O envelhecimento populacional nos municípios do Sul/Sudoeste de Minas: análise de agrupamento usando componentes principais

Larissa Gonçalves
Orientadora: Profa. Dra. Patrícia Ramos
Coorientador: Prof. Dr. Lincoln Frias

Programa de Pós-Graduação em Estatística Aplicada e Biometria
Universidade Federal de Alfenas

Sumário

Transição demográfica

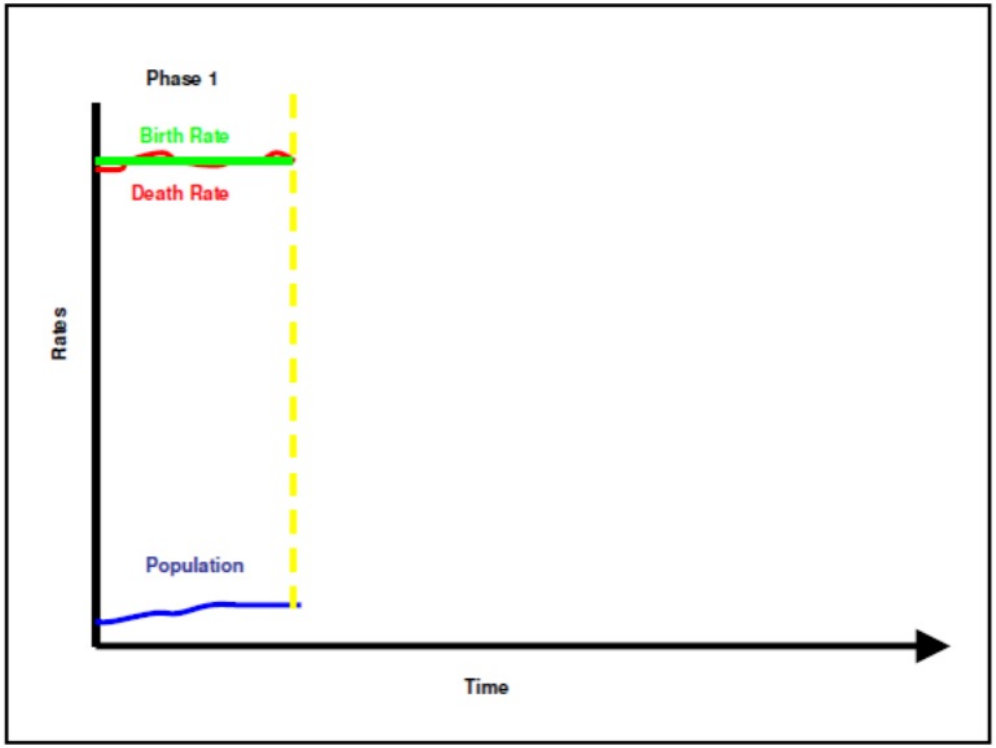


Figura: Primeira fase do processo de transição demográfica.

Transição demográfica

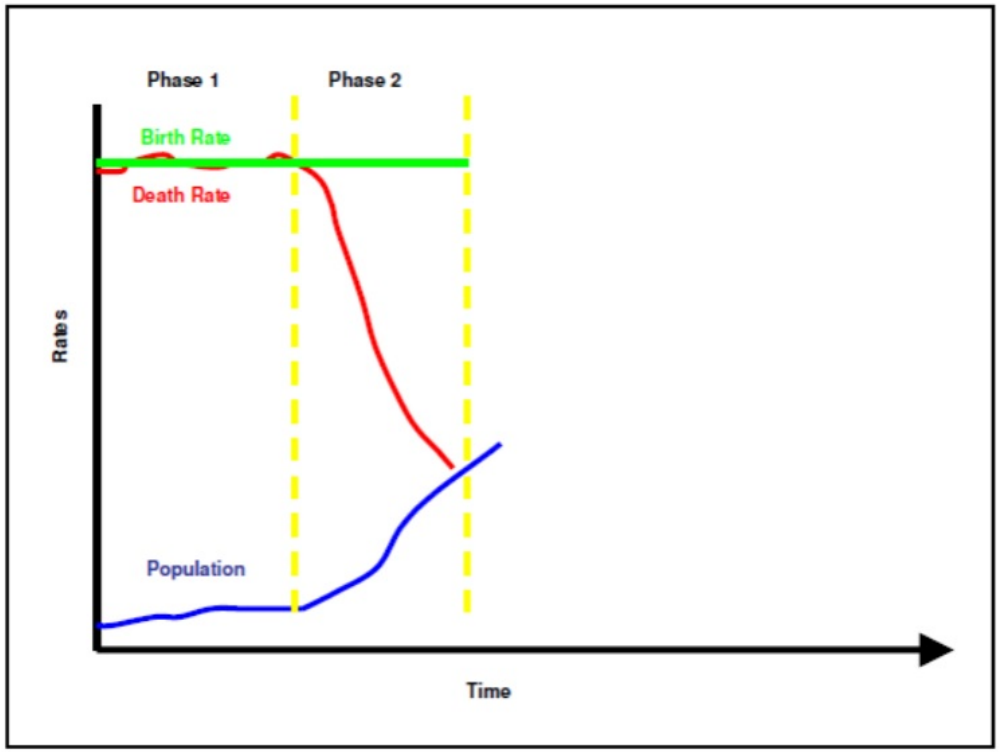


Figura: Segunda fase do processo de transição demográfica.

Transição demográfica

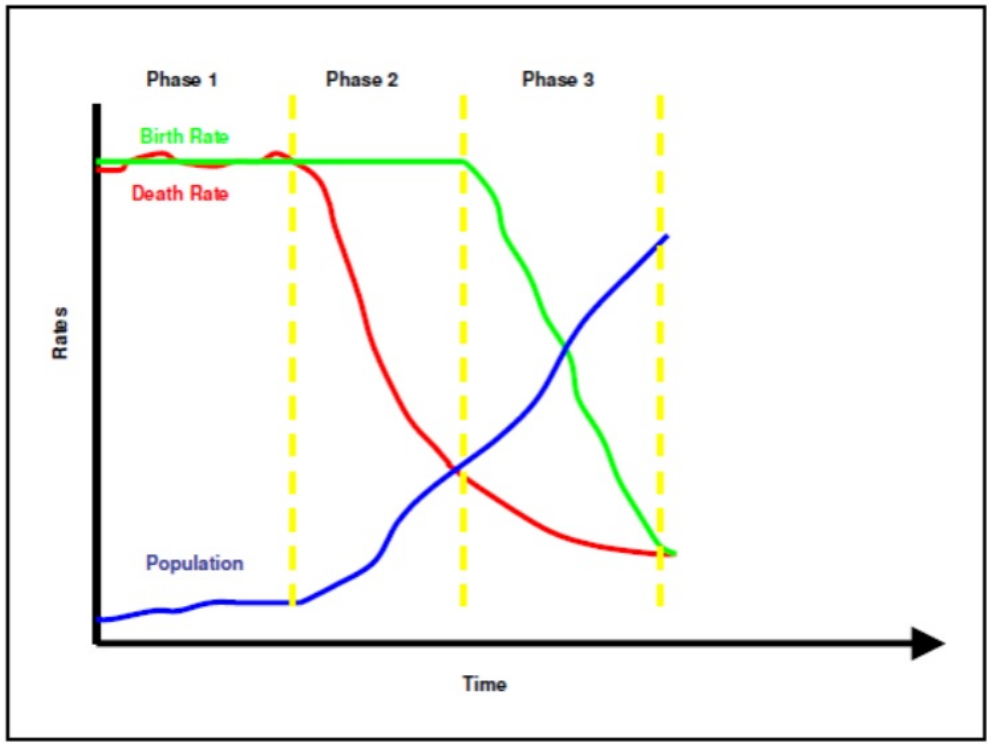


Figura: Terceira fase do processo de transição demográfica.

Transição demográfica

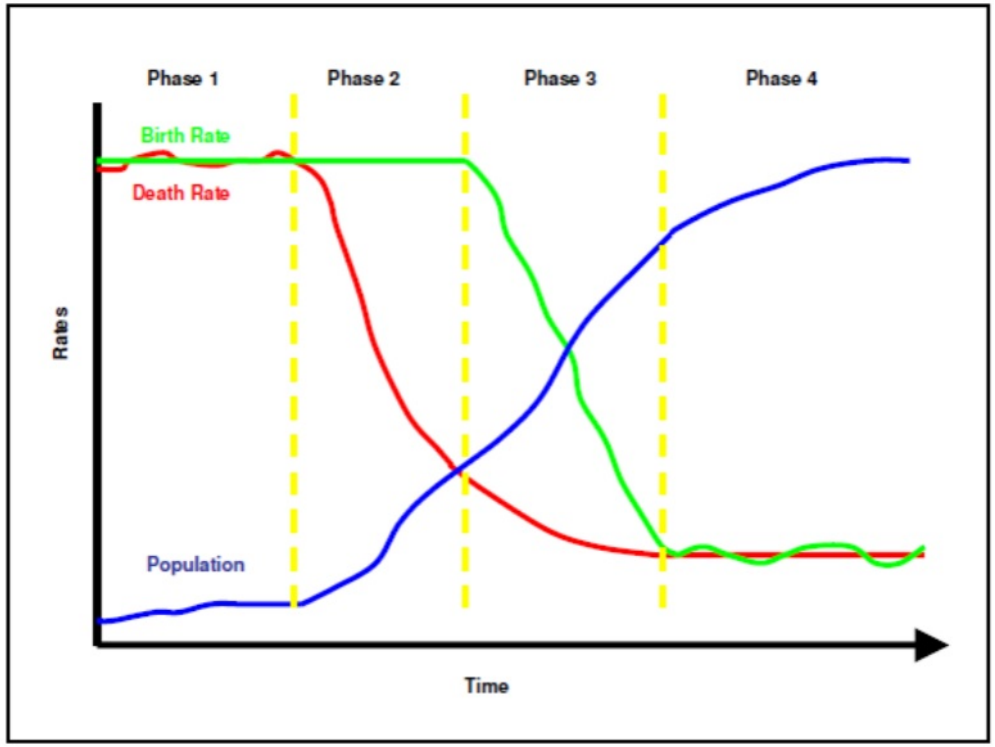


Figura: Quarta fase do processo de transição demográfica.

Envelhecimento populacional

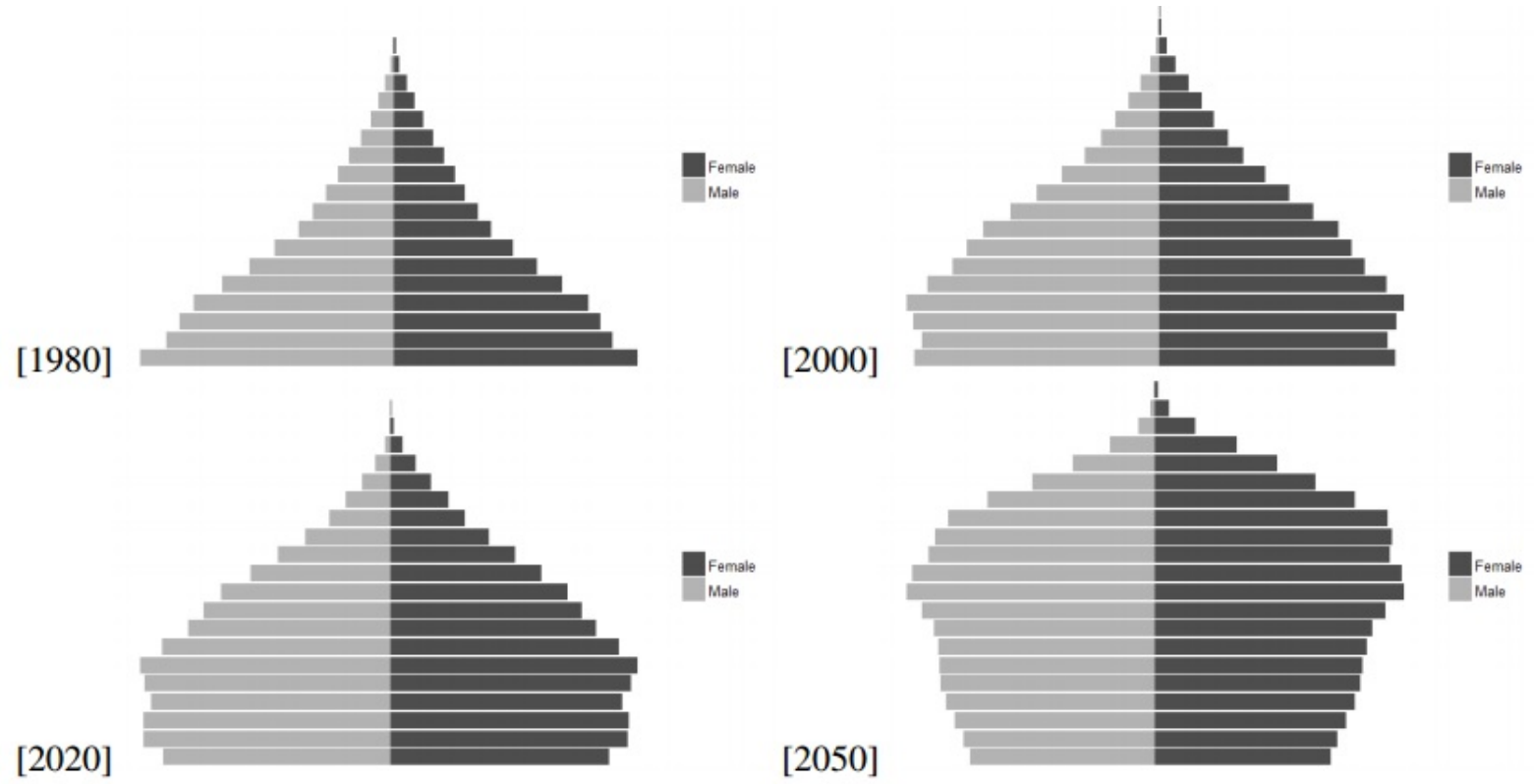


Figura: Pirâmides etárias absolutas do Brasil, 1980-2050.
Fonte: elaboração própria a partir de dados do *United States Census Bureau*, fonte disponível em: *www.census.gov/population/international/data/idb*

Envelhecimento populacional

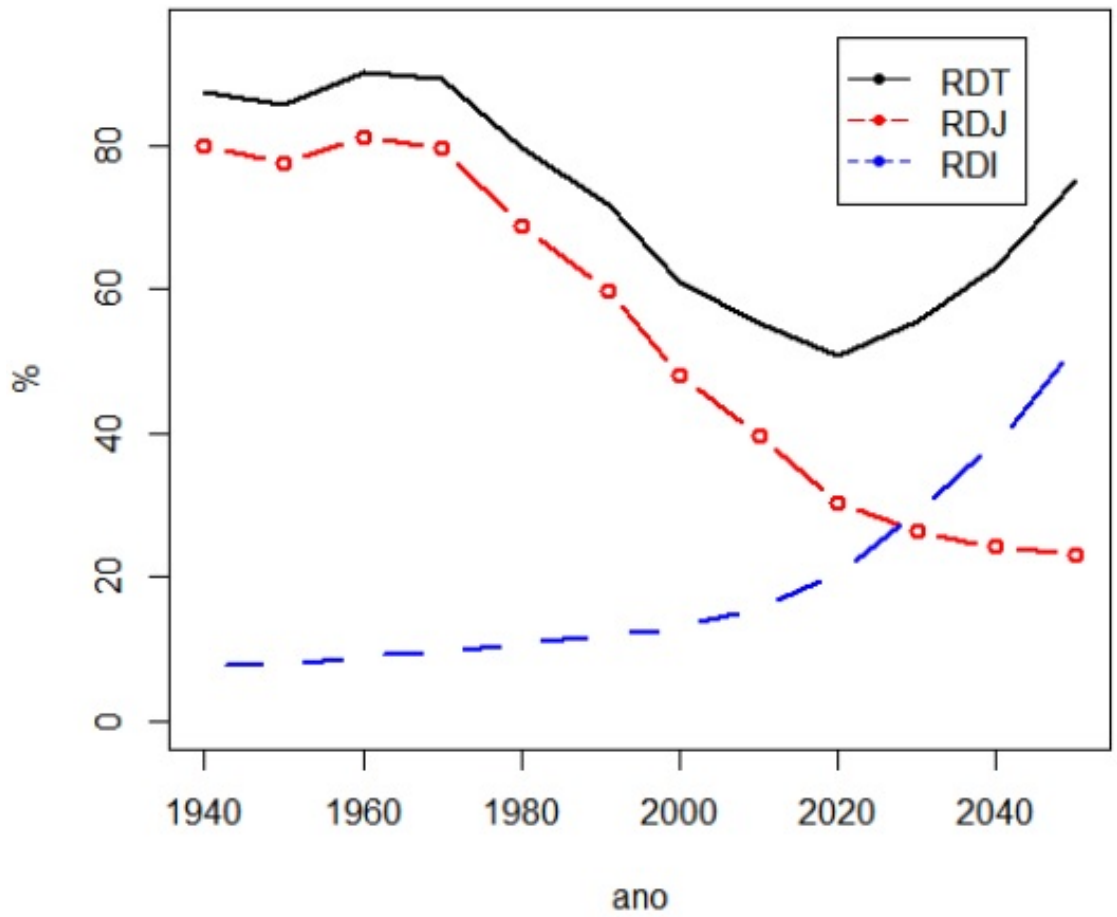


Figura: Razão de dependência do Brasil, 1940 a 2050.

Objetivos

- Propor uma classificação dos municípios mineiros, com ênfase na mesorregião Sul/Sudoeste, em relação ao processo de envelhecimento populacional.
- Utilizar diferentes métodos de agrupamento para comparação dos resultados.
- Analisar diferentes critérios para definição do número de grupos.

Análise multivariada

- Dados multivariados
- Conjunto de técnicas
- Representação de dados multivariados

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \tag{1}$$

Análise multivariada

- Vetor de médias amostral

$$\bar{\boldsymbol{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}.$$

- Matriz de covariâncias amostral

$$\boldsymbol{S} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}.$$

Análise multivariada

- Matriz de correlações amostral

$$\boldsymbol{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}, \tag{2}$$

em que $r_{ij} = \dfrac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}$

Análise multivariada

- Análise de componentes principais
- Análise de agrupamento (*clusters*)

Análise de componentes principais

- Transformar variáveis correlacionadas
- Explicar a estrutura de covariâncias das p variáveis, $\boldsymbol{X}^T = \begin{bmatrix} X_1 & X_2 & \dots & X_p \end{bmatrix}$
- Os componentes principais são as p combinações lineares obtidas, $\boldsymbol{Y}^T = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_p \end{bmatrix}$, não correlacionados entre si.

Análise de componentes principais

- Matriz de covariâncias ou da matriz de correlações das variáveis originais
- O primeiro componente principal Y_1 é a combinação linear

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$

Análise de componentes principais

A proporção da variância total de $\boldsymbol{X}$ explicada pelo i-ésimo componente principal é definida por:

$$\frac{Var(Y_i)}{\text{Variância total de } X} = \frac{\lambda_i}{\text{traço}(\boldsymbol{S})} = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

Os primeiros k componentes, em que $k < p$, explicam uma proporção da variância total dada por:

$$\frac{\sum_{i=1}^{k} Var(Y_i)}{\text{Variância total de } X} = \frac{\sum_{i=1}^{k} \lambda_i}{\text{traço}(\boldsymbol{S})} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j}. \quad (3)$$

Análise de componentes principais

Critérios para reter componentes

- Avaliar representatividade dos k primeiros componentes
- Scree plot
- Regra de Kaiser

Análise de agrupamento

Busca uma partição dos elementos de uma amostra em grupos de tal forma que:

(a) as observações de um mesmo grupo sejam similares entre si em relação às variáveis medidas

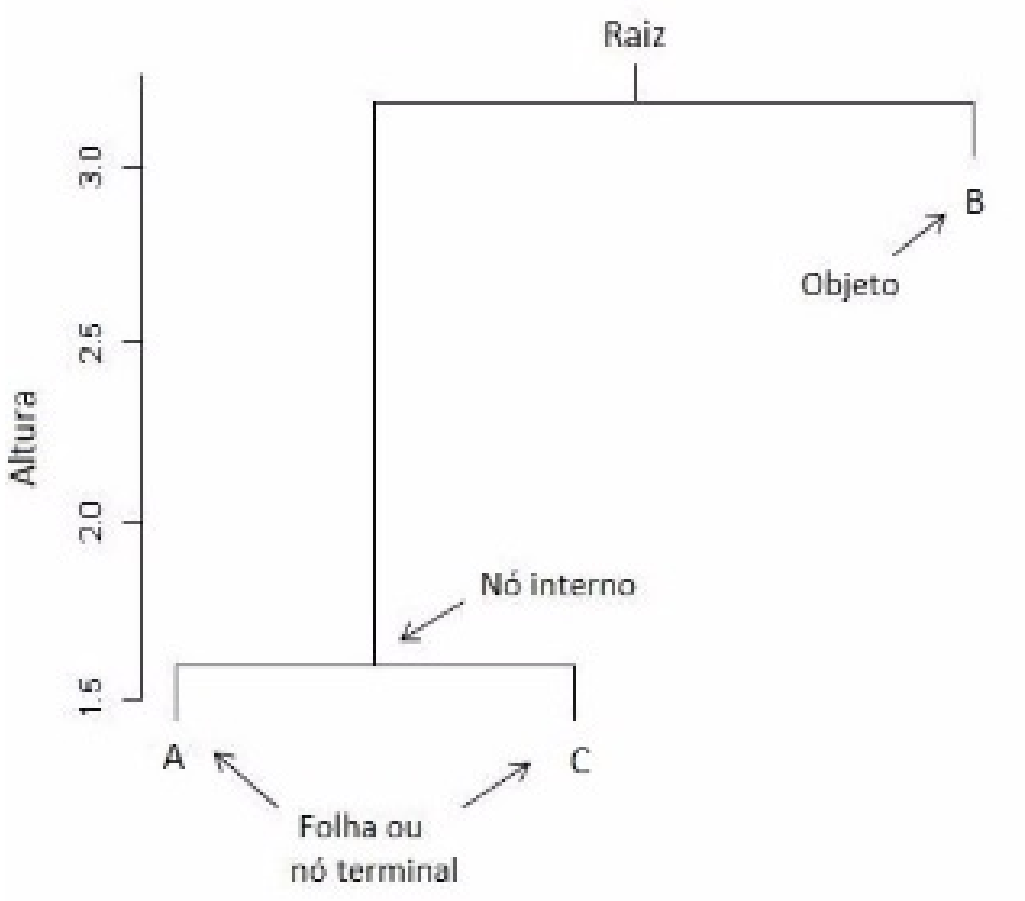
(b) as observações de grupos diferentes sejam heterogêneas entre si em relação a essas mesmas variáveis

Análise de agrupamento

- Métodos hierárquicos: aglomerativos e divisivos
- Métodos não hierárquicos

Métodos hierárquicos

- Processo tem uma hierarquia
- Dendrograma mostra a história do agrupamento



Métodos hierárquicos aglomerativos

- n grupos de tamanho 1;
- Em cada passo é formado apenas um novo grupo;

Métodos hierárquicos divisivos

- 1 grupo constituído dos *n* elementos amostrais observados;
- Grupo inicial vai sendo subdividido;

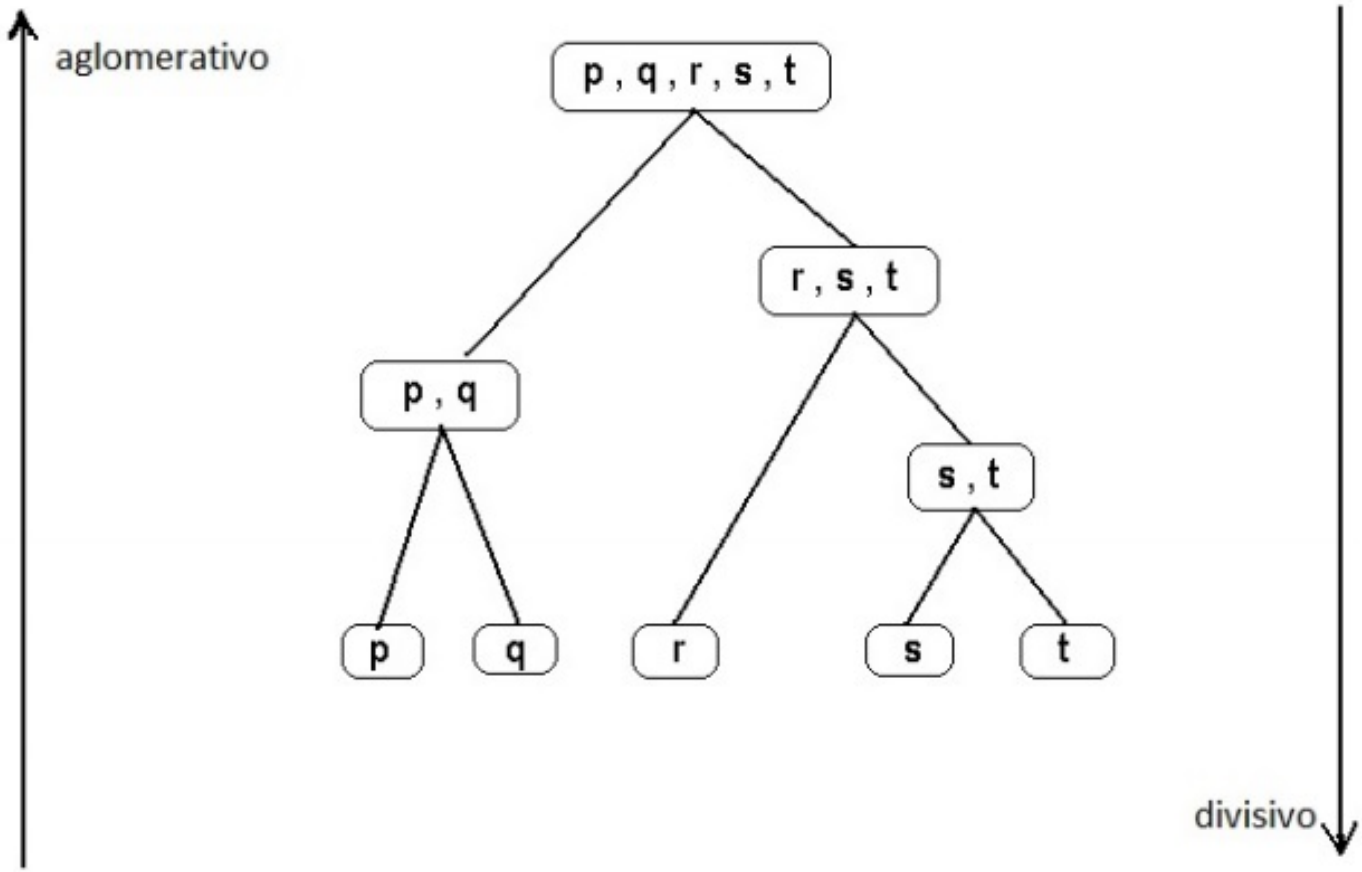


Figura: Ilustração dos métodos hierárquicos aglomerativos e divisivos

Análise de agrupamento

Medidas de distância

- Matriz de distâncias $n \times n$
- Dois elementos são considerados próximos quando sua distância é pequena.
- Exemplo de matriz de distâncias para $n = 4$

$$\begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{21} & 0 & d_{23} & d_{24} \\ d_{31} & d_{32} & 0 & d_{34} \\ d_{41} & d_{42} & d_{43} & 0 \end{bmatrix}$$

Análise de agrupamento

Medidas de distância

Distância generalizada:

$$d_{ij} = (\boldsymbol{X_{i.}} - \boldsymbol{X_{j.}})^T \boldsymbol{A} (\boldsymbol{X_{i.}} - \boldsymbol{X_{j.}})$$

- Quando $\boldsymbol{A} = \boldsymbol{I}$ temos a distância euclidiana
- Quando $\boldsymbol{A} = \boldsymbol{D}^{-1}$ temos a distância padronizada
- Quando $\boldsymbol{A} = \boldsymbol{S}^{-1}$ temos a distância de Mahalanobis

Métodos de análise agrupamento

- Métodos de agrupamentos hierárquicos aglomerativos
  1. Método da ligação simples (vizinho mais próximo);
  2. Método de ligação completa (vizinho mais distante);
  3. Método da média das distâncias;
  4. Método do centróide;
  5. Método de Ward;

Análise de agrupamento

Processo iterativo geral

Processo iterativo geral:

Passo 1. Inicialmente a distância entre dois grupos é obtida através da distância entre dois objetos:

$$d_{G_i G_j} = d_{ij}$$

Análise de agrupamento

Processo iterativo geral

Processo iterativo geral:

Passo 1. Inicialmente a distância entre dois grupos é obtida através da distância entre dois objetos:

$$d_{G_i G_j} = d_{ij}$$

Passo 2. Selecionar na matriz de distâncias os dois grupos G_i e G_j que possuem **menor distância**;

Análise de agrupamento

Processo iterativo geral

Processo iterativo geral:

Passo 1. Inicialmente a distância entre dois grupos é obtida através da distância entre dois objetos:

$$d_{G_i G_j} = d_{ij}$$

Passo 2. Selecionar na matriz de distâncias os dois grupos G_i e G_j que possuem **menor distância**;

Passo 3. Unir os dois grupos G_i e G_j em um novo;

Análise de agrupamento

Processo iterativo geral

Passo 4. Definir a distância entre o novo agrupamento e todos os agrupamentos, de acordo com a técnica hierárquica aglomerativa escolhida;

Análise de agrupamento

Processo iterativo geral

Passo 4. Definir a distância entre o novo agrupamento e todos os agrupamentos, de acordo com a técnica hierárquica aglomerativa escolhida;

Passo 5. Reiniciar o processo a partir do passo 2 até que se chegue a um agrupamento final.

Métodos de agrupamentos

1. Método da Ligação Simples (vizinho mais próximo):

A similaridade entre dois grupos é definida pelos dois elementos mais parecidos entre si. Dessa forma, a distância entre dois grupos A e B é definida como:

$$d_{AB} = min\{d_{ij}\} \quad i \in A, j \in B$$

Métodos de agrupamentos hierárquicos

2. Método de Ligação Completa (vizinho mais distante):
A distância entre dois agrupamentos é definida como o máximo entre as distâncias calculadas para os pares de grupos, de tal forma que para dois grupos A e B a distância é definida como:

$$d_{AB} = max\{d_{ij}\} \quad i \in A, j \in B$$

3. Método da média das distâncias: A distância entre dois agrupamentos é definida como a média das distâncias entre todos os pares de objetos

$$d_{A,B} = \frac{1}{n_A n_B} \sum_{i \epsilon A} \sum_{j \epsilon B} d_{ij}$$

em que n_A e n_B são os números de observações nos grupos A e B.

Métodos de agrupamentos hierárquicos

4. Método do centróide: A distância entre dois agrupamentos é definida a partir da distância entre dois centróides. Então a distância entre dois grupos A e B é definida como:

$$d_{AB} = (\overline{\boldsymbol{X_A}} - \overline{\boldsymbol{X_B}})^T (\overline{\boldsymbol{X_A}} - \overline{\boldsymbol{X_B}})$$

Métodos de agrupamentos hierárquicos

5. Método de Ward ou mínima variância:

- Busca a menor soma de quadrados mínimos dentro do grupo

Métodos de agrupamentos hierárquicos

5. Método de Ward ou mínima variância:

- Busca a menor soma de quadrados mínimos dentro do grupo

1. Cada elemento é considerado como um único agrupamento

Métodos de agrupamentos hierárquicos

5. Método de Ward ou mínima variância:

- Busca a menor soma de quadrados mínimos dentro do grupo

1. Cada elemento é considerado como um único agrupamento

2. Em cada passo do algoritmo de agrupamento calcula-se a soma de quadrados dentro de cada grupo

Métodos de agrupamentos hierárquicos

5. Método de Ward ou mínima variância:

- Busca a menor soma de quadrados mínimos dentro do grupo

1. Cada elemento é considerado como um único agrupamento

2. Em cada passo do algoritmo de agrupamento calcula-se a soma de quadrados dentro de cada grupo

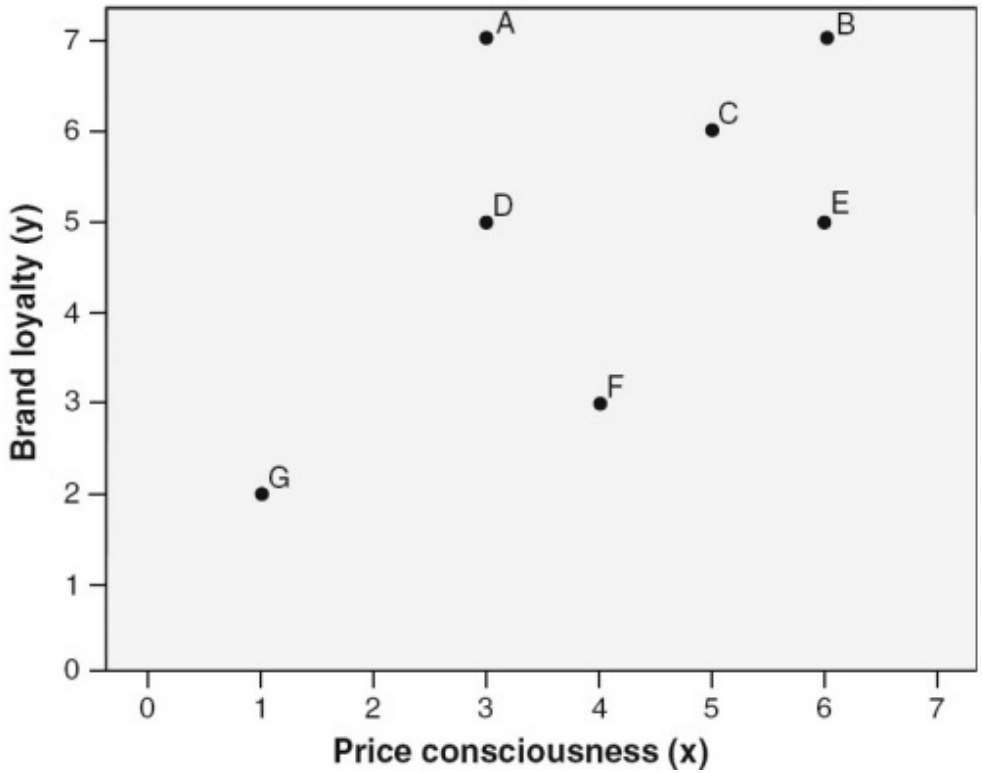
3. Combinam-se os dois grupos que resultarem no menor valor de soma de quadrados

Métodos não hierárquicos

- Definir o número k de grupos inicialmente
- As observações podem ser movidas para dentro ou fora dos grupos
- Não é possível construir dendrogramas
- Processo é aplicado à matriz de dados
- Método *Fuzzy* c-médias, Redes neurais artificiais e k-médias

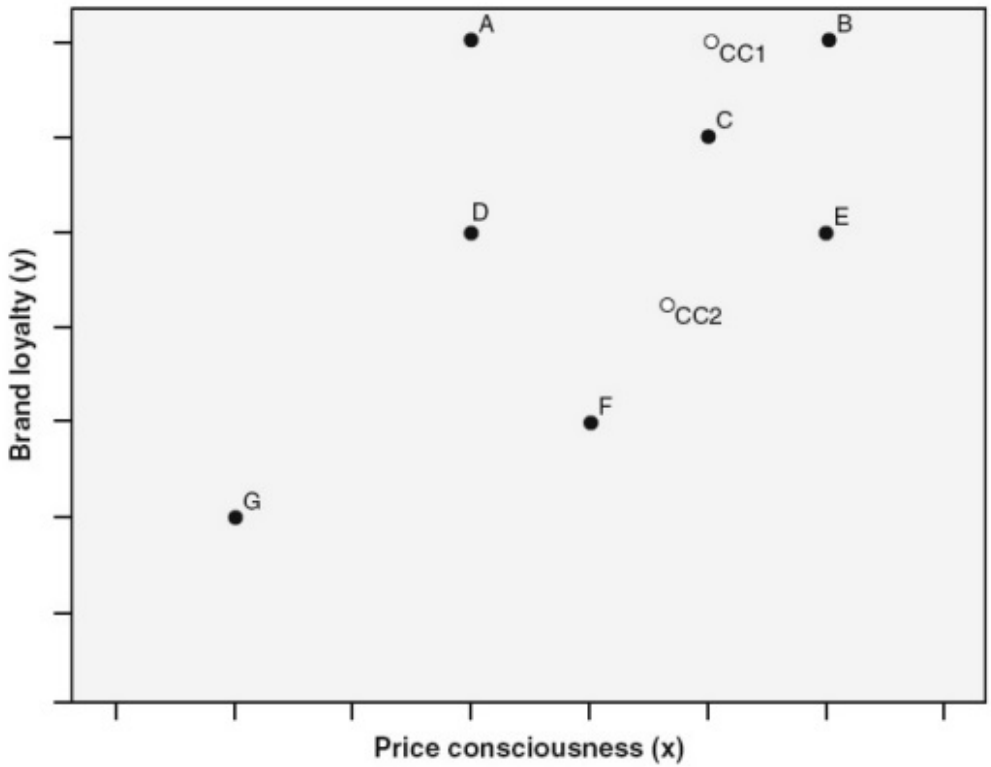
Método das k-Médias

O mercado de uma marca será separado de acordo com duas variáveis: X - consciência sobre o preço e Y - fidelidade à marca.



Método das k-Médias

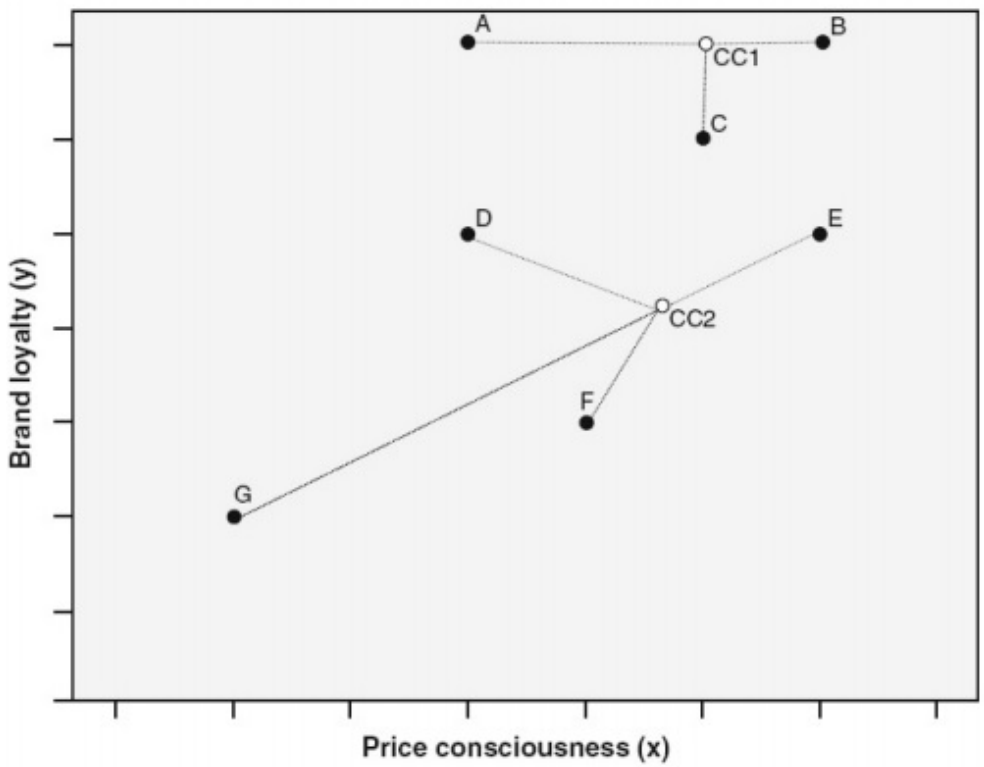
Passo 1 O valor de k definido como 2 grupos de clientes. O algoritmo seleciona aleatoriamente um centro para cada grupo - CC1 e CC2.



Método das k-Médias

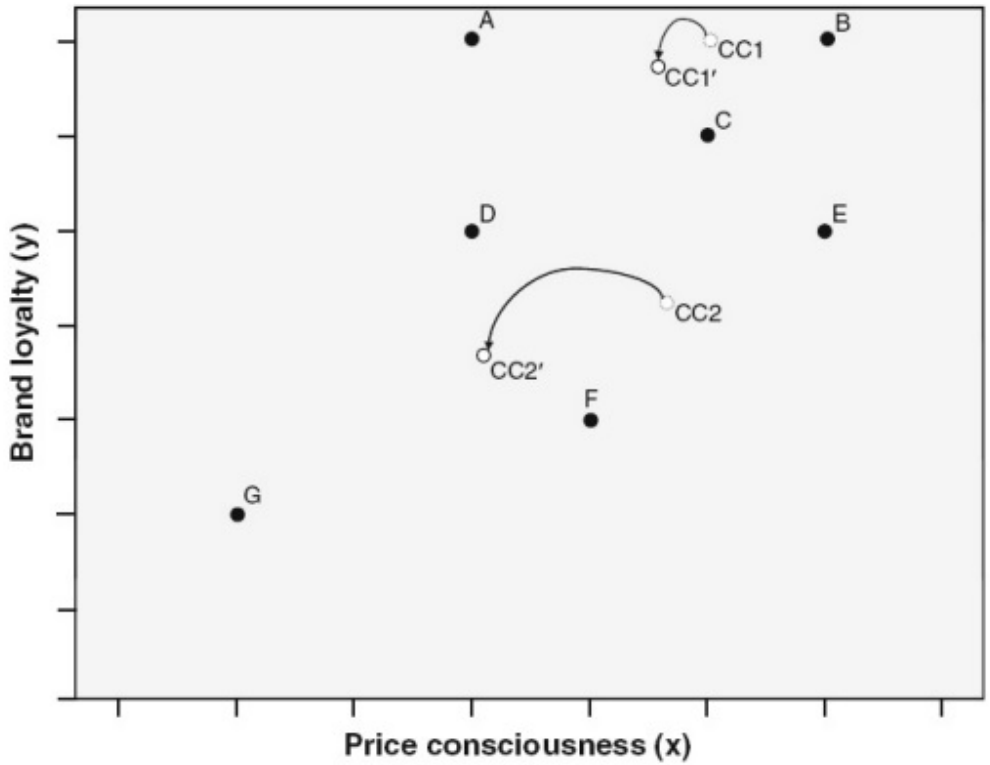
Passo 2

Distâncias são calculadas dos centros para cada observação. Cada observação ficará associada ao centro para o qual a distância é menor (A, B, C:1/ D, E, F, G:2)



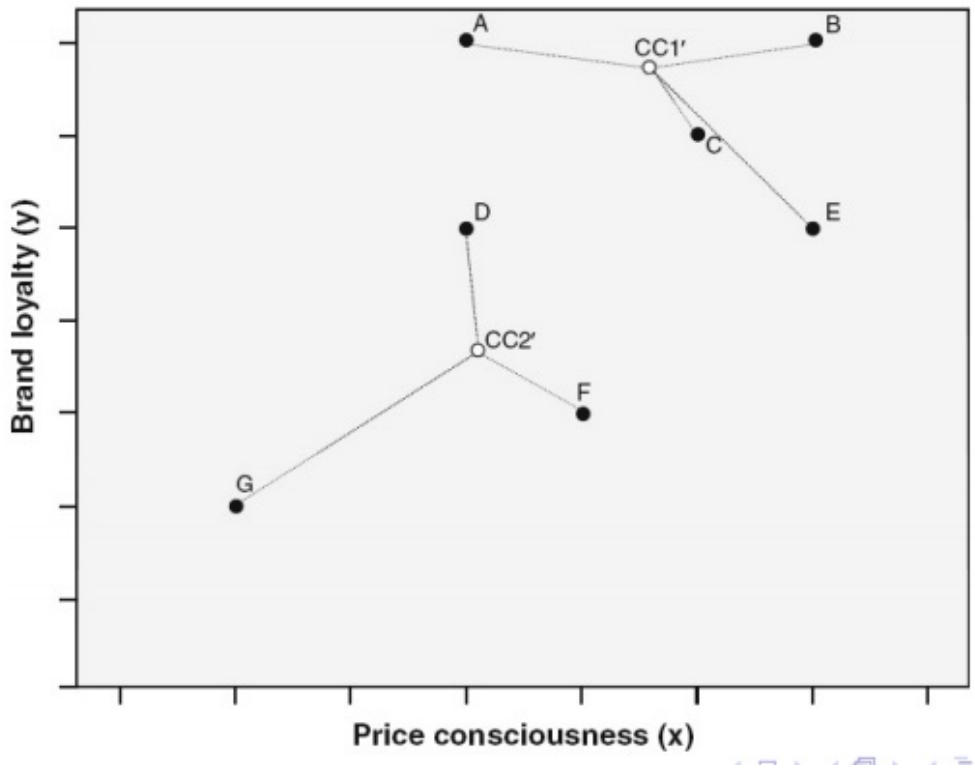
Método das k-Médias

Passo 3 Baseando-se na partição realizada, cada centróide é calculado (valores médios das obs. em cada grupo). CC1 e CC2 mudam de lugar.



Método das k-Médias

Passo 4 Distâncias das observações para os novos centros são calculadas e as observações são atribuídas aos grupos para os quais a distância para o centro é menor.



Escolha do número de grupos

- Escolha está sujeita a dois tipos de erros diferentes
- Alguns critérios

Escolha do número de grupos

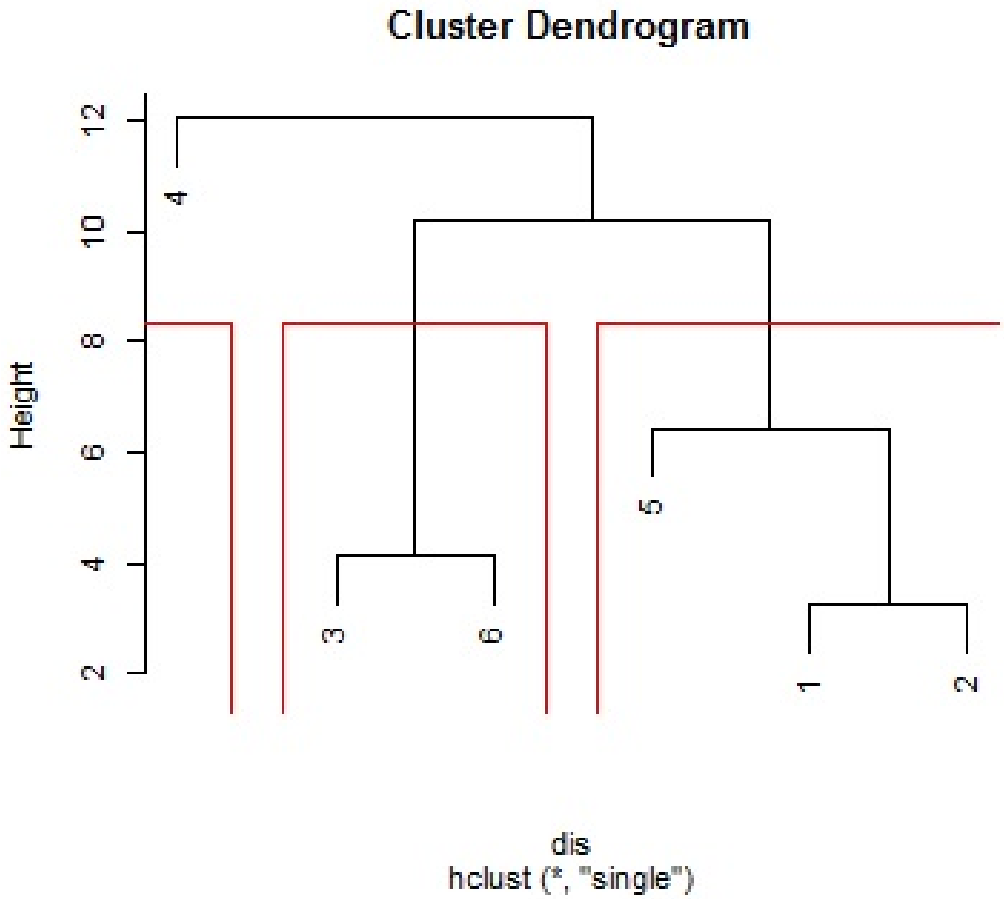


Figura: Ilustração de corte no dendrograma

Escolha do número de grupos

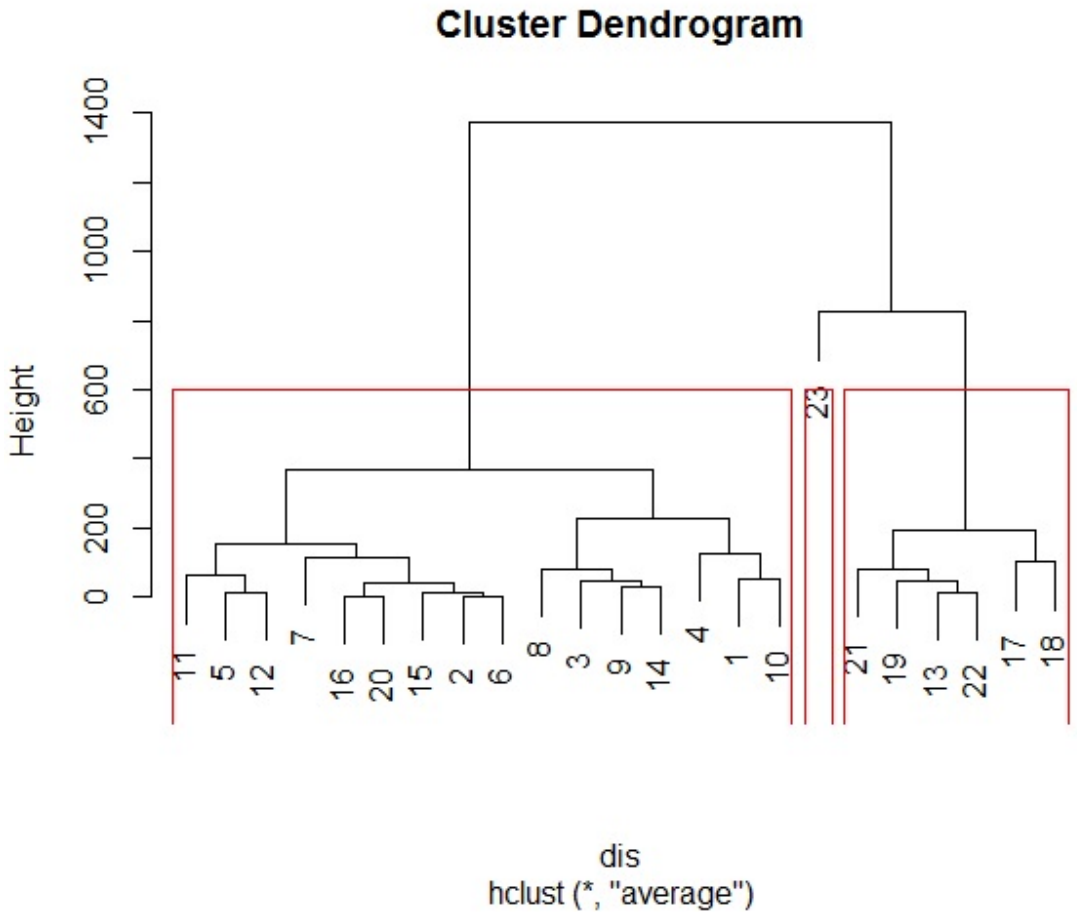


Figura: Ilustração de corte no dendrograma

Escolha do número de grupos

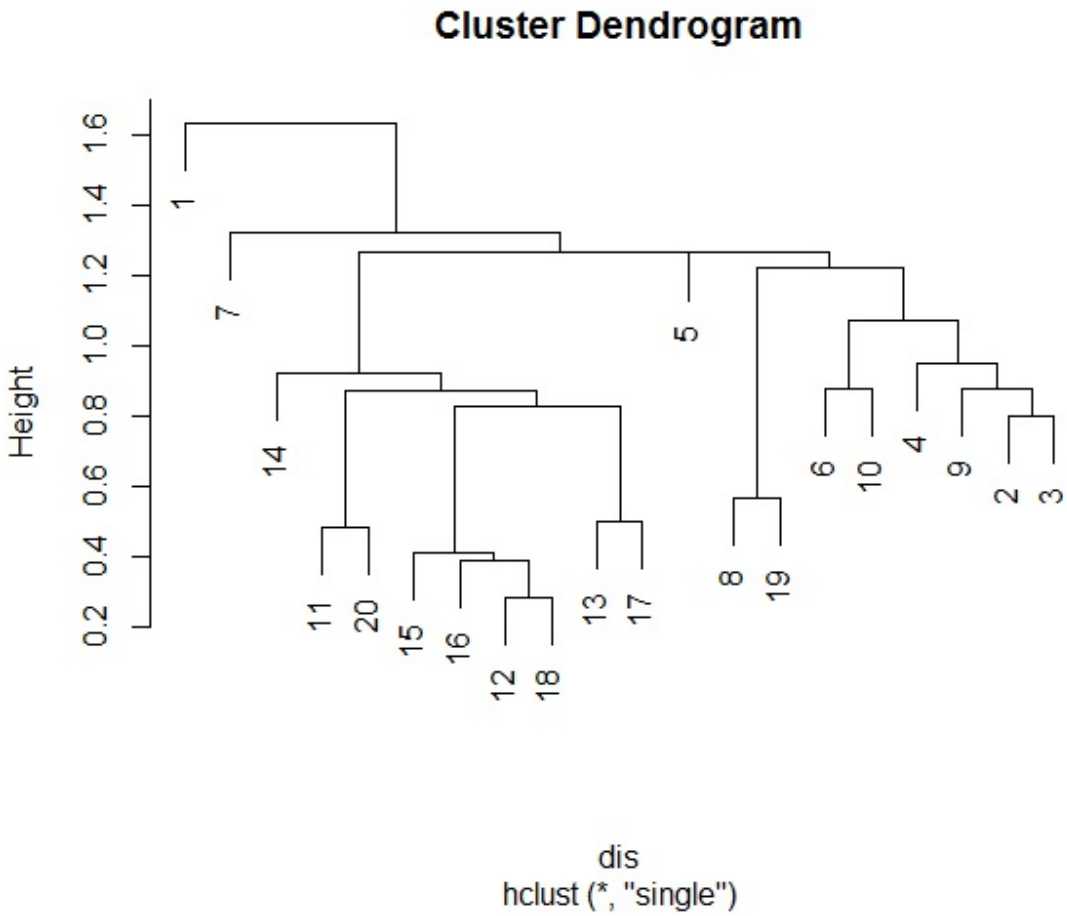


Figura: Dendrograma

Escolha do número de grupos

- Número de grupos obtidos após a aplicação de uma técnica hierárquica aglomerativa
- Pseudo F
- Pseudo T^2
- Estatística GAP

Escolha do número de grupos

Pseudo F

- Cálculo da estatística F em cada passo do processo de agrupamento

$$F = \frac{SQE/(g* - 1)}{SQD/(n - g*)},$$

em que $SQE = \sum_{i=1}^{g*} n_i(\bar{X}_{i.} - \bar{X})^T(\bar{X}_{i.} - \bar{X})$

$$SQD = \sum_{i=1}^{g*} \sum_{j=1} n_i(X_{ij} - \bar{X}_{i.})^T(X_{ij} - \bar{X}_{i.})$$

- Teste F de análise de variância em cada passo
- Comparação dos vetores de médias dos grupos
- Pseudo F não aumenta de forma monótona
- Acima de determinado k pode ocorrer decréscimo na estatística
- Quanto maior o valor de F, menor valor-p e mais H_0 é rejeitada. Portanto, maior a heterogeneidade entre grupos diferentes.

Dados e metodologia

- Mesorregião Sul/Sudoeste de Minas Gerais é formada 146 municípios
- Dados do censo demográfico 2010 do IBGE

Tabela: Lista de variáveis

Esperança de vida ao nascer
Taxa de fecundidade total
Mortalidade infantil
Mortalidade até 5 anos de idade
Razão de dependência
Probabilidade de sobrevivência até 40 anos
Probabilidade de sobrevivência até 60 anos
Taxa de envelhecimento
População total

Dados e metodologia

- Aplicar análise de componentes principais
- Diferentes métodos de agrupamento
- Critérios de escolha do número de grupos

Resultados esperados

- Como os municípios mineiros estão agrupados em relação às variáveis relacionadas ao envelhecimento populacional
- Identificar padrões nos dados e que semelhanças e diferenças entre os municípios sejam detectadas

Referências bibliográficas

BARTHOLOMEW, D. J.; STEELE, F.; GALBRAITH, J.; MOUSTAKI, I. **Analysis of multi-variate social science data**. Boca Raton: CRC press, 2008.

CAMARANO, A. A. O. **Novo regime demográfico: uma nova relação entre população edesenvolvimento?**[S.l.]: Instituto de Pesquisa Econômica Aplicada (Ipea), 2014.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster analysis**. 5. ed. UK: JohnWiley and Sons, 2011.

FERREIRA, D. F. **Estatística multivariada**. 2. ed. Lavras: Editora UFLA, 2011.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: umaabordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.