PRO2514 - Pesquisa Quantitativa em Gestão de Operações

Análise de Clusters
Análise de Agrupamentos
Análise de Conglomerados

Prof. Dr. Renato de Oliveira Moraes

Sumário

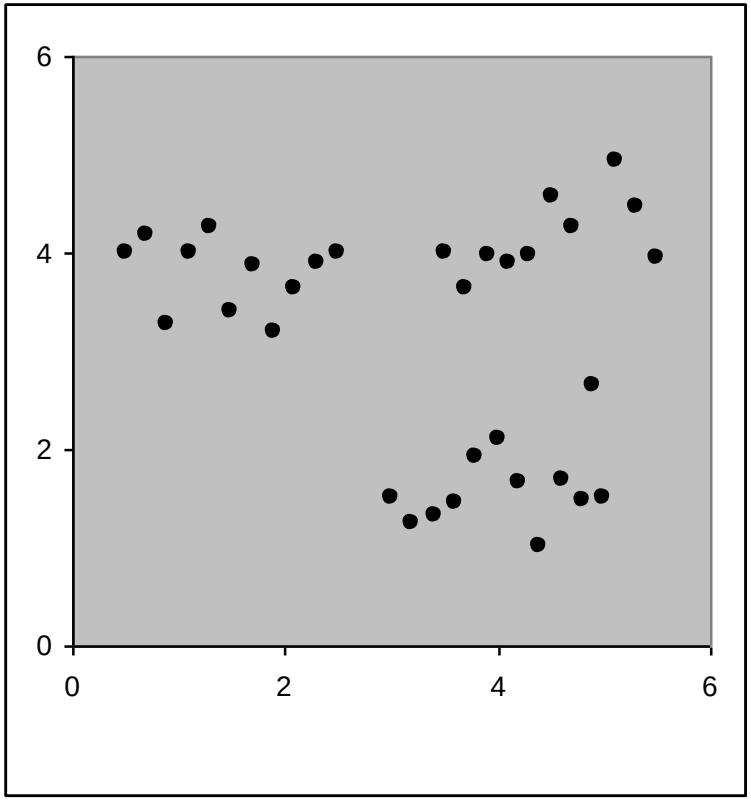
- Conceito geral (homogeneidade interna e heterogeneidade externa)
- Métodos de agrupamento: hierárquico x não hierárquico
- Medidas de (dis)similaridade
- Efeito da escala e padronização das variáveis
- Seleção das variáveis e sua influência no resultado final
- Quantidade de grupos formados
- Análise de Variância
- Significado dos grupos
- Validação dos resultados

Conceito geral

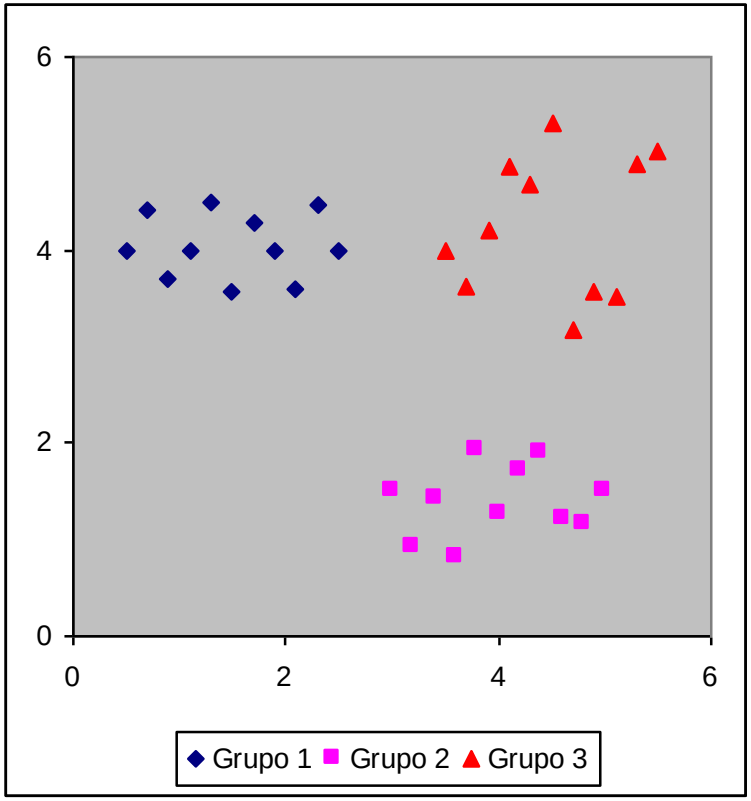
Divide a população em sub-populações que possuem características homogêneas dentro dos clusters e heterogêneas entre clusters, ou seja:

- Dentro do grupo (cluster) a variância é mínima;
- Entre grupos (clusters) a variância é máxima.

Análise de Conglomerados (*Clusters*)



Análise de Conglomerados (*Clusters*)



Métodos de agrupamento

- Hierárquico Aglomerativo:
  - Inicia-se com N grupos, cada grupo com um elemento;
  - A cada etapa, os dois grupos mais semelhantes são unidos. Ou seja, a cada etapa a quantidade de grupos se reduz em 1;
  - Em nenhum momento, os elementos de um grupo são separados.
- Hierárquico Divisivo:
  - Inicia-se com um único grupo com todos elementos;
  - A cada etapa, o grupo com menor homogeneidade interna é divido em dois grupos.
  - Ou seja, a cada etapa a quantidade de grupos aumenta em 1.
- Não hierárquico
  - Há uma definição prévia da quantidade (k) de grupos a serem formados;
  - Inicialmente os N elementos são divididos nos K grupos;
  - A cada etapa alguns elementos trocam de grupo de forma a maximizar a heterogeneidade entre os grupos.

Métodos Hierárquicos Aglomerativos

- Single linkage (SL) ou (vizinhos mais próximos): agrupa-se as pessoas com a distância mínima
- Complete linkage (CL) ou (vizinhos mais distantes): agrupa-se as pessoas com a distância máxima
- Average linkage (vizinhos comuns): é intermediária às duas anteriores (single e complete linkage), trabalha com valores ponderados das distâncias.
- Método da Centróide: distância entre dois clusters é a distância entre os centróides dos grupos
- Método Ward (mais usado): combina os indivíduos dentro dos clusters de acordo com o critério do menor incremento de soma total da distância euclidiana ao quadrado dentro do cluster

Sugestão de procedimento exploratório

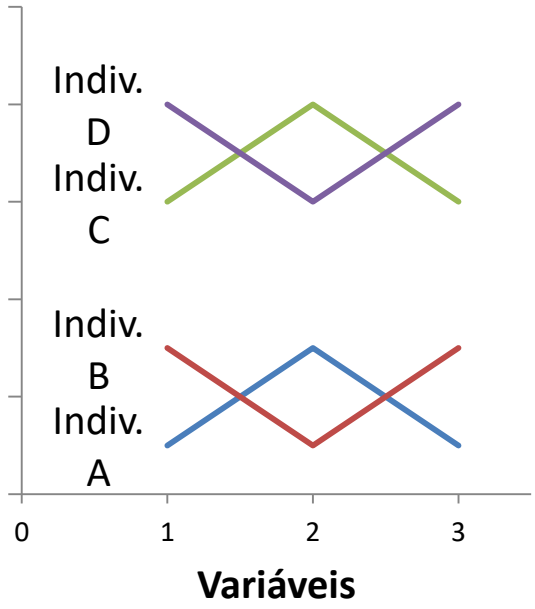
Para estudos exploratórios, onde deseja-se criar uma taxonomia (que não foi definida a priori)

1. Usar método hierárquico para determinar a quantidade K de grupos a serem formados
2. Utilizar um método não hierárquico (K-means no SPSS) para a formação dos grupos

Medidas de (dis)similaridade

É possível usar dois conceitos distintos para medida de similaridade:

- Distância
  - Grupo 1: A e B
  - Grupo 2: C e D
- Correlações.
  - Grupo 1: A e C
  - Grupo 2: B e D



Medidas de Distância

- distância euclidiana
- distância euclidiana ao quadrado
- distância euclidiana de Mahalanobis

Obs:

- Conforme o critério dois respondentes podem estar no mesmo grupo ou em grupos diferentes
- Pressupõe variáveis métricas
- Em princípio, a distância de Mahalanobis é o melhor critério, mas depende do caso, do número de variáveis e sobretudo do número de respondentes

Efeito da escala e padronização das variáveis

- É fortemente recomendável padronizar as variáveis quando diferentes escalas são utilizadas. Nestas condições, os dados padronizados se tornam adimensionais e podem comparados

- Uma opção é usar a Curva Normal (Z scores).
- Caso esteja trabalhando com fatores extraídos na análise fatorial, os dados estarão provavelmente padronizados.
- Cuidado com os outliers (valores absurdos, extremos, fora do padrão)

Efeito da escala e padronização das variáveis

Indivíduo	Peso (Kg)	Altura (m)
1	90	1,8
2	82,5	1,75
3	78	1,85
4	81	1,77

Medidas de similaridade (Distância Euclidiana)

	2	3	4
1	7,50	12,00	9,00
2		4,50	1,50
3			3,00

Indivíduo	Peso (A)	Altura (cm)
1	6	180
2	5,5	175
3	5,2	185
4	5,4	177

Medidas de similaridade (Distância Euclidiana)

	2	3	4
1	5,02	5,06	3,06
2		10,00	2,00
3			8,00

Indivíduo	Peso	Altura
1	1,40	0,17
2	-0,07	-0,98
3	-0,95	1,32
4	-0,37	-0,52

Medidas de similaridade (Distância Euclidiana)

	2	3	4
1	1,87	2,62	1,89
2		2,46	0,55
3			1,93

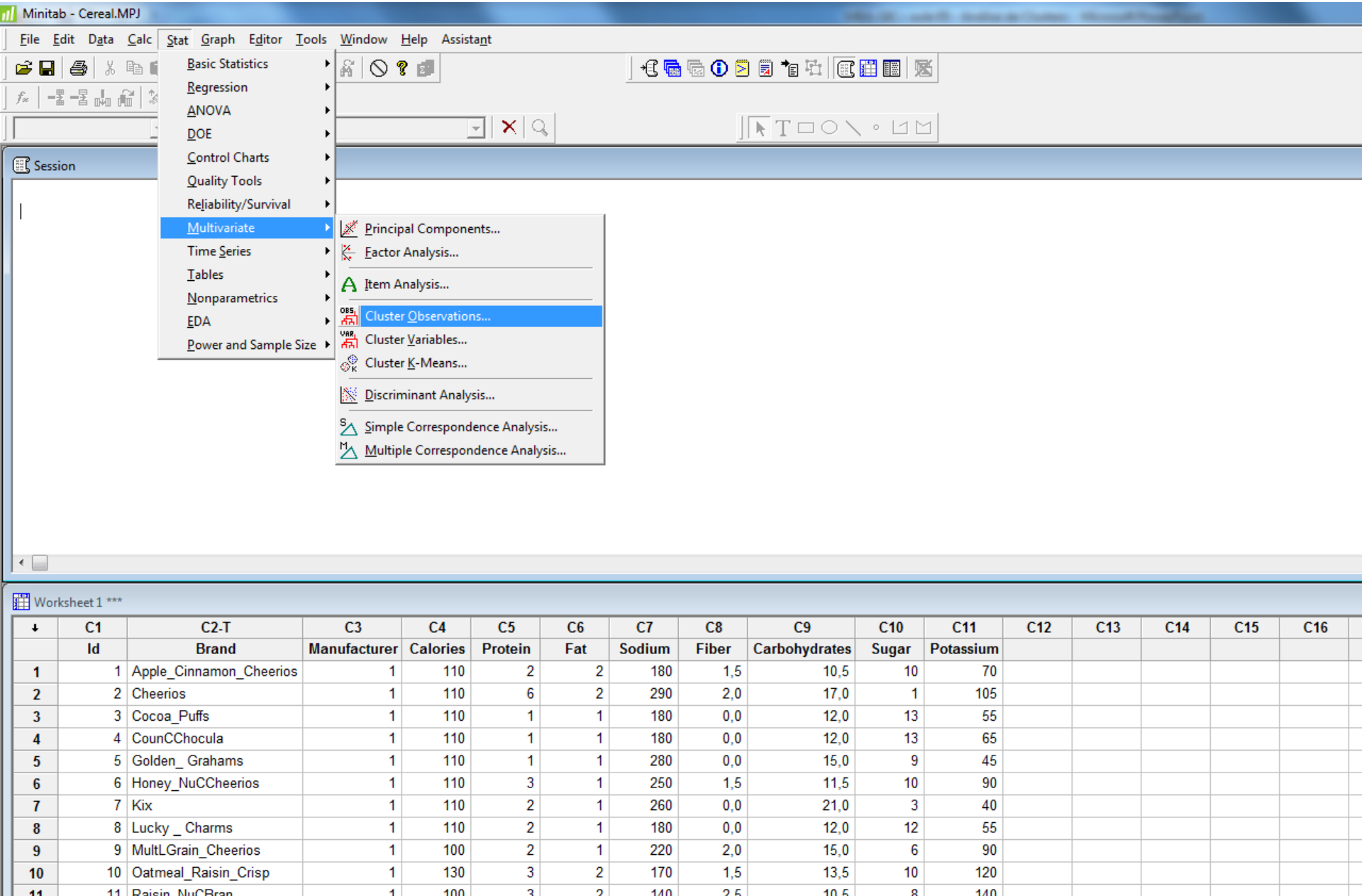
Seleção das variáveis e sua influência no resultado final

Indivíduo	Peso 1 (Kg)	Peso 2 (Kg)	Altura (m)
1	95,00	32,00	1,75
2	92,00	30,00	1,70
3	85,00	28,00	1,72
4	82,00	25,00	1,66

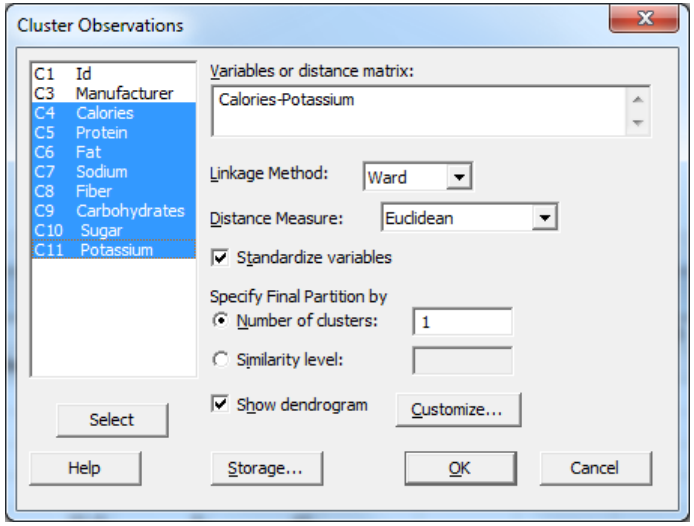
	2	3	4
1	1,57	2,28	3,98
2		1,44	2,58
3			1,95

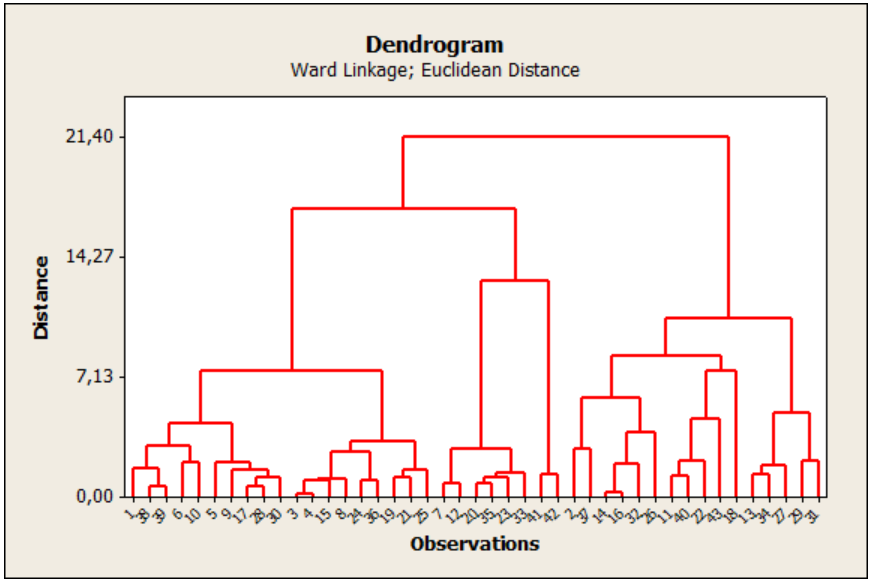
Indivíduo	Peso 1 (Kg)	Altura 1 (m)	Altura 2 (m)
1	95,00	1,75	0,98
2	92,00	1,70	0,95
3	85,00	1,72	0,97
4	82,00	1,66	0,93

	2	3	4
1	1,90	1,89	3,94
2		1,51	2,21
3			2,47



↓	C1	C2-T	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
	Id	Brand	Manufacturer	Calories	Protein	Fat	Sodium	Fiber	Carbohydrates	Sugar	Potassium					
1	1	Apple_Cinnamon_Cheerios	1	110	2	2	180	1,5	10,5	10	70					
2	2	Cheerios	1	110	6	2	290	2,0	17,0	1	105					
3	3	Cocoa_Puffs	1	110	1	1	180	0,0	12,0	13	55					
4	4	CounCChocula	1	110	1	1	180	0,0	12,0	13	65					
5	5	Golden_ Grahams	1	110	1	1	280	0,0	15,0	9	45					
6	6	Honey_NuCCheerios	1	110	3	1	250	1,5	11,5	10	90					
7	7	Kix	1	110	2	1	260	0,0	21,0	3	40					
8	8	Lucky _ Charms	1	110	2	1	180	0,0	12,0	12	55					
9	9	MultLGrain_Cheerios	1	100	2	1	220	2,0	15,0	6	90					
10	10	Oatmeal_Raisin_Crisp	1	130	3	2	170	1,5	13,5	10	120					
11	11	Raisin_NuCBran	1	100	3	2	140	2,5	10,5	8	140					


Cluster Observations
C1 Id
C3 Manufacturer
C4 Calories
C5 Protein
C6 Fat
C7 Sodium
C8 Fiber
C9 Carbohydrates
C10 Sugar
C11 Potassium
Variables or distance matrix:
Calories-Potassium
Linkage Method:
Ward
Distance Measure:
Euclidean
Standardize variables
Specify Final Partition by
Number of clusters:
1
Similarity level:
Show dendrogram
Customize...
Select
Help
Storage...
OK
Cancel


Dendrogram
Ward Linkage; Euclidean Distance
Distance
21,40
14,27
7,13
0,00
Observations

PRO

Cluster Analysis of Observations: Calories; Protein; Fat; Sodium; Fiber; ...

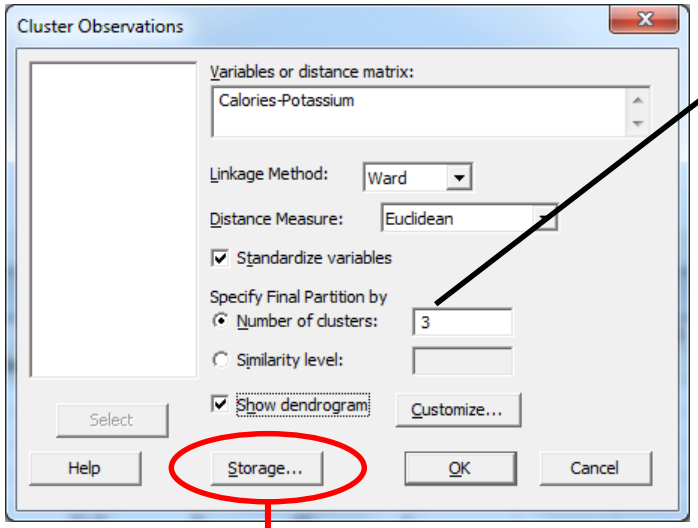
Standardized Variables, Euclidean Distance, Ward Linkage

Amalgamation Steps

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	42	98,179	0,1513	3	4	3	2
2	41	97,172	0,2349	14	16	14	2
3	40	92,574	0,6167	38	39	38	2
34	9	40,303	4,9579	13	29	13	5
35	8	29,521	5,8535	2	14	2	6
36	7	10,412	7,4404	1	3	1	19
37	6	9,418	7,5231	11	18	11	5
38	5	-1,012	8,3893	2	11	2	11
39	4	-27,962	10,6275	2	13	2	16
40	3	-54,883	12,8634	7	41	7	8
41	2	-106,291	17,1329	1	7	1	27
42	1	-157,661	21,3994	1	2	1	43

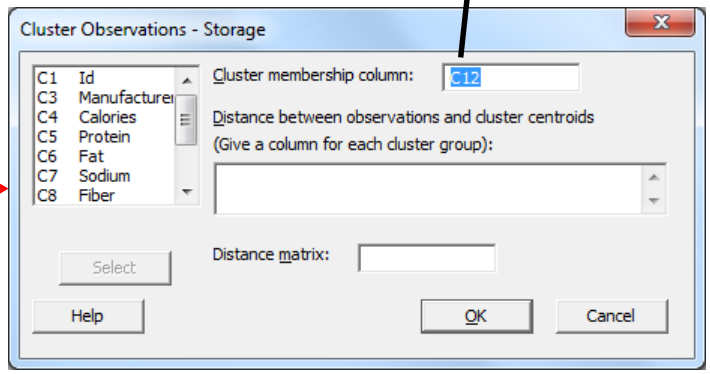
Análise do Amalgamation Steps

Step	Distance Level	Grupos	Δ	Δ%
34	4,9579	9		
35	5,8535	8	0,90	18,1%
36	7,4404	7	1,59	27,1%
37	7,5231	6	0,08	1,1%
38	8,3893	5	0,87	11,5%
39	10,6275	4	2,24	26,7%
40	12,8634	3	2,24	21,0%
41	17,1329	2	4,27	33,2%
42	21,3994	1	4,27	24,9%



Número de grupos a serem formados

Variável que conterá o número do grupo da observação



Final Partition
Number of clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	19	40,694	1,40369	2,17279
Cluster2	16	131,676	2,70878	4,99442
Cluster3	8	44,937	2,14087	3,97271

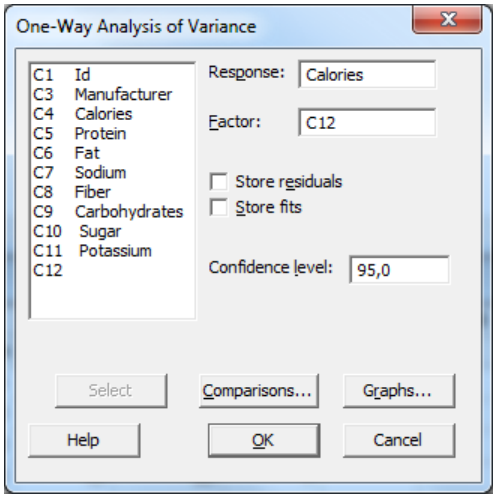
Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
Calories	0,221330	0,143292	-0,81224	-0,0000000
Protein	-0,639142	0,949322	-0,38068	0,0000000
Fat	0,094689	0,340957	-0,90680	0,0000000
Sodium	-0,025804	-0,076882	0,21505	-0,0000000
Fiber	-0,528430	0,964855	-0,67469	0,0000000
Carbohydrates	-0,325896	-0,126155	1,02631	-0,0000000
Sugar	0,713676	-0,243513	-1,20795	-0,0000000
Potassium	-0,480823	0,940012	-0,73807	-0,0000000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0,00000	2,79237	2,79370
Cluster2	2,79237	0,00000	3,47656
Cluster3	2,79370	3,47656	0,00000

Análise do perfil dos grupos

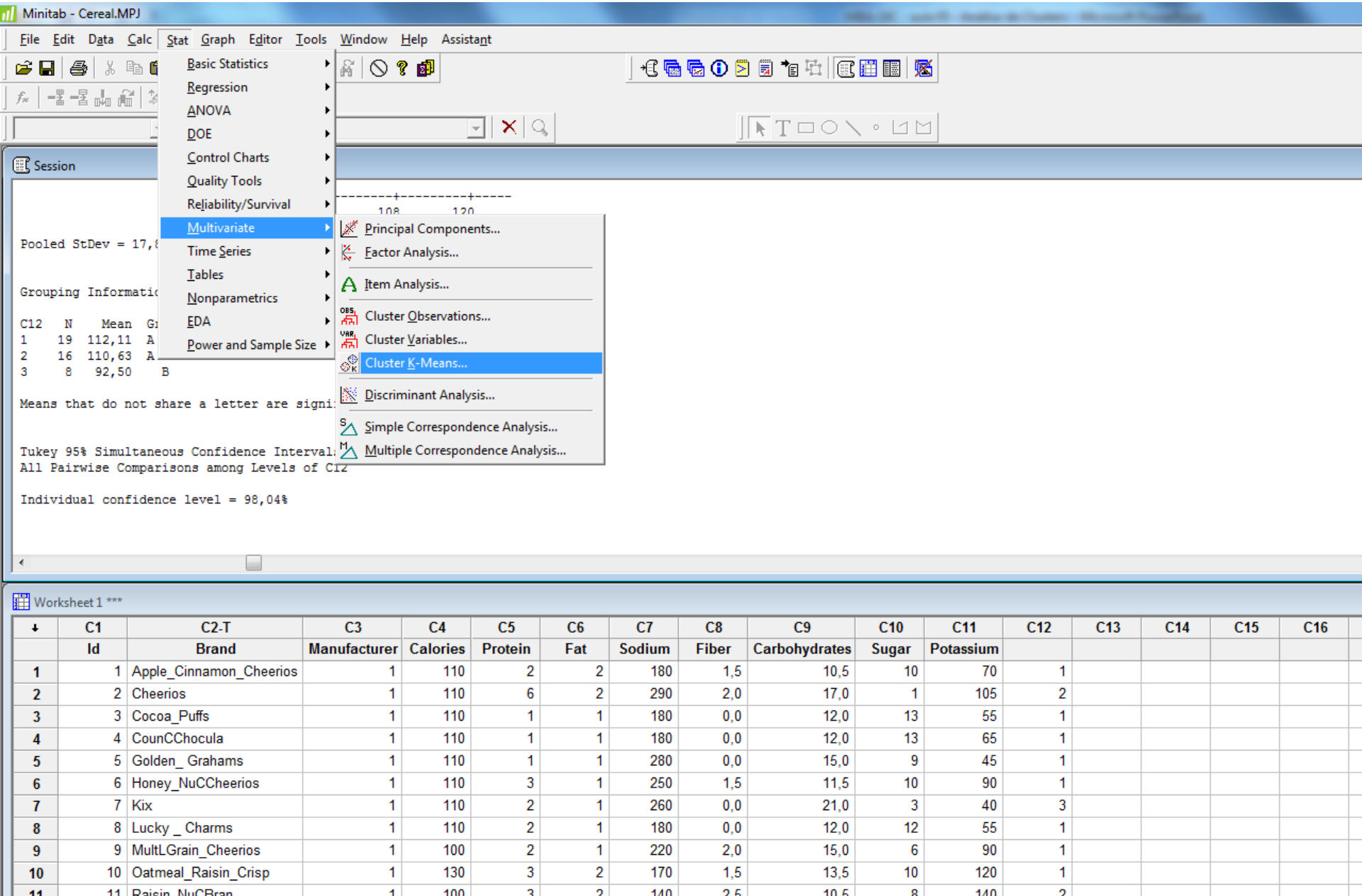


Repetir o procedimento para todas as variáveis:

- Calories
- Protein
- Fat
- Sodium
- Fiber
- Carbohydrates
- Sugar
- Potassium

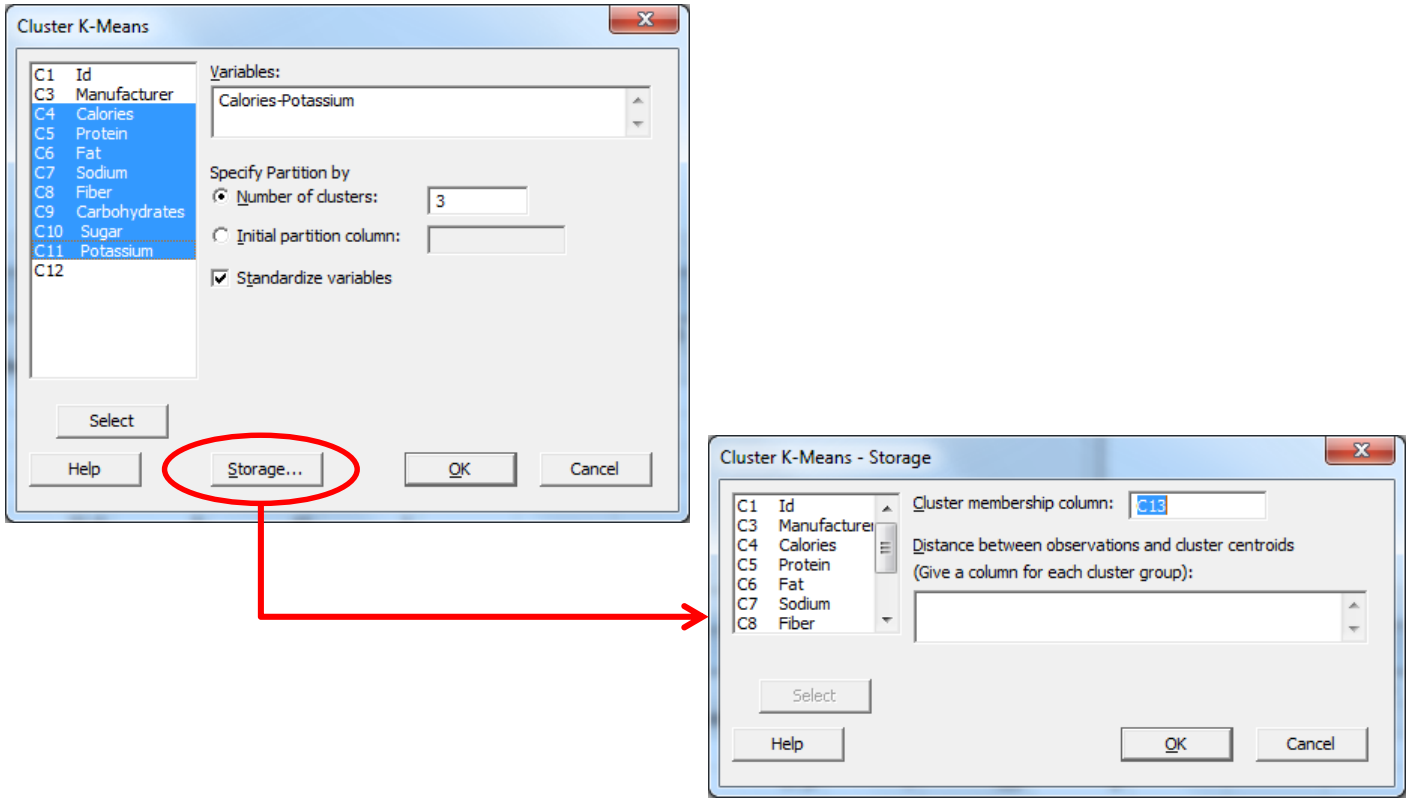
One-way ANOVA: Calories versus C12

```
Source  DF     SS    MS     F      P
C12      2   2352  1176  3,69  0,034
Error   40  12760   319
Total   42  15112

S = 17,86   R-Sq = 15,56%   R-Sq(adj) = 11,34%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean  StDev  ----+---------+---------+---------+-----
1      19  112,11   6,31                      (-----*------)
2      16  110,63  21,75                    (------*-------)
3       8   92,50  26,59  (----------*----------)
                          ----+---------+---------+---------+-----
                             84        96       108       120
Pooled StDev = 17,86

Grouping Information Using Tukey Method
C12   N    Mean  Grouping
1    19  112,11  A
2    16  110,63  A B
3     8   92,50    B

Means that do not share a letter are significantly different.
```



	C1	C2-T	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
	Id	Brand	Manufacturer	Calories	Protein	Fat	Sodium	Fiber	Carbohydrates	Sugar	Potassium					
1	1	Apple_Cinnamon_Cheerios	1	110	2	2	180	1,5	10,5	10	70	1				
2	2	Cheerios	1	110	6	2	290	2,0	17,0	1	105	2				
3	3	Cocoa_Puffs	1	110	1	1	180	0,0	12,0	13	55	1				
4	4	CounCChocula	1	110	1	1	180	0,0	12,0	13	65	1				
5	5	Golden_ Grahams	1	110	1	1	280	0,0	15,0	9	45	1				
6	6	Honey_NuCCheerios	1	110	3	1	250	1,5	11,5	10	90	1				
7	7	Kix	1	110	2	1	260	0,0	21,0	3	40	3				
8	8	Lucky _ Charms	1	110	2	1	180	0,0	12,0	12	55	1				
9	9	MultLGrain_Cheerios	1	100	2	1	220	2,0	15,0	6	90	1				
10	10	Oatmeal_Raisin_Crisp	1	130	3	2	170	1,5	13,5	10	120	1				
11	11	Raisin_NuCBran	1	100	3	2	140	2,5	10,5	8	140	2				

K-Means
Método não hierárquico



K-means Cluster Analysis: Calories; Protein; Fat; Sodium; Fiber; Carbohydrate;

Standardized Variables
Final Partition
Number of clusters: 3

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	13	76,666	2,306	4,276
Cluster2	7	51,010	2,367	5,159
Cluster3	23	107,114	1,962	4,356

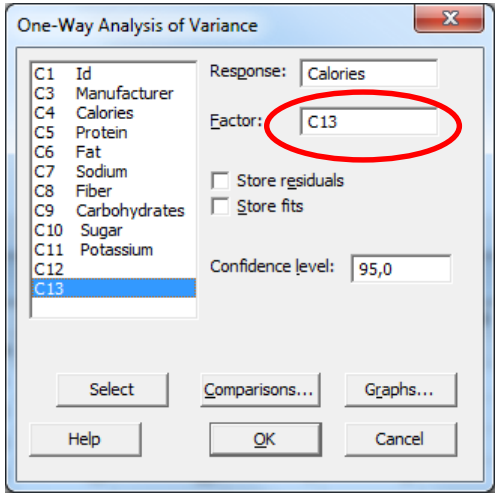
Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
Calories	0,5564	-0,5675	-0,1418	-0,0000
Protein	0,5637	1,2562	-0,7009	0,0000
Fat	0,7009	-0,3275	-0,2965	0,0000
Sodium	-0,2972	0,7335	-0,0553	-0,0000
Fiber	0,7019	0,7942	-0,6384	0,0000
Carbohydrates	-0,3853	0,3761	0,1033	-0,0000
Sugar	0,3076	-1,0466	0,1447	-0,0000
Potassium	0,8582	0,5274	-0,6456	-0,0000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0,0000	2,5287	2,7319
Cluster2	2,5287	0,0000	3,0914
Cluster3	2,7319	3,0914	0,0000

Análise do perfil dos (novos) grupos



Repetir o procedimento para todas as variáveis:

- Calories
- Protein
- Fat
- Sodium
- Fiber
- Carbohydrates
- Sugar
- Potassium

One-way ANOVA: Calories versus C13

```
Source  DF     SS    MS     F      P
C13      2   2426  1213  3,82  0,030
Error   40  12686   317
Total   42  15112

S = 17,81   R-Sq = 16,05%   R-Sq(adj) = 11,85%

                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level   N    Mean  StDev   +---------+---------+---------+---------
1      13  118,46  19,08                      (--------*-------)
2       7   97,14  13,80   (----------*----------)
3      23  105,22  18,06              (------*-----)
                          +---------+---------+---------+---------
                         84        96       108       120

Pooled StDev = 17,81

Grouping Information Using Tukey Method
C13   N    Mean  Grouping
1    13  118,46  A
3    23  105,22  A B
2     7   97,14    B

Means that do not share a letter are significantly different.
```

Quantidade de grupos formados

- Dendograma
  - É uma espécie de gráfico ou árvore, onde os indivíduos são classificados através de sua distância (entre indivíduos).
  - Facilita a visualização dos clusters extraídos.
- Roteiro de aglomeração (Amalgamation Steps)
  - Observar a variação da medida final de heterogenidade

Análise de Variância

- Compara se o comportamento de uma variável muda entre os grupos. Caso não haja mudança de comportamento, ela não deve ser utilizada na explicação/caracterizaçao dos grupos.

Significado dos grupos

- Analisar as diferenças o comportamento das variáveis (utilizadas na análise de clusters) entre os grupos

Validação dos resultados

- Divisão da amostra em dois grupos
- Repetição do procedimento
- Comparação qualitativa dos resultados observados

- Consistência teórica e conceitual do resultado observado