Ensino & Informação
Prof.: Altamir A. R. Araldi
Revista®

APRESENTAÇÃO DOS MÉTODOS DE CLASSIFICAÇÃO

Advertência:

- Só o conhecimento das propriedades lógicas dos métodos estatísticos permite evitar o uso "às cegas" dos mesmos.
- O uso "às cegas" de um método estatístico se realiza em quatro etapas (fases):

Primeira fase : utilzar-se uma tabela construída de qualquer maneira

Segunda fase: a essa tabela são aplicados quaisquer métodos estatísticos.

Terceira fase: Obtém-se assim um "resultado" [se o computador funciona.... a aplicação de um algoritmo de cálculo a uma tabela de números....dá sempre um resultado!!!]

Quarta fase : por último, o analista...

* fica perplexo... emite dúvidas sobre a utilidade da análise dos dados

ou,

* passa por cima de todos e de tudo com grande audácia (pouca seriedade e pouco profissionalismo) faz um comentário absurdo sobre "resultados" sem sentido.

Michel VOLLE, Analyse des Données, Ed.Economica, Paris, 1982.

Plano de Apresentação dos métodos de Classificação

Objetivo: mostrar as propriedades lógicas de alguns métodos estatísticos destinados a dividir em subconjuntos (classes) um conjunto de dados observados.

Apresentam-se os métodos estatísticos chamados métodos de agrupamento ("métodos de classificação"), "cluster analysis" ou "métodos de classificação automática", em relação com os métodos fatoriais de análise de dados.

1. *Que significa classificar um conjunto de dados observados?*

2. *Classes, classes empíricas e classificabilidade de um conjunto dados observados.*

3. *Como se desenvolve a aplicação de um método de classificação?*

4. *Como definir semelhanças entre "indivíduos" de uma tabela T(n,p)?*

5. *Seleção de uma distância entre os objetos a classificar.*

6. *As classificações hierárquicas ascendentes.*

7. *As classificações não hierárquicas: partição dos dados.*

8. *O tratamento de grandes tabelas de resultantes de levantamentos por amostragem ("survey"): a estratégia "análise fatorial + classificação"*

1. Que significa classificar um conjunto dados observados?

- Aplicar um método de classificação a um conjunto de observações, significa definir nesse conjunto as classes em que se distribuem os elementos do conjunto.

- Existem duas grandes famílias de métodos estatísticos que permitem classificar um conjunto de observações.

a) Os métodos de classificação propriamente ditos.

Fracionam um conjunto dado de unidades de observação em subconjuntos homogêneos.

b) Os procedimentos de classificação ou de partição

Distribuem ou designam os elementos de um conjunto de observações em classes pre-stabelecidas.

2. Classes, classes empíricas e classificabilidade de um conjunto de observações.

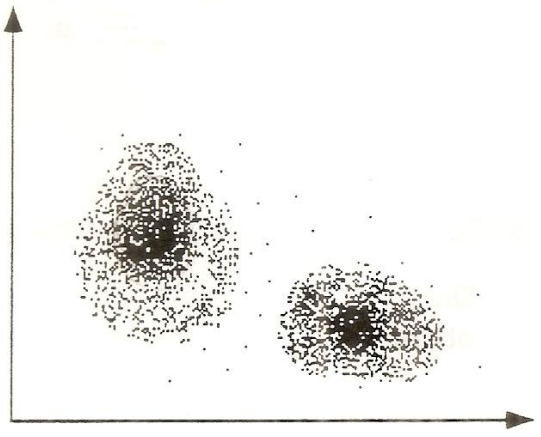
As duas famílias de métodos de classificação estão compostas de procedimentos automáticos destinados a definir "classes de indivíduos" o mais semelhantes possível.

2.1. Que são as "classes de indivíduos" semelhantes...?

Se os **n** indivíduos sobre os quais se observaram as p características estão representados num espaço de **p** dimensões....

ejemplo : p = 2

Chamam-se "classes" aos subconjuntos de indivíduos desse espaço de representação que são identificáveis porque:



* em certas zonas do espaço existe uma alta densidade de indivíduos.

* nas zonas do espaço que separa esses subconjuntos há uma baixa densidade de indivíduos.

2. Classes, classes empíricas e classificabilidade de um conjunto dado de unidades de observação.

2.2. Classificação dos elementos de uma tabela observada

- **Não se pode postular a existência de classes ("empíricas") num conjunto de observações.**
- **Só podemos verificar a existência de níveis de síntese significativos correspondentes à organização em classes e subclasses dos elementos.-**

De modo que :

➠ Os elementos de uma tabela T(n,p) qualquer não são necessariamente classificáveis....

Por isso, é necessário explorar previamente a estrutura da informação disponível, antes de orientar-se em direção a um algoritmo de classificação.

➠ A significação dos níveis de sínteses estabelecidos por um algoritmo de classificação depende da seleção de:

* uma distância adequada para avaliar a semelhança entre os elementos e grupos de elementos a comparar;

* um algoritmo adequado de classificação.

3. Como se desenvolve a aplicação de um método de classificação?

- Dispõe-se de uma tabela resumo do tipo T(n,p) [n linhas/indivíduos descritos por p caracteres]
- Os elementos de T(n,p) apresentam uma estrutura de grupo ou de hierarquia de grupos encaixados.

A aplicação se desenvolve em três etapas :

Primeira etapa : cria-se uma tabela D(n,n) que apresenta o grau de semelhança de cada indivíduo i com respeito a cada indivíduo j de T(n,p), considerando os p caracteres observados.

Segunda etapa : algoritmo de clasificação hierárquica.

1. Começa-se com uma partição do conjunto dos n indivíduos de maneira que cada um seja o único elemento de cada uma das classes de uma partição com um número de classes igual ao número de indivíduos.
2. Reune-se numa classe as duas classes mais parecidas (semelhantes) da etapa anterior. O número de classes restantes diminui de uma unidade.
3. Prossegue-se assim até não dispor mais do que uma só classe que reune todas as classes (e em consequência os indivíduos).

3. Como se desenvolve a aplicação de um método de classificação?

(continuação)

Terceira etapa : descreve-se os conteúdos dos subconjuntos de classes obtidos em cada etapa e se avalia a qualidade da classificação obtida.

Para utilizar um algoritmo de classificação hierárquica deve-se resolver dois problemas:

1. Como definir e avaliar as semelhanças entre indivíduos, isto é entre subconjuntos (classes) compostos de um indivíduo (subconjuntos de cardinalidade 1)....?
2. Como definir e avaliar a semelhança entre subconjuntos (classes) de indivíduos de cardinalidade superior a 1...?

4. ¿Como definir semelhanças entre "indivíduos" de uma tabela T(n,p)?

4.1. Índices de similaridade

A semelhança entre dois indivíduos **i** e **j** pode ser definida matematicamente por uma função S_{ij} – com valores reais – das observações correspondentes às linhas **i** e **j** de T(n,p).

Existem diferentes funções S_{ij} que variam em relação ao nível de medida das **p** variáveis de T(n,p) [quantitativas, nominais, dicotômicas, ordinais].

➠ A semelhança entre os indivíduos **i** e **j** é definida por uma função simétrica:

$s_{ij} = s_{ji} \; \forall i ; \forall j .$

Além de : $s_{ij} = s_{ij} \; \forall i \; ; \; \forall j \quad y \quad s_{ij} \leq s_{ii} = s_{jj} .$

Nesse caso, s_{ij} é um índice de similaridade.

Em geral : $0 \leq s_{ij} \leq 1.$

Mas, por exemplo, o índice de correlação: $-1 \leq s_{ij} \leq 1.$

4. Como definir semelhanças entre "indivíduos" de uma tabela T(n,p)?

4.2. Índices de similaridade

Para avaliar a "similaridade" entre os indivíduos de T(n,p) define-se "índices de dissimilaridade" que variam inversamente aos índices de similaridade.

Seja s_{ij} um índice de similaridade $0 \leq s_{ij} \leq 1$.

Então : $d_{ij} = 1 - s_{ij}$ é um índice de disimilaridade.

$$d_{ij} = d_{ji} \; \forall i; \forall j \quad y \quad d_{ii} = d_{jj} \quad y \quad 0 \leq d_{ij} \leq 1.$$

naturalmente : si $s_{ij} = 1 \Rightarrow d_{ij} = 0$.

en geral : $s_{ij} = 1 \;\; y \;\; d_{ij} = 0$ se e somente se as linhas **i** e **j** de T(n,p) **são idênticas.**

por outro lado : si $s_{ij} = 1 \;\; y \;\; d_{ij} = 0 \Rightarrow d_{ik} = d_{jk} \; \forall k$.

em particular : $s_{ii} = 1 \Rightarrow d_{ii} = 0$.

4. Como definir semelhanças entre "indivíduos" de uma tabela T(n,p)?

4.3. Distâncias

Chamamos "distância" a todo índice de similaridade que satisfaça as seguintes propriedades:

1. $d_{ii} = 0$ se e somente se **i** coincide com **j**.

$$\Rightarrow \quad d_{ii} = 0 \; y \; d_{jj} = 0$$

a tabela **D**(n,n) tem diagonal nula.

2. $d_{ij} = d_{ji} \;\; \forall i \; ; \;\; \forall j$

a tabela **D**(n,n) é simétrica.

3. $d_{ij} \leq d_{ik} + d_{kj} \;\; \forall i \, , \forall j \;\; y \;\; \forall k$

esta propriedade é chamada "desigualdade triangular".

* Se d_{ij} **satisfaz além as propriedades 1-3 d_{ij} é uma "distância".**

* Se d_{ij} é uma distância, então a semelhança entre os indivíduos i e j (para todo i e para todo j) pode ser ***representada*** num espaço euclidiano.

4. 4. Como definir semelhanças entre "indivíduos" de uma tabela T(n,p)?

4.4. Distância ultramétrica

Se d_{ij} é uma distância e também satisfaz à desigualdade ultramétrica,

$$d_{ij} \leq max\left\{d_{ik}; d_{kj}\right\} \forall i; \forall j \; y \; \forall k$$

então d_{ij} é uma distância ultramétrica

- A relação entre três indivíduos é então um triângulo isósceles com dois lados iguais ou mais compridos que o terceiro lado.

- A desigualdade ultramétrica é mais exigente que a desigualdade triangular.

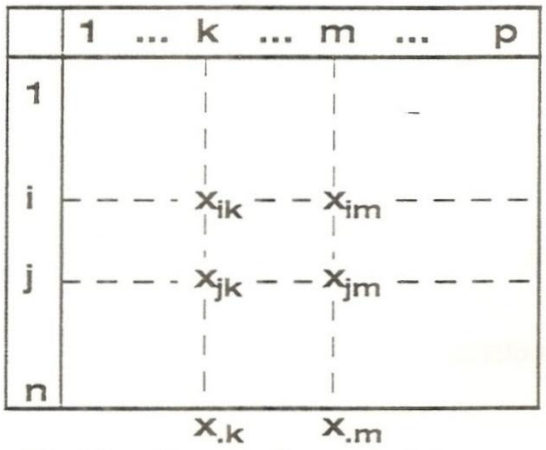
Toda ultramétrica satisfaz a desigualdade triangular.

- Toda ultramétrica é uma distância... mas toda distância não é necessariamente uma ultramétrica.

5. Seleção de uma distância entre os objetos a classificar.

A seleção de uma distância entre os objetos depende do nível de medida das características observadas, a partir das quais se quer fazer a comparação entre os objetos.

5.1. T(n,p) é uma tabela de medidas



$$x_{.k} = \sum_i x_{ik}$$

$$\bar{x}_k = \frac{1}{n}\sum_i x_{ik}$$

$$s_k^2 = \frac{1}{n}\sum_i \left(x_{ik} - \bar{x}_k\right)^2$$

Distâncias mais usuais:

1. $d_{ij}^2 = \sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2$ distância euclidiana.

2. $d_{ij}^2 = \sum_{k=1}^{p}\frac{\left(x_{ij} - x_{jk}\right)^2}{s_k^2}$ distância euclidiana reduzida

3. $d_{ij}^2 = \sum_{k=1}^{p}\frac{\left|x_{ik} - x_{jk}\right|^t}{s_k^t}$ distância de Minkowski.

4. $d_{ij} = \sum_{k=1}^{p}\frac{\left|x_{ik} - x_{jk}\right|}{s_k}$ distância de city- blocks.

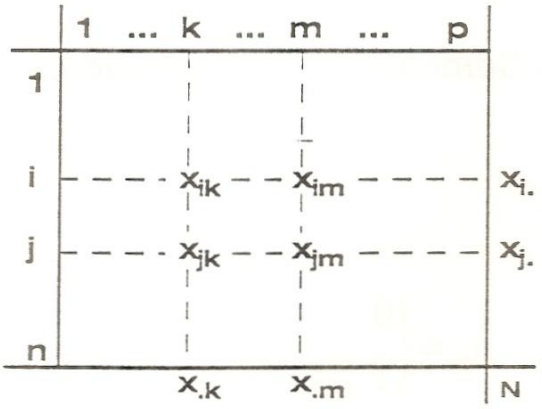
5. $d_{ij}^2 = \sum_{k=1}^{p}\frac{\left(x_{ik} - x_{jk}\right)^2}{\left(x_{ik} + x_{jk}\right)^2}$ divergência.

6. $d_{ij} = \sum_{k=1}^{p}\frac{\left|x_{ik} - x_{jk}\right|}{\left|x_{ik} + x_{jk}\right|}$ distância de Camberra.

5. Seleção de uma distância entre os objetos a classificar.

(continuacão)

5.2. T(n,p) é uma tabela de Contingência



$$d_{ij}^2 = \sum_{k=1}^{p} \frac{N}{x_{.k}} \left(\frac{x_{ik}}{x_{i.}} - \frac{x_{jk}}{x_{j.}} \right)^2$$

distância do Chi²

➪ As distâncias entre elementos em coluna de uma tabela T(n,p) se medem, com frequência, com o coeficiente de correlação.

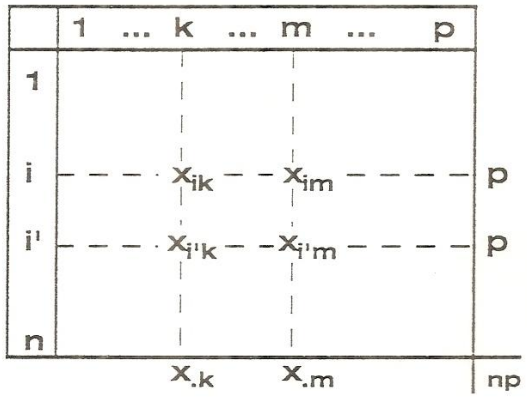
➪ No caso em que T(n,p) é uma Tabela de Contingência, a distância entre colunas da tabela se avalia também com a distância do Chi².

$$d_{km}^2 = \sum_{i=1}^{n} \frac{N}{x_{i.}} \left(\frac{x_{ik}}{x_{.k}} - \frac{x_{im}}{x_{.m}} \right)^2$$ distância do Chi² .

5. Seleção de uma distância entre os objetos a classificar.

(continuacão)

5.3. T(n,p) é uma Tabela Disjuntiva Completa ou Tabela Lógica



$$x_{ij} = \begin{cases} 0 \\ 1 \end{cases}$$

Para calcular um índice de similaridade entre as colunas **j** e **k**, constrói-se a tabela de Contingência que cruza a j-ésima coluna e a k-ésima coluna da tabela T(n,p).

K \ J	1	0	
1	a	b	$n_{.k}$
0	c	d	$n-n_{.k}$
	$n_{.j}$	$n-n_{.j}$	n

5. Seleção de uma distância entre os objetos a classificar.

(continuacão)

Índices de similaridade mais usuais :

1. $s_{jk} = \dfrac{a}{a+b+c+d}$ — Russel and Rao.

2. $s_{jk} = \dfrac{\bar{a}}{a+b+c}$ — Jaccard.

3. $s_{jk} = \dfrac{a}{a+2(b+c)}$ — Anderberg.

4. $s_{jk} = \dfrac{a(b+c)+d}{a+b+c+d}$ — Hamman.

5. $s_{jk} = \dfrac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ — Pearson.

6. $s_{jk} = \dfrac{ad-bc}{ad+bc}$ — Yule.

7. $s_{jk} = \dfrac{a+d}{a+b+c+d}$ — Sokal and Michener.

Índices de dissimilaridade correspondentes : $d_{jk} = 1 - s_{jk}$.

- Os mesmos índices podem ser calculados, *mutatis mutandis*, entre os indivíduos da tabela lógica.

- Como as tabelas lógicas podem ser consideradas como Tabelas de Contingência, a dissimilaridade entre os objetos da mesma pode ser avaliada com a distância do Chi^2.

6. As classificações hierárquicas ascendentes.

Como os métodos de análise fatorial, os métodos da classificação hierárquica são destinados a produzir uma representação gráfica da informação contida na tabela de dados.

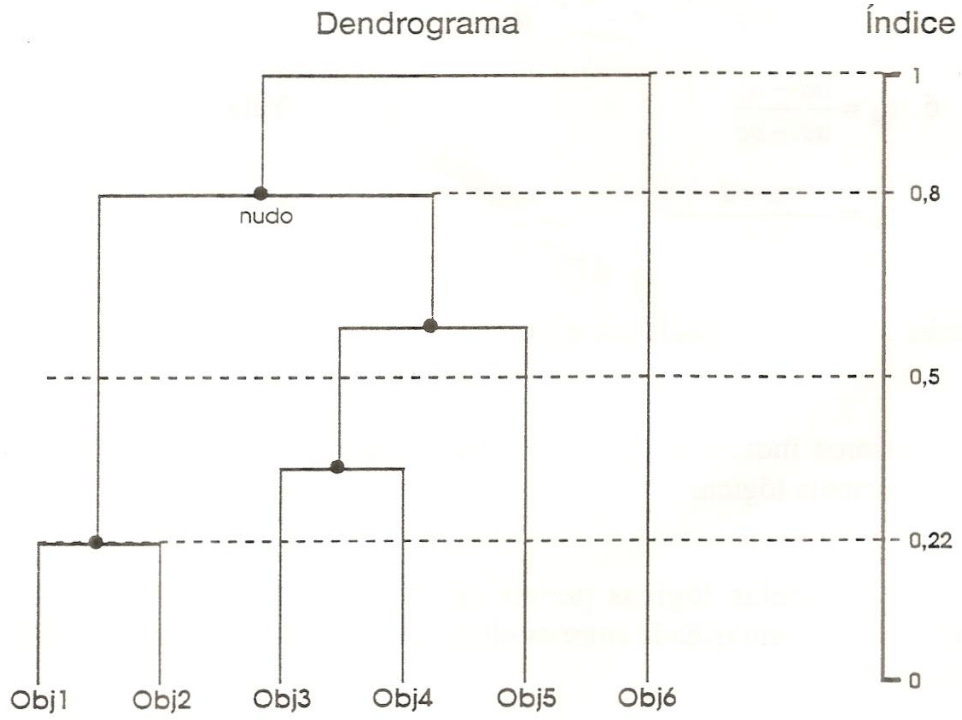
As classificações hierárquicas têm como objetivo representar, de maneira sintética, o resultado das comparações entre os objetos de uma tabela T(n,p) de observações (indivíduos, modalidades ou variáveis).

Uma classificação hierárquica é uma série de partições encaixadas.

Representação gráfica do resultado das comparações entre os indivíduos observados: "*árvore de classificação*" ou "*dendrograma*".

A essa componente analógica da mensagem corresponde uma componente digital: o índice de união.

Exemplo: resultados de uma classificação hierárquica ascendente



6. As classificações hierárquicas ascendentes..

(continuacão)

As partições encaixadas que se lêem sobre essa árvore de classificação são as seguintes:

$\{Obj1\},\{Obj2\},\{Obj3\},\{Obj4\},\{Obj5\},\{Obj6\}$

$\{Obj1,Obj2\},\{Obj3\},\{Obj4\},\{Obj5\},\{Obj6\}$

$\{Obj1,Obj2\},\{Obj3,Obj4\},\{Obj5\},\{Obj6\}$

$\{Obj1,Obj2\},\{Obj3,Obj4,Obj5\},\{Obj6\}$

$\{Obj1,Obj2,Obj3,Obj4,Obj5\},\{Obj6\}$

$\{Obj1,Obj2,Obj3,Obj4,Obj5,Obj6\}$

Diz-se que se trata de uma hierarquia indexada quando se define um índice, correspondente ao dendograma, que é sempre positivo.

Se nenhum objeto está relacionado com outro (não se parecem entre si), o índice é nulo.

O índice pode também ser normalizado. Nesse caso, quando todos os elementos se encontram ligados entre eles, a esse nó do dendograma corresponde o índice de valor 1.

A cada valor do índice corresponde uma partição na hierarquia indexada.

O índice permite avaliar a distância entre os objetos classificados. O valor da distância entre dois objetos é igual ao valor do índice correspondente ao primeiro nó que reúne ambos objetos.

6. As classificações hierárquicas ascendentes.

(continuacão)

* **Verificação: o índice utilizado no exemplo da página 16 é uma distância ultramétrica.**

=> (referenciar numéricamente)

1. A distância avaliada entre dois objetos é sempre positiva ou nula:

$$d_{(\{Obj\,x\},\{Obj\,y\})} \geq 0 \quad \forall x, y = \{1,2,3,4,5,6\}$$

2. A distância avaliada entre dois objetos idênticos é nula.

$$d_{(\{Obj\,x\},\{Obj\,x\})} = 0 \quad \forall x = \{1,2,3,4,5,6\}$$

3. A distância avaliada entre dois objetos é simétrica.

$$d_{(\{Obj\,x\},\{Obj\,y\})} = d_{(\{Obj\,y\},\{Obj\,x\})} \quad \forall x, y = \{1,2,3,4,5,6\}$$

4. A distância avaliada entre dois objetos respeita a **desigualdad e triangular**.

$$d_{(\{Obj\,x\},\{Obj\,z\})} \leq d_{(\{Obj\,x\},\{Obj\,y\})} + d_{(\{Obj\,y\},\{Obj\,z\})} \; \forall x, y, z = \{1,2,3,4,5,6\}$$

5. A distância avaliada entre dois objetos respeita a desigualdade ultramétrica.

$$d_{(\{Obj\,x\},\{Obj\,z\})} \leq Max\left[d_{(\{Obj\,x\},\{Obj\,y\})}; d_{(\{Obj\,y\},\{Obj\,z\})}\right] \forall x, y, z = \{1,2,3,4,5,6\}$$

Pode-se demonstrar que toda a árvore de classificação indexada permite definir uma distância ultramétrica e que a toda distância ultramétrica definida sobre um conjunto de objetos se pode associar uma árvore de classificação indexada.

6. 6. As classificações hierárquicas ascendentes.

(continuação)

* Que significa classificar um grupo de objetos ..?

Trata-se de construir um dendograma para um conjunto de objetos sobre os quais pode ser avaliado o grau de similaridade através de uma distância.

O problema resume-se a: Como "transformar" a distância usada numa distância ultramétrica, mudando o menos possível a distância original entre os objetos?

Se for posível transformar uma distância ultramétrica, respeitando esse critério, então é possível construir uma árvore de classificação indexada.

Essa "transformação" se faz utilizando os algoritmos de agregação de classes de objetos.

Existem diferentes algoritmos (processos iterativos) de agregação que são utilizados correntemente

- O método do vizinho mais próximo.
- O método dos centróides da distância média.
- O método baseado no crescimento mínimo do momento de **ordem dois nas classes das partições encaixadas.**

...

A seguir é mostrado como procedem esses algoritmos e suas principais propriedades...

6. As classificações hierárquicas ascendentes.

6.1. O método do vizinho mais próximo.

- Dispõe-se de uma tabela resumo do tipo T(n.p)

[n linhas/indivíduos descritos por p caracteres]

-Os elementos de T(n,p) apresentam uma estrutura de grupo ou de hierarquia de grupos encaixados.

Aplicar-se-á as etapas já vistas no processo de classificação:

Primeira etapa :

Com uma distância d_{ij} pode-se avaliar a dissimilaridade entre os objetos a classificar.

Pode-se criar uma tabela **D**(n,n), simétrica, que resume as distâncias entre os n objetos a classificar, comparados dois a dois.

Supomos que é aceitavel considerar que a distância entre duas classes que contêm um só objeto cada uma é igual à distância entre os objetos

$$d_{(\{Obj\,x\},\{Obj\,y\})} = d_{(x,y)} \quad \forall x, y \in I$$

Os termos diagonais de **D**(n,n) são nulos, já que, dij é uma distância :

$$d_{(\{Obj\,x\},\{Obj\,x\})} = d_{(x,x)} = 0 \quad \forall x \in I.$$

Segunda etapa :

Busca-se na Tabela **D**(n,n) o termo extra-diagonal de valor mínimo.

$$d_{(\{Obj\,x\},\{Obj\,y\})} = d_{(x,y)} \text{ mínimo.}$$

Forma-se uma nova classe que reagrupa esses dois objetos. $\{Obj\,x\}$ y $\{Obj\,y\}$.

Iteração :

Recomeça-se a partir da primeira etapa, mas agora só com n-1 objetos a comparar (a calcular as distâncias entre classes), já que uma classe contém atualmente dois objetos.

6. As classificações hierárquicas ascendentes.

6.1. O método do vizinho mais próximo.

Para poder calcular a Tabela **D'**(n-1, n-1) correspondente à nova situação deve-se dar um critério para calcular a distância entre uma classe que contém dois objetos e as classes restantes que só contém um objeto.

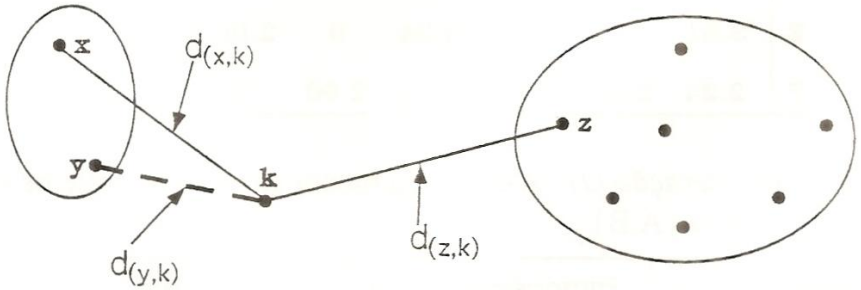
A estratégia de agregação responde a esse problema...

A estratégia do "vizinho mais próximo" consiste em escolher como distância entre a classe $\{Obj\,x, Obj\,y\}$ e a clase $\{Obj\,k\}$ a menor das duas distâncias seguintes :

$$d_{(\{Obj\,x\},\{Obj\,k\})} \text{ ou então } d_{(\{Obj\,y\},\{Obj\,k\})}$$

Em cada etapa t de iteração do processo de agregação pelo método do "vizinho mais próximo", a tabela **D**(n-t, n-t) é construída com a seguinte distância ultramétrica entre as classes :

$$d_{(\{Obj\,x.Obj\,y\},\{Obj\,k\})} = Min\left[d_{(\{Obj\,x.Obj\,k\})}; d_{.\{Obj\,y\},\{Obj\,k\}} \right]$$



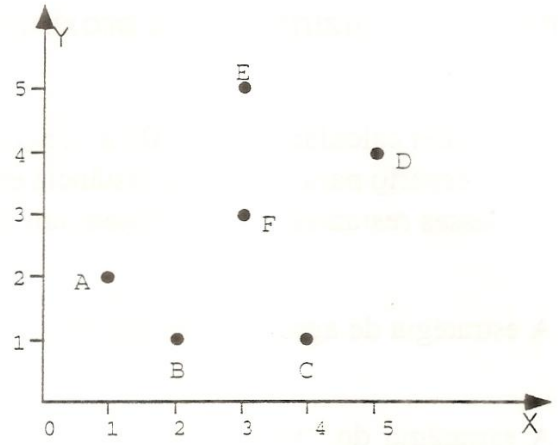
6. As classificações hierárquicas ascendentes.

6.1.O método do vizinho mais próximo: um exemplo numérico

Tabela T(n,p)

	X	Y
A	1	2
B	2	1
C	4	1
D	5	4
E	3	5
F	3	3

Representação gráfica



Primera etapa *(iteración O)* : Utilizando la distância euclidiana,

$$d_{(Obj\,A.Obj\,B)} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Pode-se calcular a matriz D(6,6) seguinte:

	A	B	C	D	E	F
A	0	1,41	3,16	4,47	3,61	2,24
B	1,41	0	2,00	4,24	4,12	2,24
C	3,16	2,00	0	3,16	4,12	2,24
D	4,47	4,24	3,16	0	2,24	2,24
E	3,61	4,12	4,12	2,24	0	2,00
F	2,24	2,24	2,24	2,24	2,00	0

Segunda etapa *(iteração O)* : A distância menor se verifica entre os objetos A e B. Forma-se a classe {A,B} :

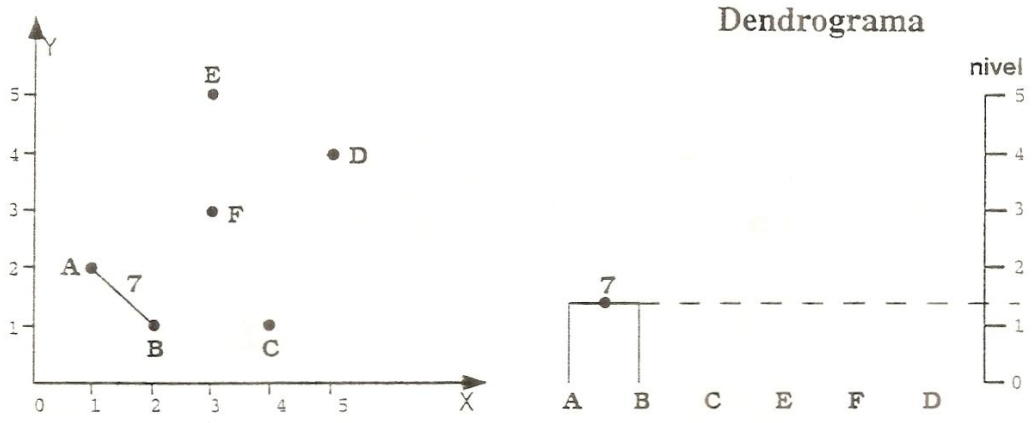
Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2

6. As classificações hierárquicas ascendentes.

6.1. O método do vizinho mais próximo: um exemplo numérico

(continuação)

Representações gráficas da primeira agregação



Primeira etapa *(iteração 1)* : Utilizando a distância ultramétrica do "vizinho mais próximo", calcula-se a tabela **D**(5,5) seguinte :

	(A,B)	C	D	E	F
(A,B)	0	2,00	4,24	3,61	2,24
C	2,00	0	3,16	4,12	2,24
D	4,24	3,16	0	2,24	2,24
E	3,61	4,12	2,24	0	2,00
F	2,24	2,24	2,24	2,00	0

Segunda etapa *(iteração 1)* : A distância menor se verifica entre os objetos E e F. Form-se a classe {E,F} :

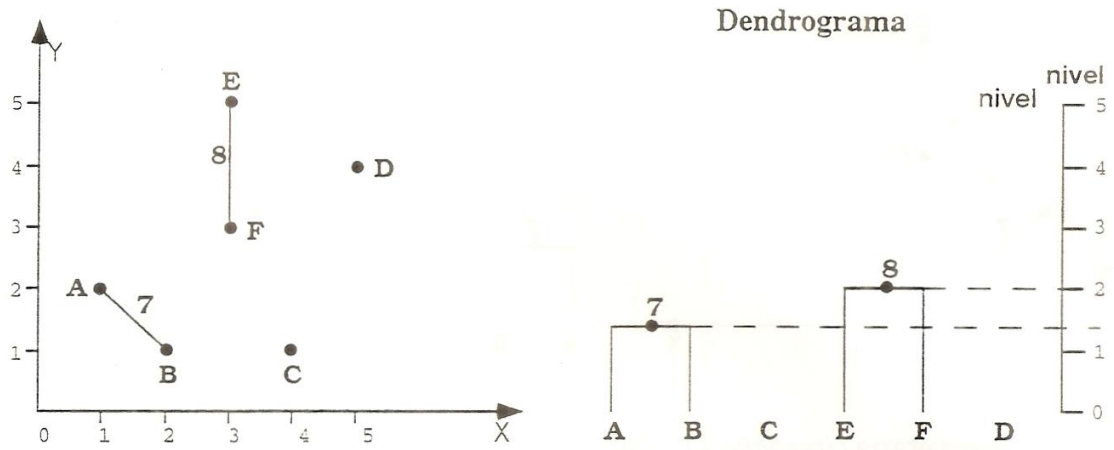
Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2
8	2,00	E	F	2

6. As classificações hierárquicas ascendentes.

6.1. O método do vizinho mais próximo: um exemplo numérico

(continuação)

Representações gráficas da segunda agregação



Primeira etapa *(iteração 2)* : Utilizando a distância ultramétrica do "vizinho mais próximo", calcula-se a tabela **D**(4,4) seguinte :

	(A,B)	C	D	(E,F)
(A,B)	0	2,00	4,24	2,24
C	2,00	0	3,16	2,24
D	4,24	3,16	0	2,24
(E,F)	2,24	2,24	2,24	0

Segunda etapa *(iteração2)* : A distância menor se verifica entre os objetos {A,B} e {A,C}. Forma-se a classe {A,B,C}:

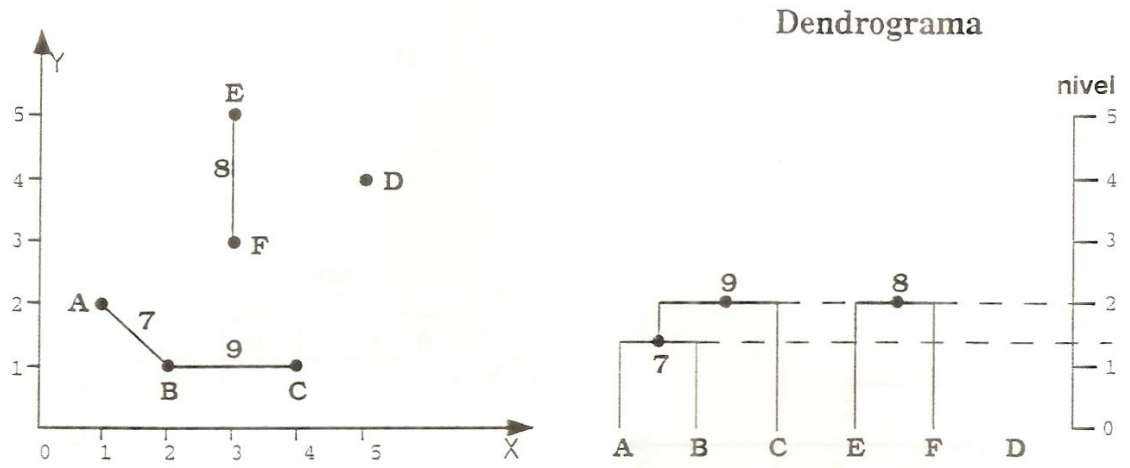
Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2
8	2,00	E	F	2
9	2,00	AB	C	3

6. As classificações hierárquicas ascendentes.

6.1. O método do vizinho mais próximo: um exemplo numérico

(continuação)

Representações gráficas da terceira agregação



Primeira etapa *(iteração 3)* : Utilizando a distância ultramétrica do "vizinho mais próximo", calcula-se a tabela D(3,3) seguinte:

	(A,B,C)	D	(E,F)
(A,B,C)	0	3,16	2,24
D	3,16	0	2,24
(E,F)	2,24	2,24	0

Segunda etapa *(iteração 3)* : A distância menor se verifica entre os objetos {A,B,C} e {E,F}. Forma-se a classe {A,B,C,E,F} :

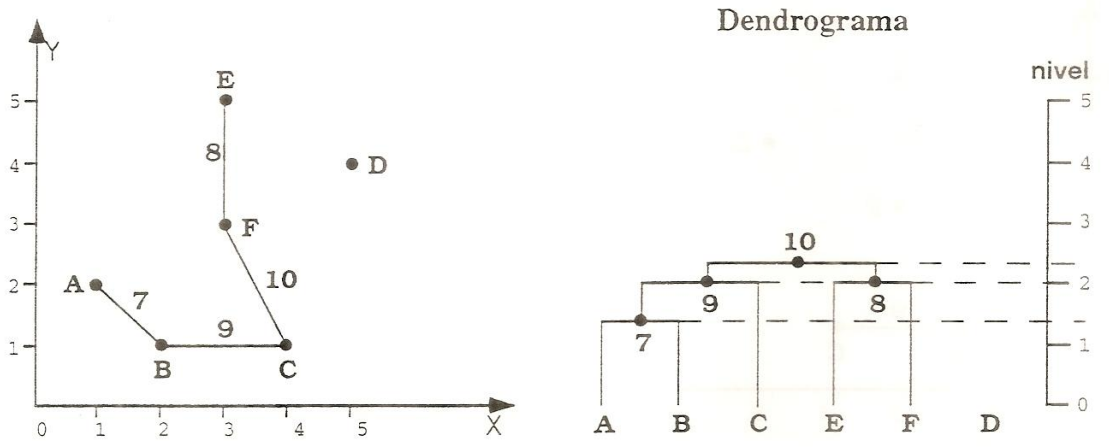
Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2
8	2,00	E	F	2
9	2,00	AB	C	3
10	2,24	ABC	EF	5

6. As classificações hierárquicas ascendentes.

6.1. O método do vizinho mais próximo: um exemplo numérico

(continuação)

Representações gráficas da quarta agregação



Primeira etapa *(iteração 4)* : Utilizando a distância ultramétrica do "vizinho mais próximo", calculamos a tabela **D**(2,2) seguinte:

	(A,B,C,E,F)	D
(A,B,C,E,F)	0	2,24
D	2,24	0

Segunda etapa *(iteração 4)* :A distância menor se verifica entre os objetos {A,B,C} e {D}. Formamos a classe {A,B,C,E,F,D} :

Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2
8	2,00	E	F	2
9	2,00	AB	C	3
10	2,24	ABC	EF	5
11	2,24	ABCEF	D	6

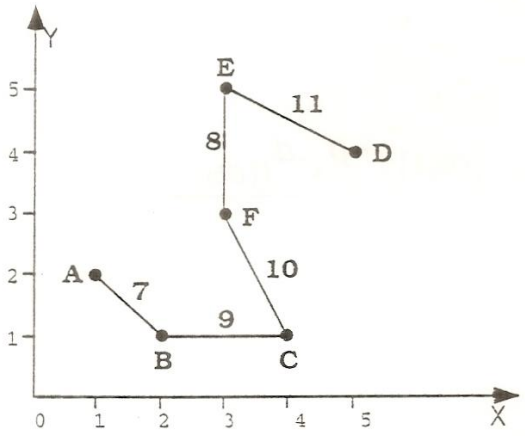
6. As classificações hierárquicas ascendentes.

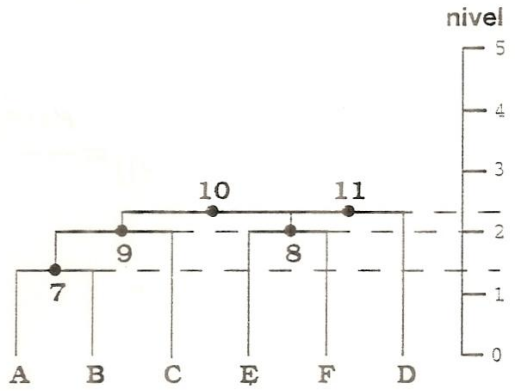
6.1.. O método do vizinho mais próximo: um exemplo numérico

(continuação)

Representações gráficas da quinta agregação

Dendrograma final





Descrição do resultado da agregação:

Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2
8	2,00	E	F	2
9	2,00	AB	C	3
10	2,24	ABC	EF	5
11	2,24	ABCEF	D	6

6. As classificações hierárquicas ascendentes.

6.2. O método dos centróides ou da distância média.

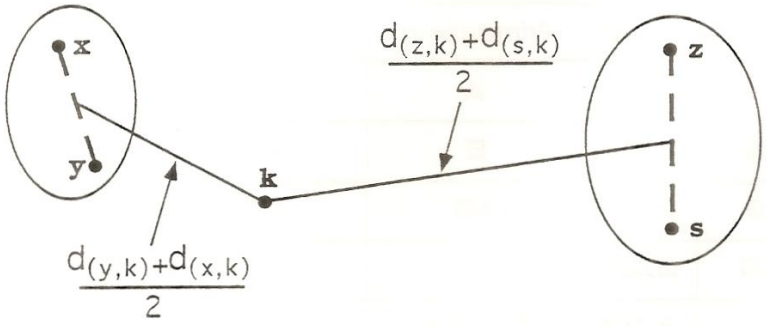
A estratégia dos **"centróides"** ou da **"distância média"** consiste em escolher como distância entre a classe $\{Obj\,x, Obj\,y\}$ e a classe $\{Obj\,k\}$ a seguinte distância:

$$d_{(\{Obj\,x,Obj\,y\},\{Obj\,k\})} = \frac{p_x.d_{(\{Obj\,x\},\{Obj\,k\})} + p_x.d_{(\{Obj\,y\},\{Obj\,k\})}}{p_x + p_y}$$

Nessa expressão :

p_x : peso ou número de objetos que contém a classe {Obj x}.

p_y :peso ou número de objetos que contém a classe {Obj y}.



6. As classificações hierárquicas ascendentes.

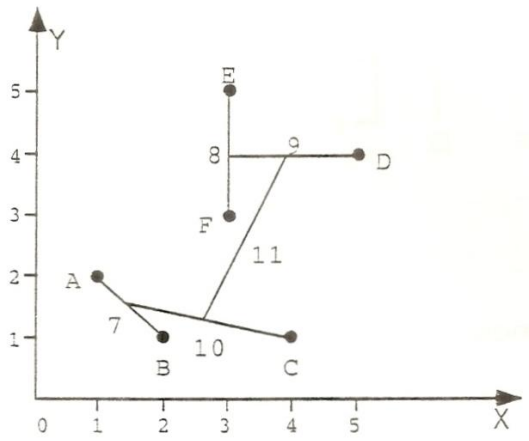
6.2. O método dos centróides ou da distância média.

Tabla de datos

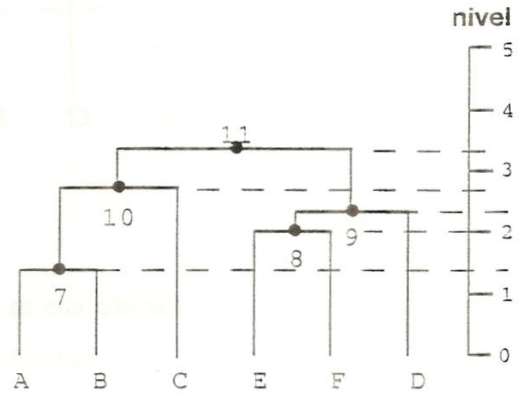
	X	Y
A	1	2
B	2	1
C	4	1
D	5	4
E	3	5
F	3	3

Matriz D(6,6)

	A	B	C	D	E	F
A	0	1,41	3,16	4,47	3,61	2,24
B	1,41	0	2,00	4,24	4,12	2,24
C	3,16	2,00	0	3,16	4,12	2,24
D	4,47	4,24	3,16	0	2,24	2,24
E	3,61	4,12	4,12	2,24	0	2,00
F	2,24	2,24	2,24	2,24	2,00	0



Dendrograma

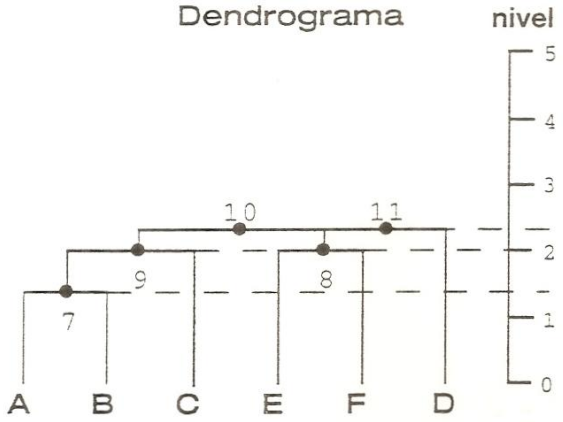


Nudo	Nivel	Primogénito	Benjamin	Peso
7	1,41	A	B	2
8	2,00	E	F	2
9	2,24	D	E,F	3
10	2,58	A,B	C	3
11	3,20	A,B,C	D,E,F	6

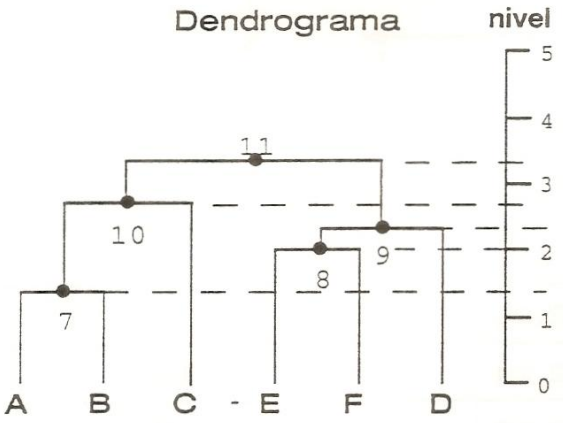
6. As classificações hierárquicas ascendentes.

6.3. Comparação dos resultados dessas três estratégias de agregação aplicadas ao mesmo exemplo numérico.

Método del vecino más próximo



Método de la distancia media



6. As classificações hierárquicas ascendentes.

6.4. Comparação da transformação da matriz de distância operada por essas três estratégias de agregação aplicadas ao mesmo exemplo numérico.

Matriz D(6,6)

	A	B	C	D	E	F
A	0	1,41	3,16	4,47	3,61	2,24
B	1,41	0	2,00	4,24	4,12	2,24
C	3,16	2,00	0	3,16	4,12	2,24
D	4,47	4,24	3,16	0	2,24	2,24
E	3,61	4,12	4,12	2,24	0	2,00
F	2,24	2,24	2,24	2,24	2,00	0

$$d_{(A,B)} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Matriz D(6,6) Ultramétrica del "vecino más próximo"

	A	B	C	D	E	F
A	0	1,41	2,00	2,24	2,24	2,24
B	1,41	0	2,00	2,24	2,24	2,24
C	2,00	2,00	0	2,24	2,24	2,24
D	2,24	2,24	2,24	0	2,24	2,24
E	2,24	2,24	2,24	2,24	0	2,00
F	2,24	2,24	2,24	2,24	2,00	0

Matriz D(6,6) Ultramétrica de la "distancia media"

	A	B	C	D	E	F
A	0	1,41	2,58	3,20	3,20	3,20
B	1,41	0	2,58	3,20	3,20	3,20
C	2,58	2,58	0	3,20	3,20	3,20
D	3,20	3,20	3,20	0	2,24	2,24
E	3,20	3,20	3,20	2,24	0	2,00
F	3,20	3,20	3,20	2,24	2,00	0

6. As classificações hierárquicas ascendentes.

6.5. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

* Princípio de funcionamento do método de agregação ou de união :

Em lugar de reunir as duas classes que apresentam a menor "distância" (segundo um critério dado para medir a semelhança entre classes), este método agrega duass classes de tal maneira que a classe resultante tenha dispersão mínima com respeito a todas as classes que possam ser formadas em uma etapa do algoritmo de agregação.

Em lugar de calcular a distância entre classes, o algoritmo calcula a dispersão de cada nova classe eventualmente constituída de duas classes originais.

Para aplicar este método, é necessário que a distância entre os objetos a classificar seja uma distância quadrática (euclidiana, euclidiana reduzida, do Chi^2, ...).

* Desenvolvimento do algoritmo de agregação:

A partir de uma distância quadrática $\mathbf{d}_{(x,y)}$ entre objetos, define-se as inércias intra-classes de todas as classes compostas por agregação de duas classes de um só objeto.

Seja $\mathbf{d}_{(x,y)}$ a distância entre o objeto x e o objeto y pertencentes ao conjunto I a classificar.

Cada classe de um objeto possui uma massa m_i associada a ela.

A inércia intra-classe da classe $\{\{x\},\{y\}\}$ é definida por :

$$\Delta_{(\{\{x\},\{y\}\})} = \frac{m_x m_y}{m_x + m_y} d^2_{(x,y)} \quad \forall x, y \in I$$

Os valores $\Delta_{(\{\{x\},\{y\}\})} \quad \forall x, y \in I$ valores constituem a matriz **D**(n,n) que fixa o ponto de partida do algoritmo.

As classes $\{x\}$e $\{y\}$ que integram a classe de dois objetos $\{\{x\},\{y\}\}$ com inércia intra-classe mínima $\left(\Delta_{(\{\{x\},\{y\}\})\ mínimo}\right)$ são reunidas.

6. As classificações hierárquicas ascendentes.

6.6.O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

(continuação)

Na etapa seguinte, calcula-se o crescimento da inércia intra-classe que resulta da criação de uma nova classe produzida pela fusão das classes {x} e {y} com cada classe {k}.

O crescimento da inércia intra-classe é definido assim $\forall x,y,k \in I$:

$$\Delta_{(\{\{x\},\{y\},\{k\}\})} = (m_x + m_k)\Delta_{(\{\{x\},\{k\}\})} + (m_y + m_k)\Delta_{(\{\{y\},\{k\}\})} - \frac{m_k}{m}\Delta_{(\{\{x\},\{y\}\})}$$

Sendo : $m = m_x + m_y + m_k$.

A inércia intra-classe do sub-conjunto de tres elementos será maior que a inércia intra-classe dos subconjuntos de dois elementos da etapa anterior.

A agregação se faz respeitando o critério de um crescimento mínimo da inércia intra-classe do novo subconjunto criado.

* Vantagens do método de agregação:

- Em cada etapa, se constrói uma partição do conjunto de elementos a classificar composta de classes homogêneas (com inércia intra-classe mínima).

- Como o critério de agregação é o crescimento da inércia intra-classe, os níveis de agregação podem ser expressos — em cada etapa — em termos de taxas de crescimento que traduzem a relação entre inércia intra-classes e inércia total.

- A formação de classes bem diferenciadas é garantida pelo critério de minimizar o crescimento da inércia intra-classe.

* Desvantagens do método de agregação:

- Os nós da hierarquia representam um crescimento de inércia proporcional ao quadrado das distâncias... Os grupos de baixo nível parecem muito mais homogêneos e diferenciados entre eles.

- Tendência de produzir grupos esféricos com massas equilibradas.

- Dificuldade de detectar objetos isolados ou grupos algo estirados.

6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K)

. Os indivíduos observados são apresentados, em linhas da tabela **TC**(J,K), em **J** subconjuntos segundo a partição correspondente à variável categórica em linha.

. A cada um dos **J** subconjuntos se pode associar uma massa $\mathbf{m}_j$, dada pela distribuição marginal da variável categórica em linha

. A comparação desses **J** objetos a classificar será feita através dos perfis desses subconjuntos nas modalidades da variável categórica em coluna.

Primeira etapa:

A matriz D(J,J) é definida com a distância do Chi² entre todas as linhas de **TC**(J,K) tomadas duas a duas...

$$d^2_{(j,j')} = \sum_{k=1}^{k} \frac{1}{f_{.k}} \left(\frac{f_{jk}}{f_{j.}} - \frac{f_{j'k}}{f_{j'.}} \right)^2 \quad \forall j, j' \in J$$

Etapas de agregação:

Agregar em uma nova classe dois elementos (duas linhas da tabela **TC**(J,K)) produzindo um crescimento mínimo da inércia cuja fusão se traduza em um decréscimo o mínimo do φ^2 da tabela **TC**(J,K).

A inércia total da tabela TC(J,K): $\varphi^2 = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{\left(f_{jk} - f^*_{jk} \right)^2}{f^*_{jk}}$

6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Em que quantidade diminui o φ^2 da tabela TC(J,K) quando se agregam duas linhas quaisquer...? -

Seja a classe $\{j\},\{j'\} \quad \forall j,j' \in J$.

O decréscimo do φ^2 da tabela TC(J,k) quando se agregam as linhas **j** e **j'** é medido por:

$$\Delta_{(\{\{j\},\{j'\}\})} = \sum_k \frac{\left(f_{jk} - f^*_{jk}\right)^2}{f^*_{jk}} + \sum_k \frac{\left(f_{j'k} - f^*_{j'k}\right)^2}{f^*_{j'k}} - \sum_k \frac{\left(f_{jk} - f_{j'k} - f^*_{jk} - f^*_{j'k}\right)^2}{f^*_{jk} + f^*_{j'k}}$$

Pode-se demostrar que :

$$\Delta_{(\{\{j\},\{j'\}\})} = \frac{f_{j.} f_{j'.}}{f_{j.} + f_{j'.}} \sum_k \frac{\left(\frac{f_{jk}}{f_{j.}} - \frac{f_{j'k}}{f_{j'.}}\right)^2}{f_{.k}}$$

A quantidade em que diminui a inércia total quando são agregadas duas linhas de uma tabela de Contingência é proporcional ao quadrado da distância do Chi^2 entre os perfis considerados, com um coeficiente que depende do peso dos dois perfis.

$$\Delta\left(\{\{j\},\{j'\}\}\right) = \frac{f_{j.} f_{j'.}}{f_{j.} + f_{j'.}} \sum_k \frac{\left(\frac{f_{jk}}{f_{j.}} - \frac{f_{j'k}}{f_{j'.}}\right)^2}{f_{.k}} = \frac{m_j m_{j'}}{m_j + m_{j'}} d^2_{\varphi^2}(j,j')$$

6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

		1° año	2°año	3°año	4°año	5°año	Total
Hombres	Estrategia A	175	160	132	145	118	730
	Estrategia B	116	126	120	95	176	633
	Estrategia C	131	135	154	185	345	950
	Estrategia D	17	21	29	44	141	252
Mujeres	Estrategia A	13	5	22	12	19	71
	Estrategia B	19	9	29	21	27	105
	Estrategia C	40	33	110	58	128	369
	Estrategia D	5	3	6	10	13	37
	Total	516	492	602	570	967	3147

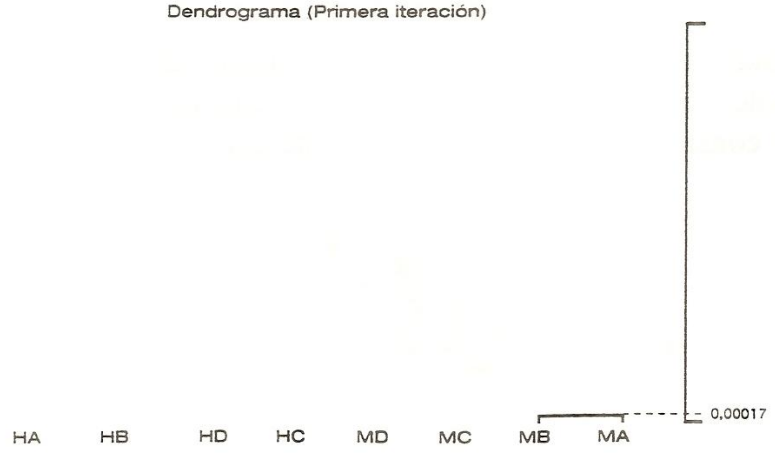
Chi^2= 276.65

φ^2 = 0.0879

Tabla $\Delta_{(J,J)}$

	HA	HB	HC	HD	MA	MB	MC	MD
HA	0							
HB	0,00857	0						
HC	0,03087	0,00867	0					
HD	0,05000	0,02621	0,01213	0				
MA	0,00595	0,00373	0,00405	0,00983	0			
MB	0,00619	0,00391	0,00410	0,01214	0,00017	0		
MC	0,03162	0,01395	0,01084	0,01591	0,00110	0,00184	0	
MD	0,00376	0,00227	0,00063	0,00237	0,00161	0,00119	0,00184	0

Dendrograma (Primera iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

		1° año	2°año	3°año	4°año	5°año	Total
Hombres	Estrategia A	175	160	132	145	118	730
	Estrategia B	116	126	120	95	176	633
	Estrategia C	131	135	154	185	345	950
	Estrategia D	17	21	29	44	141	252
Mujeres	Estrategia A-B	32	14	51	33	46	176
	Estrategia C	40	33	110	58	128	369
	Estrategia D	5	3	6	10	13	37
	Total	516	492	602	570	967	3147

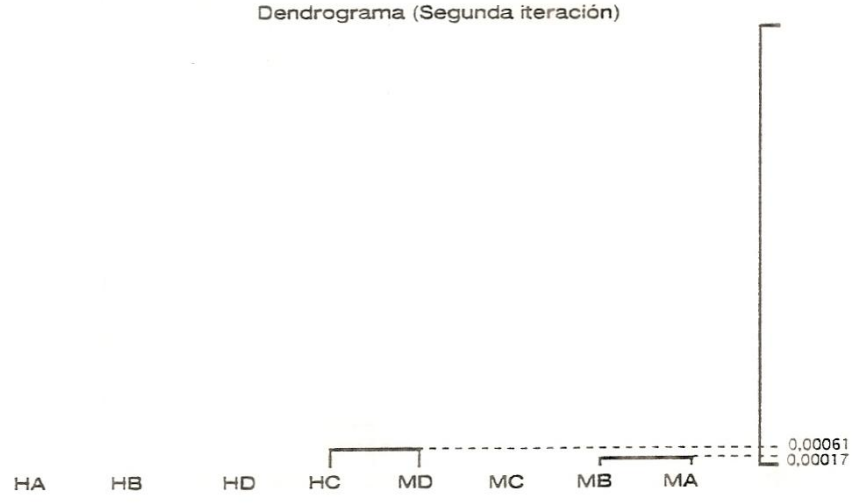
Chi^2= 276.09

φ^2 = 0.0877

Tabla $\Delta_{(J-1,J-1)}$

	HA	HB	HC	HD	MA+MB	MC	MD
HA	0						
HB	0,00857	0					
HC	0,03087	0,00867	0				
HD	0,05000	0,06907	0,01213	0			
MA-MB	0,01083	0,00965	0,00736	0,01744	0		
MC	0,03162	0,00889	0,01084	0,01591	0,00238	0	
MD	0,00376	0,01725	(0,00061)	0,00237	0,00158	0,00184	0

Dendrograma (Segunda iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

	1° año	2°año	3°año	4°año	5°año	Total
Hombres Estrategia A	175	160	132	145	118	730
Estrategia B	116	126	120	95	176	633
Est. C + Mujeres Est. D	131	135	154	185	345	950
Estrategia D	17	21	29	44	141	252
Mujeres Estrategia A-B	32	14	51	33	46	176
Estrategia C	40	33	110	58	128	369
Total	516	492	602	570	967	3147

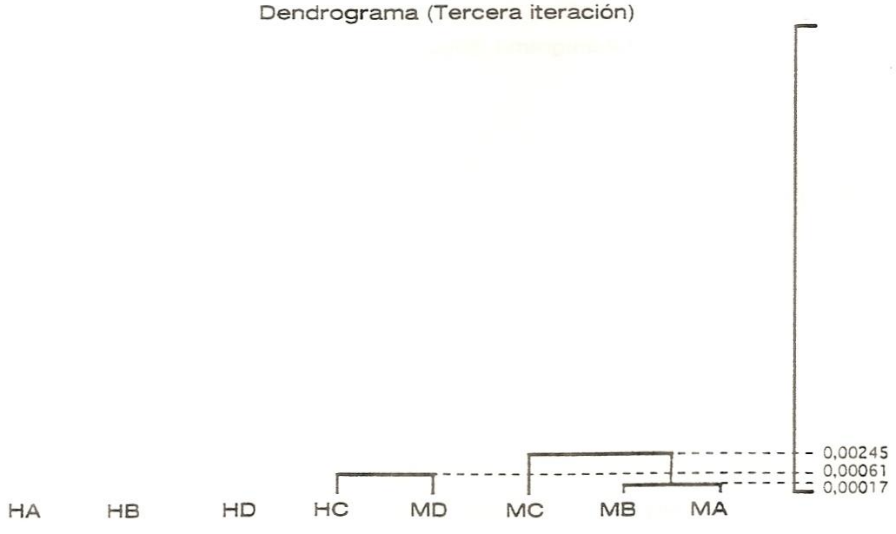
Chi^2= 274.10

φ^2 = 0.0871

Tabla $\Delta_{(J-2,J-2)}$

	HA	HB	HC-MD	HD	MA-MB	MC
HA	0					
HB	0,00857	0				
HC-MD	0,03110	0,00903	0			
HD	0,05000	0,02621	0,01209	0		
MA-MB	0,01083	0,00667	0,00728	0,01744	0	
MC	0,03365	0,01480	0,01157	0,01652	0,00245	0

Dendrograma (Tercera iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

—	1° año	2°año	3°año	4°año	5°año	Total
Hombres Estrategia A	175	160	132	145	118	730
Estrategia B	116	126	120	95	176	633
Est. C + Mujeres Est. D	131	135	154	185	345	950
Estrategia D	17	21	29	44	141	252
Mujeres Estrategia A-B-C	72	47	161	91	174	545
Total	516	492	602	570	967	3147

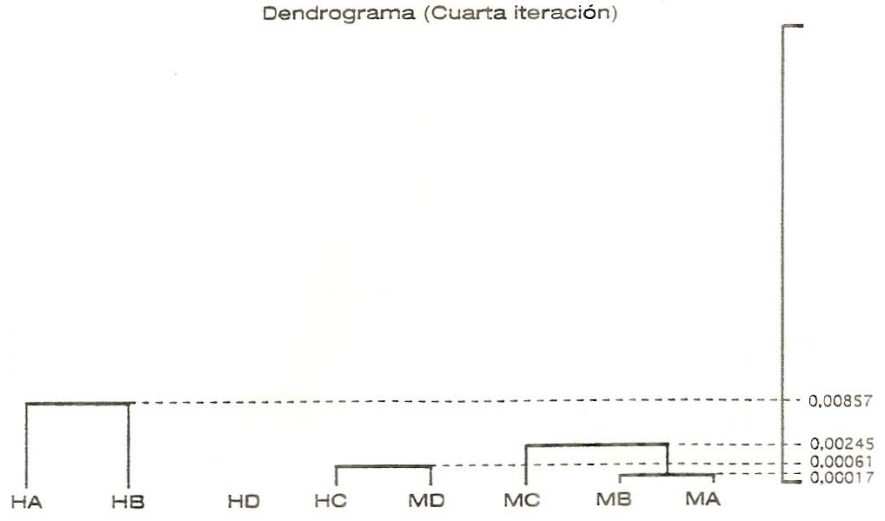
Chi^2= 266.63

φ^2 = 0.0847

Tabla $\Delta_{(J-3,J-3)}$

	HA	HB	HC-MD	HD	MA-MB-MC
HA	0				
HB	0,00857	0			
HC-MD	0,03110	0,00903	0		
HD	0,05000	0,02621	0,01209	0	
MA-MB-MC	0,03485	0,01571	0,01408	0,02144	0

Dendrograma (Cuarta iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

	1° año	2°año	3°año	4°año	5°año	Total
Hombres Estrategia A-B	291	286	252	240	294	1363
Est. C + Mujeres Est. D	131	135	154	185	345	950
Estrategia D	17	21	29	44	141	252
Mujeres Estrategia A-B-C	72	47	161	91	174	545
Total	516	492	602	570	967	3147

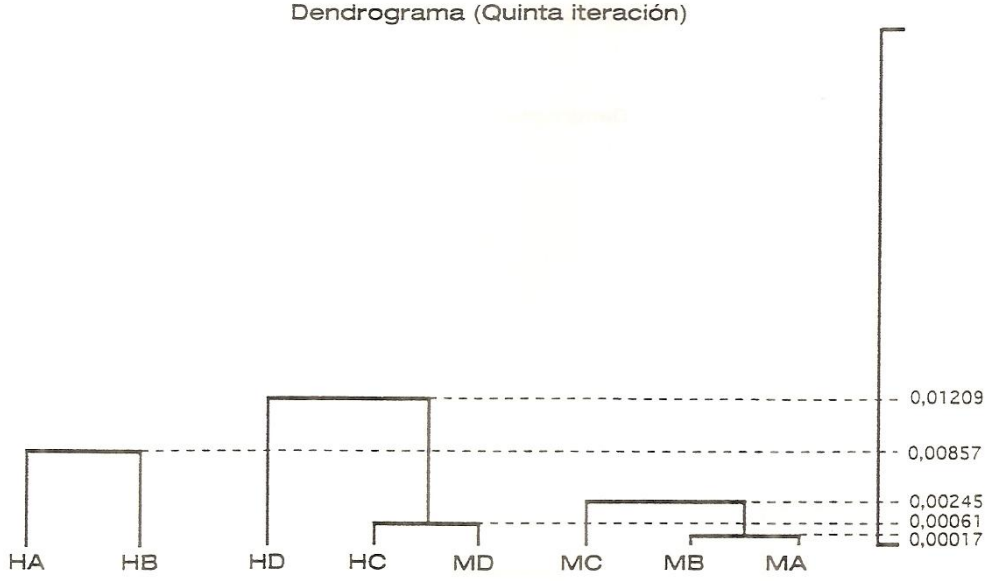
Chi^2= 239,66

φ^2 = 0,0762

Tabla $\Delta_{(J-4,J-4)}$

	HA-HB	HC-MD	HD	MA-MB-MC
HA-HB	0			
HC-MD	0,02524	0		
HD	0,04343	(0,01209)	0	
MA-MB-MC	0,03076	0,01408	0,02144	0

Dendrograma (Quinta iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

–	1° año	2°año	3°año	4°año	5°año	Total
Hombres Estrategia A-B	291	286	252	240	294	1363
Est. C + Mujeres Est. D + Hombres Est. D	131	135	154	185	345	950
Mujeres Estrategia A-B-C	72	47	161	91	174	545
Total	516	492	602	570	967	3147

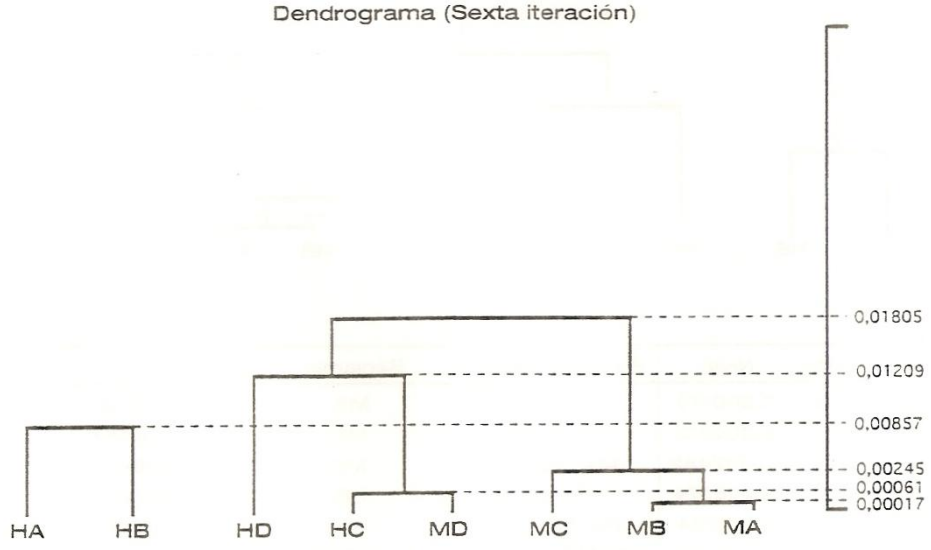
Chi^2= 201,295

φ^2 = 0,06396

Tabla $\Delta_{(J-5,J-5)}$

	HA-HB	HC-MD-HD	MA-MB-MC
HA-HB	0		
HC-MD-HD	0,04319	0	
MA-MB-MC	0,03076	0,01805	0

Dendrograma (Sexta iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método baseado no crescimento mínimo do momento de ordem dois nas classes das partições encaixadas.

Exemplo numérico: classificação hierárquica das linhas de uma tabela TC(J,K) (continuação)

Tabla de Contingencia : "Vietnam"

	1° año	2°año	3°año	4°año	5°año	Total
Hombres Estrategia A-B	291	286	252	240	294	1363
HC + M D + H D + MA + MB + MC	225	206	350	330	673	1784
Total	516	492	602	570	967	3147

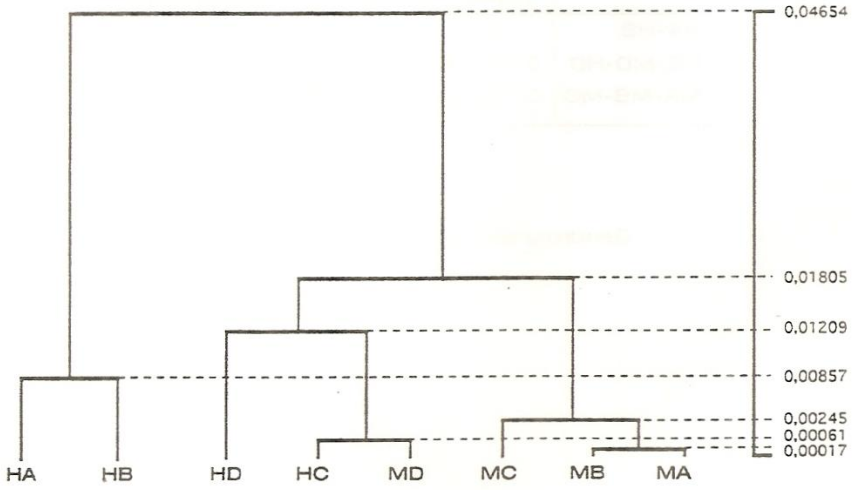
Chi^2= 146,457

φ^2 = 0.04654

Tabla $\Delta_{(J-6,J-6)}$

	HA-HB	HC-MD-HD-MA-MB-MC
HA-HB	0	
HC-MD-HD-MA-MB-MC	(0,04654)	0

Dendrograma



Nudo	Nivel	Primogénito	Benjamin	Masa
9	0,000176	MA	MB	0,0559
10	0,000613	HC	MD	0,30187
11	0,002452	MA-MB	MC	0,18494
12	0,008574	HA	HB	0,4331
13	0,012094	HC-MD	HD	0,38195
14	0,018053	HC-MD-HD	MA-MB-MC	0,56689
15	0,046538	HA-HB	HC-MD-HC-MA-MB-MC	1

6. As classificações hierárquicas ascendentes.

6.6. O método de Ward.

No caso de uma tabela T(n,p) de variáveis quantitativas, a estratégia de agregação do "crescimento mínimo do momento de ordem dois" é chamado método de Ward.

O princípio de funcionamento do método de Ward pode ser apresentado (*) como uma generalização multidimensional do modelo da Análise da Variância

$$SCD_{tot} = SCD_{res} + SCD_{fac} \qquad (1)$$

SCD_{tot} : soma dos quadrados dos desvios das observações à média geral.

SCD_{res} : soma dos quadrados dos desvios das observações em cada grupo, com respeito à média do grupo para todos os grupos.

SCD_{fac} : soma dos quadrados dos desvios das observações em cada grupo, com respeito à média geral.

Se a tabela T(n,p) contém uma só variável e são distinguidos K grupos nas observações da expressão (1) resulta que:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{k=1}^{K}\sum_{i=1}^{n}(x_i-\bar{x}_k)^2+\sum_{k=1}^{K}n_k(\bar{x}_k-\bar{x})^2 \qquad (2)$$

A inércia total é decomposta em uma soma de inércia intra-classes e da inércia inter-classes.

Se a tabela T(n,p) contém mais de uma variável e se se distinguem K grupos de observações, se substitui em (2) os desvios relativamente à média pelo quadrado das distâncias euclidianas relativamente ao centro de gravidade

$$\sum_{i=1}^{n}d^2_{(i,G)}=\sum_{k=1}^{K}\sum_{i=1}^{n_k}d^2_{(i,G_k)}+\sum_{k=1}^{K}n_k d^2_{(G_k,G)} \qquad (3)$$

(*) ROUX, M. "Classification des données d'enquêtes", in : GRANGÉ, D. et LEBART, L. "*Traitement statistiques des enquêtes*", Dunod, Paris, 1993.

6. As classificações hierárquicas ascendentes.

6.6. O método de Ward. (continuação)

Sendo :

Coordenadas de G : $\left\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p\right\}$, termo geral: : $\bar{x}_p = \frac{1}{n}\sum_{i=1}^{n} x_{ip}$.

Coordenadas de G_k : $\left\{\bar{x}_{k1}, \bar{x}_{k2}, \ldots, \bar{x}_{kp},\right\}$, termo geral : $\bar{x}_{kp} = \frac{1}{n}\sum_{i=1}^{n_k} x_{ip}$.

Coordenadas de i : $\left\{x_{i1}, x_{i2}, \ldots, x_{ip}\right\}$.

Distâncias euclidianas :

$$d_{(i,G)} = \sum_{p=1}^{P}\left(x_{ip} - \bar{x}_p\right)^2 ; \forall i = 1,\ldots I, \forall p = 1,\ldots P$$

$$d^2_{(i,G_k)} = \sum_{p=1}^{P}\left(x^k_{ip} - \bar{x}_{kp}\right)^2 ; \forall i = 1,\ldots I, \forall k = 1,\ldots K, \forall p = 1,\ldots P$$

$$d^2_{(G_k,G)} = \sum_{p=1}^{P}\left(\bar{x}_{kp} - \bar{x}_p\right)^2 ; \forall k = 1,\ldots K, \forall p = 1,\ldots P$$

Substituindo em (3) :

$$\sum_{i=1}^{n}\sum_{p=1}^{P}\left(x_{ip} - \bar{x}_p\right)^2 = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\sum_{p=1}^{P}\left(x^k_{ip} - \bar{x}_{kp}\right)^2 + \sum_{k=1}^{K}\sum_{p=1}^{P} n_k\left(\bar{x}_{kp} - \bar{x}_p\right)^2 \qquad (4)$$

Se as observações se apresentam em grupos bem diferenciados:

- as inércias intra-grupos devem ser baixas,
- as inércias inter-grupos devem ser elevadas.

Critério de agregação : minimizar o crescimento da inércia intra-grupos resultante da agregação de dois grupos numa nova classe.

6. As classificações hierárquicas ascendentes.

6.6. O método de Ward. (continuação)

Se forem agrupadas as classes {k} e {k'}, a expressão (3) permite avaliar a inércia intra-grupos da nova classe $\{K \cup K'\}$:

$$\sum_{i=1}^{n} d^2_{(i,G_{k\cup k'})} = \underbrace{\sum_{i=1}^{n_k} d^2_{(i^k,G_k)} + \sum_{i=1}^{n_{k'}} d^2_{(i^{k'},G_{k'})}}_{\text{inercias intra-grupos de las clases componentes}} + \underbrace{n_k\, d^2_{(G_k,\, G_k \cup G_{k'})} + n_{k'}\, d^2_{(G_{k'},\, G_k \cup G_{k'})}}_{\text{inercia inter-grupos}}$$

A inércia intra-grupos da classe $\{K \cup K'\}$ é mais elevada que a soma das inércias intra-grupos das classes que a compõem {k} e {k'}.

O aumento de inércia intra-grupo da nova classe é definido por :

$$\Delta_{(k\cup k')} = n_k d^2_{(G_k, G_{k\cup k'})} + n_{k'} d^2_{(G_{k'}, G_{k\cup k'})} \qquad (5)$$

Critério de agregação: reunir as classes que minizem a quantidade $\Delta_{(k\cup k')}$.

O algoritmo deve calcular, em cada etapa, os valores entre todos os pares de classes já definidas e selecionar o valor mínimo, para agregar as classes correspondentes.

Pode-se demonstrar que:

$$\Delta_{(k\cup k',s)} = \frac{n_k + n_s}{n_k + n_{k'} + n_s}\Delta_{(k,s)} + \frac{n_{k'} + n_s}{n_k + n_{k'} + n_s}\Delta_{(k',s)} + \frac{n_s}{n_k + n_{k'} + n_s}\Delta_{(k,k')} \quad (6)$$

Isto é, o crescimento da inércia intra-grupos da nova classe $\{k \cup k' \cup s\}$ pode ser avaliado sem referência ao centro de gravidade da nova classe $\left(G_{(k\cup k'\cup s)}\right)$.

Indubitavelmente, para poder aplicar a expressão (6) como critério de agregação, é necessário recalcular as distâncias euclidianas entre todas as classes definidas em cada etapa.

Limitações de capacidade de cálculo...

6. As classificações hierárquicas ascendentes.

6.6. O método de Ward. (continuação)

Pode-se demonstrar que a expressão (5) pode ser formulada assim :

$$\Delta_{(k\cup k')} = \frac{n_k . n_{k'}}{n_k + n_{k'}} d^2_{(G_k, G_{k'})} \qquad (7)$$

Esta expressão :

- não faz referência ao centro de gravidade $G_{(k\cup k')}$.

- permite trabalhar com a tabela T(n,p) substituindo, em cada etapa, os n_k indivíduos que foram agregados na classe **k** pelo centro de gravidade G_k correspondente.

6. As classificações hierárquicas ascendentes.

6.6. O método de Ward: exemplo numérico.

Datos

	X	Y
A	1	1
B	2	2
C	3,5	4,5
D	5,5	3
E	6	5
F	5	5

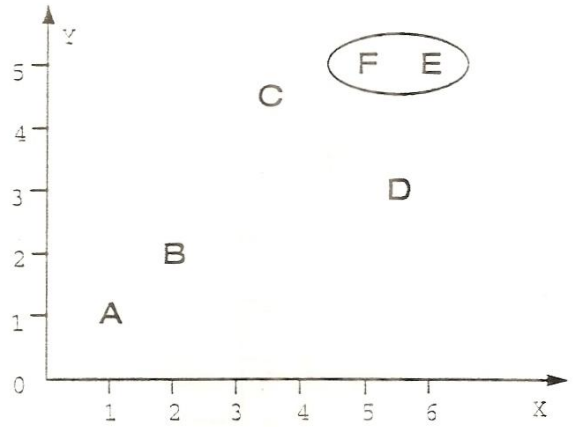
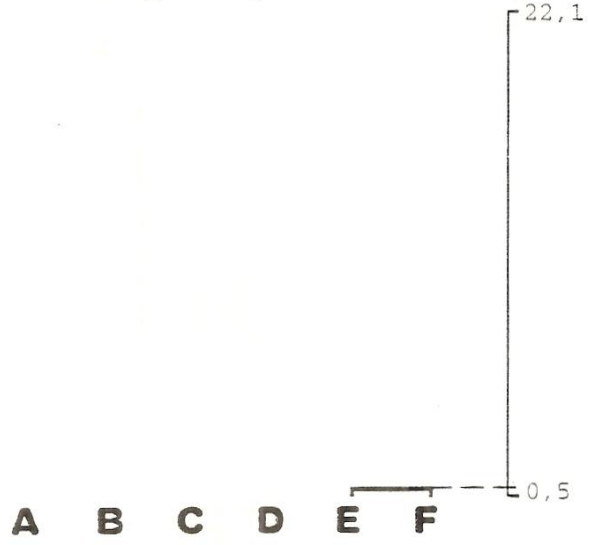


Tabla Δ(i,i)

	A	B	C	D	E	F
A	0					
B	1	0				
C	9,250	4,250	0			
D	12,125	6,625	3,125	0		
E	20,500	12,50	3,250	2,125	0	
F	16,000	9,00	1,250	2,125	0,500	0

Dendrograma (Primera iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método de Ward: exemplo numérico.

(continuação)

Datos

	X	Y	Peso
A	1	1	1
B	2	2	1
C	3,5	4,5	1
D	5,5	3	1
E-F	5,5	5	2

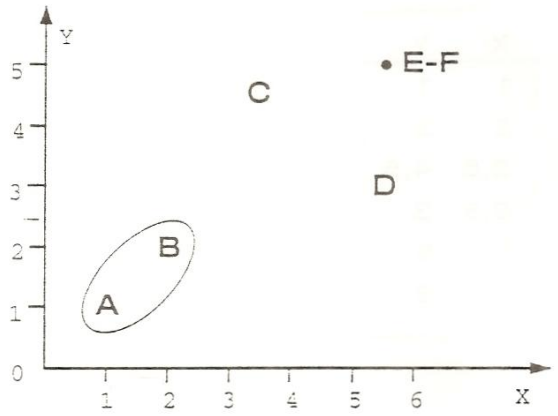
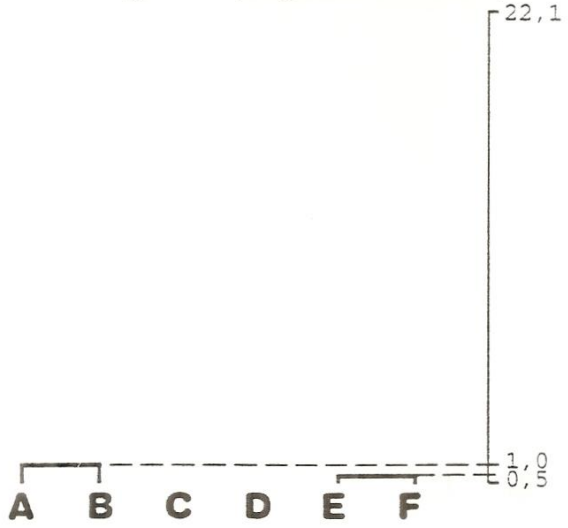


Tabla Δ(i-1,i-1)

	A	B	C	D	E-F
A	0				
B	1,000	0			
C	9,250	4,250	0		
D	12,125	6,625	3,125	0	
E-F	24,167	14,167	2,833	2,667	0

Dendrograma (Segunda iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método de Ward: exemplo numérico.

(continuação)

Datos

	X	Y	Peso
A-B	1,5	1,5	2
C	3,5	4,5	1
D	5,5	3,0	1
E-F	5,5	5	2

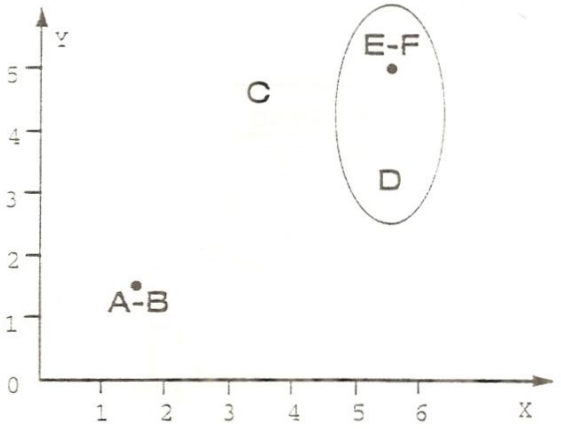
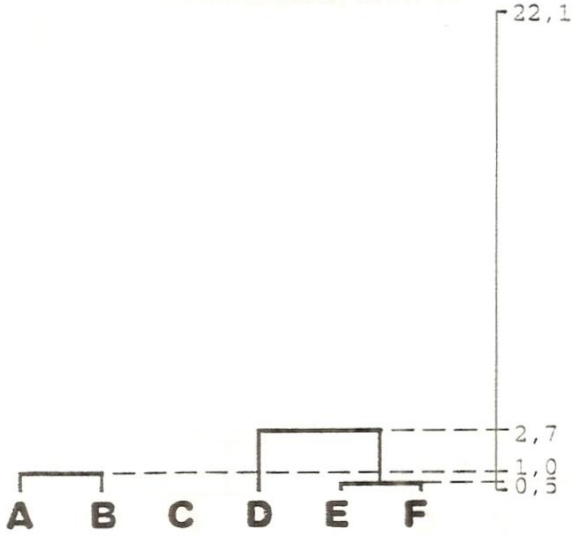


Tabla Δ(i-2,i-2)

	A-B	C	D	E-F
A-B	0			
C	8,667	0		
D	12,167	3,125	0	
E-F	28,250	2,833	2,667	0

Dendrograma (Tercera iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método de Ward: exemplo numérico.

(continuação)

Datos

	X	Y	Peso
A-B	1,5	1,5	2
C	3,5	4,5	1
E-F-D	5,5	4,3	3

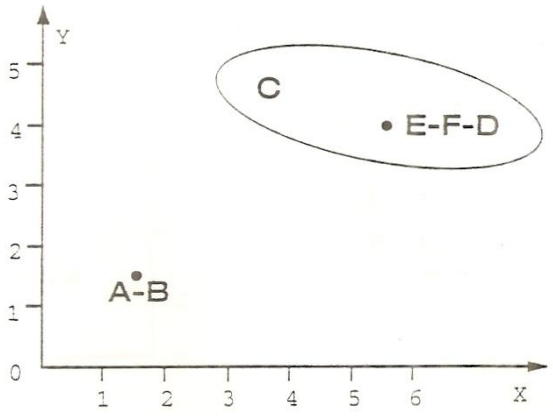
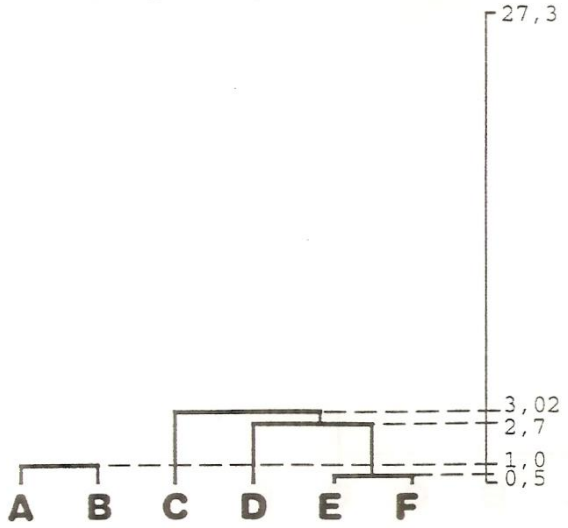


Tabla Δ(i-3,i-3)

	A-B	C	E-F-D
A-B	0		
C	8,67	0	
E-F-D	28,83	3,02	0

Dendrograma (Cuarta iteración)



6. As classificações hierárquicas ascendentes.

6.6. O método de Ward: exemplo numérico.

(continuação)

Datos

	X	Y	Peso
A-B	1,5	1,5	2
E-F-D-C	5	4,4	4

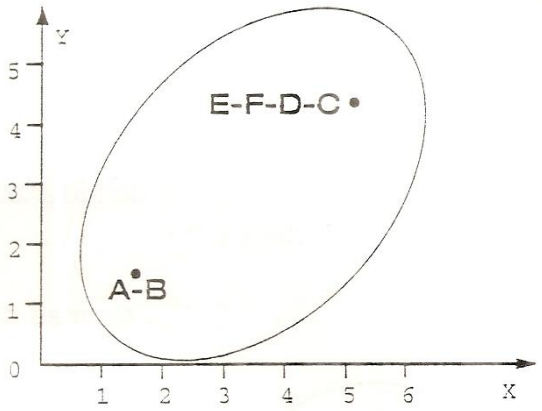
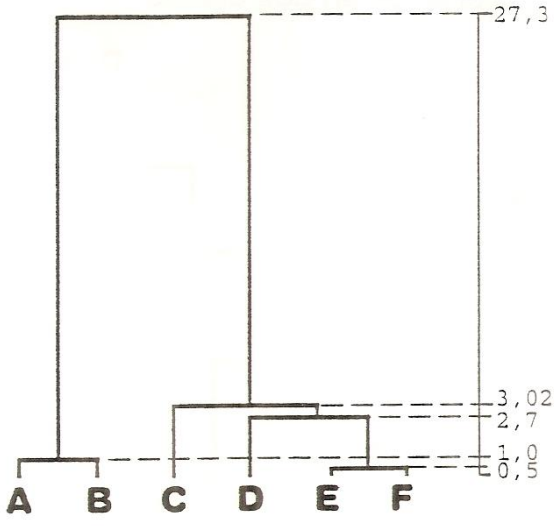


Tabla Δ(i-4,i-4)

	A-B	E-F-D-C
A-B	0	
E-F-D-C	27,35	0

Dendrograma (Quinta iteración)



6. As classificações hierárquicas ascendentes.

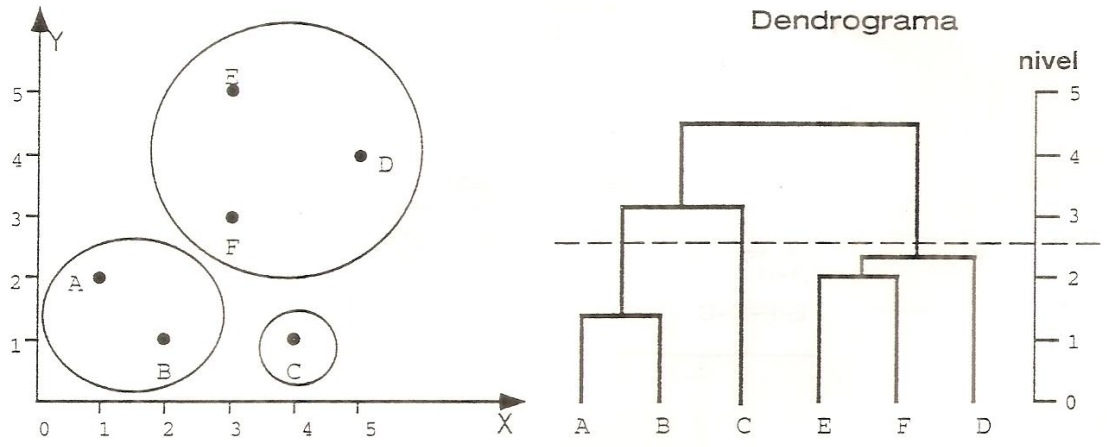
6.7. Como selecionar uma partição a partir de uma hierarquia?

A partir de uma "árvore de classificação" pode-se escolher uma 'boa" partição dos n objetos submetidos à classificação hierárquica ascendente.

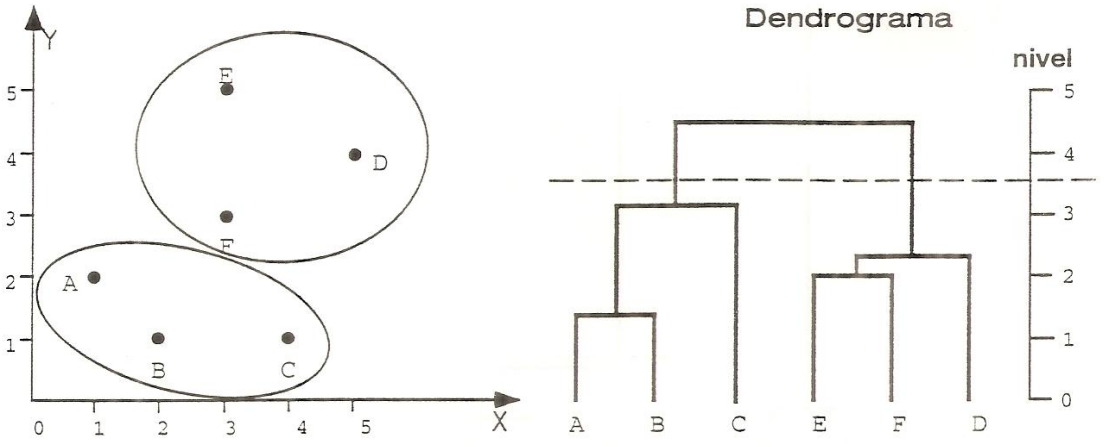
Para selecionar uma 'boa" partição deve-se escolher um nível de agregação para a qual o valor do índice não seja muito elevado ... baixa transformação das distâncias iniciais entre os objetos.

Para isso é suficiente "cortar" o dendograma com uma reta horizontal que cruze os ramos ascedentes mais compridos....

Partición en tres clases



Partición en dos clases



7.7. As classificações não hierárquicas:

7.1 Método de "agregação em torno de centros móveis".

* Procedimento:

1. **Determina-se o número de classes desejado.**
2. **Define-se uma partição inicial (eventualmente, afetando aleatoriamente os indivíduos a cada classes).**
3. **Recentrando: calcula-se o centro de gravidade de cada classe.**
4. **Realocação: distribui-se os indivíduos em cada grupo segundo sua semelhança com o novo centro de gravidade calculado.**

Repetem-se as etapas 3 e 4 até obter a estabilidade das classes.

7.2 Variantes do método de partições:

- **Variante K-médias: efetua-se o "recentro" quando um indivíduo muda de classe.**
- **Variante "nuvens dinâmicas": o "recentro" é feito quando todos os indivíduos tenham sido alocados em uma classe**
- **Variante "Isodata": impõe-se condições suplementares para impedir a formação de classes com efectivos muito baixos ou de diâmetro muito grande.**
- **Variante "indivíduos- típicos": em lugar de tomar como referência o centro de gravidade de uma classe (indivíduo médio, não observado), pode-se selecionar um indivído-típico em cada classe em formação. Os indivíduos-típicos constituem os nós de cada classe.**

7. 7. As classificações não hierárquicas:

7.3 Vantagens e inconvenientes dos métodos de partição.

* **Vantagens:**

- Estes métodos tendem a maximizar os momentos de inércia inter-classes.

- Trabalha-se com a tabela T(n,p) e uma tabela de centros de gravidade ou de nós de classe, cujo tamanho depende do número de classes escolhido.

- As distâncias a calcular são somente entre os indivíduos e os centros de gravidade.

* **Incovenientes:**

Os resultados dependem da escolha da partição inicial.

* **Solução:**

1 Com um dado número de classes, se efetua uma partição inicial aleatória seguida do algoritmo de partição.

2 Recomeça-se o processo com outras alocações aleatórias seguidas de aplicações do algoritmo de partição.

Os resultados (a composição das classes) não serão os mesmas ao final de diferentes ensaios.

* **Uma análise complementar permite ver que:**

- Certos indivíduos foram alocados, em todos os ensaios, na mesma classe. Estes nós de indivíduos "estáveis" constituem as "formas fortes" = grupos homogêneos cuja existência não depende do número de classes selecionado.

- Outros indivíduos mudam de classe em cada ensaio = "formas débeis"= indivíduos isolados ou intermédios entre "formas fortes".

8. O tratamento de grandes tabelas resultantes de levantamentos por amostragem: a estratégia "análise fatorial + classsificação".

* Dois métodos de classificação permitem o tratamento de grandes tabelas... o método de Ward e os métodos "agregação em torno a centros móveis". Ambos trabalham com os centros de gravidade das classes em formação e tendem a otimizar a inércia inter-classe da partição resultante.

  Mas esses métodos só podem ser utilizados com tabelas T(n,p) de variáveis quantitativas.

  Os dados de "survey" conduzem a tabelas T(n,p) de variáveis categóricas (eventualmente ordinais).

* A A.F.C.M. desse tipo de tabelas pode ser considerada como uma etapa preliminar à estratégia de classificação.

* As coordenadas fatoriais dos indivíduos sobre os primeiros eixos de uma A. F. C. M. constituem um "bom" resumo da tabela de dados brutos que resulta da observação.

* Dispõe-se assim de uma tabela "indivíduos x variáveis quantitativas" que pode ser submetida à classificação.

8. O tratamento de grandes tabelas resultantes de "survey": a estratégia "análise fatorial + classsificação".

* Quantos eixos devem ser conservados na nova tabela resumo..?

* Vantagens:

- Estabilidade dos eixos fatoriais relativamente à sondagem. Instabilidade dos métodos de classificação. O pré-tratamento fatorial permite remediar parcialmente a esse problema.

- O feito de conservar um baixo número de eixos pode ser considerado como uma maneira de eliminar flutuações aleatórias que escondem os fenômenos importantes presentes nos dados. O tratamento fatorial opera como um filtro da informação importante.

- O uso das coordenadas fatoriais permite utilizar algoritmos de classificação adequados, o qual dá um "ponto de vista" original sobre a estrutura dos dados.

* Estratégia de elaboração de uma partição:

1. Análise fatorial de correspondências múltiplas.
2. Construção de uma classificação hierárquica (método de Ward) sobre a tabela "indivíduos x coordenadas fatoriais" (observando um número adequado de eixos fatoriais).
3. Corte da "árvore de classificação" num número adequado de classes.
4. Elaboração de uma partição pelo método de "agregação em torno a centros móveis".

Importância desta estratégia...