Escalonamento Multidimensional

Jhessica Letícia Kirch
Universidade de São Paulo

Simpósio de Microbiologia Agrícola
11 de abril de 2023

ESCALONAMENTO MULTIDIMENSIONAL

Objetivo: construir um diagrama que mostre as relações entre um certo número de objetos, a partir de uma tabela de distâncias entre objetos.

Tipos:

- Uma dimensão – os objetos caem em uma reta;
- Duas dimensões – os objetos caem em um plano;
- Três dimensões – os objetos caem no espaço;
- Mais altas dimensões – não são possíveis por simples representação geométrica.

ESCALONAMENTO MULTIDIMENSIONAL

- É uma técnica essencialmente gráfica.
- Útil quando os dados são representados por distâncias ou tabelas de contingência.
- Não faz restrição aos tipos de dados que deram origem à matriz de distâncias. Uma vez que todas as operações são realizadas sobre tal matriz, as variáveis podem ser quantitativas, ordinais, binárias, qualitativas ou até mesmo uma mistura delas.

ESCALONAMENTO MULTIDIMENSIONAL

- Quanto maior a dimensão da matriz de distâncias, mais difícil torna-se o estudo do relacionamento entre os objetos.

- Dizemos que a análise foi bem sucedida quando a perda de informação, ao sair da dimensão n para a dimensão k é mínima.

PROCEDIMENTO PARA ESCALONAMENTO MULTIDIMENSIONAL

- Inicia-se com uma matriz de distâncias entre n objetos, sendo δ_{ij} a distância do objeto i ao objeto j;

- O número de dimensões para o mapeamento dos objetos é fixado por uma solução particular em t (1 ou mais).

PROCEDIMENTO PARA ESCALONAMENTO MULTIDIMENSIONAL

1. Uma configuração inicial é estabelecida para os n objetos em t dimensões, i.e., coordenadas $(x_1, x_2, \ldots, x_t)$, são assumidas para cada objeto em um espaço t-dimensional.

2. As distâncias euclidianas entre os objetos são calculadas para a configuração assumida. Seja d_{ij} a distância entre o objeto i e o objeto j para esta configuração.

PROCEDIMENTO PARA ESCALONAMENTO MULTIDIMENSIONAL

3. Uma regressão de d_{ij} em δ_{ij} (dados de entrada) é feita.

$$d_{ij} = \alpha + \beta\delta_{ij} + \varepsilon_{ij}$$

em que ε_{ij} é um erro de regressão e α e β são constantes.

PROCEDIMENTO PARA ESCALONAMENTO MULTIDIMENSIONAL

4. A qualidade de ajuste entre d_{ij} e $\hat{d}_{ij}$ é medida por uma estatística adequada.

Ex.: Fórmula stress de Kruskal:

$$STRESS\ 1 = \left\{ \frac{\sum \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum \hat{d}_{ij}^2} \right\}^{1/2}$$

é uma medida do quanto a configuração espacial de pontos tem que ser forçada para obter os dados de distâncias δ_{ij}.

PROCEDIMENTO PARA ESCALONAMENTO MULTIDIMENSIONAL

5. As coordenadas $(x_1, x_2, \ldots, x_t)$ de cada objeto são alteradas levemente de tal forma que o stress é reduzido.

- Os passos de 2 a 5 são repetidos até que indicação de que o stress não pode ser mais reduzido.

Resultado da análise: Coordenadas dos n objetos em t dimensões, que servem desenhar um mapa que mostra como os objetos estão relacionados.

- $t = 1, 2$ ou 3 é o ideal, pois um representação gráfica dos n objetos é então direta, mas nem sempre possível

PROCEDIMENTO PARA ESCALONAMENTO MULTIDIMENSIONAL

- O stress é utilizado para avaliar o percentual de informação não mapeada.

- Guia rústico de Kruskal e Wish (1978, p.56) para valores de STRESS:

  É questionável reduzir o número de dimensões até que STRESS1 exceda 10% ou aumentando quando já é menor do que 5%.

- É pouco importante aumentar o número de dimensões t se isto leva a um pequeno decréscimo no stress.

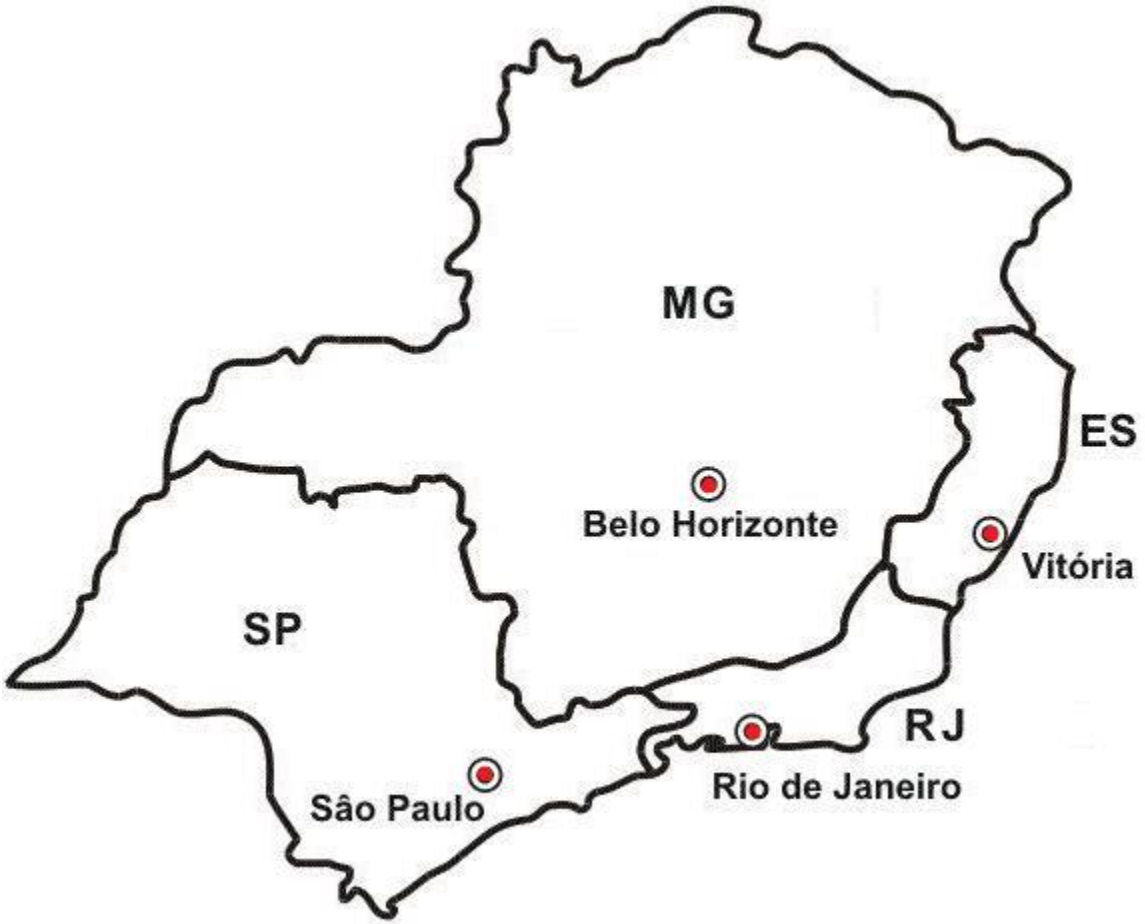
EXEMPLO ILUSTRATIVO

Exemplo Distâncias rodoviárias entre as capitais do sudeste do Brasil

Situação: Temos um mapa de distâncias rodoviárias, e queremos reproduzir com elas, o mapa de distâncias geográficas.

EXEMPLO ILUSTRATIVO

Figura 1 Capitais dos estados do sudeste do Brasil



EXEMPLO ILUSTRATIVO

Considerações:

- Se as distâncias rodoviárias fossem proporcionais às distâncias geográficas, seria possível reconstituir o verdadeiro mapa exatamente, usando uma análise bidimensional;

- As distâncias rodoviárias são em alguns casos muito maiores do que as distâncias geográficas;

- Tudo que se pode esperar é uma reconstituição bastante aproximada do verdadeiro mapa da Figura 1.

EXEMPLO ILUSTRATIVO

Tabela 1 Distâncias (δ_{ij}) rodoviárias em km entre 4 capitais do sudeste do Brasil.

	Belo Horizonte	**Rio de Janeiro**	**São Paulo**	**Vitória**
Belo Horizonte	0			
Rio de Janeiro	442	0		
São Paulo	590	433	0	
Vitória	515	518	940	0

EXEMPLO ILUSTRATIVO

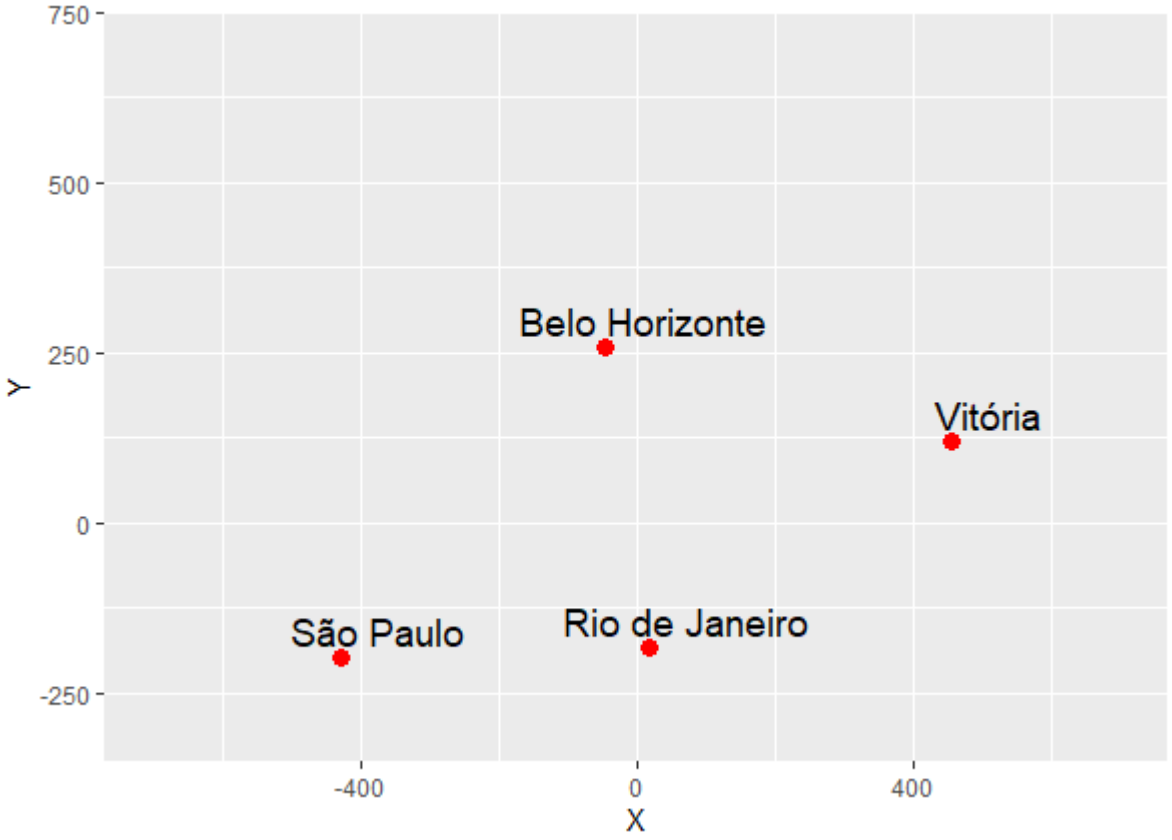
Tabela 2 Coordenadas produzidas (matriz X $(n \times t)$ por escalonamento multidimensional aplicado às distâncias entre capitais do sudeste do Brasil.

Capital	Dimensão	
	1	**2**
Belo Horizonte	58,35	257,01
Rio de Janeiro	- 54,13	- 174,41
São Paulo	- 470,83	- 15,73
Vitória	466,60	- 66,88

- As coordenadas da Tabela 2 foram rotacionadas para se assemelhar a representação gráfica da Figura 1.

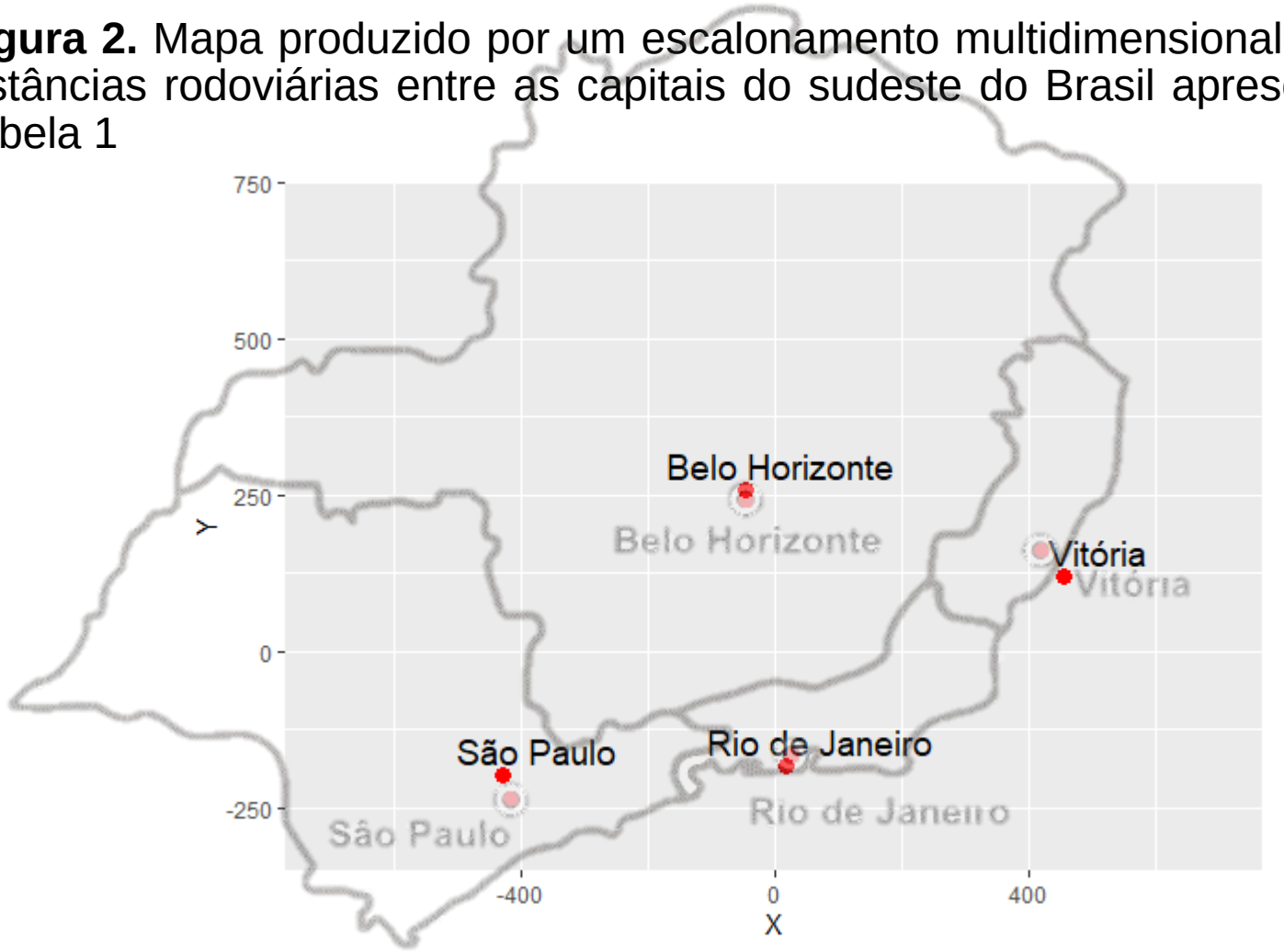
EXEMPLO ILUSTRATIVO

Figura 2 Mapa produzido por um escalonamento multidimensional usando as distâncias rodoviárias entre as capitais do sudeste do Brasil apresentadas na Tabela 1



EXEMPLO ILUSTRATIVO

Figura 2. Mapa produzido por um escalonamento multidimensional usando as distâncias rodoviárias entre as capitais do sudeste do Brasil apresentadas na Tabela 1



EXEMPLO ILUSTRATIVO

- Comparando a Figura 1 com a Figura 2 observa-se que o escalonamento multidimensional teve bastante sucesso na reconstituição do mapa real.

- **Exemplo: O comportamento de votação de parlamentares**

- A tabela a seguir mostra os votos de 15 parlamentares de Nova Jersey em 19 projetos de lei sobre meio ambiente.

- Em geral acredita-se que Republicanos e Democratas tendem a votar em consistência com a linha partidária.

Tabela 3. Votos de 15 parlamentares em 19 projetos de lei sobre meio ambiente entre 15 parlamentares de Nova Jersey.

	Congressman														
	Hunt (R)	Sandman (R)	Howard (D)	Thompson (D)	Frelinghuysen (R)	Forsythe (R)	Widnall (R)	Roe (D)	Helstoski (D)	Rodino (D)	Minish (D)	Rinaldo (R)	Maraziti (R)	Daniels (D)	Patten (D)
Bill	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 (Y)	0	0	0	1	1	1	1	0	1	1	1	1	0	1	N
2 (N)	1	0	N	0	0	0	1	0	0	0	0	0	1	0	0
3 (N)	1	N	0	N	1	1	N	1	1	1	1	1	1	1	1
4 (Y)	0	0	1	N	0	0	0	1	1	1	1	1	0	0	0
5 (Y)	0	N	1	N	0	0	0	1	1	1	1	1	1	1	0
6 (Y)	1	N	1	N	0	1	1	0	0	0	0	0	1	0	0
7 (Y)	0	0	1	1	1	1	1	1	N	1	1	1	1	1	1
8 (N)	1	1	0	0	1	0	0	N	0	0	0	0	1	0	0
9 (Y)	0	0	1	1	0	0	0	N	1	1	1	1	1	N	1
10 (Y)	0	0	1	1	0	0	0	1	1	1	1	1	0	0	1
11 (Y)	0	0	1	1	1	0	0	1	1	1	1	1	0	0	1
12 (Y)	0	N	1	N	0	0	0	1	1	1	1	1	0	1	1
13 (N)	1	N	0	1	N	0	1	0	0	0	0	0	0	0	1
14 (N)	0	N	0	0	0	0	1	0	0	0	0	0	0	0	0
15 (N)	1	N	0	0	1	0	1	0	0	0	0	0	1	0	0
16 (N)	N	N	0	N	1	1	1	0	0	N	0	0	1	N	0
17 (N)	N	N	0	0	1	0	1	0	0	0	0	0	0	0	0
18 (Y)	1	1	1	1	N	0	1	1	1	1	1	1	0	1	1
19 (Y)	0	0	1	1	N	0	0	1	1	0	0	1	0	0	1

A Tabela 4 mostra a distância entre 15 parlamentares de Nova Jersey. Ela é calculada pelo número de votos de discordância entre os parlamentares.

EXEMPLO

Tabela 4. Distância entre 15 parlamentares de Nova Jersey.

	Hunt	Sandman	Howard	Thompso	Frelingh	Forsythe	Widnall	Roe	Helstoski	Rodino	Minish	Rinaldo	Maraziti	Daniels	Pattern
Hunt (R)	0														
Sandman (R)	8	0													
Howard (D)	15	17	0												
Thompson (D)	15	12	9	0											
Frelinghuysen(R)	10	13	16	14	0										
Forsythe (R)	9	13	12	12	8	0									
Widnall (R)	7	12	15	13	9	7	0								
Roe (D)	15	16	5	10	13	12	17	0							
Helstoski (D)	16	17	5	8	14	11	16	4	0						
Rodino (D)	14	15	6	8	12	10	15	5	3	0					
Minish (D)	15	16	5	8	12	9	14	5	2	1	0				
Rinaldo (R)	16	17	4	6	12	10	15	3	1	2	1	0			
Maraziti (R)	7	13	11	15	10	6	10	12	13	11	12	12	0		
Daniels (D)	11	12	10	10	11	6	11	7	7	4	5	6	9	0	
Pattern (D)	13	16	7	7	11	10	13	6	5	6	5	4	13	9	0

EXEMPLO

- **Nota:** Os números mostrados são o número de vezes que o parlamentar votou diferentemente em 19 propostas de leis ambientais (R= Partido Republicano, D=Partido Democrata)

- **Fonte:** Romesburg, H.C. (1984), *Cluster Analysis for researchers*, Lifetime Learning Publications, Belmont, CA

- Obs: De acordo com Silva (2016), em geral valores de stress inferiores a 15% são aceitáveis.

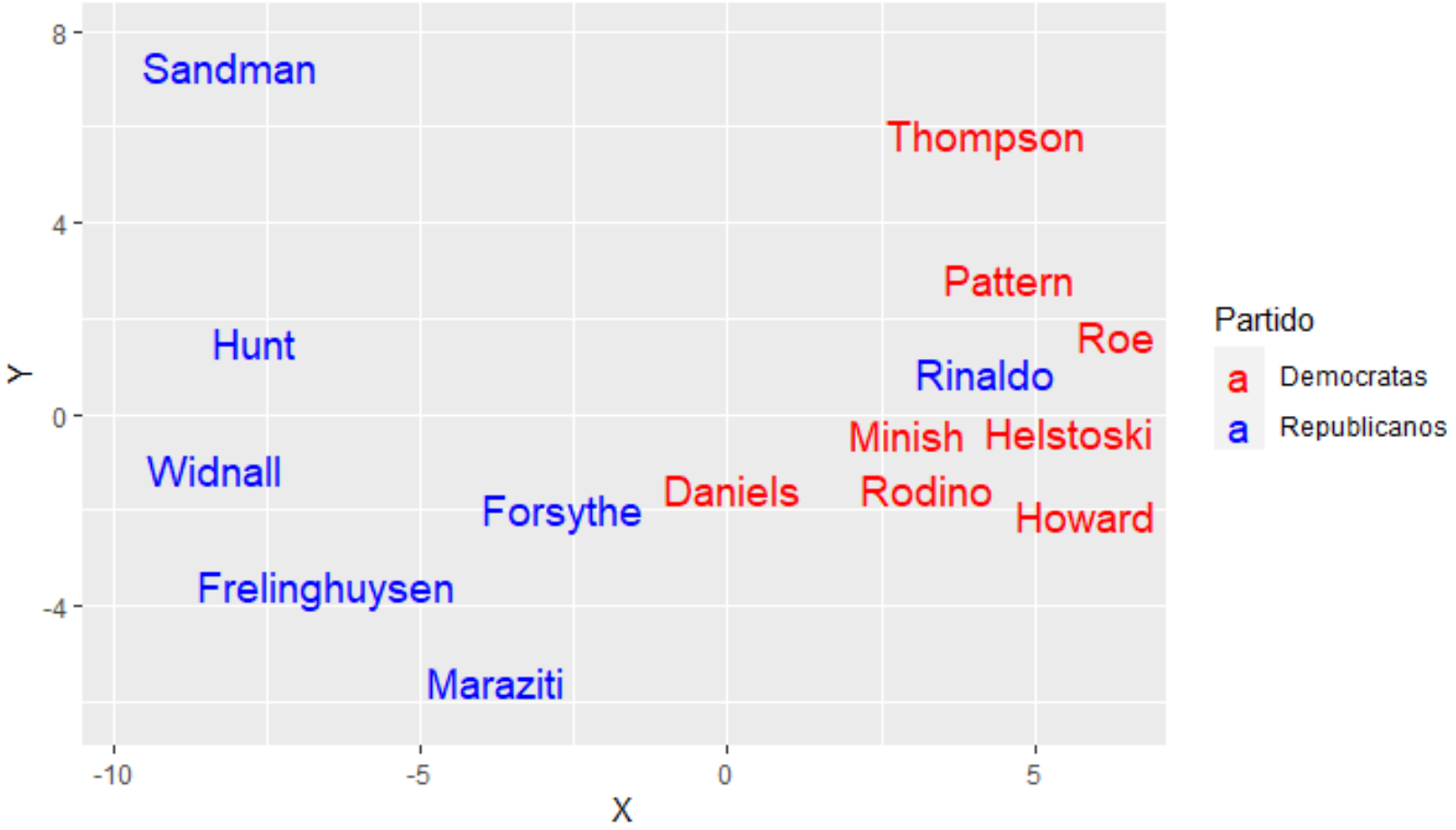
EXEMPLO

Tabela 5 Coordenadas de 15 parlamentares obtidas de um escalonamento multidimensional baseadas no comportamento de votação.

Parlamentares	Dimensão	
	1	**2**
Hunt (R)	-8,44	0,91
Sandman (R)	-7,41	7,88
Howard (D)	6,09	-1,50
Thompson (D)	3,52	5,25
Frelinghuysen(R)	-7,25	-4,18
Forsythe (R)	-3,28	-2,57
Widnall (R)	-9,71	-1,12
Roe (D)	6,34	1,04
Helstoski (D)	6,30	0,27
Rodino (D)	4,28	-0,92
Minish (D)	4,26	-0,39
Rinaldo (R)	5,03	0,27
Maraziti (R)	-4,46	-6,22
Daniels (D)	0,81	-0,94
Pattern (D)	3,89	2,23

EXEMPLO

Figura 5 Representações de parlamentares obtidas de um escalonamento multidimensional.



EXEMPLO

Leitura:

- A dimensão 1 reflete grandes diferenças entre os partidos, pois D caem do lado esquerdo e R caem do lado direito exceto Rinaldo.

Exercício:

- Considere os dados sobre a porcentagem de pessoas empregadas em diferentes indústrias em 26 países da Europa. Destes dados construa uma matriz de distâncias euclidiana entre os países e implemente um escalonamentos multidimensional.