo o somatório por $2L$, vem que

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{2L}{N}, \quad (2.6)$$

que, por sua vez, representa o *grau médio* do grafo. Assim, o grau médio de uma rede não-direcionada é dada pela equação 2.6. Retomando a discussão de grafos direcionados, onde temos

o grau de entrada e o grau de saída e verifica-se a relação 2.3, percebe-se que não há o fator $1/2$, pois apenas o sentido da aresta é levada em consideração. Desta forma, o grau médio de um grafo direcionado será:

$$\langle k^{entrada} \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{L_D}{N}, \quad (2.7)$$

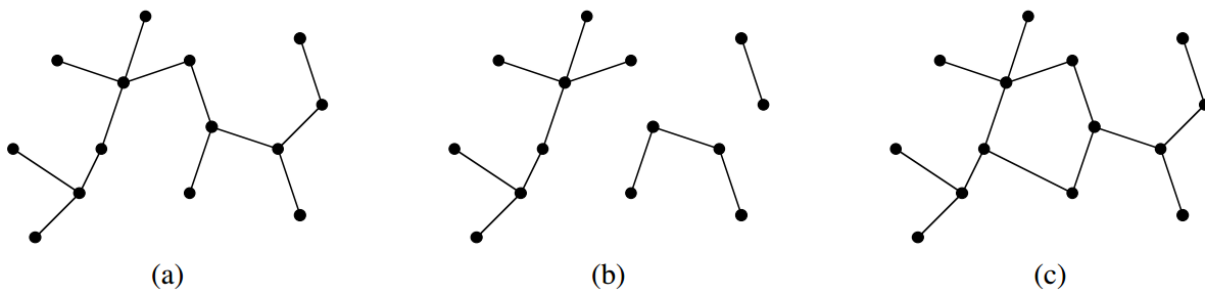
$$\langle k^{saida} \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{L_D}{N}, \quad (2.8)$$

onde L_D é o grau de um grafo direcionado.

2.1.4 Árvores

Um grafo *acíclico* é aquele que não possui ciclos (um conjunto de vértices conectados em uma rede fechada) ou *loops*. Definimos uma *árvore* como sendo um grafo acíclico e conectado. Com base nessa definição, chamamos os grafos acíclicos de *florestas*, de modo que os componentes (um conjunto isolado de um grafo) de uma floresta são árvores (figura 5). Para que o grafo seja conectado, é necessário que haja pelo menos um caminho ligando qualquer par de vértices desse grafo. É possível mostrar que árvores são os grafos conectados que obedecem a essa condição com exatamente um caminho entre quaisquer dois vértices (BONDY; MURTY, 2008).

Figura 5 – Árvore e Floresta.



Fonte: (CORMEN *et al.*, 2009). a) Uma árvore; b) Uma floresta; c) um grafo que contém um ciclo e, por isso, não é caracterizado como uma árvore ou floresta.

Seguindo essa analogia entre árvores e florestas, cada vértice de grau 1 de uma árvore pode ser considerado como uma folha. Por outro lado, se um grafo não contém folhas, ou seja, todos os seus vértices tem, pelo menos, grau 2, então esse grafo contém um ciclo. Esse conceito

é particularmente interessante pois, se retirarmos uma folha de uma árvore, ainda teremos uma árvore. Em outras palavras, o subgrafo $T - v_i$, onde T é uma árvore e $v_i \in T$ é uma folha de T , é uma árvore. Assim, o número de vértices e arestas diminui por 1 sempre que removemos uma folha de uma árvore: $|T - v_i| = |T| - 1$ e $\|T - v_i\| = \|T\| - 1$. Além disso, podemos relacionar o número de vértices e arestas de uma árvore pela equação $\|T\| = |T| - 1$. O seguinte teorema é válido:

Teorema: As seguintes afirmações são equivalentes para um grafo T (DIESTEL, 2005):

- (i) T é uma árvore;
- (ii) Quaisquer dois vértices de T são ligados por um único caminho em T ;
- (iii) T é minimamente conectado, ou seja, T é conectado mas $T - E$ é desconexo para cada aresta $E \in T$.

A equivalência entre (i) e (iii) garante que todo grafo conectado contém uma *spanning tree*.

2.1.5 Minimal Spanning Trees

Uma *spanning tree* é uma sub-árvore de um grafo conexo (existe um caminho entre qualquer par de vértices) que conecta todos os vértices do grafo com o menor peso total possível (CHARTRAND; ZHANG, 2013). Ou seja, de forma que a soma dos pesos das arestas seja mínima. Dado um grafo G que contém uma *spanning tree* T , então G é conectado uma vez que quaisquer dois vértices de G são conectados por um caminho em T , ou seja, em G . Consequentemente, pode-se mostrar que todo grafo conectado contém uma *spanning tree* (BONDY; MURTY, 2008). Ademais, valem as propriedades:

- Uma *spanning tree* tem somente um caminho entre cada par de vértices;
- Uma *spanning tree* de n vértices possui $n - 1$ arestas.

Ao resolvermos problemas reais, é interessante atribuímos pesos as arestas de um grafo e deduzirmos qual *spanning tree* possui o menor caminho. Dado um grafo G com um subgrafo H e com o peso de uma aresta dado por w_{ij} , dizemos que o peso de H é dado por

$$w(H) = \sum_{(v_i, v_j) \in \|H\|} w_{ij}, \quad (2.9)$$

onde somamos sobre todas as arestas possíveis (v_i, v_j) em H . Com isso, desejamos encontrar dentre as possíveis *spanning trees* de G aquela cujo peso seja mínimo. Tal *spanning tree*

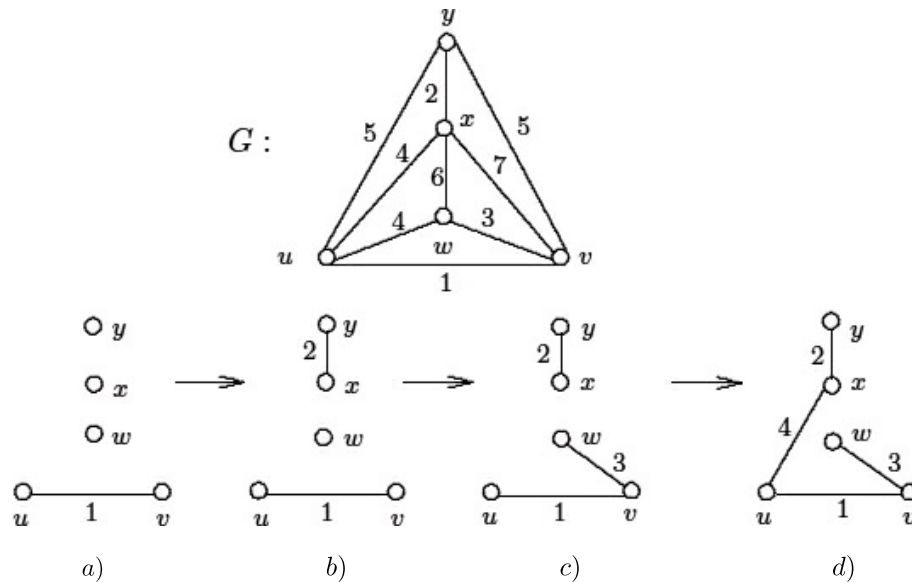
é denominada *Minimal Spanning Tree* (MST) (ou *Minimum Spanning Tree*, dependendo da referência).

Existem diferentes algoritmos para encontrar a MST, sendo os mais famosos o *algoritmo de Prim* e o *algoritmo de Kruskal*.

Algoritmo de Kruskal: Para um grafo conectado e ponderado G , uma spanning tree T é obtida da seguinte forma:

- K1:** Para a primeira aresta E_1 de T , selecionamos, dentre todas as aresta possíveis de G , aquela com o menor peso (Figura 6-a);
- K2:** Para a segunda aresta E_2 , selecionamos qualquer das arestas restantes com menor peso (Figura 6-b);
- K3:** Para a terceira aresta, repetimos o procedimento anterior tomando o cuidado de não produzir ciclos com as demais arestas selecionadas (Figura 6-c);
- K4:** Repetimos os procedimentos acima até que a spanning tree seja obtida (Figura 6-d).

Figura 6 – Algoritmo de Kruskal.



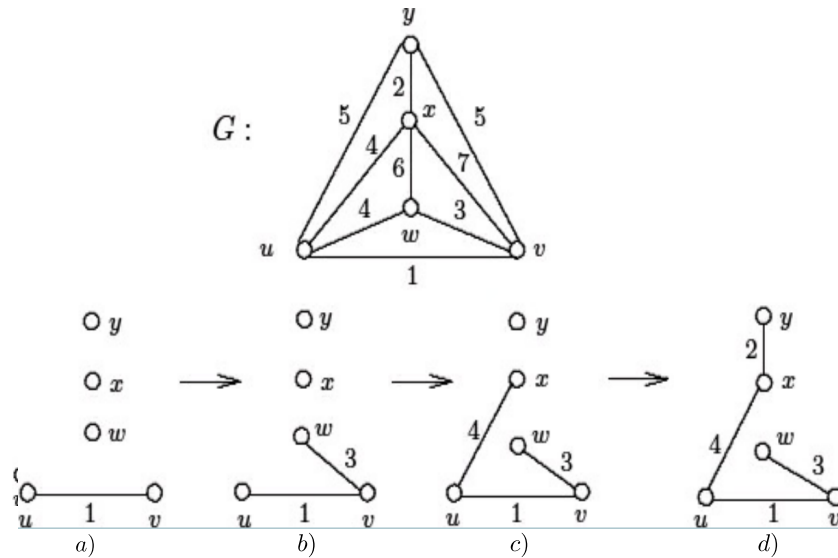
Fonte: (CHARTRAND; ZHANG, 2013). Passos executados no algoritmo de Kruskal para obter uma spanning tree a partir de um grafo conectado e ponderado.

Algoritmo de Prim: Para um grafo conectado e ponderado G , uma spanning tree T é obtida da seguinte forma:

- P1:** Dentre todas as arestas de G selecionamos aquela com menor peso e tomamos um de seus vértices, denotado por v_1 (Figura 7-a);

- P2:** A partir de v_1 , selecionamos, dentre seus vizinhos, a aresta com menor peso (Figura 7-b);
- P3:** Para as próximas ligações, selecionamos a aresta de menor peso, dentre as arestas que possuem exatamente um dos vértices de uma aresta já selecionada (Figura 7-c,d).

Figura 7 – Algoritmo de Prim.



Fonte: (CHARTRAND; ZHANG, 2013). Passos executados no algoritmo de Prim para obter uma spanning tree a partir de um grafo conectado e ponderado.

Notamos, a partir das Figuras (6-d) e (7-d), que as MSTs formadas são semelhantes para ambos os algoritmos pois, dentre todas as spanning trees de um grafo G , há apenas uma que é mínima. Desta forma, a escolha de algoritmo fica a critério pessoal. No entanto, o algoritmo de Prim é vantajoso porque, no processo de busca pelo caminho mínimo, partindo de uma aresta já selecionada, tem-se a matriz de vizinhos. Por esta razão, esse será o método utilizado neste trabalho.

2.2 Teoria da Probabilidade

Dizemos que uma teoria é determinística quando é possível prever o resultado de um experimento antecipadamente. Entretanto, para os casos não-determinísticos ou aleatórios, não se pode prever o resultado final. Quando jogamos uma moeda uma, duas, ou sucessivas vezes, por exemplo, será uma experimentação onde não se sabe o resultado prévio, mas tem-se em mente a possibilidade do resultado ser cara ou coroa. Ou seja, dado um número N de lançamentos de uma moeda qualquer, sendo esse número suficientemente grande, chega-se a um resultado já conhecido de 50% de chance para cara ou para coroa.

Se o número de lançamentos com resultado cara for dado por N_{cara} e o resultado coroa por N_{coroa} , podemos concluir que a *probabilidade* associada a cada um desses eventos é dada por $P(cara) = \frac{N_{cara}}{N}$ e $P(coroa) = \frac{N_{coroa}}{N}$. Assim, temos que

$$P(cara) + P(coroa) = 100\%. \quad (2.10)$$

É importante salientar que, sendo a moeda não viciada, os dois possíveis eventos são igualmente prováveis e a probabilidade de 50% para cada um dos eventos só ocorrerá para $N \rightarrow \infty$.

Antes de prosseguir, há alguns conceitos básicos a serem definidos. O primeiro deles é o *espaço amostral*, usualmente denotado por Ω , e que contém todos os possíveis resultados de um experimento. Cada elemento do espaço amostral é chamado de *ponto amostral*. Podemos ter subconjuntos do espaço amostral, que são denominados de *eventos* (WASSERMAN, 2004).

No exemplo da moeda, o espaço amostral é $\Omega = \{cara, coroa\}$, e *cara* e *coroa* são *pontos amostrais* desse espaço. No caso de lançarmos duas moedas seguidas, o espaço amostral é $\Omega = \{(cara, cara), (cara, coroa), (coroa, cara), (coroa, coroa)\}$. Agora, temos quatro pontos amostrais distintos, a saber, $(cara, cara), (cara, coroa)$, etc. Quando falamos da possibilidade das duas moedas terem lados iguais, ou seja $A = \{(cara, cara), (coroa, coroa)\}$, temos que A denota o evento em que as duas moedas tem lados iguais.

Nesse ponto, é interessante definirmos *experimento aleatório*, que é um experimento que apresenta resultados diferentes quando executado diversas vezes, mesmo tendo sido executado de maneira semelhante (WASSERMAN, 2004). No caso do lançamento de uma única moeda, temos dois resultados igualmente prováveis, mas não é possível determinar o resultado de cada lançamento antecipadamente.

Os conceitos até então discutidos podem ser facilmente generalizados. Dado um experimento realizado N vezes, a probabilidade de que um evento A_i ocorra é dada por

$$P(A_i) = \lim_{N \rightarrow \infty} \frac{A_i}{N}. \quad (2.11)$$

Tem-se as seguintes propriedades para uma função de conjuntos $P(A) = f : \Omega \rightarrow \mathbb{R}$, com $P[A] \in [0, 1]$, onde pode-se representar uma probabilidade por (CESAR, d):

- $P(A) \geq 0 \quad \forall A \in \Omega$;
- $P(\Omega) = 1$, Ω é chamado de evento certo;
- $P(A) = 1 - P(A')$;

onde $A + A' = \Omega$ e dizemos que A' é o evento complementar de A .

Podemos trabalhar com outra característica da probabilidade que se trata de eventos independentes. Os eventos X e Y são independentes se $P(XY) = P(X)P(Y)$. Daí, pode-se mostrar que se X e Y independem entre si, então $(X' \text{ e } Y)$, $(X \text{ e } Y')$ e $(X' \text{ e } Y')$, onde X' e Y' são os complementares de X e Y , também são independentes entre si. Isso significa que os eventos complementares X' e Y' também são independentes.

Prova: Usando $Y = XY + X'Y$, onde $X + X' = 1$ e $P(X) = 1 - P(X')$, percebemos que $P(Y) = P(XY) + P(X'Y)$. Logo, $P(X'Y) = P(Y) - P(XY)$. Como X e Y são independentes, então $P(X'Y) = P(Y) - P(X)P(Y) = P(Y)[1 - P(X)] = P(Y)P(X')$, provando que X' e Y são independentes. Chamando X' de Y' e Y de X , temos que X e Y' são independentes. Se X' e Y são independentes, então mudando Y para Y' temos que X' e Y' são independentes.

2.2.1 Valor Esperado, Momentos e Covariância

Realizar operações com funções de conjuntos torna-se complicado, sendo mais fácil o tratamento com funções numéricas. Assim, é interessante criar uma associação entre os eventos e a probabilidade deles ocorrerem a partir de uma função $f : \mathbb{R} \rightarrow [0, 1]$ que permite associar uma probabilidade a cada evento. Para tanto, definimos uma *variável aleatória* (v.a.) como sendo um mapeamento $x : \Omega \rightarrow \mathbb{R}$ que associa um número real $x(\omega)$ a cada ponto amostral ω (WASSERMAN, 2004). Desta forma, poderemos trocar $P(A)$ por $P(x)$, onde x é uma variável aleatória. A função $f(x)$ é uma função de distribuição de probabilidade, logo, obedece as seguintes propriedades (PAPOULIS, 1991):

- $f(x) \geq 0$;
- $\int_{-\infty}^{\infty} f(x)dx = 1$;
- $P(x < x_i \leq x + dx) = f(x)dx$. Portanto, $f(x)dx$ é a probabilidade de encontrar a v.a. x_i no intervalo $(x, x + dx)$.

A *Esperança* ou *Valor Esperado* de uma variável aleatória é dada por:

$$E[x] := \int_{-\infty}^{\infty} xf(x)dx, \quad (2.12)$$

sendo a notação $\langle x \rangle = E[x]$ muito usual na física.

Às vezes, também se usa $\bar{x}$ para denotar o valor esperado, embora seja necessário tomar cuidado, já que a média é obtida em relação a amostragem ao invés de todo o espaço amostral. Apesar disso, $\bar{x}$ é uma boa inferência para $E[x]$, uma vez que a média da amostra

se aproxima da média da distribuição para espaços amostrais muito grandes. O fato de que a média $\bar{x}$ converge para o valor esperado $E[x]$, é conhecido em estatística como Lei dos Grandes Números (WASSERMAN, 2004).

Resta saber qual o valor esperado de uma função $g(x)$ onde x é uma variável aleatória. Daí, partimos da seguinte variável aleatória: $y = g(x)$, então $E[y] = \int_{-\infty}^{\infty} yf(y)dy$. Mas, assumindo que a probabilidade contida numa zona diferencial deve ser invariante sob mudança de variáveis, isto é, $|f(y)dy| = |f(x)dx|$, então,

$$f_Y(y) = \left| \frac{dx}{dy} \right| f_X(x) = \left| \frac{d}{dy}(x) \right| f_X(x) = \left| \frac{d}{dy}(g^{-1}(y)) \right| f_X(g^{-1}(y)) = \frac{f_X(x)}{|g'(x)|}. \quad (2.13)$$

Como $dy = \frac{dg(x)}{dx}dx = g'(x)dx$. Temos,

$$E[g(y)] = \int_{-\infty}^{\infty} g(x) \frac{f(x)}{|g'(x)|} g'(x)dx = \int_{-\infty}^{\infty} g(x) f(x)dx. \quad (2.14)$$

Verifica-se alguns casos particulares dados por:

- $g(x) = ax$, onde a é uma constante. Então: $E[ax] = \int_{-\infty}^{\infty} axf(x)dx = a \int_{-\infty}^{\infty} f(x)dx = aE[x]$;
- $g(x) = q(x) + h(x)$. Logo $E[q(x) + h(x)] = \int_{-\infty}^{\infty} q(x)f(x)dx + \int_{-\infty}^{\infty} h(x)f(x)dx = E[q(x)] + E[h(x)]$.

Partindo da definição de esperança para uma função de uma variável aleatória, é interessante explorar o caso de uma função $g(x)$ dada por $g(x) = x^n$. Neste caso, o valor esperado leva o nome de *momento de n-ésima ordem*, sendo definido por

$$M_n = E[x^n] = \int_{-\infty}^{\infty} x^n f(x)dx. \quad (2.15)$$

Devido ao comportamento assintótico de x^n quando $x \rightarrow \pm\infty$, devemos impor que a função $f(x)$ caia com uma lei de potência do tipo $f(x) \propto x^{-m}$, de modo que somente existirão os momentos a partir da ordem $n = m - 2$ (PAPOULIS, 1991). É fácil demonstrar as seguintes propriedades dos momentos:

- $M_0 = \int_{-\infty}^{\infty} f(x)dx = 1$;
- $M_1 = \int_{-\infty}^{\infty} xf(x)dx \equiv \mu$.

Devemos notar que $M_1 = E[x]$, logo μ é a média da variável x .

Além da média de x , podemos definir a *variância* σ^2 , que indica a dispersão de uma medida com relação ao seu valor médio. Em outras palavras, μ nos fornece o valor médio de uma medida enquanto σ^2 fornece o quanto uma medida difere desta média. Afim de definir a variância σ^2 , devemos, primeiramente, definir o *momento centrado de n-ésima ordem*:

$$m_n = E[(x - \mu)^n] = \int_{-\infty}^{\infty} (x - \mu)^n f(x)dx. \quad (2.16)$$

Além do momento centrado de ordens 0 e 1, é interessante calcular o momento centrado de ordem 2:

- $m_0 = \int_{-\infty}^{\infty} f(x)dx = 1$;
- $m_1 = \int_{-\infty}^{\infty} (x - \mu)f(x)dx = \int_{-\infty}^{\infty} xf(x)dx - \mu \int_{-\infty}^{\infty} f(x)dx = \mu - \mu = 0$;
- $m_2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \equiv \sigma^2$.

Agora, vamos explorar o valor esperado de uma função $z = g(x, y)$, onde x e y são variáveis aleatórias:

$$E[g(x, y)] = \int_{-\infty}^{\infty} z f_z(z) dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy. \quad (2.17)$$

Esse é o caso mais simples de esperança multivariada. Podemos notar as seguintes propriedades

- $E[k] = k$;
- $E[ag(x, y) + bh(x, y)] = aE[g(x, y)] + bE[h(x, y)]$. Em particular, $E[x + y] = E[x] + E[y]$.

Naturalmente, podemos definir momentos no caso multivariado:

$$M_{mn} = E[x^n y^m] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^n y^m f(x, y) dx dy. \quad (2.18)$$

Segue, imediatamente, que $M_{00} = 1$. Além disso,

$$\begin{aligned} M_{10} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy = \int_{-\infty}^{\infty} x f_x(x) dx = E[x] = \mu_x, \\ M_{01} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy = \int_{-\infty}^{\infty} y f_y(y) dy = E[y] = \mu_y. \end{aligned} \quad (2.19)$$

De maneira análoga, definimos os momentos centrados multivariados

$$m_{nm} = E[(x - \mu_x)^n (y - \mu_y)^m] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^n (y - \mu_y)^m f(x, y) dx dy. \quad (2.20)$$

$m_{00} = 1$, enquanto $m_{01} = m_{10} = 0$.

Os momentos centrados m_{02} , m_{20} e m_{11} são os mais relevantes e, portanto, recebem nomes específicos (CESAR, c):

- Variâncias:

$$\begin{aligned} V(x) &= \sigma_x^2 = m_{02} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x, y) dx dy, \\ V(y) &= \sigma_y^2 = m_{20} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_y)^2 f(x, y) dx dy. \end{aligned} \quad (2.21)$$

- Covariância

$$\text{cov}(x, y) = m_{11} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy, \quad (2.22)$$

Podemos notar que a covariância de uma variável com ela mesma resulta na variância dessa variável: $V(x) = \text{cov}(x, x)$, $V(y) = \text{cov}(y, y)$. Vale citar as propriedades da covariância:

- $\text{cov}(x_1, x_2) = \text{cov}(x_2, x_1)$;
- $\text{cov}(x_1 + x_2, x_3) = \text{cov}(x_1, x_3) + \text{cov}(x_2, x_3)$;
- $\text{cov}(x_1, x_2) = E[x_1 x_2] - E[x_1]E[x_2]$;
- $\text{cov}(ax_1, bx_2) = ab\text{cov}(x_1, x_2)$;
- $\text{cov}(x, k) = 0$.

Ademais, a variância possui as seguintes propriedades:

- $V(x) = E[x^2] - (E[x])^2$;
- $V[kx] = k^2 V[x]$;
- $V[a + bx] = b^2 V[x]$;
- $V[ax_1 + bx_2] = a^2 V[x_1] + b^2 V[x_2] + 2ab\text{cov}(x_1, x_2)$. Em particular, $V[x_1 \pm x_2] = V[x_1] + V[x_2] \pm 2\text{cov}(x_1, x_2)$.

Recorrendo a definição de variáveis independentes, lembramos que se X e Y são independentes, então $P(XY) = P(X)P(Y)$. Neste caso, os eventos $x \in X$ e $y \in Y$ são independentes e tem-se $f(x, y) = f_x(x)f_y(y)$. Consequentemente

$$\begin{aligned} E[xy] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_x(x)f_y(y)dx dy \\ &= \left[\int_{-\infty}^{\infty} xf_x(x)dx \right] \left[\int_{-\infty}^{\infty} yf_y(y)dy \right] = E[x]E[y], \end{aligned} \quad (2.23)$$

e

$$\begin{aligned} \text{cov}(x, y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y)f(x, y)dx dy \\ &= E[xy] - E[x]E[y] = E[x]E[y] - E[x]E[y] = 0. \end{aligned} \quad (2.24)$$

Ou seja, a covariância é não-nula somente se as variáveis forem dependentes entre si.

2.2.2 Coeficiente e distância de correlação

Tendo em mente os conceitos de variância e covariância estabelecidos na seção anterior, vamos definir a matriz de covariância como sendo a matriz $n \times n$ cujas entradas são dadas por $V_{ij} = \text{cov}(x_i, x_j)$ e sua diagonal principal é formada pela variância de cada variável aleatória. Entretanto, quando lidamos com dados reais, é mais adequado utilizar variáveis discretas, de modo que devemos redefinir a matriz de covariância da seguinte forma (CESAR, c)

$$V_{ij} = \frac{1}{n} \sum_k (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j). \quad (2.25)$$

Podemos concluir que a matriz definida acima simétrica ($V_{ij} = V_{ji}$) e é definida positiva.

Como já visto, na seção anterior, a covariância indica o quão dois eventos são independentes. Primeiramente, devemos notar que, se os eventos x e y não são adimensionais, então a correlação entre eles também não será adimensional. Assim, torna-se necessário definir uma grandeza adimensional que esteja relacionada a correlação. Essa grandeza, denominada *coeficiente de correlação*, é definida como (BARABÁSI, 2016)

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{cov}(x, x)}\sqrt{\text{cov}(y, y)}} = \frac{\text{cov}(x, y)}{\sqrt{V[x]V[y]}} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}. \quad (2.26)$$

Além da adimensionalidade, o coeficiente de correlação tem a vantagem adicional de estar dentro do intervalo $[-1, 1]$, onde $+1$ indica que os eventos são positivamente correlacionados, 0 indica que eles são independentes e -1 indica que eles são anti-correlacionados. Para demonstrar essa propriedade, vamos partir da esperança de uma quantidade positiva e utilizar a desigualdade de Schwartz. Dado $E\{[\lambda(x - \mu_x) - (y - \mu_y)]^2\} \geq 0$, onde $\lambda \in \mathbb{R}$, temos

$$\begin{aligned} E\{[\lambda(x - \mu_x) - (y - \mu_y)]^2\} &= \lambda^2 E[(x - \mu_x)^2] - 2\lambda E[(x - \mu_x)(y - \mu_y)] + E[(y - \mu_y)^2] \\ &= \lambda^2 V[x] - 2\lambda \text{cov}(x, y) + V[y] \geq 0. \end{aligned} \quad (2.27)$$

Dada uma desigualdade $a\lambda^2 + b\lambda + c \geq 0$, sabemos que esta admite solução quando $b^2 - 4ac \leq 0$. Isso significa que, para a igualdade presente em (2.27), temos

$$4[\text{cov}(x, y)]^2 - 4V[x]V[y] \leq 0 \Rightarrow \frac{[\text{cov}(x, y)]^2}{V[x]V[y]} \leq 1 \Rightarrow -1 \leq \frac{\text{cov}(x, y)}{\sqrt{V[x]V[y]}} \leq 1 \quad (2.28)$$

Assim, fica provado que o coeficiente de correlação obedece $-1 \leq \rho_{xy} \leq 1$.

O próximo conceito a ser abordado é o de *distância de correlação*. Antes, precisamos definir o que é distância. Dado um espaço S , dotado de uma função $d(x, y) : S \times S \rightarrow \mathbb{R}$, para $\forall x, y \in S$, dizemos que o par (d, S) é um espaço métrico se $d(x, y)$ obedece aos axiomas:

- $d(x, z) \leq d(x, y) + d(y, z)$ (Desigualdade Triangular),
- $d(x, y) = 0$ se, e somente se, $x = y$,
- $d(x, y) = d(y, x)$ (Simetria).

A função $d(x, y)$ é denominada distância.

A função distância mais familiar é a euclidiana, definida por

$$d_E(x, y) = \sqrt{\sum_{i=0}^n (x_i - y_i)^2}, \quad (2.29)$$

onde x e y são vetores em $R^n := \{(x_i) = ((x_1, \dots, x_n)) | x_i \in R, i = 1, \dots, n\}$. Podemos notar que d_E obedece aos axiomas de distância de modo que (R^n, d_E) , ou simplesmente $\mathbb{R}^n$ é um espaço

métrico, conhecido como *espaço euclidiano*. O espaço euclidiano possui uma operação $\langle \cdot, \cdot \rangle$, conhecida como *produto interno*, definida como $\langle x, y \rangle := \sum_{i=0}^n x_i y_i$ e que obedece aos axiomas:

- $\langle x, y \rangle = \langle y, x \rangle$;
- $\langle ax, y \rangle = a \langle x, y \rangle$;
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$;
- $\langle x, x \rangle \geq 0$. A igualdade é válida se, e somente se, $x = 0$.

A partir do produto interno euclidiano, podemos ainda definir a norma $\| \cdot \|$ ou comprimento de um vetor x como $\|x\| = \sqrt{\langle x, x \rangle}$. A norma obedece a

- $\|xy\| \leq \|x\| \|y\|$ (Desigualdade de Cauchy);
- $\|x + y\| \leq \|x\| + \|y\|$ (Desigualdade Triangular).

Podemos notar que $d_E(x, y) = \|x - y\|$. Portanto, a distância entre dois pontos em $\mathbb{R}^n$ equivale ao comprimento do vetor que liga esses dois pontos.

O produto interno e distância podem ser generalizados a outros espaços métricos. Em particular, podemos definir o produto interno e a distância entre funções. Dados uma função peso $\omega(x) : \mathbb{R} \rightarrow \mathbb{R}_+^*$ e funções $\phi_i(x) : \mathbb{R} \rightarrow \mathbb{C}$, $\forall x \in [a, b]$, o produto interno das funções ϕ_i e ϕ_j é dado por

$$\langle \phi_i, \phi_j \rangle = \int_a^b \phi_i^*(x) \phi_j(x) \omega(x) dx. \quad (2.30)$$

A exigência de que $\omega(x) \geq 0$ é equivalente a imposição feita sobre a função densidade de probabilidade $f(x)$. Além disso, podemos exigir que a norma seja normalizada através da imposição de que $\int_a^b \omega(x) dx = 1$ ou, equivalentemente, que $\int_a^b \omega'(x) dx = 1$ onde $\omega(x) / \int_a^b \omega(x) dx = 1$. Assim, se identificarmos a função peso $\omega(x)$ como sendo uma função densidade de probabilidade $f(x)$ no intervalo $[a, b]$, podemos ver que

$$\langle \phi_i, \phi_j \rangle = \int_a^b \phi_i^*(x) \phi_j(x) f(x) dx, \quad (2.31)$$

corresponde a um produto interno. Além disso, podemos definir a distância entre as funções ϕ_i e ϕ_j como

$$d(\phi_i, \phi_j) = \int_a^b \|\phi_i(x) - \phi_j(x)\|^2 f(x) dx. \quad (2.32)$$

O produto interno entre duas funções é particularmente útil porque podemos associá-lo a covariância. Partindo da equação (2.22), vemos que $cov(x, y) = \langle x - \mu_x, y - \mu_y \rangle$. Já a equação (2.21) nos fornece $\sigma_x^2 = \langle x - \mu_x, x - \mu_x \rangle$ e $\sigma_y^2 = \langle y - \mu_y, y - \mu_y \rangle$, de modo que $\sigma_x = \|x - \mu_x\|$ e $\sigma_y = \|y - \mu_y\|$. Com isso, podemos interpretar a variância de x como sendo a distância entre x e

sua média μ_x . A covariância, por sua vez, está associada ao produto interno. Assim, quando dois eventos são independentes, eles são também perpendiculares entre si.

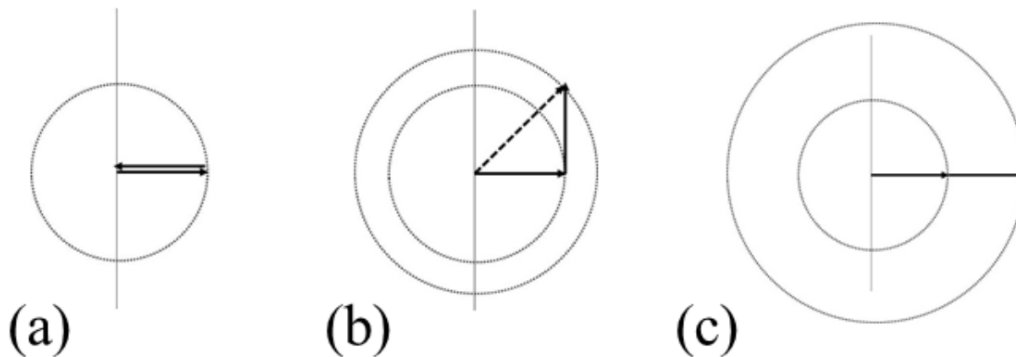
Se definirmos $p = \frac{x - \mu_x}{\sigma_x}$ e $q = \frac{y - \mu_y}{\sigma_y}$, podemos notar que $E[p] = E[q] = 0$ e $E[p^2] = E[q^2] = 1$, portanto, p e q são vetores unitários. Tendo em mente a definição de distância dada pela equação (2.32), podemos definir a distância entre x e y como

$$\begin{aligned} d_{xy}^2 &= \int_a^b \|p - q\|^2 f(x) dx = E[(p - q)^2] = E[p^2] + E[q^2] - 2E[pq] \\ &= 2(1 - \rho_{xy}), \end{aligned} \quad (2.33)$$

onde utilizamos o fato de que $\rho_{xy} = E[pq]$. Assim, temos uma grandeza $d_{xy} = \sqrt{2(1 - \rho_{xy})}$ que possui as propriedades de distância e que está contida no intervalo $[0, 2]$. Chamamos esta grandeza de *distância de correlação*.

Devemos enfatizar que, quanto maior a correlação entre x e y , menor a distância de correlação entre eles. Por ser uma distância, espera-se que $d_{xy} = 0$, quando $x = y$. Neste caso, temos $\rho_{xy} = 1$, o que indica que a correlação é positiva, mas não necessariamente que $x = y$. Na verdade, como as variáveis p e q são idênticas, podemos realizar um experimento em duas amostras diferentes e medir uma distância de similaridade entre as duas por meio da distância de correlação: $\rho_{xy} = 1$ ($d_{xy} = 0$) indica similaridade total; $\rho_{xy} = 0$ ($d_{xy} = \sqrt{2}$) indica independência (ou perpendicularidade) e $\rho_{xy} = -1$ ($d_{xy} = 2$) indica oposição entre as amostras. Esses casos estão esquematizados na Figura 8, onde vemos a subtração e soma de vetores unitários.

Figura 8 – Soma de vetores representando distâncias de correlação.



Fonte: (CESAR, c). a) para $d_{xy} = 0$; b) para $d_{xy} = \sqrt{2}$ e c) $d_{xy} = 2$.

Ao se obter a matriz de distância de correlação entre diferentes amostras, torna-se possível fazer um MST de similaridade entre todas essas amostras e agrupar as suas relações

pelo ordenamento crescente das suas distâncias. Assim, utilizaremos esse método para investigar os agrupamentos dos mercados financeiros, nos aproveitando que o termo d_{ij} nos fornece o peso da aresta entre um vértice i e um outro vértice j . Então, podemos gerar um grafo a partir da distância de correlação entre esses ativos financeiros. Sendo esse grafo completo e com todos os vértices conectados.

3 SÉRIES TEMPORAIS DO MERCADO FINANCEIRO

A análise de dados provenientes de séries temporais tem crescente importância devido a produção massiva de dados a cada instante de tempo, como por exemplo, informações vinculadas a internet: compras realizadas online, páginas acessadas; dados financeiros; fenômenos científicos e dados comportamentais. Por esses e vários outros motivos, há uma crescente busca por quantidade, qualidade e representatividade das séries temporais coletadas, com técnicas estatísticas e de aprendizado de máquina (NIELSEN, 2009).

O termo séries temporais é empregado quando se tem como objetivo extrair um resumo completo de informações a partir de pontos organizados em ordem cronológica. É feito para representar o comportamento passado, a fim de prever o comportamento futuro ou possuir um modelo com alta representatividade de certo fenômeno. As séries temporais podem ser do tipo estacionária: quando seus valores estão dispostos aleatoriamente no tempo, mas ao redor de uma média constante; ou do tipo não-estacionárias: quando estão sujeitas a um processo estocástico como, por exemplo, o movimento do preço de um ativo financeiro.

Além de termos diversas outras aplicações encontradas na medicina, na análise climática, economia e astronomia, o estudo de séries temporais e redes complexas no ramo da física estatística desenvolve-se direcionalmente para séries temporais não-estacionárias, por possuírem peculiaridades que são de difícil manipulação.

Neste trabalho abordaremos o estudo de séries temporais provenientes do mercado financeiro e para isso, este capítulo tem como finalidade a apresentação das características de séries temporais de grupos financeiros, sua definição e o comportamento do preço de um determinado ativo em função do tempo, como realizar uma análise multivariadas a partir de diferentes escalas e torná-las séries estacionárias e uma demonstração para as ações do *SP500*.

3.1 O que são ações?

Mercados financeiros são sistemas nos quais um grande número de operadores interagem entre si e reagem a informações externas para determinar o melhor preço a ser dado em um produto. Os principais operadores podem ser classificados em:

- Empreendedores: empresários que procuram o mercado financeiro como meio de financiamento de seus empreendimentos;
- Especuladores: são aqueles interessados somente no retorno financeiro gerado pela varia-

ção assimétrica dos preços de um determinado ativo;

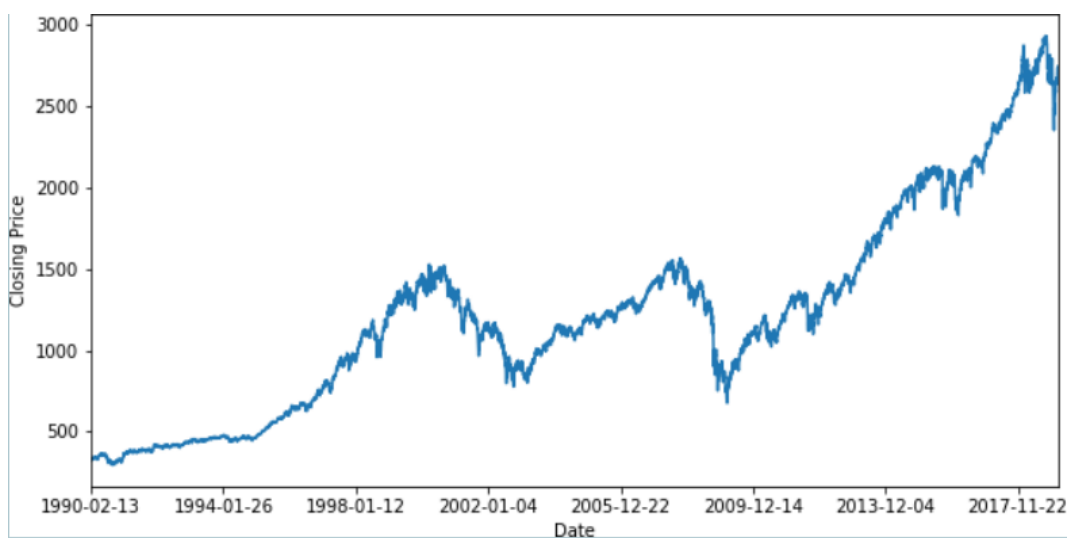
- Corretoras: instituições que recebem comissão por garantir a execução de ordens dos participantes no mercado financeiro.

Os mercados financeiros são acessíveis a operadores em todo mundo e os itens que podem ser negociados nesses mercados incluem *commodities* - produtos agrícolas como soja e milho e entre outros; ouro e outros minerais; moedas; ações e títulos de dívida pública (*Bonds*).

As ações tem como objetivo primário o financiamento de obras e empresas. Uma ação é um título que representa uma fração ou cota de participação numa companhia. Sendo assim, temos dois perfis de acionistas: aqueles que querem ser sócios de uma determinada empresa e aqueles que buscam obter lucro na valorização do ativo financeiro.

As ações, assim como outros ativos de mercado financeiro, possuem um preço dinâmico que muda constantemente no tempo e exibem tendência, ciclos, padrões sazonais e outros comportamentos não-estacionários. Quando observamos a evolução temporal do preço, volume ou número de transações de um produto financeiro, chegamos a conclusão de que essa evolução é imprevisível.

Figura 9 – Série Temporal dos preços de fechamento do índice *S&P500*.



Fonte: (NIELSEN, 2009). Plot da evolução dos preços de fechamento do índice *S&P500* durante o período de 1990 a 2017, onde é possível visualizar o movimento em tendência das séries temporais do mercado financeiro.

Na figura 9, temos o gráfico da evolução temporal do preço de fechamento diário do índice *S&P500*, que consistem em uma carteira teórica contendo quinhentas ações dentre as mais representativas do mercado norte-americano. É possível observar que, durante a

série histórica do preço, há a formação de tendências momentâneas: de alta, de baixa ou de lateralização. Entretanto, não se tem um μ constante que represente o comportamento global desse índice caracterizando, assim, uma série temporal não-estacionária. De modo geral, essa não-estacionariedade se reproduz em todos os outros ativos do mercado financeiro uma vez que as transações são realizadas por pessoas, as quais interpretam de forma particular uma notícia externa, tornando o mercado financeiro um ambiente probabilístico, pois, não é possível prever o comportamento futuro.

3.2 Retorno e Log-Retorno

Tendo em vista o fato de que as operações no mercado financeiro geram uma receita (positiva ou negativa), então temos o investimento de um capital inicial para, ao final de um determinado período de tempo Δt , obtermos essa receita. Supondo que essa receita é positiva, devemos recuperar o capital inicial $\$0$ investido mais uma *retorno* R sobre o mesmo, resultando em $\$_{\Delta t} = \$0 + R\$0 = (1 + R)\0 . Se reinvestirmos esse montante n vezes a juros simples teremos, após n períodos, $\$n = (1 + nR)\0 , enquanto que a juros compostos teremos $\$n = Z^n\0 , onde $Z = (1 + R)$. No regime de juros simples, o capital cresce de maneira aditiva e, no regime de juros compostos, de maneira multiplicativa (CESAR, b).

Podemos mostrar que o mercado de ações funciona de forma multiplicativa. Vamos supor que o preço de uma ação em um determinado dia i é dado por $\$i$. Naturalmente, ao comprarmos a ação no dia 0, pagamos $\$0$ por ela. No dia seguinte a compra, temos uma diferença de preço igual a $\$1 - \0 , de modo que o retorno R sobre a ação é dado por $(\$1 - \$0)/\$0$. Consequentemente, temos $Z_1 \equiv \$1/\0 . No n -ésimo dia, teremos (CESAR, a)

$$Z_{total} = \frac{\$n}{\$0} = \prod_{i=1}^n \frac{\$i}{\$_{i-1}} = \prod_{i=1}^n Z_i. \quad (3.1)$$

Portanto, para obtermos o Z de um dado período, devemos multiplicar os Z 's correspondentes a cada período unitário demonstrando, assim, a natureza multiplicativa do mercado financeiro.

Quando queremos analisar individualmente as variáveis aleatórias do mercado, precisamos utilizar as propriedades do *logaritmo* para converter produtórios em somatórios. Esse mecanismo é necessário para entender melhor o mercado financeiro, uma vez que esse cresce exponencialmente. Em primeiro lugar, trabalhar com retornos no lugar de preços tem como vantagem a normalização: pode-se medir todas as variáveis numa métrica compatível, possibilitando o desenvolvimento de relações entre variáveis como, por exemplo, a covariância

(equação (2.22)), mesmo que estas se originem de séries de preços com valores distintos.

Agora, definimos o *log-retorno* r como $r = \ln(1 + R)$. Para recuperar o retorno a partir do log-retorno, utilizamos $R = e^r - 1$. A utilização do log-retorno tem vantagens tanto teóricas quanto computacionais. Sempre que os retornos forem muito pequenos, podemos fazer a aproximação $r \approx R$. Adicionalmente, numa sequência de operações, temos um retorno dado pela equação (3.1), enquanto o log-retorno é dado por uma expressão aditiva

$$r_{total} = \sum_{i=1}^n r_i = \ln \$n - \ln \$0. \quad (3.2)$$

Assim, podemos facilmente obter o log-retorno após n períodos sabendo-se o preço no instante n e o preço inicial de um ativo. A vantagem computacional reside justamente em substituímos uma operação de multiplicação de n termos por uma adição, o que se mostra muito útil para valores grandes de n .

3.3 Análise de séries temporais

A análise de séries temporais aplicada a mercado financeiro possui algumas dificuldades, entre elas a interpretação das quantidades probabilísticas e o fato de o mercado não exibir séries temporais estacionárias. Vamos definir, formalmente, o que são séries temporais estacionárias. Dada uma série temporal $\{X_t\}$, dizemos que esta é estacionária se satisfaz, para todos os inteiros t e h , as seguintes condições (ZHANG, 2009)

- (a) $Var(X_t) < \infty$;
- (b) $E[X_t]$ não depende de t ;
- (c) $Cov(X_t, X_{t+h})$ não depende de t .

Como já citado, o mercado financeiro é um exemplo de séries temporais não-estacionárias. No entanto, ainda é interessante investigarmos a dinâmica relacionada a essas séries, ou seja, sua evolução no tempo, pois o mercado financeiro é um complexo sistema de interação no qual observamos períodos de alta atividade e períodos relativamente calmos. Apesar da complexidade das interações, podemos identificar algumas características universais como, por exemplo, longas tendências de alta de preço que, por sua vez, são revertidas por *crashes* sendo os mais famosos e atuais: Crise do Pontocom (2000), Crise do Subprime (2008) e Crise Sanitária do Covid-19 (2020).

A análise dos ativos é dada a partir do preço P de fechamento de cada ativo em um dia de pregão (sendo usual o *range* de 252 dias). Assim, para cada dia de negociação t , temos

que o log-retorno do i -ésimo ativo é dado pelo preço de fechamento em t e $t - 1$ (dia anterior): $r_i(t) = \ln P_i(t) - \ln P_i(t - 1)$. A partir da obtenção dos preços normalizados pelo log-retorno, podemos construir a matriz de coeficientes de correlação entre os ativos:

$$\rho_{ij}^t = \frac{\langle r_i^t r_j^t \rangle - \langle r_i^t \rangle \langle r_j^t \rangle}{\sqrt{[\langle r_i^{t2} \rangle - \langle r_i^t \rangle^2][\langle r_j^{t2} \rangle - \langle r_j^t \rangle^2]}}. \quad (3.3)$$

É importante salientar que a existência de correlação não indica uma relação de causalidade, ou seja, a correlação entre dois eventos ou fenômenos não significa que um ocorre como resultado do outro (CORRELATION...). Na prática, estabelecer relações de causa e efeito é muito mais complexo do que estabelecer correlações e é preciso uma análise cautelosa. Exemplos de situações em que a correlação claramente não corresponde a causalidade podem ser encontrados em (SPURIOUS...).

Séries temporais podem ser modeladas como redes complexas (GAO Z.-K.; SMALL, 2016; BARIGOZZI; HALLIN, 2016). Propriedades estatísticas como conectividade e caminho mínimo médio definem a topologia e a dinâmica de evolução de uma rede. Para séries temporais de preços (AN *et al.*, 2018), é comum vermos a análise de matrizes de correlação para um conjunto de ativos (CHI *et al.*, 2010; TUMMINELLO *et al.*, 2010). A partir das matrizes de correlação, podemos construir redes nas quais os ativos são tratados como vértices de um grafo, enquanto as relações entre pares são determinadas por meio de arestas. Neste sentido, as matrizes de correlação são importantes não só para a análise e visualização de redes como também servem para conectar a análise de redes de sistemas financeiros com as teorias financeiras já existentes.

Como o coeficiente de correlação (3.3) possui valores entre $[-1, 1]$, teríamos um problema na construção da MST com pesos negativos. Assim, torna-se necessário a utilização da distância de correlação,

$$d_{ij}^t = \sqrt{2(1 - \rho_{ij}^t)}, \quad (3.4)$$

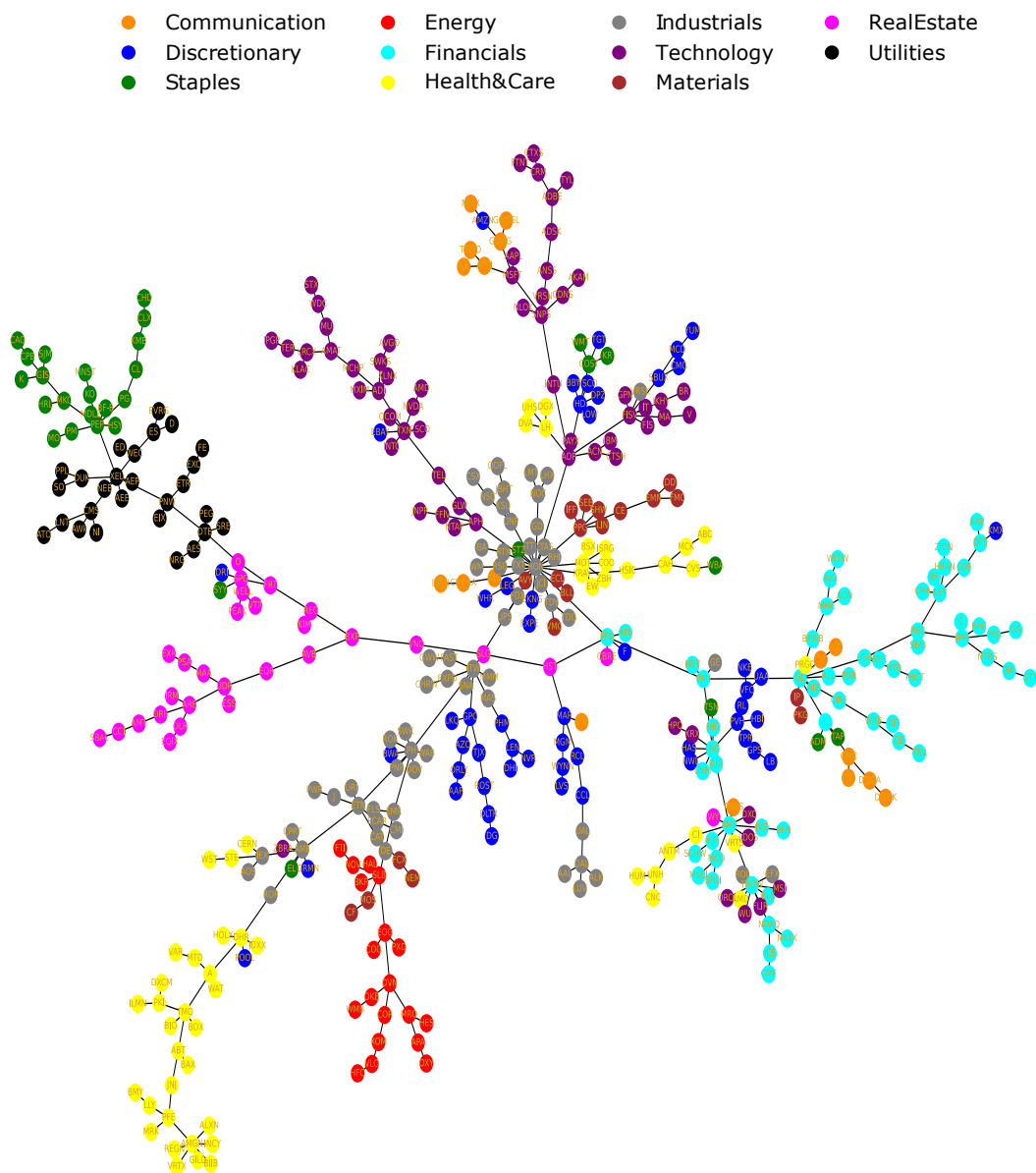
cujos valores estão contidos entre $[0, 2]$.

Por fim, obtemos a MST utilizando o algoritmo de Prim, tendo como *input* as ações representando os vértices e a distância de correlação referente as ações i e j como as arestas ponderadas. A partir da MST, podemos obter as propriedades já citadas no texto. Antes disso, no entanto, vamos seguir o mesmo procedimento para o índice SP500, na próxima seção deste capítulo.

3.4 Análise do SP500

Para a construção da MST na Figura 10, obtivemos os dados dos preços de fechamento através da API do *Yahoo! Finance* utilizando o pacote *yfinance*, pelo *Python*, como descrito em (AROUSSI,). Os dados dessa figura representam as séries temporais de ações entre 01/01/2010 – 31/12/2020, negociadas em dias úteis durante todo o período. A partir dessa informação, calculamos o log-retorno do preço de fechamento e obtivemos a matriz de distância de correlação.

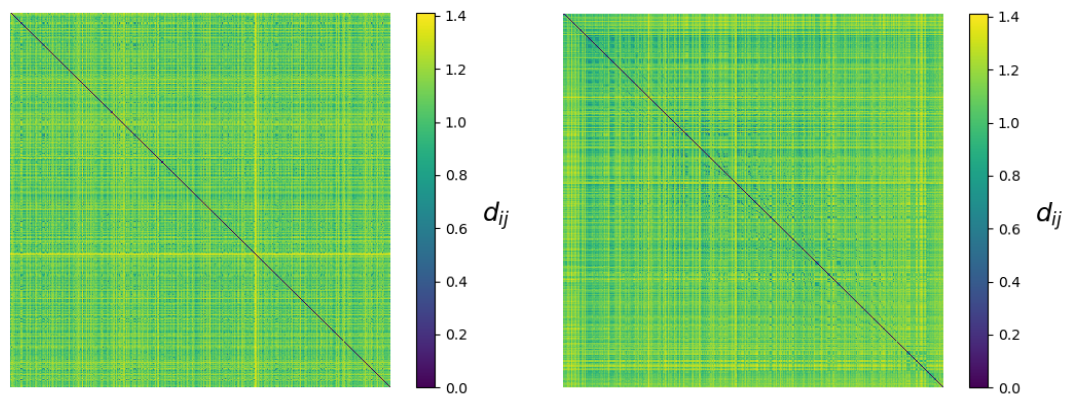
Figura 10 – MST do Índice *S&P500* para os anos de 2010 a 2020.



Fonte: O autor. MST de 431 ações do *S&P500*. A coloração dos vértices representa o setor ao qual a ação pertence.

As ações foram divididas em onze setores: setor de bens de consumo (*Staples*), setor de comunicação (*Communication*), setor imobiliário (*Real Estate*), setor de Consumo não básico (*Discretionary*), setor de energia (*Energy*), setor financeiro (*Financials*), setor de indústria (*Industrials*), setor de materiais (*Materials*), setor de saúde (*Health&Care*), setor de tecnologia (*Technology*) e Setor de Utilidades (*Utilities*).

Figura 11 – Mapas de calor da matriz de distância de correlação e da matriz de distância de correlação ordenada pela MST.



Fonte: O autor. Mapas de calor da distância de correlação das ações ordenadas em ordem alfabética (esquerda) e ordenadas pela MST (direita).

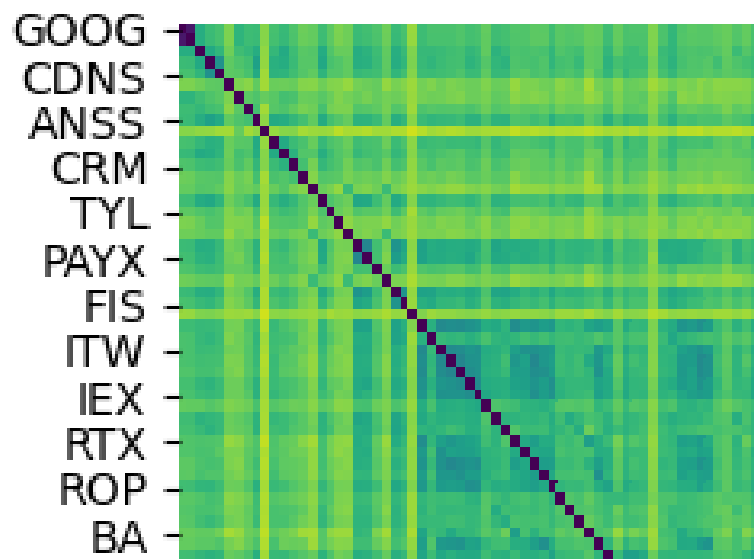
Para alguns setores, podemos notar que, mesmo partindo da matriz de distância de correlação ordenada alfabeticamente, pela MST, recuperamos o aglomerado que representa esse setor. No entanto, como há a possibilidade de crises durante o período analisado, é possível que alguns dos setores estejam espalhados pela rede da MST e com uma distância de correlação menor com ações de setores distintos. Os setores de tecnologia e comunicações são os que mais apresentaram esse tipo de comportamento.

Outra forma de visualizarmos a formação de aglomerados é através do mapa de calor. Com isso em mente, plotamos o mapa de calor da matriz de distância de correlação ordenada alfabeticamente para remover o viés de setor no qual as ações foram obtidas (Figura 11- esquerda). A partir da MST (Figura 10)

Na Figura 12, observamos a formação de alguns aglomerados onde a cor representa a distância de correlação. As regiões com cor mais próxima do azul estão com menor distância de correlação, logo, são mais correlacionadas entre si. A diagonal principal (mais escura) representa a correlação de ação com ela mesma, ou seja, temos distância de correlação nula.

É interessante analisarmos as propriedades de grau e caminho mínimo na MST e

Figura 12 – Ampliação do mapa de calor da matriz de distância de correlação ordenada pela MST.



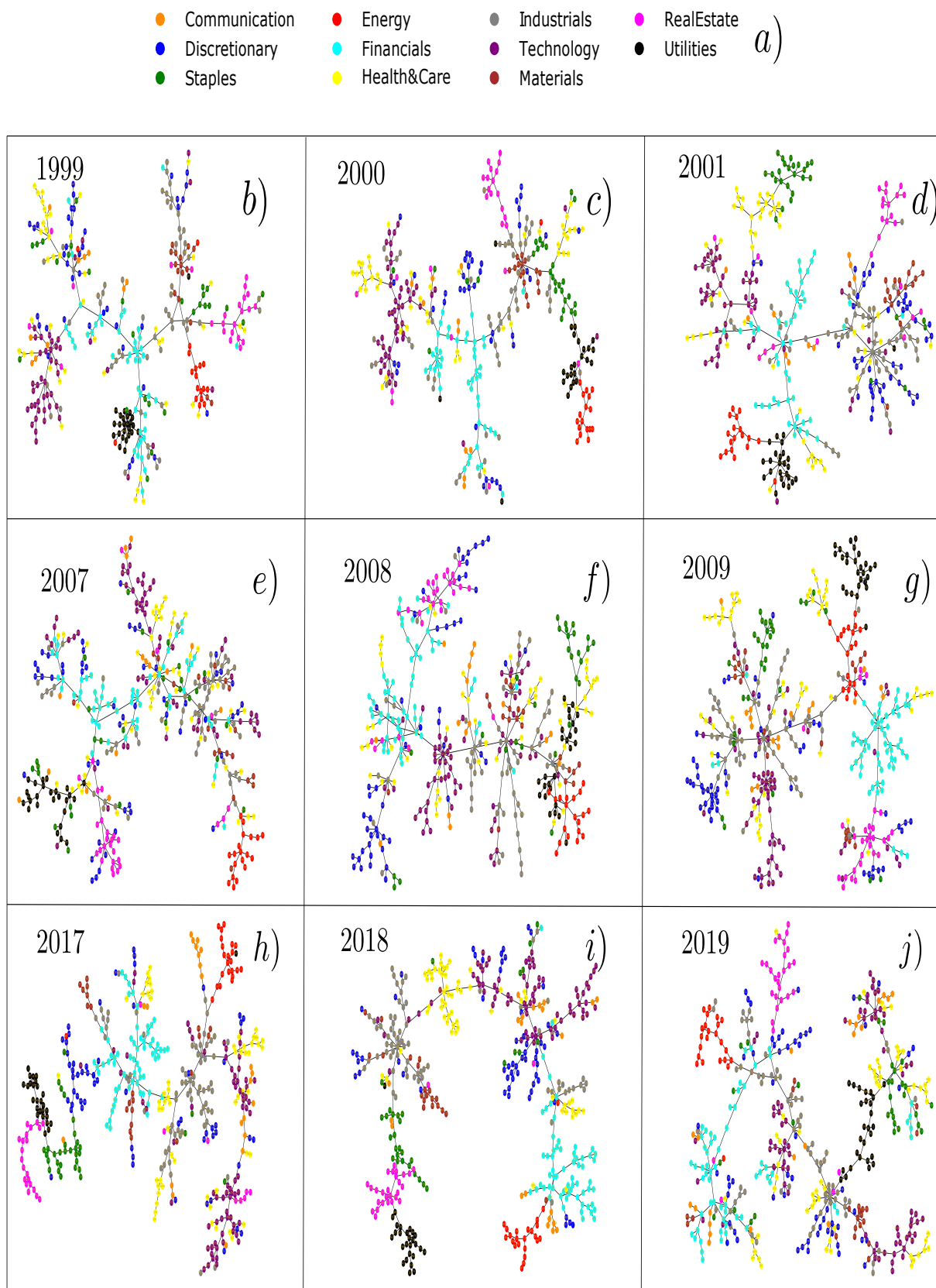
Fonte: O autor. Mapa de calor da distância de correlação das ações ordenadas pela MST.

como estes evoluem no tempo. Com esse objetivo, analisamos cada ano do período individualmente e medimos o *grau médio total* e o *caminho mínimo médio total*, por setor. O grau médio total de um determinado setor foi obtido ao analisar, individualmente, as ações desse setor e a sua vizinhança para com todas as outras ações da MST, somar sobre todos os vizinhos e normalizar esse valor pelo número de ações do setor.

A finalidade dessa análise é verificar se os aglomerados mantêm a característica de *cluster* durante a evolução temporal, se há muitas folhas – existe uma ação muito representativa para o setor e as demais que o compõem estão correlacionadas com ela, – se há muitos galhos – os aglomerados do setor são, predominantemente, lineares.

Obtivemos as MSTs de séries temporais anuais, dentro do período citado. Na Figura 13, temos a representação de algumas dessas MSTs. Podemos notar que a topologia dessas MSTs muda de ano a ano, reafirmando a natureza dinâmica do mercado.

Figura 13 – MSTs para o Índice *S&P500*.



Fonte: O autor. MSTs do *S&P500* para alguns anos.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, apresentamos os resultados e discutimos a análise dos dados coletados através da aplicação do algoritmo de *Prim* para a construção da MST contendo as séries temporais do mercado financeiro. Para isso, analisamos a formação de aglomerados para cada região ou tipo de ativo que compõe os dados macroeconômicos analisados, que chamaremos de **grupos financeiros**.

Começamos descrevendo os dados coletados e tratados, na primeira seção e, na seção, seguinte, descrevemos importantes eventos financeiros que estão compreendidos no período considerado. Na terceira seção, analisamos o comportamento médio das séries temporais dos grupos financeiros. Na quarta seção, analisamos a estrutura dos agrupamentos formados a partir da análise individual de cada ativo e como se daria a relação geral de cada grupo, para isso, utilizamos propriedades embasadas em estatísticas sobre a matriz de correlação num período de 10 anos.

4.1 Dados

Os dados utilizados foram coletados do site *TradingView* (TRACK. . . ,) através da API *TvDatafeed* (TVDATAFEED. . . ,) que habilita o *download* dos dados históricos de cada série temporal presente no *TradingView*, pelo *Python*. O conjunto de dados representam as séries temporais de: índices, índices setoriais, dados de *commodities*, moedas e *bonds*. Esses dados, juntamente com os algoritmos utilizados nesta análise, podem ser encontrados em (AVLIS,).

O estudo dos dados foi realizado no período de 10 anos, que abrange as informações entre 2 de setembro de 2012 a 2 de setembro de 2022. Os dados possuem um total de 524 semanas de cotações com o preço de fechamento semanal e que foram negociadas no mercado durante todo o período. No total, temos 602 distribuídos entre grupos financeiros. Os grupos financeiros foram divididos em doze regiões/agrupamentos financeiros:

- América do Norte (América do Norte, na sigla em inglês (NA)): Estados Unidos e Canadá. Possui dados de índices como *S&P500*, *Dow Jones Industrial* (DJI), *New York Stock Exchange* (NYSE), *NASDAQ* e *Toronto Composite Index* (SP/TSX).
- América Latina e México (América Latina e México, na sigla em inglês (LATAM)): Contém os índices do Brasil (seu índice principal IBOSVSPA e índices setoriais), México (IPC ME), Peru (SP/BVL), Chile (SP IPSA) e Colômbia (ICAP). Esses são os poucos

índices com registro válido para os países da América Latina que não possuem depreciação da moeda por descontrole inflacionário.

- Europa Ocidental (Europa Ocidental, na sigla em inglês (WE)): Possui dados das maiores empresas da União Europeia, representados pelo índice STOXX, além de dados do Reino Unido (FTSE 100), Alemanha (DAX), França (CAC40), entre outros.
- Europa Oriental (Europa Oriental, na sigla em inglês (EE)): Compreende dados da Rússia (IMOEX), Polônia (WIG20), Países Bálticos, entre outros.
- Oriente Médio e África (Oriente Médio e África, na sigla em inglês (MENA)): Inclui Emirados Árabes Unidos (FADG), Israel (TA35), Catar (GNRI), África do Sul (SA40), entre outros.
- China (China (CN)): Contém o principal índice chinês *Shanghai Stock Exchange* - SSE e outros índices setoriais do país.
- Ásia e Pacífico (Ásia e Pacífico, na sigla em inglês (APAC)): Contém dados de países da Ásia, como Japão (NI225), Coreia do Sul (KOSPI), Índia (NIFTY), Hong Kong (HSI) bem como países da Oceania, como Austrália (SP/ASX) e Nova Zelândia (NZX).
- Moedas (Abreviação para moedas (*currencies* em inglês) (CUR)): Compreende índices de moedas com maior liquidez e representatividade no mercado internacional. Tais índices são criados pra medir o valor de uma moeda em relação a outras (*benchmark*), como por exemplo, DXY (dólar americano), EXY (euro), BXY (libra esterlina), JXY (Yen japonês), etc.
- *Foreign Exchange Rates* (*Foreign Exchange Rates* (FOREX)): Mercado de negociação de pares de moedas.
- Metais (METALS): Índices que representam os mercados futuros de metais, tais como cobre, ferro, alumínio, ouro e prata.
- Commodities (Abreviação para *commodities* (COMM)): Contém índices de matérias primas que são negociadas em mercado futuro como, por exemplo: soja, milho, boi gordo, café, trigo, petróleo e outros.
- Bonds (BONDS): Composto com séries temporais das taxas de juros de títulos públicos de 10 anos, para cada país, representados pelos seus respectivos índices em NA, LATAM, WE, EE e etc.

Os dados foram escolhidos para que obtivéssemos o maior número possível de componentes, numa janela de 10 anos. Essa escolha foi feita com base nos dados disponíveis

e também para analisar o comportamento de alguns grupos macroeconômicos nos últimos dez anos. Esse período comporta um ciclo de alta, após a Grande Recessão 2008 – 2012 gerada pela crise do *Sub-Prime* (2007 – 2008) e reestruturação do sistema financeiro, além de algumas correções significativas no período de 2015, 2018 e uma grande crise sanitária (2019 - *até o momento*) – *Covid-19*.

4.2 Eventos e Ciclos

Com base no período e nos ativos escolhidos, devemos enfatizar alguns eventos financeiro importantes (KENTON, ; WILLIAMS, ; ROSS,):

- Crise (da Dívida Pública) da Zona do Euro (2009 – 2014): durante essa crise, alguns países da zona do euro tiveram dificuldade de refinarçar suas dívidas públicas. Grécia, Portugal, Itália e Espanha foram os países mais afetados. Dentre as fatores que levaram à crise, pode-se citar os efeitos da crise do *Sub-Prime*, a crise no mercado imobiliário.
- Crise Russa (2014): Crise motivada pela queda nos preços do petróleo em 2014 e sanções econômicas impostas devido a anexação da Criméia pela Rússia.
- Crise Brasileira (2014-): Crise financeira iniciada devido a queda no preço de *commodities* e exacerbada por eventos políticos que culminaram no *impeachment* da então presidente Dilma Rousseff, em 2016, e pela crise sanitária do Covid-19.
- Crise no Mercado Financeiro Chinês (2015): *Crash* iniciado em junho de 2015 como consequência do estouro da bolha financeira inflacionária gerada no mercado chinês devido a política expansionista do governo chinês no período de 2010 – 2014, pós crise do *Sub-Prime*.
- Desaleração Econômica (2018): Evento motivado por mudanças na curva de juros, incertezas no mercado de *commodities* e aumento de inflação.
- Pandemia do Covid-19 (2020-): Crise sanitária causada pelo **SARS-CoV-2** que levou a restrições de oferta e demanda, devido a *lockdowns*, com consequente escassez de alimentos e matéria prima e diminuição da cadeia produtiva global. Além disso, entre fevereiro e abril de 2020, vários mercados globais experienciaram quedas expressivas devido as incertezas e instabilidades causadas pela pandemia (Figura 14).
- Recessão Pós Pandemia e Invasão da Ucrânia pela Rússia (2022-): Devido a políticas públicas para o combate aos efeitos do Covid-19, houve um aumento drástico de impressão de moedas. Isso acarretou em descontrole inflacionário em diversos países, levando os

bancos centrais a elevarem as taxas básicas de juros, como visto no gráfico de *BONDS*, na Figura 14. Somando-se a isso, o conflito iniciado pela Rússia em fevereiro de 2022, ao invadir a Ucrânia, resultou em sanções de vários países à Rússia, além da interrupção na exportação de *commodities*, petróleo e gás dos dois países.

4.3 Análise Exploratória dos Dados

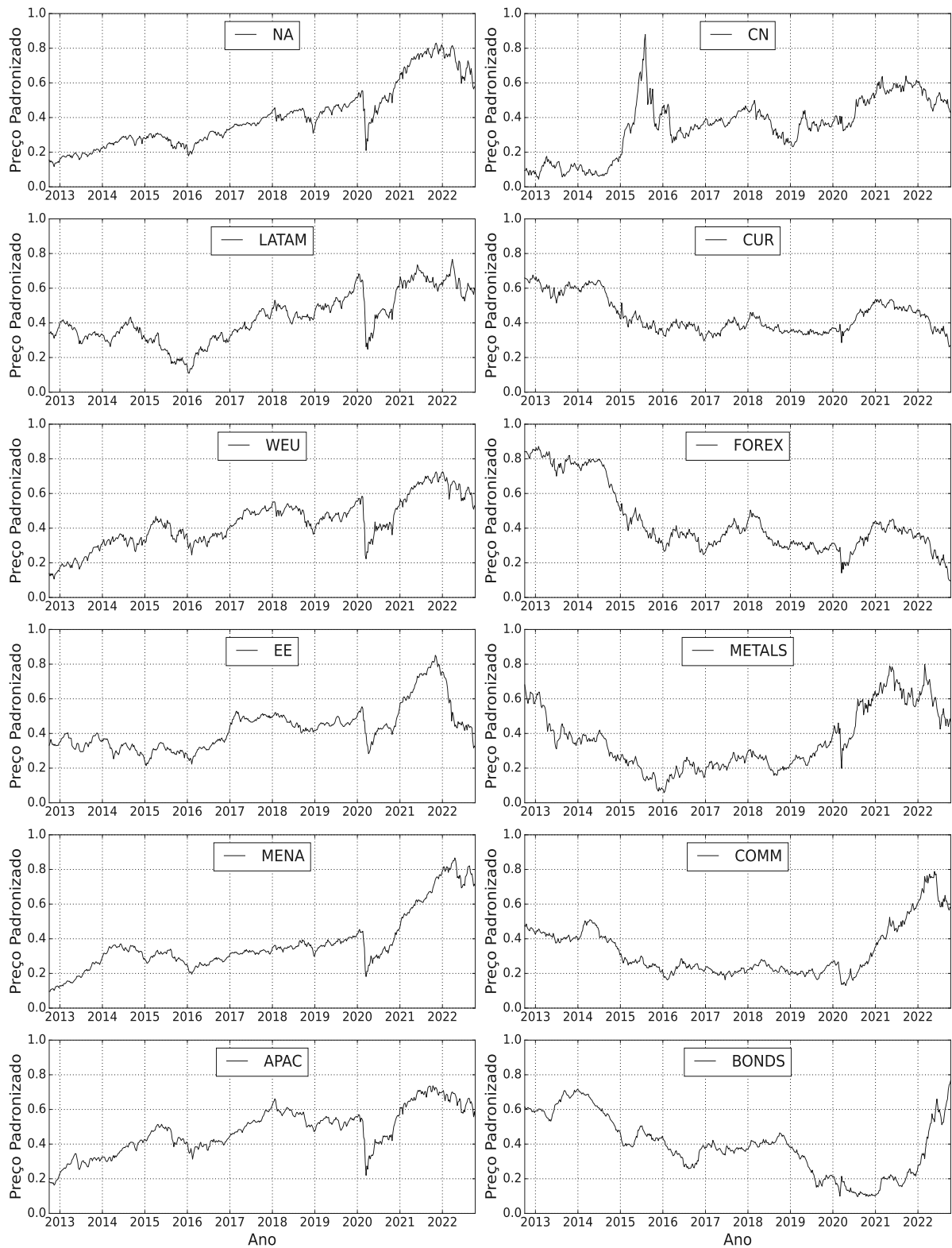
Nesta seção, vamos entender o comportamento dos dados obtidos através dos seus preço padronizados e, com isso, obter os gráficos das séries temporais médias para cada grupo financeiro. Após isso, vamos calcular as matrizes de correlação entre os grupos financeiros para todo o período. Por fim, analisamos a estrutura hierárquica desses grupos financeiros por meio de um dendrograma.

Os dados foram padronizados para que pudéssemos compará-los pois, pertencendo a grupos financeiros distintos, não possuem a mesma escala. A padronização segue a equação $P_i \rightarrow (P_i - P_{\text{MIN}})/(P_{\text{MAX}} - P_{\text{MIN}})$, onde P_{MAX} representa o preço de fechamento máximo e P_{MIN} o preço de fechamento mínimo encontrados em cada série temporal presente no *dataset* do período analisado. Assim, temos as séries temporais com a escala de preços entre $[0, 1]$. Além de evitar problemas de escala nos dados, a padronização também evita problemas de dispersão nos dados. Assim, após o processo de padronização, os desvios médio e padrão de cada dado assumem os valores 0 e 1, respectivamente.

Para iniciar a análise exploratória de dados, partiremos do estudo do comportamento médio de cada grupo financeiro citado na seção 4.1. Por mais que existam *outliers*, por estarem na mesma escala e por estarmos usando o comportamento médio, teremos uma boa aproximação da série histórica de cada grupo financeiro. Com isso, podemos confirmar alguns dos eventos que já foram apresentados na seção 4.2.

Podemos observar, pela Figura 14 que, com exceção do mercado chinês, a maioria dos mercados financeiros – especificamente, *NA*, *LATAM*, *WEU*, *EE*, *MENA*, *APAC*, – possuem um padrão similar, especialmente no período 2020 – 2022. Tal comportamento corresponde a uma queda brusca em 2020 – correspondente a crise sanitária COVID-19 – com uma subsequente recuperação até o início de 2022. Essa semelhança vai ser vista mais formalmente quando observarmos as variações no log-retorno (Figura 15) e obtivermos a correlação entre esses grupos financeiros (Figura 16). Como podemos ver, o mercado chinês exibe um padrão atípico em relação aos demais mercados. Isso também poderá ser observado nas matrizes de correlação.

Figura 14 – Séries Temporais



Fonte: O autor. Séries temporais para os grupos financeiros citados em 4.1.

Também é possível observar a proximidade nos ciclos apresentados pelas séries temporais de *CUR* e *FOREX*: temos um mercado consolidado no período 2012 – 2014, seguido de uma tendência de baixa até 2016, uma nova lateralização entre 2016 – 2020, uma queda em 2020, uma tendência de alta entre 2020 – 2021 e uma de baixa entre 2021 – 2022. Outros dois grupos com comportamento semelhante são *METALS* e *COMM*, nos quais podemos ver uma tendência comportamentos marcados por ciclos característicos desses dois ativos. Novamente, essas semelhanças de comportamento serão justificadas quando obtivermos as matrizes de correlação.

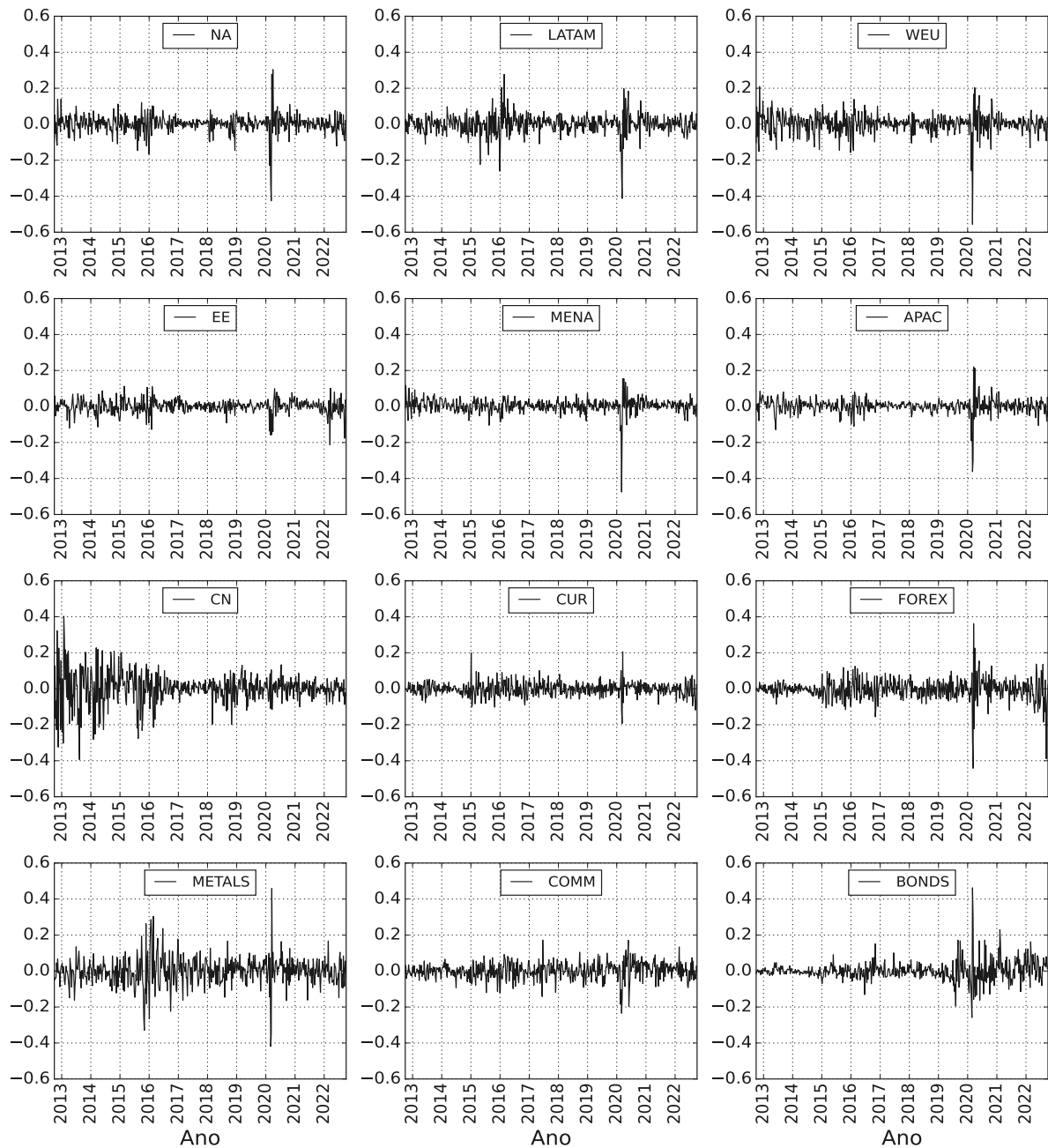
Agora, vamos analisar o log-retorno de cada grupo financeiro no período de 10 anos (Figura 15). O log-retorno é uma medida utilizada para calcular a variação percentual de um ativo financeiro em um período de tempo específico. Vale lembrar que ele é calculado utilizando a fórmula $\ln[P_i(t)/P_i(t-1)]$, onde $P_i(t)$ é o preço atual do ativo e $P_i(t-1)$ é o preço anterior. A transformação logarítmica remove a tendência linear dos preços, tornando a série de preços estacionária.

O log-retorno é amplamente utilizado na análise financeira, pois ele fornece uma medida de variação percentual consistente, independentemente do nível de preço do ativo. Além disso, seu uso permite que as distorções causadas pelas diferentes escalas de preços sejam ignoradas e, assim, obtemos uma visão mais precisa da volatilidade do ativo. O uso de log-retornos também é importante na modelagem financeira, como no caso dos modelos de risco de mercado, onde o log-retorno é utilizado para modelar a distribuição de rendimentos dos ativos financeiros (ONNELA *et al.*, 2002). Em resumo, o log-retorno é uma medida utilizada para medir a variação percentual de um ativo financeiro, e é amplamente utilizado na análise financeira e na modelagem financeira devido a sua consistência e precisão.

Podemos observar, na Figura 15, o log-retorno dos preços médios de fechamento semanal dos ativos de cada grupo financeiro. É possível notar que alguns grupos possuem oscilações de menor magnitude no seu log-retorno durante a maior parte do período. Isso pode ser visto como um sinal de estabilidade no mercado, falta de liquidez ou de um mercado menos ativo. Esse comportamento pode ser observado nos grupos *EE*, *MENA*, *CUR*, os dois primeiros provavelmente devido a baixa liquidez dos mercados e o último provavelmente devido a estabilidade desses ativos.

Por outro lado, mudanças abruptas no log-retorno de um índice de mercado pode indicar que o mercado está passando por uma volatilidade anormal, como uma crise financeira

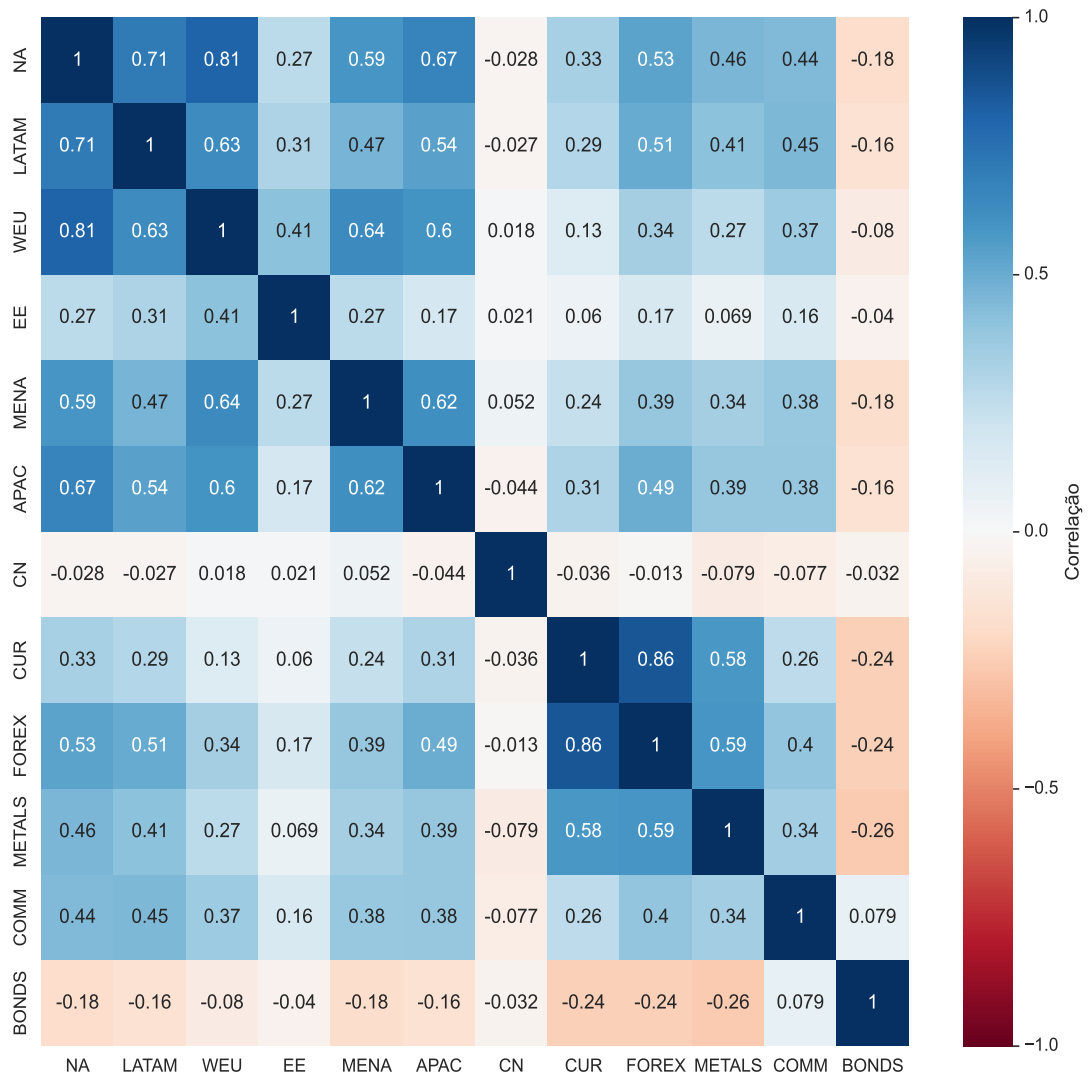
Figura 15 – Log-retorno semanal



Fonte: O autor. Log-retorno semanal dos preços padronizados médios de cada grupo financeiro num período de 10 anos.

ou uma desaceleração econômica. Isso pode ser causada por diversas razões, como eventos políticos, econômicos, empresariais, entre outros. Isso é observado em alguns mercados, como *NA*, *LATAM*, *CN*, *METALS*. Em especial, em 2020, devido a crise sanitária do COVID-19, vemos oscilações bruscas no log-retorno de diversos grupos financeiros, chegando a 40%. Portanto, essa foi a crise de maior impacto no período observado.

Figura 16 – Matriz de correlações



Fonte: O autor. Matriz de correlações entre os grupos financeiros citados em 4.1.

Neste momento, vamos analisar o mapa de calor dos doze grupos financeiros, na Figura 16. As cores frias representam uma correlação maior, enquanto as cores quentes representam uma correlação negativa. Vemos, no canto superior esquerdo, um bloco contendo os grupos financeiros *NA*, *LATAM*, *WEU*, *EE*, *MENA*, *APAC*. Vemos ainda que *EE* é menos correlacionado com os demais grupos desse bloco. Além disso, o grupo *CN* é não correlacionado com os demais grupos. No canto inferior direito, temos mais um bloco, incluindo *CUR*, *FOREX*, *METALS*, *COMM*. Embora *FOREX* e *CUR* tenham correlação positiva com todos os grupos, exceto *BONDS*, eles tem correlação mais alta entre si. Por outro lado, *METALS* e *COMM* tem

correlação positiva mas não tão significativa quanto esperado. Por fim, é possível observar que *BONDS* possui um comportamento de anti-correlação aos demais mercados financeiros, como esperado. Estes comportamentos estão relativamente de acordo com o que observamos nas séries temporais (figura 14).

Para entender melhor a formação do mapa de calor da Figura 16 e a relação de cada par da matriz de correlação, podemos analisar o valor absoluto desses coeficientes que, por sua vez, assumem valores entre $[0, 1]$. Com isso, podemos percorrer cada ligação em forma decrescente a partir de um *Threshold* que se inicia em 1.0 e assume valores menores, a um passo de 0.01, até atingir o zero. Então, podemos analisar quantas ligações estão se formando para um determinado *Threshold* e acompanhar os dois maiores agrupamentos (*clusters*) que se formam ligando-se a pares já formados anteriormente.

Na figura 17, há um comportamento de transição, quando o segundo maior componente para de crescer e os novos pares se ligam ao *cluster* dominante. Em suma, para entender o quanto cada grupo financeiro está relacionado com os demais, temos que analisar para qual valor do *threshold*, levando em consideração os coeficientes da matriz de correlação, há uma formação dominante e, principalmente, quando ela ocorrerá. Na Figura 17, vemos que para o valor 0.54, temos uma transição.

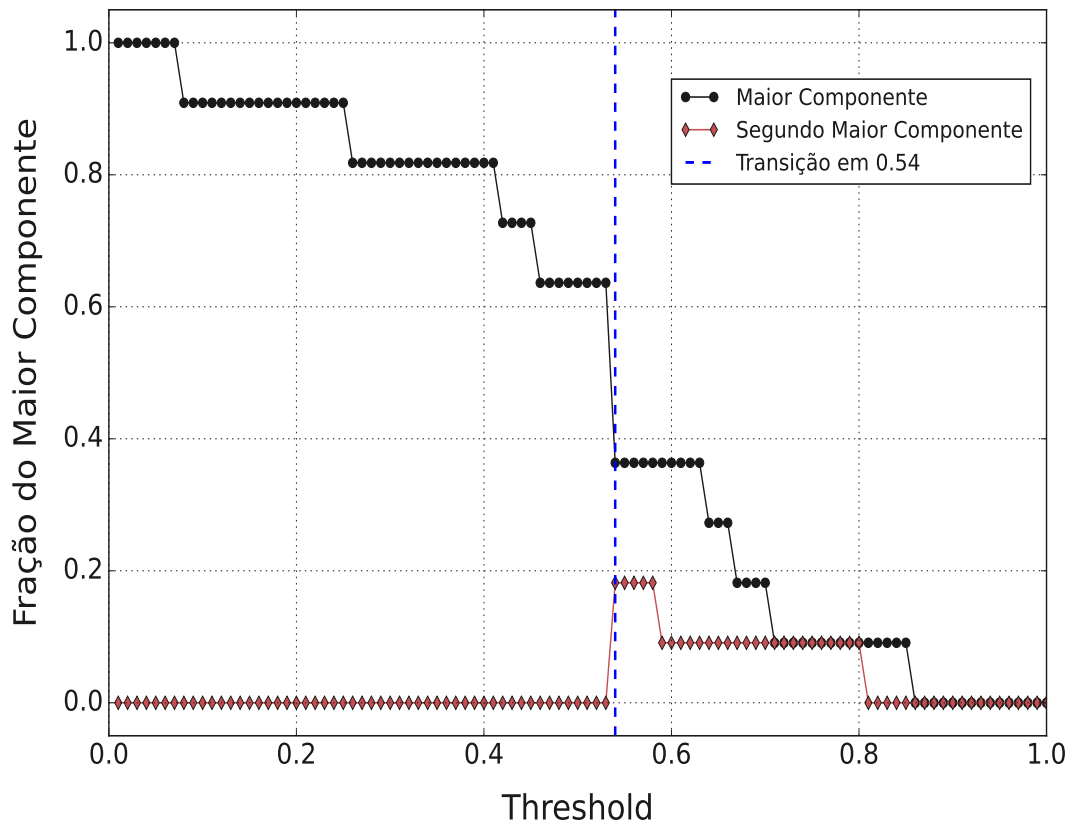
Esse resultado é significativo, mas pode ser complementado quando analisamos o agrupamento hierárquico ou dendrograma da matriz de correlação (Figura 18), o que nos permite ordenar os dados em *clusters*. Para esse dendrograma, utilizamos o método *complete linkage* da biblioteca *scipy* do *Python* (SCIPY...), que consiste em ordenar pela maior distância os pares. Para isso, calculamos uma distância da por:

$$d(X, Y) = 1 - |\rho_{X,Y}|, \quad (4.1)$$

onde $\rho_{X,Y}$ é o coeficiente de correlação de pares X, Y da Figura 16 e $d(X, Y)$ assume valores contidos no intervalo $[0, 1]$, assim como o valor absoluto de $\rho_{i,j}$ para a construção da Figura 17. Nesse caso, a distância definida acima é uma medida de dissimilaridade. Quanto menores (maiores) os valores observados, mais próximas (distantes) os pares estão.

Existem outros métodos disponíveis como *single linkage*, *average linkage*, etc, cada um com seus méritos e limitações (EVERITT *et al.*, 2011; MARTINEZ *et al.*, 2017; LEÓN *et al.*, 2017). No método *single linkage*, aglomerados distintos podem acabar se mesclando no

Figura 17 – Análise da conectividade dos grupos financeiros

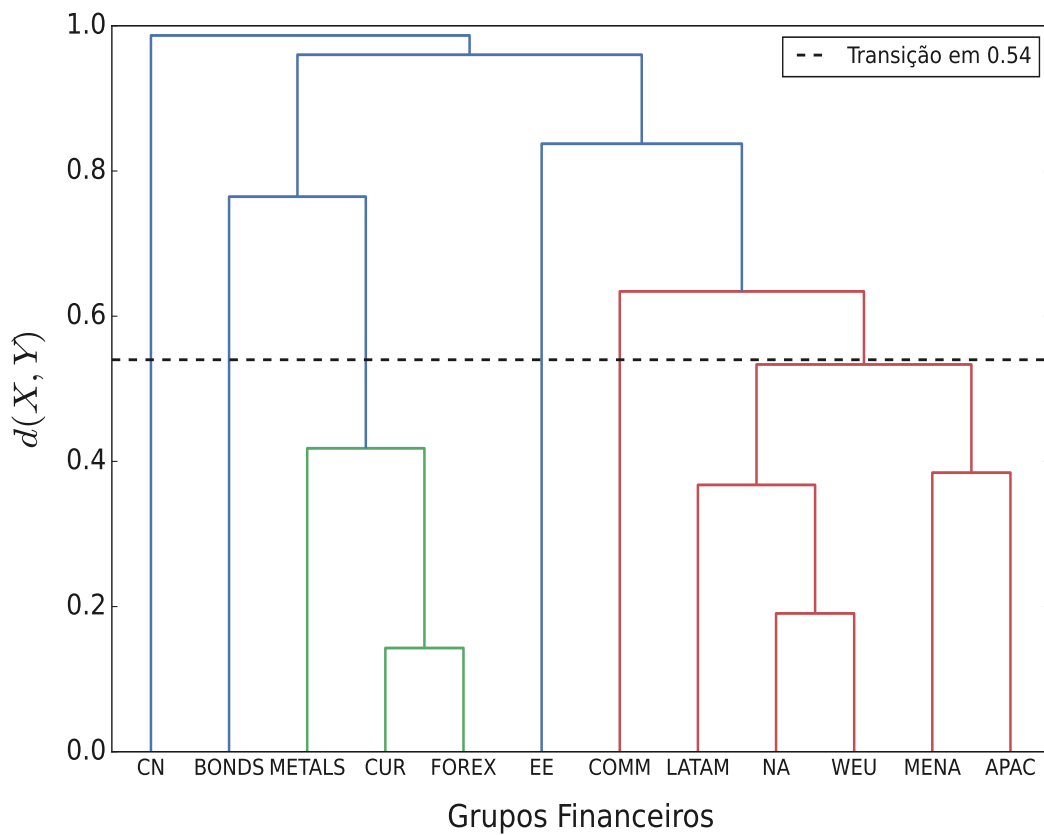


Fonte: O autor. Análise dos agrupamentos a partir do valor absoluto dos coeficientes de correlação da Figura 16 para se identificar a região de transição do maior componente.

ordenamento da menor distância entre os pares, pois, não é possível distinguir se um determinado vértice presente naquele entorno de distância mínima é ligado ou não com os outros pares em comum do valor de distância encontrado. Assim, a análise de *clusters* não é significativa. Já o método de *complete linkage* não sofre desse problema de encadeamento. Entretanto, esse método é sensível a *outliers*, gerando uma formação incompleta de determinados *clusters*. Isso não foi um problema para os dados considerados nesta seção, pois não temos *outliers*.

Na Figura 18, podemos notar que para valores abaixo do valor 0.54, temos um *cluster* formado pelos grupos *LATAM*, *NA*, *WEU*, *MENA*, *APAC*, um segundo *cluster* formado por *METALS*, *CUR*, *FOREX* e os grupos *CN*, *BONDS*, *EE* e *COMM* isolados. Temos alguns pontos interessantes a notar em relação a isso. Primeiramente, temos o grupo *EE* separados dos demais grupos financeiros, o que não era perceptível pela matriz de correlação. Além disso, o grupo *COMM* está mais hierarquicamente relacionado com *LATAM*, *NA*, *WEU*, *MENA*, *APAC* do que *METALS*, *CUR*, *FOREX*, ao contrário do que a matriz de correlação nos leva a pensar.

Figura 18 – Dendrograma da Matriz de Correlação



Fonte: O autor. Agrupamento Hierárquico (Dendrograma) para a matriz de correlação dos grupos financeiros pelo método de *complete linkage*.

Apesar de parecer contraditório, esse resultado é consistente uma vez que a dissimilaridade entre *COMM* e os grupos *LATAM*, *NA*, *WEU*, *MENA*, *APAC* é menor que a dissimilaridade entre *COMM* e *CUR*, *METALS*. Outro ponto interessante é que o grupo *CN* é o que está mais hierarquicamente distante dos demais confirmando sua não-correlação com os demais grupos e também não possui uma correlação com *BONDS*.

Até o momento, analisamos o comportamento médio do grupo para entendermos melhor o conjunto de dados financeiros com os quais estamos trabalhando. No entanto, necessitamos de uma análise complementar do comportamento individual de cada ativo e qual a influência no comportamento médio. Além disso, precisamos de uma análise mais refinada para a evolução temporal, pois janelas temporais menores do que a de 10 anos nos darão uma informação mais rica dos ciclos de cada grupo financeiro.

4.4 Análise dos agrupamentos a partir da correlação

Nessa seção, analisamos todos os ativos dos grupos financeiros da seção 4.1. Vamos começar com o mapa de calor representando a matriz de correlação (602×602) entre ativos, que é dado na Figura 19. A motivação para a construção da MST vem do fato de que esta representa muito bem e de maneira reduzida a matriz de correlação. É uma estrutura gerada a partir de um grafo ponderado, contendo como informação dos vértices cada ativo financeiro e suas arestas sendo os pesos obtidos a partir das distâncias de correlação de cada par de vértice. Outra vantagem está no fato de que o algoritmo de Prim minimiza os pesos da rede, gerando uma solução única para cada configuração de MST e guardando as informações da conexão que gerou a MST.

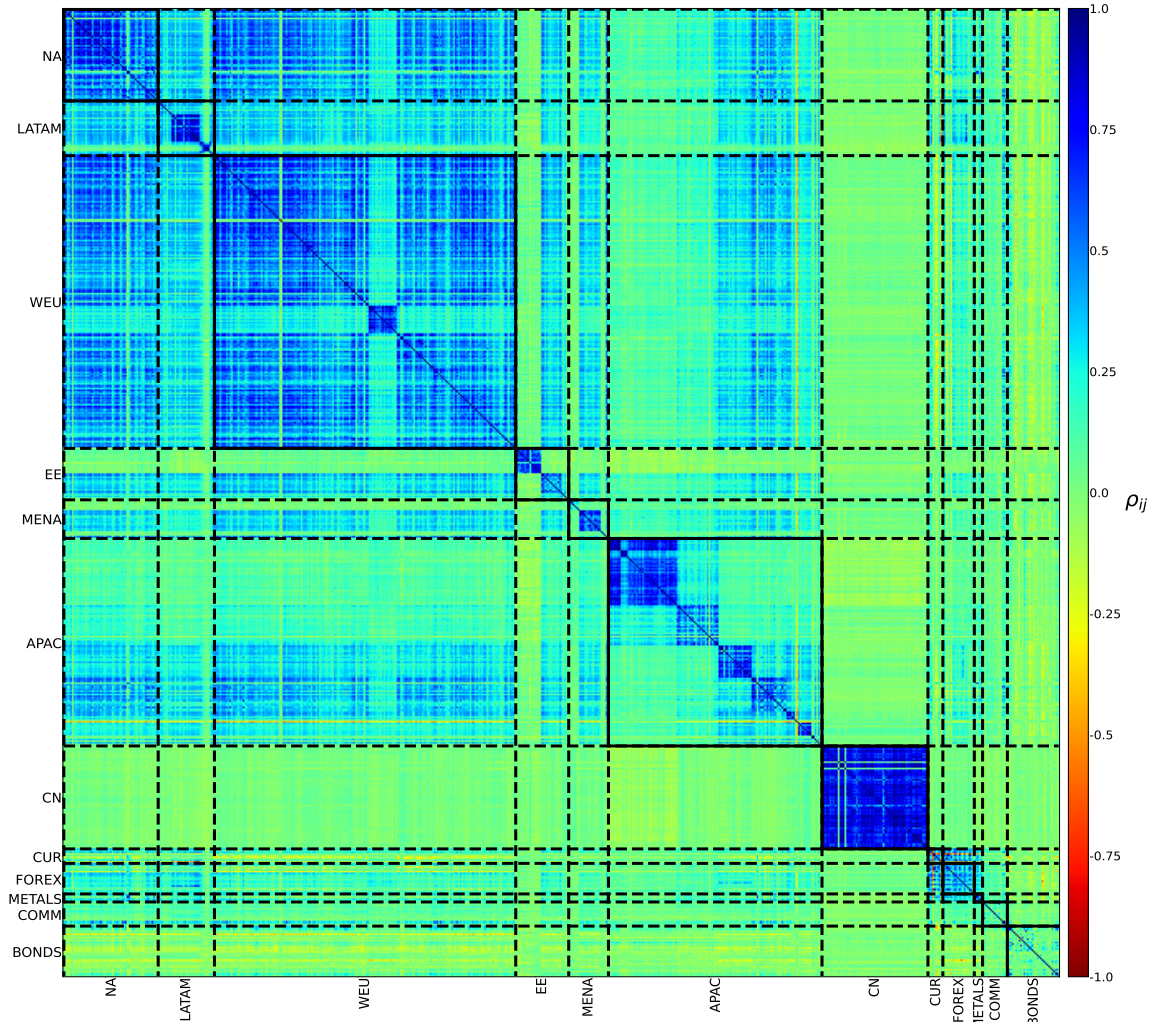
Na figura, temos diversos blocos na diagonal que representam as correlações entre ativos de um mesmo grupo e, dentro de alguns desses blocos, temos sub-blocos:

- No bloco *LATAM*, temos um sub-bloco representando ativos do mercado brasileiro e um segundo sub-bloco menor contendo ativos dos mercados peruano, chileno e colombiano.
- No bloco *WEU*, temos um sub-bloco isolado com ativos da Turquia.
- O bloco *EE* é basicamente dividido em dois sub-blocos: um contendo ativos do mercado russo e outro contendo ativos dos demais mercados do grupo financeiro.
- No bloco *MENA* o sub-bloco que se destaca contém ativos israelenses.
- Por fim, no bloco *APAC* temos dois sub-blocos, um contendo ativos japoneses e sul-coreanos e outro contendo os demais ativos.

Agora, podemos usar as informações da matriz de correlação para calcular as distâncias de correlação e, com isso, obter a MST. O resultado é mostrado na Figura 20.

Podemos observar que *CN* tem quase todos os seus ativos em um único *cluster*. Os ativos dos grupos *NA*, *WEU*, *MENA FOREX* e *BONDS* também estão, em sua maioria, concentrados em *clusters*, com exceção de alguns ativos. Os ativos do grupo *APAC* se concentram em três *clusters*, um contendo ativos do Japão e Coreia do Sul (lado inferior esquerdo), outro contendo ativos de Austrália e Nova Zelândia (lado inferior direito) e um terceiro *cluster* contendo, majoritariamente, ativos indianos. No grupo *LATAM*, temos três ramos, um com ativos do Peru, outro com ativos do México e o terceiro com os ativos restantes. O grupo *EE* tem dois *clusters*, um com ativos da Rússia (lado superior esquerdo) e outro com ativos da Polônia (lado inferior direito). Além disso, temos os ativos dos países bálticos agrupados entre si e o ativo da Sérvia isolado.

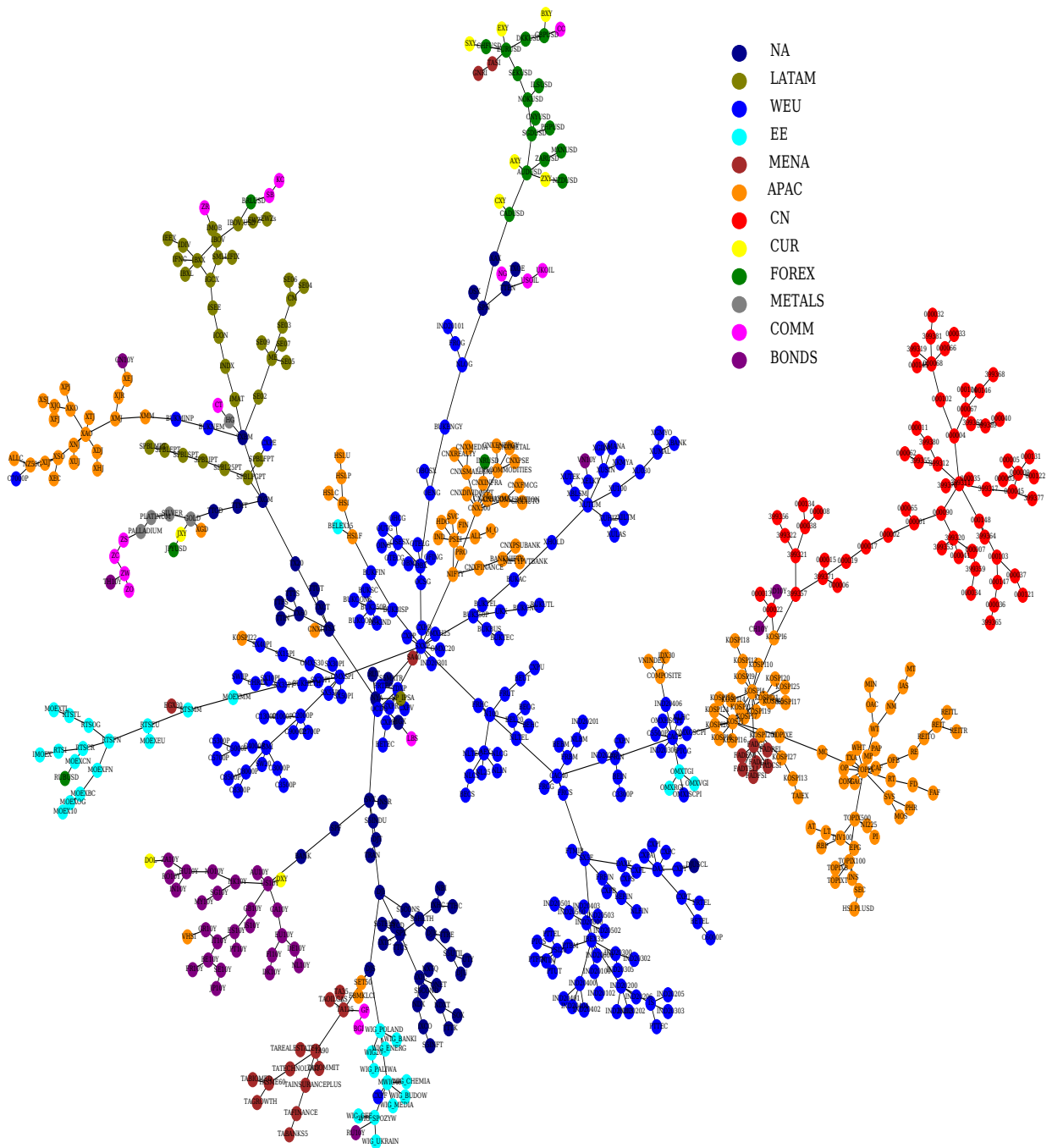
Figura 19 – Matriz de Correlação de todos os ativos



Fonte: O autor. Matriz de correlação dos ativos descritos na seção 4.1.

Podemos rearranjar os ativos da matriz de correlação 19 através da MST construída pela técnica de Prim. Com isso, obtemos uma nova representação dos dados que reflete as relações de proximidade entre os vértices. A matriz de correlação resultante é dada na Figura 21. A MST é gerada a partir de uma matriz de distância de correlação, sendo composta por vértices e arestas. A ordem dos vértices na MST reflete a proximidade dos dados, onde os vértices conectados por arestas de menor peso são mais próximos entre si. Reorganizando os dados da matriz de correlação a partir da MST, podemos obter uma nova matriz de correlação onde os

Figura 20 – MST para todos os ativos no período de 10 anos

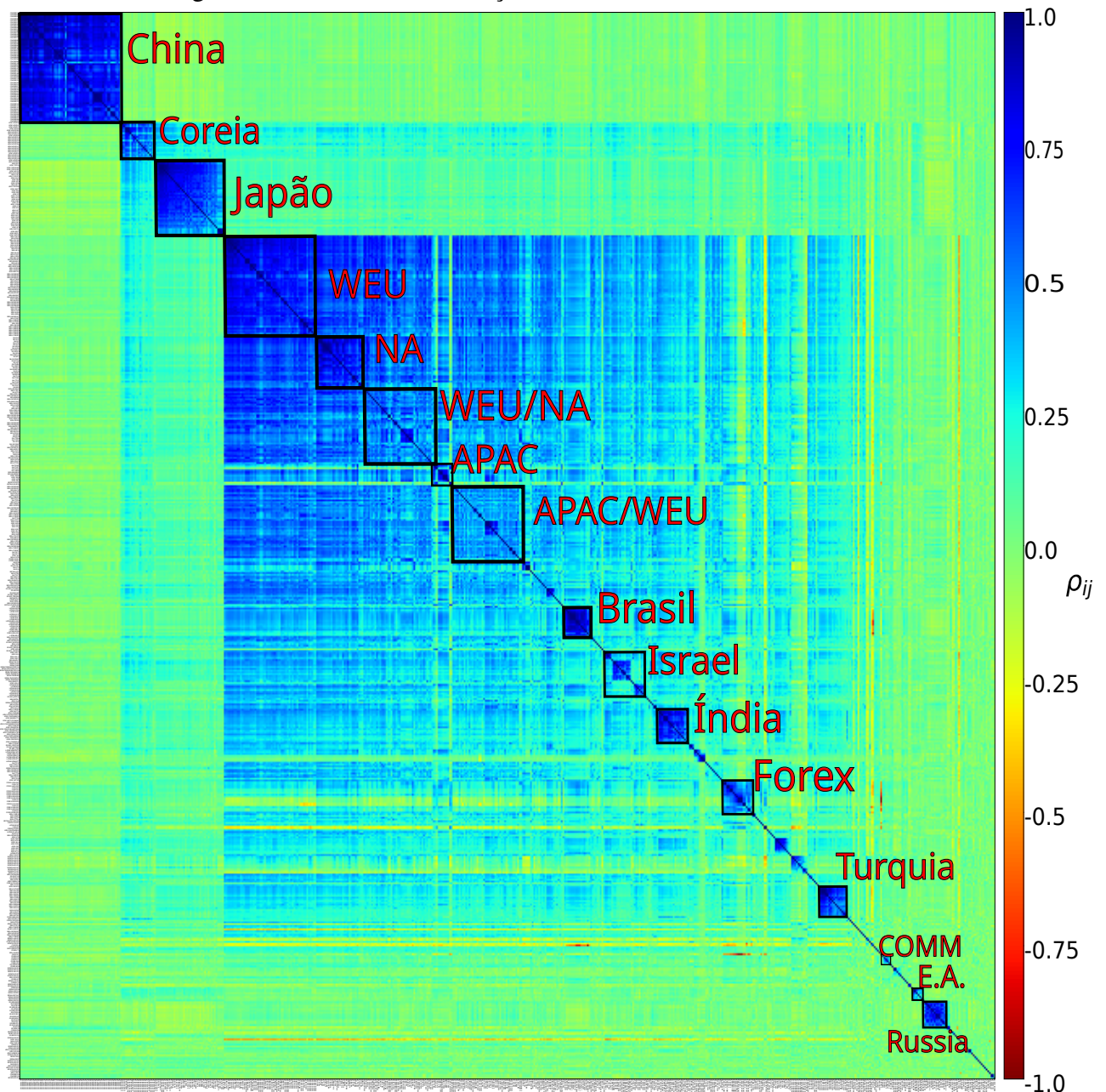


Fonte: O autor. *Minimal Spanning Tree* dos ativos descritos na seção 4.1.

dados estão organizados de acordo com sua proximidade. Isso pode ser útil para identificar padrões e tendências nos dados que podem não ser evidentes na matriz de correlação original.

Agora, podemos observar alguns *clusters* da MST (Figura 21). O primeiro bloco é composto por ativos do grupo *CN*. Logo abaixo e à direita, vemos um bloco de ativos da Coreia

Figura 21 – Matriz de Correlação de todos os ativos reordenada

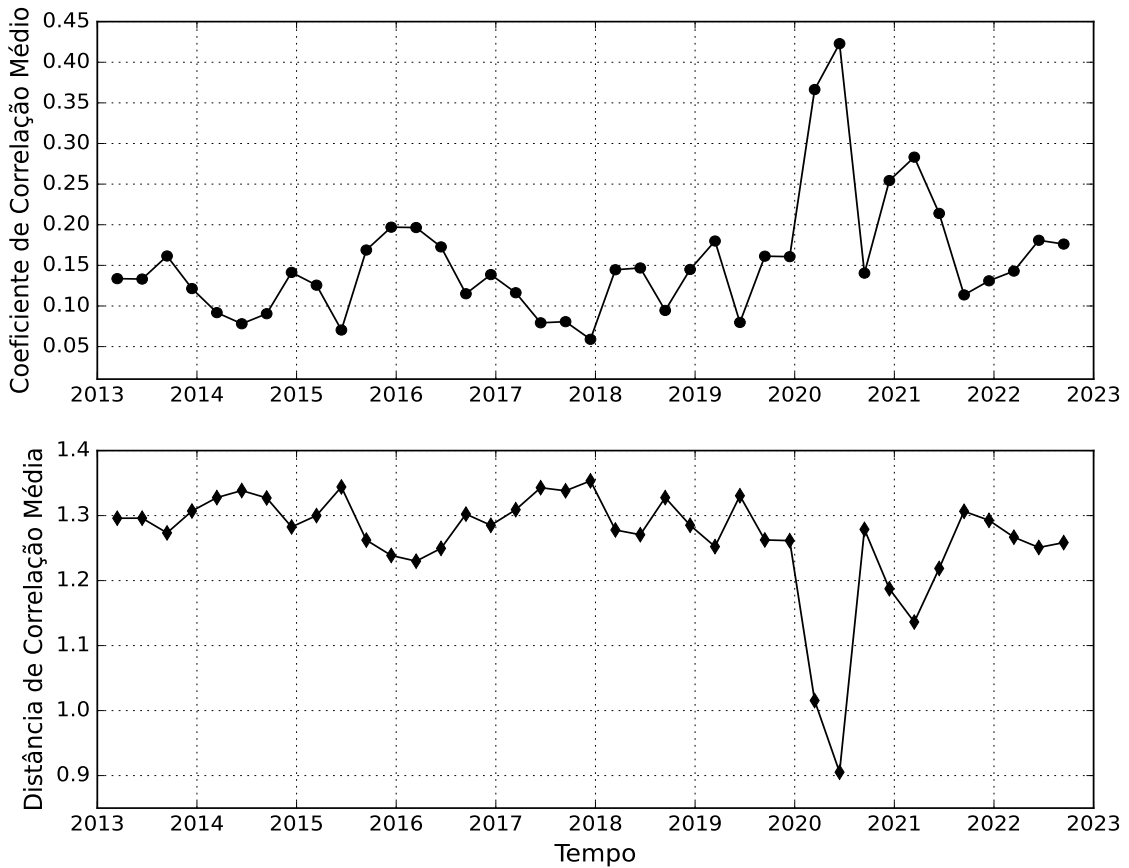


Fonte: O autor. Matriz de correlação dos ativos descritos na seção 4.1 ordenados pela ordem dos pesos dados pela MST.

de Sul, seguido de um bloco com ativos do Japão. É possível observar alguns aglomerados pertencentes aos blocos *NA* e *WEU* por serem mercados com diversos ativos financeiros e com muitos países distintos (no caso da Europa). Tanto na Figura 20 quanto nessa do mapa de calor, nota-se a separação da Rússia e da Polônia, ambos pertencentes ao grupo *EE*. Além disso, observa-se em ambas as figuras os blocos pertencentes a Israel, Brasil e Índia, assim como a separação de vários grupos oriundos do *APAC*. Diferentemente da Figura 20, onde temos um

separação clara, observamos vários ruídos no mapa de calor. Ou seja, se tomarmos a Figura 21 como base de classificação, torna-se necessária uma técnica mais robusta para classificação dos agrupamentos.

Figura 22 – Análise dos Valores Médios da Correlação e da Distância de Correlação



Fonte: O autor. Análise dos Valores Médios da Correlação e da Distância de Correlação durante o período de 10 anos com uma janela móvel de 26 semanas (um semestre) com *steps* de 13 semanas (um trimestre).

Na figura seguinte, consideramos como o coeficiente de correlação médio e a distância de correlação média evoluem no tempo. O coeficiente de correlação médio é calculado levando em conta a expressão $\frac{1}{N(N-1)} \sum_{i,j} \rho_{ij}$. Dividimos pelo fator $N(N-1)$, que corresponde ao número de termos fora da diagonal principal da matriz de correlação, para garantir que o coeficiente de correlação médio seja normalizado. A distância de correlação média foi calculada com base na expressão, $\frac{1}{N-1} \sum_{i,j} d_{ij}^t$, onde $N-1$ é o número de arestas na MST e d_{ij}^t é a distância de correlação dada por (3.4). Cada ponto nesses gráficos representa a média semestral, calculada de trimestre a trimestre. Ou seja, tem-se uma janela temporal de 26 semanas, onde cada janela

móvel é atualizada em 13 semanas (trimestralmente). O primeiro ponto no gráfico do coeficiente de correlação médio, por exemplo, representa o valor médio do coeficiente de correlação para o período setembro/2012 - março/2013.

Podemos notar que as séries temporais das duas quantidades são, aproximadamente, a imagem refletida da outra, o que confirma a característica da distância de correlação, dada por $d_{i,j} = \sqrt{2(1 - \rho_{i,j})}$, representar uma translação e reproduzir as informações da matriz de correlação. Esse fato é corroborado pelo comportamento dos gráficos da Figura 14, onde tem-se um redução da distância de correlação quando há um aumento na correlação dos ativos.

Em 2020, que marca o início da crise sanitária Covid-19, vemos um aumento abrupto do coeficiente de correlação médio no início de 2020, o que ocorre devido a atuação das forças de mercado atuando sobre os ativos afim de obter comportamento unificado do mercado. Vemos que, das crises citada na Seção 4.2, somente a de 2020 teve um efeito significativo no coeficiente e distância de correlação médios. Isso é explicado pelo fato de que as demais crises foram locais, enquanto a crise de 2020 afetou diversos mercados.

5 CONCLUSÃO

Neste trabalho, buscou-se entender como se dava o agrupamento de ativos do mercado financeiro para um período de 10 anos compreendido entre 2012 e 2022. A construção desse trabalho destinou-se a analisar o comportamento de um vasto número de ativos financeiros e como se daria a sua classificação, nesse período, a partir da matriz de correlação das séries temporais do mercado financeiro.

A partir da correlação de pares, tornou-se possível entender a formação de agregados financeiros. Estes, como vimos, não necessariamente se organizam por região ou tipo de ativos, mas por afinidade econômica que, por sua vez, é construída durante um ciclo financeiro. Como foi observado na seção 4.4 há a possibilidade de se construir uma análise de formação de agregados utilizando metodologias de hierarquia de redes. Porém, a precisão na formação desses *clusters* atinge o maior valor quando utilizamos a técnica de MST pelo algoritmo de Prim, levando em conta o fato de que essa técnica minimiza os pesos das ligações dos vértices e a necessidade de haver conexão entre os vértices que representam cada ativo financeiro presente na matriz de correlação.

Devido as características das séries temporais de mercado financeiro de serem estocásticas e ao comportamento inerente ao mercado financeiro de não possuir anti-correlação ou correlação perfeita, há uma grande dificuldade no estudo de muitos ativos financeiros individuais. Além disso, a utilização de um curto período de tempo impõe mais desafios, uma vez que a oscilação gerada nos ativos tendem a construir muitos ruídos momentâneos - que são amenizados somente em períodos de euforia ou super-ciclos de alta. Consequentemente, para se haver um estudo mais direcional na classificação dos agregados, torna-se necessário a implementação de técnicas de classificação de agregados mais eficientes. Portanto, esse trabalho assume como perspectiva a utilização de técnicas de classificação ou de predições provenientes da área de *machine learning*.

REFERÊNCIAS

- AN, S.; GAO, X.; JIANG, M.; SUN, X. Multivariate financial time series in the light of complex network analysis. **Physica A: Statistical Mechanics and its Applications**, v. 503, p. 1241–1255, 8 2018.
- ANDERSEN, T. G.; BOLLERSLEV, T.; DIEBOLD, F. X.; VEGA, C. Real-time price discovery in global stock, bond and foreign exchange markets. **Journal Of International Economics**, v. 73, n. 2, p. 251–277, 11 2007.
- AROUSSE, R. **Finance market data downloader**. Disponível em: <https://github.com/ranaroussi/yfinance>. Acesso em: 05 abr. 2021.
- AVLIS, E. **Master's Thesis-Financial Market Correlation Analyses**. Disponível em: https://github.com/EwertonAvlis/Master-s_Thesis-Financial_Market_Correlation_Analyses. Acesso em: 20 jan. 2023.
- BALDUZZI, P.; ELTON, E. J.; GREEN, T. C. Economic news and bond prices: evidence from the u.s. treasury market. **The Journal Of Financial And Quantitative Analysis**, v. 36, n. 4, p. 523, 12 2001.
- BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 10 1999.
- BARABÁSI, A. L. **Linked: The new science of networks**. Cambridge: Perseus, 2002. 288 p.
- BARABÁSI, A.-L. **Network Science**. 1. ed. Cambridge: Cambridge University Press, 2016. 475 p.
- BARIGOZZI, M.; HALLIN, M. A network analysis of the volatility of high dimensional financial series. **Journal Of The Royal Statistical Society Series C: Applied Statistics**, v. 66, n. 3, p. 581–605, 9 2016.
- BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D. U. Complex networks: structure and dynamics. **Physics Reports**, v. 424, n. 4-5, p. 175–308, 2 2006.
- BONANNO, G.; CALDARELLI, G.; LILLO, F.; MICCICHÉ, S.; VANDEWALLE, N.; MANTEGNA, R. N. Networks of equities in financial markets. **The European Physical Journal B - Condensed Matter**, v. 38, n. 2, p. 363–371, 3 2004.
- BONANNO, G.; LILLO, F.; MANTEGNA, R. N. High-frequency cross-correlation in a set of stocks. **Quantitative Finance**, v. 1, n. 1, p. 96–104, 1 2001.
- BONANNO, G.; LILLO, F.; MANTEGNA, R. N. Topology of correlation-based minimal spanning trees in real and model markets. **Physical Review E**, v. 68, n. 4, p. 046130–046130, 10 2003.
- BONDY, J. A.; MURTY, U. S. R. **Graph Theory**. S.l.: Springer Publishing Company, Incorporated, 2008.
- BONGIORNO, C.; MICCICHE, S.; MANTEGNA, R.; GURTNER, G.; LILLO, F.; POZZI, S. Adaptive air traffic network: Statistical regularities in air traffic management. In: EUROCONTROL. **USA/EUROPE AIR TRAFFIC MANAGEMENT RESEARCH AND DEVELOPMENT SEMINAR 2015, 11th**. [S. l.], 2015.

BRIDA, J. G.; GÓMEZ, D. M.; RISSO, W. A. Symbolic hierarchical analysis in currency markets: an application to contagion in currency crises. **Expert Systems With Applications**, v. 36, n. 4, p. 7721–7728, 5 2009.

CESAR, C. L. **Aula 1 - Matemática Financeira**. Disponível em: <https://www.ifi.unicamp.br/~lenz/Econofisica/>. Acesso em: 29 mar. 2021.

CESAR, C. L. **Diversificacao de Portfolios e Otimizacao**. Disponível em: <https://www.ifi.unicamp.br/~lenz/Econofisica/>. Acesso em: 29 mar. 2021.

CESAR, C. L. **Teoria da Probabilidade II**. Disponível em: <https://www.ifi.unicamp.br/~lenz/Econofisica/>. Acesso em: 25 mar. 2021.

CESAR, C. L. **Teoria dos Conjuntos + Medida + Probabilidade**. Disponível em: <https://www.ifi.unicamp.br/~lenz/Econofisica/>. Acesso em: 25/03/2021.

CHARTRAND, G.; ZHANG, P. **A First Course in Graph Theory**. New York: Dover Publications, 2013.

CHI, K. T.; LIU, J.; LAU, F. C. A network perspective of the stock market. **Journal Of Empirical Finance**, v. 17, n. 4, p. 659–667, 9 2010.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. **Introduction to Algorithms**. 3. ed. [S.L.]: The MIT Press, 2009.

CORONNELLO, C.; TUMMINELLO, M.; LILLO, F.; MICCICHÈ, S.; MANTEGNA, R. Sector identification in a set of stock return time series traded at the london stock exchange. **Acta Physica Polonica B**, v. 36, p. 2653–2679, 2005.

CORRELATION and causation. Disponível em: <https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/correlation-and-causation>. Acesso em: 12 jan. 2023.

DIESTEL, R. **Graph Theory**. 3. ed. New York: Springer-Verlag Heidelberg, 2005.

EDERINGTON, L. H.; LEE, J. H. How markets process information: news releases and volatility. **The Journal Of Finance**, v. 48, n. 4, p. 1161–1191, 9 1993.

EULER, L. **Mathematics and The Modern World: The Koenigsberg Bridges**. San Francisco: W.H. Freeman and Company, 1968.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. 5. ed. Chichester: John Wiley Sons, 2011. 352 p. (Wiley Series in Probability and Statistics).

GABAIX, X.; GOPIKRISHNAN, P.; PLEROU, V.; STANLEY, H. E. A theory of power-law distributions in financial market fluctuations. **Nature**, v. 423, n. 6937, p. 267–270, 5 2003.

GAO Z.-K.; SMALL, M. K. J. Complex network analysis of time series. **Epl (Europhysics Letters)**, v. 116, n. 5, p. 50001, 12 2016.

GILMORE, C.; LUCEYB, B.; BOSCIA, M. An ever-closer union? examining the evolution of linkages of european equity markets via minimum spanning trees. **Physica A**, v. 387, n. 25, p. 6319–6329, 11 2008.

JUNG, W.-S.; CHAE, S.; YANG, J.-S.; MOON, H.-T. Characteristics of the Korean stock market correlations. **Physica A: Statistical Mechanics and its Applications**, v. 361, n. 1, p. 263–271, 2 2006.

KENTON, W. **Financial Crisis: Definition, Causes, and Examples**. Disponível em: <https://www.investopedia.com/terms/f/financial-crisis.asp>. Acesso em: 10 dez. 2022.

KUMAR, S.; DEO, N. Correlation and network analysis of global financial indices. **Physical Review E**, v. 86, n. 2, p. 026101, 8 2012.

LEÓN, C.; KIM, G. Y.; MARTÍNEZ, C.; LEE, D. Equity markets' clustering and the global financial crisis. **Quantitative Finance**, v. 17, n. 12, p. 1905–1922, 8 2017.

MANTEGNA, R. N. Hierarchical structure in financial markets. **The European Physical Journal B**, v. 11, n. 1, p. 193–197, 9 1999.

MANTEGNA, R. N.; STANLEY, H. E. **An Introduction to Econophysics**. Cambridge: Cambridge University Press, 2000.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate Analysis**. San Diego: Academic Press, 1979. 521 p.

MARTINEZ, W. L.; MARTINEZ, A. R.; SOLKA, J. **Exploratory Data Analysis with MATLAB**. 3. ed. New York: Chapman And Hall/Crc, 2017. 616 p.

MCDONALD, M.; SULEMAN, O.; WILLIAMS, S.; HOWISON, S.; JHONSON, N. F. Detecting a currency's dominance or dependence using foreign exchange network trees. **Physical Review E**, v. 72, n. 4, p. 046106, 10 2005.

MICCICHÈ, S.; BONANNO, G.; LILLO, F.; MANTEGNA, R. N. Degree stability of a minimum spanning tree of price return and volatility. **Physica A: Statistical Mechanics and its Applications**, v. 324, n. 1-2, p. 66–73, 6 2003.

NIELSEN, A. **Practical Time Series Analysis**. Sebastopol: O'Reilly Media, Inc., 2009.

NOBI, A.; LEE, S.; KIM, D. H.; LEEA, J. W. Correlation and network topologies in global and local stock indices. **Physics Letters A**, v. 378, n. 34, p. 2482–2489, 7 2014.

NOBI, A.; MAENG, S. E.; HA, G. G.; LEE, J. W. Effects of global financial crisis on network structure in a local stock market. **Physica A: Statistical Mechanics and its Applications**, v. 407, p. 135–143, 8 2014.

ONNELA, J. P.; CHAKRABORTI, A.; KASKI, K.; KERTIÉSZ, J. Dynamic asset trees and portfolio analysis. **The European Physical Journal B - Condensed Matter**, v. 30, n. 3, p. 285–288, 2002.

ONNELA, J. P.; CHAKRABORTI, A.; KASKI, K.; KERTÉSZ, J.; KANTO, A. Dynamics of market correlations: taxonomy and portfolio analysis. **Physical Review E**, v. 68, n. 5, p. 056110, 11 2003.

PAPOULIS, A. **Random variables and stochastic processes**. [S. l.]: McGraw Hill, 1991.

POZZI, F.; MATTEO, T. D.; ASTE, T. Centrality and peripherality in filtered graphs from dynamical financial correlations. **Advances In Complex Systems**, v. 11, n. 06, p. 927–950, 12 2008.

ROSS, S. **3 Financial Crises in the 21st Century**. Disponível em: <https://www.investopedia.com/articles/investing/011116/3-financial-crises-21st-century.asp>. Acesso em: 10 dez. 2022.

SCIPY.CLUSTER.HIERARCHY.LINKAGE. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html>. Acesso em: 12 dez. 2022.

SENSOY, A.; YUKSEL, S.; ERTURK, M. Analysis of cross-correlations between financial markets after the 2008 crisis. **Physica A: statistical mechanics and its applications**, v. 392, n. 20, p. 5027–5045, 10 2013.

SHIN, K.-H.; LIM, G.; MIN, S. Dynamics of the global stock market networks generated by decca methodology. **Applied Sciences**, v. 10, n. 6, p. 2171, 3 2020.

SIECZKA, P.; HOLYST, J. A. Correlations in commodity markets. **Physica A: statistical mechanics and its applications**, v. 388, n. 8, p. 1621–1630, 4 2009.

SINHA, S.; CHATTERJEE, A.; CHAKRABORTI, A.; CHAKRABARTI, B. K. **Econophysics: an introduction**. [S.L.]: Wiley-VCH Verlag, 2011. 369 p.

SPURIOUS Correlations. Disponível em: <https://www.tylervigen.com/spurious-correlations>. Acesso em: 12 jan. 2023.

STAUFFER, P.; OLIVEIRA, P. M. C.; BERNARDES, A. T. Monte carlo simulation of volatility clustering in market model with herding. **IJTAF**, v. 02, n. 01, p. 83–94, 1 1999.

STROGATZ, S. H. Exploring complex networks. **Nature**, v. 410, n. 6825, p. 268–276, 3 2001.

TABAK, B. M.; THIAGO, R. S.; CAJUEIRO, D. O. Topological properties of stock market networks: the case of brazil. **Physica A: statistical mechanics and its applications**, v. 389, n. 16, p. 3240–3249, 8 2010.

TANG, Y.; XIONG, J. J.; JIA, Z. Y.; ZHANG Y, C. Complexities in financial network topological dynamics: modeling of emerging and developed stock markets. **Complexity**, v. 2018, p. 1–31, 11 2018.

TRACK All Markets. Disponível em: <https://www.tradingview.com/>. Acesso em: 02 jul. 2022.

TUMMINELLO, M.; LILLO, F.; MANTEGNA, R. N. Correlation, hierarchies, and networks in financial markets. **Journal Of Economic Behavior & Organization**, v. 75, n. 1, p. 40–58, 7 2010.

TVDATAFEED - A simple TradingView historical Data Downloader. Disponível em: <https://github.com/StreamAlpha/tvdatafeed>. Acesso em: 02 jul. 2022.

WASSERMAN, L. **All of statistics: a concise course in statistical inference**. New York: Springer, 2004. v. 26.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 6 1998.

WILLIAMS, W. **Timeline of U.S. Stock Market Crashes**. Disponível em: <https://www.investopedia.com/timeline-of-stock-market-crashes-5217820>. Acesso em: 10 dez. 2022.

ZHANG, Y. **Stock Market Network Topology Analysis Based on a Minimum Spanning Tree Approach**. Dissertação, S.L., 2009.