Fundação Oswaldo Cruz
Escola Nacional de Saúde Pública
Programa de Pós-graduação em Epidemiologia em Saúde Pública
Doutorado Acadêmico em Epidemiologia em Saúde Pública

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

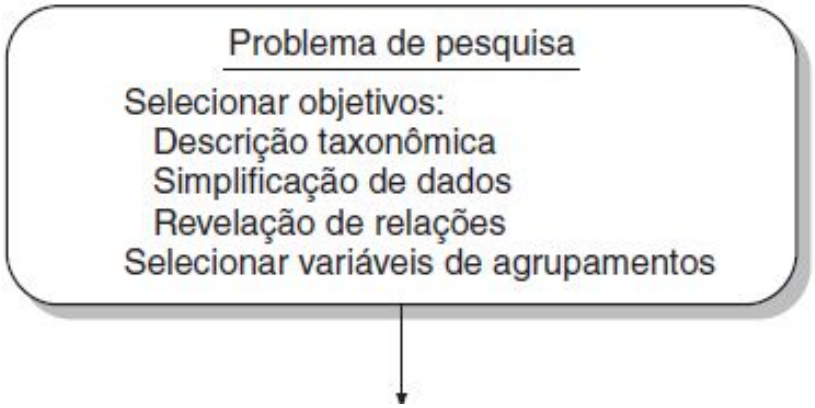
Análise de agrupamentos

Elizabeth Leite Barbosa
Isiyara Taverna Pimenta

DEFINIÇÃO

Grupos de técnicas multivariadas cuja finalidade principal é agregar objetos/indivíduos com base nas características que eles possuem. A ideia é maximizar a homogeneidade de objetos dentro dos grupos e a heterogeneidade entre os grupos.

ESTÁGIO 1

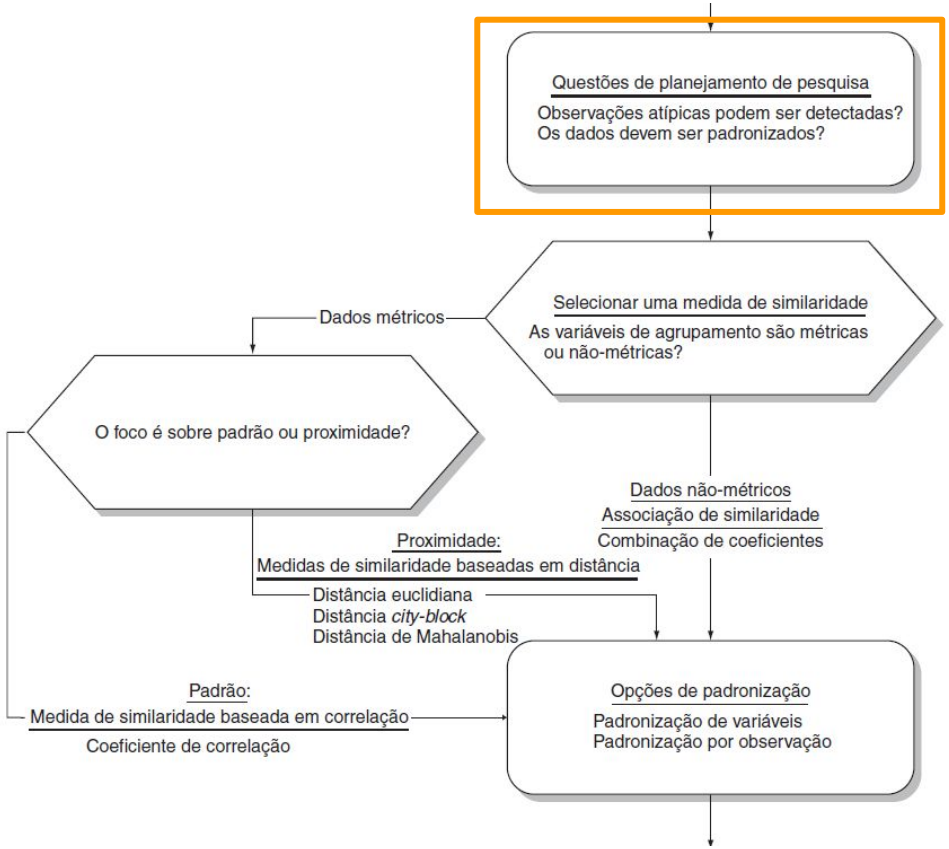


Por quê deve existir a estrutura?
Quais tipos de medidas que devem ser usadas para caracterizar os objetos?

Descrição taxonômica -> fins exploratórios- classificação de objetos com base empírica.

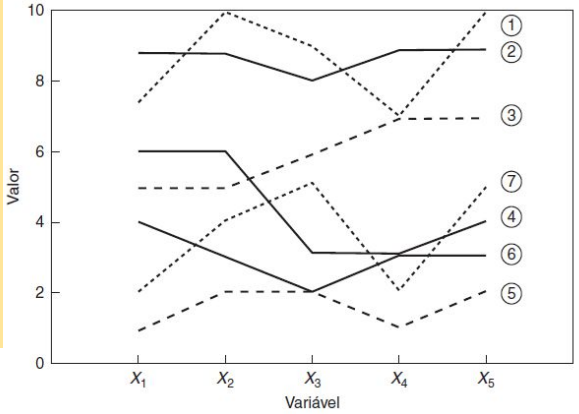
Tipologia proposta (classificação com base teórica) pode ser comparada com aquela obtida pela análise de agrupamentos -> fins confirmatórios

ESTÁGIO 2



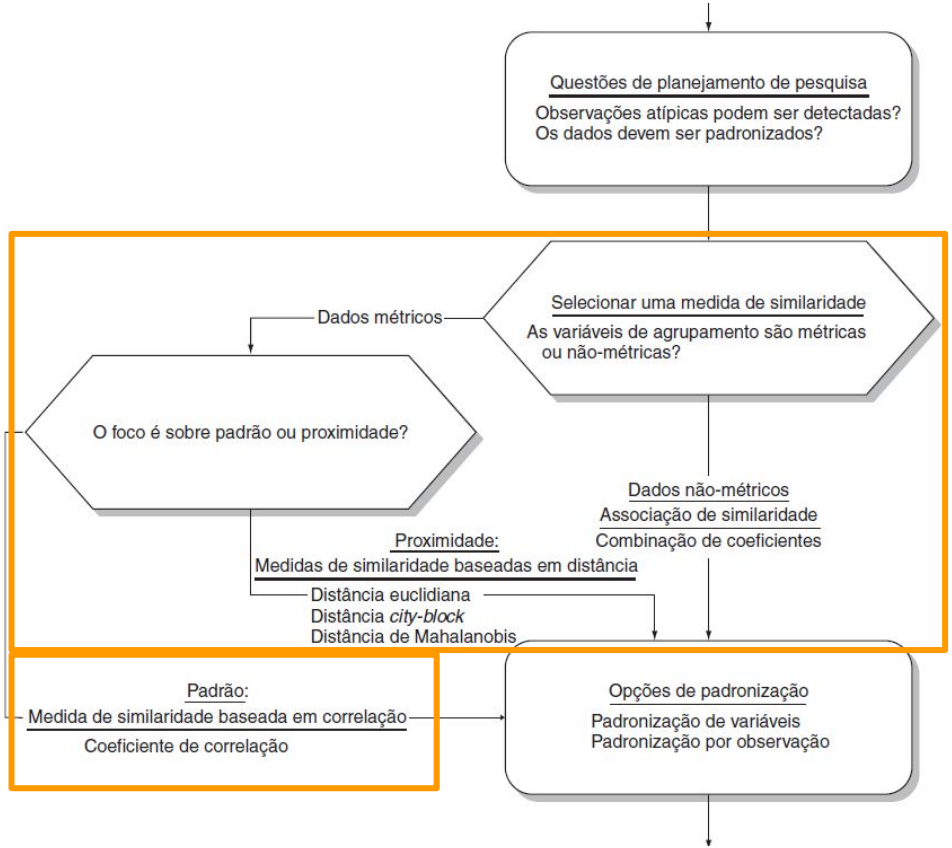
Detecção bi e multivariada ou Medidas de similaridade

FIGURA 8-4 Diagrama de perfil.

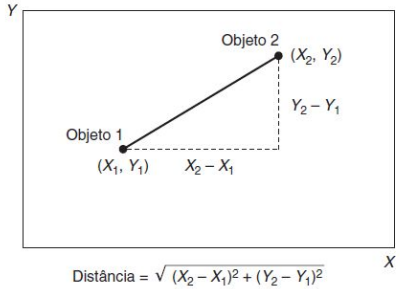


D² de Mahalanobis: mede a distância de cada observação em um espaço multidimensional a partir do centro médio de todas as observações. **Valores elevados indicam observações afastadas.**

ESTÁGIO 2

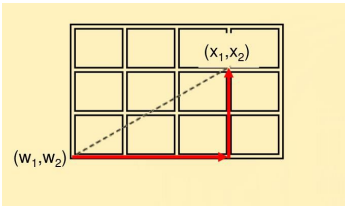


Distância euclidiana

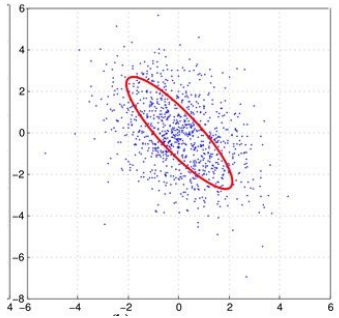


Distância = $\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$

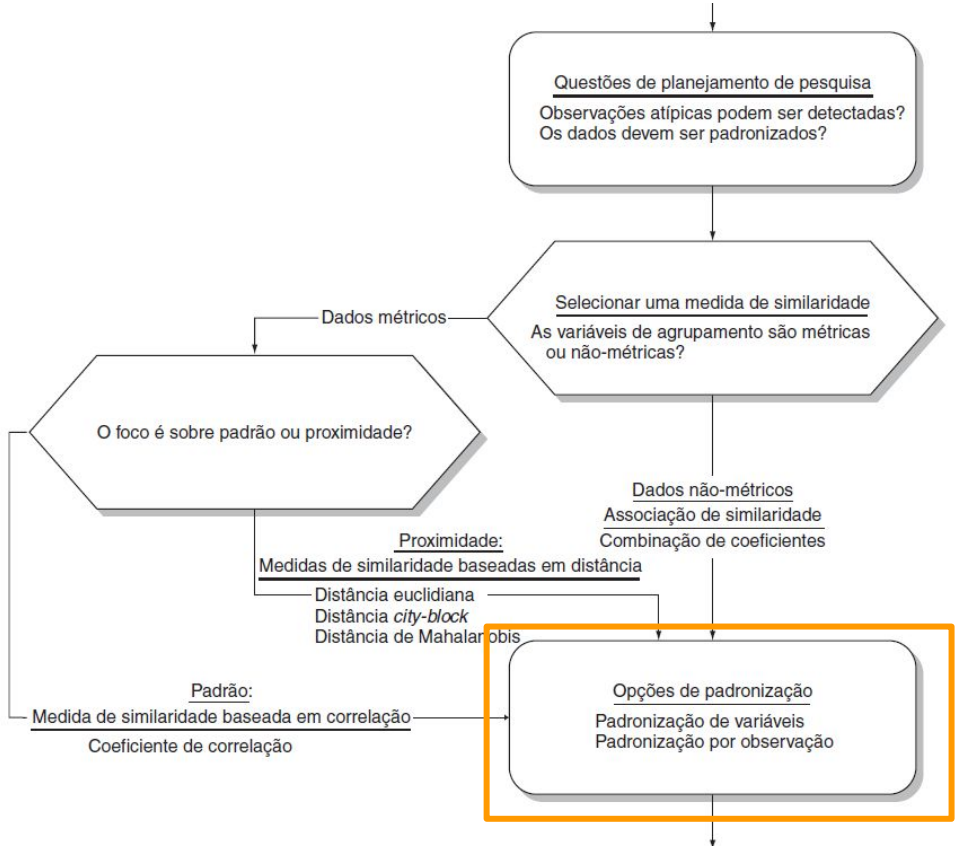
Distância de city-block



Distância de Mahalanobis



ESTÁGIO 2



Diferentes escalas ou magnitudes entre as variáveis

Cenário 1: Medidas de distância baseadas em probabilidade de compra e minutos de comercial assistido

	Distância euclidiana simples		*Distância euclidiana ao quadrado ou absoluta*		*Distância city-block*	
Par de objetos	Valor	Ordem	Valor	Ordem	Valor	Ordem
A-B	5,025	3	25,25	3	5,5	3
A-C	3,162	2	10,00	2	4,0	2
B-C	2,062	1	4,25	1	2,5	1

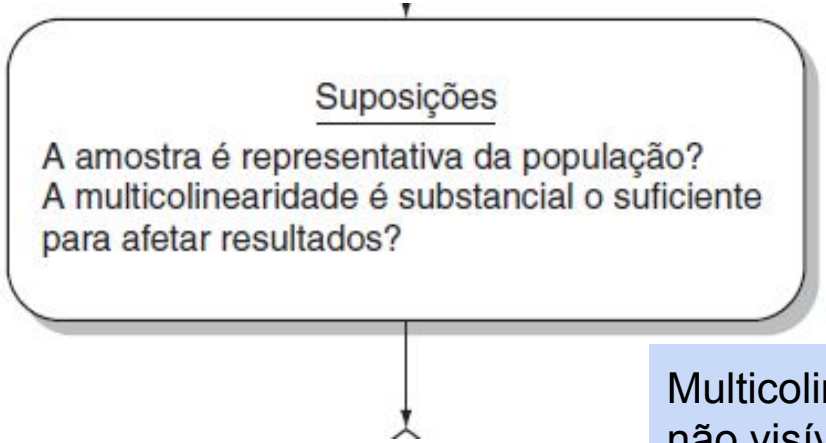
Cenário 2: Medidas de distância baseadas em probabilidade de compra e segundos de comercial assistido

	Distância euclidiana simples		*Distância euclidiana ao quadrado ou absoluta*		*Distância city-block*	
Par de objetos	Valor	Ordem	Valor	Ordem	Valor	Ordem
A-B	30,41	2	925	2	35	3
A-C	60,07	3	3,609	3	63	2
B-C	30,06	1	904	1	32	1

Cenário 3: Medidas de distância baseadas em valores padronizados de probabilidade de compra e minutos ou segundos de comercial assistido

	Valores padronizados		*Distância euclidiana simples*		*Distância euclidiana ao quadrado ou absoluta*		*Distância city block*	
Par de objetos	Probabilidade de compra	Minutos/segundos de comercial assistido	Valor	Ordem	Valor	Ordem	Valor	Ordem
A-B	–1,06	–1,0	2,22	2	4,95	2	2,99	2
A-C	0,93	0,0	2,33	3	5,42	3	3,19	3
B-C	0,13	1,0	1,28	1	1,63	1	1,79	1

ESTÁGIO 3



As exigências de normalidade, linearidade e homocedasticidade, que eram tão importantes em outras técnicas, realmente têm pouco peso na análise de agrupamentos.

Multicolinearidade atua como um processo de ponderação não visível para o observador, mas que afeta a análise.

- Reduzir as variáveis a números iguais em cada conjunto
- Usar uma das medidas de distância, como a de Mahalanobis, que compensa essa correlação.

ESTÁGIO 4

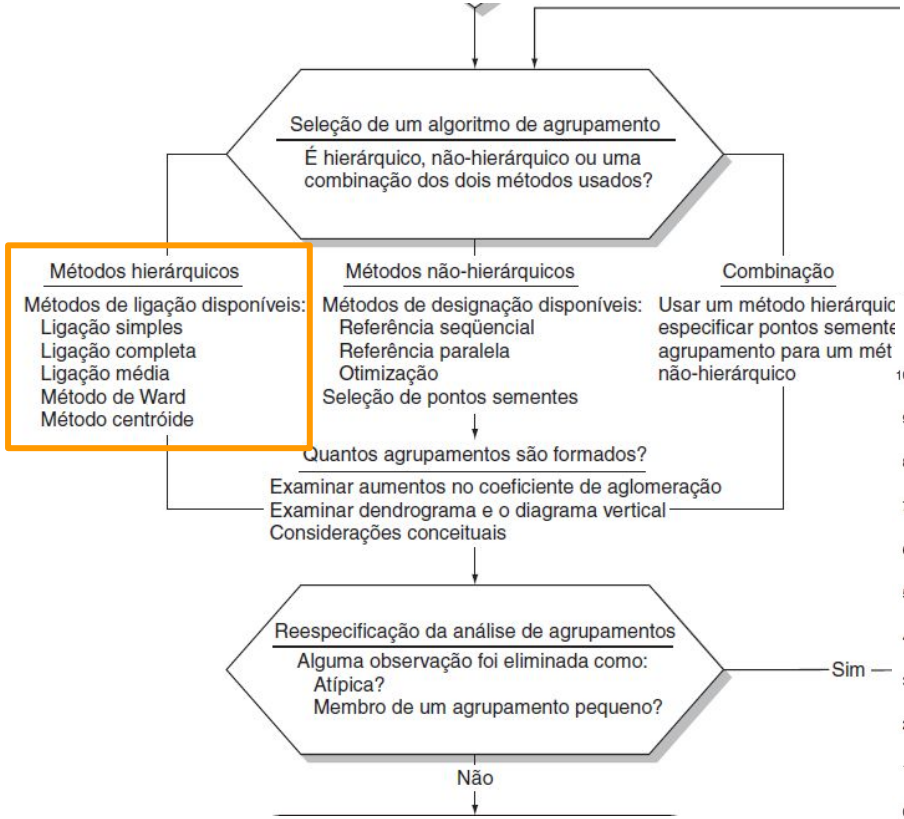
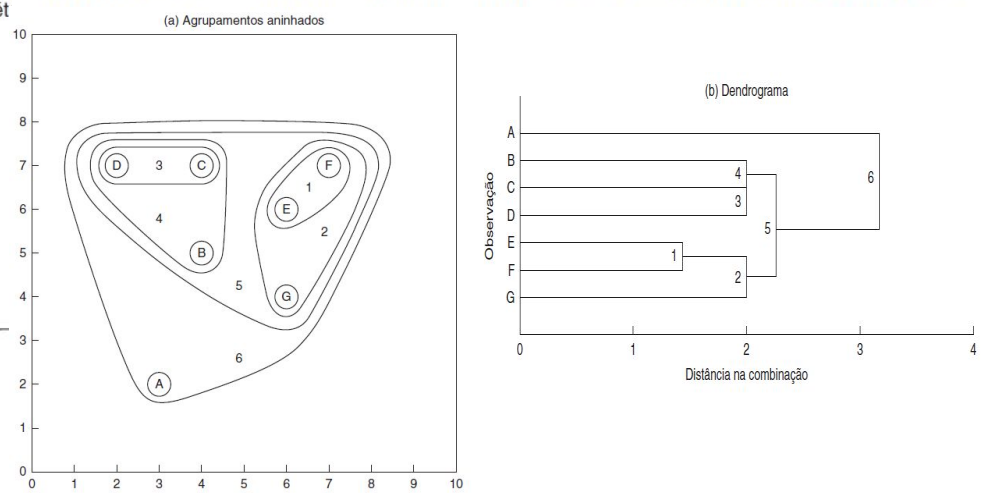


TABELA 8-2 Processo de agrupamento hierárquico aglomerativo

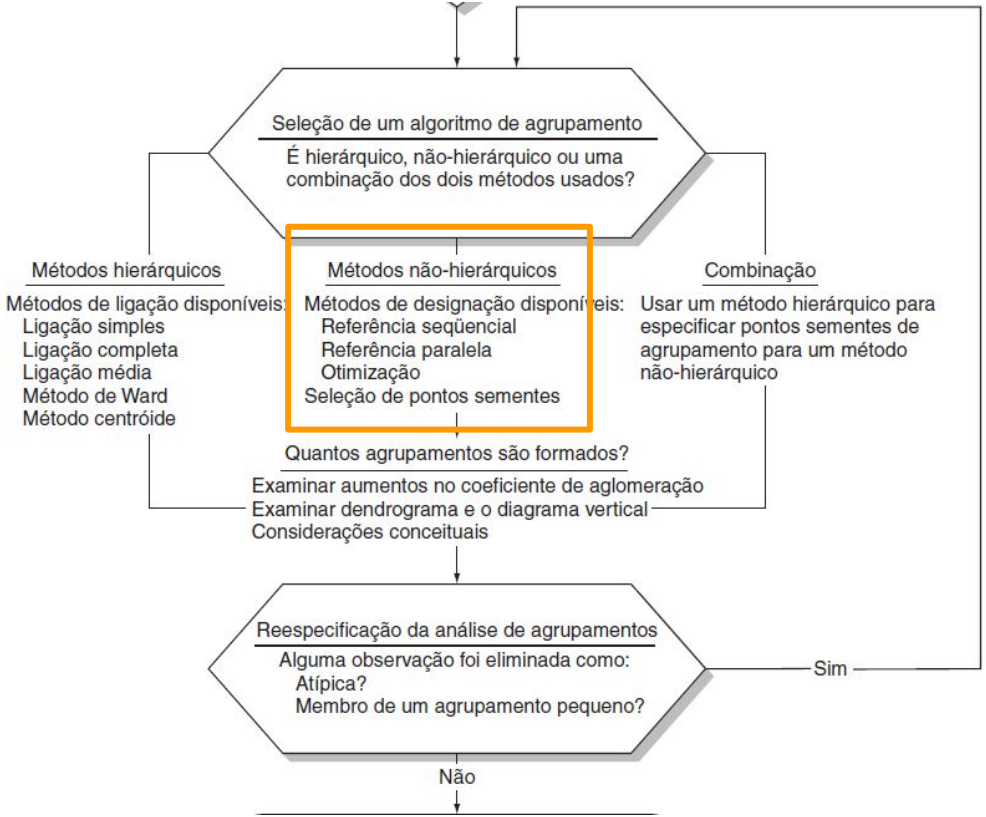
	PROCESSO DE AGLOMERAÇÃO		SOLUÇÃO DE AGRUPAMENTO		
Passo	*Distância mínima entre observações não-agrupadas*[a]	*Par de observações*	*Pertinência a agrupamento*	*Número de agrupamentos*	*Medida de similaridade geral (distância média dentro do agrupamento)*
	Solução inicial		**(A) (B) (C) (D) (E) (F) (G)**	**7**	**0**
1	1,414	E-F	(A) (B) (C) (D) (E-F) (G)	6	1,414
2	2,000	E-G	(A) (B) (C) (D) (E-F-G)	5	2,192
3	2,000	C-D	(A) (B) (C-D) (E-F-G)	4	2,144
4	2,000	B-C	(A) (B-C-D) (E-F-G)	3	2,234
5	2,236	B-E	(A) (B-C-D-E-F-G)	2	2,896
6	3,162	A-B	(A-B-C-D-E-F-G)	1	3,420

[a]Distância euclidiana entre observações.

FIGURA 8-2 Descrições gráficas do processo de agrupamento hierárquico.

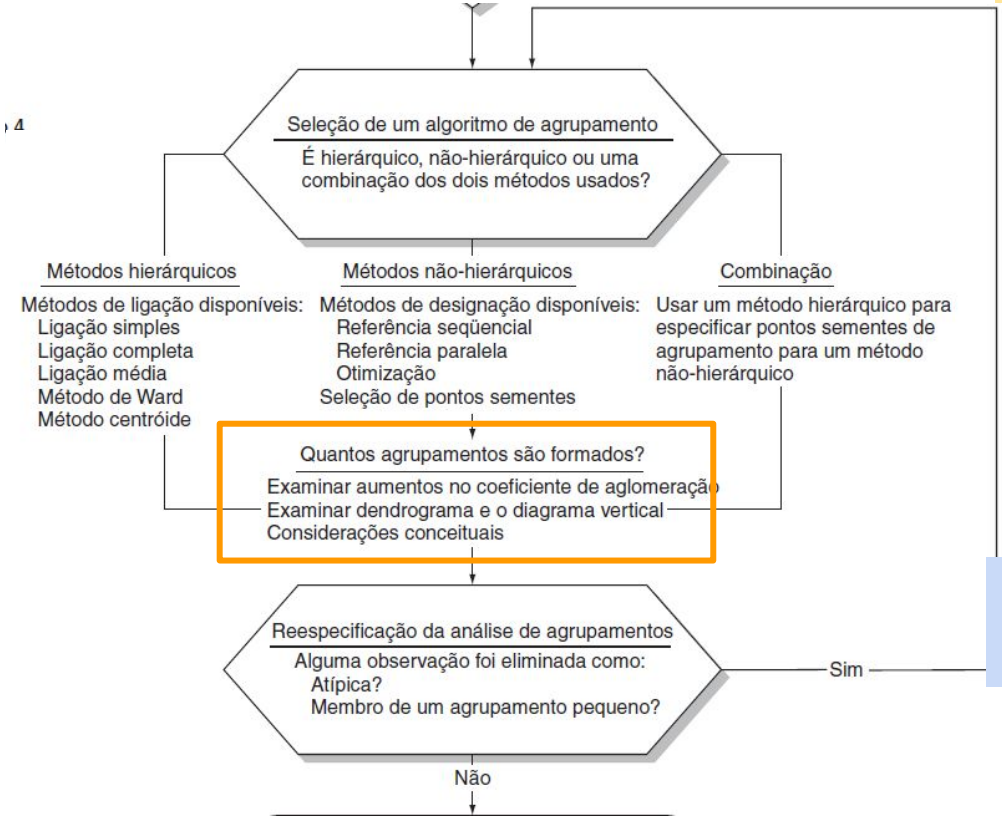


ESTÁGIO 4



- Especificar sementes de agrupamento
- Designação:
  - Sequencial-> começa pela seleção de uma semente de agrupamento e inclui todos os objetos dentro de uma distância pré-especificada. Depois passa para a segunda semente.
  - Paralela->considera todas as sementes simultâneamente
  - Otimização-> permite a redesignação de observações.

Estágio 4

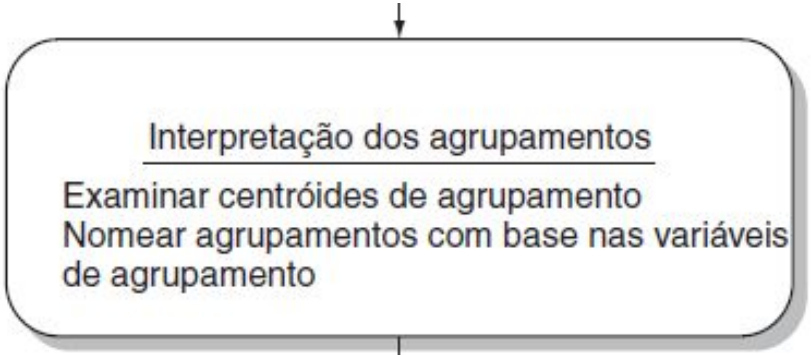


REGRAS DE PARADA

- Medidas de mudanças de heterogeneidade
- Variações percentuais de heterogeneidade -> coeficiente de aglomeração
- Medidas de variação de variância -> raiz do desvio padrão quadrático médio (RMSSTD) -> Grandes aumentos na RMSSTD sugerem a união de dois agrupamentos bastante distintos
- Medidas estatísticas de variação de heterogeneidade
- Medidas diretas de heterogeneidade

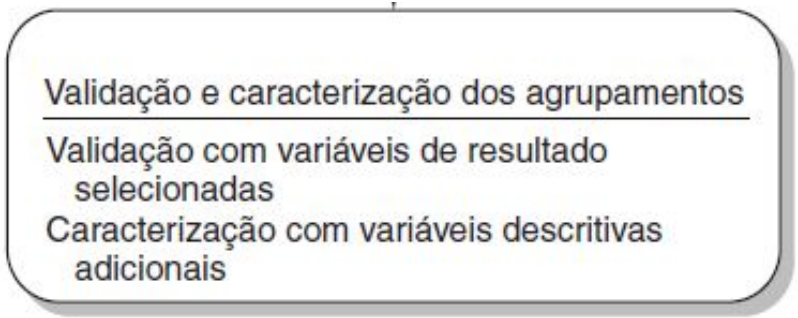
Empregar diversas regras de parada e procurar por uma solução que seja consenso.

Estágio 5



O estágio de interpretação envolve o exame de cada agrupamento em termos da variável estatística de agrupamento para nomear ou designar um rótulo que descreva precisamente
a natureza dos agregados.

Estágio 6



Tentativas do pesquisador para garantir que a solução de agrupamentos seja representativa da população geral, e assim seja generalizável para outros objetos e estável com o passar do tempo.

ANÁLISE DE AGRUPAMENTOS NO R

Peter S. Fader & Leonard M. Lodish

A Cross-Category Analysis of Category Structure and Promotional Activity for Grocery Products

➔ **Objeto:** 331 categorias de produtos de mercearia

➔ **Variáveis promocionais**

FEAT: porcentagem do volume vendido do produto durante as semanas em que a marca foi anunciada no panfleto da loja ou no jornal local

DISP: porcentagem do volume vendido do produto durante as semanas em que a marca comprada recebeu exibição (suporte final) do varejista

PCUT: porcentagem do volume vendido do produto com uma redução temporária de preço

➔ Variáveis promocionais

MCOUP: porcentagem do volume vendido do produto com cupom promocional do fabricante

SCOUP: porcentagem do volume vendido do produto com cupom promocional da loja varejista

Objetivo: analisar se é possível agrupar as categorias de produtos em políticas promocionais diferentes

➔ **Passo a passo:**

1. Análise das variáveis e dos objetos a serem agrupados
2. Seleção da medida de similaridade
3. Seleção do algoritmo de agrupamento
4. Definição do número de agrupamentos
5. Interpretação dos agrupamentos

1. Análise das variáveis e dos objetos a serem agrupados

	FEAT	DISP	PCUT	SCOUP	MCOUP
BEER	1.44902441	3.41156077	2.09969557	0.6229513	-1.02344041
WINE	0.77624847	2.57488280	-0.44938134	-0.6908370	-1.02344041
FRESH_BRE	0.50713809	-0.49293641	0.48975226	0.6229513	-0.91482367
CUPCAKES	-0.56930341	0.34374156	-0.18105745	0.6229513	-0.69759019
MUFF/BAGE	1.31446922	1.04097320	1.16056197	1.9367396	-0.91482367
PIES/CAKE	-0.70385860	-0.49293641	-0.58354328	-0.6908370	-0.80620693
PASTRY	-0.70385860	-0.49293641	-0.44938134	-0.6908370	-0.69759019
SNACK_PAS	-0.43474822	0.34374156	-0.18105745	-0.6908370	-0.69759019
SDWCH_RO	1.04535884	-0.21404376	0.62391420	0.6229513	-1.02344041

Padronizar as variáveis

1. Análise das variáveis e dos objetos a serem agrupados

Há outliers?

Distância de Mahalanobis:

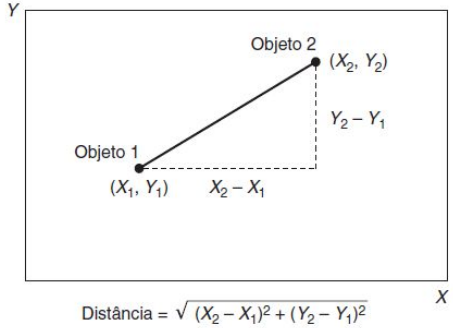
$$d_{ij} = \sqrt{(X_i - X_j)'.S^{-1}.(X_i - X_j)}$$

X_i = vetor de atributos do objeto i
X_j = vetor de atributos do objeto j
S = matriz de covariâncias

CANNED_HA	EGGS	SUGAR	DRY_SLD_M	CRB_SFT_D	FLOUR	DISP_DIAP
119.3003744	65.3686562	27.3553722	24.7555290	23.7249857	22.8301442	22.6813983
WINE	BCN/MT_ST	FRANKS	LNDRY_SAN	PIE_FILLI	DSHWSH_DE	CAN/FRG_F
21.0138776	21.0063215	20.5174066	19.4024653	17.4499023	16.7355275	16.3788766

2. Seleção da medida de similaridade

Distância euclideana



Distância = $\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$

	BEER	WINE	FRESH_BRE	CUPCAKES
BEER	0.00	3.06	4.33	4.34
WINE	3.06	0.00	3.48	2.95
FRESH_BRE	4.33	3.48	0.00	1.53
CUPCAKES	4.34	2.95	1.53	0.00

3. Seleção do algoritmo de agrupamento (método hierárquico)

Qual o melhor algoritmo?

Comparar os resultados usando a **correlação cofenética** - mensura quão bem o agrupamento reflete os dados

Compara as distâncias efetivamente observadas entre os objetos e as distâncias previstas a partir do processo de grupamento. Espera-se observar correlações altas

3. Seleção do algoritmo de agrupamento (método hierárquico)

Método de Ward

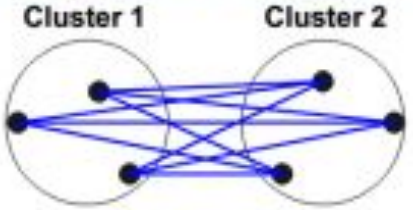
cor = 0,53

Medida de distância entre dois clusters é a soma das distâncias ao quadrado entre eles - agrupa clusters cuja soma das distâncias ao quadrado seja a menor possível.

Método da ligação média

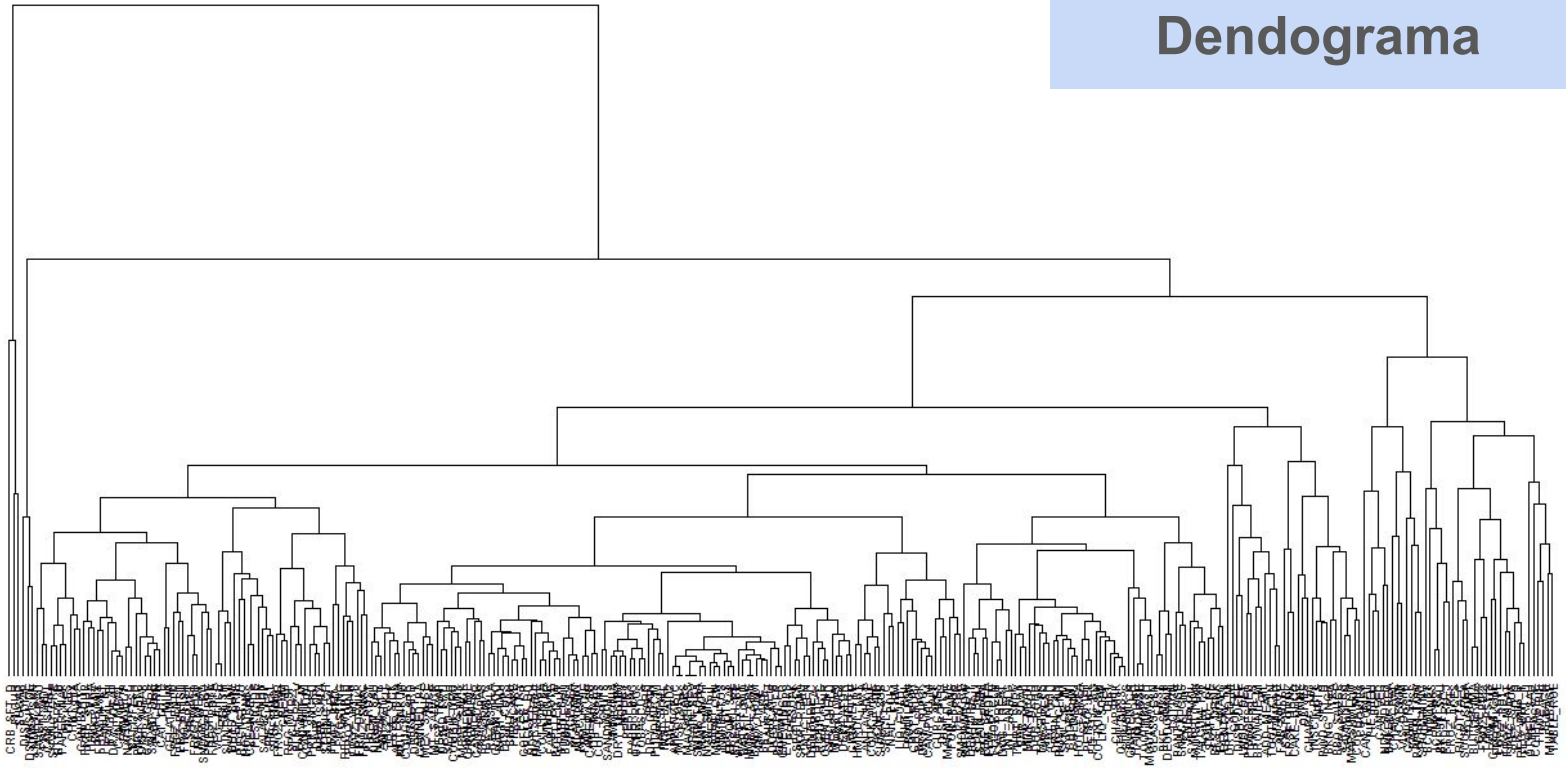
cor = 0,82

Considera a distância média de todos os objetos de um cluster com todos os objetos de outro cluster.

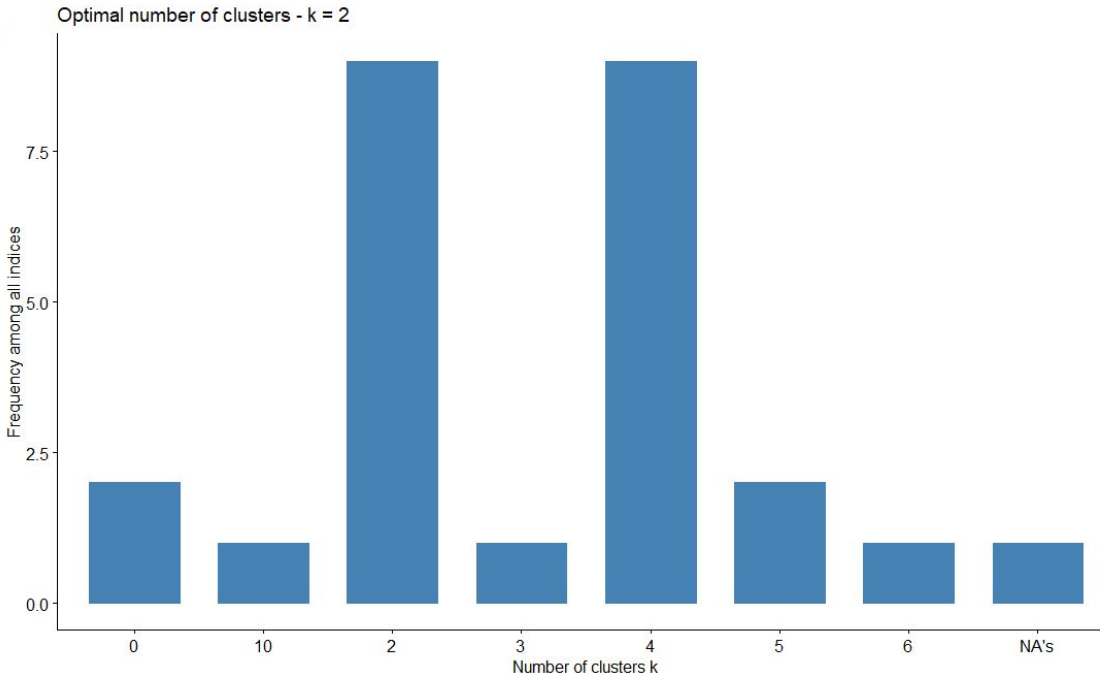


4. Definição do número de agrupamentos

Dendograma

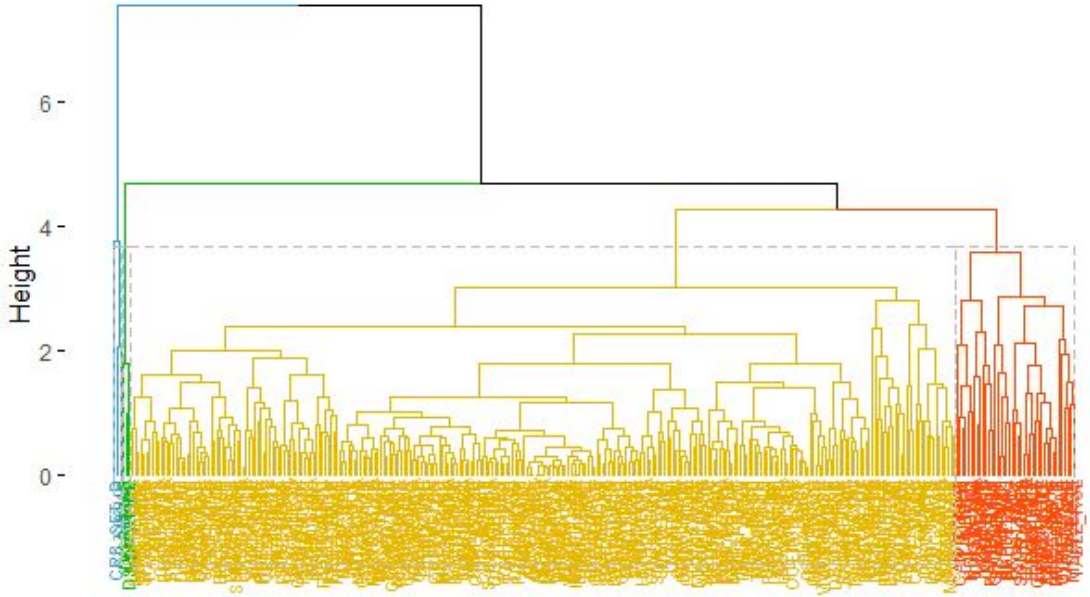


4. Definição do número de agrupamentos



Alternativa: NbClust

4. Interpretação dos agrupamentos



Obtenção de valores centróides

valor médio dos objetos contidos no agrupamento em cada variável

	[,1]	[,2]	[,3]	[,4]
FEAT	22.3	5.9	8.0	30.3
DISP	14.3	6.4	6.7	24.3
PCUT	24.7	9.2	9.3	28.7
SCOUP	1.4	0.4	0.3	4.7
MCOUP	6.5	10.6	49.0	5.0

DOI: 10.1590/1413-81232020253.01202018

Padrões alimentares de idosos no Brasil: Pesquisa Nacional de Saúde, 2013

Dietary patterns of the elderly in Brazil: National Heath Survey, 2013

Ingrid Freitas da Silva Pereira (https://orcid.org/0000-0001-6863-4227) [1]
Diôgo Vale (https://orcid.org/0000-0003-2636-4956) [1]
Mariana Silva Bezerra (https://orcid.org/0000-0002-5095-5804) [1]
Kenio Costa de Lima (https://orcid.org/0000-0002-5668-4398) [2]
Angelo Giuseppe Roncalli (https://orcid.org/0000-0001-5311-697X) [2]
Clélia de Oliveira Lyra (https://orcid.org/0000-0002-1474-3812) [3]