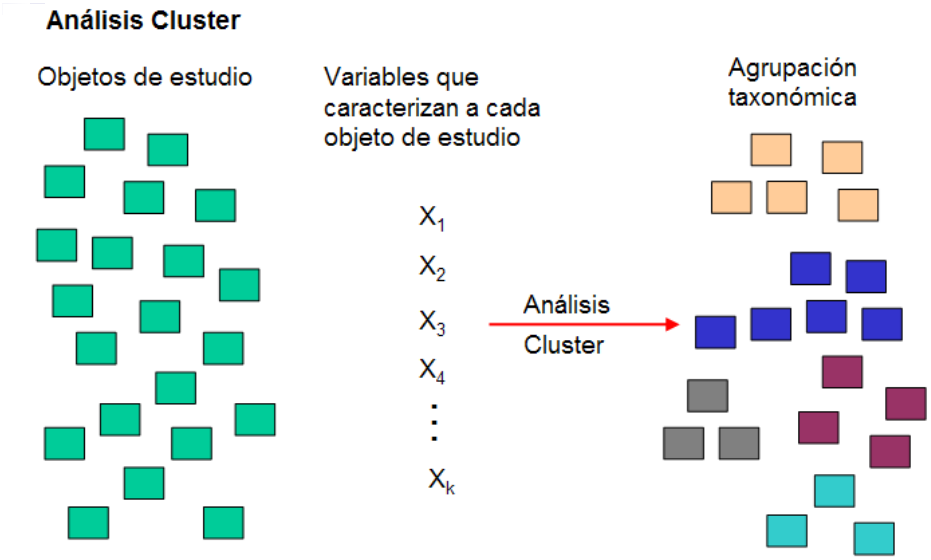
Universidad Nacional Agraria La Molina
Dpto. de Estadística e Informática
Curso: Técnicas Multivariadas I

ANÁLISIS CLUSTER

I. Introducción[1]

El análisis cluster es una técnica diseñada para clasificar tantas observaciones en grupos de tal forma que:

- Cada grupo (conglomerado o cluster) sea homogéneo respecto a las variables utilizadas para caracterizarlos; es decir, que cada observación contenida en él sea parecida a todas las que estén incluidas en ese grupo.
- Que los grupos sean lo más distintos posible unos de otros respecto a las variables consideradas.



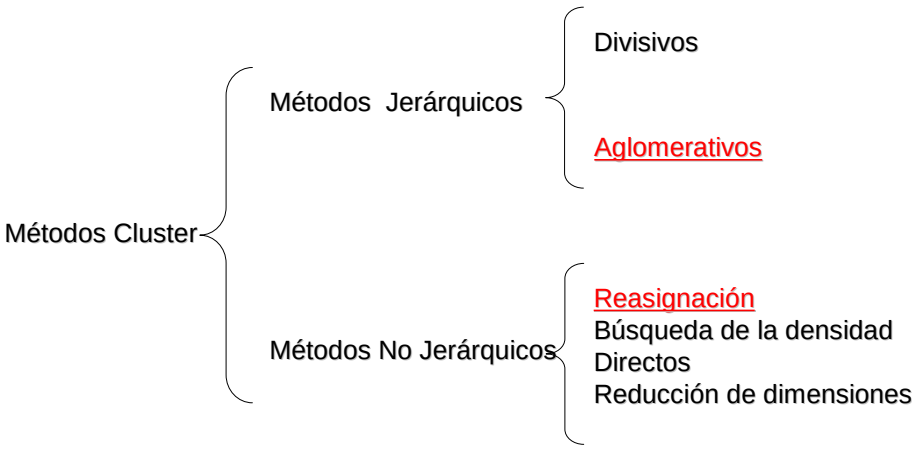
Agrupación de objetos basándose en las características que poseen

Ejemplo:

- El responsable de marketing tiene una BDD con las características sociodemográficas de sus clientes: edad, nivel educativo, nivel de ingresos, estado civil, ocupación, número de hijos, etc.
- Se plantea si pudiera dividir a sus clientes en subgrupos que tuvieran características sociodemográficas similares entre sí, pero que fueran lo más diferentes posible unos subgrupos de otros.
- Si fuera posible, se podría diseñar campañas de publicidad distintas para cada grupo, con creatividades diferentes o utilizando diarios, revistas o cadenas de televisión distintas según el grupo al que fuera dirigida la campaña

[1] Basado en Uriel, Ezequiel & Aldas, Joaquín. "Análisis Multivariante Aplicado. Aplicaciones al marketing, investigación de mercados, economía, dirección de empresas y turismo".

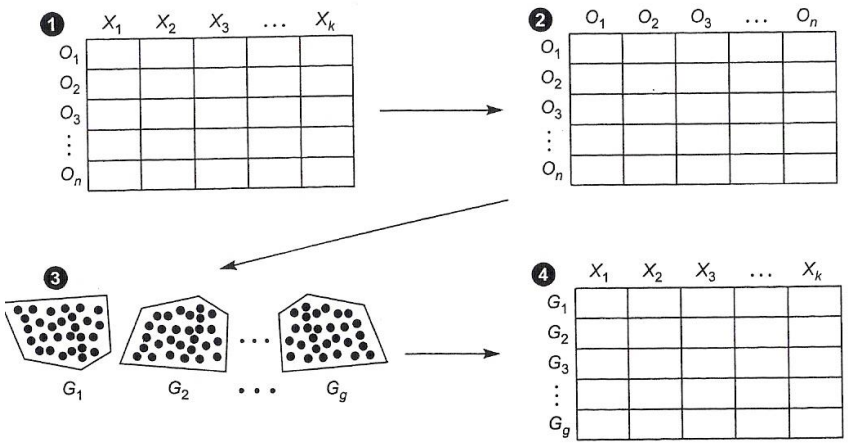
II. Clasificación de los métodos de Análisis Cluster[2]



III. Análisis Cluster Jerárquico

Procedimiento

1. Se tienen n observaciones (individuos, empresas, etc.) de los que se tiene información sobre p variables (edad, estado civil, número de hijos, etc)
2. Se establece un indicador que nos diga en qué medida cada par de observaciones se parece entre sí. A este medida se le denomina distancia o similaridad.
3. Se crean grupos, de forma que cada grupo contenga aquellas observaciones que más se parezcan entre sí. Hay dos tipos de AC: jerárquico y no jerárquico. A su vez, en cada tipo se pueden utilizar distintos métodos de agrupación y conglomeración.
4. Se debe describir los grupos que se ha obtenido y compararlos unos con los otros. Para ello bastará con ver qué valores promedio toman las p variables utilizadas en el AC en cada uno de los g grupos obtenidos ($g \leq n$)



[2] Basado en Pérez, César. "Técnicas de Análisis Multivariante de Datos: aplicaciones con SPSS". Pearson Prentice Hall. 2004. España.

Estandarización de los datos

- Las medidas de similaridad son muy sensibles a las unidades que estén medidas dichas variables.
- Para evitar esta influencia no deseable de una variable debida exclusivamente a la unidad en que viene medida, es necesario corregir el efecto de los datos recurriendo a un proceso de estandarización.

Medidas de Distancia

- **Distancia euclidiana**: es la raíz cuadrada de la suma de las diferencias al cuadrado entre los dos elementos en la variable o variables consideradas

$$D(X,Y) = \sqrt{\sum (X_i - Y_i)^2}$$

- **Distancia euclidiana al cuadrado**

$$D^2(X,Y) = \sum (X_i - Y_i)^2$$

- **Distancia métrica de Chebychev**: es la referencia máxima en valores absolutos entre los valores de los elementos

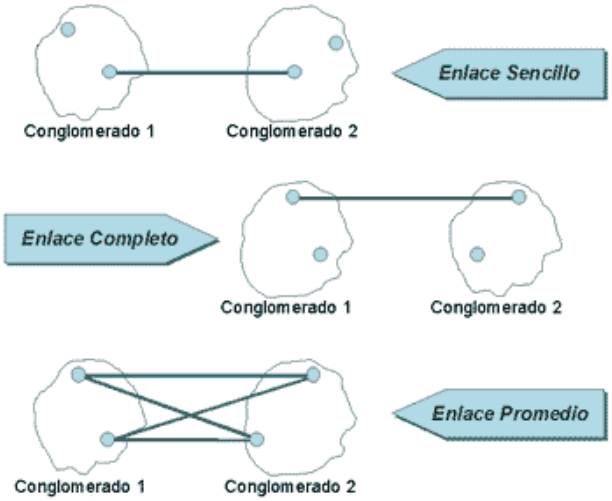
$$D(X,Y) = Max_i \left| X_i - Y_i \right|$$

Formación de los grupos

- Una vez que mediante la matriz de distancias, se sabe que observaciones están más próximas entre sí, y más distantes de otras, es necesario formar los grupos.
- Ello implica tomar dos decisiones:
  - Selección del algoritmo de agrupación que se elige
  - Determinación de un número de grupos o clusters.

Algoritmos de agrupamiento

- Método del vecino más cercano (vinculación simple)
- Método del vecino más lejano (vinculación completa)
- Método de la vinculación intergrupos (vinculación promedio)
- Método del centroide (vinculación de centroides)
- Método de Ward



Ejemplo de Aplicación Nº 1[3]

	X_1	X_2
A	1	2
B	2	1
C	4	1
D	5	4
E	3	5
F	3	3

	A	**B**	**C**	**D**	**E**	**F**
A	0	1.41	3.16	4.47	3.61	2.24
B	1.41	0	2	4.24	4.12	2.24
C	3.16	2	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
F	2.24	2.24	2.24	2.24	2	0

[3] Basado en Pérez, César. "Técnicas de Análisis Multivariante de Datos: aplicaciones con SPSS". Pearson Prentice Hall. 2004. España.

	(A,B)	C	D	E	F
(A,B)	0	2	4.24	3.61	2.24
C	2	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
F	2.24	2.24	2.24	2	0

	(A,B)	C	D	(E,F)
(A,B)	0	2	4.24	2.24
C	2	0	3.16	2.24
D	4.24	3.16	0	2.24
(E,F)	2.24	2.24	2.24	0

	(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

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

Salida con SPSS

Matriz de distancias

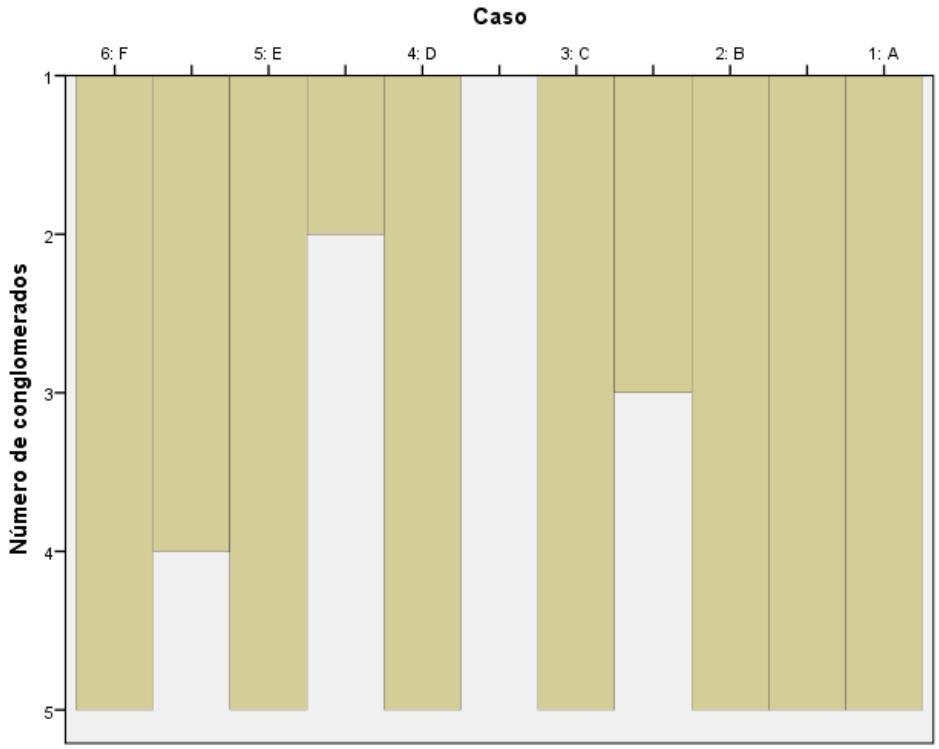
Caso	distancia euclídea					
	1:A	2:B	3:C	4:D	5:E	6:F
1:A	,000	1,414	3,162	4,472	3,606	2,236
2:B	1,414	,000	2,000	4,243	4,123	2,236
3:C	3,162	2,000	,000	3,162	4,123	2,236
4:D	4,472	4,243	3,162	,000	2,236	2,236
5:E	3,606	4,123	4,123	2,236	,000	2,000
6:F	2,236	2,236	2,236	2,236	2,000	,000

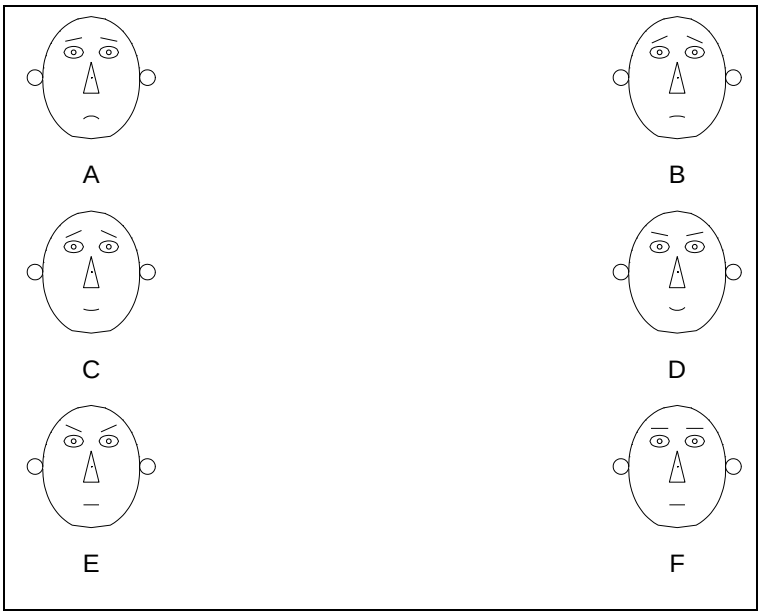
Esta es una matriz de disimilaridades

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	1	2	1,414	0	0	3
2	5	6	2,000	0	0	4
3	1	3	2,000	1	0	5
4	4	5	2,236	0	2	5
5	1	4	2,236	3	4	0

Diagrama de témpanos vertical





IV. Selección del número de clusters[4]

El SPSS sólo ofrece el dendograma como herramienta de apoyo.

- Debe detenerse el proceso de fusión cuando los grupos que se han de unir están a una distancia significativamente mayor de los que previamente se han fusionado.
- Se debe realizar el cálculo de las tasas de variación entre los coeficientes de aglomeración obtenidos entre etapas sucesivas. Cuando una tasa de variación sea drásticamente superior a la anterior, será el momento de detener las fusiones.

Otros Indicadores

- Raíz cuadrada de la media de las desviaciones típicas del nuevo conglomerado (RMSSTD)
- R^2 semiparcial (SPR)
- R cuadrado (RS)
- Distancia entre los conglomerados (DC)

Estadístico	Concepto Medido	Comentarios
RMSSTD	Homogeneidad del nuevo conglomerado	El valor debe ser pequeño
SPR	Homogeneidad de los conglomerados fusionados	El valor debe ser pequeño
RS	Heterogeneidad entre conglomerados	El valor debe ser grande
CD	Homogeneidad de los conglomerados fusionados	El valor debe ser pequeño

[4] Basado en Programme PRESTA. "Métodos Estadísticos Aplicados a la Investigación en Ciencias Sociales y Huamanas". 1997

Ejemplo de Aplicación Nº 2[5]

A un grupo de 21 personas se le medirá una serie de atributos de tipo métrico, y conforme a estos atributos se van a clasificar a estas personas en grupos o categorías de tal forma que dentro de cada grupo las unidades muestrales sean lo más homogénea posible, y entre los grupos estas unidades, comparativamente, sean lo más heterogénea posibles.

La información que se recolectó de estas 21 personas (usando una escala de Likert del 1 al 7, donde 1 es desacuerdo y 7 de acuerdo), fue su grado de conformidad a las siguientes afirmaciones:

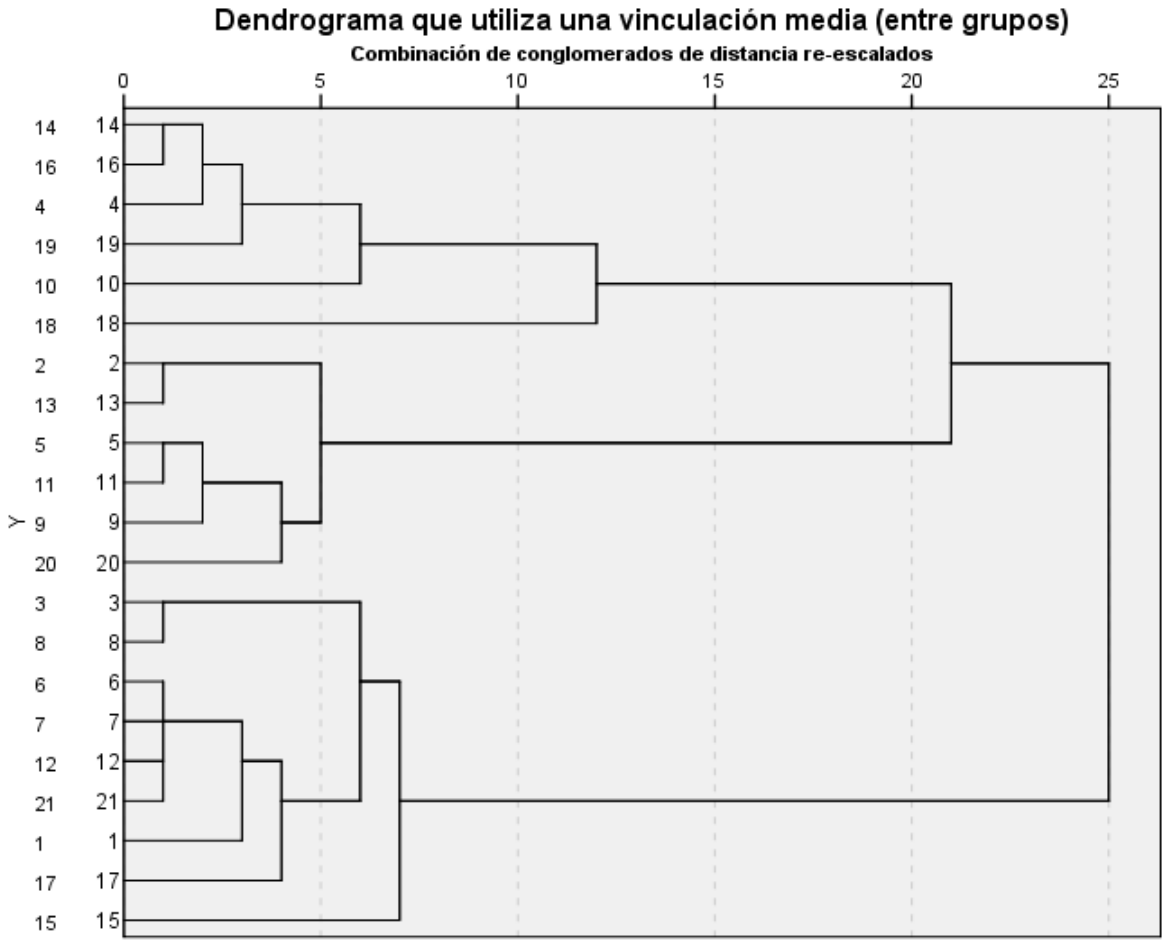
- Salir de compras es divertido
- Salir de compras afecta el presupuesto
- Al salir de compras aprovecho de comer fuera
- Al salir a comprar trato de hacer las mejores
- No me importa salir de compras
- Al salir de compra voy a ahorrar si comparo precios

Salida con SPSS

Historial de conglomeración

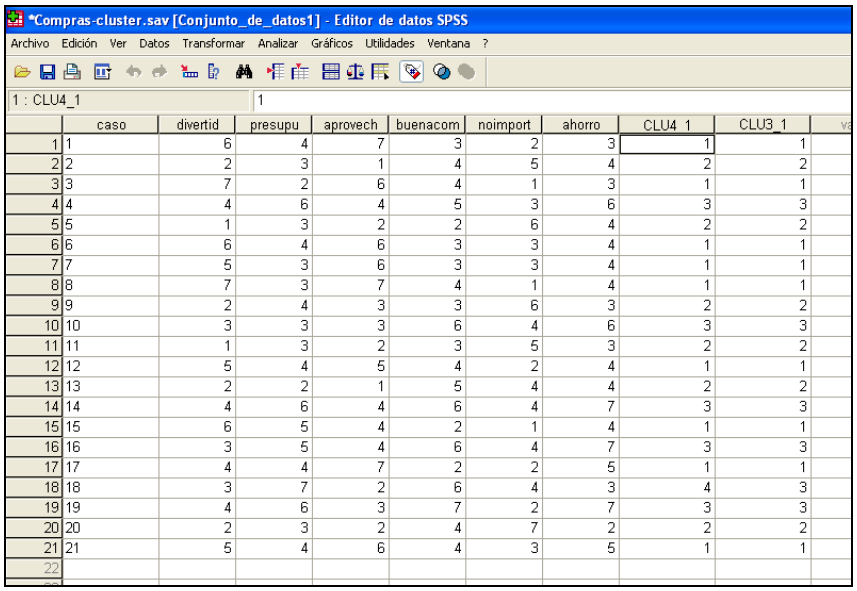
Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 2	Conglomerado 1	
1	14	16	2.000	0	0	8
2	6	7	2.000	0	0	7
3	12	21	3.000	0	0	7
4	2	13	3.000	0	0	14
5	5	11	3.000	0	0	9
6	3	8	3.000	0	0	15
7	6	12	3.500	2	3	10
8	4	14	4.000	0	1	11
9	5	9	4.500	5	0	12
10	1	6	5.750	0	7	13
11	4	19	7.000	8	0	16
12	5	20	7.333	9	0	14
13	1	17	8.000	10	0	15
14	2	5	10.750	4	12	19
15	1	3	11.667	13	6	17
16	4	10	11.750	11	0	18
17	1	15	14.125	15	0	20
18	4	18	22.600	16	0	19
19	2	4	37.944	14	18	20
20	1	2	46.389	17	19	0

[5] Basado en Gondar, Emilio. Data Mining Institute www.estadistico.com. 2004



Análisis de la Tabla de Frecuencias

	Average Linkage (Between Groups)	Average Linkage (Between Groups)	Average Linkage (Between Groups)	Average Linkage (Between Groups)	Average Linkage (Between Groups)	Average Linkage (Between Groups)	Average Linkage (Between Groups)
	Count	Count	Count	Count	Count	Count	Count
1	6	6	8	8	9	9	9
2	2	6	6	6	6	6	12
3	2	2	4	5	5	6	
4	4	4	1	1	1		
5	4	1	1	1			
6	1	1	1				
7	1	1					
8	1						


*Compras-cluster.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

1 : CLU4_1 1

	caso	divertid	presupu	aprovech	buenacom	noimport	ahorro	CLU4_1	CLU3_1
1	1	6	4	7	3	2	3	1	1
2	2	2	3	1	4	5	4	2	2
3	3	7	2	6	4	1	3	1	1
4	4	4	6	4	5	3	6	3	3
5	5	1	3	2	2	6	4	2	2
6	6	6	4	6	3	3	4	1	1
7	7	5	3	6	3	3	4	1	1
8	8	7	3	7	4	1	4	1	1
9	9	2	4	3	3	6	3	2	2
10	10	3	3	3	6	4	6	3	3
11	11	1	3	2	3	5	3	2	2
12	12	5	4	5	4	2	4	1	1
13	13	2	2	1	5	4	4	2	2
14	14	4	6	4	6	4	7	3	3
15	15	6	5	4	2	1	4	1	1
16	16	3	5	4	6	4	7	3	3
17	17	4	4	7	2	2	5	1	1
18	18	3	7	2	6	4	3	4	3
19	19	4	6	3	7	2	7	3	3
20	20	2	3	2	4	7	2	2	2
21	21	5	4	6	4	3	5	1	1
22									

Estadística tipo F de Beale[6]

Otro procedimiento fue sugerido por Beale. Suponga que se tienen dos agrupaciones posibles, en donde la primera consta de c_1 agrupamientos y la segunda de c_2, y suponga que $c_2 < c_1$. Es decir, la segunda agrupación contiene menos agrupamientos que la primera. Sean W_1 y W_2 las sumas de cuadrados correspondientes de las distancias dentro de cada agrupamiento, calculadas desde las medias de éstos últimos. Esto es, suponga que se tienen n_r puntos datos en el r-ésimo agrupamiento, r=1,2,……., c_1. Si x_{rq} representa el q-ésimo vector de observaciones en el r-ésimo agrupamiento, entonces

$$W_1 = \sum_{r=1}^{c_1} \sum_{q=1}^{n_r} (x_{rq} - \bar{x}_r)'((x_{rq} - \bar{x}_r)$$

De manera semejante, se puede definir W_2. Si W_1 y W_2 son casi del mismo tamaño, entonces la agrupación que consta del menor número de agrupamientos es tan buena como la que consta del número más grande y, por sencillez, se seleccionaría aquella que consta del menor número. Sin embargo, si W_1 es mucho menor que W_2, entonces se podría decir que la primera agrupación es una mejora sobre la segunda y se seleccionaría la agrupación con el mayor número de agrupamientos.

Para determinar si la primera agrupación es mejor que la segunda, Beale sugirió calcular una pseudoestadística tipo F por medio de

[6] Basado en Dallas, Johnson. “Métodos multivariados aplicados al análisis de Datos”. 2000.

El estadístico F de Beale se calcula con la siguiente expresión:

$$F^* = \frac{(W_2 - W_1)}{W_1} \frac{(N - c_1)k_1}{(N - c_2)k_2 - (N - c_1)k_1}, \quad \text{donde } k_1 = c_1^{-2/p} \text{ y } k_2 = c_2^{-2/p}$$

Si F* es mayor que un punto crítico F, con $k_2(N-c_2)-k_1(N-c_1)$ grados de libertad para el numerador y $k_1(N-c_1)$ para el denominador, entonces se elegiría la primera agrupación (aquellas con más agrupamientos) sobre la segunda (aquella con menos agrupamientos)

Ejemplo

En este se va a probar si es significativa la diferencia entre la opción de 3 clusters con la de 4 clusters

```
Cluster Centroids
                                                Grand
Variable   Cluster1   Cluster2   Cluster3   centroid
divertid    5.66667    1.66667    3.50000    3.90476
presupu     3.66667    3.00000    5.50000    4.00000
combino     6.00000    1.83333    3.33333    4.04762
bestbuy     3.22222    3.50000    6.00000    4.09524
noimport    2.00000    5.50000    3.50000    3.42857
ahorro      4.00000    3.33333    6.00000    4.38095
```

Para el caso de 3 clusters, W2 tiene c2 = 3, el valor de W2 se obtiene:

caso	divertid	presupu	combino	bestbuy	noimport	ahorro	cl_3	div_pro	pre_pro	com_pro	bes_pro	noim_pro	aho_pro	W_2
1	6	4	7	3	2	3	1	5.667	3.667	6.000	3.222	2.000	4.000	2.271
2	2	3	1	4	5	4	2	1.667	3.000	1.833	3.500	5.500	3.300	1.795
3	7	2	6	4	1	3	1	5.667	3.667	6.000	3.222	2.000	4.000	7.161
4	4	6	4	5	3	6	3	3.500	5.500	3.333	6.000	3.500	6.000	2.195
5	1	3	2	2	6	4	2	1.667	3.000	1.833	3.500	5.500	3.300	3.463
6	6	4	6	3	3	4	1	5.667	3.667	6.000	3.222	2.000	4.000	1.271
7	5	3	6	3	3	4	1	5.667	3.667	6.000	3.222	2.000	4.000	1.939
8	7	3	7	4	1	4	1	5.667	3.667	6.000	3.222	2.000	4.000	4.827
9	2	4	3	3	6	3	2	1.667	3.000	1.833	3.500	5.500	3.300	3.063
10	3	3	3	6	4	6	3	3.500	5.500	3.333	6.000	3.500	6.000	6.861
11	1	3	2	3	5	3	2	1.667	3.000	1.833	3.500	5.500	3.300	1.063
12	5	4	5	4	2	4	1	5.667	3.667	6.000	3.222	2.000	4.000	2.161
13	2	2	1	5	4	4	2	1.667	3.000	1.833	3.500	5.500	3.300	6.795
14	4	6	4	6	4	7	3	3.500	5.500	3.333	6.000	3.500	6.000	2.195
15	6	5	4	2	1	4	1	5.667	3.667	6.000	3.222	2.000	4.000	8.381
16	3	5	4	6	4	7	3	3.500	5.500	3.333	6.000	3.500	6.000	2.195
17	4	4	7	2	2	5	1	5.667	3.667	6.000	3.222	2.000	4.000	6.383
18	3	7	2	6	4	3	3	3.500	5.500	3.333	6.000	3.500	6.000	13.527
19	4	6	3	7	2	7	3	3.500	5.500	3.333	6.000	3.500	6.000	4.861
20	2	3	2	4	7	2	2	1.667	3.000	1.833	3.500	5.500	3.300	4.329
21	5	4	6	4	3	5	1	5.667	3.667	6.000	3.222	2.000	4.000	3.161
													W_2	89.896

Para el caso de 4 clusters, W_1 tiene $c_1 = 4$, el valor de W_1 se obtiene previamente ejecutando el Análisis Clusters con 4 clusters para obtener los promedios por variable

```
Cluster Centroids

                                                  Grand
Variable   Cluster1   Cluster2   Cluster3   Cluster4   centroid
divertid    5.66667    1.66667        3.6          3    3.90476
presupu     3.66667    3.00000        5.2          7    4.00000
combino     6.00000    1.83333        3.6          2    4.04762
bestbuy     3.22222    3.50000        6.0          6    4.09524
noimport    2.00000    5.50000        3.4          4    3.42857
ahorro      4.00000    3.33333        6.6          3    4.38095
```

Con los centroides del cluster se procede a calcular el valor de W_1

caso	divertid	presupu	combino	bestbuy	noimport	ahorro	cl_4	div_pro	pre_pro	com_pro	bes_pro	noim_pro	aho_pro	W_1
1	6	4	7	3	2	3	1	5.667	3.667	6.000	3.222	2.000	4.000	2.271
2	2	3	1	4	5	4	2	1.667	3.000	1.833	3.500	5.500	3.333	1.750
3	7	2	6	4	1	3	1	5.667	3.667	6.000	3.222	2.000	4.000	7.161
4	4	6	4	5	3	6	3	3.600	5.200	3.600	6.000	3.400	6.600	2.480
5	1	3	2	2	6	4	2	1.667	3.000	1.833	3.500	5.500	3.333	3.417
6	6	4	6	3	3	4	1	5.667	3.667	6.000	3.222	2.000	4.000	1.271
7	5	3	6	3	3	4	1	5.667	3.667	6.000	3.222	2.000	4.000	1.938
8	7	3	7	4	1	4	1	5.667	3.667	6.000	3.222	2.000	4.000	4.827
9	2	4	3	3	6	3	2	1.667	3.000	1.833	3.500	5.500	3.333	3.084
10	3	3	3	6	4	6	3	3.600	5.200	3.600	6.000	3.400	6.600	6.280
11	1	3	2	3	5	3	2	1.667	3.000	1.833	3.500	5.500	3.333	1.083
12	5	4	5	4	2	4	1	5.667	3.667	6.000	3.222	2.000	4.000	2.161
13	2	2	1	5	4	4	2	1.667	3.000	1.833	3.500	5.500	3.333	6.750
14	4	6	4	6	4	7	3	3.600	5.200	3.600	6.000	3.400	6.600	1.480
15	6	5	4	2	1	4	1	5.667	3.667	6.000	3.222	2.000	4.000	8.382
16	3	5	4	6	4	7	3	3.600	5.200	3.600	6.000	3.400	6.600	1.080
17	4	4	7	2	2	5	1	5.667	3.667	6.000	3.222	2.000	4.000	6.382
18	3	7	2	6	4	3	4	3.000	7.000	2.000	6.000	4.000	3.000	0.000
19	4	6	3	7	2	7	3	3.600	5.200	3.600	6.000	3.400	6.600	4.280
20	2	3	2	4	7	2	2	1.667	3.000	1.833	3.500	5.500	3.333	4.416
21	5	4	6	4	3	5	1	5.667	3.667	6.000	3.222	2.000	4.000	3.161
													W_1	73.656

El estadístico F de Beale se calcula con la siguiente expresión:

$$F^* = \frac{(W_2 - W_1)}{W_1} \frac{(N - c_1)k_1}{(N - c_2)k_2 - (N - c_1)k_1}, \quad \text{donde } k_1 = c_1^{-2/p} \text{ y } k_2 = c_2^{-2/p}$$

Para nuestro caso el valor de F* es:

$$F^* = \frac{(89.896 - 73.566)}{73.566} \frac{(21-4)4^{-2/6}}{(21-3)3^{-2/6} - (21-4)4^{-2/6}} = 1.34$$

Esto se compararía con un punto crítico de la distribución F, con los grados de libertad del numerador iguales a

$$(N - c_2)k_2 - (N - c_1)k_1 = (21-3)3^{-2/6} - (21-4)4^{-2/6} = 1.77$$

y los grados de libertad del denominador iguales a

$$(N - c_1)k_1 = (21-4)4^{-2/6} = 10.71$$

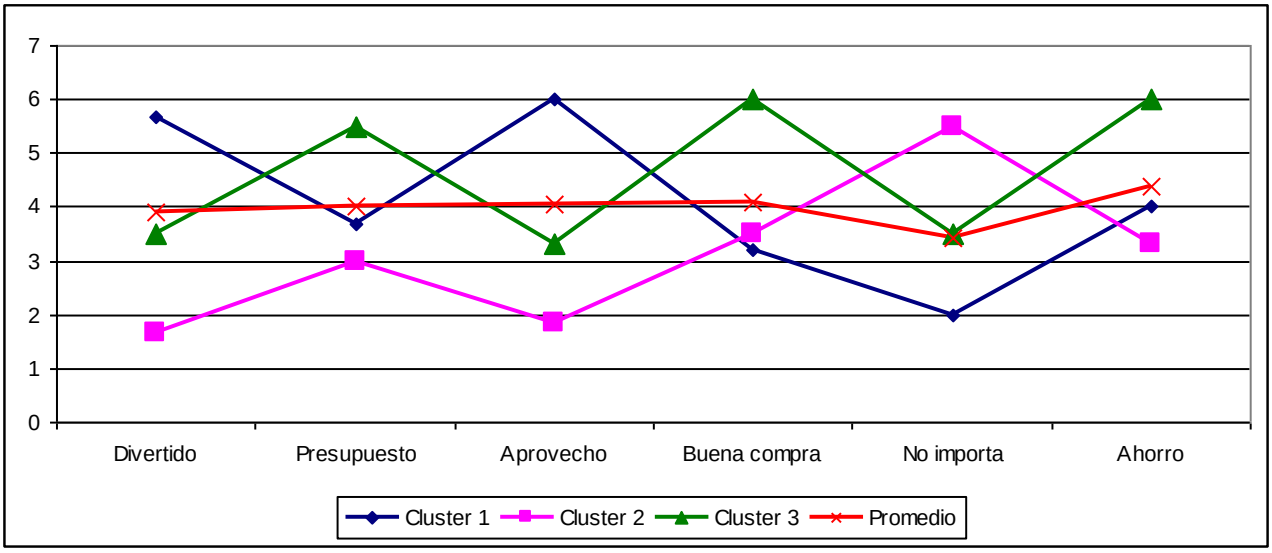
Se concluirá que no hay diferencias significativas entre las sumas de cuadrados de ambos grupos, por lo que la agrupación del menor tamaño (3) es tan buena como la que consta del número más grande (4) y, por sencillez, se selecciona aquella que consta del menor número.

V. Caracterización de los Clusters

Caracterización de los clusters usando variables activas

Consiste en analizar los centros de gravedad de cada grupo (promedios) una vez finalizada la clasificación de todos los individuos en base a las variables utilizadas para efectuar la participación, que suelen llamarse "variables activas".

		Divertido	Presupuesto	Aprovecho	Buena compra	No importa	Ahorro
		Media	Media	Media	Media	Media	Media
Average Linkage (Between Groups)	1	5.667	3.667	6.000	3.222	2.000	4.000
	2	1.667	3.000	1.833	3.500	5.500	3.333
	3	3.500	5.500	3.333	6.000	3.500	6.000
	Total	3.905	4.000	4.048	4.095	3.429	4.381



1. Los casos del cluster 1 tienen valores altos en las variables **divertido, aprovecho**, medios en las variables **presupuesto**, **buena compra**, **ahorro** y bajo en **no importa.**
2. Los casos del cluster 2 tienen valores altos sólo en la variables **no importa**, medios en las variables **presupuesto**, **buena compra**, **ahorro** y bajo en **divertido**, **aprovecho**.
3. Los casos del cluster 3 tienen valores altos en las variables **presupuesto**, **buena compra**, **ahorro**, medios en las variables **divertido**, **aprovecho**, **no importa** y no tiene valores bajos.

Caracterización de los clusters usando variables pasivas

Es el análisis del perfil de cada grupo en base a variables pasivas, es decir en variables que no han intervenido a la hora de efectuar la partición pero pueden aportar información interesante para definir mejor los grupos. Por ejemplo, variables de tipo sociodemográfico (sexo, edad, nivel de estudios, hábitat, etc.) o socioeconómico (clase social, profesión, nivel de renta, etc.)

ANÁLISIS CLUSTER JERÁRQUICO CON R

```
##################################
# ANÁLISIS CLUSTER JERÁRQUICO #
#          EJEMPLO COMPRAS          #
##################################
```

Librerías

```
library(ag)
library(cluster)
library(aplpack)
library(tcltk)
library(fpc)
```

Ingreso de datos

```
datosc=read.csv("compras-cluster.csv")
datosc=datosc[,-1]
str(datosc)
datosc
```

Cálculo de la matriz de distancia euclidiana al cuadrado

```
d <- dist(datosc, method = "euclidean")
d
d=d^2
d
```

Método jerárquico usando el método de enlace promedio

```
fit <- hclust(d, method="average")
str(fit)
```

Proceso de agrupamiento indicando los individuos

```
fit$merge
```

```
     [,1] [,2]
 [1,]   -6   -7
 [2,]  -14  -16
 [3,]   -2  -13
 [4,]   -3   -8
 [5,]   -5  -11
 [6,]  -21    1
 [7,]  -12    6
 [8,]   -4    2
 [9,]   -9    5
[10,]   -1    7
[11,]  -19    8
[12,]  -20    9
[13,]  -17   10
[14,]    3   12
[15,]    4   13
[16,]  -10   11
[17,]  -15   15
[18,]  -18   16
[19,]   14   18
[20,]   17   19
```

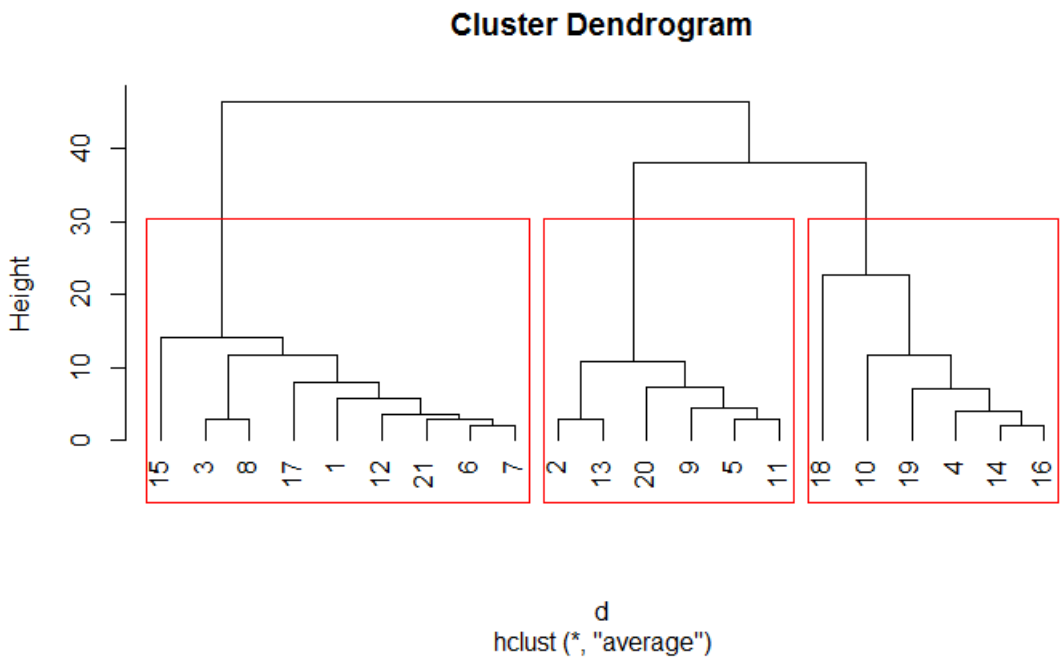
Proceso de agrupamiento indicando los distancias

```
fit$height
```

```
[1]  2.000000  2.000000  3.000000  3.000000  3.000000  3.000000  3.666667
4.000000  4.500000
[10]  5.750000  7.000000  7.333333  8.000000 10.750000 11.666667 11.750000
14.125000 22.600000
[19] 37.944444 46.388889
```

Dendograma

```
plot(fit)
plot(fit,hang=-1)
plclust(fit,hang=-1) # opción alternativa en versiones anteriores
rect.hclust(fit, k=3, border="red")
```



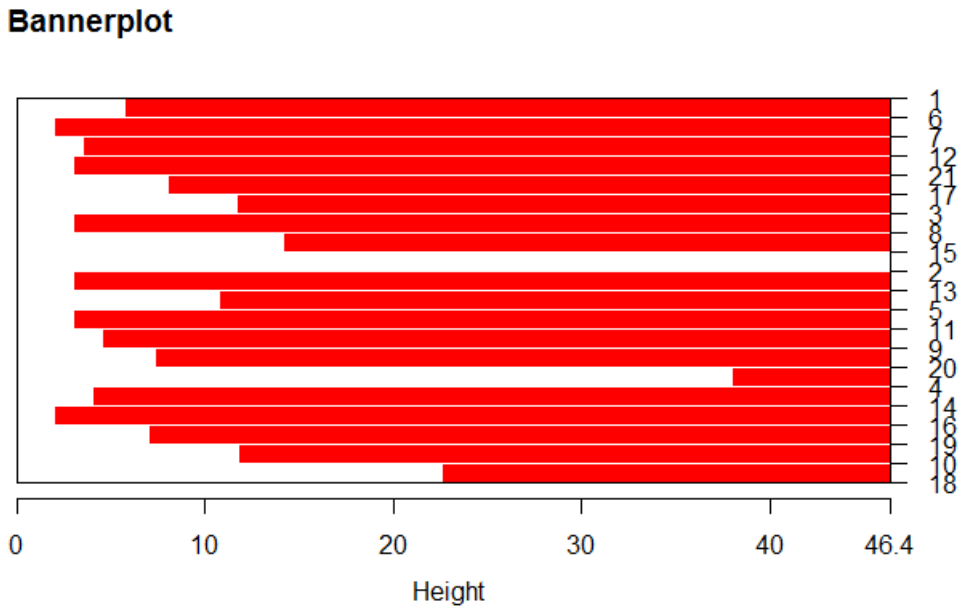
Genera clusters a una distancia de 25

```
cut<-cutree(fit,h=25)
cut
```

```
[1] 1 2 1 3 2 1 1 1 2 3 2 1 2 3 1 3 1 3 3 2 1
```

Gráfica de Témpanos - BannerPlot

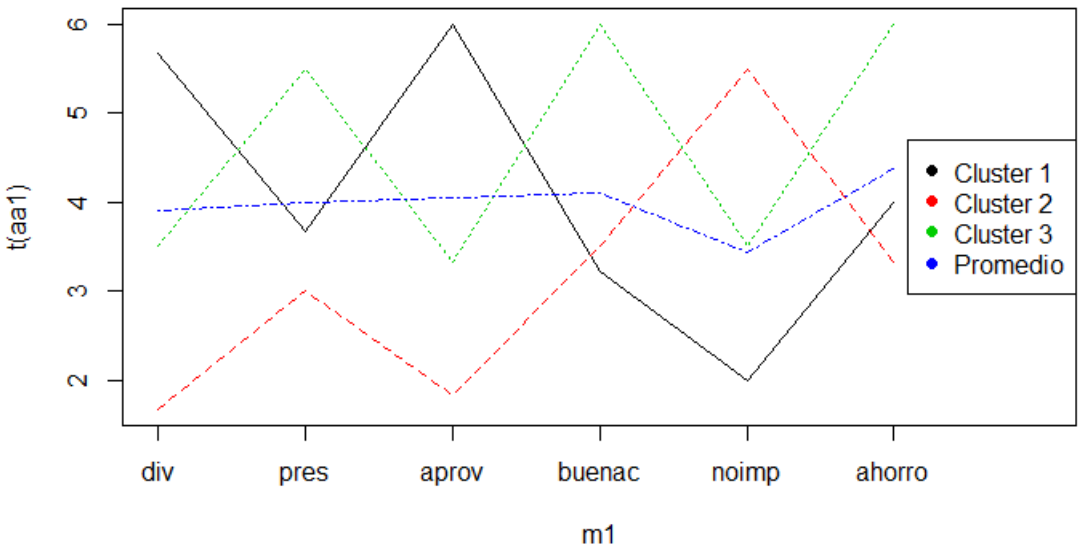
```
bannerplot(agnes(d,metric="euclidean",method="average"), main = "Bannerplot")
```



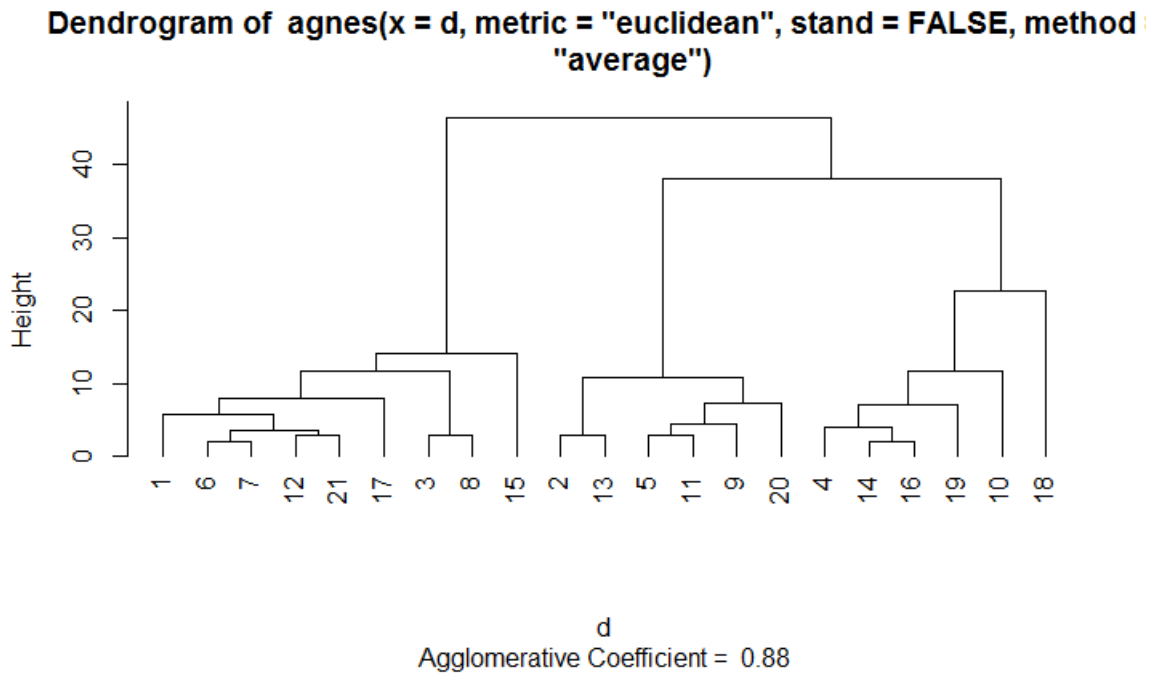
```
# Genera k=3 clusters
cut2<-cutree(fit,k=3)
cut2
```

```
[1] 1 2 1 3 2 1 1 1 2 3 2 1 2 3 1 3 1 3 3 2 1
```

```
# Junta el archivo de datos con la columna de cluster
datosf=cbind(datosc,cut2)
datosf
str(datosf)


# Diagrama de Cajas de variable Divertid según Cluster
boxplot(datosc$divertid~cut,col="gray" )


# Diagrama de líneas de promedios por cluster
pcl1=subset(datosf[,-7],cut2==1)
pcl1
pcl2=subset(datosf[,-7],cut2==2)
pcl2
pcl3=subset(datosf[,-7],cut2==3)
pcl3
ppcl1=apply(pcl1,2,mean)
ppcl1
ppcl2=apply(pcl2,2,mean)
ppcl2
ppcl3=apply(pcl3,2,mean)
ppcl3
pgen=apply(datosf[,-7],2,mean)
pgen

aa1=rbind(ppcl1,ppcl2,ppcl3,pgen)
aa1
m1<-matrix(rep(c(1,2,3,4,5,6),4),6,4)
m1
matplot(m1,t(aa1),type="l")
matplot(xlim=c(1,7),m1,t(aa1),type="l",xaxt = "n",col=1:4)
legend("right", c("Cluster 1", "Cluster 2", "Cluster 3", "Promedio"), pch=c(19,19,19,19),
       col=1:4)
axis(1, at=1:6, labels=c("div","pres","aprov","buenac","noimp","ahorro"))
```

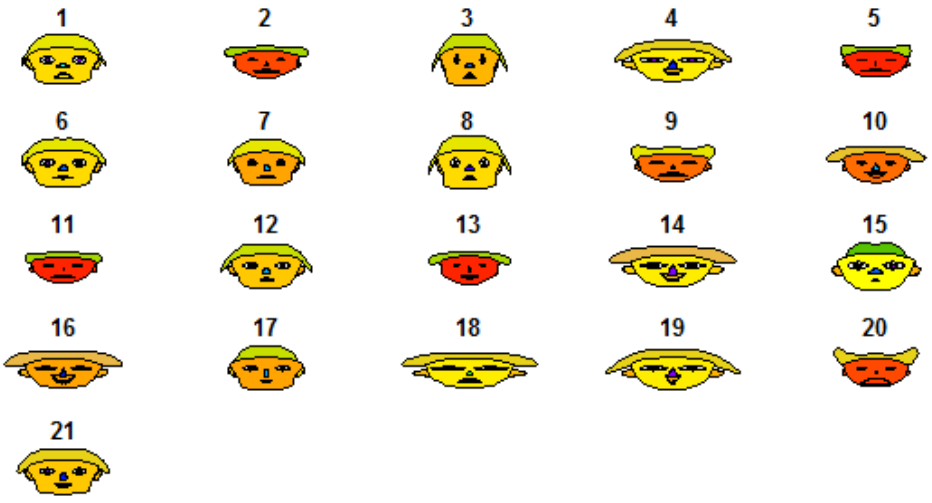


# Otra forma realizar Cluster Jerárquico con la función agnes
library(cluster)
agn1 = agnes(d, metric="euclidean", method="average", stand = FALSE)
agn1
plot(agn1,hang=-1)



Caras de Chernoff

library(aplpack)
library(tcltk)
datosf=cbind(datosc,cut2)
datosf
str(datosf)
faces(datosf[,1:6])



VI. Análisis Cluster No Jerárquico

Es aquel donde se conoce a priori el número de grupos "k" que se desea, y las observaciones son asignadas a cada uno de esos "k" clusters de forma tal que maximiza la homogeneidad de los sujetos asignados a un mismo grupo y la heterogeneidad entre los distintos clusters.

Procedimiento: (Método k-Means)

1. Se determinan los centroides iniciales de los k grupos, estos es, los valores medios de las variables que caracterizan las observaciones en cada uno de esos grupos. Estos centroides se conocen como semillas.
2. Cada observación se asigna a aquel cluster, de entre los k existentes, cuyo centroide esté más cercano a esa observación en términos de distancia euclídea.
3. Se recalculan los centroides de los k grupos de acuerdo con las observaciones que han sido clasificadas en casa uno de ellos. Si el cambio en los centroides es mayor que un valor criterio de convergencia preestablecido, se vuelve al paso 2, finalizando el proceso cuando se cumpla el criterio de convergencia o se supere un número prefijado de iteraciones.

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

	Coordenadas del centroide	
Cluster	X1	X2
(AB)	2	2
(CD)	-1	-2

Cluster	A	B
(AB)	10	10
(CD)	61	9

	Coordenadas del centroide	
Cluster	X1	X2
(A)	5	3
(BCD)	-1	-1

Cluster	A	B	C	D
(A)	0	40	41	89
(BCD)	52	4	5	5

Estudio de la Asociación entre los 2 análisis cluster

Una vez creadas las dos variables de agrupamiento, es el momento de estudiar la relación (asociación) entre las mismas. Este estudio se lleva a cabo con el objetivo de comprobar la coincidencia entre los resultados del **ACJ** y del **ACNJ**.

Tabla de contingencia Average Linkage (Between Groups) * Cluster Number of Case

Recuento

		Cluster Number of Case			
		1	2	3	Total
Average Linkage (Between Groups)	1	0	0	9	9
	2	0	6	0	6
	3	6	0	0	6
Total		6	6	9	21

Pruebas de chi-cuadrado

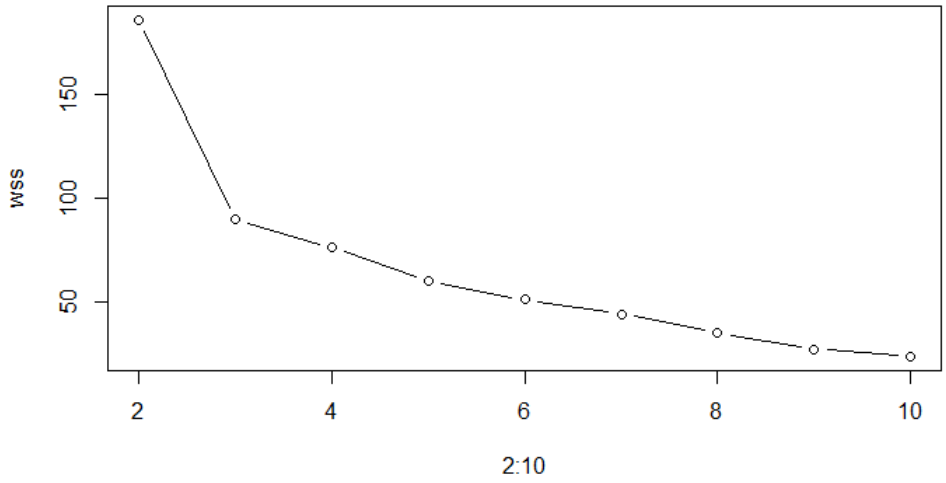
	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	42.000[a]	4	.000
Razón de verosimilitudes	45.318	4	.000
Asociación lineal por lineal	20.000	1	.000
N de casos válidos	21		

a. 9 casillas (100.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 1.71.

VII. Otras técnicas para determinar número de clusters

Criterio de la Suma de Cuadrados entre Cluster

```
# Criterio de la Suma de Cuadrados entre Clusters
wss<-numeric()
for(h in 2:10){
  b<-kmeans(datosc,h)
  wss[h-1]<-b$tot.withinss
}
plot(2:10,wss,type="b")
```



VIII. Análisis Cluster con Análisis Factorial

Cuando el número de variables es mayor a 2 se puede llevar a cabo un análisis de componentes principales o un análisis factorial para ver, si en realidad, caen dentro de un espacio de dimensiones reducidas. Es importante que al utilizar cualquiera de estas técnicas no trabajen con calificaciones estandarizadas.

El análisis factorial para el presente estudio es el siguiente:

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2.665	44.423	44.423	2.665	44.423	44.423
2	2.030	33.828	78.251	2.030	33.828	78.251
3	.581	9.685	87.935			
4	.439	7.312	95.248			
5	.223	3.712	98.960			
6	.062	1.040	100.000			

Método de extracción: Análisis de Componentes principales.

Matriz de componentes[a]

	Componente	
	1	2
Divertido	.961	-.022
Presupuesto	.085	.761
Aprovecho	.915	-.140
Buena compra	-.172	.841
No importa	-.922	-.129
Ahorro	.131	.840

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

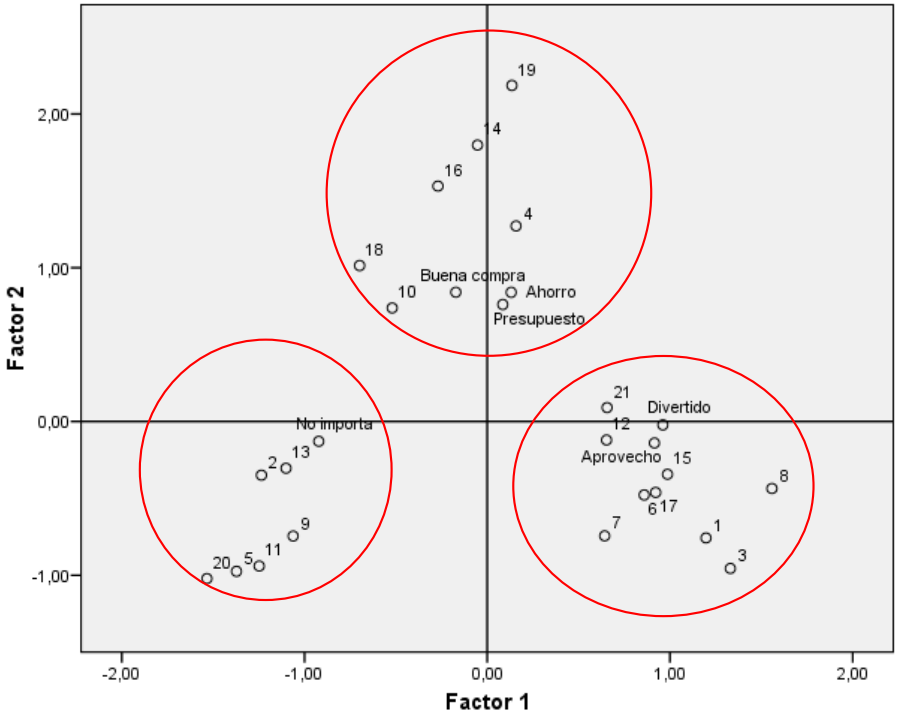
En el que se aprecia que con dos factores se está reteniendo el 78.251% como variabilidad común.

Las puntuaciones o scores de los individuos es el siguiente:

Individuos	Factor 1	Factor 2
1	1.19704	-0.75734
2	-1.23571	-0.34886
3	1.33062	-0.95596
4	0.15777	1.2717
5	-1.37218	-0.97453
6	0.85876	-0.47748
7	0.64251	-0.74387

Individuos	Factor 1	Factor 2
8	1.55801	-0.43577
9	-1.06232	-0.7447
10	-0.52022	0.73873
11	-1.24808	-0.93998
12	0.65268	-0.12054
13	-1.1013	-0.304
14	-0.05338	1.79739
15	0.98646	-0.34329
16	-0.26963	1.53101
17	0.92154	-0.46128
18	-0.69861	1.01417
19	0.13472	2.18546
20	-1.53438	-1.0216
21	0.65572	0.09074

Y el gráfico (ploteo) respectivo es:



A la vista de lo observado en este gráfico, se desprende que existen 3 clusters bien diferenciados:

Cluster 1 - casos: 14, 16, 10, 4, 19 y 18

Cluster 2 - casos: 2, 13, 5, 11, 9 y 20

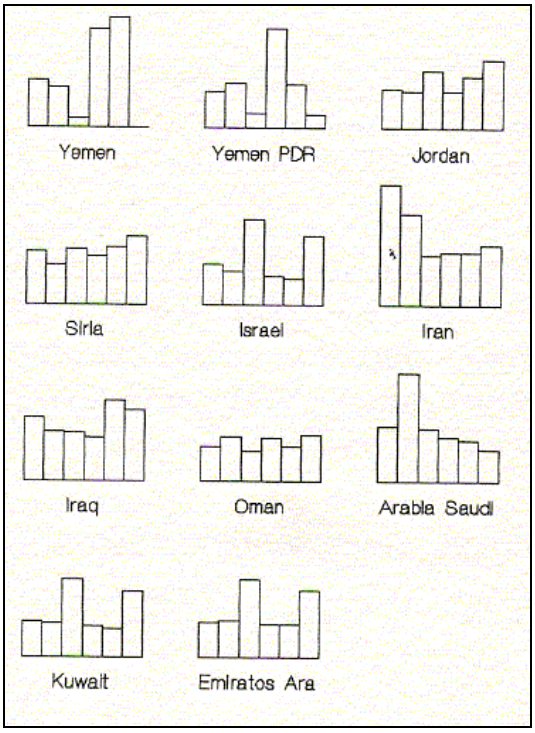
Cluster 3 - casos: 3, 8, 6, 7, 12, 21, 1, 17 y 15

VIII. Métodos gráficos para el Análisis Cluster

Los gráficos multivariantes se caracterizan por no ajustarse al habitual sistema de coordenadas cartesianas y por posibilitar la representación multivariante de un conjunto de variables cuantitativas clasificadas en función de una variable cualitativa. Una característica común para la correcta ejecución de estos gráficos es la conveniencia de que las variables estén en una escala común; de lo contrario, la interpretación del gráfico estará distorsionada. Si las variables no cumplen este requisito, será preciso tipificarlas.

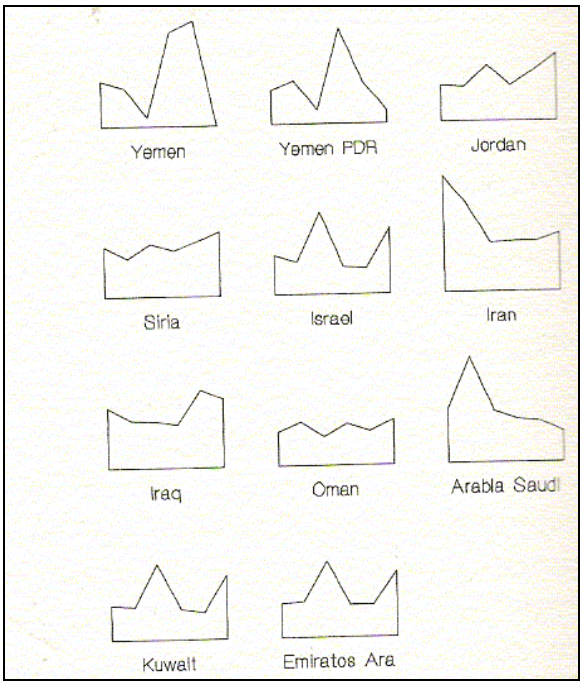
Iconos de histogramas

Se puede construir un histograma con p barras para cada una de las categorías de una variable cualitativa, donde cada barra del histograma representa el nivel de cada una de p variables cuantitativas.



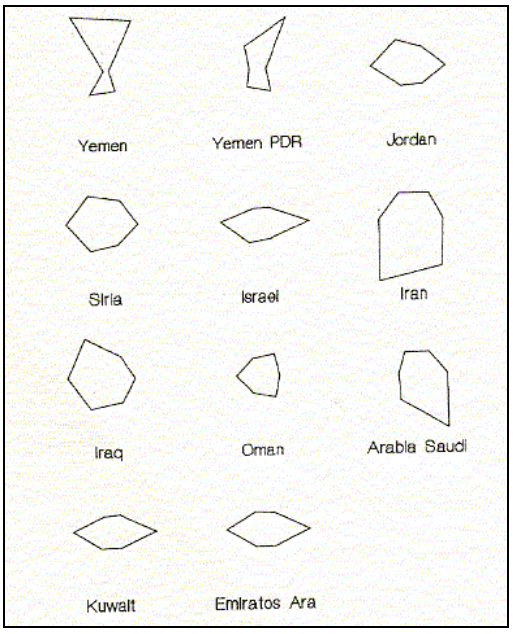
Gráficos de perfiles

Un gráfico muy parecido al ícono de histogramas es el gráfico de perfiles. Consiste en unir los extremos de las barras de cada histograma.



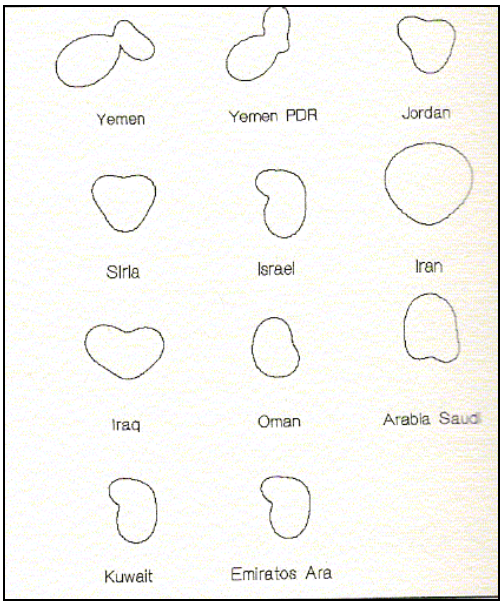
Gráficos en estrella

Si un gráfico de perfiles se construye en coordenadas polares, el resultado es un gráfico en estrella. Las variables que se pretenden representar se suponen formando vectores con origen en el centro de un círculo y longitud igual al nivel de cada una de ellas, si se unen los extremos de cada vector, el resultado es una estrella cuya forma estará en función de los niveles de las variables.



Gráficos de Fourier en coordenadas polares

Una técnica para examinar cómo se agrupan entre sí un conjunto de variables consiste en aplicar la transformación de Fourier y representar gráficamente los resultados.



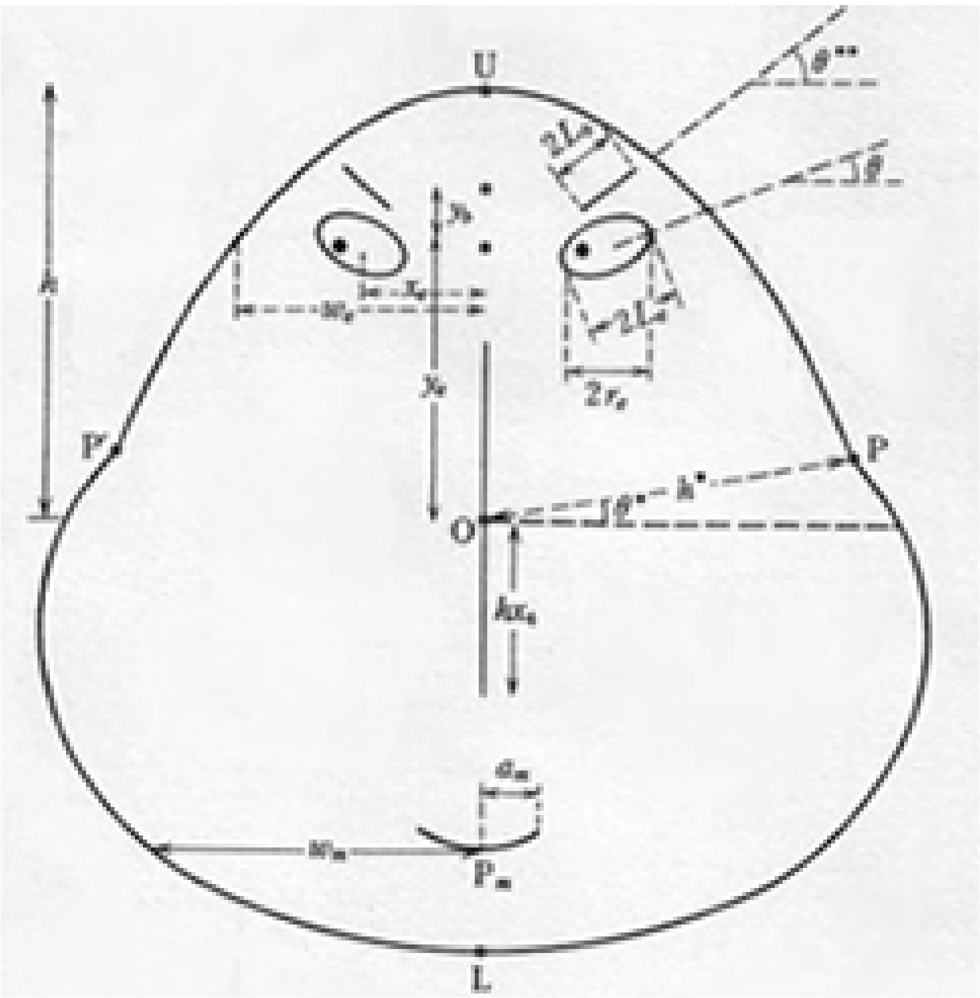
Si se construye en coordenadas cartesianas, este gráfico se conoce como gráfico de Fourier o también curvas de Andrews y tiene el inconveniente que las curvas pueden solaparse

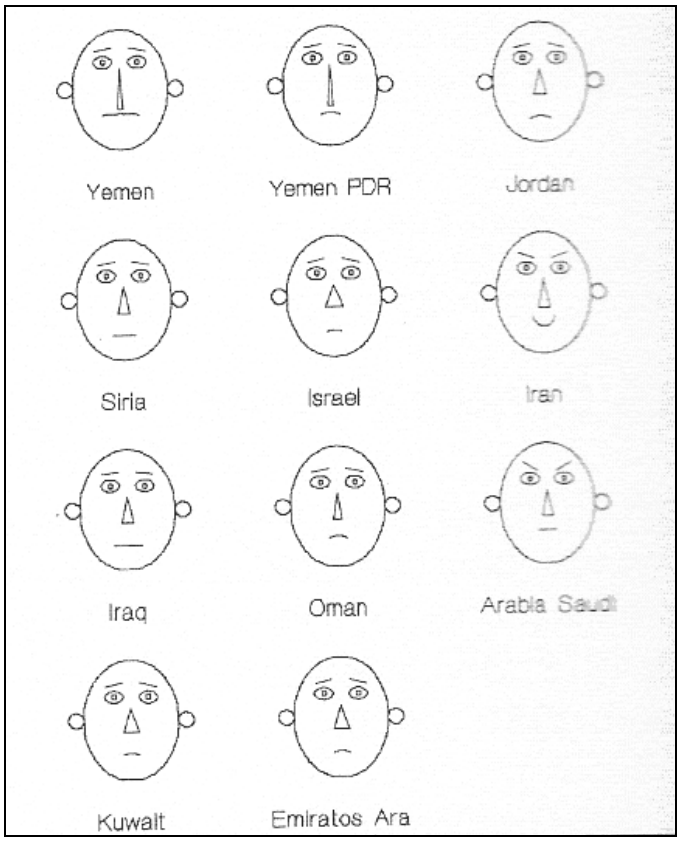
Caras de Chernoff

Chernoff (1973) propuso utilizar diferentes rasgos del rostro humano para representar múltiples variables cuantitativas. Cada variable es asignada a un rasgo del rostro, de forma que el resultado es un rostro que refleja el equilibrio de las diversas variables implicadas

Este gráfico permite representar hasta 20 variables conjuntamente, asignándolas a una del as siguiente características del rostro:

1. Curvatura de la boca
2. Angulo de la ceja
3. Anchura de la nariz
4. Longitud de la nariz
5. Longitud de la boca
6. Altura del centro de la boca
7. Separación entre los ojos
8. Altura del centro de los ojos
9. Inclinación de los ojos
10. Excentricidad de los ojos
11. Cuerpo medio de los ojos
12. Posición de las pupilas
13. Altura de las cejas
14. Longitud de la ceja
15. Altura de la cara
16. Excentricidad de la elipse superior de la cara
17. Excentricidad de la elipse inferior de la cara
18. Nivel de las orejas
19. Radio de la oreja
20. Longitud del cabello


U
P
P'
O
L
P_m
w_m
d_m
$2r_e$
h^*
θ^*
θ^{**}
θ


Yemen
Yemen PDR
Jordan
Siria
Israel
Iran
Iraq
Oman
Arabia Saudi
Kuwait
Emiratos Ara