UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

CURSO DE GRADUAÇÃO EM FÍSICA

ELIEZER BARBOSA LIMA NETO

ANÁLISE DA REDE "MINIMUM SPANNING TREE" DE CORRELAÇÃO ECONÔMICA ENTRE AS PRINCIPAIS AÇÕES MUNDIAIS NO RECENTE PERÍODO DE INSTABILIDADE GLOBAL

FORTALEZA-CE

2024

ELIEZER BARBOSA LIMA NETO

ANÁLISE DA REDE “MINIMUM SPANNING TREE” DE CORRELAÇÃO ECONÔMICA ENTRE AS PRINCIPAIS AÇÕES MUNDIAIS NO RECENTE PERÍODO DE INSTABILIDADE GLOBAL

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Física.

Orientador: Prof.Dr. Carlos Lenz Cesar.

FORTALEZA-CE

2024

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)

L697a Lima Neto, Eliezer Barbosa.
Análise da rede “minimum spanning tree” de correlação econômica entre as principais ações mundiais no recente período de instabilidade global / Eliezer Barbosa Lima Neto. – 2024.
73 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Física, Fortaleza, 2024.
Orientação: Prof. Dr. Carlos Lenz Cesar.

1. Minimum Spanning Tree. 2. Lockdown. 3. Guerra na Ucrânia. I. Título.

CDD 530

ELIEZER BARBOSA LIMA NETO

ANÁLISE DA REDE “MINIMUM SPANNING TREE” DE CORRELAÇÃO ECONÔMICA ENTRE AS PRINCIPAIS AÇÕES MUNDIAIS NO RECENTE PERÍODO DE INSTABILIDADE GLOBAL

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Física do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Física.

Aprovada em: 24/09/2024.

BANCA EXAMINADORA

Prof.Dr. Carlos Lenz Cesar (Orientador)
Universidade Federal do Ceará (UFC)

Prof.Dr. Wandemberg Paiva Ferreira
Universidade Federal do Ceará (UFC)

Prof.Dr. Francisco Nepomuceno Filho
Universidade Federal do Ceará (UFC)

A minha mãe, Claudene Lima.

AGRADECIMENTOS

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo financiamento da minha bolsa de iniciação científica, que possibilitou a conclusão do meu curso.

Ao Departamento de Física da Universidade Federal do Ceara (UFC) e todos os seus profissionais pela minha formação como físico e pelo ambiente propicio a isso.

Agradeço ao Prof. Dr. Carlos Lenz Cesar pela ótima orientação e tempo gasto com a minha formação profissional.

Aos professores participantes da banca examinadora Wandemberg Paiva Ferreira e Francisco Nepomuceno Filho pelo tempo e pelas valiosas colaborações e sugestões.

Aos professores João Miltom Pereira, Ricardo Renan Landim, Roberto Vinhaes Maluf que me ajudaram na formação do meu caráter profissional e dedicação estudantil .

Agradeço à minha noiva, Keliny Alves, pelo apoio emocional e incetivo profissional.

A minha vó, por parte da minha educação e apoio financeiro.

A minha irmã Aneliz Lima, por me ajudar com meus trabalhos e me dar apoio nos meu projetos.

A Ana Lúcia, por divertir até os dias mais tristes.

Ao meu amigo Lucas Sievers, pela indicação para a bolsa e por me ajudar a abrir a mente para as oportunidades além do pre-conceito.

Agradeço aos amigos Matheus Macedo e Gustavo Fraklin, pelos alegres dias e passeios pelo departamento de física.

Aos meus colegas Anderson Alves, Letícia Maranhão, Eduardo Bino, Mauro pela ajuda na minha graduação e por deixarem tais dias mais alegres.

RESUMO

Neste trabalho utilizaremos a 'Minimum Spanning Tree' (MST) usando Distância de Correlação para estudar as relações entre as principais ações globais no período de 2018 a 2024. A matriz Distância de Correlação é dada por d(i,j) = $\sqrt{2(1-r(i,j))}$, onde r(i,j) é o coeficiente de correlação entre as vértices x_i e x_j. O coeficiente de correlação varia entre -1, anti correlação perfeita, e +1, correlação perfeita, e r(ij) = 0, significa variáveis descorrelacionadas, ou independentes. Correlação de uma variável consigo mesma é sempre igual à +1, $r_(i,i)$ = 1. Dessa forma, d(i,j) varia entre 0 para r(i,j)= +1, e 2 para r(i,j) = -1. Com a matriz de distâncias calculamos a MST, a rede que conecta todas as variáveis, sem ciclos, com a menor distância. Para isso usaremos o algoritmo de PRIM, que procura o vértice mais próximo de qualquer elemento da rede já encontrada. Podemos mostrar a MST através de grafos, ou reordenar a matriz de correlação pela ordem do MST. Com isso, visualizamos os 'clusters' formados por proximidade de variáveis. Utilizamos os dados do mercado financeiro composto por os principais índices de ações dos países obtidos do banco de dados da 'Yahoo Finance' no período 2018 – 2024, que incluiu os anos de 2020-2022 com pandemia e lockdown, permitindo uma comparação com os anos de 2018-2019 para entender os efeitos econômicos do lockdown. Por outro lado, no período de 2022 e 2023 aconteceu a guerra Ucrânia-Rússia, com fortes impactos nos preços de energia e alimentos, e, consequentemente, nos valores das ações de várias empresas.

Palavras-chave: Minimum Spanning Tree; lockdown; guerra na Ucrânia.

ABSTRACT

In this work we will use the Minimum Spanning Tree (MST) using Correlation Distance to study the relationships between the main global stocks in the period from 2018 to 2024. The Correlation Distance matrix is given by d(i,j) = $\sqrt{2(1 - r(i, j))}$, where r(i,j) is the surface coefficient between the vertices x_i and x_j. The brightness coefficient varies between -1, perfect anti-transparency, and +1, perfect transparency, and r(ij) = 0, meaning uncorrelated, or independent, variables. Correlation of a variable with itself is always equal to +1, $r_(i, i)$ = 1. Thus, d(i,j) varies between 0 for r(i,j)= +1, and 2 for r(ij,j) = -1. Using the distance matrix, we calculate the MST, the network that connects all variables, without cycles, with the shortest distance. To do this, we will use the PRIM algorithm, which searches for the closest vertex to any element of the network already found. We can show the MST through graphs, or reorder the expansion matrix in the order of the MST. With this, we visualize the clusters formed by proximity of variables. We use financial market data composed of the main stock indexes of the countries obtained from the 'Yahoo Finance' database for the period 2018 - 2024, which included the years 2020-2022 with pandemic and lockdown, allowing a comparison with the years 2018-2019 to understand the economic effects of the lockdown. On the other hand, in the period 2022 and 2023, the Ukraine-Russia war took place, with strong impacts on energy and food prices, and, consequently, on the stock values of several companies.

Keywords: Minimum Spanning Tree ; lockdown ; war in Ukraine.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AEX	Amsterdam Exchange Index
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSI 300	China Securities Index 300
dim	dimensão
EUA	Estados Unidos da América
ME	Moscow Exchange
MST	Minimum Spanning Tree
SMI	Swiss Market Index
UFC	Universidade Federal do Ceara
v.a's	variáveis aleatórias

LISTA DE SÍMBOLOS

$C(v)$	Custo mais barato de uma conexão com o vértice v
$E(v)$	Aresta que fornece a conexão mais barata para o vértice v
V	Conjunto de vértices
E	Valor esperado
n	Número de nós
X, Y	Variáveis aleatórias
σ^2	Variância
$\mathbb{E}(X)$	Esperança matemática
μ_X	Média
A, B, C	Matrizes genéricas
P	Probabilidade
$\operatorname{cov}(X, Y)$	Covariância explícita
σ	Desvio padrão
ρ	Coeficiente de correlação de Pearson da população
$L_1(x), \ldots, L_n(x)$	Funções de base de Langrange
λ_i	Autovalores
$w_n(x)$	Base de Newton
$d(x)$	distancia de correlação
φ	Produto interno

SUMÁRIO

1 INTRODUÇÃO

Na publicação do livro que deu origem ao termo Econofísica por Stanley, H Eugene e Mantegna, Rosario N (Mantegna (1999)) (Stanley e Mantegna (2000)) surgiu a ideia de buscar relações de correlação entre ações negociadas em bolsas de valores. Usando a distância de correlação $d(i,j) = \sqrt{2-(1-r(i,j))}$,onde $r_ij = \frac{cov_(i,j)}{\sqrt{cov_(i,i)cov_(j,j)}}$ é o coeficiente de correlação de Pearson, utilizaram a Minimal Spanning Tree (MST) para buscar relações entre as ações. A distância de correlação definida dessa forma obedece aos 3 axiomas para uma distância. Até onde sabemos todos os trabalhos utilizando essa técnica se aplicaram à ações de bolsas de determinados países em certas datas específicas, ou na correlação entre os índices de bolsas de valores de diferentes países. Com a disponibilidade de dados atual, entretanto, é possível fazer um trabalho mais completo em dois sentidos: (1) expandir as correlações para principais ações das bolsas de diversos países economicamente importantes, e (2) incluir períodos de instabilidade global para caracterizar a robustez das relações de correlações. Para isso coletamos dados desde 2018 até 2024. Lembrando que em 2020 a pandemia de COVID e o consequente lockdown foi um evento global com impacto em diversos setores da economia, com quedas no PIB da maioria dos países em torno de 4,5%. Logo depois surgiu a guerra Rússia-Ucrânia e a Guerra Israel-Hamas. Todos esses eventos têm impacto na economia global que desejamos observar através da organização das matrizes de correlação. Para capturar os eventos no tempo utilizamos um período de 6 meses para cada ano considerando 2018-2019 um período de normalidade a ser comparado com os períodos posteriores.

Na internet atual é comum encontrar as expressões do tipo: Correlação Não é Causalidade, ou da falácia Post hoc ergo procter hoc, sobre causalidade e correlação. Entretanto, todo o campo da Epidemiologia foi baseado nas correlações entre doenças e suas possíveis causas. Desde John Snow, que fez uma correlação espacial entre poços de água e epidemia de cólera em Londres, que a correlação tem um papel fundamental na descoberta das causas das doenças. Se a correlação não é obrigatoriamente uma relação de causalidade, por outro lado a ausência de correlação aponta fortemente para uma ausência de causalidade. Exemplos de falta de correlação mesmo na presença da causalidade são forçados incluindo forças em direções contrárias que se anulam. As condições de causalidade são muito mais restritas, incluindo sempre a condição de que a causa vem antes do efeito, e requer um estudo mais profundo entre as variáveis que estudos de correlação apontam como mais prováveis para a relação de causa e efeito. No entanto as utilizações da matriz de correlação nos estudos do mercado financeiro

não objetivam a extração de relações de causa e efeito, mas de similaridades entre grupos de ações e clusterização das mesmas. A MST tem sido uma das técnicas preferidas entre as diversas técnicas de clusterização.

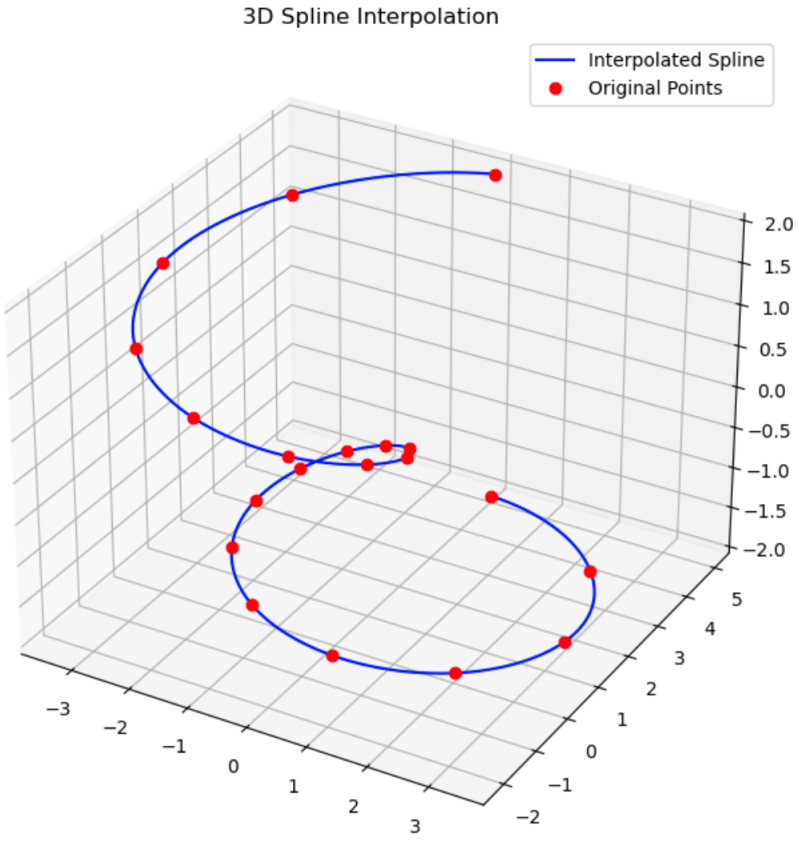
De uma forma geral, este trabalho pega os fechamentos dos valores das ações dos principais índices mundiais e monta um banco de dados, indexando-os em uma matriz. Sobre esses dados aplicamos um processo de limpeza para (1) completar dados que faltam. Feriados, por exemplo, são diferentes em países diferentes. Aplicamos a técnica de interpolação nos casos de dados ausentes em períodos curtos. (2) Empresas que entram e saem da bolsa de valores. Empresas com ações na bolsa de valores que passou um semestre em aberto ou continuou vendendo após ter fechado foram excluídas.

Com a matriz de correlação apenas já foi possível observar mudanças nos períodos de instabilidade global distribuição geográfica por países. Com a matriz de correlação obtivemos a matriz de distância na qual aplicamos o algoritmo de Prim para encontrar a MST. São dois os algoritmos mais utilizados para encontrar a MST, o de Kruskal e o de Prim. Ambos chegam na mesma MST no final. Entretanto, embora o algoritmo de Kruskal seja mais rápido, o de Prim é especialmente adequado para identificação de clusters, porque ele vai construindo a MST buscando os vértices mais próximos da sub rede já encontrada, ou seja, que pertencem a um cluster. Já Kruskal vai encontrando as sub redes em paralelo, não conectadas, para conectá-las apenas no final. Preocupados em observar e definir regiões e setores mais impactados pela instabilidade global aplicamos a MST para o reordenamento das matrizes de correlação em nível global, e, em dois períodos, incluindo o reordenamento interno de cada país. Dessa forma, nesse trabalho discutimos os conceitos de interpolação; os conceitos de teoria da probabilidade, covariância, correlação de Pearson e distância de correlação de forma axiomática; a aquisição dos dados globais; e finalmente apresentamos os resultados obtidos com essas análises.

2 INTERPOLAÇÃO

Interpolação é um método matemático de estimar novos pontos ou preencher dados faltosos a partir do conjunto de dados anteriores e posteriores a esse, ou seja, é baseado no intervalo de um conjunto discreto de pontos de dados conhecidos(Steffensen (2013)).

Figura 1 – Interpolação de um conjunto finito de pontos em uma curva. Os pontos em vermelho são conectados por curvas de spline interpoladas azuis deduzidas apenas dos pontos vermelhos. As curvas interpoladas têm fórmulas polinomiais muito mais simples do que a da curva original.



Fonte: Feito pelo autor(2024).

2.1 Interpolação linear

A interpolação linear (Stern *et al.* (2015)) é um método empregado para ajustar curvas usando polinômios lineares, ou seja, usaremos para reconstruir dados faltosos de nossa matriz, já que não podemos retirar toda uma coluna ou linha por causa de 1 ou 3 dados faltosos, excluindo assim uma variável importante de nosso trabalho.

2.1.1 Método

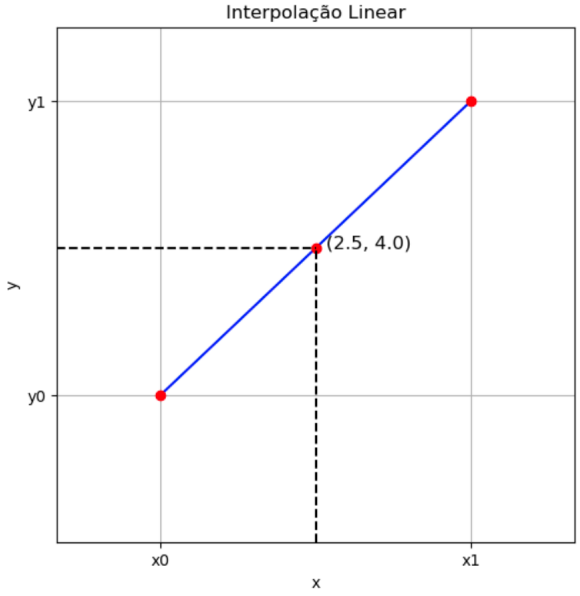
Dado introdução a seção, imagine dois pontos (x_0, y_0) e (x_1, y_1), onde o interpolante linear é a linha que passa por esses dois pontos , ou seja, se queremos encontrar y do nosso ponto desconhecido (x, y) temos que,

$$\begin{aligned} y &= y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} = \\ &= \frac{y_0(x_1 - x_0)}{x_1 - x_0} + \frac{y_1(x - x_0) - y_0(x - x_0)}{x_1 - x_0} = \\ &= \frac{y_1 x - y_1 x_0 - y_0 x + y_0 x_0 + y_0 x_1 - y_0 x_0}{x_1 - x_0} = \\ &= \frac{y_0(x_1 - x) + y_1(x - x_0)}{x_1 - x_0}, \end{aligned} \tag{2.1}$$

onde podemos estender para uma média ponderada, onde os pesos estão inversamente relacionado á distancia dos pontos finais ao ponto desconhecido, ou seja, o ponto mais próximo tem mais influencia do que o ponto mais distante. onde os pesos são,

$$1 - \frac{x - x_0}{x1 - x0} \qquad \text{e} \qquad 1 - \frac{(x_1 - x)}{x_1 - x_0} \tag{2.2}$$

Figura 2 – Dado os dois pontos vermelhos, a linha azul é o interpolante linear entre os pontos, e o valor y em x pode ser encontrado por interpolação linear.



Fonte: Feito pelo autor (2024).

onde, são distancias normalizadas entre o ponto desconhecido e cada um dos pontos finais, dado que estes somam para 1,

$$\begin{aligned} y &= y_0\left(1-\frac{x-x_0}{x_1-x_0}\right)+y_1\left(1-\frac{x_1-x}{x_1-x_0}\right)= \\ &= y_0\left(1-\frac{x-x_0}{x_1-x_0}\right)+y_1\left(\frac{x-x_0}{x_1-x_0}\right)= \\ &= y_0\left(\frac{x_1-x}{x_1-x_0}\right)+y_1\left(\frac{x-x_0}{x_1-x_0}\right) \end{aligned} \tag{2.3}$$

2.2 Interpolação polinomial

Vamos generalizar para o caso polinomial, a interpolação polinomial é a interpolação de um conjunto de dados bi-variados pelo polinômio de menor grau possível que passa pelo pontos do conjunto de dados (Kress (2012)).

2.2.1 Teorema da interpolação

Para quaisquer $n+1$pontos de dados bivariadas (Davis (1963)) $(x_0,y_0),\ldots,(x_n,y_n)\in\mathbb{R}^2$, onde não $x_j$há dois iguais, existe um polinômio único $p(x)$de grau no máximo $n$que interpola esses pontos, ou seja, $p(x_0)=y_0,\ldots,p(x_n)=y_n$

Equivalentemente, para uma escolha fixa de nós de interpolação x_j, a interpolação polinomial define uma bijeção linear $L_n$entre os (n +1)-tuplas de valores de número real $(y_0,\ldots,y_n)\in\mathbb{R}^{n+1}$e o espaço vetorial $P(n)$de polinômios reais de grau no máximo n :

$$L_n:\mathbb{R}^{n+1}\xrightarrow{\sim}P(n). \tag{2.4}$$

Prova pela funções base de Lagrange

Considere as funções de base Lagrange (Pletzer e Fillmore (2015)) $L_1(x),\ldots,L_n(x)$por:

$$L_j(x)=\prod_{i\neq j}\frac{x-x_i}{x_j-x_i}=\frac{(x-x_0)\cdots\left(x-x_{j-1}\right)\left(x-x_{j+1}\right)\cdots(x-x_n)}{\left(x_j-x_0\right)\cdots\left(x_j-x_{j-1}\right)\left(x_j-x_{j+1}\right)\cdots\left(x_j-x_n\right)} \tag{2.5}$$

Observe que $L_j(x)$ é um polinômio de grau n, e nós temos $L_j(x_k)=0$para cada um $j\neq k$, enquanto $L_k(x_k)=1$. Segue-se que a combinação linear:

$$p(x)=\sum_{j=0}^{n}y_jL_j(x) \tag{2.6}$$

assim,

$$p(x_k) = \sum_j y_j L_j(x_k) = y_k \tag{2.7}$$

$p(x)$um polinômio interpolante de grau n.

Prova por sistema de equações lineares

Tendo o polinômio na forma,

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0. \tag{2.8}$$

Substituindo-se isso nas equações de interpolação $p(x_j) = y_j$, temos um sistema de equações lineares nos coeficientes a a_j, que se lê em forma matricial-vetor como a seguinte multiplicação :

$$\begin{bmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \dots & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \dots & x_1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \dots & x_n & 1 \end{bmatrix} \begin{bmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{2.9}$$

Um interpolante $p(x)$corresponde a uma solução $A = (a_n, \dots, a_0)$da equação da matriz acima $X \cdot A = Y$. A matriz X à esquerda é uma matriz de Vandermonde, cujo determinante é conhecido $\det(X) = \prod_{1 \le i < j \le n} (x_j - x_i)$,por ser que não é diferente de zero, uma vez que os nós $x_j$são todos distintos. Isso garante que a matriz seja invertível e a equação tenha a solução única $A = X^{-1} \cdot Y$; isto é, $p(x)$existe e é única.

Corolário

Se $f(x)$é um polinômio de grau no máximo n, então o polinômio interpolar de $f(x)$pontos $n+1$distintos é $f(x)$ele próprio.

Interpolação de Lagrange

Escrevendo o polinômio em termos de polinômios de Lagrange, temos,

$$
\begin{aligned}
p(x) = & \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)} y_0 + \\
& + \frac{(x-x_0)(x-x_2)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_n)} y_1 + \\
& + \cdots + \\
& + \frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})} y_n + \\
= & \sum_{i=0}^{n} \left(\prod_{\substack{0 \le j \le n+ \\ j \neq i}} \frac{x-x_j}{x_i-x_j} \right) y_i = \sum_{i=0}^{n} \frac{p(x)}{p'(x_i)(x-x_i)} y_i
\end{aligned}
\tag{2.10}
$$

Para a forma matricial se utiliza a formula de Sylvester (Jones (1999)) e os polinômios de Lagrange com valor matricial são os covariantes de de Frobenius. Onde, dada uma matriz diagonalizável A expressa por uma função analítica $f(A)$, em termos de autovetores e autovalores de A, temos,

$$
f(A) = \sum_{i=1}^{k} f(\lambda_i) A_i \tag{2.11}
$$

onde os $-\lambda_i$ são os autovalores de A, e a matriz,

$$
A_i \equiv \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{1}{\lambda_i - \lambda_j} \left(A - \lambda_j I \right) \tag{2.12}
$$

são os covariantes de Frobenius correspondentes de A, que nada mais são do que a projeção da matriz dos polinômios de Lagrange de A.

Não obstante, há uma generalização baseada nos polinômios interpolantes de Hermite, tal que,

$$
f(A) = \sum_{i=1}^{s} \left[\sum_{j=0}^{n_i-1} \frac{1}{j!} \phi_i^{(j)}(\lambda_i) (A - \lambda_i I)^j \prod_{j=1, j \neq i}^{s} \left(A - \lambda_j I \right)^{n_j} \right] \tag{2.13}
$$

onde,

$$
\phi_i(t) := f(t) / \prod_{j \neq i} \left(t - \lambda_j \right)^{n_j} \tag{2.14}
$$

ou, também temos,

$$f(A) = \sum_{i=1}^{s} A_i \sum_{j=0}^{n_i-1} \frac{f^{(j)}(\lambda_i)}{j!} (A - \lambda_i I)^j \tag{2.15}$$

onde A_i são os covariantes correspondentes de A.

Formula restante de Lagrange

Ao interpolar uma determinada função f por um polinômio P n de grau n nos nós $x_0, ..., x_n$ teremos o erro,

$$f(x) - p_n(x) = f[x_0, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i) \tag{2.16}$$

onde $f[x_0, ..., x_n, x]$ é a $(n+1)^a$ diferença dividida dos pontos dados,

$$(x_0, f(x_0)), \ldots, (x_n, f(x_n)), (x, f(x)) \tag{2.17}$$

Além disso, há uma forma restante de Lagrange do erro, para uma função f que é n + 1 vezes continuamente direcionável em um intervalo fechado I, e um polinômio $p_n(x)$de grau na maioria n que interpola f em n + 1 pontos distintos $x_0, \ldots, x_n \in I$. Para cada um $x \in I$existe $\xi \in I$ tal que,

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{2.18}$$

Este limite de erro sugere a escolha dos pontos de interpolação para minimizar o produto $|\prod(x - x_i)|$, o que é conseguido pelos nós de Chebyshev.

Prova do resto de Lagrange

Definindo o termo erro como $R_n(x) = f(x) - p_n(x)$ e uma função auxiliar,

$$Y(t) = R_n(t) - \frac{R_n(x)}{W(x)} W(t) \tag{2.19}$$

onde,

$$W(t) = \prod_{i=0}^{n} (t - x_i) \tag{2.20}$$

assim,

$$Y^{(n+1)}(t) = R_n^{(n+1)}(t) - \frac{R_n(x)}{W(x)}(n+1)! \tag{2.21}$$

tal que $P_n(x)$ que é um polinomio de grau máximo de n, temos que $R_n^{(n+1)}(t) = f^{(n+1)}(t)$, e,

$$Y^{(n+1)}(t) = f^{(n+1)}(t) - \frac{R_n(x)}{W(x)}(n+1)! \tag{2.22}$$

Agora, uma vez que x i são raízes de $R_n(t)$ e $W(t)$, temos $Y(x) = Y(x_j) = 0$, o que significa que Y tem pelo menos n + 2 raízes. Do teorema de Rolle, $Y'(t)$tem pelo menos n + 1 raízes, e iterativamente $Y^{(n+1)}(t)$ tem pelo menos uma raiz - no intervalo I. Assim:

$$Y^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{R_n(x)}{W(x)}(n+1)! = 0 \tag{2.23}$$

e,

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{2.24}$$

Não obstante, também temos a **interpolação de Newton** (Pletzer e Hayek (2019)) que tem o teorema:

Para um polinômio p_n de grau menor ou igual a n, que se interpola f nos nós $x_i$onde $i = 0, 1, 2, 3, \cdots, n$. Seja p_{n+1}o polinômio de grau menor ou igual a n+1 que se interpola f nos nós $x_i$onde $i = 0, 1, 2, 3, \cdots, n, n+1$. Então p_{n+1}é dado por,

$$p_{n+1}(x) = p_n(x) + a_{n+1} w_n(x) \tag{2.25}$$

onde $w_n(x) := \prod_{i=0}^{n}(x - x_i)$também conhecido como base de Newton e,

$$a_{n+1} := \frac{f(x_{n+1}) - p_n(x_{n+1})}{w_n(x_{n+1})} \tag{2.26}$$

Onde será apenas elucidado para mostrar o arsenal que temos para a parte de interpolação de dados . No caso de duvida do leitor, esse pode utilizar o diagrama de Lozenge para ver uma formula que se encaixa ao seu conjunto de dados.

2.3 Interpolação Spline

Para este caso (Hall e Meyer (1976)), temos uma forma diferente de interpolar, onde o interpolante é um polinômio fragmentado onde o denotamos de Spline. Não obstante, em vez de ajustar um único polinômio de alto grau a todos os valores de uma vez, a interpolação Spline ajusta um polinômio de baixo grau a um pequeno subconjunto de valores, por exemplo, ajustando nove cúbicas entre cada um dos dez pares de polinômios de pontos em vez de ajustar um polinômio de grau único para todos os polinômios. A interpolação spline é geralmente preferida à interpolação polinomial porque o erro de interpolação pode ser pequeno mesmo quando polinômios de baixo grau são usados no spline. A interpolação também evita o problema do **fenômeno de Runge**, onde ocorrem oscilações entre pontos ao interpolar com polinômios de ordem superior.

2.3.1 Modelo matemático

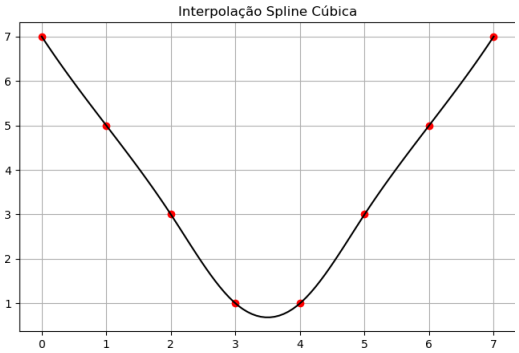
Queremos modelar tipos semelhantes de curvas usando um conjunto de equações matemáticas. Suponha que temos uma sequência de $n+1$ nós, (x_0, y_0) através (x_n, y_n). Haverá um polinômio cúbico $q_i(x) = y$ entre cada par sucessivo de nós (x_{i-1}, y_{i-1}) e (x_i, y_i) conexão com ambos, onde $i = 1, 2, \ldots, n$. Assim, haverá n polinômios, com o primeiro polinômio começando em (x_0, y_0), e o último termo nulo em (x_n, y_n).

A curvatura de uma curva genérica $y = y(x)$ é definida por,

$$\kappa = \frac{y''}{(1+y'^2)^{3/2}} \tag{2.27}$$

tal que, para minimizar a flexão(sob a restrição de passar por todos os nós), definiremos y' e y'' contínuos em todos os lugares, inclusive nos nós. Cada polinômio deve ter um valor igual ao y correspondente, ou seja,

Figura 3 – Interpolação com splines cúbicos entre oito pontos.



Fonte: Feito pelo autor (2024).

$$\begin{cases} q_i(x_i) = q_{i+1}(x_i) = y_i \\ q_i'(x_i) = q_{i+1}'(x_i) \qquad 1 \leq i \leq n-1. \\ q_i''(x_i) = q_{i+1}''(x_i) \end{cases} \tag{2.28}$$

Só é possível para polinômios de grau 3 ou superior, onde se é mais utilizado, tanto antigamente por engenheiro navais e projetistas como hoje por programadores e cientistas, é utilizar os splines cúbicos.

Exemplo

Considere a função,

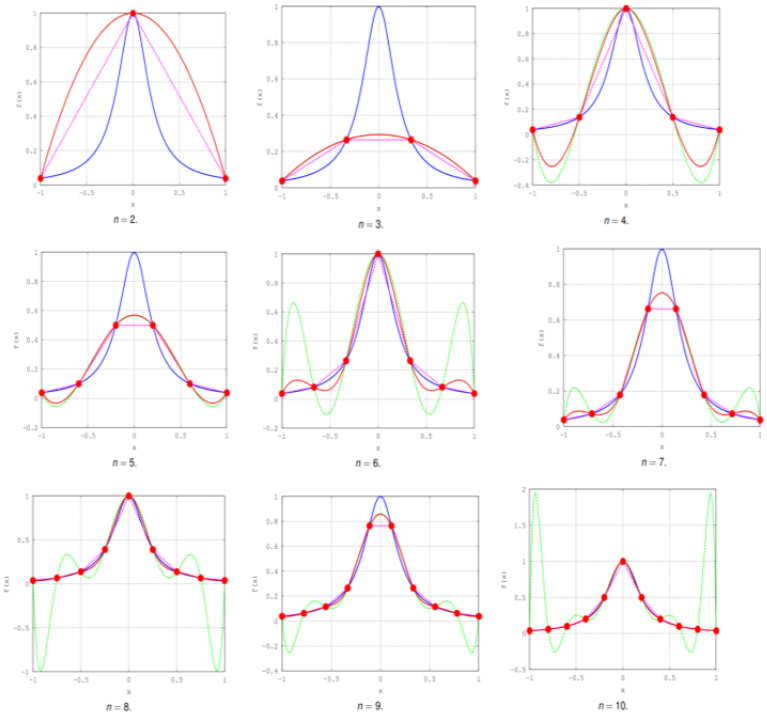
$$f(x) = \frac{1}{1+25x^2}, \quad x \in [-1,+1] \tag{2.29}$$

e nós,

$$x_k = -1 + \frac{2}{n}k, \quad k = 0,1,\ldots,n \tag{2.30}$$

igualmente espaçados no intervalo [1,-1]. Onde, as figuras 4 mostraram f(x) de azul, seu polinômio interpolador de verde, interpolador linear de magenta e a spline cubica de vermelho.

Figura 4 – f(x) azul, polinomial verde, linear magenta, spline vermelha.



Fonte: Prof. Dr. Marcos Eduardo Valle -IMECC - Unicamp

3 ELEMENTOS DA TEORIA DE PROBABILIDADE

3.1 Definição Axiomática da probabilidade

Seja Ω o conjunto do espaço amostral e A um subconjunto conjunto mensurável desse espaço, ou seja, A é um evento. A probabilidade do evento é uma função de conjunto dada por $P(A) = f : \Omega \to \Re$ que associa cada conjunto evento à um número real. Para ser uma probabilidade essa função deve obedecer à apenas três axiomas:

1. $P(A) \geq 0 \quad \forall A \in \Omega$
2. $P(\Omega) = 1$ Onde Ω é chamado de evento certo.
3. Se $A \cap B = \emptyset$ então $P(A \cup B) = P(A) + P(B)$

Com esses axiomas é possível mostrar, como teorema, que o contradomínio da função probabilidade está restrito ao conjunto $[0,1]$. Seja $\bar{A}$ o conjunto complementar de A, ou seja, $A \cup \bar{A} = \Omega$. Se $\bar{A}$ é o complementar de A, então $A \cap \bar{A} = \emptyset$. Pelo axioma (3), então

$$P(A \cup \bar{A}) = P(\Omega) = 1 \tag{3.1}$$

e pelo axioma (2), temos

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}), \tag{3.2}$$

logo, com o axioma (1), temos que

$$P(A) \geq 0 \quad \text{e} \quad P(\bar{A}) \geq 0 \quad \Rightarrow \quad P(A) + P(\bar{A}) = 1, \tag{3.3}$$

o que implica que

$$0 \leq P(A) \leq 1. \tag{3.4}$$

Um ponto importante da teoria da probabilidade para esse trabalho se refere aos eventos independentes. Os eventos A e B são independentes se

$$P(A \cap B) = P(A)P(B). \tag{3.5}$$

3.2 Variável Aleatória

Uma nova função de conjuntos $f : \Omega \to \mathbb{R}$ no espaço amostral associando eventos a números reais é utilizada para construir uma variável aleatória. Por exemplo, em um dado, as faces são associadas a números entre 1 e 6. Jogo de 2 dados simultâneos pode ser associado à variável soma das faces dos dados. As cores podem ser associadas a um conjunto de 3 números no sistema RGB. Essas variáveis aleatórias podem ser discretas, mistas ou contínuas. É comum fazer tratamentos diferenciados para variáveis contínuas e discretas, no entanto, a utilização das funções Delta de Dirac permite generalizar o tratamento para qualquer tipo de variável.

3.3 Função Distribuição de Probabilidade denotada por F maiúsculo $F_{(x)}$

Seja o conjunto $A = \{x \leq s\}$ um evento ao qual associamos a probabilidade $P(A)$. A Função Distribuição de Probabilidade é definida por:

$$F(x) = P(x \leq s), \quad x \in \mathbb{R}$$

Essa função tem as seguintes propriedades:

$$F(-\infty) = 0 \quad \text{e} \quad F(+\infty) = 1 \tag{3.6}$$

e,

$$F(+\infty) = P(x \leq +\infty) = P(\Omega) = 1 \tag{3.7}$$

tambem,

$$F(-\infty) = P(x \leq -\infty) = P(\emptyset) = 0 \tag{3.8}$$

se

$$x_2 > x_1 \tag{3.9}$$

então

$$F(x_2) \geq F(x_1) \tag{3.10}$$

pois

$$P(x_2 \leq x_1) < P(x_1 \leq x \leq x_2) \tag{3.11}$$

logo

$$F(x_1) \leq F(x_2) \tag{3.12}$$

Vale a pena notar que essa função é adimensional, apenas um número entre 0 e 1, ou entre 0 e 100% em porcentagem.

3.4 Função Densidade de Probabilidade denotada por f minúsculo $f(x)$

Aqui é onde as diferenças entre variáveis contínuas e discretas aparecem. Utilizando variável contínua, a função densidade de probabilidade é a derivada da função distribuição de probabilidade:

$$f(x) = \frac{dF(x)}{dx} \quad \Rightarrow \quad F(x) = \int_{-\infty}^{x} f(x')dx' \tag{3.13}$$

Note que a dimensão de f(x)é

$$f(x) = \frac{1}{x} \tag{3.14}$$

Como $F(x)$ não é decrescente, então

$$f(x) \geq 0 \tag{3.15}$$

e como

$$F(+\infty) = 1 \tag{3.16}$$

então

$$F(x) = \int_{-\infty}^{x} f(x')dx = 1 \tag{3.17}$$

Ou seja, para ser uma função densidade de probabilidade, a função deve ser sempre positiva ou nula e a área abaixo dela tem que valer 1.

O problema com as variáveis discretas é que $F(x)$ tem descontinuidades nas quais a função não seria diferenciável, logo, $f(x)$ não existiria nesses pontos. Aqui entra a função Delta

de Dirac. Dirac criou uma função com área unitária cuja largura tende a zero, e a altura tende a infinito para manter a área constante. Daí mostrou que a derivada da função Degrau definida por:

$$degrau(x-x_0) = \begin{cases} 1 & \text{se } x \geq x_0, \\ 0 & \text{se } x < x_0 \end{cases} \tag{3.18}$$

Trata-se de uma função descontínua, logo, não diferenciável. Entretanto, ela pode ser aproximada por uma função contínua que se torna cada vez mais fina, por exemplo, usando a logística:

$$L_n(x) = \frac{1}{1+e^{-n(x-x_0)}} \Rightarrow L_n(-\infty) = 0, \quad L_n(+\infty) = 1 \tag{3.19}$$

Quanto maior o parâmetro n, mais rápido a função salta de zero para 1. Mas essa função admite derivada dada por

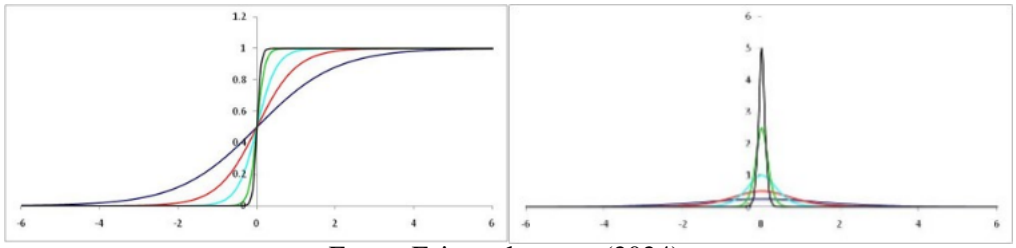
$$\frac{dL_n}{dx} = \frac{ne^{-n(x-x_0)}}{\left(1+e^{-n(x-x_0)}\right)^2} \tag{3.20}$$

cuja área vale 1, pois

$$\int_{-\infty}^{+\infty} \frac{dL_n}{dx} dx = L_n(+\infty) - L_n(-\infty) = 1 \tag{3.21}$$

Figura 5 mostra as curvas $L_n(x)$ para vários valores de n. Se $n \to \infty$, a logística se torna a função degrau enquanto $\frac{dL_n}{dx}$ se torna a função Delta de Dirac, cuja área sempre vale 1, a altura cresce para infinito e a largura tende a zero.

Figura 5 – Curvas de L para varios valores de n



Fonte: Feito pelo autor (2024).

As duas propriedades da função Delta de Dirac que utilizaremos são:

$$\int_a^b f(x)\delta(x-x_0)dx = \begin{cases} f(x_0), & \text{se } x_0 \in [a,b] \\ 0, & \text{se } x_0 \notin [a,b] \end{cases} \quad \Rightarrow \quad \int f(x)\delta'(x-x_0)dx = -f'(x_0) \tag{3.22}$$

$$\int f(x)\delta'(x-x_0)dx' = \text{Degrau}(x-x_0) \quad \Rightarrow \quad \frac{d}{dx}\text{Degrau}(x-x_0) = \delta(x-x_0) \tag{3.23}$$

Uma boa representação da função Delta de Dirac é uma seta no ponto de descontinuidade.

Derivando funções descontínuas. Seja a função descontínua

$$f(x) = \begin{cases} f_<(x), & \text{se } x < x_0, \\ f_>(x), & \text{se } x \geq x_0 \end{cases} \tag{3.24}$$

Em termos da função degrau, ela pode ser escrita como:

$$f(x) = f_<(x) + (f_>(x) - f_<(x))\,\text{Degrau}(x - x_0) \tag{3.25}$$

Sua derivada pela regra do produto é dada por:

$$f'(x) = f_1'(x) + [(f_2'(x) - f_1'(x))]Degrau(x - x_0) + [f_2(x) - f_1(x)]\delta(x - x_0) \tag{3.26}$$

Dessa forma, foi possível generalizar o conceito de função densidade de probabilidade para os casos discretos e contínuos. Vale notar que a dimensão da função Delta de Dirac vale

$$\delta(x) = \frac{1}{x} \tag{3.27}$$

pois

$$\int \delta(x)dx = 1 \tag{3.28}$$

Então a Delta de Dirac tem dimensão de densidade de probabilidade.Um bom exemplo é a função densidade de probabilidade de um dado honesto com probabilidade $1/6$. Nesse caso:

$$f(x) = \frac{1}{6}(\delta(x-1) + \delta(x-2) + \delta(x-3) + \delta(x-4) + \delta(x-5) + \delta(x-6)) \tag{3.29}$$

Note que a $F(x)$ sai automaticamente de $f(x)$ por integração:

$$\begin{aligned} F(x) = \frac{1}{6}\Bigg(& \int_{-\infty}^{x} \delta(x'-1)dx' + \int_{-\infty}^{x} \delta(x'-2)dx' \\ & + \int_{-\infty}^{x} \delta(x'-3)dx' + \int_{-\infty}^{x} \delta(x'-4)dx' \\ & + \int_{-\infty}^{x} \delta(x'-5)dx' + \int_{-\infty}^{x} \delta(x'-6)dx' \Bigg) \end{aligned} \tag{3.30}$$

$$F(x) = (\frac{1}{6})\,(\text{Degrau}(x-1) + \text{Degrau}(x-2) \\ +\text{Degrau}(x-3)\text{Degrau}(x-4) \\ +\text{Degrau}(x-5) + \text{Degrau}(x-6) \tag{3.31}$$

3.5 Operação esperança

.

Esperança de uma variável aleatória é definida por:

$$E[x] = \int x f(x) dx \tag{3.32}$$

Também usamos a notação $\langle x \rangle = E[x]$. Note que como o caso discreto aparece automaticamente para função densidade de probabilidade

$$f(x) = \sum_i P_i \delta(x - x_i) \tag{3.33}$$

Aqui P_i é adimensional, com dimensão de probabilidade $\sum P = 1$. Nesse caso:

$$E[x] = \int x \sum_i P_i \delta(x - x_i) dx = \sum_i P_i \int x \delta(x - x_i) dx = \sum p_i x_i \tag{3.34}$$

A operação esperança para uma função de variável aleatória $\langle G(x) \rangle$ é dada por

$$\langle G(x) \rangle = \int G(x) f(x) dx \tag{3.35}$$

3.6 Momentos de uma distribuição de probabilidade

O momento de ordem n é definido por:

$$M_n = \langle x^n \rangle = \int_{-\infty}^{\infty} x^n f(x) dx. \tag{3.36}$$

Dessa definição, notamos que

$$M_0 = \int_{-\infty}^{\infty} f(x) dx = 1 \tag{3.37}$$

e que

$$M_1 = \langle x \rangle = \int_{-\infty}^{\infty} x f(x) dx. \tag{3.38}$$

A esperança de uma variável é denotada usualmente pela letra grega $\mu = \langle x \rangle$.Já o momento centrado de uma distribuição é definido por:

$$m_n = \langle (x-\mu)^n \rangle = \int_{-\infty}^{\infty} (x-\mu)^n f(x) dx. \tag{3.39}$$

O momento centrado de ordem 2 é chamado de variância, denotado por σ^2, dado por:

$$m_2 = \sigma^2 = \langle (x-\mu)^2 \rangle = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx, \tag{3.40}$$

enquanto $\sigma = \sqrt{\sigma^2}$ é chamado de Desvio Padrão. Vale notar a identidade da variância:

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx = \langle x^2 \rangle - \langle x \rangle^2. \tag{3.41}$$

3.7 Análise Multivariada

Para sistemas multivariados, a função densidade de probabilidade é uma função

$$f(x_1, x_2, \cdots, x_n) : \mathbb{R}^n \to \mathbb{R} \tag{3.42}$$

tal que

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n = 1. \tag{3.43}$$

No caso em que as variáveis x e y são independentes, então

$$f(x,y) = g(x)h(y). \tag{3.44}$$

Os momentos agora devem ser generalizados para:

$$M_{k_1 k_2 \cdots k_n} = \langle x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n} \rangle = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n} f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n. \tag{3.45}$$

$$m_{k_1 k_2 \cdots k_n} = \langle (x_1 - x_1)^{k_1} (x_2 - x_2)^{k_2} \cdots (x_n - x_n)^{k_n} \rangle. \tag{3.46}$$

Com os quais é comum a utilização da seguinte notação:

$$\mu_i = \langle x_i \rangle = \int_{-\infty}^{\infty} x_i f(x_1, x_2, \cdots, x_n) dx_1 \cdots dx_n, \to \sigma_i^2 = \langle (x_i - \mu_i)^2 \rangle. \tag{3.47}$$

Se as variáveis x e y são independentes, então:

$$M_{1,1} = \langle xy \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x,y) dx dy = \int_{-\infty}^{\infty} x g(x) dx \int_{-\infty}^{\infty} y h(y) dy = \langle x \rangle \langle y \rangle. \tag{3.48}$$

Para duas variáveis x e y, usamos as notações:

$$M_{10} = \langle x \rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)dxdy = \mu_x, \tag{3.49}$$

$$M_{01} = \langle y \rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)dxdy = \mu_y, \tag{3.50}$$

$$m_{20} = \langle (x-\mu_x)^2 \rangle = \sigma_x^2 \quad \rightarrow \quad m_{02} = \langle (y-\mu_y)^2 \rangle = \sigma_y^2, \tag{3.51}$$

$$m_{11} = \langle (x-\mu_x)(y-\mu_y) \rangle = \mathrm{COV}(x,y). \tag{3.52}$$

Entre os momentos de ordem 2, m_{20}, m_{02} e m_{11}, os momentos 02 e 20 são as variâncias de x e y, respectivamente, e o momento de ordem 11 é chamado de covariância. Em uma matriz de covariância

$$m = \begin{pmatrix} m_{20} & m_{11} \\ m_{11} & m_{02} \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_y^2 \end{pmatrix}, \tag{3.53}$$

os termos da diagonal são as variâncias e fora da diagonal as covariâncias.

A covariância é simétrica $\mathrm{cov}(x,y) = \mathrm{cov}(y,x)$ e tem a seguinte propriedade:

$$\mathrm{cov}(x,y) = \langle xy - x\mu_y - \mu_x y + \mu_x\mu_y \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle. \tag{3.54}$$

Isso significa que, se as variáveis forem independentes, a covariância será nula

$$\mathrm{cov}(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle = 0, \tag{3.55}$$

tendo que ela poderia ser usada como um bom indicador de independência estatística entre variáveis aleatórias.

4 COEFICIENTE E DISTÂNCIA DE CORRELAÇÃO

4.1 Coeficiente de correlação

A medida da covariância (Rice (2007)) como uma medida da independência entre duas variáveis aleatórias (v.a's), entretanto, apresenta alguns problemas. Primeiro trata-se de uma medida com dimensão, dimensão (dim),

$$\dim\left(\operatorname{cov}(x,y)\right) = \dim\left(x\right)\dim\left(y\right) \tag{4.1}$$

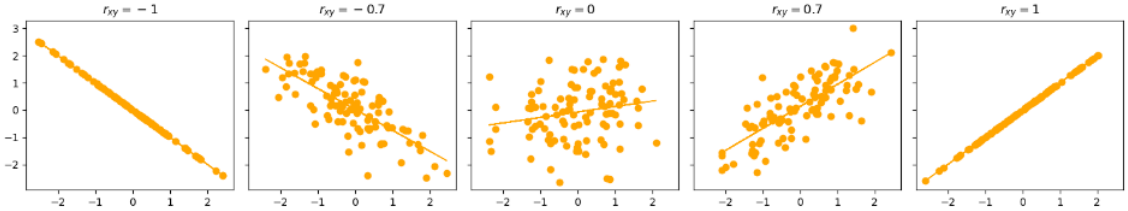
Se x e y têm dimensão de distância, ou massa, por exemplo, a covariância terá dimensão de área, ou massa ao quadrado. Precisamos de uma grandeza adimensional relacionada à covariância para ser utilizada como um grau de independência entre v.a.s (Ablowitz e Fokas (2003)). Então vamos construir o coeficiente de correlação adimensional definido por:

$$r_{xy} = r\left(x,y\right) = \frac{\operatorname{cov}(x,y)}{\sqrt{\mathrm{V}(x)\mathrm{V}(y)}} = \frac{\operatorname{cov}(x,y)}{\sqrt{\mathrm{V}[x]\,\mathrm{V}[y]}} \tag{4.2}$$

Com essa definição (Park (2017)) ganhamos mais do que simplesmente a obtenção de uma grandeza adimensional, porque podemos mostrar que se trata de um número que varia entre +1 e -1, com zero significando independência estatística, +1 correlação positiva perfeita e -1 correlação negativa, ou anti-correlação, perfeita.

$$-1 \leq r_{xy} \leq 1 \tag{4.3}$$

Figura 6 – Exemplos de diagramas de dispersão com diferentes valores de coeficiente de correlação



Fonte: Feito pelo autor (2024).

4.1.1 Teorema do coeficiente de correlação

Prova usando a desigualdade de *Schwartz*:

$$E\left[(\lambda(x-\mu_x)-(y-\mu_y))^2\right] \geq 0 \quad \forall\lambda \in \mathbb{R} \tag{4.4}$$

pois se trata da esperança (Ye (2010)) de uma quantidade positiva. Desenvolvendo o quadrado temos:

$$E[\lambda(x-\mu_x)-(y-\mu_y)]^2 = \lambda^2 E(x-\mu_x)^2 - 2\lambda E[(x-\mu_x)(y-\mu_y)] + E[(y-\mu_y)^2] \tag{4.5}$$

Logo,

$$E\left[\lambda(x-\mu_x)-(y-\mu_Y)\right] = \lambda E\left[(x-\mu_x)\right] - 2\lambda E\left[(x-\mu_x)(y-\mu_y)\right] + E\left[(y-\mu_y)\right] \tag{4.6}$$

e pode ser escrito em termos das variâncias como:

$$E\left[(\lambda(x-\mu_x)-(y-\mu_y))^2\right] = \lambda^2 \mathrm{V}(x) - 2\lambda \mathrm{cov}(x,y) + \mathrm{V}(y) \tag{4.7}$$

Isso nos leva à desigualdade da equação quadrática em λ dada por:

$$V[x]\lambda^2 - 2\mathrm{cov}_(x,y)\lambda + \mathrm{V}[y] \geq 0 \tag{4.8}$$

com $\mathrm{V}(x) \geq 0$ e $\mathrm{V}(y) \geq 0$.

A desigualdade $a\lambda^2 + b\lambda + c \geq 0$ com $a > 0$ só pode ser satisfeita se $a\lambda^2 + b\lambda + c = 0$ não admite raízes reais ou apenas uma raiz que toca o eixo λ. Essa condição implica que $b^2 - 4ac \leq 0$. Agora fazendo $a = \mathrm{V}(x)$, $b = -2\mathrm{cov}(x,y)$, e $c = \mathrm{V}(y)$ percebe-se que $4\mathrm{cov}^2(x,y) - 4\mathrm{V}(x)\mathrm{V}(y) \leq 0$ ou sej,

$$\frac{\mathrm{cov}^2(x,y)}{\mathrm{V}(x)\mathrm{V}(y)} \leq 1 \tag{4.9}$$

tambem,

$$-1 \leq \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{V}(x)\mathrm{V}(y)}} \leq 1 \tag{4.10}$$

Esse teorema pode ser generalizado (Helwig (2017)) e utilizado para definir ortogonalidade entre v.a.s.,

$$-1 \leq \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} \leq 1 \tag{4.11}$$

4.1.2 Espaços métricos e distância de correlação:

Um espaço é métrico (Vakhania (1999)) se existe uma função distância $d(x,y) : E \times E \to \mathbb{R}$ para $\forall x, y \in E$, satisfazendo aos axiomas:

1. Desigualdade triangular:
$$d(x,z) \leq d(x,y) + d(y,z) \tag{4.12}$$

2. Se
$$d(x,y) = 0 \tag{4.13}$$

   então $x = y$

3.
$$d(x,y) = d(y,x) \tag{4.14}$$

Com esses axiomas podemos demonstrar o teorema:

$$d(x,y) \geq 0 \tag{4.15}$$

Fazer $z = x$ no axioma 1:

$$d(x,x) \leq d(x,y) + d(y,x) \tag{4.16}$$

Usando os axiomas (2) e (3)

$$2d(x,y) \geq 0 \tag{4.17}$$

logo $d(x,y) \geq 0$.

Então a função distância deve ser um número real e positivo. Se essa função existe então ela é a métrica do espaço e podemos medir distâncias entre os elementos do conjunto E. Nesse caso dizemos que o espaço é métrico. A distância Euclidiana entre os vetores

$$x = (x_1, x_2, \ldots, x_n) \text{ e } y = (y_1, y_2, \ldots, y_n) \tag{4.18}$$

é definida como

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{j=1}^{n} (x_j - y_j)^2} \tag{4.19}$$

Esse foi o modelo de distância para definição dos axiomas de uma distância.

4.1.3 Definição de produto interno

O produto interno pode ser definido de forma generalizada através dos seguintes axiomas. Suponha um conjunto de elementos $\varphi_k | k \in \mathbb{Z} = 0, 1, 2, 3, ...$, que podem ser funções, vetores, matrizes, etc., e que admite uma operação denominada produto interno $\langle \varphi_i, \varphi_j \rangle : E \times E \to \mathbb{C}$ definida através de 3 axiomas:

1. Simetria conjugada

$$\langle \varphi_i | \varphi_j \rangle = \overline{\langle \varphi_j | \varphi_i \rangle} \tag{4.20}$$

2. Linearidade

$$\langle a\varphi_i + b\varphi_j | \varphi_k \rangle = a\langle \varphi_i | \varphi_k \rangle + b\langle \varphi_j | \varphi_k \rangle \tag{4.21}$$

3. Positiva definida

$$\langle \varphi_i, \varphi_i \rangle \geq 0 \tag{4.22}$$

e, se $\langle \varphi_i, \varphi_i \rangle = 0$, então $\varphi_i = 0$

Do axioma (1) podemos mostrar que $\langle \varphi_i | \varphi_i \rangle \in \mathbb{R}$ pois $\langle \varphi_i | \varphi_i \rangle = \overline{\langle \varphi_i | \varphi_i \rangle}$. Além disso $\langle \varphi_i | \varphi_i \rangle \geq 0$ pelo axioma (3). O produto interno então pode ser usado na definição de uma norma:

$$\| \varphi \| = \sqrt{\langle \varphi | \varphi \rangle} \tag{4.23}$$

4.1.4 Distância entre funções

Vejamos se a operação

$$\langle \varphi | \varphi_j \rangle = \int_a^b \varphi_i(x)\varphi_j(x)w(x)dx \tag{4.24}$$

satisfaz as condições de produto interno de funções $\varphi(x) : \mathbb{R} \to \mathbb{C}$, com uma função peso $w(x) : \mathbb{R} \to \mathbb{R}^+ \quad \forall x \in [a,b]$. Para a definição de produto interno, a restrição da função peso ser positiva é suficiente. Entretanto, como a média ponderada pode ser extraída através de pesos normalizados, sempre podemos exigir, sem perda de generalidade, que

$$\int_a^b w(x)dx = 1. \tag{4.25}$$

Se $\int_a^b w(x)dx \neq 1$, redefinimos

$$w(x) = \frac{w(x)}{\int_a^b w(x)dx}, \tag{4.26}$$

de modo que $\int_a^b w(x)dx = 1$. Notamos que as exigências de que $w(x) \geq 0 \forall x \in [a,b]$ coincidem com a exigência de que $w(x)$ seja uma função densidade de probabilidade no intervalo $[a,b]$. Note que dessa definição temos que

$$\langle \varphi | \varphi_j \rangle = \int_a^b \varphi_i(x)\varphi_j(x)w(x)dx \tag{4.27}$$

e

$$d(\varphi_i, \varphi_j) = \int_a^b (\varphi_i(x) - \varphi_j(x))^2 w(x)dx \tag{4.28}$$

é um produto interno legítimo, e que

$$d(\varphi_i, \varphi_j) = \int_a^b (\varphi_i(x) - \varphi_j(x))^2 w(x)dx \tag{4.29}$$

é uma distância.

A covariância (Zar (1972)), dada pelo momento de ordem 11:

$$\mathrm{cov}(x,y) = m_{11} = \int_{-\infty}^{+\infty} (x-\mu_x)(y-\mu_y)f(x,y)dxdy \tag{4.30}$$

é um produto interno:

$$\mathrm{cov}(x,y) = \langle (x-\mu_x), (y-\mu_y) \rangle f(x,y)dxdy, \tag{4.31}$$

uma vez que $f(x,y) \geq 0 \forall x,y$. Ou seja, a covariância é semelhante ao produto escalar entre dois vetores no espaço euclidiano. Nesse contexto, variáveis independentes são ortogonais. Também podemos associar a variância em termos da covariância na forma:

$$\mathrm{V}(x) = \mathrm{cov}(x,x) = \int_{-\infty}^{+\infty} (x-\mu_x)^2 f(x,y) dxdy \tag{4.32}$$

e

$$\mathrm{V}(y) = \mathrm{cov}(y,y) = \int_{-\infty}^{+\infty} (y-\mu_y)^2 f(x,y) dxdy. \tag{4.33}$$

Ou seja $\sigma_x = \langle x-\mu_x | x-\mu_x \rangle$ e $\sigma_y = \langle y-\mu_y | y-\mu_y \rangle$, logo temos,

$$\sigma_x = ||\langle x-\mu_x | x-\mu_x \rangle|| \tag{4.34}$$

e,

$$\sigma_y = ||\langle y-\mu_y | y-\mu_y \rangle|| \tag{4.35}$$

4.1.5 Distância de Correlação

Definindo as v.a's padronizadas

$$p = \frac{X-\mu_x}{\sigma_x} \quad \text{e} \quad q = \frac{Y-\mu_y}{\sigma_y}, \tag{4.36}$$

notamos que

$$E[p] = E[q] = 0, \tag{4.37}$$

assim como

$$E\left[p^2\right] = E\left[q^2\right] = 1. \tag{4.38}$$

Isso significa que se trata de vetores unitários. Nesse caso

$$r_{xy} = \mathrm{cov}(p,q) = E[pq]. \tag{4.39}$$

Podemos definir uma distância (Ratner (2009)) entre as variáveis X e Y através de

$$d_{xy}^2 = E\,(p-q)^2 = E\left[p^2\right] + E\left[q^2\right] - 2E[pq]. \tag{4.40}$$

Mas

$$E[pq] = \frac{E\left[(X-\mu_x)(Y-\mu_y)\right]}{\sigma_x \sigma_y} = r_{xy}, \tag{4.41}$$

e

$$E\left[p^2\right] = E\left[q^2\right] = 1, \tag{4.42}$$

logo

$$d_{xy}^2 = 2(1-r_{xy}). \tag{4.43}$$

Então a grandeza

$$d_{xy} = \sqrt{2(1-r_{xy})} \tag{4.44}$$

se comporta como uma distância, chamada de distância de correlação.
Como $-1 \leq r_{xy} \leq +1$, a distância de correlação varia entre

$$0 \leq d_{xy} \leq 2. \tag{4.45}$$

Quanto maior a correlação, menor a distância. Vale notar um ponto importante aqui. Para ser uma distância exigimos que $d(x,y) = 0$ então $x = y$. Mas

$$d(x,y) = 0 \quad \Leftrightarrow \quad r_{xy} = +1. \tag{4.46}$$

Duas v.a's relacionadas por uma transformação linear

$$X_j = aY_j + b \quad (\text{com } a > 0) \tag{4.47}$$

apresentam correlação $r_{xy} = +1$, embora $X_j \neq Y_j$. Entretanto, as duas variáveis padronizadas

$$p = \frac{X - \mu_x}{\sigma_x} \quad \text{e} \quad q = \frac{Y - \mu_y}{\sigma_y} \tag{4.48}$$

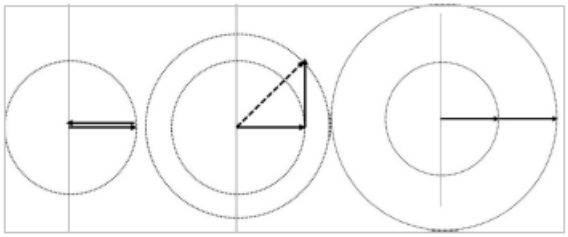
são idênticas.

Assim, ao realizar um mesmo experimento em duas amostras diferentes podemos medir uma distância de similaridade entre as duas através da distância de correlação dada por

$$d_{xy} = \sqrt{2(1 - r_{xy})} \quad \text{com } d_{xy} = 0 \tag{4.49}$$

significando similaridade total, e $d_{xy} = 2$, significando independência entre as amostras, $r_{xy} = 0$, e a oposição total entre as amostras, $r_{xy} = -1$. Os 3 casos podem ser apresentados como a subtração (soma) de vetores no círculo unitário da figura figura 7.

Figura 7 – Exemplo de correlação

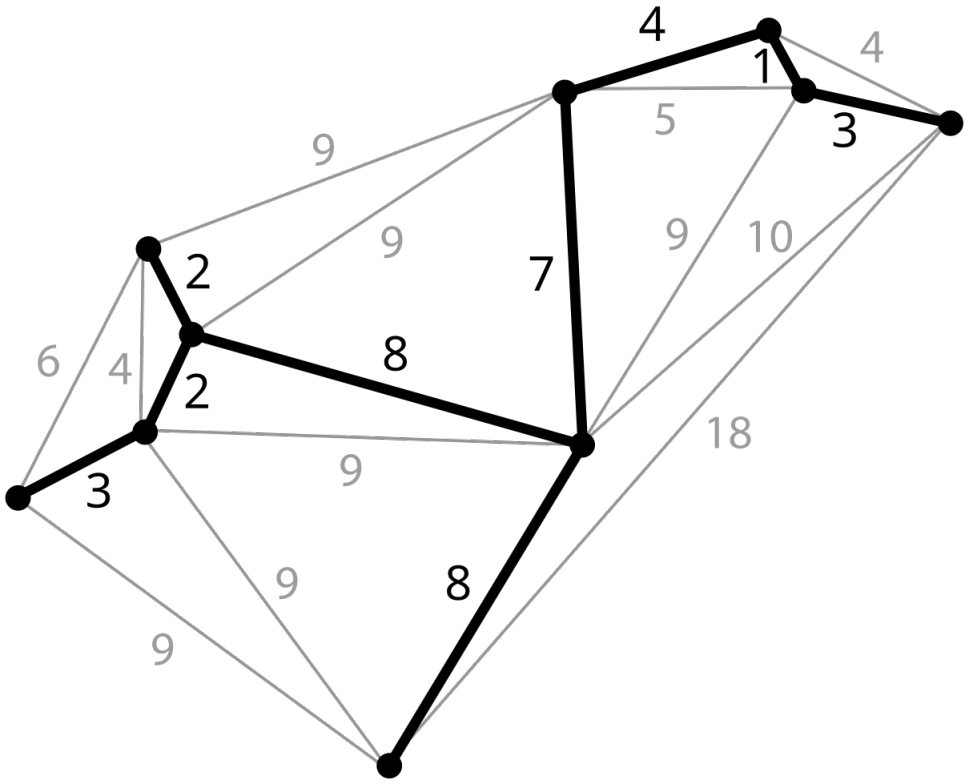


Fonte: index Lenz-Carlos Lenz Cesar-Unicamp

5 MINIMUM SPANNING TREE

5.1 Árvore de abrangência miníma

Figura 8 – Exemplo de conexão MST.



Fonte: wikipedia.org.

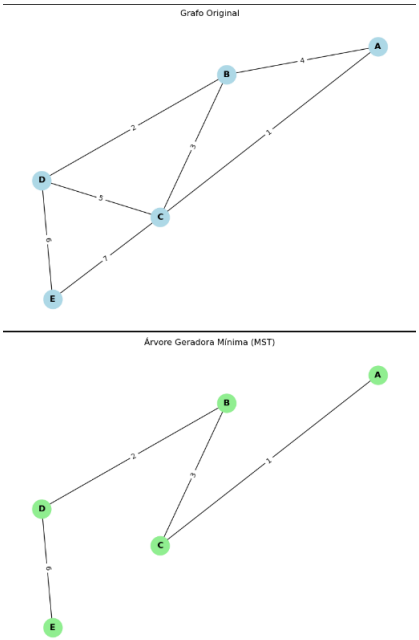
Árvore de peso mínimo (Sedgewick e Wayne (2011)) é um subconjunto das bordas de um grafo não direcionado, ponderado de borda e conectado a todos os vértices, sem ciclos e com o menor peso de borda possível. Em outras palavras, é uma árvore de extensão cujos pesos de borda somados são os menores possíveis. A maioria das vezes, qualquer grafo não direcionado à borda, embora não seja necessariamente conectado, tem uma união de floresta mínima.

A MST é usado em muitas situações. Uma empresa de telecomunicações que tenta instalar cabos em um novo bairro é um exemplo disso. Se o cabo for enterrado apenas ao longo de caminhos específicos (por exemplo, estradas), haveria um grafo contendo os pontos (por exemplo, casas) que estão conectados por esses caminhos. Devido ao fato de serem mais longos ou exigir que o cabo seja enterrado mais profundamente, alguns dos caminhos podem ser mais caros; esses caminhos seriam indicados por bordas com pesos maiores. O peso de borda da

moeda pode ser medido em unidades, pois não há necessidade de comprimentos de borda para atender a regras normais de geometria, como a desigualdade do triângulo. Para esse grafo, uma árvore de abrangência seria um subconjunto desses caminhos que não têm ciclos, mas conecta todas as casas, pode haver várias árvores se estendendo. Uma árvore de abrangência mínima é a mais barata e representa a melhor maneira de colocar o cabo.

Propriedades

Figura 9 – Exemplos de uma rede 'MST'



Fonte: Feito pelo autor (2024).

Se houver n nós na MST então haverá n-1 arestas conectando-as. Pode haver várias MST mínimas do mesmo peso, em particular, se todos os pesos de arestas de um determinado gráfico forem os mesmos, então cada árvore desse grafo é mínimo.

Singularidade

Existirá apenas um 'MST' se cada extremidade tiver seu próprio peso. Isso é verdade em vários cenários realistas (Heineman *et al.* (2009)), como o exemplo da empresa de telecomunicações mencionado anteriormente. Nesses casos, é pouco provável que os dois caminhos tenham os mesmos custos. Isso também se aplica à extensão de bosques. Assuma o oposto, que os 'MSTs' A e B são diferentes. Devido ao fato de que A e B têm os mesmos nos, há pelo menos uma borda que pertence à primeira, mas não à segunda. Seja e1 a aresta com o

menor peso entre estas. Esta escolha é especial porque os pesos das arestas são todos diferentes. Assuma que e1 está em A, sem perder generalidade.

Como B é um MST, e1 U B deve conter e1. Como um árvore, A não possui ciclos, então C deve ter uma borda e2 que não está em A. Como e1 foi selecionado como a única aresta de menor peso entre aquelas pertencentes a precisamente um de A e B, o peso de e2 deve ser maior que o peso de e1. Como e1 e e2 fazem parte do ciclo C, substituir e2 em B cria um com menos peso.Isso contradiz a crença de que B é um MST.

De forma mais geral, se os pesos das arestas não forem todos distintos, então apenas o (multi)conjunto de pesos nas árvores de abrangência mínima é certamente único, é o mesmo para todas as árvores de abrangência mínima.

Se os pesos forem positivos, então uma árvore de extensão mínima é, de fato, um subgrafo de custo mínimo conectando todos os vértices, uma vez que se um subgrafo contiver um ciclo, remover qualquer vantagem ao longo desse ciclo diminuirá seu custo e preservará a conectividade.

subgrafo

Subgrafo de custo mínimos os pesos forem positivos, então uma árvore geradora mínima é, de fato, um subgrafo de custo mínimo conectando todos os vértices, já que se um subgrafo contém um ciclo, remover qualquer aresta ao longo desse ciclo diminuirá seu custo e preservará a conectividade.

propriedade cíclica

Para qualquer ciclo C no gráfico, se o peso de uma aresta e de C for maior do que qualquer um dos pesos individuais de todas as outras arestas de C, então esta aresta não pode pertencer a um MST.

Assuma o contrário, ou seja, que e pertence a um MST T1. Então, deletar e quebrará T1 em duas subárvores com as duas extremidades de e em subárvores diferentes. O restante de C reconecta as subárvores, portanto, há uma aresta f de C com extremidades em subárvores diferentes, ou seja, ela reconecta as subárvores em uma árvore T2 com peso menor que o de T1, porque o peso de f é menor que o peso de e.

propriedade de corte

Para qualquer corte C do gráfico, se o peso de uma aresta e no conjunto de corte de C for estritamente menor que os pesos de todas as outras arestas do conjunto de corte de C, então essa aresta pertence a todos os 'MST's do gráfico.

Suponha que haja um MST T que não contém e. Adicionar e a T produzirá um ciclo, que cruza o corte uma vez em e e cruza de volta em outra aresta e'. Excluindo e', obtemos uma árvore de abrangência T/e' U e de peso estritamente menor que T. Isso contradiz a suposição de que T era um MST.

Por um argumento semelhante, se mais de uma aresta tiver peso mínimo em um corte, então cada uma dessas arestas estará contida em alguma árvore de abrangência mínima.

Aresta de custo mínimo

Se a aresta de custo mínimo e de um grafo for única, então essa aresta é incluída em qualquer MST.

Se e não fosse incluído no MST, remover qualquer uma das arestas (de maior custo) no ciclo formado após adicionar e ao MST produziria uma árvore de abrangência de menor peso.

Contração

Se T for uma árvore de arestas MST, então podemos contrair T em um único vértice, mantendo a invariante de que o MST do grafo contraído mais T fornece o MST para o grafo antes da contração.

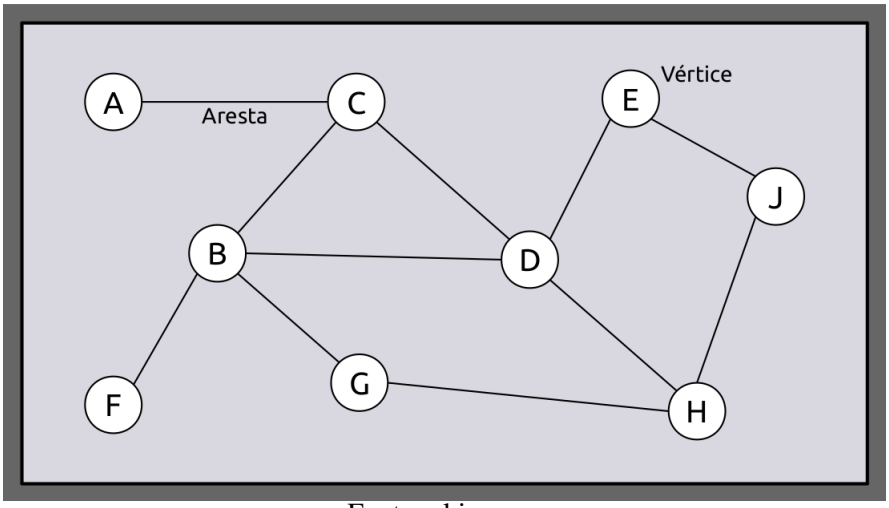
5.1.1 Algoritmo de Dijkstra

Vamos escolher um nó inicial, e deixar que a distância do nó N seja a distância do nó inicial até N. O algoritmo de Dijkstra (Dijkstra (2022)) começará inicialmente com distâncias infinitas e tentará melhorá-las passo a passo.

1 : Marque todos os nós como não visitados. Crie um conjunto de todos os nós não visitados chamado conjunto não visitado.

2 : Atribua a cada nó um valor de distância do início: para o nó inicial, é zero, e para todos os outros nós, é infinito, já que inicialmente nenhum caminho é conhecido para esses nós. Durante a execução do algoritmo, a distância de um nó N é o comprimento do caminho mais curto descoberto até agora entre o nó inicial e N.

Figura 10 – Rede MST conectada com o algoritmo de Dijkstra



Fonte: akira.org.

3 : Do conjunto não visitado, selecione o nó atual para ser aquele com a menor distância finita; inicialmente, este será o nó inicial, que tem distância zero. Se o conjunto não visitado estiver vazio ou contiver apenas nós com distância infinita (que são inalcançáveis), o algoritmo termina indo para a etapa 6. Se estivermos preocupados apenas com o caminho para um nó de destino, podemos terminar aqui se o nó atual for o nó de destino. Caso contrário, podemos continuar a encontrar os caminhos mais curtos para todos os nós alcançáveis.

4 : Para o nó atual, considere todos os seus vizinhos não visitados e atualize suas distâncias através do nó atual; compare a distância recém-calculada com a atualmente atribuída ao vizinho e atribua a ele a menor. Por exemplo, se o nó atual A for marcado com uma distância de 6, e a aresta que o conecta com seu vizinho B tiver comprimento 2, então a distância para B através de A será 6 + 2 = 8. Se B foi marcado anteriormente com uma distância maior que 8, atualize-o para 8 (o caminho para B através de A é mais curto). Caso contrário, mantenha sua distância atual (o caminho para B através de A não é o mais curto).

5 : Quando terminarmos de considerar todos os vizinhos não visitados do nó atual, marque o nó atual como visitado e remova-o do conjunto não visitado. Isso é para que um nó visitado nunca seja verificado novamente, o que é correto porque a distância registrada no nó atual é mínima (conforme garantido na etapa 3) e, portanto, final. Volte para a etapa 3.

6 : Assim que o “loop”sair (etapas 3–5), cada nó visitado conterá sua menor distância do nó inicial.

5.1.2 Algoritmo de Kruskal

O algoritmo(Kruskal (1956)) executa as seguintes etapas:

Cria uma floresta (um conjunto de árvores) consistindo inicialmente de uma árvore de vértice único separada para cada vértice no gráfico de entrada. Classifique as arestas do gráfico por peso. Faça um "loop"pelas arestas do gráfico, em ordem crescente classificada por seu peso. Para cada aresta: Teste se adicionar a aresta à floresta atual criaria um ciclo. Caso contrário, adicione a aresta à floresta, combinando duas árvores em uma única árvore.

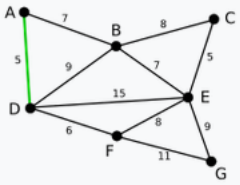
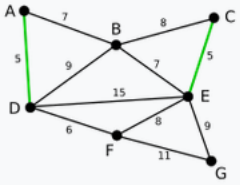
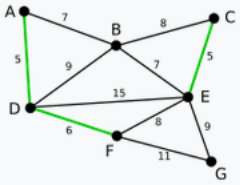
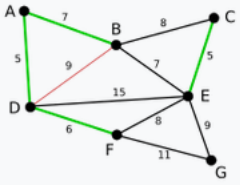
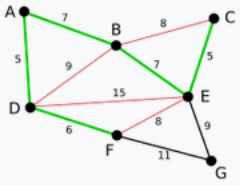
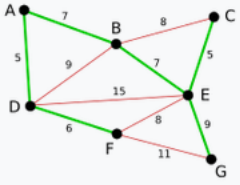
No término do algoritmo, a floresta forma uma floresta de abrangência mínima do gráfico. Se o gráfico estiver conectado, a floresta tem um único componente e forma uma árvore de abrangência mínima.

Para um gráfico com E arestas e V vértices, o algoritmo de Kruskal pode ser mostrado para rodar em tempo O(E log E), com estruturas de dados simples. Aqui, O expressa o tempo em notação O grande, e log é um logaritmo para qualquer base (já que dentro da notação O os logaritmos para todas as bases são equivalentes, porque são os mesmos até um fator constante). Este limite de tempo é frequentemente escrito como O(E log V), que é equivalente para gráficos sem vértices isolados, porque para esses gráficos $\frac{V}{2} <$ E < V2 e os logaritmos de V e E estão novamente dentro de um fator constante um do outro.

Para atingir este limite, primeiro classifique as arestas por peso usando uma classificação de comparação em tempo O(E log E). Uma vez classificadas, é possível percorrer as arestas em ordem classificada em tempo constante por aresta. Em seguida, use uma estrutura de dados de conjunto disjunto, com um conjunto de vértices para cada componente, para manter o controle de quais vértices estão em quais componentes. Criar essa estrutura, com um conjunto separado para cada vértice, leva V operações e tempo O(V). A iteração final por todas as arestas realiza duas operações de busca e possivelmente uma operação de união por aresta. Essas operações levam tempo amortizado $O(\alpha(V))$ tempo por operação, dando o pior caso de tempo total $O(E\alpha(V))$ para esse loop, onde α é a função Ackermann inversa de crescimento extremamente lento. Essa parte do limite de tempo é muito menor do que o tempo para a etapa de classificação, então o tempo total para o algoritmo pode ser simplificado para o tempo para a etapa de classificação.

Nos casos em que as arestas já estão classificadas, ou onde elas têm peso inteiro pequeno o suficiente para permitir que algoritmos de classificação de inteiros, como classificação por contagem ou classificação por radia, as classifiquem em tempo linear, as operações de conjunto disjunto são a parte restante mais lenta do algoritmo e o tempo total é $O(E\alpha(V))$.

Figura 11 – Exemplo de uso do Algoritmo Kruskal

	**AD** e **CE** são as arestas mais curtas, com comprimento 5, e **AD** foi arbitrariamente escolhido, por isso é destacado.
	**CE** é agora a aresta mais curta que não forma um ciclo, com comprimento 5, por isso é destacado como a segunda borda.
	A próxima borda, **DF** com comprimento 6, é realçada usando o mesmo método.
	As bordas mais curtas são **AB** e **BE**, ambas com comprimento 7. **A AB** é escolhida arbitrariamente e é destacada. A borda **BD** foi destacada em vermelho, porque já existe um caminho (em verde) entre **B** e **D**, por isso forma um ciclo (**ABD**) se fosse escolhido.
	O processo continua a destacar a próxima borda mais pequena, **BE** com comprimento 7. Muitas mais arestas são destacadas em vermelho nesta fase: **BC** porque formaria o ciclo **aC**, **DE** porque formaria o ciclo **DEBA**, e **FE** porque formaria **FE**.
	Finalmente, o processo termina com a borda **EG** de comprimento 9, e a árvore de abrangação mínima é encontrada.

Fonte: wikipedia.org.

5.1.3 Algoritmo de Prim

O algoritmo(Prim (1957)) pode ser descrito como realizando as seguintes etapas:

Inicializar uma árvore com um único vértice, escolhido arbitrariamente do gráfico.

Aumentar a árvore por uma aresta: Das arestas que conectam a árvore a vértices que ainda não estão na árvore, encontrar a aresta de peso mínimo e transferi-la para a árvore. Repita a etapa (até que todos os vértices estejam na árvore).

Associe a cada vértice v do gráfico um número C[v] (o custo mais barato de uma conexão com v) e uma aresta E[v] (a aresta que fornece essa conexão mais barata). Para inicializar esses valores, defina todos os valores de $C[v]$ como $+\infty$ (ou qualquer número maior que o peso máximo da aresta) e defina cada E[v] como um valor de sinalizador especial indicando que não há aresta conectando v a vértices anteriores.

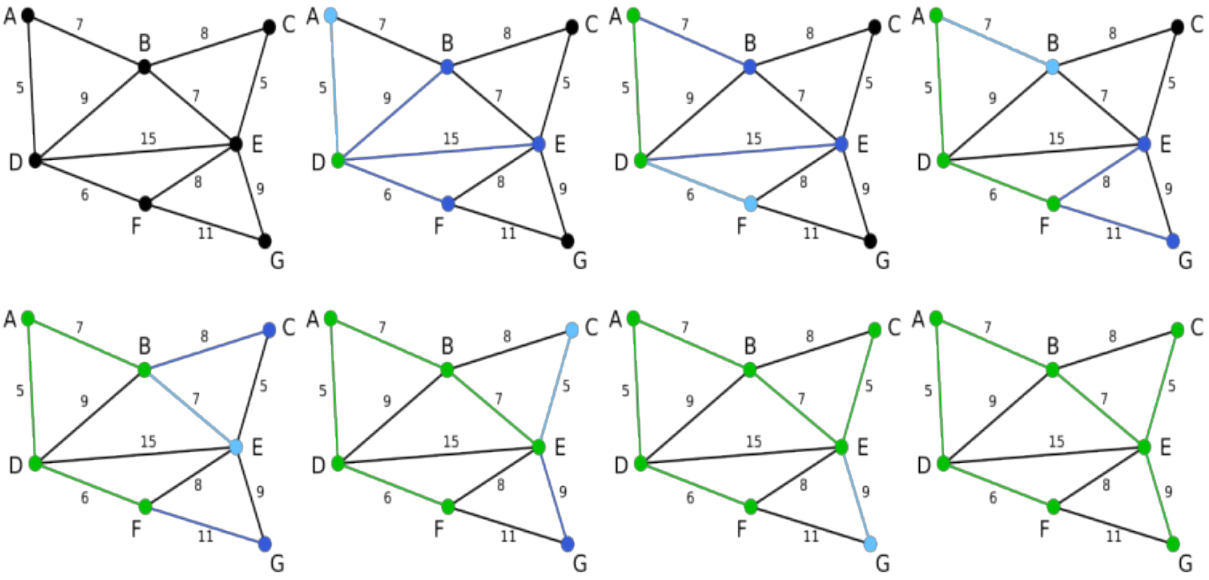
Inicializar uma floresta vazia F e um conjunto Q de vértices que ainda não foram incluídos em F (inicialmente, todos os vértices). Repita os seguintes passos até que Q esteja vazio: Encontre e remova um vértice v de Q tendo o valor mínimo possível de C[v]

Adicione v a F. Faça um loop sobre as arestas vw conectando v a outros vértices w. Para cada aresta, se w ainda pertencer a Q e vw tiver peso menor que C[w], execute os seguintes passos: Defina C[w] para o custo da aresta vw Defina E[w] para apontar para a aresta vw. Retorne F, que inclui especificamente as arestas correspondentes em E

Conforme descrito acima, o vértice inicial para o algoritmo será escolhido arbitrariamente, porque a primeira iteração do 'loop' principal do algoritmo terá um conjunto de vértices em Q que têm pesos iguais, e o algoritmo iniciará automaticamente uma nova árvore em F quando completar uma árvore de abrangência de cada componente conectado do gráfico de entrada. O algoritmo pode ser modificado para começar com qualquer vértice s em particular, definindo C[s] como um número menor que os outros valores de C (por exemplo, zero), e pode ser modificado para encontrar apenas uma única árvore de abrangência em vez de uma floresta de abrangência inteira (correspondendo mais de perto à descrição informal) parando sempre que encontrar outro vértice sinalizado como não tendo aresta associada.

Diferentes variações do algoritmo diferem umas das outras em como o conjunto Q é implementado: como uma lista encadeada simples ou matriz de vértices, ou como uma estrutura de dados de fila de prioridade mais complicada. Essa escolha leva a diferenças na complexidade de tempo do algoritmo. Em geral, uma fila de prioridade será mais rápida em encontrar o vértice v com custo mínimo, mas implicará em atualizações mais caras quando o valor de C[w] mudar.

Figura 12 – Exemplo de uso do Algoritmo de Prim



Fonte: Wikipedia.org.

6 DADOS FINANCEIROS

Para este trabalho, os dados foram obtidos do 'Yahoo Finance' com consulta no 'Investing.com'. Esse procedimento foi adotado devido à facilidade para baixar os dados. Os bancos de dados do 'Yahoo Finance' nos permitem acessar várias informações ao mesmo tempo com poucas linhas de código. O 'Yahoo Finance' coleta seus dados diretamente das fontes onde são gerados, como índices, ações e commodities retirados de lugares como a 'New York Stock Exchange (NYSE)', 'National Association of Securities Dealers Automated Quotations (NASDAQ)', Bovespa, entre outros.

Osprovedores de dados financeiros colaboram com fontes como 'Morningstar', 'ICE Data Services', 'Refinitiv (anteriormente Thomson Reuters)' e 'SP Global', que oferecem uma vasta gama de informações, incluindo preços, relatórios e análises de mercado. Informações financeiras e relatórios de desempenho são coletados diretamente dos relatórios das empresas. As notícias e análises são fornecidas por fontes como 'Reuters', 'Bloomberg', entre outras.

Com base nessas informações e após uma análise da veracidade de alguns dos dados baixados, comparando-os com outras fontes, foi possível verificar que esses canais são seguros para a obtenção de dados.

6.1 Ações de capital

A natureza deste trabalho é entender como o comércio de cada país se relaciona entre si e com o comércio de outros países. Dessa forma, não queremos apenas compreender a rede mundial, mas também cada parte da mesma. As ações selecionadas para este trabalho são, de acordo com a maior bolsa de cada país, as maiores ações de cada país. Cada um desses países tem um índice que representa suas maiores empresas. Um exemplo bem conhecido é o S&P 500, que se refere às maiores empresas americanas. As empresas incluídas nesses índices representam uma grande fração dos PIBs dos países. Dessa forma incluímos as ações que participam dos índices das bolsas dos países estudados, representando os países que queremos analisar em nossa matriz.

Ações de capital consistem em todas as ações pelas quais a propriedade de uma corporação ou empresa é dividida. Uma única parte da ação representa a propriedade fracionária da corporação, proporcional ao número total de ações. Isso normalmente dá ao acionista o direito a essa fração dos lucros da empresa, ao produto da liquidação de ativos (após o pagamento

de todos os créditos prioritários, como dívida garantida e não garantida) ou ao poder de voto, muitas vezes dividido em proporção ao valor investido por cada acionista. Nem todas as ações são necessariamente iguais, pois certas classes de ações podem ser emitidas, por exemplo, sem direito a voto, com direitos de voto reforçados ou com prioridade no recebimento de lucros ou na liquidação de ativos, antes ou depois de outras classes de acionistas.

Uma pessoa que possui uma percentagem das ações tem a propriedade da corporação proporcional à sua participação. As ações formam o capital social de uma empresa. O capital social de uma corporação é dividido em ações, cujo total é indicado no momento da criação do negócio. Ações adicionais podem ser posteriormente autorizadas pelos acionistas existentes e emitidas pela empresa. Em algumas jurisdições, cada ação tem um certo valor nominal, que é um valor contábil utilizado para representar o patrimônio líquido no balanço da corporação.

As ações representam uma fração de propriedade em uma empresa. Uma empresa pode emitir diferentes tipos (ou classes) de ações, cada uma com regras de propriedade, privilégios ou valores distintos. A propriedade de ações pode ser formalizada pela emissão de um certificado de ações. Um certificado de ações é um documento legal que especifica o número de ações possuídas pelo acionista e outras especificidades, como o valor nominal, se houver, ou a classe das ações.

Normalmente, cada índice tem um nome relacionado ao país que está negociando e ao número de ações incluídas naquele índice. Nos EUA, por exemplo, temos o , cujo nome tem uma explicação histórica ligada ao 'S&P', e o número 500 significa que o índice inclui 500 ações, ou seja, as 500 maiores empresas listadas naquele índice. Isso é uma boa forma de quantificar e verificar o andamento do mercado nacional. A tabela abaixo mostra as 10 maiores economias mundiais e seus respectivos índices.

6.2 Limpeza, indexação e montagem do banco de dados

A indexação de dados nada mais é do que a junção de todos os dados seguindo uma lógica de série temporal. Ou seja, queremos agrupar todas as ações em uma mesma série temporal, como elucidaremos com um exemplo visual abaixo, juntando o índice 'Amsterdam Exchange Index (AEX)' da Holanda e 'Swiss Market Index (SMI)' da Suíça.

Um problema que precisou ser resolvido foi a questão da moeda. Cada bolsa de valores cota de acordo com a moeda de seu país, mas ao unir todas as ações em uma matriz, teremos várias ações sendo cotadas em moedas diferentes. Para resolver esse problema, a solução

Figura 13 – Exemplo das 10 maiores economias e seus indicadores de maiores empresas internas.

País	Índice de Ações Representativo
Estados Unidos	S&P 500
China	CSI 300
Japão	Nikkei 225
Alemanha	DAX (Deutscher Aktienindex)+DMAX 100
Reino Unido	FTSE 100 (Financial Times Stock Exchange 100 Index) + FTSE250
França	CAC 40 (Cotation Assistée en Continu)+CACNEXT20
Índia	Nifty50 + NiftyNext50
Itália	FTSE MIB (Financial Times Stock Exchange Milano Indice di Borsa)50
Brasil	Ibovespa (Índice Bovespa)88
Canadá	S&P/TSX Composite Index (Standard & Poor's/Toronto Stock Exchange Composite Index) 250

Fonte: elaborado pelo autor (2024).

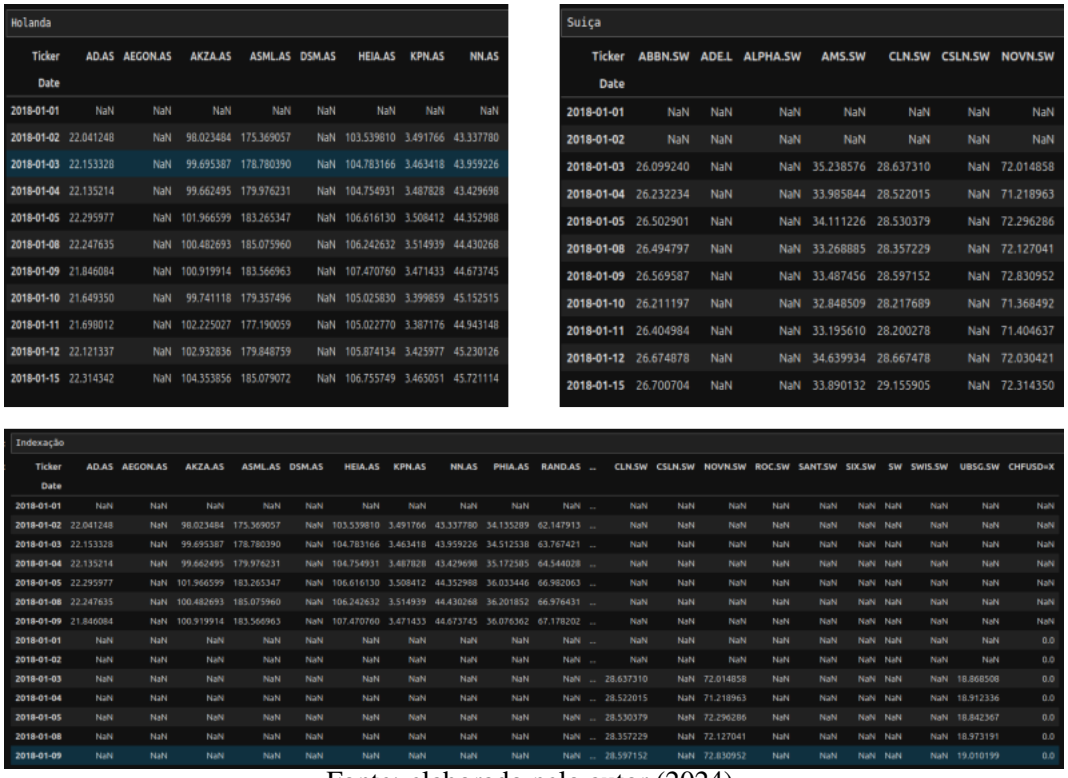
foi simples: como baixamos a taxa de câmbio, apenas multiplicamos o valor da moeda do país pela taxa de câmbio em relação à moeda que desejamos padronizar.

$$(\mathbf{Xi}Real) \cdot (\mathbf{Yi}\frac{Dolar}{Real}) = \mathbf{Zi}Dolar \tag{6.1}$$

Onde cada elemento da coluna dólar/real multiplica cada elemento da coluna da ação econômica cotada em real, no caso da B3, a bolsa de valores brasileira. Repetindo esse processo para todos os países abordados neste trabalho, teremos nossa matriz toda em dólar, que, no contexto global, é a moeda mais negociada em transações internacionais. Além disso, todas as nossas variáveis macroeconômicas já foram baixadas em dólar. Posteriormente, resolvido o problema da moeda utilizada, resta apenas lidar com os dados faltantes.

Agora, precisamos remover a maioria dos 'NaN', ou seja, dados faltantes. Foi

Figura 14 – Holanda e Suíça sendo indexadas.



Holanda

Ticker	AD.AS	AEGON.AS	AKZA.AS	ASML.AS	DSM.AS	HEIA.AS	KPN.AS	NN.AS
Date								
2018-01-01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-02	22.041248	NaN	98.023484	175.369057	NaN	103.539810	3.491766	43.337780
2018-01-03	22.153328	NaN	99.695387	178.780390	NaN	104.783166	3.463418	43.959226
2018-01-04	22.135214	NaN	99.662495	179.976231	NaN	104.754931	3.487828	43.429698
2018-01-05	22.295977	NaN	101.966599	183.265347	NaN	106.616130	3.508412	44.352988
2018-01-08	22.247635	NaN	100.482693	185.075960	NaN	106.242632	3.514939	44.430268
2018-01-09	21.846084	NaN	100.919914	183.566963	NaN	107.470760	3.471433	44.673745
2018-01-10	21.649350	NaN	99.741118	179.357496	NaN	105.025830	3.399859	45.152515
2018-01-11	21.698012	NaN	102.225027	177.190059	NaN	105.022770	3.387176	44.943148
2018-01-12	22.121337	NaN	102.932836	179.848759	NaN	105.874134	3.425977	45.230126
2018-01-15	22.314342	NaN	104.353856	185.079072	NaN	106.755749	3.465051	45.721114

Suíça

Ticker	ABBN.SW	ADE.L	ALPHA.SW	AMS.SW	CLN.SW	CSLN.SW	NOVN.SW
Date							
2018-01-01	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-02	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-03	26.099240	NaN	NaN	35.238576	28.637310	NaN	72.014858
2018-01-04	26.232234	NaN	NaN	33.985844	28.522015	NaN	71.218963
2018-01-05	26.502901	NaN	NaN	34.111226	28.530379	NaN	72.296286
2018-01-08	26.494797	NaN	NaN	33.268885	28.357229	NaN	72.127041
2018-01-09	26.569587	NaN	NaN	33.487456	28.597152	NaN	72.830952
2018-01-10	26.211197	NaN	NaN	32.848509	28.217689	NaN	71.368492
2018-01-11	26.404984	NaN	NaN	33.195610	28.200278	NaN	71.404637
2018-01-12	26.674878	NaN	NaN	34.639934	28.667478	NaN	72.030421
2018-01-15	26.700704	NaN	NaN	33.890132	29.155905	NaN	72.314350

Indexação

Ticker	AD.AS	AEGON.AS	AKZA.AS	ASML.AS	DSM.AS	HEIA.AS	KPN.AS	NN.AS	PHIA.AS	RAND.AS	...	CLN.SW	CSLN.SW	NOVN.SW	ROC.SW	SANT.SW	SIX.SW	SW	SWIS.SW	UBSG.SW	CHFUSD=X
Date																					
2018-01-01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-02	22.041248	NaN	98.023484	175.369057	NaN	103.539810	3.491766	43.337780	34.135289	62.147913	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-03	22.153328	NaN	99.695387	178.780390	NaN	104.783166	3.463418	43.959226	34.512538	63.767421	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-04	22.135214	NaN	99.662495	179.976231	NaN	104.754931	3.487828	43.429698	35.172585	64.544028	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-05	22.295977	NaN	101.966599	183.265347	NaN	106.616130	3.508412	44.352988	36.033446	66.982063	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-08	22.247635	NaN	100.482693	185.075960	NaN	106.242632	3.514939	44.430268	36.201852	66.976431	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-09	21.846084	NaN	100.919914	183.566963	NaN	107.470760	3.471433	44.673745	36.076362	67.178202	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2018-01-01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0
2018-01-02	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0
2018-01-03	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	28.637310	NaN	72.014858	NaN	NaN	NaN	NaN	NaN	18.868508	0.0
2018-01-04	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	28.522015	NaN	71.218963	NaN	NaN	NaN	NaN	NaN	18.912336	0.0
2018-01-05	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	28.530379	NaN	72.296286	NaN	NaN	NaN	NaN	NaN	18.842367	0.0
2018-01-08	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	28.357229	NaN	72.127041	NaN	NaN	NaN	NaN	NaN	18.973191	0.0
2018-01-09	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	28.597152	NaN	72.830952	NaN	NaN	NaN	NaN	NaN	19.010199	0.0

Fonte: elaborado pelo autor (2024).

removida qualquer linha ou coluna que contivesse mais de 50% de dados faltantes. Além disso, eliminamos aqueles dias em que apenas algumas ações estavam sem dados ou em que certas ações não foram cotadas. Posteriormente, a matriz com os dados restantes foi interpolada linearmente para preencher os dados faltantes remanescentes, concluindo assim o processamento para aquele período. Nosso banco de dados foi montado semestralmente, de 2018 a 2024, utilizando esses métodos para cada semestre, a fim de manter a verossimilhança dos dados e evitar influências de outros semestres.

Figura 15 – Exemplo de limpeza de banco de dados.

Ticker	AD.AS	AKZA.AS	ASML.AS	HEIA.AS	KPN.AS	NN.AS	PHIA.AS	RAND.AS	VPK.AS
Date									
2018-01-02	22.041248	98.023484	175.369057	103.539810	3.491766	43.337780	34.135289	62.147913	43.241685
2018-01-03	22.153328	99.695387	178.780390	104.783166	3.463418	43.959226	34.512538	63.767421	44.055736
2018-01-04	22.135214	99.662495	179.976231	104.754931	3.487828	43.429698	35.172585	64.544028	44.150321
2018-01-05	22.295977	101.966599	183.265347	106.616130	3.508412	44.352988	36.033446	66.982063	44.304712
2018-01-08	22.247635	100.482693	185.075960	106.242632	3.514939	44.430268	36.201852	66.976431	44.045069
2018-01-09	21.846084	100.919914	183.566963	107.470760	3.471433	44.673745	36.076362	67.178202	44.111132
2018-01-10	21.649350	99.741118	179.357496	105.025830	3.399859	45.152515	35.871247	66.422514	44.806568
2018-01-11	21.698012	102.225027	177.190059	105.022770	3.387176	44.943148	35.973569	67.241359	44.620331
2018-01-12	22.121337	102.932836	179.848759	105.874134	3.425977	45.230126	36.510774	67.508010	44.796610
2018-01-15	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-03	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-04	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-05	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-08	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-09	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-10	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-11	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-12	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039
2018-01-15	22.314342	104.353856	185.079072	106.755749	3.465051	45.721114	36.501596	69.057169	45.209039

Fonte: elaborado pelo autor (2024).

6.3 Retorno e Log-retorno

Suponha que um investidor emprestou S_0, a ser paga com taxa R_0 no período 1, R_1 no período 2 e R_n no período n com juros compostos. No momento 1 ele teria:

$$S_1 = S_0 + R_1 S_0 = (1 + R_1) S_0, \tag{6.2}$$

no período 2:

$$S_2 = S_1 + R_2 S_1 = (1 + R_2) S_1, \tag{6.3}$$

no período n:

$$S_n = (1 + R_1)(1 + R_2) \cdots (1 + R_n) S_0. \tag{6.4}$$

Essa é uma operação multiplicativa. Existem bons argumentos para extrair o logaritmo desses valores e definir o log-retorno através de:

$$r = \ln(1 + R). \tag{6.5}$$

O primeiro, para transformar operação de multiplicação em adição:

$$\ln[(1 + R_1)(1 + R_2) \cdots (1 + R_n)] = \ln(1 + R_1) + \ln(1 + R_2) + \cdots + \ln(1 + R_n). \tag{6.6}$$

O segundo, porque no caso contínuo $S_n = e^{rt} S_0$. O terceiro vem por conta do Teorema Central do Limite que afirma que se uma variável tem esperança e variâncias finitas então a adição de número grande dessas variáveis independentes deve seguir a distribuição normal. Se o preço de uma ação no período anterior foi P_{t-1} e no presente P_t, o retorno do investidor foi:

$$R = \frac{P_t - P_{t-1}}{P_{t-1}}, \tag{6.7}$$

e o ganho relativo:

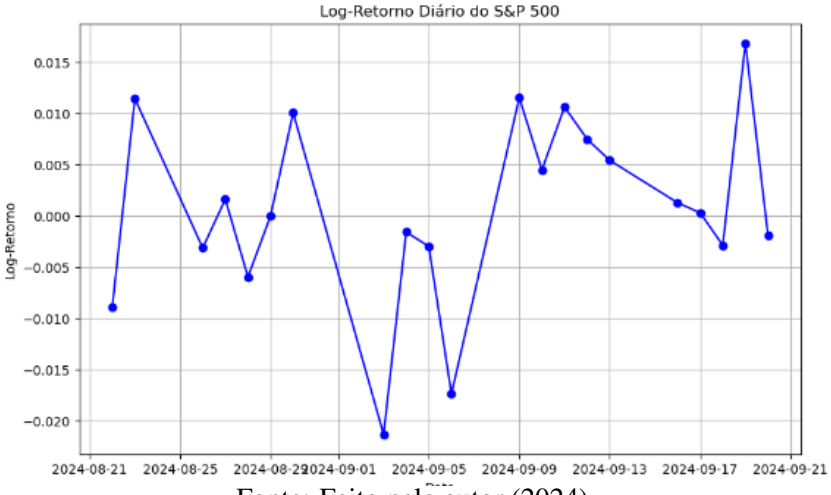
$$1 + R = \frac{P_t}{P_{t-1}}. \tag{6.8}$$

Então o log-retorno é dado por:

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right). \tag{6.9}$$

Como mostrado na 16, aplicada ao índice S&P 500 para um período recente à produção deste trabalho. Dessa forma, com a variação dos dados, podemos aplicar nosso arsenal estatístico.

Figura 16 – Exemplo de aplicação do Log-retorno na S&P 500.



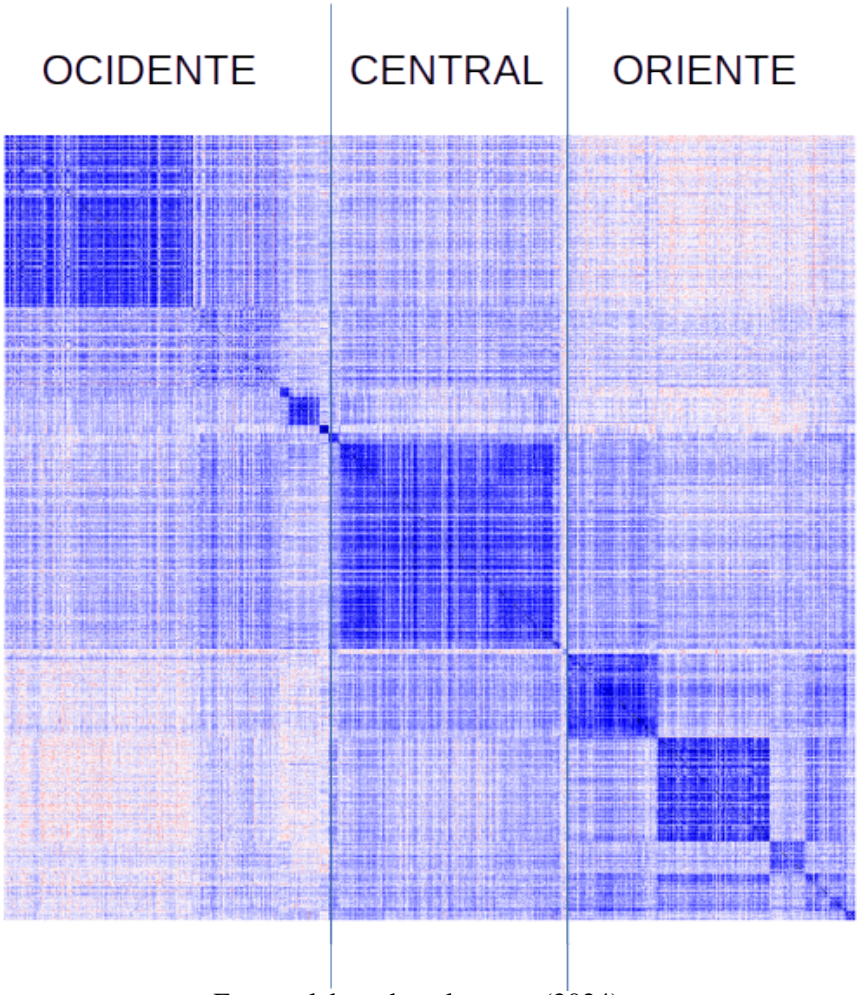
Fonte: Feito pelo autor (2024)

7 ANÁLISE DOS DADOS

7.1 matriz de correlação organizada pela geografia mostrados como mapas de calores

A primeira organização dos índices da matriz foi feita com base na localização geográfica dos países, iniciando no ocidente e indo em direção ao oriente. As ações de cada país, nas figuras que serão apresentadas, estão organizadas de forma alfabética. Assim, começamos pela América do Norte, com os 'Estados Unidos da América (EUA)', e seguimos até o Chile. Posteriormente, no meio da matriz, ao longo da diagonal principal, encontramos a Europa e a África. Por fim, no final da matriz, estão os países do oriente, como os da Ásia e Oceania.

Figura 17 – Mapa de calor da correlação das maiores ações globais organizadas pela posição geográfica no período de 2018.1



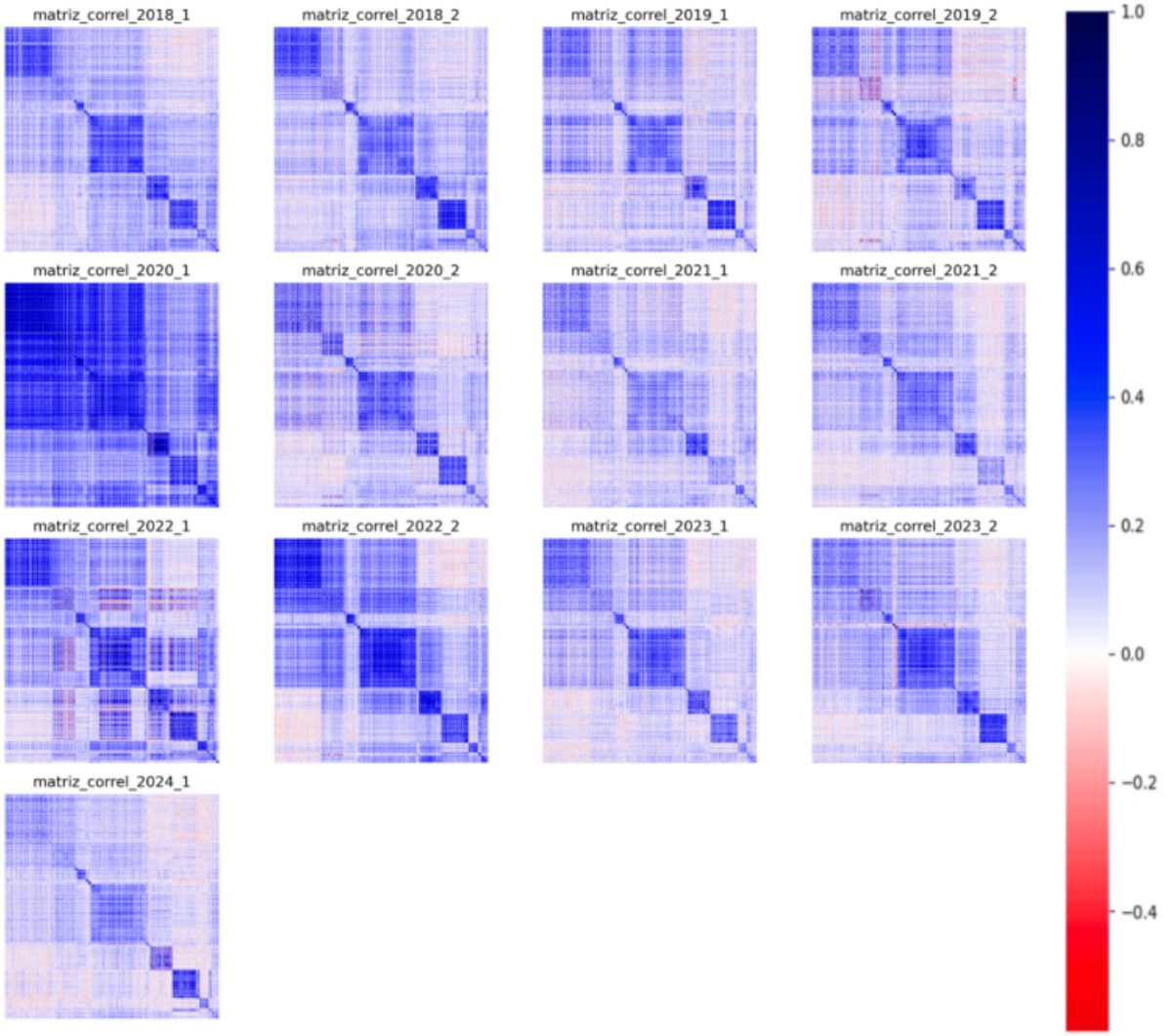
Fonte: elaborado pelo autor (2024).

Usamos afigura acima como modelo. A Figura 17oi obtida a partir da correlação do primeiro semestre de 2018, utilizado como parâmetro de normalidade econômica. Podemos

observar na diagonal principal que cada país está relacionado consigo mesmo.

O primeiro agrupamento (’cluster’) é o SP 500 dos ’EUA’. Já o segundo é o do Canadá, que se mostra como o menos correlacionado consigo mesmo. Contudo, após esse caso excepcional do Canadá, observamos que nenhum outro país apresenta uma baixa correlação consigo. Observando o período de 2018 a 2024, temos:

Figura 18 – Mapa de calor da correlação das maiores ações globais organizadas pela posição geográfica no período de 2018 a 2024



Fonte: elaborado pelo autor (2024).

Destacamos os semestres de 2020.1 e 2022.1. Na análise do semestre 2020.1, podemos ver que a parte Oriental e Central do globo está altamente correlacionada, e todo o mapa parece mais azul que os demais, indicando que as ações estão se correlacionando mais intensamente do que nos outros semestres, principalmente as do ocidente. Esse período coincide com o início do ’lockdown’ durante a pandemia global de ’Covid-19’.

Além disso, observamos que, de 2018 a 2019, as áreas fora dos agrupamentos estão ficando mais vermelhas, sugerindo que o mundo estava se descorrelacionando, exceto pelo primeiro agrupamento (’EUA’) e pela Europa, além da própria Europa e dos países orientais.

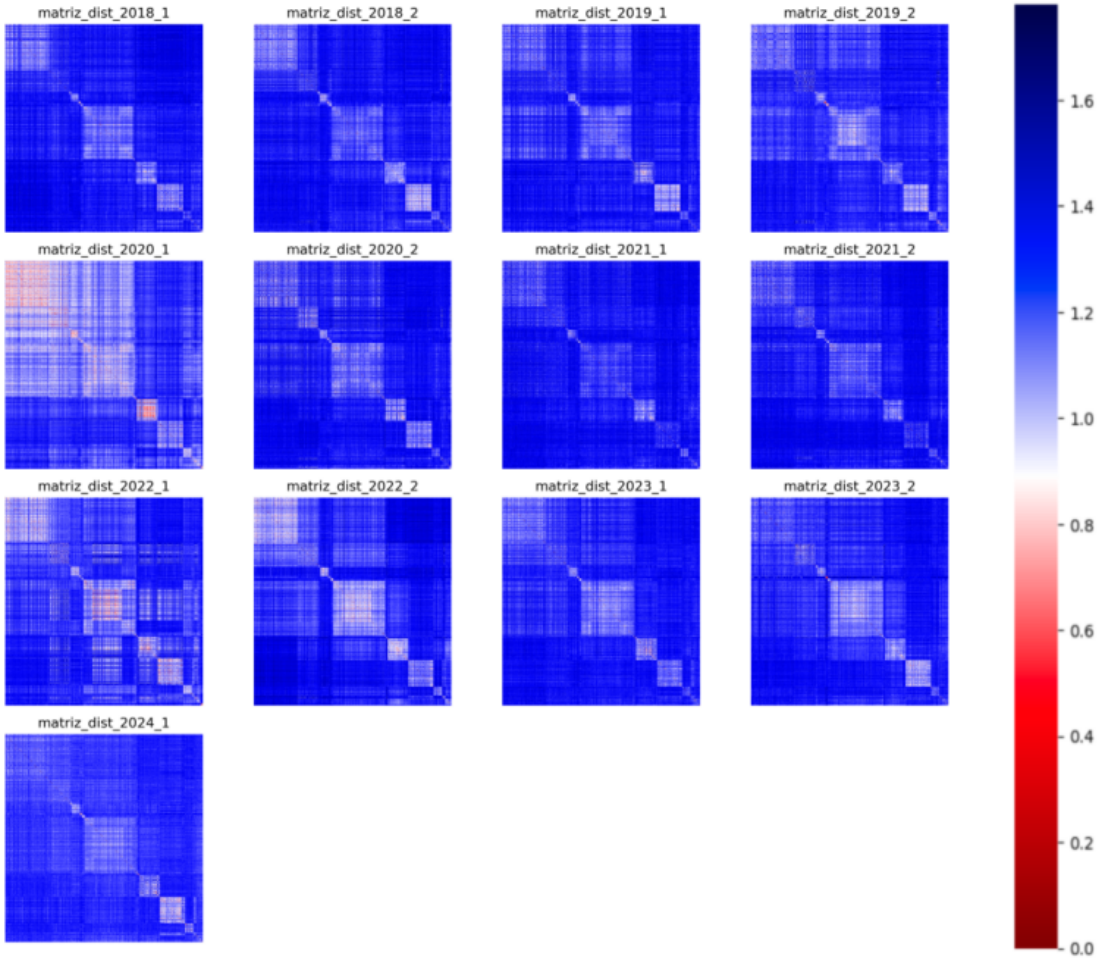
Já no período de 2020.2, vemos o retorno ao padrão observado nos períodos anteriores, mesmo ainda estando no período da pandemia. No entanto, pode-se notar a flexibilidade do mercado em relação ao momento conturbado.

Em2022.1, com o início da guerra na Ucrânia, nota-se que, fora da diagonal principal, as correlações se tornam muito mais fracas. Observa-se também a formação de vários subagrupamentos que se correlacionam e descorrelacionam de maneira mais intensa e volátil, evidenciando um período de grande instabilidade econômica.

Outra forma de analisar as correlações e identificar os agrupamentos ocultos, ou agrupamentos dentro de agrupamentos, é através da matriz de distância. Nessa matriz, os valores variam de 0 a 2, e, ao colocar a cor branca no valor 1, podemos ver Figura 19 que, quanto mais avermelhado, mais próxima é a correlação, ou seja, mais correlacionado. Também é possível identificar momentos de maior instabilidade econômica, já que a maior parte da matriz apresenta coloração azul com linhas brancas. Nos períodos de correlação muito alta, a distância diminui, resultando em áreas vermelhas de correlação intensa.

No mapa abaixo, podemos ver que os períodos de menor distância de correlação coincidem exatamente com os momentos já mencionados: o ’lockdown’ e a guerra na Ucrânia. No mapa de calor de 2020.1, o primeiro agrupamento (’EUA’) está bem próximo, embora os demais também estejam relativamente próximos, embora menores do que os ’EUA’. Observamos ainda que a China está bem correlacionada consigo mesma, algo que pode ser facilmente visto na parte oriental.

Figura 19 – Mapa de calor da distancia de correlação das maiores ações globais organizadas pela posição geográfica no período de 2018 a 2024



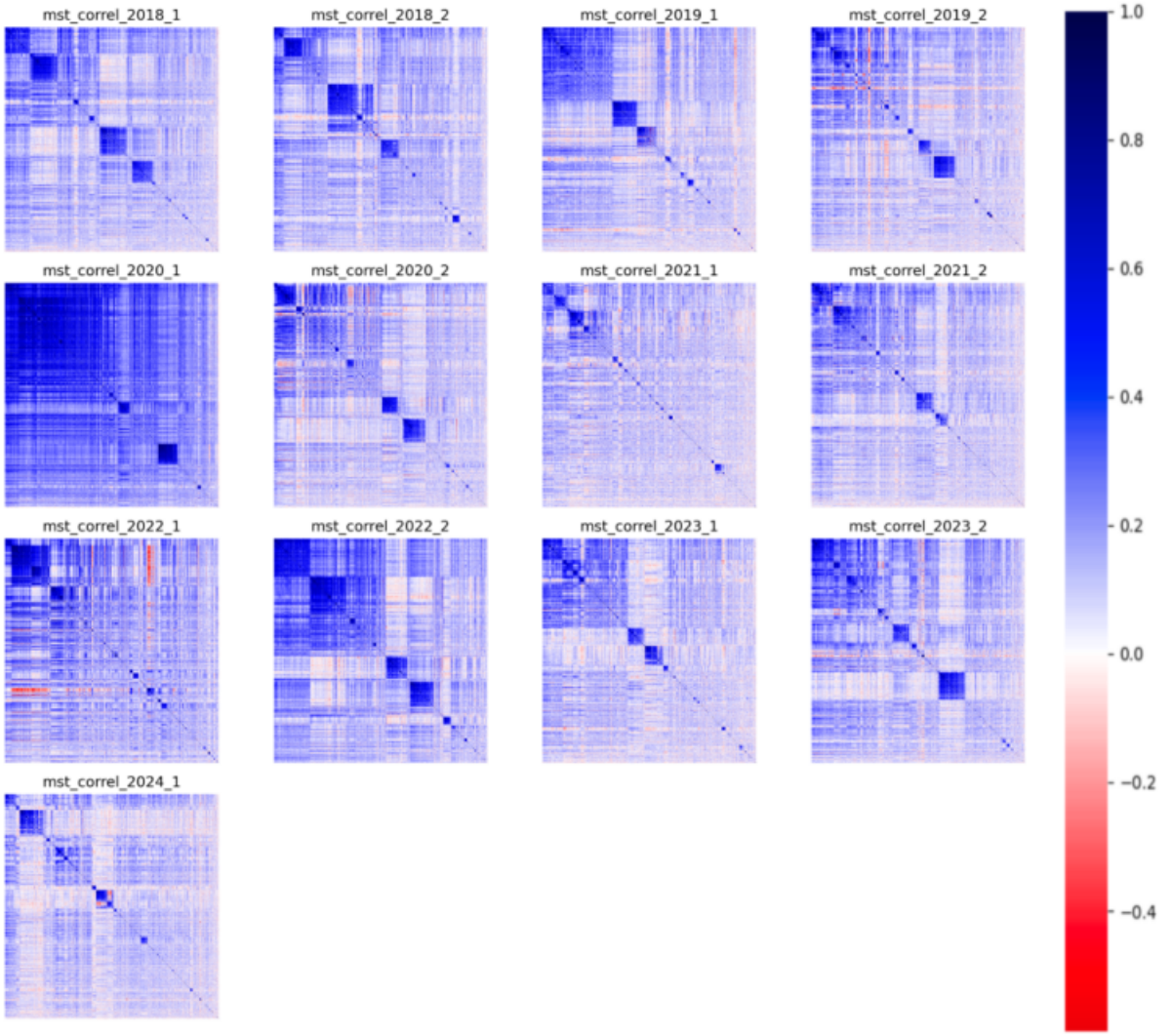
Fonte: elaborado pelo autor (2024).

7.2 Matriz de correlação organizada pela "MST"Global mostrados como mapas de calores

pós organizar os índices geograficamente, finalmente podemos organizá-los pela árvore mínima de expansão ("MST"). Na figura 20 ,percebemos que não faz mais sentido pensar em cada país individualmente, mas sim na economia global como um todo. Nesta nova configuração, torna-se impossível localizar e entender cada país ou um conjunto de dados de forma lógica. Observamos que as ações também formam agrupamentos bem definidos, e é fácil identificar esses padrões. Nos semestres de instabilidade, há uma diferença notável em relação aos demais, como no semestre 2020.1, em que o mundo continua bem correlacionado, e em 2022.1, quando ocorre uma descorrelação mais acentuada.

No mapa de calor abaixo, figura 21 que mostra a distância de correlação, observamos

Figura 20 – Mapa de calor da correlação das maiores ações globais organizadas pela 'minimum spanning tree' no período de 2018 a 2024

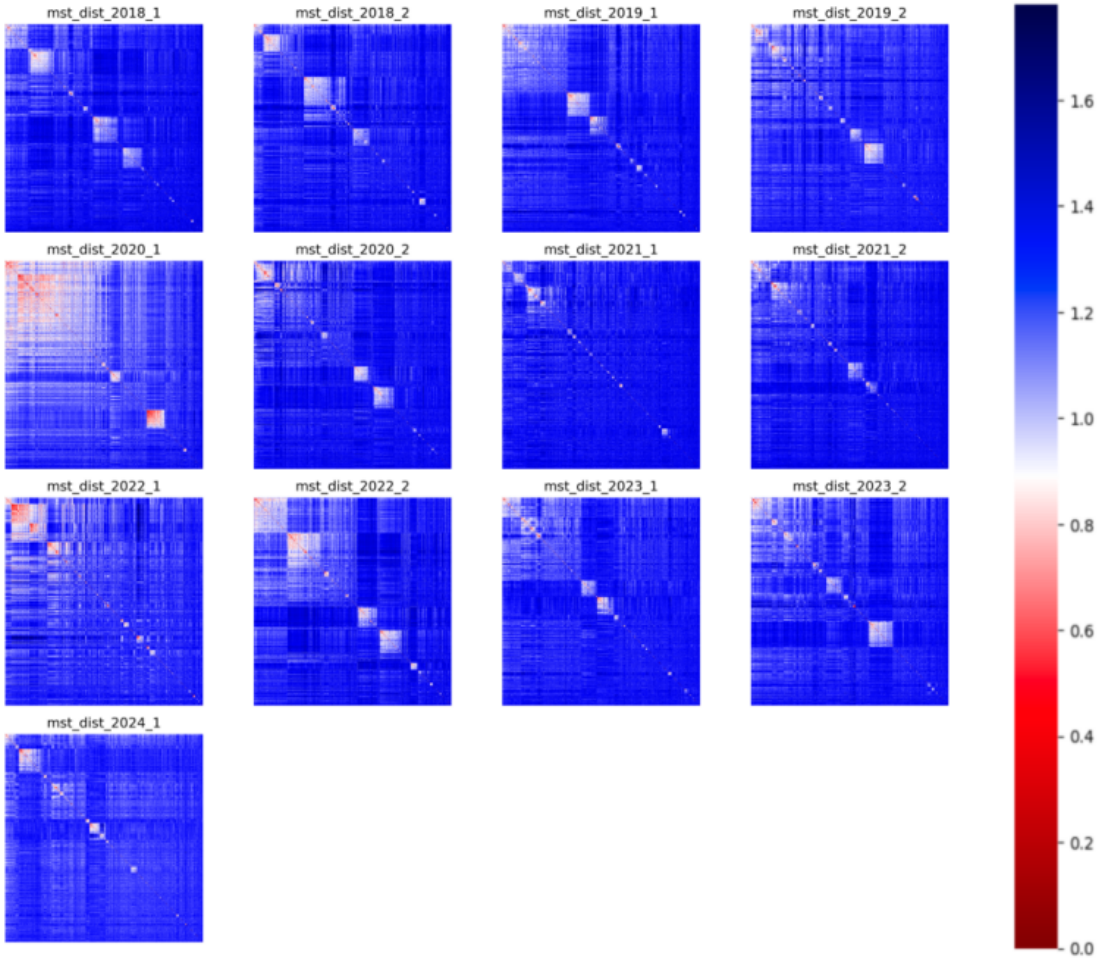


Fonte: elaborado pelo autor (2024).

os "clusters" ocultos que revelam os agrupamentos mais correlacionados dentro da matriz de correlação organizada pela "MST". Notamos que existem pontos muito pequenos de proximidade, indicando uma supercorrelação entre certas ações. Nos semestres de 2020.1 e 2022.1, a intensidade do vermelho é significativamente maior, mostrando que várias ações se aproximaram mais, ou seja, estão bem mais correlacionadas.

O interessante aqui é que não há fronteiras definidas, e a localização dessas correlações é determinada pela proximidade econômica. O código de Prim organiza todas elas de acordo com o seu valor de fechamento diário. Assim podemos identificar os setores e países aos quais essas ações pertencem, evidenciando quais são essas proximidades.

Figura 21 – Mapa de calor da distancia de correlação das maiores ações globais organizadas pela 'minimum spanning tree' no período de 2018 a 2024



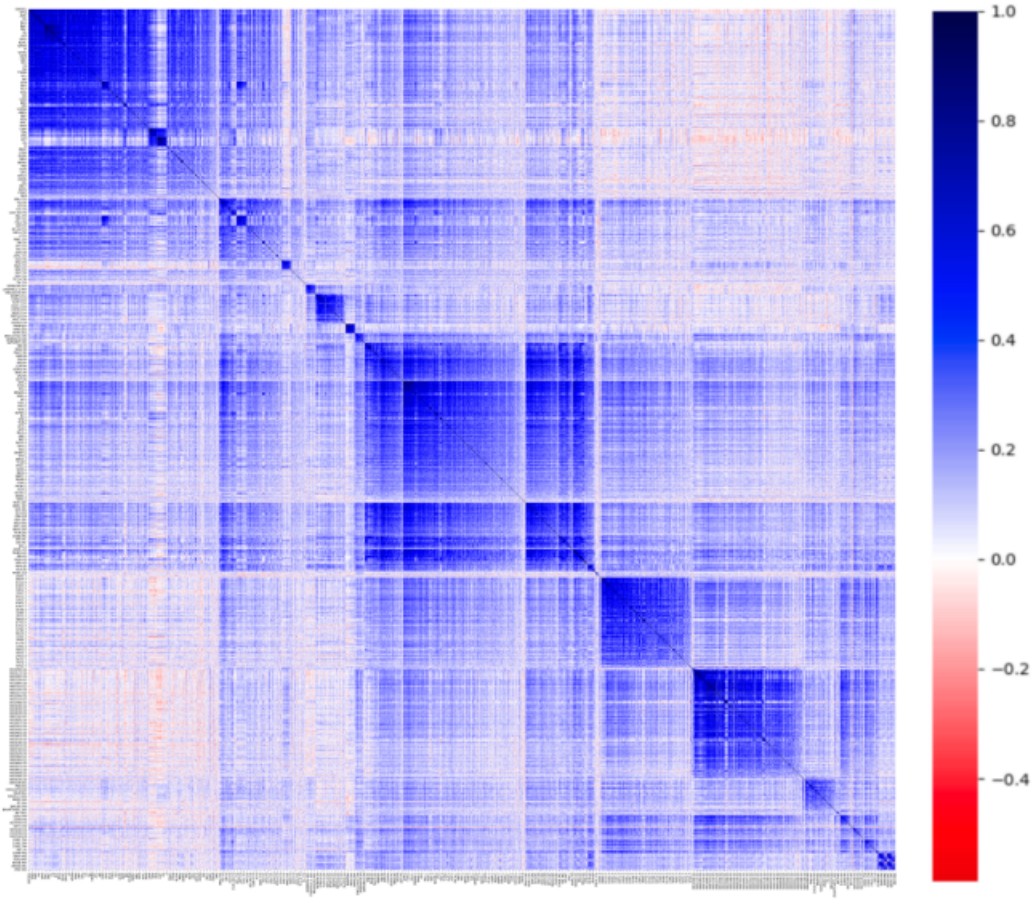
Fonte: elaborado pelo autor (2024).

7.3 Matriz organizada pela 'MST' de cada pais

Nesta seção, vamos organizar as figuras pela 'MST' de cada país, começando pela correlação, figura 22 e focando no semestre de normalidade de 2018.1. Agora, em vez de organizar toda a matriz pela 'MST', vamos apenas organizar pequenas frações dela, ou seja, as ações de cada país. À primeira vista, percebe-se a formação de gradientes em torno de cada agrupamento, e nota-se também um vermelho nas pontas, indicando uma descorrelação dos 'EUA' com o Oriente e uma correlação maior com o Ocidente. Ademais, observamos 'clusters' se formando na região do Canadá, onde o que anteriormente parecia quase impossível agora revela que essas ações se correlacionam entre si e com uma certa proximidade das correlações americanas.

Vemos no mapa abaixo, figura 23, que, logicamente, os agrupamentos do mapa anterior se tornaram agrupamentos mais claros no mapa de distância.

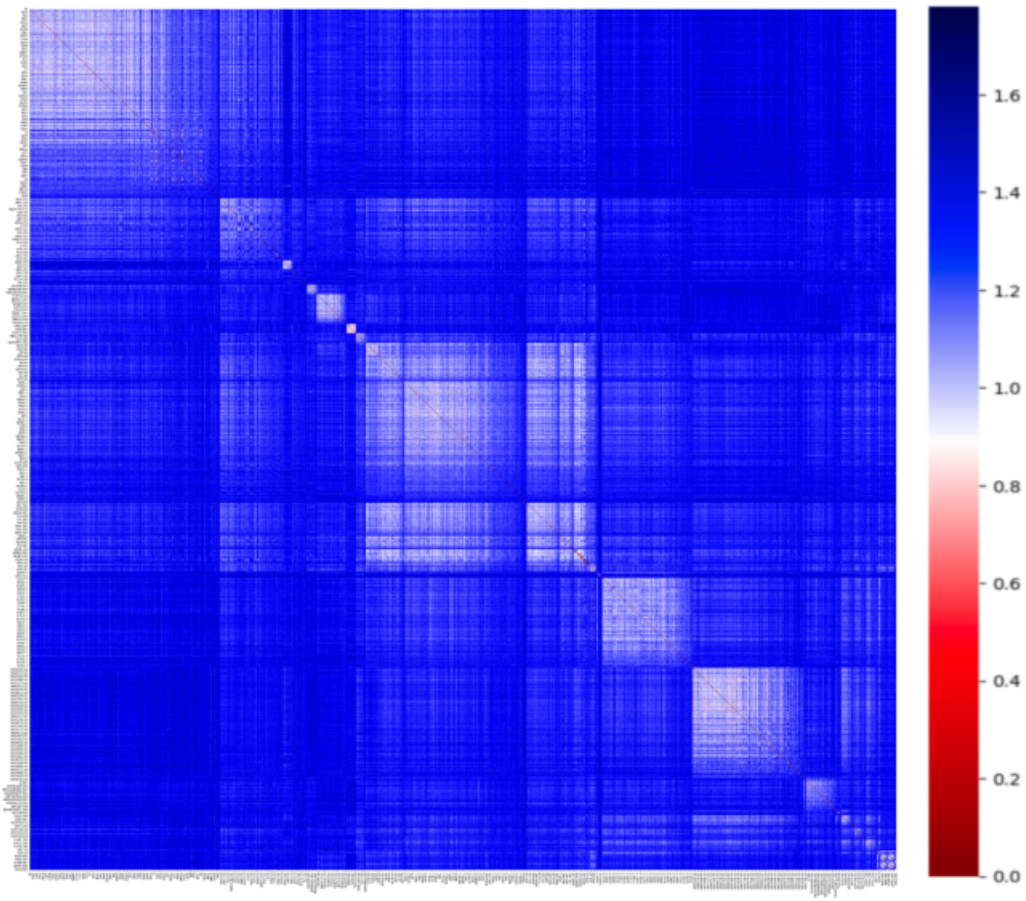
Figura 22 – Mapa de calor da correlação onde as ações de cada país foi organizado pela 'MST' no período do primeiro semestre de 2018



Fonte: elaborado pelo autor (2024).

Para o período de instabilidade global utilizamos o semestre de 2020.1. Observa-se que há linhas que repartem grandes agrupamentos, formando assim uma correlação muito alta entre os países do Ocidente. Contudo, é possível notar países que estão altamente correlacionados entre si, mas não com outros, assim como países que possuem correlações tanto internas quanto externas. No que diz respeito à parte ocidental, o Oriente apresenta grandes agrupamentos de vários países organizados em blocos contínuos ao longo da diagonal principal, em contraste com o Ocidente, onde os agrupamentos se mostram quase indistinguíveis.

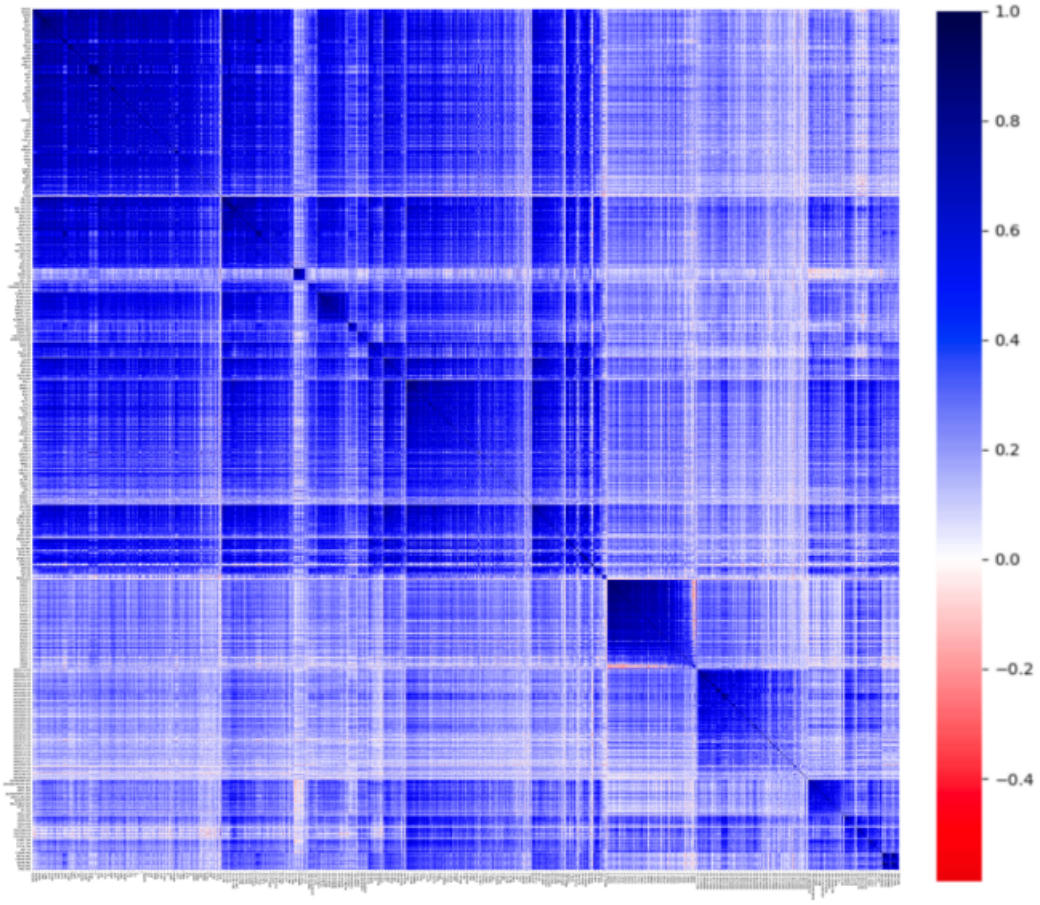
Figura 23 – Mapa de calor da distancia de correlação onde as ações de cada país foram organizadas pela 'MST' no período do primeiro semestre de 2018



Fonte: elaborado pelo autor (2024).

Observando a matriz de distância na figura 25, notamos que a SP 500 continua bem correlacionada consigo mesma. Entretanto, é evidente que os países da América do Sul também apresentam uma forte correlação entre si; por exemplo, Brasil, Chile e Argentina estão agrupados em proximidade na diagonal da tabela, logo após o Canadá. Em seguida, temos a Inglaterra, que também se mostra bastante correlacionada, seguida pelo restante da Europa. Após a Europa, a China, representada pela sua China Securities Index 300 (CSI 300), apresenta uma proximidade acentuada. Por último, e não menos importante, no final do mapa de calor, encontramos a Rússia com o índice Moscow Exchange (ME), que também demonstra uma forte correlação consigo mesma.

Figura 24 – Mapa de calor da correlação onde as ações de cada país foram organizadas pela ’MST’ no período do primeiro semestre de 2020

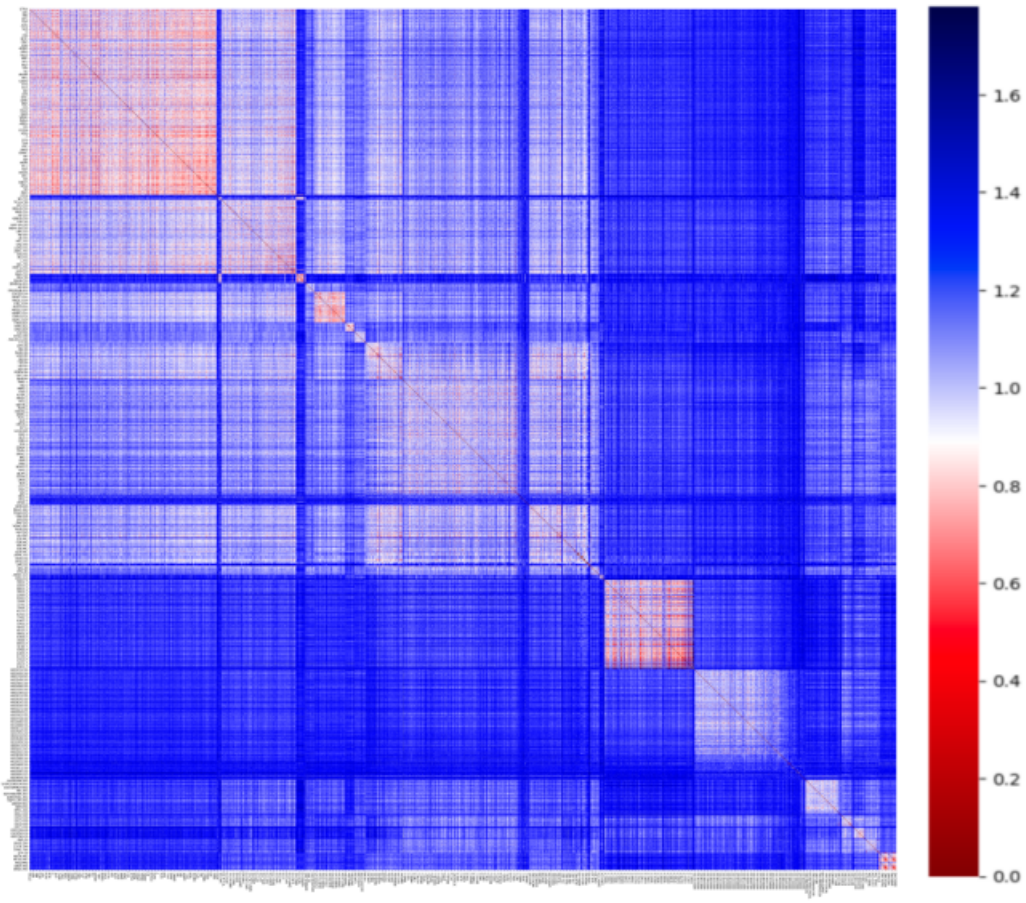


Fonte: elaborado pelo autor (2024).

7.4 Analisando a ’MST’ Global pelos setores

Como a nossa ’Minimum Spanning Tree (MST)’ global tornava quase impossível a análise de partes individuais, mas facilitava a visão do todo, vamos agora organizá-la de uma forma que permita extrair outras informações. No nosso caso, primeiramente abordaremos os setores. As ações são divididas em 11 setores, e, ao separar apenas essa informação, poderemos realizar uma análise minuciosa da situação. Para analisar outros fatores na matriz ordenada pela MST usamos a seguinte estratégia – sem mudar a ordem da matriz, se ação x possuir o mesmo fator que a ação y então a célula xy é pintada na cor daquele fator, já se tiverem fatores diferentes a célula xy é pintada de branco. Obviamente que a diagonal não deve ter cores brancas.

Figura 25 – Mapa de calor da distancia de correlação onde as ações de cada país foram organizadas pela 'MST' no período do primeiro semestre de 2020

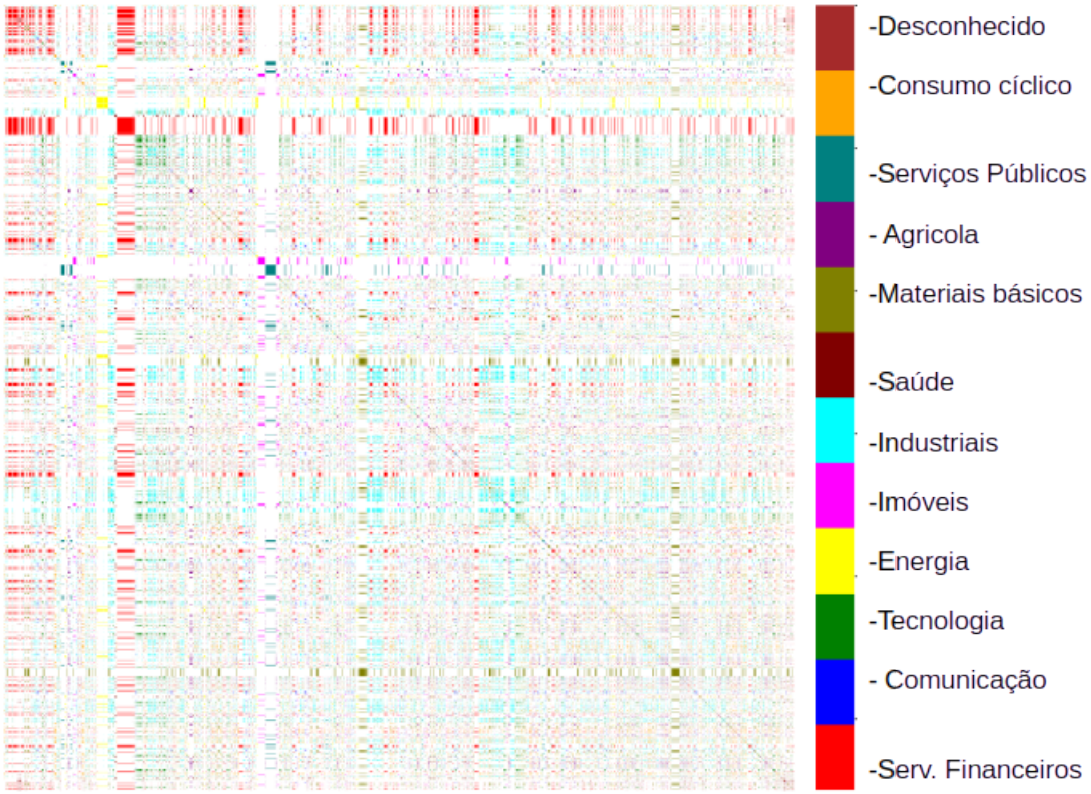


Fonte: elaborado pelo autor (2024).

Observa-se, pela matriz na figura 26, que há uma parte em branco e outras coloridas. É importante notar que as colunas e linhas foram substituídas pelos setores correspondentes às ações. Embora tenhamos mais de 2000 ações, existem apenas 11 setores, portanto, ao trocar os setores nas colunas e linhas, podemos montar um mapa de calor que colore apenas os lugares onde o setor for igual nas linhas e colunas (i=j).

Observamos que o setor financeiro está muito bem agrupado e domina quase todo o mapa de calor, seguido pelo setor industrial. Além disso, nota-se a formação de um quadrado na diagonal com o setor de materiais básicos, assim como outros pequenos agrupamentos dos demais setores. Algo que pode ser difícil de perceber é que alguns agrupamentos maiores contêm vários agrupamentos menores, criando, assim, um grupo de agrupamentos locais com setores

Figura 26 – Mapa de calor da 'MST' Global com apenas os setores como indicadores do primeiro semestre 2018

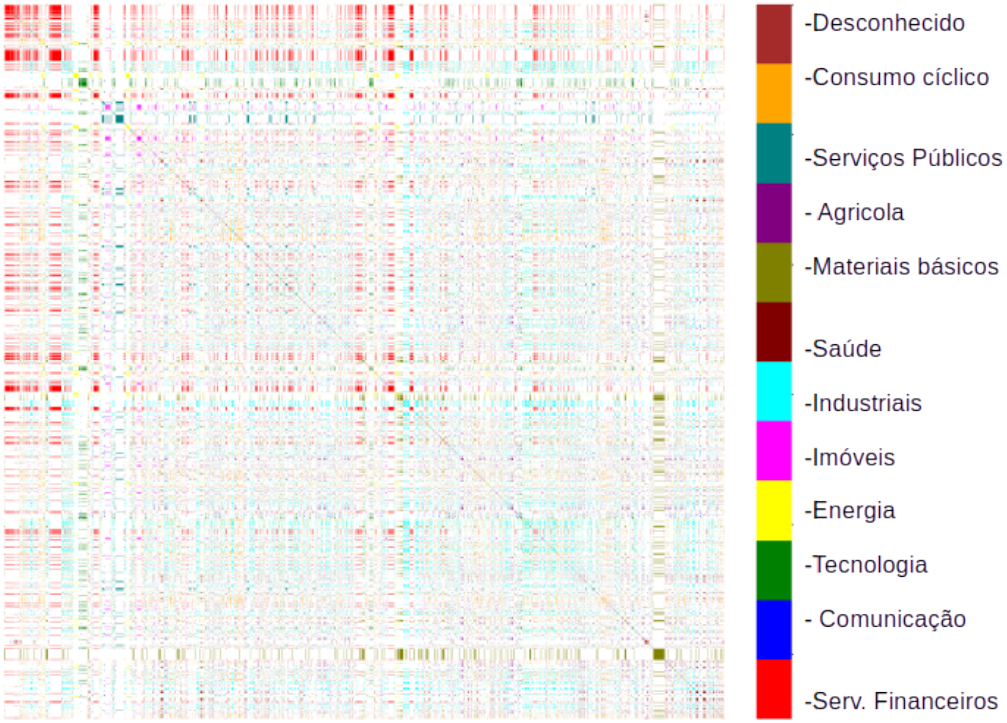


Fonte: elaborado pelo autor (2024).

extremamente correlacionados pela MST global.

Já para o período de instabilidade, como mostrado na figura 27, observamos que o ramo financeiro se torna maior do que em 2018, enquanto os outros clusters diminuem. O quadrado que se formava na diagonal principal desaparece, e é possível notar que os agrupamentos se expandem, terminando no final da matriz.

Figura 27 – Mapa de calor da 'MST' Global com apenas os setores como indicadores do primeiro semestre 2020



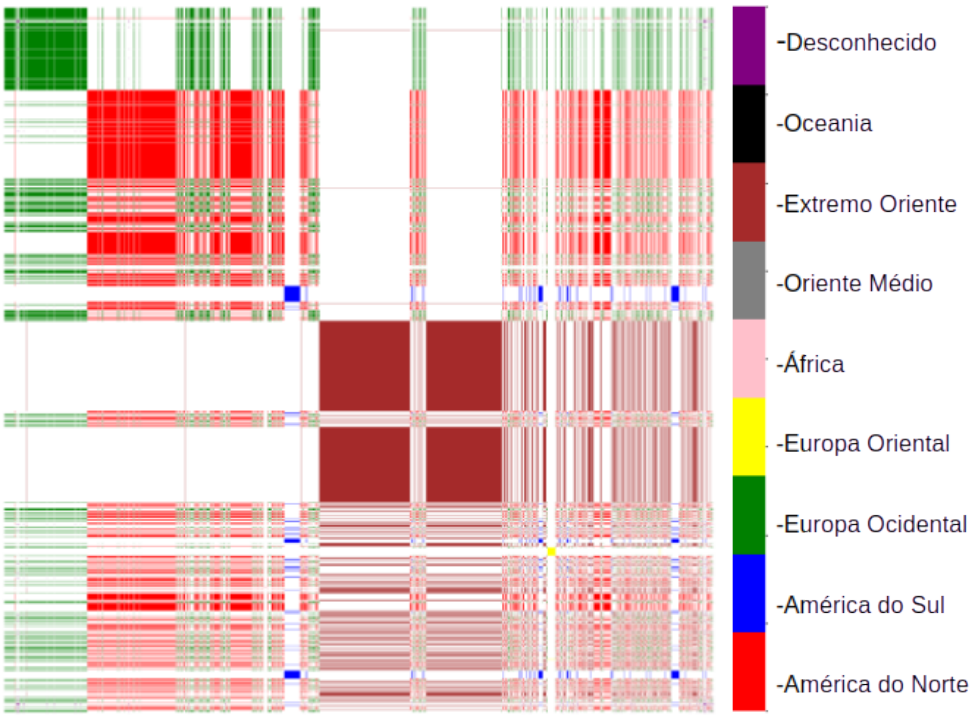
Fonte: elaborado pelo autor (2024).

7.5 Analisando a 'MST' Global pela região

Nestaseção, faremos o mesmo que na seção anterior, mas agora com foco nas regiões, onde observamos vários agrupamentos grandes dominando boa parte do mapa de calor. Começando pelo Extremo Oriente, que está no centro da tabela e forma um 'cluster' considerável. Observa-se também que os pequenos 'clusters' da América do Sul estão relacionados aos setores de Tecnologia e Indústria, diferenciando-se do restante do mundo. Além disso, aquele grande agrupamento observado na 26 agora revela a presença da Europa Ocidental, com dispersões lineares nas laterais. No que se refere ao centro da matriz, notamos a significativa atuação da América do Norte, que se destaca visualmente. Entretanto, na parte superior, composta pela América do Norte e Europa, identificamos pequenos agrupamentos relacionados aos setores de tecnologia, serviços públicos, energia e imóveis—algo não observado nas demais regiões.

Já na figura 29, a Europa Ocidental apresenta uma diminuição de tamanho, enquanto a América do Norte experimenta um crescimento significativo. Nesse contexto, o Extremo Oriente

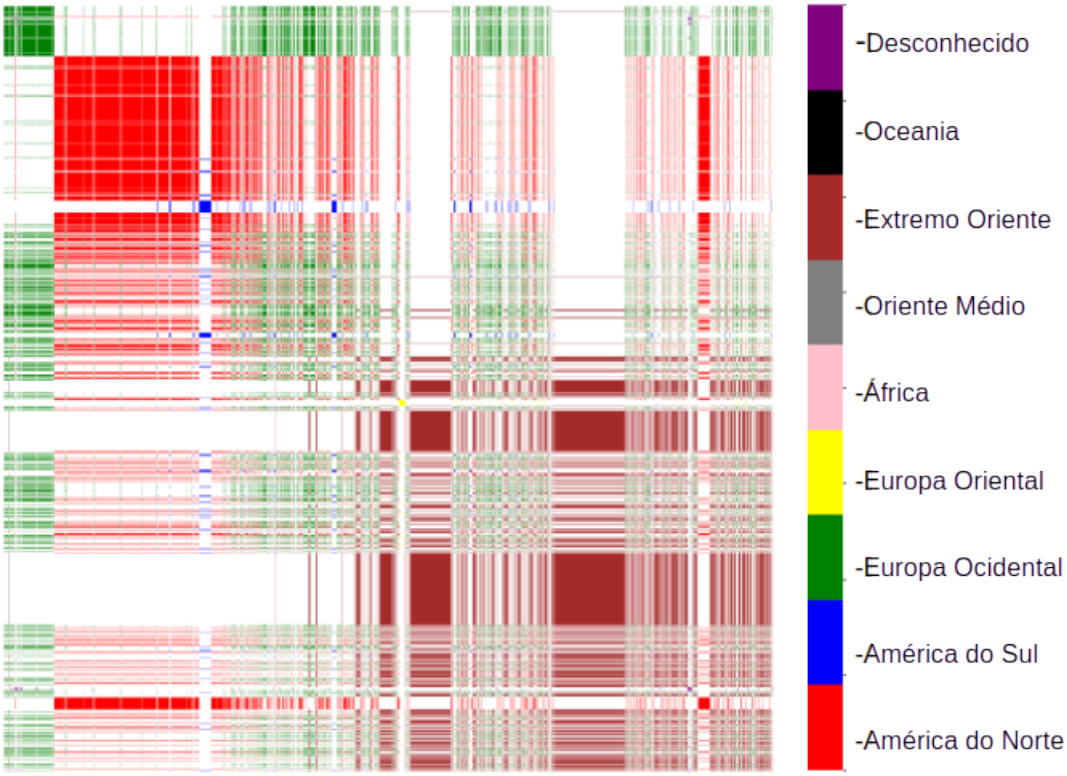
Figura 28 – Mapa de calor da 'MST' Global com apenas as regiões como indicadores do primeiro semestre 2018



Fonte: elaborado pelo autor (2024).

parece se dissipar, e é possível observar agrupamentos da América do Sul dentro do agrupamento da América do Norte, estendendo-se pelo restante do mundo. Vale ressaltar que a Europa Ocidental também exibe um comportamento semelhante, embora com uma proximidade maior em relação ao Extremo Oriente, estabelecendo correlações com essa região. Em contrapartida, a América do Norte demonstra pequenos agrupamentos nesse contexto, onde anteriormente a América do Sul estava mais visivelmente representada.

Figura 29 – Mapa de calor da 'MST' Global com apenas as regiões como indicadores do primeiro semestre 2020



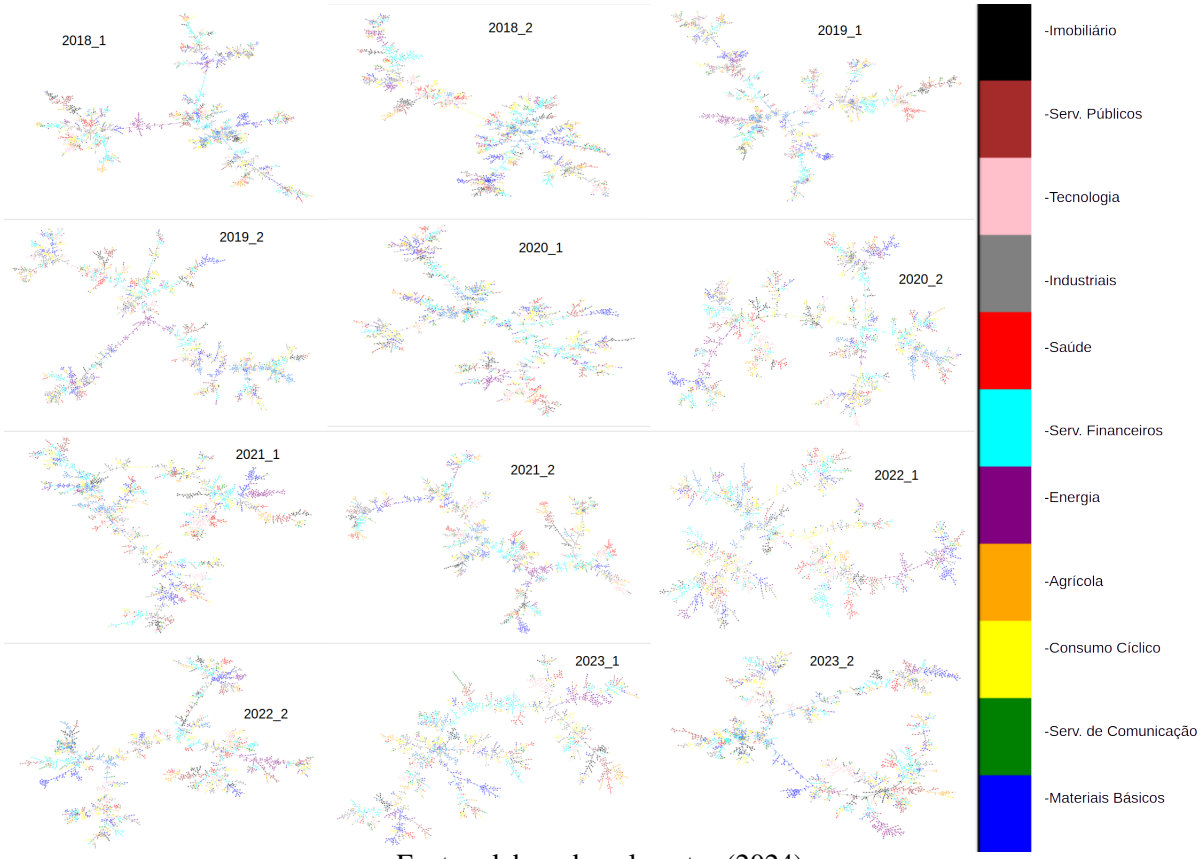
Fonte: elaborado pelo autor (2024).

7.6 Rede 'MST' de correlação Global

Por fim, em nossa análise, a rede "MST" colorida pelos setores, durante o período de normalidade econômica, revela agrupamentos bem definidos, abrangendo diversos setores. Nos anos que antecederam o "lockdown", o setor financeiro está presente nos ramos e em alguns troncos da nossa árvore geradora mínima. À medida que nos aproximamos dos períodos de instabilidade, esses setores se concentram nos troncos, formando "clusters"no centro da árvore.

Além disso, observamos junções dos setores de saúde com serviços públicos isolados, bem como a interconexão entre setores imobiliários e setores públicos. Um grande agrupamento do setor de energia em 2018.1 sustenta um ramo de outros setores, conectando-os ao restante da árvore. Também notamos o consumo cíclico se espalhando pela árvore nos períodos de 2020.2 e 2021.1, uma variedade de interações que agora deixo a cargo do leitor para a observação.

Figura 30 – Redes da ’MST’ da distancia correlação para o período de 2018 a 2024



Fonte: elaborado pelo autor (2024).

8 CONCLUSÃO

No período de instabilidade, observou-se uma diferença acentuada na correlação e na distância de correlação em relação aos demais períodos. Durante a fase de "normalidade" econômica, os agrupamentos dos países na diagonal principal foram bem evidenciados, incluindo o Canadá, que não apresentava uma formação clara antes da organização da matriz de correlação geográfica pela "MST" de cada país.

Analisando o semestre de 2020.1, notamos que o globo estava altamente correlacionado, com destaque para o Ocidente, que apresentava uma acentuação significativa. A matriz de distância de correlação revelou certos agrupamentos ocultos, ligados aos "EUA", à China e a alguns outros países na diagonal principal. Isso se torna ainda mais evidente nas figuras 24 e 25.

Quando comparamos as seções 6.4 e 6.5, observamos a organização das regiões mais correlacionadas e os setores mais interligados. A Europa Ocidental apresenta uma grande correlação com os serviços financeiros, que se espalham pelo mapa, tanto em termos de região quanto de setor. A América do Norte, por sua vez, mostra agrupamentos relacionados aos serviços financeiros e ao ramo da tecnologia. Para o Extremo Oriente, observamos uma relação dispersa com os setores industriais, uma correlação agrupada em 2018.1 com a América do Sul e, posteriormente, em 2020.1 com a América do Norte. Em ambos os períodos, há uma relação com a Europa Ocidental, que se adentra pelo Extremo Oriente, além da América do Norte se espalhando pela parte superior desse último.

Para finalizar a parte de análise, visualizamos a rede "MST", na qual diversos agrupamentos são formados por setores distintos do globo. Notamos que, nos períodos de instabilidade, o setor financeiro se aproxima do tronco da árvore, enquanto os demais formam agrupamentos nos ramos. Além disso, foram observadas diversas relações de proximidade, como a interligação entre os setores da saúde e imobiliário com os serviços públicos, bem como entre o setor de energia e os setores industriais e de serviços públicos.

Esse trabalho mostra a possibilidade de estudar a dinâmica da economia global, utilizando como dados os fechamentos das ações e empregando técnicas de matrizes de correlação, distância de correlação e Minimum Spanning Tree é possível visualizar a relação entre as economias dos países. Para trabalhos futuros, propõe-se adicionar e catalogar de forma mais eficiente as regiões, a fim de observar todas elas em nosso mapa de calor e construir a rede colorindo-as de acordo com essas regiões.

REFERÊNCIAS

ABLOWITZ, M. J.; FOKAS, A. S. **Complex variables:** introduction and applications. 2. ed. Cambridge: Cambridge university press, 2003.

DAVIS, P. **Interpolation and approximation**. New York: Dover Publications, 1963. Disponível em: https://books.google.com.br/books?id=228PAQAAMAAJ. Acesso em: 10 set. 2024.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. **Numerische Mathematik**, *[s. l.]*, v. 1, p. 269-271, 1959.

HALL, C. A.; MEYER, W. Optimal error bounds for cubic spline interpolation. **Journal of Approximation Theory**, *[s. l.]*, v. 16, n. 2, p. 105-122, feb. 1976. Disponível em: https://www.sciencedirect.com/science/article/pii/002190457690040X. Acesso em: 12 set. 2024.

HEINEMAN, G. T.; POLLICE, G.; SELKOW, S. **Algorithms in a nutshell**. Beijing: O'Reilly, 2009.

HELWIG, N. E. **Data, covariance, and correlation matrix**. Minnesota, 2017. 40 slides. Disponível em: http://users.stat.umn.edu/~helwig/notes/datamat-Notes.pdf. Acesso em: 12 set. 2024.

JONES, P. W. First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates. **Monthly Weather Review**, Boston, v. 127, n. 9, p. 2204-2210, sep. 1999. Disponível em: https://journals.ametsoc.org/view/journals/mwre/127/9/ 1520-0493_1999_127_2204_fasocr_2.0.co_2.xml. Acesso em: 14 set. 2024.

KRESS, R. **Numerical analysis**. New York: Springer, 1998. Disponível em: https://books.google.com.br/books?id=Jv_ ZBwAAQBAJ. Acesso em: 16 set. 2024.

KRUSKAL, J. B. On the shortest spanning subtree of a graph and the traveling salesman problem. **Proceeding of the Amerian Mathematical Society**, *[s. l.]*, v. 7, n. 1, p. 48-50, feb. 1956.

MANTEGNA, R. N. Hierarchical structure in financial markets. **The European Physical Journal B**: condensed matter and complex systems, *[s. l.]*, v. 11, p. 193-197, 1999.

PARK, K. **Fundamentals of probability and stochastic processes with applications to communications**. Holmdel, New Jersey: Springer, 2018. Disponível em: https://books.google.com.br/books?id=OmRADwAAQBAJ. Acesso em: 12 set. 2024.

PLETZER, A.; FILLMORE, D. Conservative interpolation of edge and face data on n dimensional structured grids using differential forms. **Journal of Computational Physics**, *[s. l.]*, v. 302, p. 21-40, dec. 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/ S0021999115005562. Acesso em: 18 set. 2024.

PLETZER, A.; HAYEK, W. Mimetic Interpolation of Vector Fields on Arakawa C/D Grids. **Monthly Weather Review**, Boston, v. 147, n. 1, p. 3-16, jan. 2019. Disponível em: https://journals.ametsoc.org/view/journals/mwre/147/1/mwr-d-18-0146.1.xml. Acesso em: 18 set. 2024.

PRIM, R. C. Shortest connection networks and some generalizations. **The Bell System Technical Journal**, *[s. l.]*, v. 36, n. 6, p. 1389-1401, nov. 1957.

RATNER, B. The correlation coefficient: Its values range between+ 1/- 1, or do they? Journal of Targeting, **Measurement and Analysis for Marketing**, *[s. l.]*, v. 17, p. 139-142, may 2009.

RICE, J. **Mathematical statistics and data analysis**. 3th. ed. Australia: Thomson Brooks/Cole, 2007. Disponível em: https://books.google.com.br/books?id= KfkYAQAAIAAJ. Acesso em: 20 set. 2024.

SEDGEWICK, R.; WAYNE, K. **Algorithms**. Upper Saddle River, NJ: Addison-Wesley, 2011.

STANLEY, H. E.; MANTEGNA, R. N. **An introduction to econophysics:** correlations and complexity in finance. Cambridge: Cambridge University Press, 2000.

STEFFENSEN, J. **Interpolation.** 2nd. ed. Mineola, NY: Dover Publications, 2013. Disponível em: https://books.google.com.br/books?id= 0xV7rO28veQC. Acesso em:

STERN, A.; TONG, Y.; DESBRUN, M.; MARSDEN, J. E. Geometric Computational Electrodynamics with Variational Integrators and Discrete Differential Forms. **Geometry, Mechanics and Dynamics Fields Institute Communications**, New York, p. 437-475, 2015. Disponível em: http://dx.doi.org/10.1007/978-1-4939-2441-7_19. Acesso em: 20 set. 2024.

VAKHANIA, N. N. Random vectors with values in quaternion Hilbert spaces. **Theory of Probability & Its Applications**, *[s. l.]*, v. 43, n. 1, 1999.

YE, J. Fuzzy decision-making method based on the weighted correlation coefficient under intuitionistic fuzzy environment. **European Journal of Operational Research**, *[s. l.]*, v. 205, n. 1, p. 202-204, aug. 2010.

ZAR, J. H. Significance testing of the Spearman rank correlation coefficient. **Journal of the American Statistical Association**, *[s. l.]*, v. 67, n. 339, p. 578-580, 1972.