

# Métodos em Aprendizado Não Supervisionado de Máquina - Trabalho III

Daniel Lopes Toso

Junho 2024

## 1 Introdução

O objeto deste trabalho é utilizar os requerimentos de passaporte entre países para encontrar classificações latentes entre os mesmos através de técnicas de aprendizado de máquina não-supervisionado direcionadas à dados de rede.

Os dados foram obtidos através do webscraping do site [Passport Index](#), que contém todas as relações atualizadas de exigências de visto entre países.

Os códigos para o webscraping em Python e para os gráficos, clusters, manipulação de dados e análises feitas com o R estão disponíveis no Github.

## 2 Materiais e Métodos

### 2.1 Estrutura de Relações

Após a coleta das informações relativas aos 199 países encontrados, foi construída uma matriz de relações entre os países, como pode ser observado na Tabela 1. As células representam o tipo de visto necessário para que pessoas associadas ao país da linha entrem no país contido na coluna. Por exemplo, algerianos precisam de visto para qualquer viagem ao Afeganistão mas podem permanecer em Angola por até 30 dias apenas com o passaporte.

Tabela 1: Tabela de requerimentos de visto entre países

Países	Afghanistan	Albania	Algeria	Andorra	Angola	...
Afghanistan		visa required	visa required	visa required	visa required	...
Albania	visa required		visa-free / 90 days	visa required	visa required	...
Algeria	visa required	eVisaApply now		visa required	visa-free / 30 days	...
Andorra	visa required	visa-free / 90 days	visa required		visa required	...
Angola	visa required	eVisaApply now	visa required	visa required		...
...	...	...	...	...	...	...

É possível perceber que existem diversos níveis para as mesmas categorias, como por exemplo "*visa-free*" que pode se estender por 30 ou 90 dias. No total, são encontradas 73 categorias diferentes nos dados. Para simplificar a análise, esses 73 possíveis estados foram condensados em 6 "notas" diferentes para a relação dos países:

- Nota 0: Entrada não admitida / Conexão de um país com si próprio.
- Nota 1: Visto necessário.
- Nota 2: Programas de visto temporário facilitado para turistas.
- Nota 3: Assinatura de visto pós-chegada no país.
- Nota 4: Programas de visto online/automatizado facilitado.
- Nota 5: Circulação livre (com passaporte ou outros documentos).

A partir disso, chega-se em uma estrutura de dados de rede. As observações são direcionadas pois as relações burocráticas entre os países são assimétricas, a exemplo da exigência do visto para os brasileiros que viajam para os Estados Unidos, enquanto estadunidenses precisam apresentar apenas o passaporte em viagens para o Brasil desde 1º de Dezembro de 2023. Além disso, as observações também tem peso, a depender da nota relacionada a facilidade de ingresso no país.

Os países com maior "poder" foram encontrados a partir da soma do nível de seus passaportes para todos os demais países.

## 2.2 Algoritmo Walktrap

Para produzir as visualizações e realizar a clusterização a partir dos dados de rede neste trabalho, foi utilizado o pacote *igraph* do R. A clusterização se deu a partir do [algoritmo Walktrap](#).

O Walktrap funciona do seguinte modo:

1. Definir um número N de passos.
2. Realizar diversos [Passeios Aleatórios](#) com N passos entre os vértices do grafo.
3. Estimar a probabilidade de transição dos estados de uma cadeia de Markov de um tempo para outro para todas as conexões entre os vértices.
4. Determinar a distância entre os vértices através da probabilidade de transição.
5. Inicializar um cluster para cada ponto do grafo.
6. Mesclar clusters com base na distância entre eles, utilizando o [Método de Ward](#).
7. Avaliar a qualidade da partição através da diferença entre a variação dos pontos dentro e fora da mesma.
8. Repetir (6-7) até que todos os pontos estejam no mesmo cluster.
9. Determinar o número e etiqueta dos clusters com base no máximo da avaliação de qualidade das partições.

Assim, o algoritmo calcula, ao mesmo tempo, o melhor número de clusters e suas etiquetas.

## 3 Resultados

### 3.1 Análise dos Dados

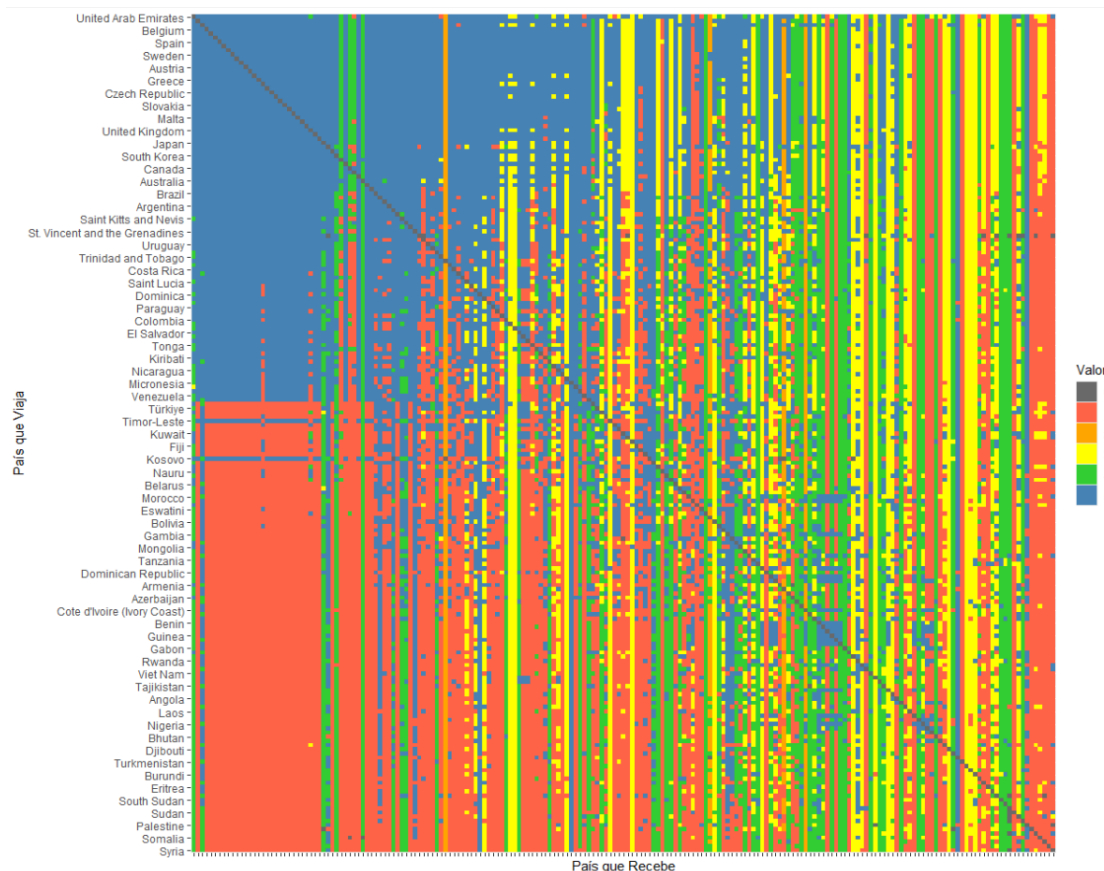


Figura 1: Relações entre países, ordenadas pelos passaportes com maior "poder".

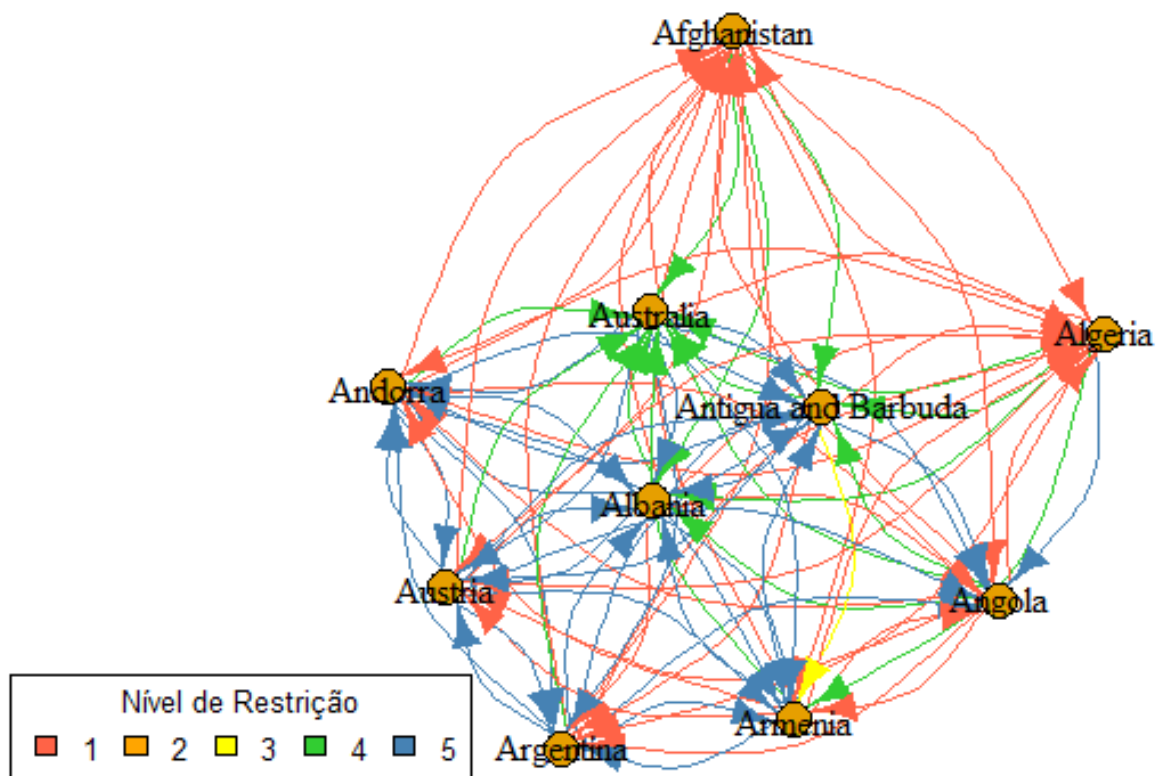


Figura 2: Grafo dos 10 primeiros países em ordem alfabética

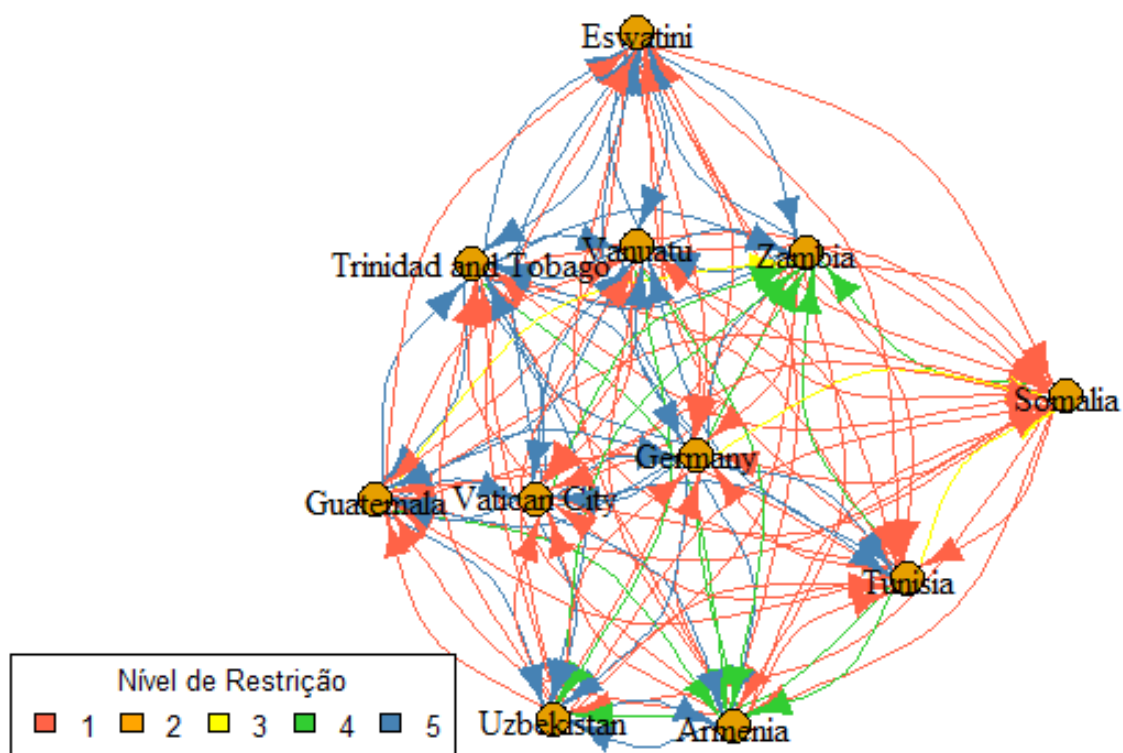


Figura 3: Grafo com 10 países escolhidos aleatoriamente

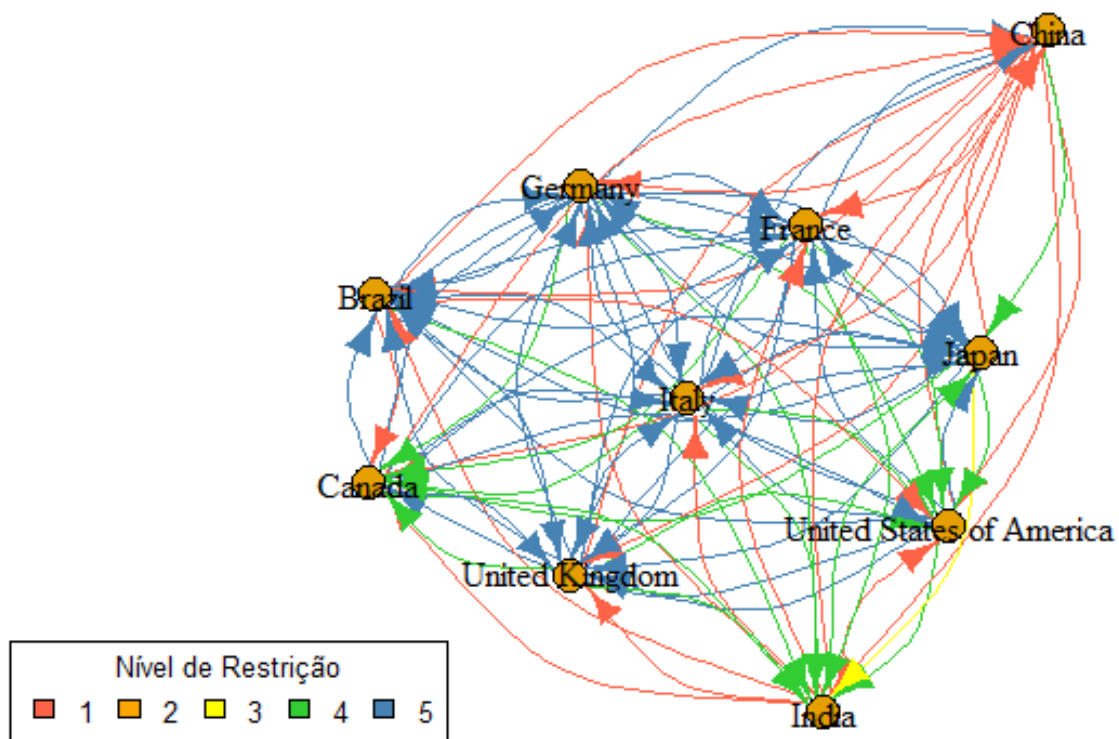


Figura 4: Grafo com os 10 países com o maior PIB

Devido a alta dimensão dos dados, o grafo com todas as conexões é pouco informativo devido ao grande número de vértices sobrepostos. Portanto, para representar a rede, foram feitas essas pequenas amostras dos dados totais.

Vale ressaltar que na Figura 1 alguns nomes de países foram obtidos para possibilitar a visualização.

### 3.2 Clusterização com Walktrap

Países em cada Cluster - Walktrap com 4 passos

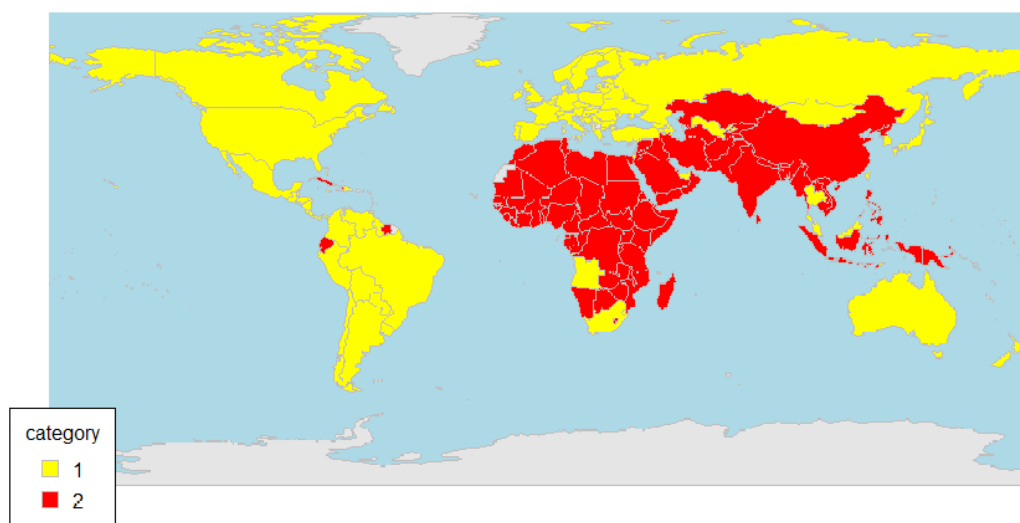


Figura 5: Mapa mundi com os Clusters - Walktrap com 4 passos

## Países em cada Cluster - Walktrap com 10 passos

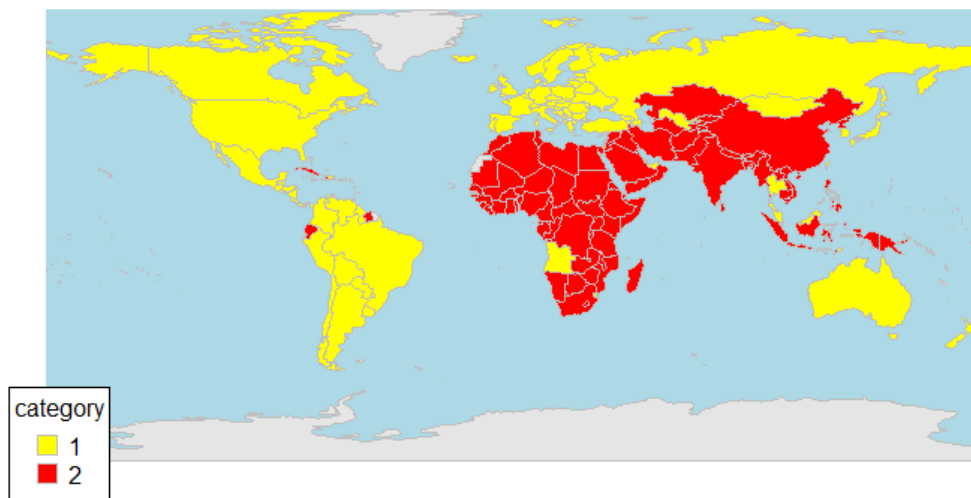


Figura 6: Mapa mundi com os Clusters - Walktrap com 10 passos

## 4 Discussão

### 4.1 Análise Descritiva

A partir da figura 1, é possível perceber algumas características interessantes sobre as diferentes políticas internacionais dos países. A região quadrada azul no canto superior esquerdo indica que, em geral, os países com passaportes de maior acesso tendem a estar contidos em bolhas de acesso mútuo, o que sugere seu pertencimento a estruturas que facilitam a migração e circulação de pessoas como a [União Europeia](#) ou o [Mercosul](#). Também a partir da figura 1 é possível perceber que os Emirados Árabes Unidos foram classificados como donos do passaporte com maior poder, mesmo resultado encontrado pelo site [PassportIndex](#) a partir da métrica própria utilizada por eles.

Analisando as Figuras 2 e 3 é possível perceber a disparidade das políticas migratórias entre os países da Europa e Américas e os países da África e Oriente Médio. O acesso desse último grupo é muito mais controlado e também seus passaportes parecem ser mais restritos.

### 4.2 Análise de Perfil

As figuras 5 e 6 mostram as clusterizações obtidas a partir do Walktrap com dois parâmetros distintos para a heurística do número de passos. É possível perceber que os resultados encontrados se mostram robustos quanto a esse parâmetro.

Vale notar que o aumento indefinido do número de passos possibilita que os passeios aleatórios do algoritmo tenham acesso a um número maior dos vértices do grafo, o que resulta que o cluster ideal encontrado sempre conterá todos os pontos que estão conectados entre si (o que é o caso para esses dados).

A clusterização encontrou dois grupos que podem ser relacionados aos conceitos políticos de "Ocidente" e "Oriente". Os países no cluster 1 possuem fortes laços com a cultura europeia e o modo de produção capitalista, a perceber pela presença dos países asiáticos como Japão e Coreia do Sul, que receberam influência da política externa estadunidense durante a guerra fria. Outros países fora do Ocidente geográfico que se encontram no grupo 1 são a Austrália e Nova Zelândia.

## 5 Conclusão

Em resumo, a análise descritiva dos dados permitiu o apontamento de relações intrínsecas aos países que se baseiam em fenômenos bem definidos do mundo real. Além disso, a clusterização foi bem sucedida, destacando a atualidade da disputa política entre os países de duas bolhas diferentes formadas historicamente a partir do colonialismo europeu, modos de produção divergentes, entre outros fatores.