Análise de Conglomerados

Ludmilla Viana Jacobson
ludmilla@est.uff.br

Objetivo

Dividir os elementos da amostra, ou população, em grupos de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características.

Aplicações

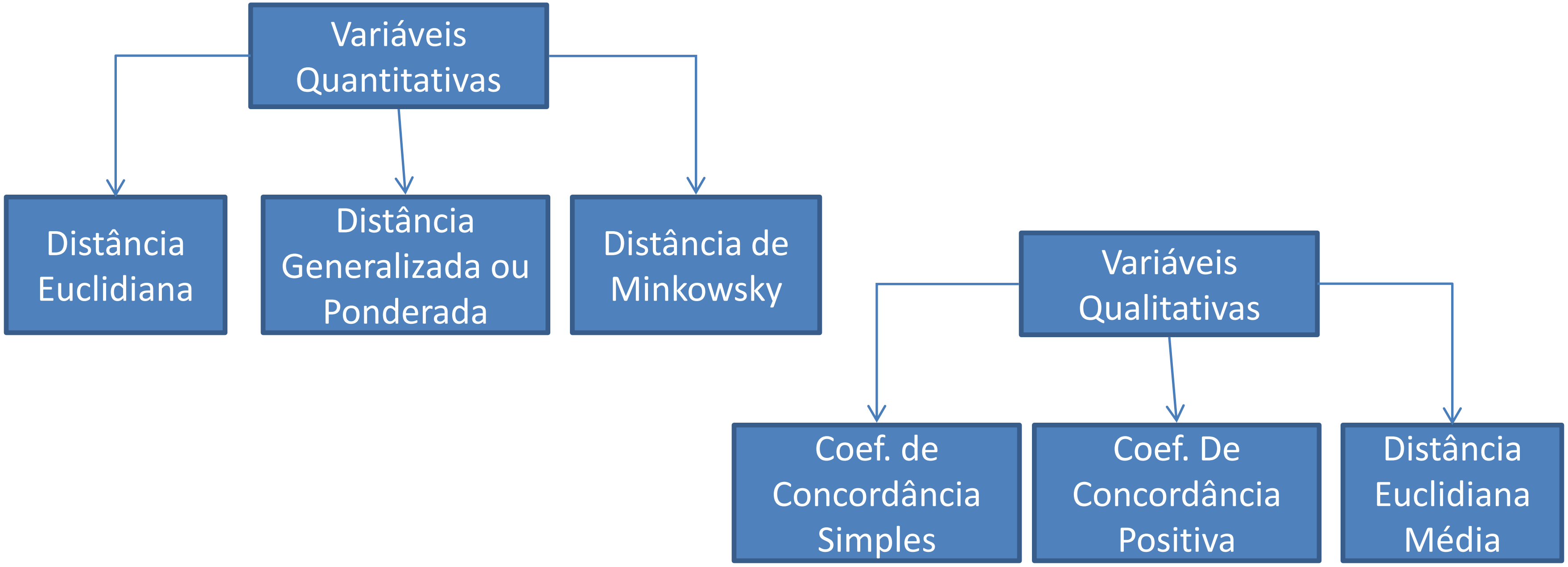
- Psicologia – é utilizada na classificação de pessoas de acordo com seus perfis de personalidade;
- Pesquisa de Mercado – segmentação de clientes de acordo com perfis de consumo;
- Ecologia – na classificação de espécies;
- Geografia – na classificação de cidades, estados ou regiões de acordo com variáveis físicas, demográficas e econômicas.

Passos importantes para a definição dos Grupos

- Variáveis para a formação dos grupos;
- Medida de similaridade ou dissimilaridade (distância), que expressam o grau de semelhança entre os objetos;
- Critérios ou técnicas para a construção dos conglomerados.

Medida de Distância

A escolha da medida adequada vai depender da natureza qualitativa ou quantitativa dos atributos que caracterizam os objetos/elementos/indivíduos.



Medida de Distância: Distância Euclidiana

Suponha que se tenha disponível um conjunto de dados constituído de n elementos amostrais, tendo-se medido p-variáveis aleatórias em cada um deles. O objetivo é agrupar esses elementos em g grupos. Para cada elemento amostral j, tem-se, portanto, o vetor de medidas definido por (j = 1, 2,..., n) :

$$\underset{\sim}{X}_j = \left[X_{1j} \quad X_{2j} \quad \ldots \quad X_{pj} \right]^T$$

$$d(X_l, X_k) = \left[(X_l - X_k)^T (X_l - X_k) \right]^{1/2} = \left[\sum_{i=l}^{p} (X_{il} - X_{ik})^2 \right]^{1/2}$$

Medida de Distância: Distância Euclidiana

Exemplo:

Indivíduo	Renda	Idade
A	9,60	28
B	8,40	31
C	2,40	42
D	18,20	38
E	3,90	25
F	6,40	41

Fonte: Mingoti, 2005

$$d(X_A, X_B) = \sqrt{(9{,}60 - 8{,}40)^2 + (28 - 31)^2} = 3{,}23$$

Quanto menor os seus valores, mais similares serão os elementos que estão sendo comparados.

Medida de Distância: Distância Euclidiana

Exemplo:

Matriz de Distâncias entre os seis elementos amostrais

$$\begin{array}{cc}
 & \begin{matrix} A & B & C & D & E & F \end{matrix} \\
\begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} &
\begin{bmatrix}
0 & & & & & \\
3{,}23 & 0 & & & & \\
15{,}74 & 12{,}53 & 0 & & & \\
13{,}19 & 12{,}04 & 16{,}29 & 0 & & \\
6{,}44 & 7{,}50 & 17{,}06 & 19{,}33 & 0 & \\
13{,}39 & 10{,}19 & 4{,}12 & 12{,}18 & 16{,}19 & 0
\end{bmatrix}
\end{array}$$

Medida de Distância: Distância Generalizada ou Ponderada Ou Distância Estatística

$$d(X_l, X_k) = \left[(X_l - X_k)^T A \ (X_l - X_k)\right]^{1/2}$$

Onde A_{pxp} é uma matriz de ponderação.

- Quando A_{pxp} é uma matriz identidade, a distância generalizada é a distância euclidiana;
- Quando A_{pxp} é igual a S^{-1}_{pxp} , tem-se a distância de *Mahalanobis*;
- Quando A_{pxp} = diag (1/p), tem-se distância euclidiana média.

Medida de Distância: Distância de Minkowsky

$$d(X_l, X_k) = \left[\sum_{i=l}^{p} w_i \left| X_{il} - X_{ik} \right|^m \right]^m$$

Onde w_i`s são os pesos de ponderação para as variáveis.

- Casos mais usados m=1 e m=2.
- Para m=1 esta distância é conhecida como *city-block* ou *Manhattan*, e para m=2 tem-se a distância euclidiana.

Medida de Distância: Coeficiente de Concordância Simples

Exemplo:

Variável	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Elemento 1	0	1	1	1	1	0	1	0	0	0
Elemento 2	0	0	1	1	1	0	1	1	0	0

$$S(1,2) = \frac{\textit{número de pares concordantes}}{\textit{número total de pares}} = \frac{8}{10} = 0{,}80$$

Quanto maior o valor de s(.), maior a similaridade entre os elementos que estão sendo comparados.

Medida de Distância:
Coeficiente de Concordância Positiva

Exemplo:

Variável	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Elemento 1	0	1	1	1	1	0	1	0	0	0
Elemento 2	0	0	1	1	1	0	1	1	0	0

$$S(1,2) = \frac{\textit{número de pares concordantes do tipo } (1 \;\; 1)}{\textit{número total de pares}} = \frac{4}{10} = 0{,}40$$

Quanto maior o valor de s(.), maior a similaridade entre os elementos que estão sendo comparados.

Medida de Distância: Euclidiana Média

Exemplo:

Variável	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Elemento 1	0	1	1	1	1	0	1	0	0	0
Elemento 2	0	0	1	1	1	0	1	1	0	0

$$d(1,2) = \left(\frac{\text{número de pares disconcordantes}}{\text{número total de pares}} \right)^{1/2} = \left(\frac{2}{10} \right)^{1/2} = (0{,}20)^{1/2} = 0{,}45$$

Quanto menor o valor da distância, maior será a similaridade dos elementos comparados.

Métodos de Análise de *Cluster*

- Métodos Hierárquicos – são utilizadas em análises exploratórias dos dados com o intuito de identificar possíveis agrupamentos e o valor provável do número de grupos g. Os clusters são formados através de um processo iterativo.

- Métodos Não-hierárquicos – é necessário que o valor do número de grupos já esteja pré-especificado pelo pesquisador.

Métodos Hierárquicos

Métodos Hierárquicos

- Aglomerativo – cada elemento amostral do conjunto de dados observado é considerado como sendo um *cluster*, que sucessivamente sofre uma série de fusões com outros *clusters* até que no final todos os elementos estejam em um único *cluster*.

- Divisivo – no início há apenas um *cluster* formado pelo conjunto de elementos que é dividido sucessivamente até que no final cada *cluster* contenha apenas um elemento.

Método Hierárquico Aglomerativo

Algoritmo para formar *clusters* em um conjunto de n elementos:

- 1) Inicie com n clusters, cada um contendo apenas um objeto e construa a matriz de distâncias de ordem n.

- 2) Identifique o menor elemento da matriz de distâncias para encontrar o par de clusters mais similares.

- 3) Reúna os dois clusters identificados na etapa 2 em um único cluster e atualize a matriz de distâncias, retirando as linhas e colunas relativas aos dois clusters identificados em 2 e incluindo a linha e coluna com as distâncias entre os demais clusters e o novo cluster formado. Note que a ordem da matriz de distâncias diminui de uma unidade a cada vez que a etapa 3 é executada.

- 4) Repita os passos 2 e 3 até que reste apenas um cluster. A cada iteração guarde a identificação dos clusters que foram fundidos e também a distância entre eles, estas informações serão utilizadas na montagem de um gráfico conhecido como dendograma, que mostra a seqüência de aglomeração dos clusters.

Método Hierárquico Aglomerativo

Critérios para Determinação dos *Clusters* ou Métodos de Encadeamento

- Distância Mínima (*Single Linkage*)
- Distância Máxima (*Complete Linkage*)
- Método de Ward (variância mínima)
- Distância Média (*Average Linkage*)

Método Hierárquico Aglomerativo
Distância Mínima

A distância entre dois clusters é dada pela distância entre os dois elementos, um em cada cluster, mais próximos.

$$d(C_1, C_2) = \min\{d(X_l, X_k, l \neq k)\}$$

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}\begin{bmatrix} A & B & C & D & E & F \\ 0 & & & & & \\ 3{,}23 & 0 & & & & \\ 15{,}74 & 12{,}53 & 0 & & & \\ 13{,}19 & 12{,}04 & 16{,}29 & 0 & & \\ 6{,}44 & 7{,}50 & 17{,}06 & 19{,}33 & 0 & \\ 13{,}39 & 10{,}19 & 4{,}12 & 12{,}18 & 16{,}19 & 0 \end{bmatrix}$$

Menor distância = d_{AB}= 3,23

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$
\begin{array}{c|cccccc}
 & A & B & C & D & E & F \\
A & 0 & & & & & \\
B & 3{,}23 & 0 & & & & \\
C & 15{,}74 & 12{,}53 & 0 & & & \\
D & 13{,}19 & 12{,}04 & 16{,}29 & 0 & & \\
E & 6{,}44 & 7{,}50 & 17{,}06 & 19{,}33 & 0 & \\
F & 13{,}39 & 10{,}19 & 4{,}12 & 12{,}18 & 16{,}19 & 0
\end{array}
$$

$$d(AB,C) = \min\{d(A,C), d(B,C)\} = \min\{15{,}74; 12{,}53\} = 12{,}53$$

$$d(AB,D) = \min\{d(A,D), d(B,D)\} = \min\{13{,}19; 12{,}04\} = 12{,}04$$

$$d(AB,E) = \min\{d(A,E), d(B,E)\} = \min\{6{,}44; 7{,}50\} = 6{,}44$$

$$d(AB,F) = \min\{d(A,F), d(B,F)\} = \min\{13{,}39; 10{,}19\} = 10{,}19$$

Método Hierárquico Aglomerativo Distância Mínima

Exemplo:

$$d(AB,C)=\min\{d(A,C),d(B,C)\}=\min\{15,74;12,53\}=12,53$$

$$d(AB,D)=\min\{d(A,D),d(B,D)\}=\min\{13,19;12,04\}=12,04$$

$$d(AB,E)=\min\{d(A,E),d(B,E)\}=\min\{6,44;7,50\}=6,44$$

$$d(AB,F)=\min\{d(A,F),d(B,F)\}=\min\{13,39;10,19\}=10,19$$

$$\begin{bmatrix} & \{AB\} & C & D & E & F \\ \{AB\} & 0 & & & & \\ C & 12,53 & 0 & & & \\ D & 12,04 & 16,29 & 0 & & \\ E & 6,44 & 17,06 & 19,33 & 0 & \\ F & 10,19 & 4,12 & 12,18 & 16,19 & 0 \end{bmatrix}$$

Menor distância = d_{CF}=4,12

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$
\begin{array}{c}
 \\ \{AB\} \\ C \\ D \\ E \\ F
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \{AB\} & C & D & E & F \end{array} \\
\begin{bmatrix}
0 & & & & \\
12{,}53 & 0 & & & \\
12{,}04 & 16{,}29 & 0 & & \\
6{,}44 & 17{,}06 & 19{,}33 & 0 & \\
10{,}19 & 4{,}12 & 12{,}18 & 16{,}19 & 0
\end{bmatrix}
\end{array}
$$

$$d(CF, AB) = \min\{d(C, AB), d(F, AB)\} = \min\{12{,}53; 10{,}19\} = 10{,}19$$

$$d(CF, D) = \min\{d(C, D), d(F, D)\} = \min\{16{,}29; 12{,}18\} = 12{,}18$$

$$d(CF, E) = \min\{d(C, E), d(F, E)\} = \min\{17{,}06; 16{,}19\} = 16{,}19$$

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$d(CF,AB)=\min\{d(C,AB),d(F,AB)\}=\min\{12{,}53;10{,}19\}=10{,}19$$

$$d(CF,D)=\min\{d(C,D),d(F,D)\}=\min\{16{,}29;12{,}18\}=12{,}18$$

$$d(CF,E)=\min\{d(C,E),d(F,E)\}=\min\{17{,}06;16{,}19\}=16{,}19$$

$$\begin{bmatrix} & \{AB\} & \{CF\} & D & E \\ \{AB\} & 0 & & & \\ \{CF\} & 10{,}19 & 0 & & \\ D & 12{,}04 & 12{,}18 & 0 & \\ E & 6{,}44 & 16{,}19 & 19{,}33 & 0 \end{bmatrix}$$

Menor distância = d_{ABE}= 6,44

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$
\begin{array}{lcccc}
 & \{AB\} & \{CF\} & D & E \\
\{AB\} & 0 & & & \\
\{CF\} & 10{,}19 & 0 & & \\
D & 12{,}04 & 12{,}18 & 0 & \\
E & 6{,}44 & 16{,}19 & 19{,}33 & 0
\end{array}
$$

$$d(ABE, CF) = \min\{d(AB, CF), d(E, CF)\} = \min\{10{,}19; 16{,}19\} = 10{,}19$$

$$d(ABE, D) = \min\{d(AB, D), d(E, D)\} = \min\{12{,}04; 12{,}18\} = 12{,}04$$

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$d(ABE,CF) = \min\{d(AB,CF), d(E,CF)\} = \min\{10{,}19;\ 16{,}19\} = 10{,}19$$

$$d(ABE,D) = \min\{d(AB,D), d(E,D)\} = \min\{12{,}04\ ;\ 12{,}18\} = 12{,}04$$

$$\begin{bmatrix} & \{ABE\} & \{CF\} & D \\ \{ABE\} & 0 & & \\ \{CF\} & 10{,}19 & 0 & \\ D & 12{,}04 & 12{,}18 & 0 \end{bmatrix}$$

Menor distância = d_{ABECF}= 10,19

Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

$$\begin{array}{cc} & \begin{matrix}\{ABE\} & \{CF\} & D\end{matrix} \\ \begin{matrix}\{ABE\} \\ \{CF\} \\ D\end{matrix} & \begin{bmatrix}0 & & \\ 10{,}19 & 0 & \\ 12{,}04 & 12{,}18 & 0\end{bmatrix}\end{array}$$

$$d(ABECF, D) = \min\{d(ABE, D), d(CF, D)\} = \min\{12{,}04; 12{,}18\} = 12{,}04$$

$$\begin{array}{cc} & \begin{matrix}\{ABECF\} & D\end{matrix} \\ \begin{matrix}\{ABECF\} \\ D\end{matrix} & \begin{bmatrix}0 & \\ 12{,}04 & 0\end{bmatrix}\end{array}$$

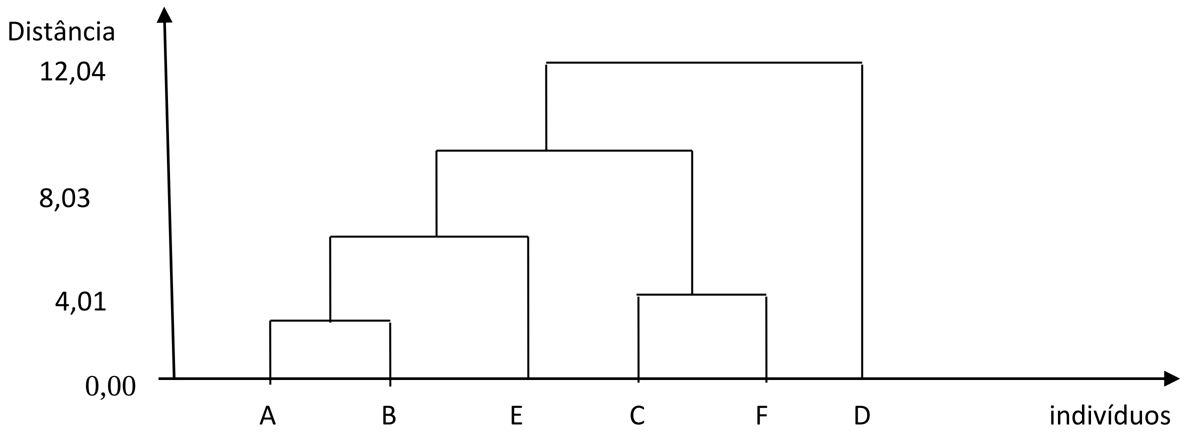
Método Hierárquico Aglomerativo
Distância Mínima

Exemplo:

Tabela. Histórico do agrupamento

Passo	N de grupos	Fusão	Distância
1	5	A e B	3,23
2	4	C e F	4,12
3	3	AB e E	6,44
4	2	ABE e CF	10,19
5	1	ABECF e D	12,04

Dendrograma



Método Hierárquico Aglomerativo
Distância Máxima

A distância entre dois clusters é dada pela distância entre os dois elementos, um em cada cluster, mais distantes.

$$d\left(C_1, C_2\right) = \max\left\{d\left(X_l, X_k, l \neq k\right)\right\}$$

Método Hierárquico Aglomerativo
Distância Máxima

Exemplo:

Menor distância = d_{AB}=3,23

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array}\left[\begin{array}{cccccc} A & B & C & D & E & F \\ 0 & & & & & \\ 3,23 & 0 & & & & \\ 15,74 & 12,53 & 0 & & & \\ 13,19 & 12,04 & 16,29 & 0 & & \\ 6,44 & 7,50 & 17,06 & 19,33 & 0 & \\ 13,39 & 10,19 & 4,12 & 12,18 & 16,19 & 0 \end{array}\right]$$

$$d(AB,C) = \max\{d(A,C), d(B,C)\} = \max\{15,74;12,53\} = 15,74$$

$$d(AB,D) = \max\{d(A,D), d(B,D)\} = \max\{13,19;12,04\} = 13,19$$

$$d(AB,E) = \max\{d(A,E), d(B,E)\} = \max\{6,44;7,50\} = 7,50$$

$$d(AB,F) = \max\{d(A,F), d(B,F)\} = \max\{13,39;10,19\} = 13,39$$

Método Hierárquico Aglomerativo
Distância Máxima

Exemplo:

$$d(AB,C) = \max\{d(A,C), d(B,C)\} = \max\{15{,}74; 12{,}53\} = 15{,}74$$

$$d(AB,D) = \max\{d(A,D), d(B,D)\} = \max\{13{,}19; 12{,}04\} = 13{,}19$$

$$d(AB,E) = \max\{d(A,E), d(B,E)\} = \max\{6{,}44; 7{,}50\} = 7{,}50$$

$$d(AB,F) = \max\{d(A,F), d(B,F)\} = \max\{13{,}39; 10{,}19\} = 13{,}39$$

$$\begin{bmatrix} & \{AB\} & C & D & E & F \\ \{AB\} & 0 & & & & \\ C & 15{,}74 & 0 & & & \\ D & 13{,}19 & 16{,}29 & 0 & & \\ E & 7{,}50 & 17{,}06 & 19{,}33 & 0 & \\ F & 13{,}39 & 4{,}12 & 12{,}18 & 16{,}19 & 0 \end{bmatrix}$$

Menor distância = d_{CF}=4,12

Método Hierárquico Aglomerativo Distância Média

A distância entre dois clusters é dada pela média das distância entre os pares de elementos.

$$d(C_1, C_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2} \right) d(X_l, X_k)$$

Método Hierárquico Aglomerativo
Distância Média

Exemplo:

Menor distância = d_{AB}=3,23

$$
\begin{array}{c}
\\ A \\ B \\ C \\ D \\ E \\ F
\end{array}
\begin{bmatrix}
A & B & C & D & E & F \\
0 & & & & & \\
3{,}23 & 0 & & & & \\
15{,}74 & 12{,}53 & 0 & & & \\
13{,}19 & 12{,}04 & 16{,}29 & 0 & & \\
6{,}44 & 7{,}50 & 17{,}06 & 19{,}33 & 0 & \\
13{,}39 & 10{,}19 & 4{,}12 & 12{,}18 & 16{,}19 & 0
\end{bmatrix}
$$

$$d(AB,C)=\{d(A,C)+d(B,C)\}/2=\{15{,}74+12{,}53\}/2=14{,}13$$

$$d(AB,D)=\{d(A,D)+d(B,D)\}/2=\{13{,}19+12{,}04\}/2=12{,}62$$

$$d(AB,E)=\{d(A,E)+d(B,E)\}/2=\{6{,}44+7{,}50\}/2=6{,}97$$

$$d(AB,F)=\{d(A,F)+d(B,F)\}/2=\{13{,}39+10{,}19\}/2=11{,}79$$

Método Hierárquico Aglomerativo
Distância Média

Exemplo:

$$d(AB,C)=\{d(A,C)+d(B,C)\}/2=\{15,74+12,53\}/2=14,13$$

$$d(AB,D)=\{d(A,D)+d(B,D)\}/2=\{13,19+12,04\}/2=12,62$$

$$d(AB,E)=\{d(A,E)+d(B,E)\}/2=\{6,44+7,50\}/2=6,97$$

$$d(AB,F)=\{d(A,F)+d(B,F)\}/2=\{13,39+10,19\}/2=11,79$$

$$\begin{bmatrix} & \{AB\} & C & D & E & F \\ \{AB\} & 0 & & & & \\ C & 14,13 & 0 & & & \\ D & 12,62 & 16,29 & 0 & & \\ E & 6,97 & 17,06 & 19,33 & 0 & \\ F & 11,79 & 4,12 & 12,18 & 16,19 & 0 \end{bmatrix}$$

Menor distância = d_{CF}=4,12

Método Hierárquico Aglomerativo
Método de Ward

Variância mínima – considerado o melhor método hierárquico

- Agrupa os clusters que resultam na menor "perda de informação", dada pelo incremento da soma dos quadrados dos erros (SQE).
- No início quando o número de objetos (N) é igual ao número de clusters (K), SQE=0.
- A medida que os clusters vão sendo formados (K=N-1; K=N-2; ...), SQE cresce (K↓ SQE↑)

Método Hierárquico Aglomerativo
Método de Ward

Para uma partição em K cluster tem-se que:

$SQE = SQE_1 + SQE_2 + ... + SQE_K$ Onde:

$$SQE_i = \sum_{j=1}^{N_i} \left(\underset{\sim}{X}_j - \underset{\sim}{\overline{X}}_i \right)^T \left(\underset{\sim}{X}_j - \underset{\sim}{\overline{X}}_i \right)$$

N_i = total de objetos pertencentes ao cluster i ($N_1 + N_2 + ... + N_k = N$)

$\overline{X}_i$ = é o centro de gravidade (Médio) do cluster i

X_j (j=1,N) = objetos (vetores) classificados no cluster i.

No final quando todos os objetos são alocados em um único cluster:

$$SQE = \sum_{j=1}^{N} \left(\underset{\sim}{X}_j - \underset{\sim}{\overline{X}} \right)^T \left(\underset{\sim}{X}_j - \underset{\sim}{\overline{X}} \right)$$

Onde $\overline{X}$ é o centro de gravidade de todo o conjunto de dados

Método Hierárquico Aglomerativo
Método de Ward

Idéia principal:

- Escolher grupamentos que tem pouca variância, pois os clusters têm que ser homogêneos.
- A medida que aumenta o número de clusters, a SQE também aumenta porque estamos juntado coisas heterogêneas.

Método Hierárquico Aglomerativo
Método de Ward

Como encontrar o número de Clusters?

$$\underbrace{\sum_{j=1}^{N}\left(X_j-\overline{X}\right)^T\left(X_j-\overline{X}\right)}_{\text{Total}}=\underbrace{\sum_{i=1}^{K}\sum_{j=1}^{N_i}\left(X_j-\overline{X}_i\right)^T\left(X_j-\overline{X}_i\right)}_{\text{Intra}}+\underbrace{\sum_{i=1}^{K}N_i\left(\overline{X}_i-\overline{X}\right)^T\left(\overline{X}_i-\overline{X}\right)}_{\text{Inter}}$$

$$\textit{Índice de Comparcidade de Separação}\,(ICS)=\frac{Intra}{N\times mín\left(d_i,d_j\right)}$$

Mín (di, dj) – menor distância entre os centros de gravidade dos k clusters.

A partição ideal é a que minimizar ICS

Métodos Não-hierárquicos

Métodos Não-hierárquicos

Os métodos não-hierárquicos diferem dos hierárquicos em vários aspectos:

- Requerem que o usuário tenha especificado previamente o número de conglomerados (k) desejado;

- Em cada estágio do agrupamento, os novos grupos podem ser formados através da divisão ou junção de grupos já combinados em passos anteriores. Por isso não é possível construir dendrogramas.

- Podemos lidar com conjunto de dados maiores, pois não é necessário armazenar a matriz de distâncias.

Métodos Não-hierárquicos k-Means

Passos para a construção dos clusters:

- Escolha um conjunto de K centróides iniciais ou particione os elementos em K conglomerados iniciais.
- Percorra a lista de elementos e aloque cada elemento ao centróide mais próximo. Usar uma medida de distância (distância euclidiana).
- A partir dos clusters formados calcule novos centróides para cada cluster. Recalcule o centróide do conglomerado que recebeu um novo elemento e do conglomerado que perdeu o elemento.
- Repita o passo 2 até não mais haver alocações novas.

Método Não-hierárquico
k-Means

Exemplo:

Indivíduo	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-1

Partição arbitrária AB e CD

Indivíduo	Média(X1)	Média(X2)
AB	(5+(-1))/2	(3+1)/2
CD	(1+(-3))/2	(-2+(-2))/2

Indivíduo	Média(X1)	Média(X2)
AB	2	2
CD	-1	-2

Método Não-hierárquico k-Means

Exemplo:

Indivíduo	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-1

Indivíduo	Média(X1)	Média(X2)
AB	2	2
CD	-1	-2

Passo2: Distância de cada item ao centróide e realoca cada item ao grupo mais próximo.

d^2 (A,(AB))=$(5-2)^2+(3-2)^2=10$

d^2 (A,(CD))=$(5+1)^2+(3+2)^2=61$

Como A é mais próximo de AB, não é realocado

d^2 (B,(AB))=$(-1-2)^2+(1-2)^2=10$

d^2 (B,(CD))=$(-1+1)^2+(1+2)^2=9$

logo B passa para CD dando origem a (BCD) e atualizamos os centróides

Análise de *Clusters* em dois estágios

Análise de *Clusters* em dois estágios

- Utilizada para bancos de dados muito extensos.
- O primeiro passo é a formação de pré-*clusters*, com o objetivo de reduzir o tamanho da matriz das distâncias entre todos os possíveis pares de casos.
- Todos os casos de um pré-*cluster* são tratados como uma única entidade.
- O segundo passo é a realização de uma análise hierárquica de *cluster*

Seleção das variáveis para formar os *clusters*

- 1) Calcular a Matriz de Correlação
- 2) Calcular uma matriz de distâncias segundo a fórmula:

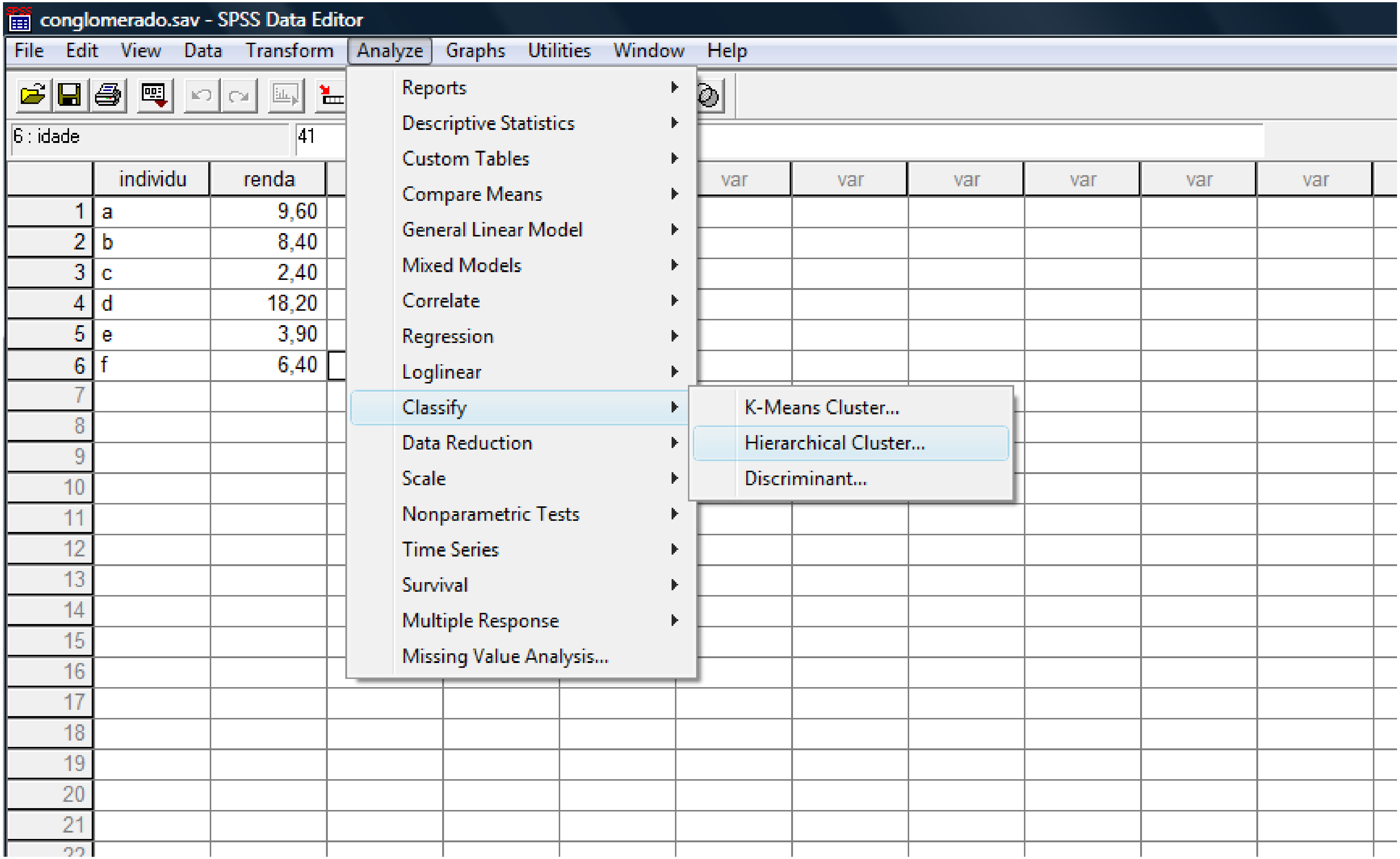
$$D_{pxp} = 1_{pxp} - ABS\left(R_{pxp}\right)$$

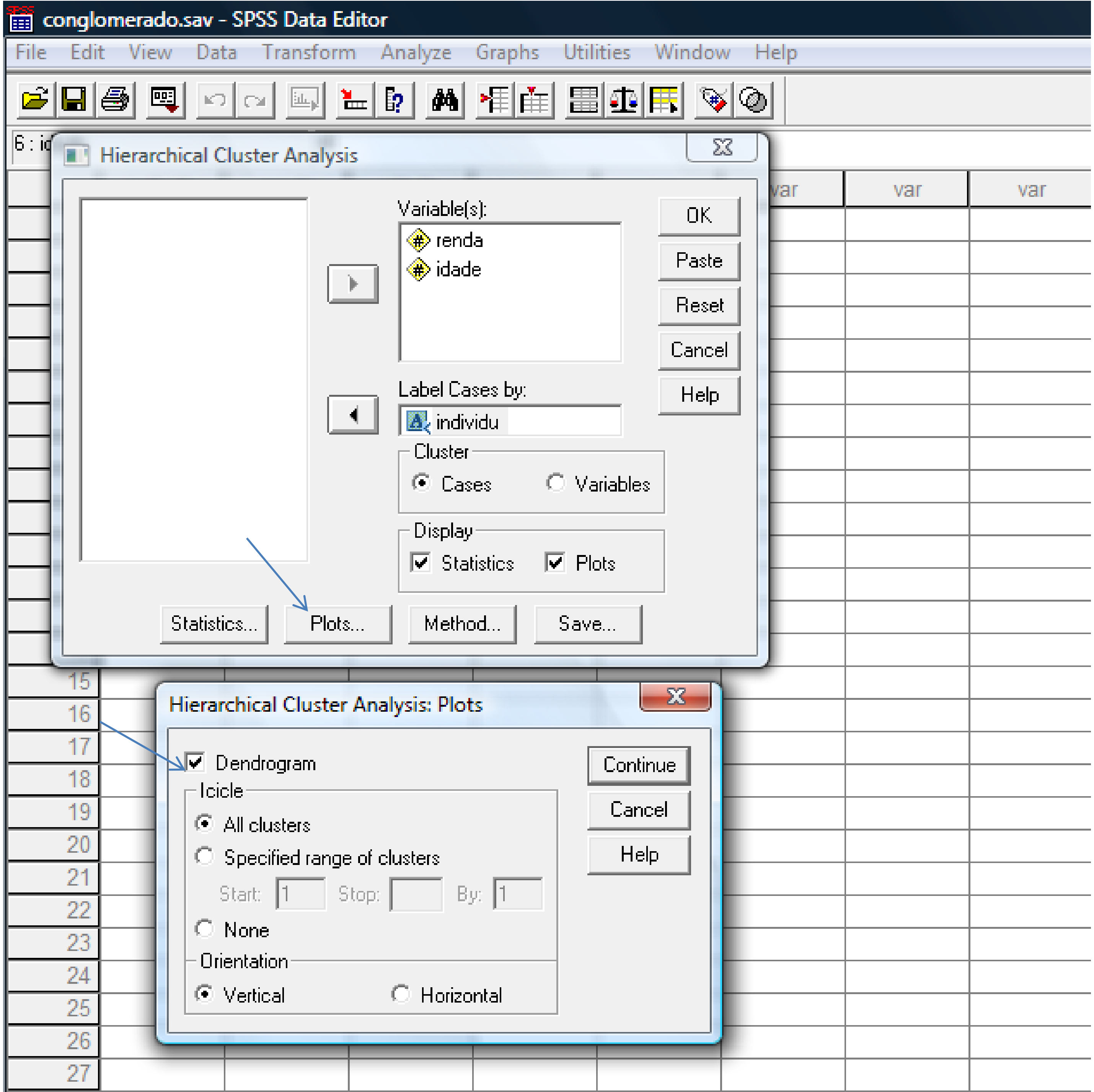
Onde 1_{pxp} é uma matriz com p-linhas e p-colunas, todas iguais ao número 1; $ABS(R_{pxp})$ é a matriz contendo os valores absolutos da matriz R_{pxp}.

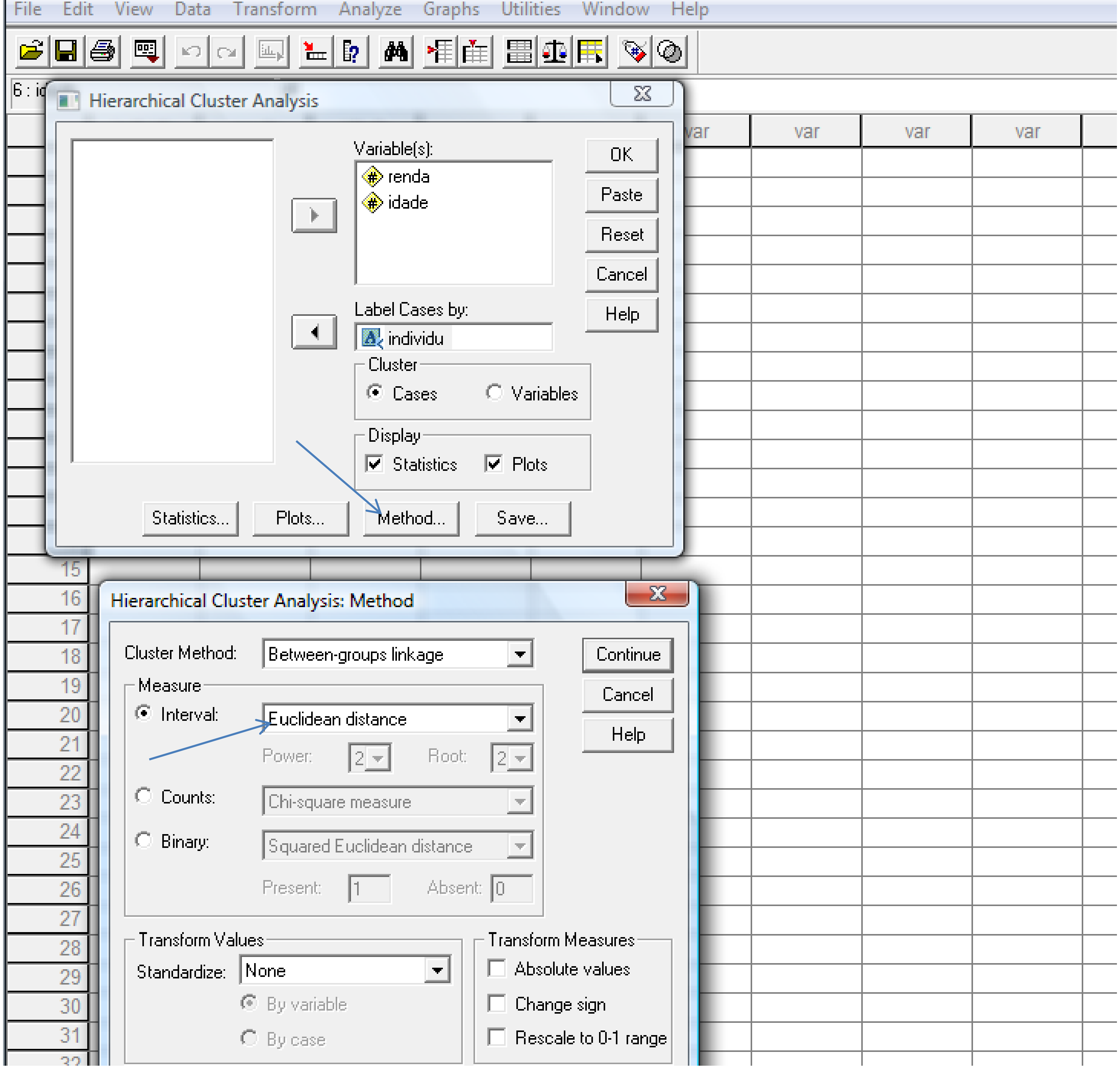
- 3) A partir de 2 aplicar um método de ligação, para formar clusters.

- 4) Selecionar apenas uma variável de cada cluster.

SPSS


conglomerado.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
6 : idade
41
individu
renda
var
a 9,60
b 8,40
c 2,40
d 18,20
e 3,90
f 6,40
Reports
Descriptive Statistics
Custom Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
K-Means Cluster...
Hierarchical Cluster...
Discriminant...


conglomerado.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
renda
idade
OK
Paste
Reset
Cancel
Help
Label Cases by:
individu
Cluster
Cases
Variables
Display
Statistics
Plots
Statistics...
Plots...
Method...
Save...
var
var
var
Hierarchical Cluster Analysis: Plots
Dendrogram
Icicle
All clusters
Specified range of clusters
Start: 1
Stop:
By: 1
None
Orientation
Vertical
Horizontal
Continue
Cancel
Help


File Edit View Data Transform Analyze Graphs Utilities Window Help
Hierarchical Cluster Analysis
Variable(s):
renda
idade
Label Cases by:
individu
Cluster
Cases
Variables
Display
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Method...
Save...
var
var
var
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Hierarchical Cluster Analysis: Method
Cluster Method:
Between-groups linkage
Continue
Cancel
Help
Measure
Interval:
Euclidean distance
Power:
2
Root:
2
Counts:
Chi-square measure
Binary:
Squared Euclidean distance
Present:
1
Absent:
0
Transform Values
Standardize:
None
By variable
By case
Transform Measures
Absolute values
Change sign
Rescale to 0-1 range

Output1 - SPSS Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

- Output
  - Frequencies
    - Title
    - Notes
    - Statistics
    - SEXO
  - Frequencies
    - Title
    - Notes
    - Statistics
    - Frequency Table
      - Title
      - IDADE_CA
      - IDADE
  - Frequencies
    - Title
    - Notes
    - Statistics
    - Frequency Table
      - Title
      - IDADE_CA
      - SEXO
  - T-Test
    - Title
    - Notes
    - Group Statistics
    - Independent Samples Test
  - T-Test
    - Title
    - Notes
    - Group Statistics
    - Independent Samples Test
  - Cluster
    - Title
    - Notes
    - Case Processing Summary
    - Proximity Matrix
    - Average Linkage (Between Groups
      - Title
      - Agglomeration Schedule
      - Vertical Icicle
      - Dendrogram
        - Title
        - Dendrogram

	Euclidean Distance					
Case	1:a	2:b	3:c	4:d	5:e	6:f
1:a	,000	3,231	15,743	13,189	6,441	13,388
2:b	3,231	,000	12,530	12,043	7,500	10,198
3:c	15,743	12,530	,000	16,298	17,066	4,123
4:d	13,189	12,043	16,298	,000	19,326	12,175
5:e	6,441	7,500	17,066	19,326	,000	16,194
6:f	13,388	10,198	4,123	12,175	16,194	,000

This is a dissimilarity matrix

Average Linkage (Between Groups)

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	1	2	3,231	0	0	3
2	3	6	4,123	0	0	4
3	1	5	6,971	1	0	4
4	1	3	14,187	3	2	5
5	1	4	14,606	4	0	0

Vertical Icicle

	Case										
Number of clusters	4:d		6:f		3:c		5:e		2:b		1:a
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X	X	X		X		X	X	X
5	X		X		X		X		X	X	X

SPSS Processor is ready H: 83 , W: 363 pt.

Output1 - SPSS Viewer

File Edit View Insert Format Analyze Graphs Utilities Window Help

- Output
  - Frequencies
    - Title
    - Notes
    - Statistics
    - SEXO
  - Frequencies
    - Title
    - Notes
    - Statistics
    - Frequency Table
      - Title
      - IDADE_CA
      - IDADE
  - Frequencies
    - Title
    - Notes
    - Statistics
    - Frequency Table
      - Title
      - IDADE_CA
      - SEXO
  - T-Test
    - Title
    - Notes
    - Group Statistics
    - Independent Samples Test
  - T-Test
    - Title
    - Notes
    - Group Statistics
    - Independent Samples Test
  - Cluster
    - Title
    - Notes
    - Case Processing Summary
    - Proximity Matrix
    - Average Linkage (Between Groups
      - Title
      - Agglomeration Schedule
      - Vertical Icicle
      - Dendrogram
        - Title
        - Dendrogram

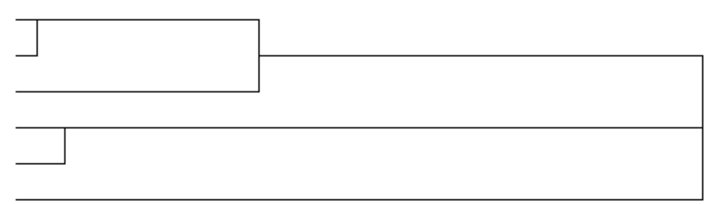
5	1	4	14,606	4	0	0

Vertical Icicle

Number of clusters	Case										
	4:d		6:f		3:c		5:e		2:b		1:a
1	X	X	X	X	X	X	X	X	X	X	X
2	X		X	X	X	X	X	X	X	X	X
3	X		X	X	X		X	X	X	X	X
4	X		X	X	X		X		X	X	X
5	X		X		X		X		X	X	X

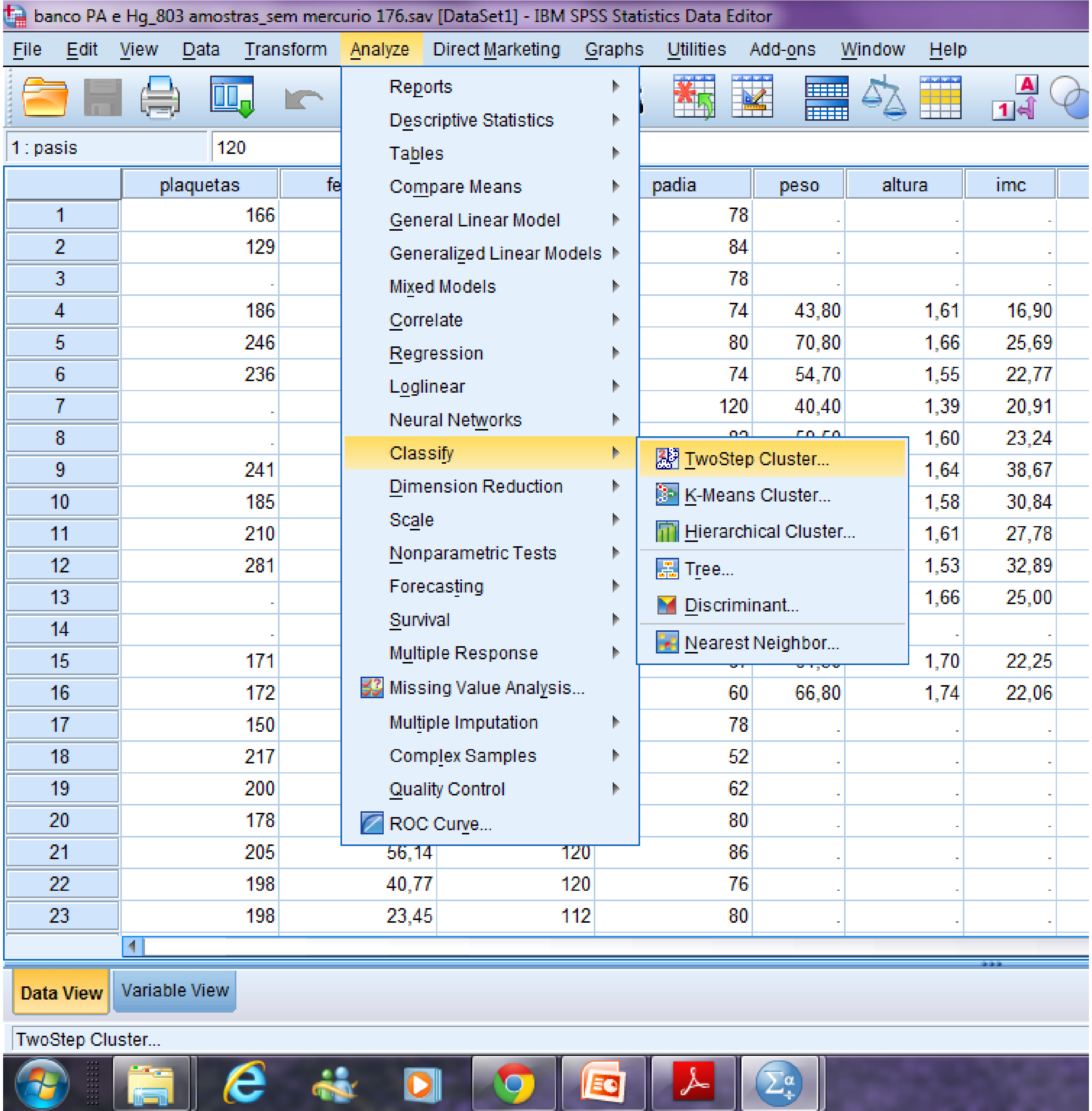
Dendrogram

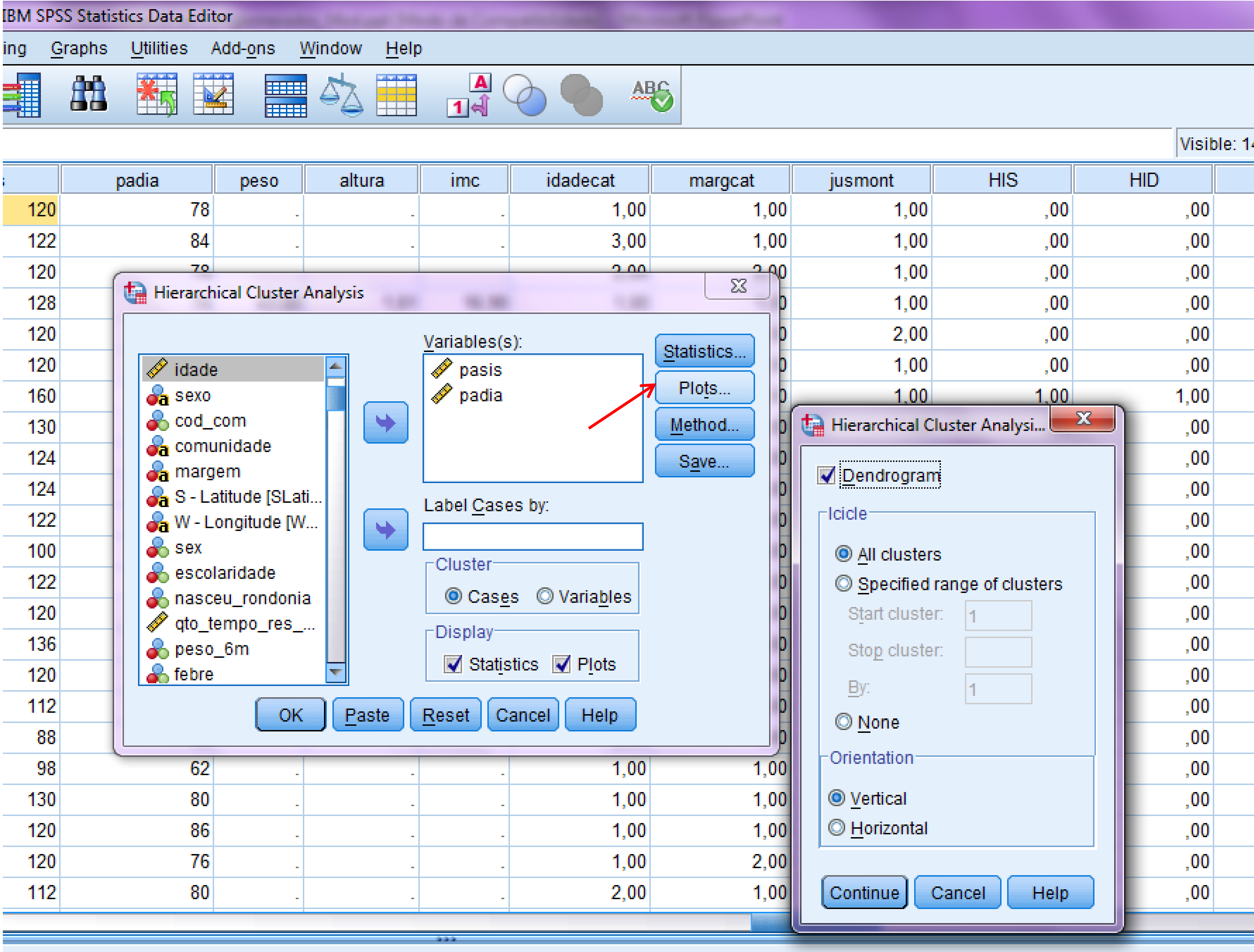
```
* * * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * * *

 Dendrogram using Average Linkage (Between Groups)

                         Rescaled Distance Cluster Combine

    C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+

  a           1
  b           2
  e           5
  c           3
  f           6
  d           4
```

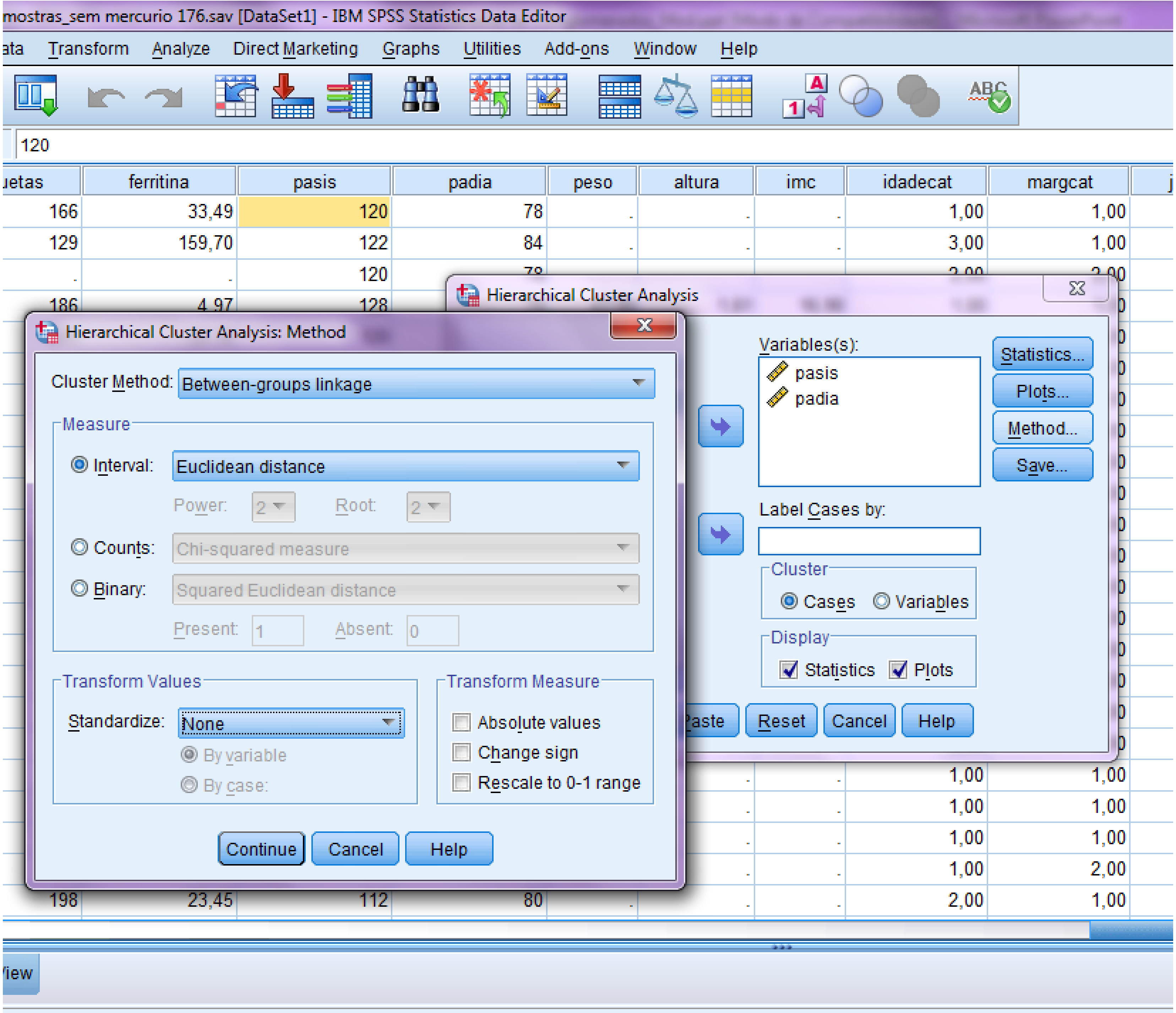


SPSS Processor is ready | H: 83 , W: 363 pt.

Outro Exemplo:
SPSS versão 19


banco PA e Hg_803 amostras_sem mercurio 176.sav [DataSet1] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Help
1 : pasis
120
plaquetas
padia
peso
altura
imc
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Quality Control
ROC Curve...
TwoStep Cluster...
K-Means Cluster...
Hierarchical Cluster...
Tree...
Discriminant...
Nearest Neighbor...
Data View
Variable View
TwoStep Cluster...


IBM SPSS Statistics Data Editor
ing Graphs Utilities Add-ons Window Help
Visible: 145
padia peso altura imc idadecat margcat jusmont HIS HID
Hierarchical Cluster Analysis
idade
sexo
cod_com
comunidade
margem
S - Latitude [SLati...
W - Longitude [W...
sex
escolaridade
nasceu_rondonia
qto_tempo_res_...
peso_6m
febre
Variables(s):
pasis
padia
Label Cases by:
Cluster
Cases
Variables
Display
Statistics
Plots
Statistics...
Plots...
Method...
Save...
OK
Paste
Reset
Cancel
Help
Hierarchical Cluster Analysi...
Dendrogram
Icicle
All clusters
Specified range of clusters
Start cluster:
Stop cluster:
By:
None
Orientation
Vertical
Horizontal
Continue
Cancel
Help


mostras_sem mercurio 176.sav [DataSet1] - IBM SPSS Statistics Data Editor
ata Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
120
uetas ferritina pasis padia peso altura imc idadecat margcat
166 33,49 120 78 1,00 1,00
129 159,70 122 84 3,00 1,00
120
186 4,97 128
198 23,45 112 80 2,00 1,00
Hierarchical Cluster Analysis
Variables(s):
pasis
padia
Statistics...
Plots...
Method...
Save...
Label Cases by:
Cluster
Cases
Variables
Display
Statistics
Plots
aste
Reset
Cancel
Help
Hierarchical Cluster Analysis: Method
Cluster Method:
Between-groups linkage
Measure
Interval:
Euclidean distance
Power:
2
Root:
2
Counts:
Chi-squared measure
Binary:
Squared Euclidean distance
Present:
1
Absent:
0
Transform Values
Standardize:
None
By variable
By case:
Transform Measure
Absolute values
Change sign
Rescale to 0-1 range
Continue
Cancel
Help
iew

Dendograma

RESULTADO

Pelo AC Hierarquica Salvei 2 grupos e 3 grupos.

Cruzei este resultado com a variável Hipertensão de acordo com os critérios já estabelecidos pela medicina. O resultado está abaixo.

Average Linkage (Between Groups) * PHAS Crosstabulation

Count

		PHAS		Total
		Não	Sim	
Average Linkage (Between Groups)	1	594	161	755
	2	0	47	47
Total		594	208	802

Average Linkage (Between Groups) * PHAS Crosstabulation

Count

		PHAS		Total
		Não	Sim	
Average Linkage (Between Groups)	1	568	160	728
	2	0	47	47
	3	26	1	27
Total		594	208	802

Average Linkage (Between Groups) * Average Linkage (Between Groups) Crosstabulation

Count

		Average Linkage (Between Groups)		Total
		1	2	
Average Linkage (Between Groups)	1	728	0	728
	2	0	47	47
	3	27	0	27
Total		755	47	802

Exemplo: K-means

As mesmas variáveis e definidos 2 grupos

Cluster Number of Case

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	441	55,0	55,0	55,0
	2	361	45,0	45,0	100,0
	Total	802	100,0	100,0	

Cluster Number of Case * PHAS Crosstabulation

Count

		PHAS		Total
		Não	Sim	
Cluster Number of Case	1	420	21	441
	2	174	187	361
Total		594	208	802

Cluster Number of Case * Average Linkage (Between Groups) Crosstabulation

Count

		Average Linkage (Between Groups)		Total
		1	2	
Cluster Number of Case	1	441	0	441
	2	314	47	361
Total		755	47	802

Cluster Number of Case * Average Linkage (Between Groups) Crosstabulation

Count

		Average Linkage (Between Groups)			Total
		1	2	3	
Cluster Number of Case	1	414	0	27	441
	2	314	47	0	361
Total		728	47	27	802

Referências Bibliográficas

- Sueli Aparecida Mingoti . Análise de Dados Através de Métodos de Estatística Multivariada. Uma Abordagem Aplicada. Belo Horizonte: Editora UFMG, 2005.

- Hair JF, Anderson RE, Tatham RL, Black WC. Multivariate Data Analysis. 5th edition, Prentice-Hall, 1998.

- Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. 6th edition, Pearson Education, 2007.