

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

**Um estudo estatístico sobre as redes
de aeroportos no Brasil**

Daniele Gentile Amancio

Trabalho de Conclusão de Curso

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

Um estudo estatístico sobre as redes
de aeroportos no Brasil

Daniele Gentile Amancio
Orientadora: Andressa Cerqueira

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Novembro de 2021

FEDERAL UNIVERSITY OF SÃO CARLOS
EXACT AND TECHNOLOGY SCIENCES CENTER
DEPARTMENT OF STATISTICS

A statistical study of Brazilian airport networks

Daniele Gentile Amancio

Advisor: **Andressa Cerqueira**

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
November 2021

Daniele Gentile Amancio

Um estudo estatístico sobre as redes
de aeroportos no Brasil

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Daniele Gentile Amancio e aprovado pela banca examinadora.

Aprovado em 08 de novembro de 2021.

Banca Examinadora:

- Andressa Cerqueira, DEs-UFSCar
- Rafael Izbicki, DEs-UFSCar
- Ricardo Felipe Ferreira, DEs-UFSCar

Agradecimentos

A Deus pela vida, pela força, fé e por me ajudar a ultrapassar todos os obstáculos encontrados ao longo do curso.

Aos meus pais, Augustinho e Eliane, por desde sempre me apoiarem, de forma incondicional, para que se tornasse realidade a minha formação em uma universidade pública. Aos meus irmãos, Heitor e Isabeli, minha avó Clarinda e tia Elirian, que sempre me incentivaram nos momentos difíceis e entenderam minha ausência enquanto me dedicava à realização deste trabalho de graduação.

Ao meu namorado Flávio pela parceria, incentivo, conforto e torcida pelo meu sucesso.

À minha querida professora orientadora Andressa Cerqueira pelos ensinamentos, pela confiança, pela motivação e compreensão. Um exemplo de pessoa e profissional. À ela, minha eterna gratidão.

De forma especial, às minhas amigas e aqueles que se tornaram amigos nesses anos de graduação.

*“A vida me ensinou a nunca desistir.
Nem ganhar, nem perder mas procurar evoluir.”*

(Charlie Brown Jr.)

Resumo

Neste Trabalho de Conclusão de Curso, propomos a análise da evolução das redes aeroportuárias do Brasil a fim de verificar seu comportamento e se houveram mudanças na sua estrutura durante a pandemia causada pelo *COVID-19*. Para isto, estudamos métodos de detecção de comunidades em redes aleatórias, isto é, como dividir os nós (aeroportos) da rede em grupos que apresentam similaridades em sua estrutura de conexão. Comparamos as comunidades obtidas para as redes aéreas construídas antes e durante a pandemia com o objetivo de determinar se ocorreram mudanças na estrutura da rede.

Palavras-chave: *Agrupamento Espectral Esférico, Clusters, COVID-19, Grafos, Redes Aeroportuárias, Redes Aleatórias.*

Abstract

In this work, we study the evolution of the Brazilian airport networks in order to describe their connectivity structure and to analyze if their structure changed during the pandemic caused by *COVID-19*. To this end, we study methods of community detection on random networks, that is, how to divide the nodes (airports) of the network into groups that have similar connection patterns. We compare the communities obtained for the airport networks built before and during the pandemic in order to determine if there are changes in their structure.

Keywords: *Airport Networks, Clusters, COVID-19, Graphs, Spherical Spectral Clustering, Random Networks..*

Lista de Figuras

2.1	Exemplo de grafos.	4
2.2	Exemplo de grafos para análise dos graus.	7
3.1	Redes aeroportuárias dos anos de 2018, 2019, 2020 e 2021.	13
3.2	Histograma dos graus de entrada e saída referentes ao 2 ^o trimestre das redes sem pesos.	15
3.3	Histograma dos graus de entrada e saída referentes ao 3 ^o trimestre das redes sem pesos.	16
3.4	Histograma dos graus de entrada e saída referentes ao 3 ^o bimestre das redes sem pesos.	18
3.5	Histograma dos graus de entrada e saída referentes ao 4 ^o bimestre das redes sem pesos.	19
3.6	Histograma dos graus de entrada e saída referentes ao 5 ^o bimestre das redes sem pesos.	20
3.7	Histograma dos graus de entrada e saída referentes ao 3 ^o bimestre das redes com pesos.	22
3.8	Histograma dos graus de entrada e saída referentes ao 4 ^o bimestre das redes com pesos.	23
5.1	Agrupamento das redes aeroportuárias em cinco grupos ($k = 5$) nos anos de 2018, 2019, 2020 e 2021.	30
5.2	Agrupamento das redes aeroportuárias com 5 grupos ($k = 5$) dos 6 ^o bimestres dos anos 2019 e 2020	34

Lista de Tabelas

2.1	Graus de entrada e saída do grafo direcionado	8
3.1	Quantidade de observações em cada rede analisada	12
3.2	Descritivas das redes por trimestres	14
3.3	Descritivas das redes por bimestres	17
3.4	Descritivas das redes por bimestres com pesos	21
5.1	Descritivas dos agrupamentos da rede sem pesos e sem direção	32
5.2	Informação mútua normalizada	33
5.3	Descritivas dos agrupamentos das redes do 6 ^o bimestre de 2019 e 2020 . . .	35

Sumário

1	Introdução	1
2	Redes	3
2.1	Grafos	3
2.2	Matriz de Adjacência	4
2.3	Redes com pesos	5
2.4	Graus e sua distribuição	6
2.5	Reciprocidade	9
2.6	Transitividade	9
3	Análise Descritiva e Exploratória	11
3.1	Estrutura dos dados	11
3.2	Análise descritiva e exploratória das redes	12
3.2.1	Redes sem pesos	14
3.2.2	Redes com pesos	21
4	Agrupamentos	25
4.1	Métodos de Agrupamentos	25
4.1.1	Agrupamento Espectral	26
4.1.2	Algoritmo de Agrupamento Espectral Esférico	27
4.2	Comparações de agrupamentos	28
4.2.1	Informação Mútua Normalizada	28
5	Resultados dos Agrupamentos	29
5.1	Estrutura dos dados	29
5.2	Resultados: Redes sem pesos e sem direção	30

6	Considerações Finais	37
A	Códigos Utilizados - <i>Software R</i>	39

Capítulo 1

Introdução

As redes aleatórias têm recebido grande atenção da comunidade estatística devido à sua vasta aplicação na modelagem de dados, por exemplo: redes de internet, transporte, telefônicas, elétricas, entrega e distribuição, entre outras. Da mesma forma, muitos estudos também foram realizados a fim de verificar como se comportam as redes de transportes que podem ser aéreas, rodoviárias e ferroviárias.

As redes aéreas podem ser reconstruídas a partir de dados publicados de companhias aéreas e as redes ferroviárias e rodoviárias podem ser obtidas a partir de mapas, como discutido por Newman (2012). Geralmente, a estrutura dessas redes não é difícil de determinar, apesar de que a compilação dos dados possa ser trabalhosa. Uma observação interessante, segundo o estudo de Colizza *et al.* (2006), é que a malha aeroportuária mundial é uma rede significativa para a proliferação de diversos vírus que causam epidemias.

Em síntese, os aeroportos de um país e as conexões entre diferentes aeroportos formam uma rede que pode ser representada via grafos, onde os nós do grafo são os aeroportos em estudo e as arestas representam que pelo menos um voo ocorreu de um aeroporto para outro. A dinâmica de uma rede aérea pode ser estudada em um longo período de tempo, quando o pesquisador está interessado na rede aérea mensal ou em um curto período de tempo, quando há interesse na rede aérea construída a cada hora.

A evolução da rede aeroportuária brasileira foi estudada por Rocha (2009) que demonstrou que, apesar de algumas flutuações em algumas propriedades estruturais, a rede passou por uma forte dinâmica de religação durante alguns períodos, que a princípio pareceu uma desaceleração no setor de aviação, mas, na verdade, foi uma reorganização da estrutura conforme a demanda mudava. Alguns eventos influenciam na mudança da estrutura de redes aéreas como, por exemplo, os atentados terroristas de 11 de Setembro

nos Estados Unidos como estudado em Siozos-Rousoulis *et al.* (2021).

Desse modo, nosso objetivo é o estudo da rede aérea brasileira com foco na comparação entre o período antes e durante a pandemia causada pelo *COVID-19* (*Coronavirus Disease 2019*), consideraremos o início da pandemia o dia 25 de Fevereiro de 2020, que foi o primeiro caso registrado na cidade de São Paulo no Brasil.

Neste contexto, neste Trabalho de Conclusão de Curso, focaremos nas redes aeroportuárias que tem seus dados publicados e disponíveis no site da Agência Nacional de Aviação Civil (ANAC, 2021). O objetivo principal, proposto neste estudo, é estudar a evolução das redes aeroportuárias do Brasil a fim de verificar seu comportamento e se houveram mudanças após o início da pandemia, nosso foco são as redes construídas em um longo período de tempo, isto é, redes aéreas obtidas mensalmente.

Neste âmbito, consideraremos pares de cidades (ou aeroportos) formando uma rede em que os vértices (nós) serão representados pelas siglas dos aeroportos; e, se houveram voos entre dois aeroportos, então haverá uma linha que liga esses nós, denominada aresta. Também é possível utilizar pesos nas arestas, por exemplo, o número de voos. Essa estrutura nos ajudará a representar a complexidade das redes de transporte aéreo, nos ajudando a entender nossa rede de estudo.

Para atingir o objetivo do trabalho será realizada uma revisão bibliográfica definindo os métodos, através de conceitos, medidas e definições a serem utilizados. Esse relatório contém, além do Capítulo 1, mais cinco capítulos. No Capítulo 2 falaremos sobre Redes, explicaremos como representá-las a partir de grafos, matriz de adjacência, redes com pesos, graus e suas distribuições, transitividade e reciprocidade que são medidas utilizadas no estudo de redes. No Capítulo 3, faremos a análise descritiva das redes de acordo com o que estudamos no capítulo anterior, com ela definiremos qual a metodologia adequada para utilizarmos adiante. No Capítulo 4, entenderemos como funciona o agrupamento para redes, entenderemos o que é o agrupamento espectral, em especial, o algoritmo do agrupamento espectral esférico e, por fim, conheceremos a definição de informação mútua normalizada, que é um método utilizado para comparação de agrupamentos. No Capítulo 5, verificaremos se, de fato, houveram mudanças nas redes de aeroportos no Brasil devido a pandemia causada pelo *COVID-19*, primeiramente será realizada um nova estruturação dos dados para, enfim, tirarmos conclusões sobre o estudo. E, finalmente, no Capítulo 6, relataremos as conclusões deste trabalho, assim como as contribuições e sugestões para estudos futuros.

Capítulo 2

Redes

Neste capítulo, estudaremos redes e apresentaremos alguns conceitos, medidas e definições. No contexto deste trabalho, consideraremos pares de cidades (ou aeroportos) formando uma rede em que os vértices (nós) serão representados pelos locais; e, se houveram voos entre dois locais, então haverá uma linha que liga esses nós, denominada aresta. Essa estrutura nos ajudará a representar a complexidade da rede de transporte aéreo, portanto, nesse capítulo, focaremos nos conceitos de grafo, matriz de adjacência, rede com peso, grau e sua distribuição, transitividade e reciprocidade.

2.1 Grafos

Na literatura, redes também são representadas por grafos, sendo eles um objeto matemático caracterizado pelo par (V, A) ; sendo V o conjunto de vértices (objetos de estudo) e A o conjunto de arestas (conexões entre esses objetos), sendo um conjunto de pares ordenados - somente se o grafo for direcionado, em que a primeira coordenada é um vértice e a segunda coordenada é outro vértice. Por exemplo, se estamos interessados em estudar as redes de transporte aéreo brasileiro, os vértices são aeroportos e as arestas, ou conexões, são os voos entre esses aeroportos. Na literatura matemática, é comum que o número de vértices de uma rede seja denotado por n e o número de arestas por m .

Os grafos podem ser direcionados ou não direcionados. Nos grafos direcionados a relação entre os vértices possui uma direção. Por exemplo, se o aeroporto A tem conexão com o aeroporto B, essa conexão pode ser representada no grafo com uma seta compondo a aresta apontando de A para B, entretanto, se o aeroporto B não tem conexão com o aeroporto A, não teremos uma seta retornando para ele. Nos grafos simples não temos

esse direcionamento, apenas arestas que significam que A tem conexão com B e vice-versa. Além disso, em grafos simples um vértice não pode se conectar a si mesmo.

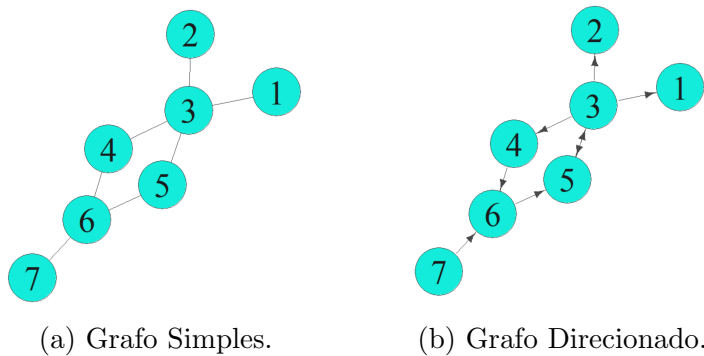


Figura 2.1: Exemplo de grafos.

Na Figura 2.1, exibimos um exemplo de dois grafos com número de vértices $n = 7$. Através da Figura 2.1a observamos um grafo simples, em que os vértices $i, j \in V$ e as arestas não possuem direcionamento, ou seja, se i vai para j , j vai para i . Já quando observamos a Figura 2.1b, temos um exemplo de grafo direcionado, ou seja, se i vai para j , não necessariamente j vai para i . Exemplificando, se observarmos os vértices 1 e 3, notamos que, na Figura 2.1a, temos uma aresta ligando esses dois vértices, ou seja, se 1 vai para 3, 3 vai para 1, por outro lado, se observarmos a Figura 2.1b utilizando os mesmos vértices, notamos que 3 vai para 1, mas 1 não vai para o 3. Veremos na próxima seção a representação matricial desses grafos.

2.2 Matriz de Adjacência

Uma das maneiras de representar uma rede matematicamente é através de uma matriz de adjacência \mathbf{A} . Essa matriz possui entradas 0 ou 1, dimensão $n \times n$ e é definida para $1 \leq i, j \leq n$, como

$$A_{ij} = \begin{cases} 1, & \text{se houver uma aresta de } i \text{ para } j; \\ 0, & \text{caso contrário.} \end{cases}$$

Por exemplo, a partir do grafo da Figura 2.1a obtemos a matriz de adjacência de um grafo simples, dada por

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \quad (2.1)$$

Baseado na matriz de adjacência (2.1) observamos uma simetria que ocorre em grafos simples, pois se temos uma aresta de i para j , necessariamente temos uma aresta indo de j para i . Em contrapartida, com base na Figura 2.1b temos a matriz de adjacência (2.2), sendo que, nesse caso, não observamos simetria e a matriz de adjacência é dada por

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \quad (2.2)$$

Com essas definições em mente, vamos introduzir um novo conceito, conhecido como redes com pesos.

2.3 Redes com pesos

Estudamos redes que possuem conexões ou não entre os vértices, definidas por grafos simples ou direcionados, no entanto, em algumas situações, gostaríamos de representar redes com pesos nas arestas. A matriz de adjacência das redes com pesos é representada, em cada entrada, com valores iguais aos pesos de cada conexão correspondente.

Exemplificando, nas redes de transporte aéreo, temos o interesse de verificar a quantidade de passageiros que vão de um aeroporto A a um aeroporto B ou a quantidade de voos entre esses aeroportos, assim, essas quantidades representariam os pesos de cada conexão.

Temos, como exemplo, a matriz de adjacência (2.3) que representa uma rede ponderada, em que a conexão entre os vértices 2 e 3 é duas vezes mais forte do que entre os vértices 1 e 3, que por sua vez, é mais forte do que entre os vértices 1 e 2.

$$\mathbf{A} = \begin{pmatrix} 0 & 0.5 & 1 \\ 0.5 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}. \quad (2.3)$$

2.4 Graus e sua distribuição

O grau do vértice i de um grafo é o número de arestas conectadas a ele, denotaremos o grau do vértice i por k_i . Para um grafo não direcionado com n vértices, o grau pode ser escrito em termos de matriz de adjacência como

$$k_i = \sum_{j=1}^n A_{ij}.$$

O número de arestas, para um grafo não direcionado, que pode ser obtido através da matriz de adjacência que é dado por

$$\sum_{i=1}^n \sum_{j>i}^n A_{ij} = m.$$

Se somarmos todos os graus dos vértices, temos

$$\sum_{i=1}^n k_i = \sum_{i=1}^n \sum_{j=1}^n A_{ij}.$$

Mas, como estamos somando os graus de um grafo simples, ou seja, não direcionado, sabemos que a matriz de adjacência é simétrica, então obtemos

$$\sum_{i=1}^n k_i = 2 \sum_{i=1}^n \sum_{j>i}^n A_{ij} = 2m.$$

Portanto, concluímos que o número total de arestas do grafo é calculado por

$$\frac{1}{2} \sum_{i=1}^n k_i.$$

O grau médio \bar{k} é a média dos graus de todos os vértices dado por

$$\bar{k} = \frac{1}{n} \sum_{i=1}^n k_i.$$

Estudaremos agora o conceito da distribuição da frequência do grau de um vértice, lembrando que o grau de um vértice é o número de arestas conectadas a ele, temos que a fração dos vértices de uma rede com grau k é definida por p_k . Essas quantidades representam a distribuição dos graus da rede e são interpretadas como a probabilidade de um vértice na rede possuir grau k . Vamos considerar os grafos da Figura 2.2.

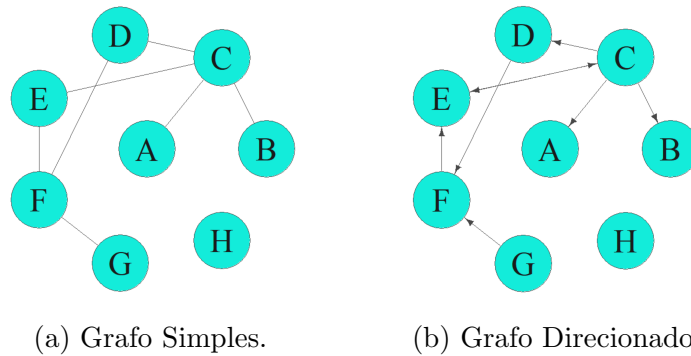


Figura 2.2: Exemplo de grafos para análise dos graus.

Analisando a Figura 2.2a, que corresponde a grafo simples com $n = 8$ vértices, temos que: o vértice H tem grau 0, os vértices A, B e G têm grau 1, os vértices D e E têm grau 2, o vértice F tem grau 3 e, por fim, o vértice C tem grau 4. Assim, temos que, para os graus $k = 0, 1, 2, 3$ e 4, os valores de p_k são dados por

$$p_0 = \frac{1}{8}, p_1 = \frac{3}{8}, p_2 = \frac{2}{8}, p_3 = \frac{1}{8}, p_4 = \frac{1}{8} \text{ e } p_k = 0, \text{ para todo } k > 4.$$

Podemos também estudar os graus de um grafo direcionado, que possuem dois tipos de graus diferentes para cada vértice, sendo eles o grau de entrada e o grau de saída de grau, basicamente, são o número de arestas entrando ou saindo no vértice de interesse. Observando então a Figura 2.2b, que corresponde a um grafo direcionado com $n = 8$ vértices, temos descritos na Tabela 2.1 os graus de entrada e saída de cada vértice.

Tabela 2.1: Graus de entrada e saída do grafo direcionado

	Grau de Entrada	Grau de Saída
Vértice A	1	0
Vértice B	1	0
Vértice C	1	4
Vértice D	1	1
Vértice E	2	1
Vértice F	2	1
Vértice G	0	1
Vértice H	0	0

A frações dos graus de entrada (p_{e_k}) são dadas por

$$p_{e_0} = \frac{2}{8}, p_{e_1} = \frac{4}{8}, p_{e_2} = \frac{2}{8} \text{ e } p_{e_k} = 0, \text{ para todo } k > 3.$$

A frações dos graus de saída (p_{s_k}) são dadas por

$$p_{s_0} = \frac{3}{8}, p_{s_1} = \frac{4}{8}, p_{s_2} = \frac{0}{8} = 0, p_{s_3} = \frac{0}{8} = 0, p_{s_4} = \frac{1}{8} \text{ e } p_{s_k} = 0, \text{ para todo } k > 4.$$

Podemos também, ao invés da fração de vértices com um determinado grau, encontrar o número total desses vértices com grau k . Um modo simples de calcular isso é a partir da distribuição de graus, sendo dado por np_k , onde n é o número total de vértices.

O conceito de grau é estendido aos grafos com pesos, sendo definido \bar{w} como grau médio das redes com pesos, que nada mais é que a média da soma dos pesos de cada aresta da rede. Por exemplo, a matriz de adjacência

$$\mathbf{A} = \begin{pmatrix} 0 & 5 & 2 \\ 4 & 0 & 2.5 \\ 1 & 3 & 0 \end{pmatrix}.$$

Possui grau médio (\bar{w}) dado por $\bar{w} = \frac{0+5+2+4+0+2.5+1+3+0}{9} = 1.94$

2.5 Reciprocidade

Também é interessante observarmos a frequência de ciclos (*loops*) de comprimento dois, onde há um par de vértices, entre os quais existem arestas direcionadas correndo em ambas as direções, ou seja, quando estamos no caso de uma rede direcionada, medimos essa frequência pela reciprocidade.

Se há uma aresta direcionada do vértice i para o vértice j e também houver uma aresta de j para i , então dizemos que a aresta de i para j é recíproca. A reciprocidade R é definida como a fração de arestas que são recíprocas. Analisando o produto de elementos da matriz de adjacência temos que: $A_{ij} \times A_{ji}$ assume o valor 1 se, e somente se, houver uma aresta de i para j e uma aresta de j para i ; e 0 caso contrário. Podemos então somar todos os pares de vértices i, j para obter uma expressão dada para a reciprocidade

$$R = \frac{1}{m} \sum_{ij} A_{ij} A_{ji} = \frac{1}{m} \text{Tr} \mathbf{A}^2,$$

onde m é o número total de arestas direcionadas e $\text{Tr}(\mathbf{A}^2)$ é o traço da matriz \mathbf{A}^2 . A reciprocidade assume valores entre 0 e 1, sendo que quanto mais próximo de 1, maior é a reciprocidade da rede.

2.6 Transitividade

Transitividade é uma propriedade interessante em teoria das redes não direcionadas, ela mede o grau com que os nós de um grafo tendem a se agrupar. Utilizaremos, neste trabalho, o **coeficiente de agrupamento** (*clustering coefficient*, em inglês), que nada mais é do que a frequência que ciclos (*loops*) de comprimento três aparecem em uma rede.

Um clique consiste em um subgrafo no qual todos os vértices estão conectados entre si. A transitividade perfeita ocorre quando um subgrafo está totalmente conectado. A transitividade parcial é dada, por exemplo, em redes sociais, quando u conhece v e v conhece w , mas não necessariamente u conhece w , porém é mais provável que u conheça w , do que qualquer outro vértice. Quando u realmente conhece w é formado um ciclo (*loop*) de comprimento três, então há um triângulo na rede, quando isso ocorre u, v e w são denominados como uma tríade fechada. Quando temos três nós conectados por apenas duas por arestas, temos então a tríade aberta.

O coeficiente de agrupamento mede o número total de tríades fechadas sobre o número total de tríades (abertas e fechadas). Por definição, temos então

$$\bar{C} = \frac{\text{número de tríades fechadas}}{\text{número de tríades abertas e fechadas}}.$$

Segundo Newman (2012), as redes sociais tendem a ter valores elevados do coeficiente de agrupamento. Tais como as redes discutidas pelo autor, como por exemplo: a rede de colaborações de atores de cinema que tem $\bar{C} = 0.2$; a rede de colaborações entre biólogos apresenta $\bar{C} = 0.09$; a rede de e-mails enviados entre membros de uma grande universidade tem $\bar{C} = 0.16$; a Internet no nível do sistema autônomo, por exemplo, possui um coeficiente de agrupamento (\bar{C}) em torno de 0.01.

Capítulo 3

Análise Descritiva e Exploratória

Nesse capítulo, faremos um breve estudo dos dados, para que possamos verificar qual o comportamento das redes aéreas brasileiras de acordo com as medidas descritivas que foram abordadas no Capítulo 2.

3.1 Estrutura dos dados

A base de dados contém informações sobre os voos nacionais e regulares das redes aeroportuárias que tem seus dados publicados e disponíveis no site da Agência Nacional de Aviação Civil (ANAC, 2021).

Verificaremos a evolução das redes aeroportuárias do Brasil dos anos de 2018, 2019, 2020 e parte do ano de 2021 (janeiro a agosto), com os dados que foram disponibilizados pela ANAC, a fim de observar o comportamento dessas redes e se houveram mudanças após o início da pandemia causada pelo *COVID-19*.

Para auxiliar a interpretação e uso dos dados estatísticos, são descritas as variáveis utilizadas no estudo da mesma forma que foram definidas pela ANAC:

Ano e Mês: Referente ao ano e ao mês que ocorreu o voo.

Aeroporto de Origem: Nesse caso, temos 4 variáveis pertencentes ao aeroporto de origem, sendo elas: **Sigla**, **Nome**, **UF** e **Região**: referentes a sigla, cidade, estado e região, respectivamente, em que os aeroportos de origem estão localizados.

Aeroporto de Destino: Nesse caso, temos 4 variáveis também, mas são pertencentes ao aeroporto de destino, sendo elas: **Sigla**, **Nome**, **UF** e **Região**: referentes a sigla, cidade, estado e região, respectivamente, em que os aeroportos de destino estão localizados.

Natureza do Voo - Domésticos: voos que tenham a decolagem e o pouso realizadas no Brasil e sejam operadas por empresas brasileiras.

Tipo de Voo: O tipo de voo é referente ao tipo de operação, nesse trabalho será utilizado apenas os voos **Regulares**, que ocorrem conforme previsão no Sistema de Registro de Operações e recebem esse nome, pois possuem a característica de serem realizadas regularmente.

Decolagens: Referente ao número de decolagens que ocorreram entre os aeroportos de origem e destino da etapa.

3.2 Análise descritiva e exploratória das redes

Para verificar o comportamento das variáveis do banco de dados estudado sobre as redes aeroportuárias, serão realizadas algumas análises descritivas comumente usadas em redes, conforme estudadas no Capítulo 2, que funcionarão como uma prévia para a futura análise de agrupamento.

Na Tabela 3.1, averiguamos a quantidade de observações referentes a diferentes rotas em diferentes anos e o número de voos por cada ano. Vale ressaltar que o ano de 2021 só possui dados disponibilizados até o mês de agosto.

Tabela 3.1: Quantidade de observações em cada rede analisada

Ano	Observações	Número de rotas	Número de voos
2018	17919	1389	783886
2019	16336	1498	765442
2020	11158	1446	377147
2021	8281	1203	302472

Averiguando os valores calculados da Tabela 3.1 percebemos que houve uma grande diferença no número de observações e na quantidade de voos no ano de 2020, em que a quantidade de voos diminuiu aproximadamente 50% em relação aos anos de 2018 e 2019, o que nos dá indícios de que a rede sofreu alterações no ano que iniciou-se a pandemia causada pelo *COVID-19*.

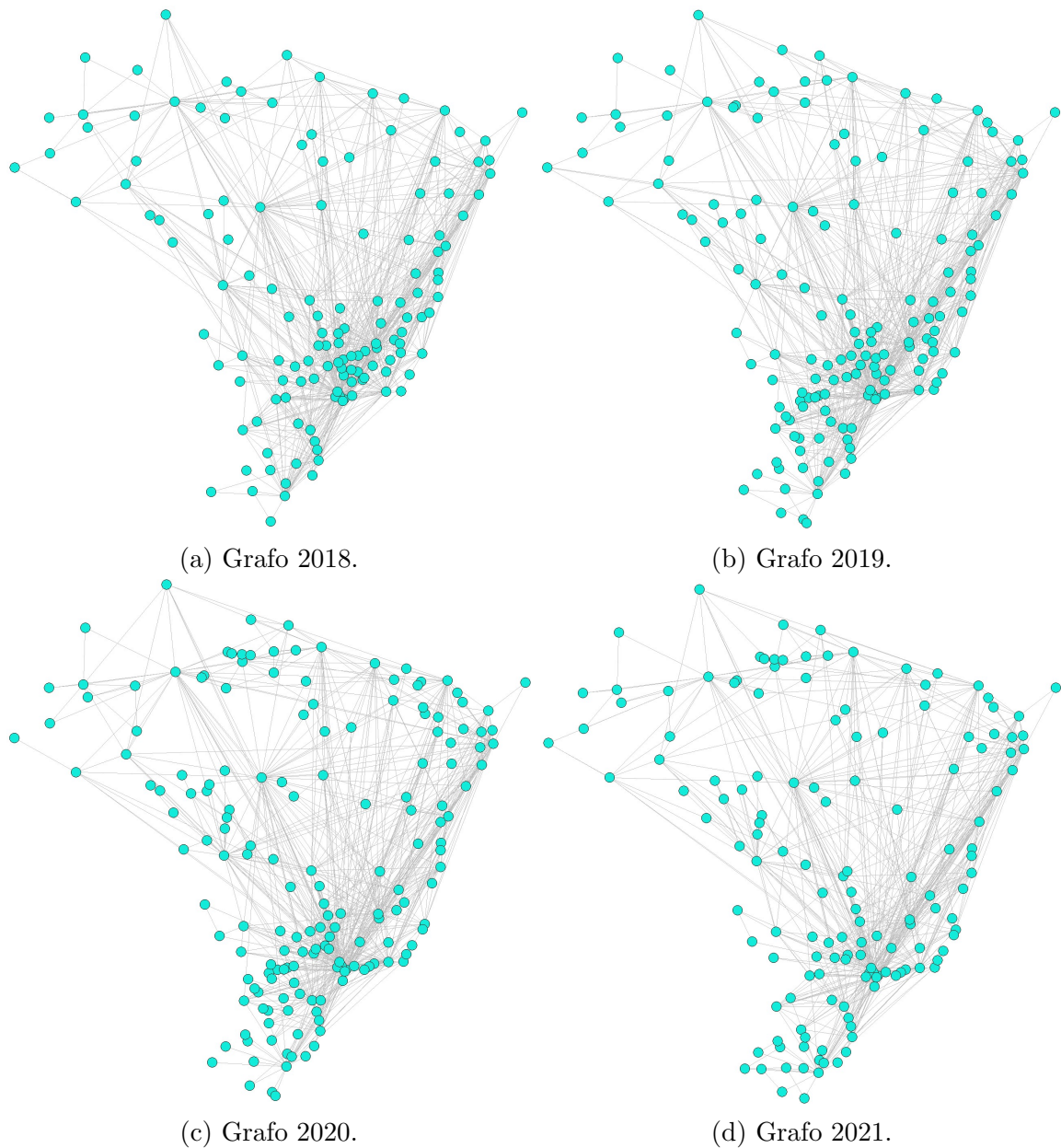


Figura 3.1: Redes aeroportuárias dos anos de 2018, 2019, 2020 e 2021.

Observamos, a partir da Figura 3.1, como é curioso que apesar do número de voos ter diminuído drasticamente nos anos de 2020 e 2021, o número de rotas se manteve muito próximo em todos os anos. Observamos também a configuração da rede mudou um pouco em 2020 e, aparentemente, tem retornado a configuração de 2018/2019 no ano de 2021.

Para realizar o estudo descritivo usaremos duas abordagens. Na primeira abordagem, analisaremos redes sem pesos, onde cada aresta representa que pelo menos um voo ocorreu. Na segunda, analisaremos redes com pesos, onde cada aresta entre os aeroportos A e B, representa o número de voos do aeroporto A para o aeroporto B. Em ambas abordagens utilizaremos redes direcionadas.

3.2.1 Redes sem pesos

Nesta seção, começaremos analisando redes sem pesos, para isso, calcularemos algumas medidas descritivas, sendo elas: Número de aeroportos (N), Número de arestas (A) (ou seja, número de rotas), Grau médio (\bar{k}), Grau médio de entrada/saída (\bar{k}_{es}), Transitividade (\bar{C}) e Reciprocidade (R).

Tabela 3.2: Descritivas das redes por trimestres

	Ano	N	A	\bar{k}	\bar{k}_{es}	R	\bar{C}
1º Trimestre	2018	140	883	12.61	6.31	0.92	0.36
	2019	165	938	11.37	5.68	0.87	0.37
	2020	183	989	10.81	5.40	0.86	0.33
	2021	150	911	12.15	6.07	0.88	0.34
2º Trimestre	2018	140	890	12.71	6.36	0.88	0.37
	2019	165	900	10.91	5.45	0.87	0.37
	2020	183	414	4.52	2.26	0.83	0.30
	2021	150	716	9.55	4.77	0.91	0.30
3º Trimestre	2018	140	903	12.90	6.45	0.89	0.36
	2019	165	848	10.28	5.14	0.92	0.34
	2020	183	612	6.69	3.34	0.89	0.30
	2021	150	794	10.59	5.29	0.92	0.32
4º Trimestre	2018	140	922	13.17	6.59	0.87	0.36
	2019	165	949	11.50	5.75	0.86	0.32
	2020	183	840	9.18	4.59	0.89	0.34

Podemos observar, a partir da Tabela 3.2, que o número de aeroportos entre os anos de 2018, 2019 e 2020 foram aumentando gradativamente, entretanto, vimos que em 2021 esse número diminuiu, o motivo é que foi disponibilizado os dados desse ano somente até o mês de agosto.

Notamos que o número de rotas diminuiu drasticamente nos 2º e 3º trimestres de 2020 comparado com as demais redes, com isso, o grau médio também diminuiu extremamente nesses dois trimestres, consequentemente, o grau médio de entrada/saída também.

Percebemos também que a reciprocidade (R) é bem alta, mas não assume o valor 1, a partir disso, sabemos que é possível voar de um aeroporto para outro, mas não

necessariamente seria possível voltar pela mesma rota. Verificamos, considerando agora a rede não direcionada, que a transitividade (\bar{C}) entre todas as redes é considerada alta, o que é esperado em uma rede de aeroportos.

Notamos que os trimestres que mais se destacaram foram o 2º e 3º, analisaremos melhor essas diferenças, a partir dos histogramas dos graus de entrada e saída de cada uma dessas redes por meio das Figuras 3.2 e 3.3.

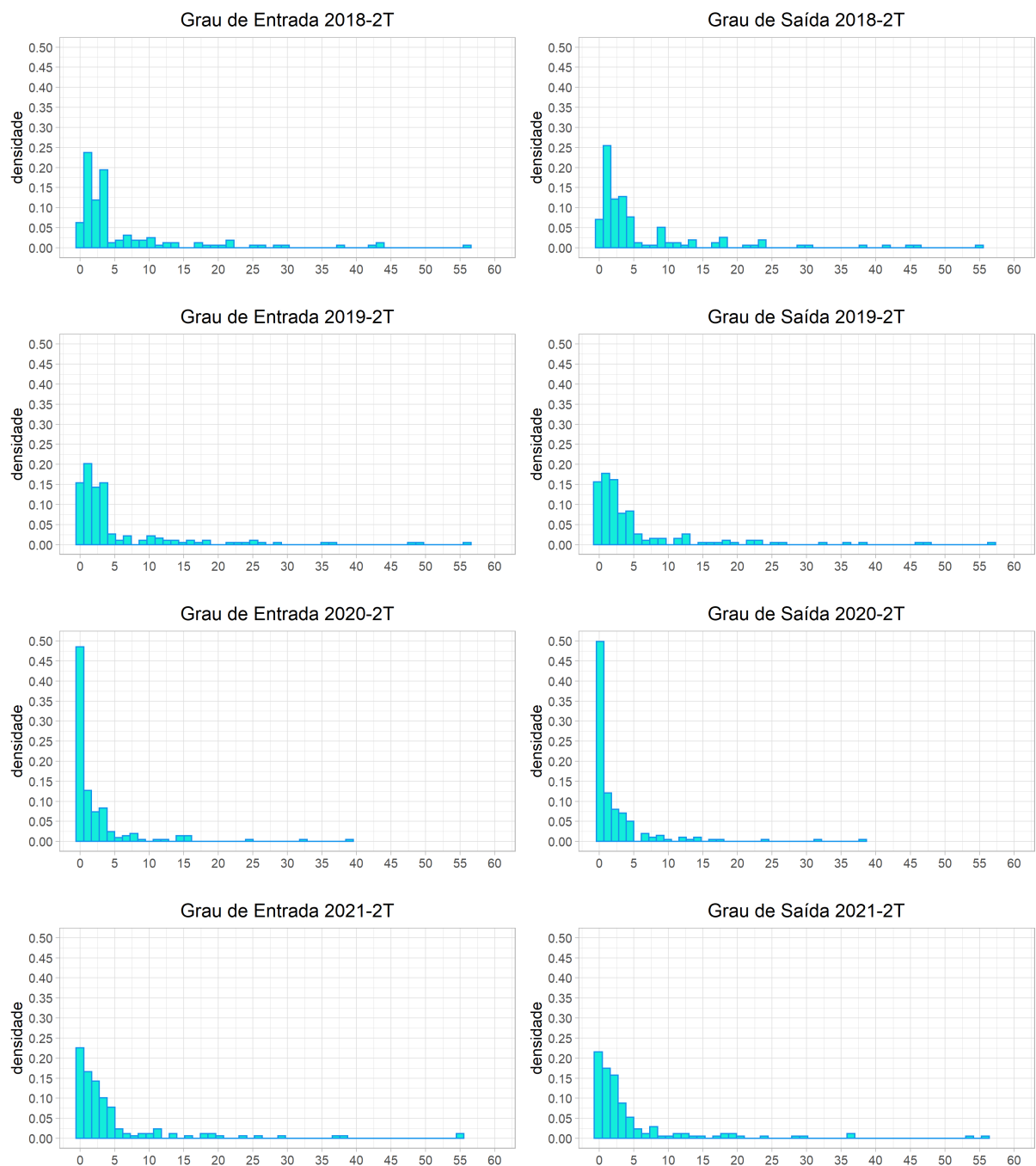


Figura 3.2: Histograma dos graus de entrada e saída referentes ao 2º trimestre das redes sem pesos.

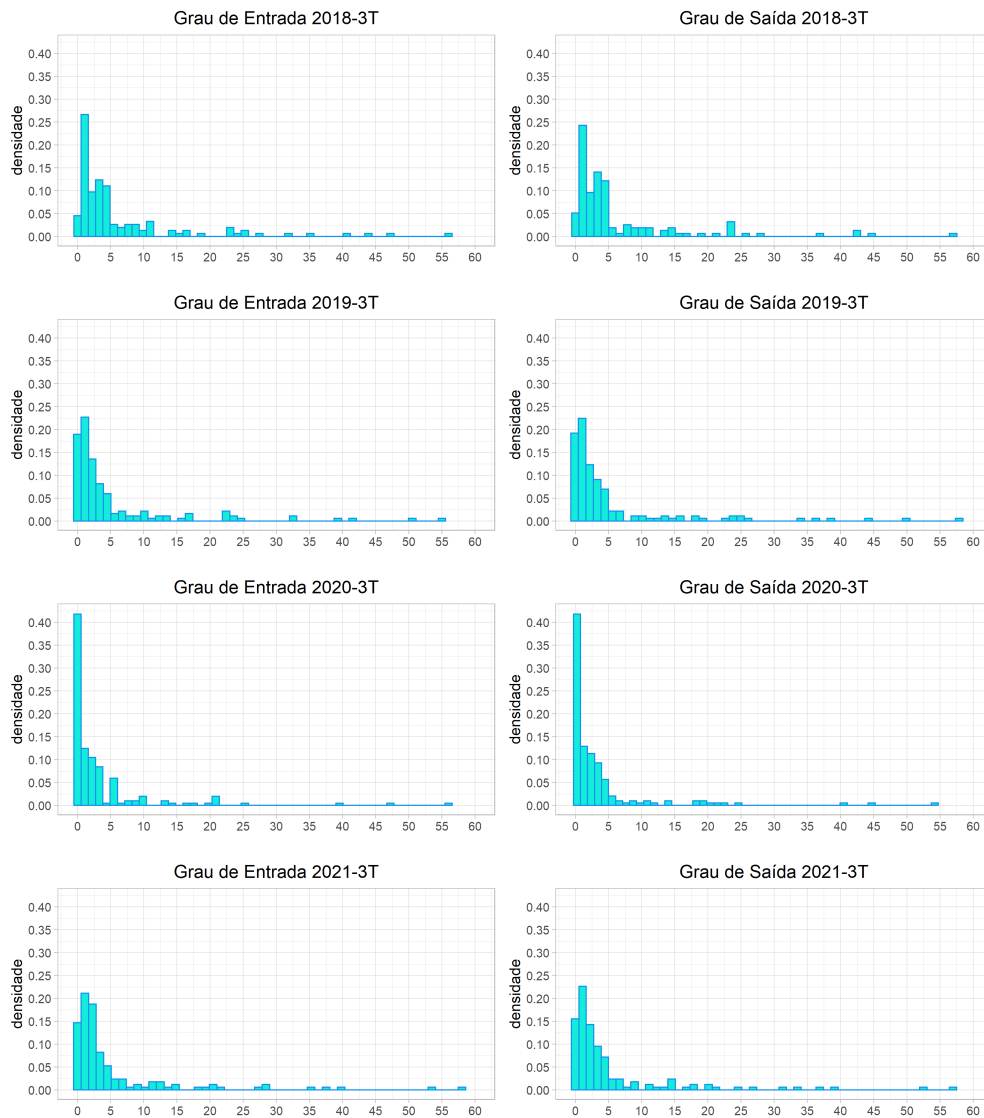


Figura 3.3: Histograma dos graus de entrada e saída referentes ao 3^o trimestre das redes sem pesos.

Observamos, a partir dos dos histogramas dos graus de entrada e saída, que todas as redes observadas no 3^o e 4^o trimestres possuem distribuição assimétrica a esquerda, melhor dizendo, os histogramas que representam os graus de entrada e saída são heterogêneos, ou seja, temos muitos vértices com graus pequenos (entre 0 e 5) e poucos vértices com graus altos (em torno de 50). Notamos que no ano de 2020 há uma alta concentração nos graus nos mostrando, de fato, todas as diferenças que foram observadas na Tabela 3.2, ou seja, os aeroportos que tiveram até 5 voos em 2020 receberam mais voos do que nos anos anteriores, fazendo então com que o grau médio fosse menor. Entretanto, separando a rede por trimestre não fica tão clara onde estão localizadas essas desigualdades, por esse motivo, decidimos “afunilar” e analisar as redes bimestralmente a fim de detectar melhor essas diferenças.

Tabela 3.3: Descritivas das redes por bimestres

	Ano	N	A	\bar{k}	\bar{k}_{es}	R	\bar{C}
1º Bimestre	2018	140	826	11.80	5.90	0.93	0.36
	2019	165	890	10.79	5.39	0.89	0.36
	2020	183	945	10.33	5.16	0.88	0.32
	2021	150	866	11.55	5.77	0.89	0.34
2º Bimestre	2018	140	830	11.86	5.93	0.92	0.34
	2019	165	876	10.62	5.31	0.89	0.35
	2020	183	858	9.38	4.69	0.91	0.29
	2021	150	748	9.97	4.99	0.91	0.30
3º Bimestre	2018	140	862	12.31	6.16	0.89	0.37
	2019	165	826	9.98	4.99	0.90	0.35
	2020	183	380	4.15	2.08	0.84	0.29
	2021	150	698	9.31	4.65	0.91	0.30
4º Bimestre	2018	140	850	12.14	6.07	0.91	0.35
	2019	165	804	9.75	4.87	0.94	0.34
	2020	183	567	6.20	3.10	0.90	0.30
	2021	150	794	10.59	5.29	0.92	0.32
5º Bimestre	2018	140	863	12.33	6.16	0.90	0.37
	2019	165	841	10.19	5.10	0.91	0.31
	2020	183	630	6.89	3.44	0.92	0.31
6º Bimestre	2018	140	877	12.53	6.26	0.89	0.35
	2019	165	901	10.92	5.46	0.88	0.31
	2020	183	807	8.82	4.41	0.90	0.34

Assim como observado na análise das redes trimestrais, novamente detectamos diferenças entre o ano de 2020 e os demais. Podemos observar, a partir da Tabela 3.3, que o número de rotas diminuiu drasticamente nos 3º, 4º e 5º bimestres de 2020 comparada com as demais redes. Com isso, o grau médio também diminuiu nesses bimestres, consequentemente, o grau médio de entrada/saída também.

Percebemos também que a reciprocidade (R) é bem alta, mas não é próxima de 1, entretanto, a partir desse alto valor sabemos que é possível voar de um aeroporto para

outro, mas não necessariamente é possível voltar pela mesma rota exatamente como foi observado nas redes trimestrais. Verificamos, considerando agora a rede não direcionada, que a transitividade (\bar{C}) entre todas as redes é considerada alta, o que é esperado em uma rede de aeroportos.

Vamos verificar como se comporta a distribuição dos graus de entrada e saída nas Figuras 3.4, 3.5 e 3.6 das redes que mais se destacaram na análise anterior.

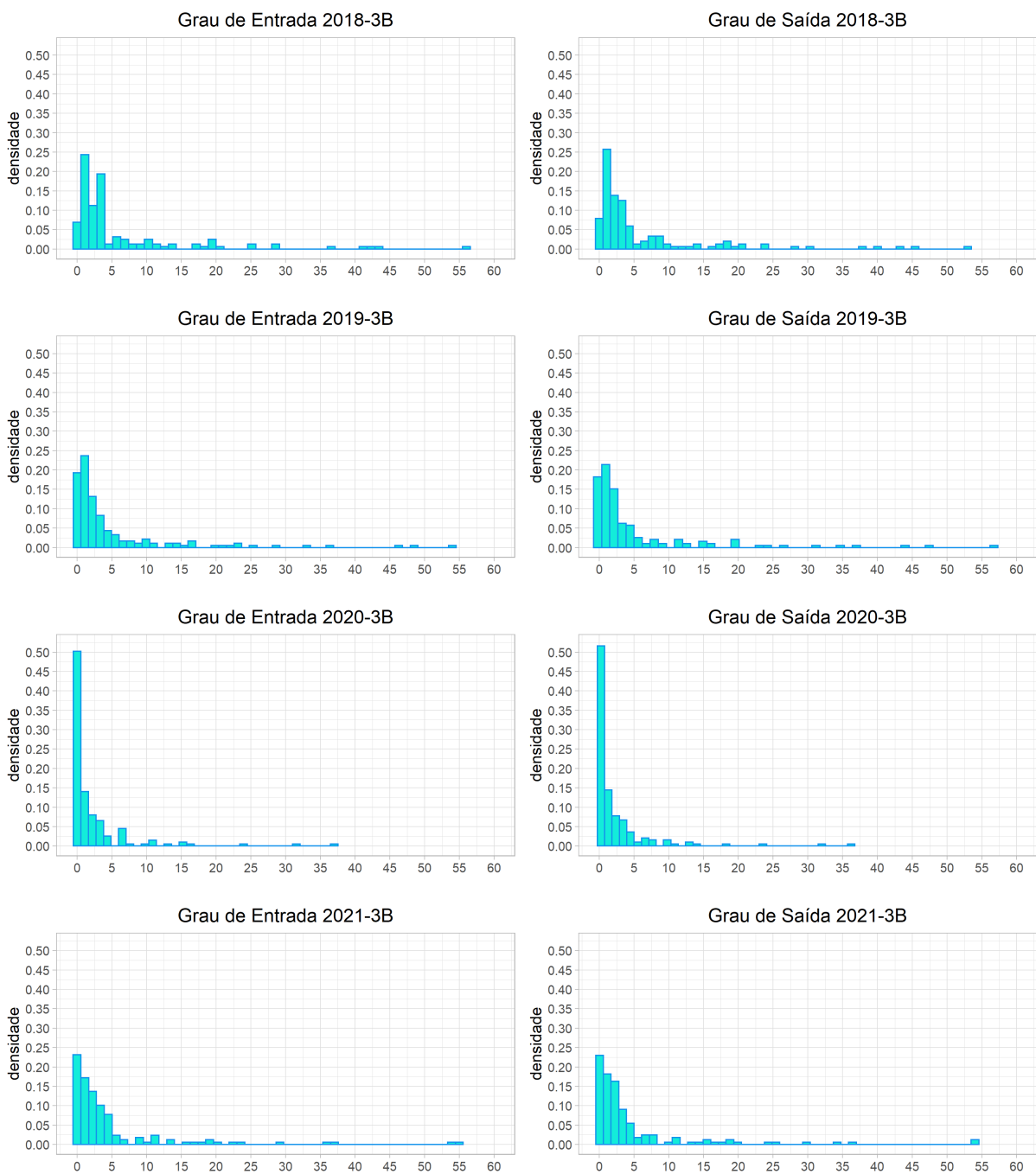


Figura 3.4: Histograma dos graus de entrada e saída referentes ao 3^o bimestre das redes sem pesos.

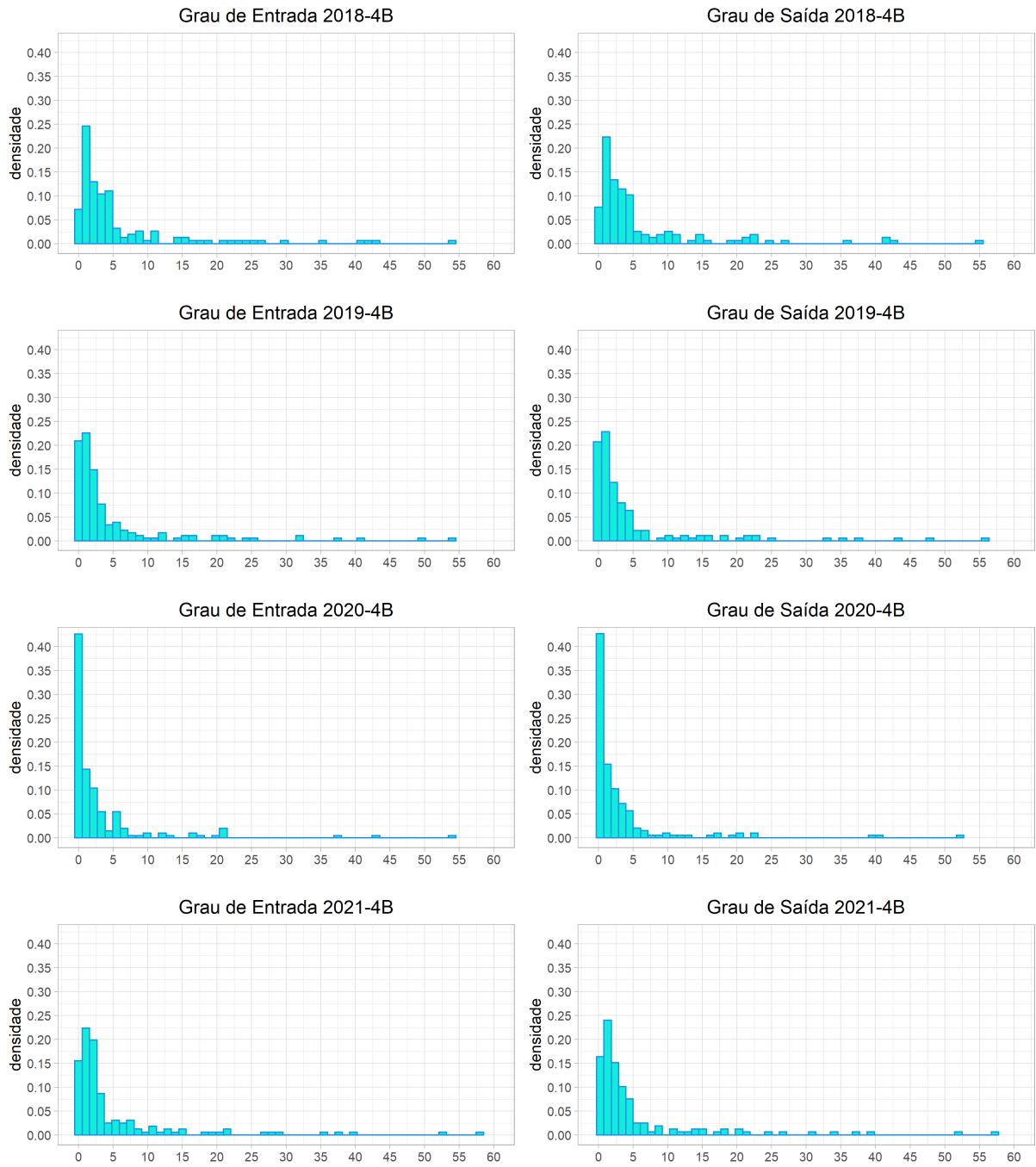


Figura 3.5: Histograma dos graus de entrada e saída referentes ao 4º bimestre das redes sem pesos.

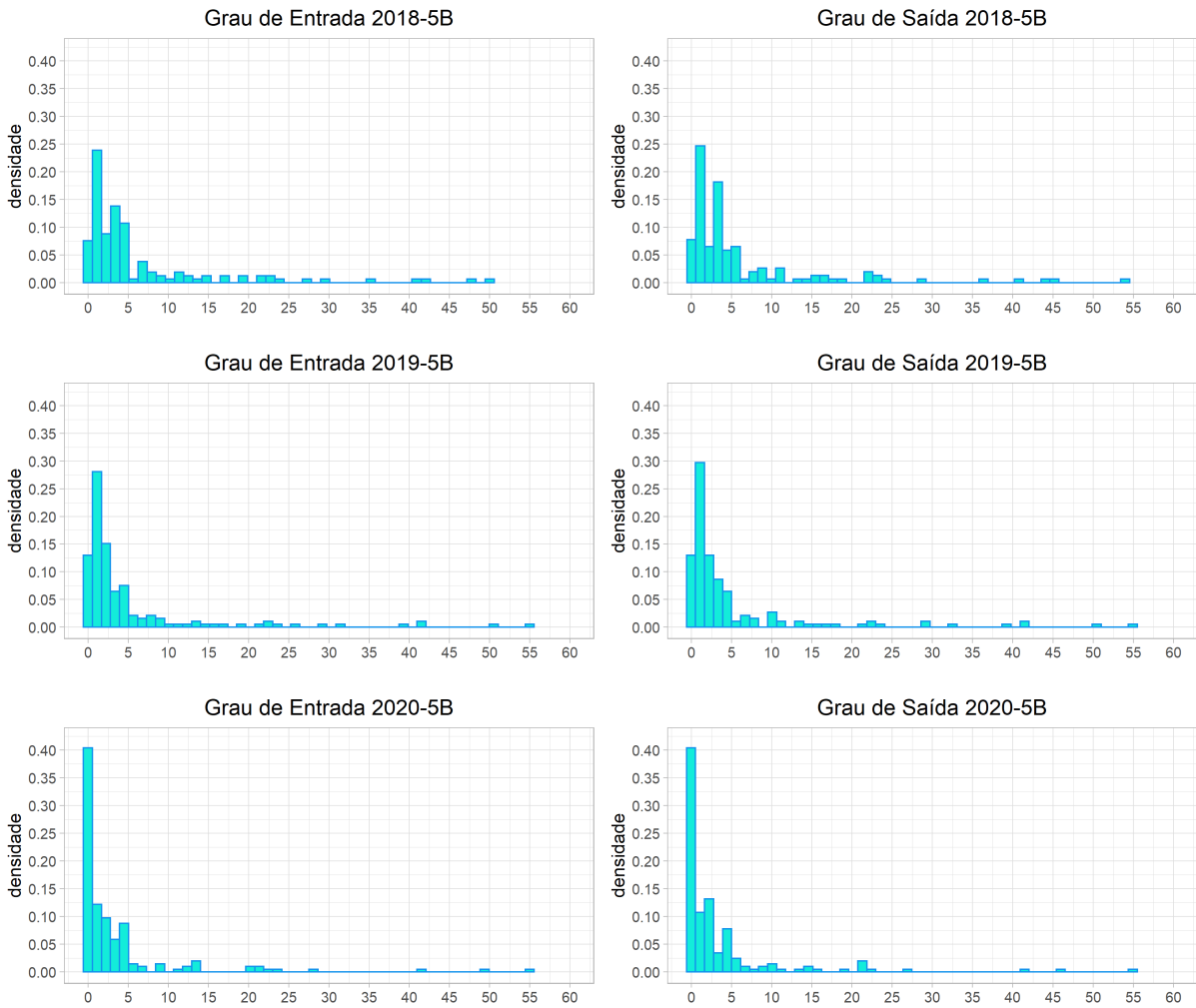


Figura 3.6: Histograma dos graus de entrada e saída referentes ao 5^o bimestre das redes sem pesos.

Assim como nas análises das redes trimestrais, observamos que todos os histogramas são assimétricos a esquerda, ou seja, são heterogêneos, em outras palavras, há muitos vértices com graus pequenos (entre 0 e 1) e poucos vértices com graus altos (em torno de 50).

Notamos também que no ano de 2020 há uma alta concentração nos graus e é interessante observar que na Figura 3.4, a amplitude dos histogramas diminuiu, dado que os vértices obtiveram graus de entrada e saída em torno de 35 (aproximadamente 35% a menos do que os graus observados nos anos anteriores.).

Entretanto, podemos analisar de uma forma ainda mais assertiva se adicionarmos pesos nas arestas, veremos mais detalhes na Subseção 3.2.2

3.2.2 Redes com pesos

Agora analisaremos as redes/grafos com pesos nas arestas. Definimos que os pesos em cada aresta serão o número de voos, ou seja, o peso será considerado como a quantidade de voos de um aeroporto A para o aeroporto B, sendo diferente do caso das redes/grafos da Subseção 3.2.1, em que eram considerados apenas se houveram ou não conexões entre os aeroportos.

Tabela 3.4: Descritivas das redes por bimestres com pesos

	Ano	N	\bar{w}
1º Bimestre	2018	140	1810.04
	2019	165	1559.03
	2020	183	1449.92
	2021	150	1119.81
2º Bimestre	2018	140	1837.09
	2019	165	1528.81
	2020	183	574.87
	2021	150	767.59
3º Bimestre	2018	140	1860.04
	2019	165	1417.32
	2020	183	156.51
	2021	150	391.40
4º Bimestre	2018	140	1963.99
	2019	165	1549.43
	2020	183	391.40
	2021	150	1260.15
5º Bimestre	2018	140	1871.73
	2019	165	1597.16
	2020	183	652.49
6º Bimestre	2018	140	1854.99
	2019	165	1626.04
	2020	183	896.44

Obviamente o N (número de aeroportos) se manteve o mesmo das análises anteriores. É muito interessante observarmos que, a partir da Tabela 3.4, fica evidente o quão precisa se torna a análise quando usamos pesos, sendo que, o grau médio do peso dado por \bar{w} no ano de 2020 é definitivamente menor comparada com os anos de 2018 e 2019, com exceção do primeiro bimestre - em que basicamente não havia sido detectado nenhum caso de *COVID-19* no Brasil - porém, ainda assim o valor foi relativamente menor que os outros anos, no qual os valores observados dos graus médios com pesos (ou seja, número médio de voos entre os aeroportos) foi quase 12 vezes menor no 3º bimestre comparado com o ano de 2018 e aproximadamente 10 vezes menor que o ano de 2019.

Vale ressaltar também que no primeiro bimestre de 2021 os graus médios com pesos estavam voltando a “normalidade”, todavia, a pandemia ainda estava longe de acabar. Entretanto, é interessante observar que no 2º e 3º bimestre o peso médio do ano de 2021 caiu novamente e voltou a crescer no 4º bimestre. Não temos os dados a partir do 5º bimestre.

Agora, limitando a abcissa dos gráficos para 500 voos para que possamos analisar melhor a distribuição dos graus de entrada/saída das redes ponderadas.

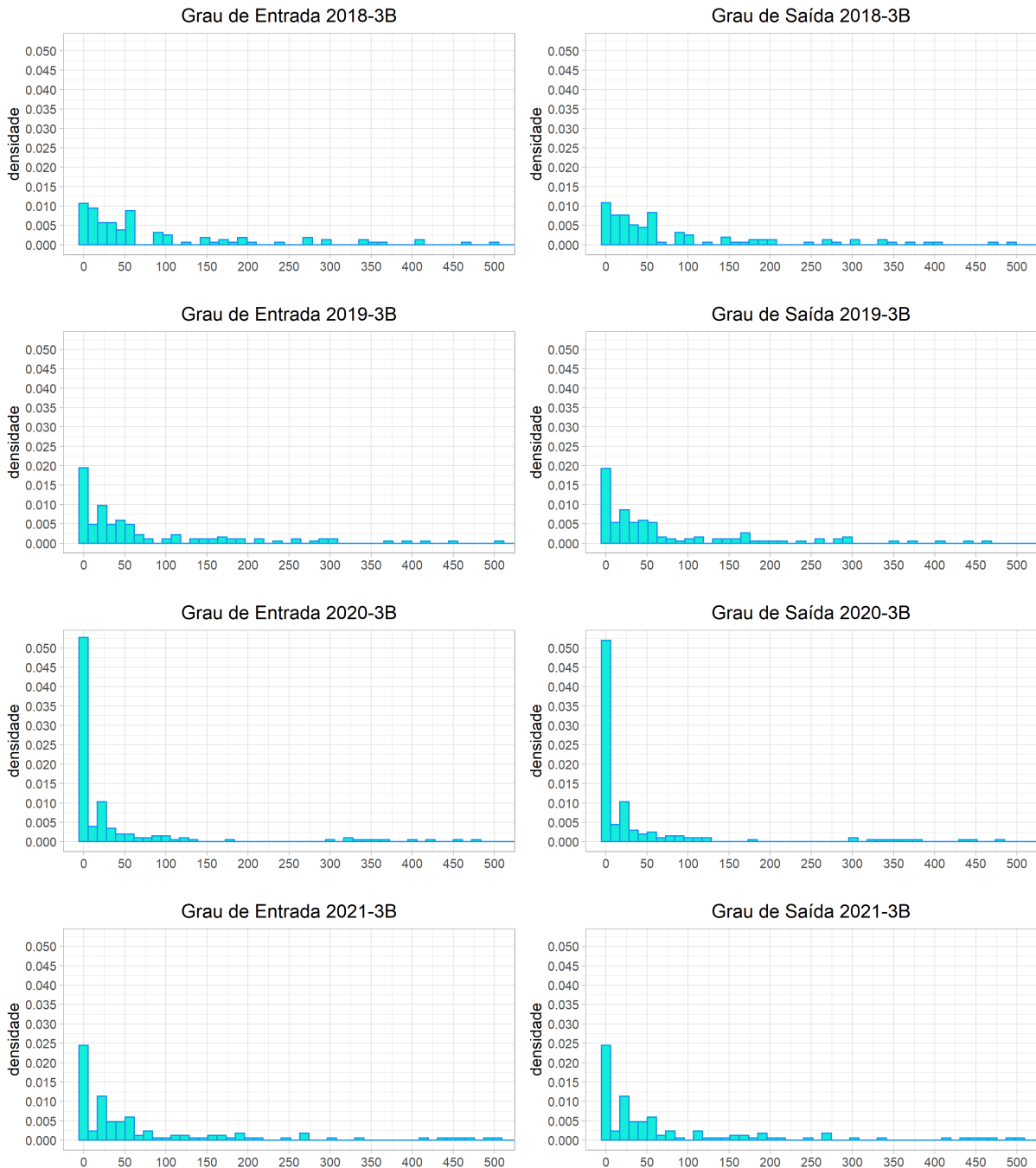


Figura 3.7: Histograma dos graus de entrada e saída referentes ao 3º bimestre das redes com pesos.

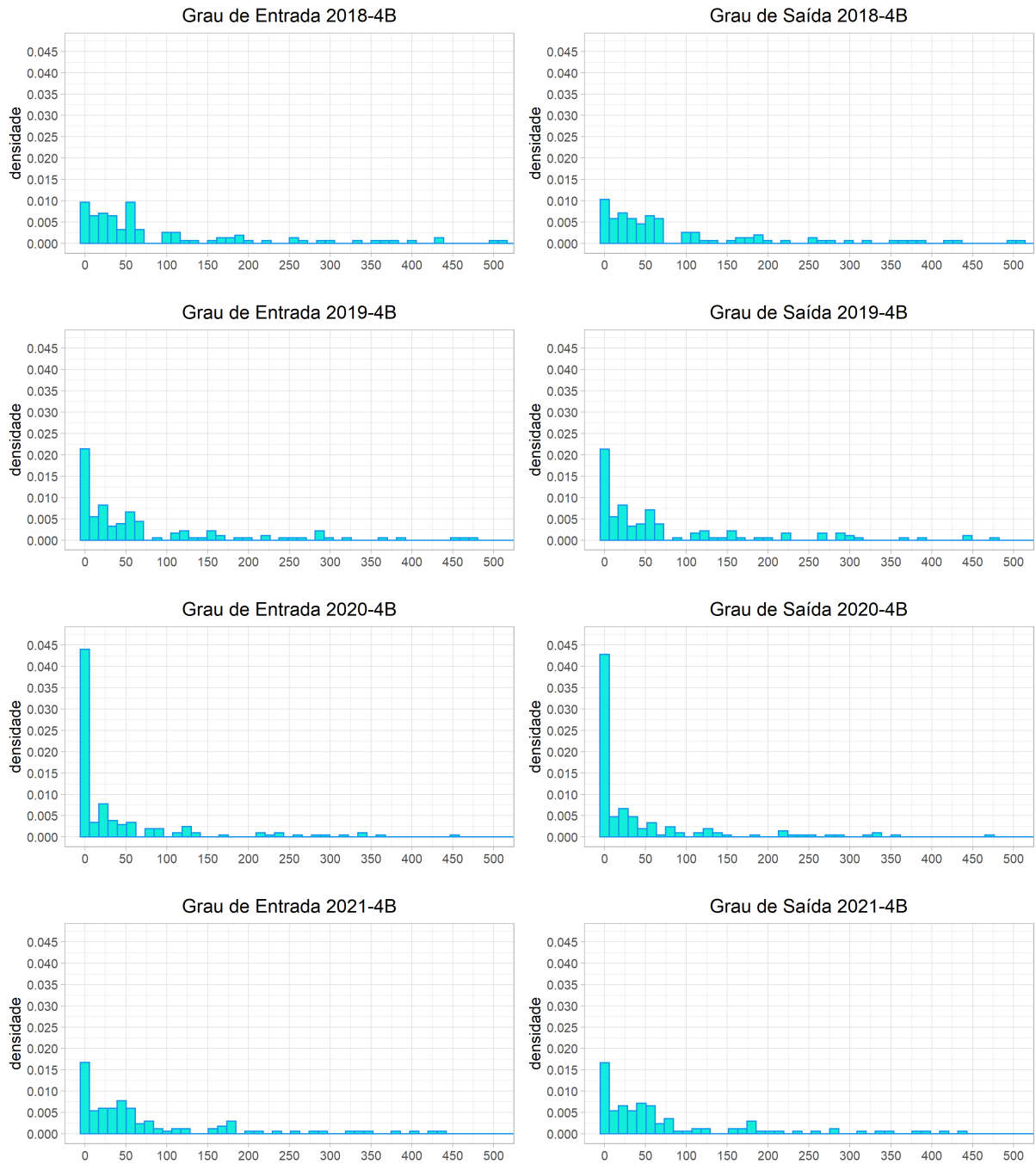


Figura 3.8: Histograma dos graus de entrada e saída referentes ao 4º bimestre das redes com pesos.

É interessante observar que nos histogramas dos graus de entrada e saída, não temos mais o número de conexões entre os aeroportos nas arestas, mas qual o peso que elas possuem, ou seja, qual o número total de voos que está entrando (ou saindo) em cada aeroporto. Como a barra do histograma em 2020 é mais alta em relação aos outros anos, significa que obtivemos mais aeroportos que tiveram até 50 voos, essa “parte” que está sobrando em 2020 foi “diluída” nos outros anos, exatamente como esperávamos.

Capítulo 4

Agrupamentos

Neste capítulo, estudaremos mais sobre métodos de agrupamentos. O objetivo desse trabalho é definir se, após a pandemia causada pelo *COVID-19*, houve diferenças nas redes de aeroportos do Brasil. Para isso, a metodologia proposta é o método de agrupamento com base na estrutura de comunidades, também conhecido como Método de Agrupamento Espectral. Veremos também algumas especificações, tais como o Laplaciano Normalizado de um grafo, que é capaz de combinar a informação do vértice e a informação da matriz de adjacência, de maneira a fornecer mais informações sobre nossas redes, e, aliado a ele, o método de Agrupamento Espectral Esférico. Por fim, será abordada uma técnica utilizada para comparação de agrupamentos, também usada como medida de similaridade, conhecida como Informação Mútua Normalizada.

4.1 Métodos de Agrupamentos

Na literatura, os métodos de agrupamentos consistem em diversas técnicas estatísticas multivariadas cujo objetivo se trata de agrupar objetos com base em alguns padrões que eles apresentam. Quando trabalhamos com redes aleatórias, uma característica fundamental é a formação de estruturas de comunidades em uma rede. Em outras palavras, deseja-se agrupar vértices que possuem características semelhantes em uma mesma comunidade. Desse modo, almejamos que haja uma alta concentração de arestas dentro de cada comunidade e poucas arestas entre as comunidades.

A estrutura de comunidades em redes está relacionada com a existência de K grupos de vértices tais que: os vértices dentro de uma comunidade estão mais conectados e os vértices que pertencem a comunidades diferentes estão menos conectados.

Assim, o objetivo consiste em, dada uma rede com estrutura de comunidades, estimar as comunidades dessa rede. Dentre a abundância de métodos de inferência de comunidades que existe na literatura, para este trabalho, foi selecionado o método espectral, que baseia-se no espectro do grafo. Escolhemos esse método, pois ele apresenta bons resultados para redes com a distribuição dos graus assimétrica, assim como visto na Subseção 3.2.1 e na Subseção 3.2.2.

4.1.1 Agrupamento Espectral

Em métodos de agrupamento, desejamos dividir um conjunto de n objetos x_1, x_2, \dots, x_n em grupos. Para isso, utilizamos uma função de similaridade de pares de objetos, denotada por S , que é simétrica e não negativa (isto é, $S(x_i, x_j) = S(x_j, x_i) \geq 0, \forall i, j = 1, \dots, n$). Para grafos, em particular, o conjunto de objetos é o conjunto de vértices do grafo e a matriz de similaridade é a própria matriz de adjacência \mathbf{A} .

Segundo Fortunato (2010), o agrupamento espectral (em inglês, *Spectral Clustering*) inclui todos os métodos e técnicas que particionam o conjunto $\{x_1, x_2, \dots, x_n\}$ em grupos a partir dos autovetores da matriz S ou de outras matrizes derivadas dela. Em outras palavras, ele consiste em uma transformação do conjunto de objetos $\{x_1, x_2, \dots, x_n\}$ em um conjunto de pontos no espaço, cujas coordenadas são elementos dos autovetores da matriz S , de modo que o conjunto de pontos é agrupado por meio de técnicas conhecidas como, por exemplo, agrupamento via *k-médias*.

De acordo com Amini *et al.* (2013) e Jin (2015), quando temos uma rede esparsa (quando os vértices da rede possuem um número pequeno de conexões, considerando o número total de nós da rede) e que possuem grande heterogeneidade nos graus do grafo, o agrupamento espectral “clássico” - que foi citado acima - não obtém os melhores resultados, cometendo falhas no agrupamento da rede. Em vista disso, Chaudhuri *et al.* (2012) propôs o **Laplaciano Normalizado** de um grafo que é definido como:

$$\mathcal{L}_{n \times n} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$$

onde $\mathbf{D}_{n \times n}$ é uma matriz diagonal dos graus do grafo (k_i) e, como já citado anteriormente, a matriz de similaridade $\mathbf{A}_{n \times n}$ é a matriz de adjacência, resultando em uma matriz simétrica

$\mathcal{L}_{n \times n}$. Logo, temos que:

$$\mathcal{L}_{n \times n} = \begin{cases} 0, & \text{se } i = j \\ \frac{1}{\sqrt{k_i k_j}}, & \text{se } A_{ij} = 1 \text{ e } i \neq j \\ 0, & \text{se } A_{ij} = 0 \text{ e } i \neq j \end{cases}$$

Uma observação interessante levantada por Shighalli e Shettar (2011) é que o Laplaciano Normalizado é a única representação que combina a informação do grau de cada vértice e a informação da matriz de adjacência.

4.1.2 Algoritmo de Agrupamento Espectral Esférico

Qin e Rohe (2013) apresentaram um método que considera a heterogeneidade dos graus da rede para determinar as comunidades, utilizando o Laplaciano Normalizado de um grafo. Para tal, o algoritmo proposto está descrito abaixo:

Entrada: Matriz \mathbf{A} , com dimensão $n \times n$, e o número de comunidades K .

Saída: $\hat{\mathbf{z}}_n$, sendo os grupos associados a cada vértice.

1. Determinar o Laplaciano Normalizado da rede: $\mathcal{L} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$;
2. Ordenar o valor absoluto dos autovalores de \mathcal{L} em ordem crescente, melhor dizendo:
 $\lambda_{(1)} \geq \lambda_{(2)} \geq \dots \geq \lambda_{(n)}$;
3. Selecionar os K maiores autovalores;
4. Definir \mathbf{U} como a matriz, $n \times K$, de modo que cada coluna de \mathbf{U} é o autovetor associado aos autovalores ordenados;
5. Regularização: Obter $\tilde{\mathbf{U}}$ normalizando cada linha de \mathbf{U} de modo que a norma do vetor \mathbf{U}_i seja igual a 1. Sendo $\tilde{\mathbf{U}}_{ij} = \frac{\mathbf{U}_{ij}}{(\sum_{j=1}^K \mathbf{U}_{ij}^2)^{\frac{1}{2}}}$;
6. Considerar cada linha de $\tilde{\mathbf{U}}$, $\tilde{\mathbf{U}}_i$, $1 \leq i \leq n$, como um ponto em \mathbb{R}^k e aplicar o método *k-médias* em $\tilde{\mathbf{U}}$ para obter $\hat{\mathbf{z}}_n$ (grupos associados a cada vértice i).

4.2 Comparações de agrupamentos

Neste trabalho, usaremos um método muito utilizado na comunidade estatística para a comparação entre grupos que foram obtidos através de métodos de agrupamentos. Este método é realizado a partir de uma medida conhecida como Informação Mútua Normalizada.

4.2.1 Informação Mútua Normalizada

A Informação mútua normalizada (em inglês: *Normalized Mutual Information* - NMI) utiliza a matriz de confusão \mathbf{R} entre dois agrupamentos “ \mathbf{c} ” e “ \mathbf{e} ”, com dimensão $K \times K$ e entrada R_{ka} , que é dada por

$$R_{ka} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(e_i = k, c_i = a)$$

onde $i = 1, \dots, n$, $a = 1, \dots, K$ e $k = 1, \dots, K$, sendo K o número de grupos definido nos agrupamentos. Logo, cada entrada R_{ka} é a proporção de vezes que um elemento foi alocado no grupo k do agrupamento “ \mathbf{e} ” e foi alocado no grupo a do agrupamento “ \mathbf{c} ”.

Podemos pensar na matriz de confusão \mathbf{R} como uma distribuição de probabilidade bivariada, em que as somas das linhas e colunas R_{k+} e R_{+a} são as probabilidades marginais. Então, a NMI, que foi definida por Yao (2003), é dada por

$$NMI(\mathbf{c}, \mathbf{e}) = - \sum_{k,a=1}^K R_{ka} \log \frac{R_{ka}}{R_{k+}R_{+a}} \sum_{k,a=1}^K (R_{ka} \log(R_{ka}))^{-1}.$$

A informação mútua normalizada é um valor observado entre 0 e 1, sendo que 1 indica que os agrupamentos são iguais. Uma vez que essa medida é normalizada, podemos medi-la e compará-la entre diferentes agrupamentos com quantidades diferente de grupos. Amini *et al.* (2013) apresentam alguns valores de referência da NMI, por exemplo, considerando n grande, se temos que 50%, 70% e 90% dos rótulos correspondentes aos mesmos elementos são iguais, os valores da NMI são de aproximadamente 0.12, 0.26 e 0.58, respectivamente. Portanto, como a informação mútua normalizada demonstrou ser uma medida eficaz para determinar a qualidade dos agrupamentos, ela será utilizada no Capítulo 5.

Capítulo 5

Resultados dos Agrupamentos

Nesse capítulo, será utilizada uma nova estrutura dos dados, para que possamos verificar qual o comportamento dos agrupamentos das redes aéreas brasileiras de acordo com o método de agrupamento abordado no Capítulo 4.

5.1 Estrutura dos dados

Como discutido no Capítulo 4, propomos utilizar o método de Agrupamento Espectral Esférico para dividir os nós da rede em comunidades, ou seja, os vértices - aeroportos - da rede são divididos de modo que, dentro de cada grupo, a probabilidade de conexão entre os vértices sejam maiores do que entre nós de grupos distintos.

Para detectar o número de agrupamentos, o primeiro passo é fixar $K = 5$ comunidades. Definiu-se que este seria um bom número ao considerarmos que o Brasil é dividido em 5 regiões geográficas. Vale ressaltar que poderíamos ter abordado diferentes métodos para estimação do número de comunidades. Contudo, esse não é o escopo deste trabalho.

Com o objetivo de comparar os agrupamentos, fixamos todos os aeroportos que apareceram nos anos de 2018, 2019, 2020 e 2021, de modo que $n = 218$ aeroportos (vértices do grafo fixos) foram detectados. Ressaltamos que, para esse capítulo do trabalho, a matriz de adjacência (\mathbf{A}) da rede é sem pesos e não direcionada, em virtude das particularidades e da aplicabilidade do algoritmo proposto para as análises da rede.

5.2 Resultados: Redes sem pesos e sem direção

Nesta seção, começaremos verificando, de maneira visual, o comportamento dos grupos formados sobre o banco de dados das redes aeroportuárias. Para uma visualização mais eficiente, cada comunidade será representada por cores diferentes.

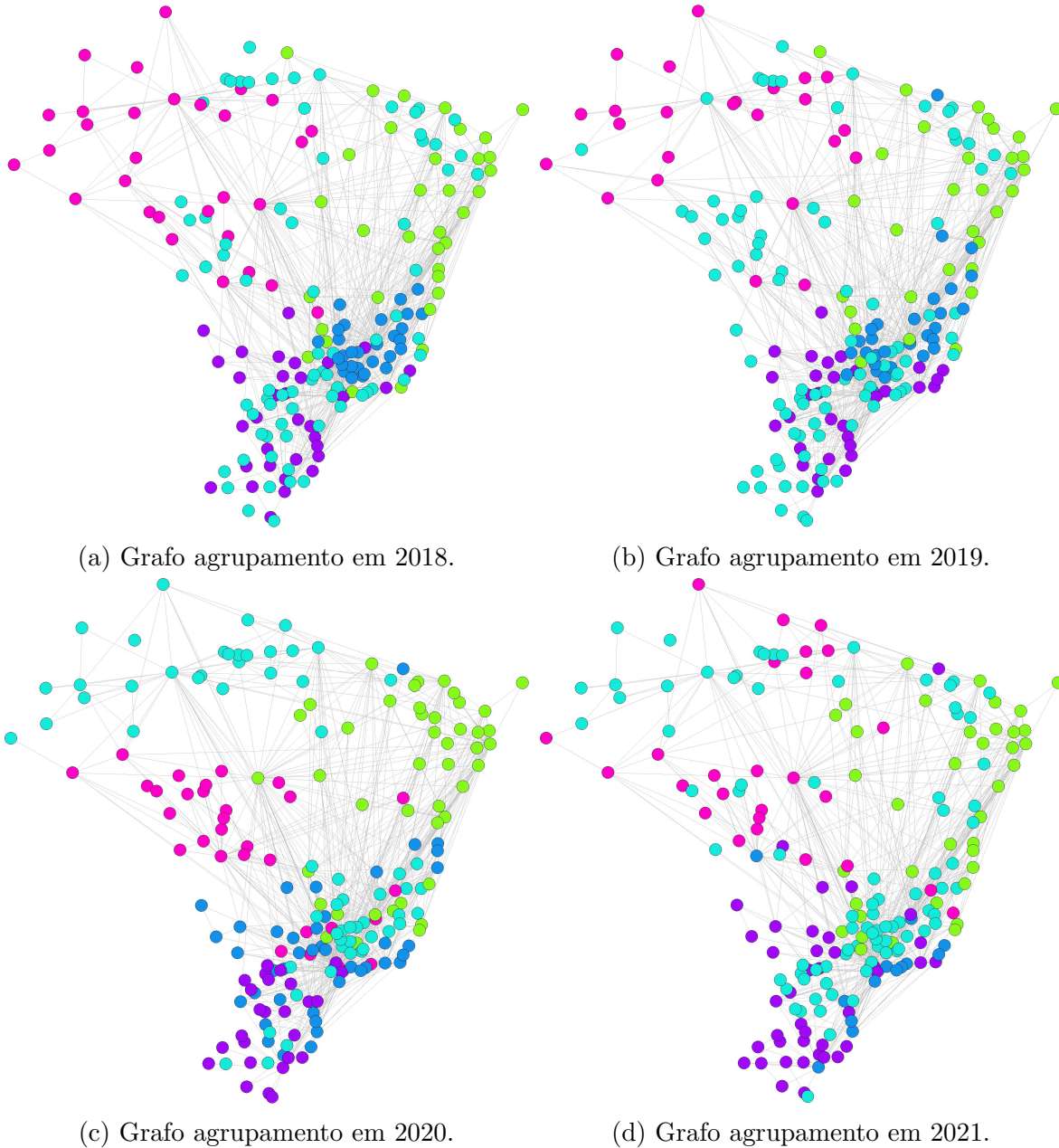


Figura 5.1: Agrupamento das redes aeroportuárias em cinco grupos ($k = 5$) nos anos de 2018, 2019, 2020 e 2021.

Observamos, a partir da Figura 5.1, que as redes em 2018 (Figura 5.1a) e 2019 (Figura 5.1b) apresentam muitas similaridades, especialmente ao observar os grupos formados pela cores rosa, roxo, verde-ciano e verde-lima. O único grupo destoante é formado pela cor azul, que ano de 2018 permaneceu somente na região sudeste. Contudo, apesar de, em

2019, esse grupo ainda apresentasse grande concentração na região sudeste, ele também abrangeu alguns aeroportos da região nordeste, sendo essa a diferença mais chamativa entre a distribuição dos agrupamentos entre 2018 e 2019.

Em relação ao ano de 2020 (Figura 5.1c) observamos que a predominância da região norte se dá pelo grupo formado pela cor verde-ciano, o que não ocorre nos anos de 2018 e 2019, cujo domínio se dá pelo grupo formado pela cor rosa. O inverso aconteceu com a região centro-oeste, em que prevaleceu o grupo representado pela cor rosa em 2020 e, nos anos anteriores, sua maior parte era representada pelo grupo verde-ciano, além da mescla dos grupos representados pelas cores verde-lima, rosa e roxo. Observamos também que a região nordeste é majoritariamente representada pela cor verde-lima em todos os anos (2018, 2019, 2020 e 2021). Já a região sudeste, em 2018 e 2019 era mais representada pela cor azul mas, em 2020 e 2021, a representação ficou em sua grande parte representada pela cor verde-ciano. A região sul, em todos os anos, obteve uma grande concentração do grupo representado pela cor roxa. Entretanto, em 2018 e 2019, seu domínio se dava pelo grupo da cor verde-ciano e, em 2020 e 2021, sua predominância era o grupo da cor roxa. Em outras palavras, observamos que o grupo formado pela cor verde-ciano nos anos de 2018, 2019, 2020 e 2021 tem uma mescla de todas as regiões do Brasil. Ainda, percebe-se que, em especial no ano de 2020, ele possui uma alta concentração nas regiões norte e sudeste.

É interessante observarmos que, na região norte de 2021 (Figura 5.1d), ainda há uma predominância do grupo formado pela cor verde-ciano (assim como no ano de 2020), mas também que o grupo formado pela rosa volta a abranger mais alguns vértices dessa região. A região nordeste continua com maior predominância do grupo formado pela cor verde-lima mas, dessa vez, observamos que o grupo formado pela cor roxa - que em todos os anos foi representado majoritariamente pela região sul - tem um único vértice na região nordeste. Notamos também que a região sudeste é bem representada pelos grupos formados pela cor verde-ciano, roxo e verde lima, assim como alguns vértices da cor azul - sendo que, nesse ano, o grupo foi formado por uma quantidade menor de aeroportos comparado com os demais anos. Por fim, a região sul, assim como nos anos de 2018 e 2019, foi bem representada pelo grupo formado pela cor roxa.

Através dessa descrição visual, acreditamos que, de fato, houve diferenças entre os agrupamentos e as redes de aeroportos no Brasil. Entretanto, tal análise é difícil e maçante de ser realizada de maneira gráfica. Para isso, faremos uma tabela com algumas medidas

descritivas para facilitar o entendimento acerca das interpretações.

Calcularemos algumas medidas descritivas, sendo elas: Número de aeroportos na formação de cada grupo (N^o), Grau médio \bar{k} (dentro de cada grupo) e o número de aeroportos presente em cada região do país, sendo elas: Norte (**N**), Nordeste (**NE**), Centro-Oeste (**CO**), Sudeste (**SE**), e Sul (**S**).

Tabela 5.1: Descritivas dos agrupamentos da rede sem pesos e sem direção

Ano	Grupo	Cores	N^o	\bar{k}	Regiões				
					N	NE	CO	SE	S
2018	1	verde-ciano	83	0.07	13	12	14	20	24
	2	verde-lima	37	4.97	2	27	1	7	0
	3	rosa	34	4.53	25	0	8	1	0
	4	roxo	35	5.54	0	0	6	9	20
	5	azul	29	3.66	0	0	0	29	0
2019	1	verde-ciano	92	0.35	13	10	20	21	28
	2	verde-lima	33	7.64	2	24	2	5	0
	3	rosa	30	5.13	25	0	3	2	0
	4	roxo	33	6.30	0	0	4	13	16
	5	azul	30	2.60	0	5	0	25	0
2020	1	verde-ciano	65	1.51	31	1	1	26	6
	2	verde-lima	46	6.78	4	31	2	9	0
	3	rosa	32	3.12	5	0	19	8	0
	4	roxo	28	2.71	0	0	0	4	24
	5	azul	47	1.15	0	7	7	19	14
2021	1	verde-ciano	91	0.42	21	11	6	36	17
	2	verde-lima	42	4.95	5	26	1	10	0
	3	rosa	28	2.21	14	1	12	1	0
	4	roxo	43	1.63	0	1	9	11	22
	5	azul	14	1.0	0	0	1	8	5

Assim como na Figura 5.1, nota-se a partir da Tabela 5.1 que o grupo 1, representado pela cor verde-ciano, abrangeu todas as regiões do Brasil em todos os anos, com uma maior concentração na região sul, sudeste e centro-oeste nos anos de 2018 e 2019; já em 2020,

sua maior concentração foi nas regiões norte e sul e, em 2021, seu domínio foi composto pelas regiões norte, sudeste e sul. Seu grau médio \bar{k} foi o menor em todos os anos pois, em geral, os métodos de agrupamentos tendem a agrupar vértices que tem características em comum. Nesse caso, a comunidade detectada foram os aeroportos que receberam menos voos (arestas). Em outras palavras, esse grupo é formado pelos aeroportos que possuem menos conexões entre si.

O grupo 2, representado pela cor verde-lima, teve sua maior concentração na região nordeste e, em todos os anos, não obteve nenhum aeroporto da região sul. O grupo 3, formado pela cor rosa, foi representado pela região norte em 2018 e 2019, e passou a ser representado pela região centro-oeste em 2020, sendo que ficou mesclado entre as regiões norte e centro-oeste em 2021. Já o grupo 4, representado pela cor roxa, sempre se dispôs majoritariamente na região sul, mas obtinha alguns aeroportos da região sudeste e centro-oeste; a partir de 2020, agrupou somente aeroportos nas regiões sul e sudeste e, em 2021, obteve um aeroporto na região nordeste.

Por fim, o grupo 5, representado pela cor azul, obteve apenas aeroportos da região sudeste em 2018, agregou mais alguns aeroportos da região nordeste em 2019, expandiu para todas as regiões, com exceção à região norte, em 2020 e agrupou somente aeroportos das regiões centro-oeste, sudeste e sul em 2021 sendo que, nesse ano em especial, esse foi o grupo que foi agrupado com muito menos aeroportos em comparação com os anos anteriores.

Na Tabela 5.2, são apresentados os valores da medida dois-a-dois da Informação Mútua Normalizada em cada bimestre e nos anos de 2018, 2019, 2020 e 2021.

Tabela 5.2: Informação mútua normalizada

	18-19	18-20	18-21	19-20	19-21	20-21
1° Bimestre	0.330	0.211	0.177	0.202	0.128	0.153
2° Bimestre	0.307	0.197	0.251	0.206	0.210	0.160
3° Bimestre	0.291	0.106	0.206	0.105	0.228	0.152
4° Bimestre	0.327	0.171	0.186	0.138	0.198	0.193
5° Bimestre	0.234	0.170	-	0.147	-	-
6° Bimestre	0.168	0.150	-	0.089	-	-
Anual	0.344	0.151	0.155	0.167	0.149	0.167

Através da Tabela 5.2, notamos que os menores valores observados da NMI são referentes aos anos de 2020 e 2021. Reitera-se que, se o valor for próximo de 1, os agrupamentos das redes são considerados iguais. Ao observar a NMI anual de 2018 e 2019, encontrou-se o valor de 0.344, que significa que a rede obteve aproximadamente 70% dos rótulos combinados. Já nas redes anuais dos anos combinados entre 2020 e 2021, esse valor passou de 70 para aproximadamente 50% dos rótulos combinados (dado que, quando o valor observado da NMI for 0.12, indica que as redes estão combinadas em aproximadamente 50% dos rótulos).

É interessante perceber como, em 2021, as redes foram voltando ao “normal” até o 3º bimestre, período em que se iniciaram as vacinações em massa. Porém, no 4º bimestre, elas voltaram a diminuir o valor de NMI, indicando que o agrupamento volta a ser diferente, dado que uma informação relevante é que, no final de junho, foi registrado o primeiro caso da variante delta em São Paulo - SP.

Notamos também que a rede de 2020 e 2021 gerou um valor de NMI anual de 0.167 que, assim como os valores bimestrais, é considerado baixo, sendo mais uma evidência de que o agrupamento dos anos de 2020 e 2021 são diferentes entre si.

Agora, vamos observar as redes que mais se destacaram na Tabela 5.2, que foram referentes ao 6º bimestre dos anos de 2019 e 2020.

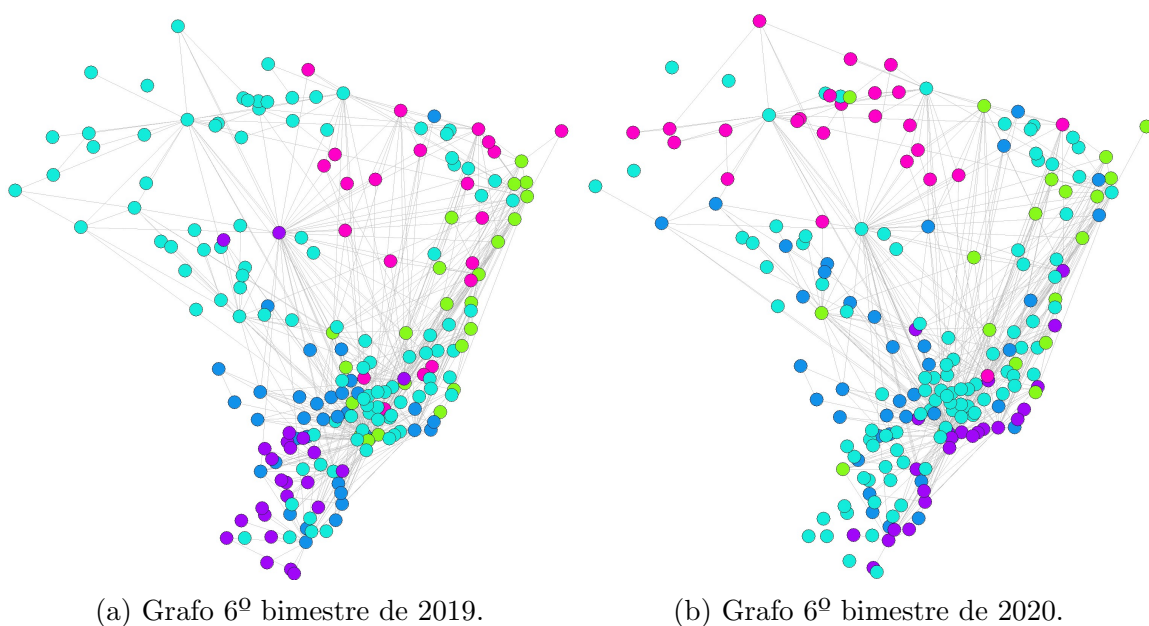


Figura 5.2: Agrupamento das redes aeroportuárias com 5 grupos ($k = 5$) dos 6º bimestres dos anos 2019 e 2020

Tabela 5.3: Descritivas dos agrupamentos das redes do 6º bimestre de 2019 e 2020

	Grupo	Cores	Nº	\bar{k}	Regiões				
					N	NE	CO	SE	S
6º bimestre de 2019	1	verde-ciano	110	0.89	34	12	17	36	11
	2	verde-lima	24	5.92	0	14	1	9	0
	3	rosa	21	0.76	6	11	0	4	0
	4	roxo	27	0.29	0	0	4	1	22
	5	azul	36	1.28	0	2	7	16	11
6º bimestre de 2020	1	verde-ciano	105	0.21	12	15	12	39	27
	2	verde-lima	22	0.27	3	14	1	3	1
	3	rosa	27	1.26	22	2	1	2	0
	4	roxo	24	4.58	0	2	1	13	8
	5	azul	40	0.55	3	6	14	9	8

Analisando o 6º bimestre por regiões do Brasil, através da Figura 5.2 e Tabela 5.3, notamos que o grupo 1, representado pela cor verde-ciano, abrangeu todas as regiões do Brasil em 2019 e 2020, mas em 2019 sua maior concentração era na região norte e sudeste; entretanto, em 2020 sua maior concentração foi nas regiões sudeste e sul. Na região norte, observamos a alta concentração de aeroportos formados pelo grupo 1 (verde-ciano) no ano de 2019, entretanto, em 2020, observamos que há uma grande mescla de grupos que representam essa região, sendo formada também pelos grupos 3, 5 e 2 (rosa, azul e verde-lima).

A região nordeste se manteve bem representada pelo grupo 2, formado pela cor verde-lima, porém, em 2019 havia uma boa mescla entre os grupos 1, 2 e 3, formado pelas cores verde-ciano, verde-lima e rosa, respectivamente; já em 2020 ela foi bem representada pela cor verde-ciano e verde-lima com 15 e 14 vértices (aeroportos) cada, mas observamos que há alguns vértices dos grupos 3, 4 e 5 também. A região centro-oeste, foi bem representada pelo grupo 1 (verde-ciano) em 2019 e em 2020 pelos grupos 5 e 1, sendo representados pelas cores verde-ciano e azul, respectivamente. A região sudeste, assim como as outras, foi bem representada pelo grupo 1 nos dois anos, porém em 2019 ela também foi representada pelo grupo 5 (azul) e em 2020 pelo grupo 4 (roxo).

O grau médio \bar{k} da rede do 6º bimestre de 2019 dos grupos 1, 3, 4 e 5 foi um grau baixo, assim como dito anteriormente, em geral, os métodos de agrupamentos tendem

a agrupar vértices que tem características em comum, nesses casos, as comunidades que foram detectadas foram os aeroportos que receberam menos voos (arestas). O mesmo aconteceu com os grupos 1, 2, 3 e 5 formados pela rede do 6^o bimestre de 2020 que obtiveram um grau médio considerado baixo. Em contra partida, notamos que o grupo 2, formado pela cor verde-lima, obteve um grau médio de 5.92 em 2019, sendo que esse grupo representa bem a região nordeste e obtém alguns aeroportos na região sudeste e 1 aeroporto da região centro-oeste. Já em 2021, o grupo 4, formado pela cor roxa, obteve grau médio de 4.58 e esse grupo é formado majoritariamente pelas regiões sudeste e sul, porém também é composto por 2 aeroportos da região nordeste e 1 da região centro-oeste, o grau médio mais elevado representa que esses agrupamentos receberam mais conexões entre si.

Capítulo 6

Considerações Finais

O foco principal desse trabalho consistia em estudar a evolução das redes aeroportuárias do Brasil, a fim de analisar seu comportamento e verificar se houve mudanças após o início da pandemia. Inicialmente, no Capítulo 2, foi realizada uma revisão bibliográfica, onde foram introduzidos conceitos, medidas e definições de redes.

Logo após, no Capítulo 3, representamos os dados de voos nacionais e regulares por meio de redes/grafos para cada ano a ser estudado. Com as redes construídas, fizemos uma análise descritiva para entender melhor a estrutura da rede de aeroportos, isto é, analisar o número de nós da rede (número de aeroportos), número de arestas, grau de entrada e saída, transitividade e reciprocidade para descrever a rede de aeroportos.

No Capítulo 4, com base nos resultados apresentamos, pudemos definir qual o melhor método de agrupamento a ser utilizado, sabendo que a estrutura de comunidades em redes está relacionada com a existência de K grupos de vértices tal que: os vértices dentro de uma comunidade estão mais conectados e os vértices que pertencem a comunidades diferentes estão menos conectados, para essa problemática, foi apresentado o algoritmo de agrupamento espectral esférico, proposto por Qin e Rohe. Com base nesse algoritmo, no Capítulo 5, podemos cumprir nosso objetivo, que era estimar quais comunidades da rede de transporte aéreo brasileira seriam detectadas e verificar se houvessem diferenças nessas redes.

Tanto o Capítulo 3 (Análise descritiva e exploratória) quanto o Capítulo 5 (Resultados dos agrupamentos), sugerem que, a partir de nossas análises, houve algum tipo de reestruturação da rede no ano de 2020, por consequência da pandemia causada pelo *COVID-19*, no entanto, aparentemente, ela parece estar em 2021 voltando a configuração das redes anteriores (2018 e 2019).

Claro que, para que possamos fazer uma análise estatística que, de fato, comprove que a rede de 2020 é diferente das demais, poderíamos utilizar algum outro método estatístico como, por exemplo, um teste de hipóteses para redes. Além disso, ao invés de utilizar os dados mensais, talvez seria interessante utilizar os dados diários dos voos entre os aeroportos e/ou usar a rede aeroportuária mundial. Por fim, também seria interessante a utilização de algum método que considere o peso como, por exemplo, o número de voos e/ou número de passageiros para obter as comunidades da rede.

Apêndice A

Códigos Utilizados - *Software R*

```
#####  
##-- EXEMPLOS DO TEXTO - GRAFOS --##  
#####  
library(igraph)  
set.seed(10)  
  
#####  
## grafo simples ##  
#####  
m1 = matrix(c (0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,  
              1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0,  
              0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1,  
              0, 0, 0, 0, 0, 1, 0), ncol = 7)  
grafo = graph_from_adjacency_matrix(m1,mode = "undirected")  
  
#Cor do vértice e da aresta  
V(grafo)$color = "#13ECDA"  
E(grafo)$color = "grey28"  
#Tamanho do vértice e da aresta  
V(grafo)$size = 40  
E(grafo)$size = 50  
#Tamanho e cor da letra dentro da aresta  
V(grafo)$label.cex = 4  
V(grafo)$label.color = "gray10"  
l = layout_nicely(grafo)  
plot(grafo, layout = l, vertex.frame.color = "gray28")
```

```

#####
## grafo direcionado ##
#####
m2 = matrix(c (0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
              0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0,
              0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1,
              0, 0, 0, 0, 0, 0, 0, 0), ncol = 7)
grafo2 = graph_from_adjacency_matrix(m2,mode = "directed")

V(grafo2)$color = "#13ECDA"
E(grafo2)$color = "grey28"
V(grafo2)$size = 40
E(grafo2)$size = 50
V(grafo2)$label.cex = 4
V(grafo2)$label.color = "gray10"
plot(grafo2, layout = 1, vertex.frame.color = "gray28",edge.arrow.size = 1.2)

#####
##-- EXEMPLO DO TEXTO - GRAUS --##
#####

#####
## grafo simples ##
#####
m1 = matrix(c (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
              1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0,
              0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0,
              0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),
              ncol = 8)
grafo = graph_from_adjacency_matrix(m1,mode = "undirected")

V(grafo)$color = "#13ECDA"
E(grafo)$color = "grey28"
V(grafo)$size = 50
E(grafo)$size = 50
V(grafo)$name = c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H')
V(grafo)$label.cex = 4
V(grafo)$label.color = "gray10"
l = layout_as_star(grafo)
plot(grafo, layout = 1, vertex.frame.color = "gray28")

```

```
#####
## grafo direcionado ##
#####
m2 = matrix(c (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
              0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
              0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0,
              0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), ncol = 8)
grafo2 = graph_from_adjacency_matrix(m2,mode = "directed")

V(grafo2)$color = "#13ECDA"
E(grafo2)$color = "grey28"
V(grafo2)$size = 50
E(grafo2)$size = 50
V(grafo2)$name = c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H')
V(grafo2)$label.cex = 4
V(grafo2)$label.color = "gray10"
plot(grafo2, layout = 1, vertex.frame.color = "gray28", edge.arrow.size = 1)
```

```
#####
#####
##----- LIMPEZA DOS DADOS DA ANAC -----##
#####
#####
```

```
#####
##-- PACOTES NECESSÁRIOS --##
#####
library(readr)
library(dplyr)

dados18 = read_delim("Dados/resumo_anual_2018.csv", ";", escape_double = FALSE,
                    locale = locale(encoding = "latin1"),
                    trim_ws = TRUE) %>%
  select(4:9, 12:15, 18:21,30,31,35) %>%
  filter(NATUREZA == "DOMÉSTICA")

a <- is.na(dados18[])
dados18[][a] = 0
```

```

datos19 = read_delim("Datos/resumo_anual_2019.csv", ";", escape_double = FALSE,
                    locale = locale(encoding = "latin1"),
                    trim_ws = TRUE) %>%
  select(4:9, 12:15, 18:21,30,31,35) %>%
  filter(NATUREZA == "DOMÉSTICA")
a <- is.na(datos19[])
datos19[][a] = 0
datos20 = read_delim("Datos/resumo_anual_2020.csv", ";", escape_double = FALSE,
                    locale = locale(encoding = "latin1"),
                    trim_ws = TRUE) %>%
  select(4:9, 12:15, 18:21,30,31,35) %>%
  filter(NATUREZA == "DOMÉSTICA")
a <- is.na(datos20[])
datos20[][a] = 0
datos21 = read_delim("Datos/resumo_anual_2021.csv", ";", escape_double = FALSE,
                    locale = locale(encoding = "latin1"),
                    trim_ws = TRUE) %>%
  select(4:9, 12:15, 18:21,30,31,35) %>%
  filter(NATUREZA == "DOMÉSTICA")
a = is.na(datos21[])
datos21[][a] <- 0

datos18 = subset(datos18,datos18$'GRUPO DE VOO' == "REGULAR")
datos19 = subset(datos19,datos19$'GRUPO DE VOO' == "REGULAR")
datos20 = subset(datos20,datos20$'GRUPO DE VOO' == "REGULAR")
datos21 = subset(datos21,datos21$'GRUPO DE VOO' == "REGULAR")

#####
##-- SALVANDO OS DADOS --##
#####
write.table(datos18, file = "datos_domestic_Regular18.txt")
datos18 = read.table("datos_domestic_Regular18.txt")
write.table(datos19, file = "datos_domestic_Regular19.txt")
datos19 = read.table("datos_domestic_Regular19.txt")
write.table(datos20, file = "datos_domestic_Regular20.txt")
datos20 = read.table("datos_domestic_Regular20.txt")
write.table(datos21, file = "datos_domestic_Regular21.txt")
datos21 = read.table("datos_domestic_Regular21.txt")

```



```
#####
##-- OBTENDO OS SIGLAS DOS AEROPORTOS --##
#####

# coluna 3: Sigla aeroporto origem
# coluna 7: sigla aeroporto destino

list18 <- names(table( c(names(table(dados18$AEROPORTO.DE.ORIGEM..SIGLA.)),
                           names(table(dados18$AEROPORTO.DE.DESTINO..SIGLA.))))))
list19 <- names(table( c(names(table(dados19$AEROPORTO.DE.ORIGEM..SIGLA.)),
                           names(table(dados19$AEROPORTO.DE.DESTINO..SIGLA.))))))
list20 <- names(table( c(names(table(dados20$AEROPORTO.DE.ORIGEM..SIGLA.)),
                           names(table(dados20$AEROPORTO.DE.DESTINO..SIGLA.))))))
list21 <- names(table( c(names(table(dados21$AEROPORTO.DE.ORIGEM..SIGLA.)),
                           names(table(dados21$AEROPORTO.DE.DESTINO..SIGLA.))))))

# número de aeroportos
length(list18)
length(list19)
length(list20)
length(list21)

#####
##-- SALVANDO AS SIGLAS DOS AEROPORTOS POR ANO --##
#####

write.table(list18 , file = "Siglas_Aeroportos18.txt")
write.table(list19 , file = "Siglas_Aeroportos19.txt")
write.table(list20 , file = "Siglas_Aeroportos20.txt")
write.table(list21 , file = "Siglas_Aeroportos21.txt")

#####
##-- OBTENDO OS NOMES DOS AEROPORTOS PARA CADA ANO USANDO AS SIGLAS --##
#####

# Selecionando os nomes das cidades desses aeroportos
# coluna 3: Sigla aeroporto origem
# coluna 4: aeroporto origem nome
# coluna 5: UF aeroporto origem
# coluna 7: sigla aeroporto destino
# coluna 8: aeroporto destino nome
# coluna 9: aeroporto destino UF
```

```
#2018
UF <- vector()
cidades <- vector()
for(i in 1:length(list18)){
  j <- which(dados18[,3]==list18[i])
  if(length(j)>=1){
    cidades[i] <- dados18[j[1],4]
    UF[i] <- dados18[j[1],5]
  }
  j <- which(dados18[,7]==list18[i])
  if(length(j)>=1){
    cidades[i] <- dados18[j[1],8]
    UF[i] <- dados18[j[1],9]
  }
}
posicao <- paste(cidades, UF, "Brazil")
write.table(posicao,"nomes_cidades18.txt")
```

```
#2019
UF <- vector()
cidades <- vector()
for(i in 1:length(list19)){
  j <- which(dados19[,3]==list19[i])
  if(length(j)>=1){
    cidades[i] <- dados19[j[1],4]
    UF[i] <- dados19[j[1],5]
  }
  j <- which(dados19[,7]==list19[i])
  if(length(j)>=1){
    cidades[i] <- dados19[j[1],8]
    UF[i] <- dados19[j[1],9]
  }
}
posicao <- paste(cidades, UF, "Brazil")
write.table(posicao,"nomes_cidades19.txt")
```

```
#2020
UF <- vector()
cidades <- vector()
for(i in 1:length(list20)){
```

```

j <- which(dados20[,3]==list20[i])
if(length(j)>=1){
  cidades[i] <- dados20[j[1],4]
  UF[i] <- dados20[j[1],5]
}
j <- which(dados20[,7]==list20[i])
if(length(j)>=1){
  cidades[i] <- dados20[j[1],8]
  UF[i] <- dados20[j[1],9]
}
}

posicao <- paste(cidades, UF, "Brazil")
write.table(posicao,"nomes_cidades20.txt")

#2021
UF <- vector()
cidades <- vector()
for(i in 1:length(list21)){
  j <- which(dados21[,3]==list21[i])
  if(length(j)>=1){
    cidades[i] <- dados21[j[1],4]
    UF[i] <- dados21[j[1],5]
  }
  j <- which(dados21[,7]==list21[i])
  if(length(j)>=1){
    cidades[i] <- dados21[j[1],8]
    UF[i] <- dados21[j[1],9]
  }
}

posicao <- paste(cidades, UF, "Brazil")
write.table(posicao,"nomes_cidades21.txt")

#####
##-- SALVANDO AS SIGLAS DOS AEROPORTOS DE TODOS OS ANOS --##
#####
Aeroportos <- names(table(c(list18,list19,list20,list21)))
length(Aeroportos)
write.table(Aeroportos, file = "Aeroportos.txt")

```

```
#####
##-- OBTENDO OS NOMES DOS AEROPORTOS PARA TODOS OS ANOS --##
#####
# Selecionando os nomes das cidades desses aeroportos
# coluna 3: Sigla aeroporto origem
# coluna 4: aeroporto origem nome
# coluna 5: UF aeroporto origem
# coluna 7: sigla aeroporto destino
# coluna 8: aeroporto destino nome
# coluna 9: aeroporto destino UF
UF <- vector()
cidades <- vector()
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados18[,3]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados18[j[1],4]
    UF[i] <- dados18[j[1],5]
  }
  j <- which(dados18[,7]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados18[j[1],8]
    UF[i] <- dados18[j[1],9]
  }
}
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados19[,3]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados19[j[1],4]
    UF[i] <- dados19[j[1],5]
  }
  j <- which(dados19[,7]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados19[j[1],8]
    UF[i] <- dados19[j[1],9]
  }
}
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados20[,3]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados20[j[1],4]
  }
}
```

```

    UF[i] <- dados20[j[1],5]
  }
  j <- which(dados20[,7]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados20[j[1],8]
    UF[i] <- dados20[j[1],9]
  }
}
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados21[,3]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados21[j[1],4]
    UF[i] <- dados21[j[1],5]
  }
  j <- which(dados21[,7]==Aeroportos[i,])
  if(length(j)>=1){
    cidades[i] <- dados21[j[1],8]
    UF[i] <- dados21[j[1],9]
  }
}

posicao <- paste(cidades, UF,"Brazil")
write.table(posicao,"nome_cidades_todos.txt")

#####
##-- OBTENDO AS REGIÕES DOS AEROPORTOS PARA TODOS OS ANOS --##
#####

# Selecionando os nomes das cidades desses aeroportos
# coluna 3: Sigla aeroporto origem
# coluna 6: Região origem
# coluna 5: UF aeroporto origem
# coluna 9: Região destino UF
regiao <- vector()
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados18[,3]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados18[j[1],6]
  }
  j <- which(dados18[,7]==Aeroportos[i,])
  if(length(j)>=1){

```

```
    regiao[i] <- dados18[j[1],10]
  }
}
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados19[,3]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados19[j[1],6]
  }
  j <- which(dados19[,7]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados19[j[1],10]
  }
}
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados20[,3]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados20[j[1],6]
  }
  j <- which(dados20[,7]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados20[j[1],10]
  }
}
for(i in 1:dim(Aeroportos)[1]){
  j <- which(dados21[,3]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados21[j[1],6]
  }
  j <- which(dados21[,7]==Aeroportos[i,])
  if(length(j)>=1){
    regiao[i] <- dados21[j[1],10]
  }
}
regiao <- paste(regiao, "Brazil")
write.table(regiao,"regiao_todos.txt")
```

```
#####
##----- EXEMPLO DO CÓDIGO PARA A REDE DO 3 BIMESTRE DE 2018 -----##
##----- AS OUTRAS 62 REDES (ANUAIS, TRIMESTRAIS E BIMESTRAIS) -----##
##----- SEGUEM EXATAMENTE O MESMO RACIOCÍNIO -----##
#####
rm(list=ls()) #LIMPA TUDO QUE ESTÁ NA MEMÓRIA DO R
setwd("D:/Desktop/Dani/TCC") #DEFINE UM DIRETÓRIO

dados18 = read.table("dados_domestic_Regular18.txt")
list18 = read.table(file = "Siglas_Aeroportos18.txt")
list18 = t(as.vector(list18))

#####
##-- CONSTRUINDO A MATRIZ DE ADJACÊNCIA POR BIMESTRE --##
#####
dados18 <- subset(dados18, MÊS>4&MÊS<=6) #3 bimestre

adj.mat18 <- matrix(0, ncol = length(list18), nrow = length(list18) )
for(i in 1:dim(dados18)[1]){
  #aeroporto origem é a coluna3 (SIGLA)
  k <- which(list18==dados18[i,3])
  #aeroporto destino é a coluna 7 (SIGLA)
  l <- which(list18==dados18[i,7])
  #selecionar o número de voos
  #numero de voos é a coluna 16
  adj.mat18[k,l] <- adj.mat18[k,l] + dados18[i,16]
  print(c(dim(dados18)[1],i))
}

# SALVANDO A MATRIZ COM PESOS E COM DIREÇÃO
write.table(adj.mat18, file= "adj_pesos_dir_18_b3.txt")

#####
##-- CRIANDO A MATRIZ SEM DIREÇÃO E SEM PESOS --##
#####
# Criar um grafo binário a partir de sua matriz de adjacencia
rownames(adj.mat18) <- colnames(adj.mat18) <- seq(1:dim(adj.mat18)[1])
adj.mat18 <- as.matrix(adj.mat18)
# Fazer a matriz ser simétrica
adj.mat18 <- adj.mat18 + t(adj.mat18)
```

```

# Agora transformar em binária
adj.mat18[which(adj.mat18!=0)] <- 1
sum(adj.mat18) #qtidade de arestas
# Salvando a matriz
write.table(adj.mat18, file= "adj_bin_undir_18_b3.txt")

#####
##-- ESTATÍSTICAS DESCRITIVAS DAS REDES --##
#####

#####
##-- PACOTES NECESSÁRIOS --##
#####
library(readr)
require(graph)
require(RBGL)
require(gRbase)
require(Rgraphviz)
require(igraph)
require(tidyverse)

adj.mat18.b3 = read.table("adj_pesos_dir_18_b3.txt")
#Criar um grafo a partir de sua matriz de adjacencia
rownames(adj.mat18.b3) <- colnames(adj.mat18.b3) <- seq(1:dim(adj.mat18.b3)[1])
adj.mat18.b3 <- as.matrix(adj.mat18.b3)
adj.mat18.b3
#aqui temos que pensar em usar redes não binárias
adj.mat18.b3[is.na(adj.mat18.b3)] <- 0
adj.mat18.b3[which(adj.mat18.b3!=0)] <- 1
#grafo
gb13 = as(adj.mat18.b3, "igraph")

#####
## MEDIDAS DESCRITIVAS --##
#####
summary(gb13)
mean(degree(gb13))
mean(degree(gb13, mode="in"))
transitivity(gb13)
reciprocity(gb13))

```



```
#####
##-- HISTOGRAMAS --##
#####
degree_inb13 = degree.gb13, mode = 'in')

dados2018b=data.frame(degree_inb13)

a=ggplot(data=dados2018b, aes(x=degree_inb13, y=..density..))+
  geom_histogram(bins=50,fill="#13ECDA", col = "#1392EC") +
  coord_cartesian(xlim=c(0,60), ylim = c(0,0.52))+
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10))+
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10))+
  labs(y="densidade", x="", title="Grau de Entrada 2018-3B")+
  theme_light() + theme(plot.title = element_text(hjust = 0.5),
    text = element_text(size = 12))

# Graus de Saída segue o mesmo raciocínio

#####
##---- GRAFOS BIMESTRAIS COM PESOS ----##
#####
adj.mat18.b3 = read.table("adj_pesos_dir_18_b3.txt")
#Criar um grafo a partir de sua matriz de adjacencia
rownames(adj.mat18.b3) <- colnames(adj.mat18.b3) <- seq(1:dim(adj.mat18.b3)[1])
adj.mat18.b3 <- as.matrix(adj.mat18.b3)

gb13=graph_from_adjacency_matrix(adj.mat18.b3, mode = c("directed"), weighted = TRUE,
  diag = TRUE)

degree_inb13=strength.gb13, vids = V.gb13, mode = c("in"),loops = FALSE)
degree_outb13=strength.gb13, vids = V.gb13, mode = c("out"),loops = FALSE)
degree_13=strength.gb13, vids = V.gb13, loops = FALSE)

# Medida descritiva
mean(degree_13)

# Os histogramas seguem a mesma estrutura da rede sem pesos
```

```
#####
##----- AGRUPAMENTO DAS REDES -----##
##-- EXEMPLO PARA A REDE SEM DIREÇÃO E SEM PESO DE 2018 PARA OS AEROPORTOS FIXOS --##
##----- AS OUTRAS 25 REDES (ANUAIS E BIMESTRAIS) SEGUEM O MESMO RACIOCÍNIO -----##
#####

#####
##-- CONSTRUINDO A MATRIZ DE ADJACÊNCIA USANDO TODOS OS ANOS --##
#####

# Construindo a matriz de adjacência usando os aeroportos que aparecem em TODOS os anos
adj.mat18.todos <- matrix(0, ncol = length(Aeropertos), nrow = length(Aeropertos))
for(i in 1:dim(dados18)[1]){
  #selecionara sigla do aeroporto de origem
  k <- which(Aeropertos==dados18[i,3])
  #selecionar a sigla do aeroporto de destino
  l <- which(Aeropertos==dados18[i,7])
  #selecionar o número de voos
  #numero de voos é a coluna 16
  adj.mat18.todos[k,l] <- adj.mat18.todos[k,l] + dados18[i,16]
  print(c(dim(dados18)[1],i))
}

#Criar um grafo a partir de sua matriz de adjacência
rownames(adj.mat18.todos) <- colnames(adj.mat18.todos) <- seq(1:dim(adj.mat18.todos)[1])
adj.mat18.todos <- as.matrix(adj.mat18.todos)
#Fazer a matriz ser simétrica
adj.mat18.todos <- adj.mat18.todos + t(adj.mat18.todos)
#Agora transformar em binária
adj.mat18.todos[which(adj.mat18.todos!=0)] <- 1
sum(adj.mat18.todos) #qtidade de arestas
write.table(adj.mat18.todos, file="adj_bin_undir_18_todos_anos.txt")

#####
##-- PARTE DO AGRUPAMENTO DAS REDES --##
#####

# Limpar a memória do R
rm(list=ls())

# Definir um diretório
setwd("D:/Desktop/Dani/TCC/TG2")
```

```
#####
##-- PACOTES NECESSÁRIOS --##
#####

library(readr)
library(dplyr)
library(igraph)
library(RgoogleMaps)
library(randnet) # pacote para estimação

#####
##-- LEITURA DOS GRAFOS --##
#####

g2018 <- read.table(file= "adj_bin_undir_18_todos_anos.txt")
# Transformando em matrizes
g2018 <- as.matrix(g2018)
# Fixando uma semente, pois o k-means é "aleatório"
# assim posso verificar os grafos em qualquer momento, caso seja necessário
set.seed(10)
# Estimar as comunidades usando Agrupamento Espectral Regularizado Esférico
z.g2018 <- reg.SSP(g2018,K=5)$cluster

#####
##-- GRAFOS --##
#####

# Definir cores
a=table(z.g2018)
order(a) # ordem decrescente
col2018 <- vector()
col2018[which(z.g2018==order(a)[5])] <- "#13ECDA" #verde-ciano
col2018[which(z.g2018==order(a)[3])] <- "#a107f8" #roxo
col2018[which(z.g2018==order(a)[4])] <- "#87FA1B" #verde-lima
col2018[which(z.g2018==order(a)[2])] <- "#FF00C9" #pink
col2018[which(z.g2018==order(a)[1])] <- "#1392EC" #azul

#####
##-- TABELA REDES --##
#####

# quais cluster de cada aeroportos
z.g2018
```

```

# quantos aeroportos tenho em cada cluster
table(z.g2018==1) #rosa
table(z.g2018==2) #verde-lima
table(z.g2018==3) #verde-ciano
table(z.g2018==4) #roxo
table(z.g2018==5) #azul

# grau médio de cada clusters
# options(digits = 3) #qtidade de casas no valor observado
B=g2018[which(z.g2018==1),which(z.g2018==1)] #rosa
G=graph.adjacency(B,mode="undirected",diag=FALSE)
mean(degree(G))
B=g2018[which(z.g2018==2),which(z.g2018==2)] #verde-lima
G=graph.adjacency(B,mode="undirected",diag=FALSE)
mean(degree(G))
B=g2018[which(z.g2018==3),which(z.g2018==3)] #verde-ciano
G=graph.adjacency(B,mode="undirected",diag=FALSE)
mean(degree(G))
B=g2018[which(z.g2018==4),which(z.g2018==4)] #roxo
G=graph.adjacency(B,mode="undirected",diag=FALSE)
mean(degree(G))
B=g2018[which(z.g2018==5),which(z.g2018==5)] #azul
G=graph.adjacency(B,mode="undirected",diag=FALSE)
mean(degree(G))

# Lendo a região dos aeroportos que selecionamos para todas as redes
regiao= read.table("regiao_todos.txt")
# quantos aeroportos tenho em cada região
table(regiao[which(z.g2018==1),1]) #rosa
table(regiao[which(z.g2018==2),1]) #verde-lima
table(regiao[which(z.g2018==3),1]) #verde-ciano
table(regiao[which(z.g2018==4),1]) #roxo
table(regiao[which(z.g2018==5),1]) #azul

#####
##-- COORDENADAS DO GRAFO --##
#####
# nomes.todos <- read.table("nome_cidades_todos.txt")
# coord <- matrix(0,nrow=dim(nomes.todos)[1],ncol=2)
# for(i in 1:dim(nomes.todos)[1]){

```

```

#       lat_long <- getGeoCode(nomes.todos[i,])
#       coord[i,] <-c(lat_long[2],lat_long[1])
#       print(c(i,dim(nomes.todos)[1]))
# }
# write.table(coord, file="coord_aeroportos.txt")
coord = read.table(file= "coord_aeroportos.txt")
coord = as.matrix(coord)

rownames(g2018) <- colnames(g2018) <- seq(1:dim(g2018)[1])
g2018 <- as.matrix(g2018)
G <- graph.adjacency(g2018,mode="undirected",diag=FALSE)
plot(G,
      layout=coord,
      vertex.size=5,
      vertex.color = col2018,
      vertex.label=NA,
      edge.color = "gray",
      main="2018")

#####
##-- NORMALIZED MUTUAL INFORMATION --##
#####
# Depois de refazer a parte de agrupamento para cada ano/bimestre
# verificamos o NMI dois-a-dois
NMI(z.g2018, z.g2019)

```


Referências Bibliográficas

- Amini, A. A., Chen, A., Bickel, P. J. e Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, **41**(4), 2097–2122.
- ANAC (2021). Agência nacional de aviação brasileira. <https://www.gov.br/anac/pt-br/assuntos/dados-e-estatisticas>. Acesso em: 26 set. 2021.
- Chaudhuri, K., Chung, F. e Tsias, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 35–1. JMLR Workshop and Conference Proceedings.
- Colizza, V., Barrat, A., Barthélemy, M. e Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences*, **103**(7), 2015–2020.
- Fortunato, S. (2010). Community detection in graphs. *Physics reports*, **486**(3-5), 75–174.
- Jin, J. (2015). Fast community detection by score. *The Annals of Statistics*, **43**(1), 57–89.
- Newman, M. (2012). Networks: An introduction. 2010: Oxford university press. *Artificial Life*, **18**, 241–242.
- Qin, T. e Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic block-model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3120–3128. Curran Associates Inc.
- Rocha, L. E. (2009). Structural evolution of the brazilian airport network. *Journal of Statistical Mechanics: Theory and Experiment*, **2009**(04), P04020.
- Shigehalli, V. S. e Shettar, V. M. (2011). Spectral techniques using normalized adjacency matrices for graph matching. *International Journal of Computational Science and Mathematics*, **3**(4), 371–378.
- Siozos-Rousoulis, L., Robert, D. e Verbeke, W. (2021). A study of the us domestic air transportation network: temporal evolution of network topology and robustness from 2001 to 2016. *Journal of Transportation Security*, pages 1–24.
- Yao, Y. (2003). Information-theoretic measures for knowledge discovery and data mining. *Entropy measures, maximum entropy principle and emerging applications*, pages 115–136.