Tema 5: Análisis de Cluster y Multidimensional Scaling

Introducción

El análisis de cluster es una técnica cuya idea básica es agrupar un conjunto de observaciones en un número dado de *clusters* o grupos. Este agrupamiento se basa en la idea de *distancia* o similitud entre las observaciones.

La obtención de dichos clusters depende del criterio o distancia considerados. Por ejemplo, una baraja de cartas españolas se podría dividir de distintos modos: en cuatro clusters (los cuatro palos), en ocho clusters (los cuatro palos y según sean figuras o números), en dos clusters (figuras y números). Es decir, todo depende de lo que consideremos como *similar*.

El número posible de combinaciones de grupos y de elementos que integran los posibles grupos se hace intratable desde el punto de vista computacional, aún con un número escaso de observaciones.

Se hace necesario, pues, encontrar métodos o algoritmos que infieran el número y componentes de los clusters más aceptable, aunque no sea el óptimo absoluto.

Previamente es necesario considerar el concepto de medida de similitud.

Medidas de similitud

En realidad, es bastante subjetivo el hecho de elegir una medida de similitud ya que depende de las escalas de medida. Se pueden agrupar observaciones según la similitud

expresada en términos de una distancia. Si se agrupan variables, es habitual utilizar como medida de similitud los coeficientes de correlación en valor absoluto. Para variables categóricas existen también criterios basados en la posesión o no de los atributos (tablas de presencia-ausencia).

Dados dos vectores $\mathbf{x}_i$, $\mathbf{x}_j$ pertenecientes a $\mathbb{R}^k$, diremos que hemos establecido una distancia entre ellos si definimos una función d con las propiedades siguientes:

1. $d : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^+$, es decir $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$;
2. $d(\mathbf{x}_i, \mathbf{x}_i) = 0 \quad \forall i$, la distancia entre un elemento y sí mismo es cero.
3. $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$, la distancia es simétrica
4. $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_p) + d(\mathbf{x}_p, \mathbf{x}_j)$, la distancia verifica la propiedad triangular.

Estas propiedades generalizan la noción intuitiva de distancia euclídea entre dos puntos.

Ejemplos de distancias entre objetos

Distancia euclídea

Dados dos objetos I_1 y I_2 medidos según dos variables x_1 y x_2, la distancia euclídea entre ambos es:

$$d_{I_1 I_2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}.$$

Con más dimensiones (o variables que se miden) es equivalente a:

$$d_{I_1 I_2} = \sqrt{\sum_{k=1}^{p} (x_{1k} - x_{2k})^2}$$

En notación vectorial se expresa como

$$d^2_{I_i I_j} = (x_i - x_j)'(x_i - x_j).$$

Si se consideran n objetos para $i, j \in \{1, \ldots, n\}$, la distancia total es

$$\mathbf{d} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2}.$$

Distancia de Minkowski

$$d_{I_iI_j} = \left[\sum_k |x_{ik} - x_{jk}|^m\right]^{1/m}$$

donde $m \in \mathbb{N}$.

Si $m = 1$, se tiene la distancia en valor absoluto y si $m = 2$, la euclídea.

Distancia de Mahalanobis

Se define como

$$d^2_{I_iI_j} = (x_i - x_j)'W^{-1}(x_i - x_j)$$

donde W es la matriz de covarianzas entre las variables. De este modo, las variables se ponderan según el grado de relación que exista entre ellas, es decir, si están más o menos correlacionadas. Si la correlación es nula y las variables están estandarizadas, se obtiene la distancia euclídea.

Ejemplos de distancias entre variables

Coeficiente de correlación de Pearson

Se define como:

$$r = \frac{S_{xy}}{S_x S_y}$$

donde S_{xy} es la covarianza muestral entre x e y, S_x y S_y son las desviaciones estándar de x e y respectivamente.

Coeficiente de correlación de rangos de Kendall

Se comparan las ordenaciones que dan dos variables, es decir, los datos se ordenan según dos criterios o características y se establece el número de concordancias y discordancias.

Método:

1. Calculo todas las posibles parejas. Tomo una pareja (i, j). Si están ordenados igual según las dos variables o criterios, se marca una *concordancia* (es decir, si el elemento i está delante del elemento j según ambas variables o criterios). Si no lo están, se establece una *discordancia*.

2. El número total de parejas distintas que se pueden hacer con n elementos es $\binom{n}{2} = \frac{n(n-1)}{2}$. Se cuenta, además

   $a =$ número total de concordancias,

   $b =$ número total de discordancias,

3. Se define el coeficiente de correlación de rangos como:

$$\tau = \frac{a-b}{\frac{n(n-1)}{2}}$$

Coeficiente de correlación de rangos de Spearman

Se consideran, igual que antes, n objetos clasificados según dos variables o criterios.

Por ejemplo, supongamos dos variables x e y que toman n valores emparejados (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n). Se definen los rangos sobre cada una de las variables, de modo que se emparejan (r_{x_1}, r_{y_1}), (r_{x_2}, r_{y_2}), $\cdots$, (r_{x_n}, r_{y_n}) :

$$\begin{array}{cc|cc} x_1 & y_1 & r_{x_1} & r_{y_1} \\ x_2 & y_2 & r_{x_2} & r_{y_2} \\ \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & r_{x_n} & r_{y_n} \end{array}$$

Se definen las diferencias $d_i = (r_{x_i} - r_{y_i})$, es decir, las diferencias de la posición del individuo i-ésimo según la clasificación (rango) dada por x y la clasificación (rango) dada por y.

El coeficiente de correlación se define, entonces, como

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Coeficientes de asociación (*matching types*)

Se consideran variables dicotómicas que toman como posibles valores 0 ó 1, del tipo *presencia – ausencia*. Existen diferentes formas de medir las coincidencias.

Ejemplo: Se tienen dos observaciones en las que se consideran 5 variables dicotómicas (*sí / no*).

Sea Sí = 1 y No = 0

individuos\variables	x_1	x_2	x_3	x_4	x_5
A	1	1	0	0	1
B	0	1	0	1	0

Un posible coeficiente de similitud sería: m/N donde m = número de variables comunes a los dos elementos y M es el número total de variables. En este ejemplo, sería 2/5.

Antes de mostrar una serie de medidas habituales, se tienen que definir los siguientes términos para 2 individuos dados.

X_{Aj} = valor del individuo A en la variable j-ésima $\in \{1, 0\}$.

X_{Bj} = valor del individuo B en la variable j-ésima $\in \{1, 0\}$.

$V = \sum_j X_{A_j}\left(1 - X_{B_j}\right)$ Nº de atributos donde A es 1 y B es 0

$R = \sum_j X_{A_j} X_{B_j}$ Nº de atributos donde A y B son 1

$S = \sum_j \left(1 - X_{A_j}\right)\left(1 - X_{B_j}\right)$ Nº de atributos donde A y B son 0

$T = \sum_j \left(1 - X_{A_j}\right) X_{B_j}$ Nº de atributos donde A es 0 y B es 1

$U = R + S + T + V$ Nº total de atributos o variables

En el ejemplo anterior,

$$\begin{aligned} V &= 1(1-0)+1(1-1)+0(1-0)+0(1-1)+1(1-0)=2 \\ R &= 1 \\ S &= 1 \\ T &= 1 \\ U &= 5 \end{aligned}$$

Esto da lugar a distintos índices de similaridad,.por ejemplo,

Indice de Russel-Rao

$$C = \frac{R}{U}$$

En el ejemplo es 1/5.

Indice de Kendall

$$C = 1 - \frac{V+T}{U}$$

En el ejemplo es 2/5.

Indice de Jaccard

$$C = \frac{R}{R+T+V}$$

En el ejemplo es 1/4.

Indice de Dice-Sorensen

$$C = \frac{2R}{2R+T+V}$$

En el ejemplo es 2/5.

Los índices más habituales son los de Jaccard y Dice-Sorensen.

Cuando se consideran variables categóricas otra posible medida de distancia se construye considerando la tabla de asociación entre variables como una tabla de contingencia y calculando el valor de la chi-cuadrado, χ^2, de modo que se puede definir la distancia como el *coeficiente de contingencia*:

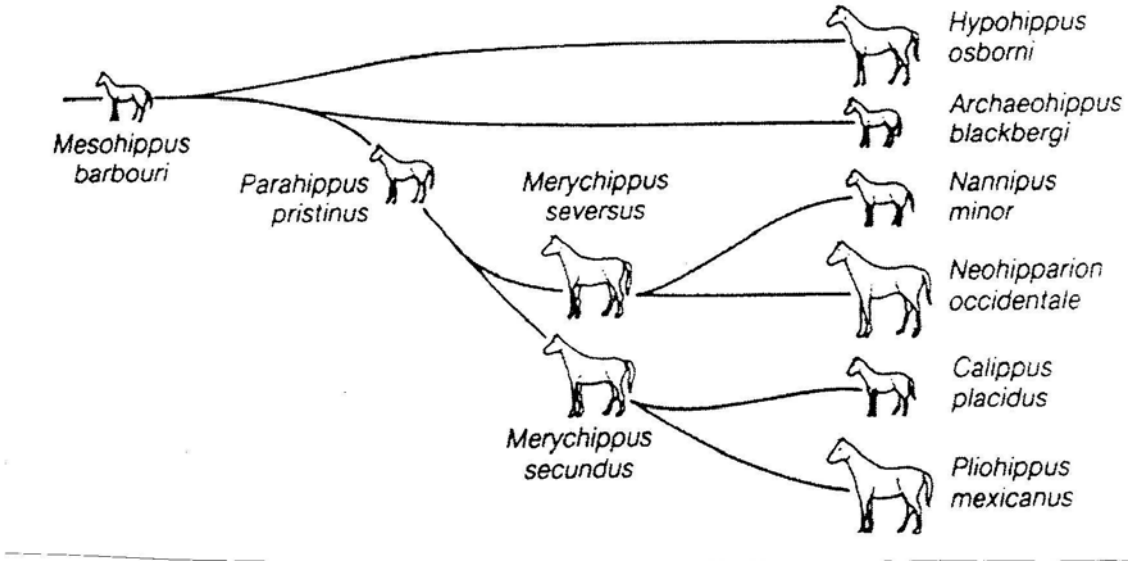
$$d_{ij} = 1 - \sqrt{\frac{\chi^2}{n}}.$$

Métodos de cluster jerárquicos

En la práctica, no se pueden examinar todas las posibilidades de agrupar los elementos, incluso con los ordenadores más rápidos. Una solución se encuentra en los llamados métodos jerárquicos. Se tienen dos posibles formas de actuar:

Métodos jerárquicos aglomerativos: se comienza con los objetos o individuos de modo individual; de este modo, se tienen tantos clusters iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único cluster.

Métodos jerárquicos divididos: se actúa al contrario. Se parte de un grupo único con todas las observaciones y se van dividiendo según lo *lejanos* que estén.



En cualquier caso, de ambos métodos se deriva un *dendograma*, que es un gráfico que ilustra cómo se van haciendo las subdivisiones o los agrupamientos, etapa a etapa.

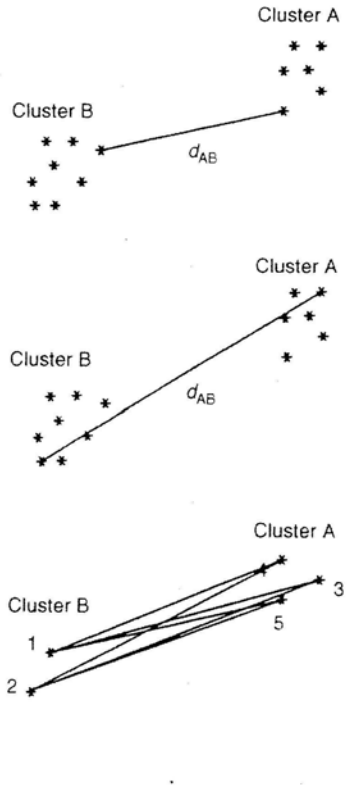
Consideramos aquí los métodos aglomerativos con diferentes métodos de unión (*linkage methods*). Los más importantes son:

(*i*) *Mínima distancia* o vecino más próximo.

(*ii*) *Máxima distancia* o vecino más lejano.

(*iii*) *Distancia media* (average distance).

Se puede observar que, de este modo, se define una posible distancia entre dos clusters: la correspondiente a la pareja de elementos más cercana, la más lejana o la media de todas las posibles parejas de elementos de ambos clusters:



Definidas las distancias anteriores, se puede considerar el algoritmo básico, dados N objetos o individuos:

1. Empezar con N clusters (el número inicial de elementos) y una matriz $N \times N$ simétrica de distancias o similitudes. $D = [d_{ik}]_{ik}$.

2. Dentro de la matriz de distancias, buscar aquella entre los clusters U y V (más próximos, más distantes o en media más próximos) que sea la menor entre todas, d_{uv}.

3. Juntar los clusters U y V en uno solo. Actualizar la matriz de distancias:

   (i) Borrando las filas y columnas de los clusters U y V.

   ii) Formando la fila y columna de las distancias del nuevo cluster (UV) al resto de clusters.

4. Repetir los pasos (2) y (3) un total de $(N - 1)$ veces.

Al final, todos los objetos están en un único cluster cuando termina el algoritmo. Además, se guarda la identificación de los clusters que se van uniendo en cada etapa, así como las distancias a las que se unen. Finalmente se construye un *dendograma*.

Ejemplo con *mínima distancia*:

Sea la matriz de distancias entre 5 objetos la dada por:

$$D = [d_{ik}]_{ik} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \mathbf{2} & 8 & 0 \end{bmatrix} \end{array}$$

Cada uno de los objetos comienza siendo un cluster. Como $\text{mín}_{i,k}\, d_{ik} = d_{53} = 2$ los objetos 3 y 5 se unen para formar el cluster (35). Para construir el siguiente nivel, calculo la distancia entre el cluster (35) y los restantes objetos 1, 2 y 4. Así:

$$\begin{aligned} d_{(35),1} &= \text{mín}\{d_{31}, d_{51}\} = \text{mín}\{3, 11\} = 3 \\ d_{(35),2} &= \text{mín}\{d_{32}, d_{52}\} = \text{mín}\{7, 10\} = 7 \\ d_{(35),4} &= \text{mín}\{d_{34}, d_{54}\} = \text{mín}\{9, 8\} = 8 \end{aligned}$$

Reconstruyo la matriz de distancias:

$$D = [d_{ik}]_{ik} = \begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ \mathbf{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

Cojo la menor distancia, $d_{(35),1} = 3$ y junto, así, el cluster (35) con el 1.

Calculo ahora las distancias del nuevo cluster a los dos elementos que quedan:

$$\begin{aligned} d_{(351),2} &= \text{mín}\{d_{(35),2}, d_{12}\} = \text{mín}\{7, 9\} = 7 \\ d_{(351),4} &= \min\{d_{(35),4}, d_{14}\} = \min\{8, 6\} = 6 \end{aligned}$$

La matriz de distancias queda como:

$$D = [d_{ik}]_{ik} = \begin{matrix} \\ (351) \\ 2 \\ 4 \end{matrix} \begin{matrix} \begin{matrix} (351) & 2 & 4 \end{matrix} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \mathbf{5} & 0 \end{bmatrix} \end{matrix}$$

La mínima distancia se alcanza entre los clusters 2 y 4 $d_{24} = 5$. Se obtienen así dos clusters: (351) y (24). La distancia que los separa es:

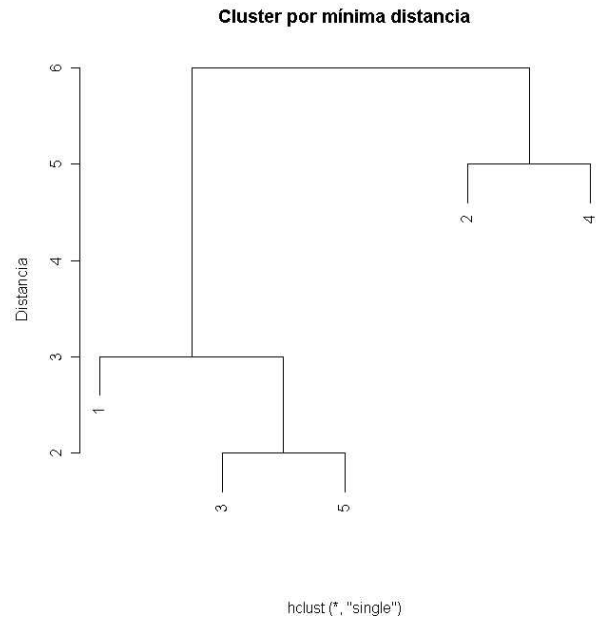
$$d_{(351),(24)} = \text{mín}\{d_{(351),2}, d_{(351),4}\} = \text{mín}\{7, 6\} = 6$$

Así, la matriz de distancias queda como:

$$D = [d_{ik}]_{ik} = \begin{matrix} \\ (351) \\ (24) \end{matrix} \begin{matrix} \begin{matrix} (351) & (24) \end{matrix} \\ \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{matrix}$$

Cuando la distancia es iguala 6, todos los objetos se unen en un único cluster.

Se pueden dibujar dendogramas:



Este tipo de distancia no funciona bien cuando los objetos están próximos.

Se obtienen dendogramas similares si se utiliza la distancia máxima, o la distancia media, aunque las distancias a las que se van uniendo los objetos en los clusters varían en cada caso.

Problemas

- Las fuentes de error y variación no entran en consideración con los métodos jerárquicos. Esto implica una gran sensibilidad a observaciones anómalas o *outliers*.

- Si un objeto se ha colocado erróneamente en un grupo al principio del proceso, ya no se puede arreglar en una etapa posterior.

- Un sistema de trabajo conveniente es usar varias distancias o similitudes con los mismos objetos y observar si se mantienen los mismos clusters o grupos. Así, se comprueba la existencia de grupos naturales.

Estos métodos se pueden usar para clasificar no sólo observaciones, sino también variables usando como medida de similitud algún coeficiente de correlación.

Métodos no jerárquicos

Se usan para agrupar objetos, pero no variables, en un conjunto de k clusters ya predeterminado. No se tiene que especificar una matriz de distancias ni se tienen que almacenar las iteraciones. Todo esto permite trabajar con un número de datos mayor que en el caso de los métodos jerárquicos.

Se parte de un conjunto inicial de clusters elegidos al azar, que son los *representantes* de todos ellos; luego se van cambiando de modo iterativo. Se usa habitualmente el método de las k-medias.

Método de las k-medias

Es un método que permite asignar a cada observación el cluster que se encuentra más próximo en términos del centroide (media). En general, la distancia empleada es la euclídea.

Pasos:

1. Se toman al azar k clusters iniciales.

2. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clusters y se reasignan a los que estén más próximos. Se vuelven a recalcular los centroides de los k clusters después de las reasignaciones de los elementos.

3. Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.

Usualmente, se especifican k centroides iniciales y se procede al paso (2) y, en la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.

Ejemplo Supongamos dos variables x_1 y x_2 y 4 elementos: A, B, C, D. con la siguiente tabla de valores:

	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Se quiere dividir estos elementos en dos grupos ($k = 2$).

De modo arbitrario, se dividen los elementos en dos clusters (AB) y (CD) y se calculan los centroides de los dos clusters.

Cluster (AB) :

$\bar{x}_1$	$\bar{x}_2$
$\frac{5-1}{2} = 2$	$\frac{3+1}{2} = 2$

Cluster (CD) :

$\bar{x}_1$	$\bar{x}_2$
$\frac{1-3}{2} = -1$	$\frac{-2-2}{2} = -2$

En el paso (2), calculamos las distancias euclídeas de cada observación al grupo de centroides y reasignamos cada una al grupo más próximo. Si alguna observación se mueve de grupo, hay que volver a calcular los centroides de los grupos. Así, las distancias son:

$$\begin{aligned} d^2(A,(AB)) &= (5-2)^2 + (3-2)^2 = 10 \\ d^2(A,(CD)) &= (5+1)^2 + (3+2)^2 = 61 \end{aligned}$$

Como A está más próximo al cluster (AB) que al cluster (CD), no se reasigna.

Se hace lo mismo para el elemento B:

$$\begin{aligned} d^2(B,(AB)) &= (-1-2)^2 + (1-2)^2 = 10 \\ d^2(B,(CD)) &= (-1+1)^2 + (1+2)^2 = 9 \end{aligned}$$

Por lo cual, el elemento B se reasigna al cluster (CD) dando lugar al cluster (BCD). A continuación, se vuelven a calcular los centroides:

Cluster	$\bar{x}_1$	$\bar{x}_2$
A	5	3
(BCD)	-1	-1

Nuevamente, se vuelven a calcular las distancias para cada observación para ver si se producen cambios con respecto a los nuevos centroides:

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

Como no se producen cambios, entonces la solución para $k = 2$ clusters es: A y (BCD).

Si se quiere comprobar la estabilidad de los grupos, es conveniente volver a correr el algoritmo con otros clusters iniciales (una nueva partición inicial).

Una vez considerados los clusters finales, es conveniente interpretarlos; para ello, se pueden cruzar con otras variables categóricas o se pueden ordenar de modo que los objetos del primer cluster aparezcan al principio y los del último cluster al final.

Tablas de análisis de la varianza

El objetivo que se persigue al formar los clusters es que los centroides estén lo más separados entre sí como sea posible y que las observaciones dentro de cada cluster estén muy próximas al centroide. Lo anterior se puede medir con el estadístico F de Snedecor:

$$F_{n,m} = \frac{\chi^2_n / n}{\chi^2_m / m},$$

y equivale al cociente de dos distribuciones chi-cuadrado divididas entre sus grados de libertad.

El estadístico F se calcula, así, como un cociente de *medias de cuadrados*. En el caso del análisis de cluster:

$$F = \frac{\text{medias de cuadrados entre clusters}}{\text{medias de cuadrados dentro de clusters}}$$

Si $F > 1$, las distancias entre los centroides de los grupos son mayores que las distancias de los elementos dentro de los grupos. Esto es lo que se pretende para que los clusters estén suficientemente diferenciados entre sí.

Problemas que surgen al fijar k clusters iniciales

(*i*) Si dos centroides iniciales caen por casualidad en un único cluster natural, entonces los clusters que resultan están poco diferenciados entre sí.

(*ii*) Si aparecen outliers, se obtiene por lo menos un cluster con sus objetos muy dispersos.

(iii) Si se imponen previamente k clusters puede dar lugar a grupos artificiales o bien a juntar grupos distintos.

Una posible solución es considerar varias elecciones del número k de clusters comparando luego sus coeficientes de la F de Snedecor.

Ejemplos

Se considera primero una muestra de los años de vida esperados por país, edad y sexo procedentes de Keyfitz y Flieger (1971) que ya se consideró en el tema 4 sobre Análisis Factorial.

Se considera otra muestra de 48 objetos de cerámica romana donde se miden diferentes tipos de oxidación (ver `http://biostatistics.iop.kcl.ac.uk/publications/everitt/`):

	AL2O3	FE2O3	MGO	CAO	NA2O	K2O	TIO2	MNO	BAO
1	1.76	1.11	0.30	0.46	0.50	1.02	1.29	0.48	1.07
2	1.58	0.85	0.25	0.49	0.50	0.97	1.27	0.41	1.29
3	1.70	0.89	0.27	0.45	0.50	0.98	1.26	0.54	1.00
...	...	...	...	...	...	...	...	...	...
43	1.56	0.11	0.08	0.01	0.06	0.56	1.17	0.02	0.93
44	1.38	0.32	0.10	0.02	0.06	0.68	1.72	0.02	1.07
45	1.79	0.19	0.09	0.06	0.04	0.56	1.33	0.04	1.29

Multidimensional Scaling (MDS) (Escalamiento Multidimensional)

Las técnicas de MDS tratan sobre el siguiente problema: para un conjunto de similitudes (o distancias) observadas entre un par de objetos de un total de N, se trata de encontrar una representación gráfica de estos en pocas dimensiones, de modo que sus posiciones *casi* ajusten las similitudes (o distancias) originales.

Con N objetos, se buscan configuraciones de $q < (N-1)$ dimensiones, de modo que el ajuste entre las posiciones originales y las posiciones en las q dimensiones sea el más preciso posible; esto se mide mediante el concepto del *stress*.

Si se usan las magnitudes originales de las distancias (o similitudes), se tiene el llamado escalamiento multidimensional *métrico*. Si se usan rangos (orden de las observaciones), en vez de distancias, se tiene el MDS *no métrico*.

Procedimiento básico

Dados N objetos, existen $M = \frac{N(N-1)}{2}$ distancias (o similitudes) entre pares de diferentes objetos. Alternativamente, se pueden usar rangos ordenados. Las similitudes se pueden ordenar en orden creciente como:

$$s_{i_1 k_1} < s_{i_2 k_2} < \ldots < s_{i_m k_m}$$

Aquí $s_{i_1 k_1}$ es la menor de las M similitudes, donde i_1, k_1 es el par de observaciones que son menos similares y, del mismo modo, i_m, k_m, las más similares. Buscamos una configuración de dimensión q tal que las distancias entre los N objetos mantengan el orden expresado en la relación anterior. Es decir, tiene que cumplirse:

$$d^{(q)}_{i_1 k_1} > d^{(q)}_{i_2 k_2} > \ldots > d^{(q)}_{i_m k_m}$$

Lo importante es que se mantenga el orden, no las magnitudes en sí.

Para un número dado de dimensiones (q), puede que no se encuentre una configuración como la anterior que conserve las similitudes anteriores. Kruskal dio una medida de la

adecuación de la representación en q dimensiones a las similitudes originales; dicha medida se denomina *stress*. Se buscan representaciones geométricas en q dimensiones de modo que el stress sea mínimo. Empíricamente, se considera que si el stress es alrededor de 0,2, la bondad del ajuste es pobre; si es del 0,05, la bondad del ajuste es buena y a partir de 0,025 es excelente.

La idea es minimizar el stress para un número fijo q de dimensiones mediante un proceso iterativo.

Relación con otras técnicas multivariantes

Las técnicas de escalamiento multidimensional están relacionadas con el Análisis Factorial y el Análisis de Cluster. Tanto el Análisis Factorial como el MDS usan una matriz (en el primer caso, de covarianzas o de correlaciones y en el segundo, de similitudes) y generan un espacio con el mínimo número de dimensiones posible donde se representan los datos. En general, el MDS necesita menos dimensiones que el Análisis Factorial para representar los datos o las variables. Por otro lado, el MDS proporciona una descripción dimensional cuantitativa de las variables, mientras que el Análisis Factorial permite, además, una descripción de los objetos o individuos en forma de sus puntuaciones factoriales.

Con respecto a las técnicas de Análisis de Cluster, el MDS comparte con ellas las siguientes características: investigan la estructura de un conjunto de variables, el punto de partida es una matriz de proximidades y en la representación gráfica que se obtiene se pueden interpretar las distancias.

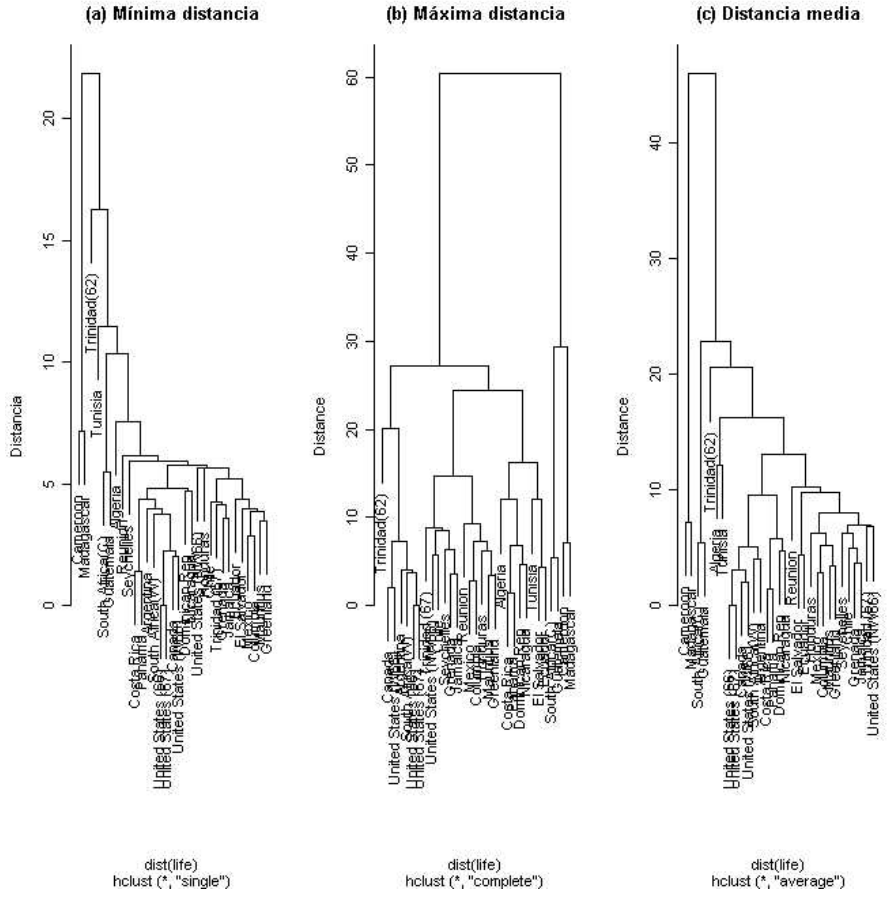
Ejemplo

Se consideran las distancias en relación a vuelos entre 10 ciudades norteamericanas:

	Atlanta	Chicago	Denver	Houston	L. Angeles	Miami	N York	S Francisco	Seattle	Washington
Atlanta	0.00	587.00	1212.00	701.00	1936.00	604.00	748.00	2139.00	218.00	543.00
Chicago	587.00	0.00	920.00	940.00	1745.00	1188.00	713.00	1858.00	1737.00	597.00
Denver	1212.00	920.00	0.00	879.00	831.00	1726.00	1631.00	949.00	1021.00	1494.00
Houston	701.00	940.00	879.00	0.00	1374.00	968.00	1420.00	1645.00	1891.00	1220.00
L Angeles	1936.00	1745.00	831.00	1374.00	0.00	2339.00	2451.00	347.00	959.00	2300.00
Miami	604.00	1188.00	1726.00	968.00	2339.00	0.00	1092.00	2594.00	2734.00	923.00
N York	748.00	713.00	1631.00	1420.00	2451.00	1092.00	0.00	2571.00	2408.00	205.00
S Francisco	2139.00	1858.00	949.00	1645.00	347.00	2594.00	2571.00	0.00	678.00	2442.00
Seattle	218.00	1737.00	1021.00	1891.00	959.00	2734.00	2408.00	678.00	0.00	2329.00
Washington	543.00	597.00	1494.00	1220.00	2300.00	923.00	205.00	2442.00	2329.00	0.00

Análisis de Cluster (con R)

```
# Se dibujan los dendogramas según los tres tipos de linkages empleados
par(mfrow=c(1,3))

plclust(hclust(dist(life),method="single"),labels=row.names(life),ylab="Distancia")
title("(a) Mínima distancia")

plclust(hclust(dist(life),method="complete"),labels=row.names(life),ylab="Distancia")
title("(b) Máxima distancia")

plclust(hclust(dist(life),method="average"),labels=row.names(life),ylab="Distancia")
title("(c) Distancia media")
```



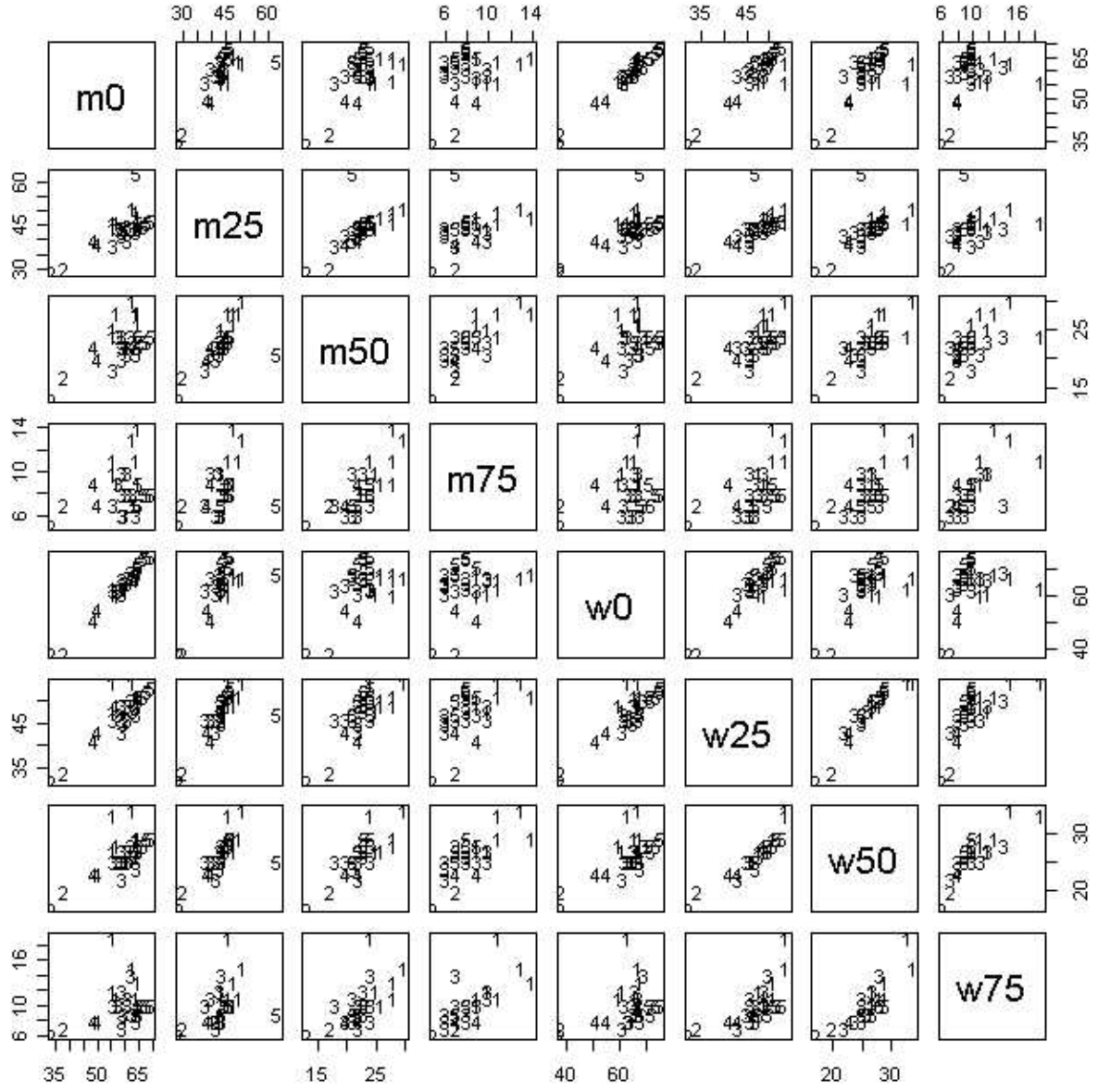

```
# Se determinan los paises que pertenecen a cada cluster
# usando el linkage del maximo, cortando a una distancia de 21
> cuantos <- cutree(hclust(dist(life),method="complete"),h=21)
> pais.clus <- lapply(1:5, function(eso){row.names(life)[cuantos==eso]})
> pais.clus

[[1]]
[1] "Algeria"       "Tunisia"       "Costa Rica"    "Dominican Rep"
[5] "El Salvador"   "Nicaragua"     "Panama"        "Ecuador"

[[2]]
[1] "Cameroon"   "Madagascar"
```

```
[[3]]
 [1] "Mauritius"           "Reunion"             "Seychelles"
 [4] "Greenland"           "Grenada"             "Honduras"
 [7] "Jamaica"             "Mexico"              "Trinidad (67)"
[10] "United States (NW66)" "Chile"               "Columbia"

[[4]]
[1] "South Africa(C)" "Guatemala"


[[5]]
[1] "South Africa(W)"     "Canada"              "Trinidad(62)"
[4] "United States (66)"  "United States (W66)" "United States (67)"
[7] "Argentina"
```

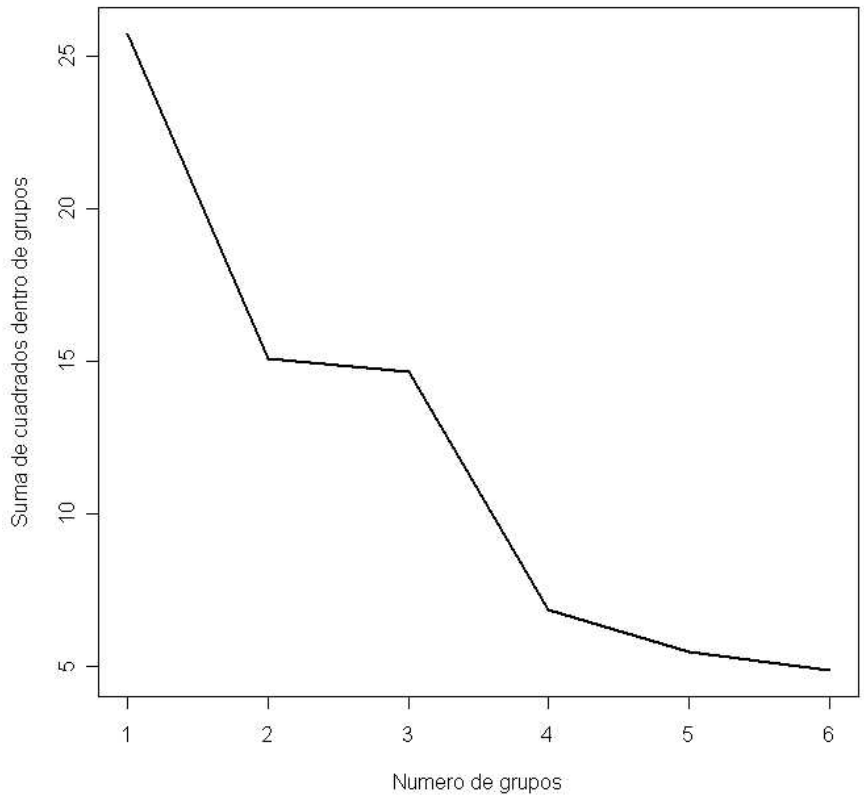
 # Calculo las medias de cada una de las variables dentro de cada cluster

```
> pais.medias <- lapply(1:5,function(eso){apply(life[cuantos==eso,],2,mean)})
> pais.medias


[[1]]
    m0    m25    m50    m75     w0    w25    w50    w75
61.375 47.625 26.875 10.750 65.000 50.750 29.250 12.625

[[2]]
  m0  m25  m50  m75   w0  w25  w50  w75
36.0 29.5 15.0  6.0 38.0 33.0 18.5  6.5

[[3]]
       m0       m25       m50       m75        w0       w25       w50       w75
60.083333 42.750000 22.000000  7.583333 64.916667 46.833333 25.333333  9.666667

[[4]]
  m0  m25  m50  m75   w0  w25  w50  w75
49.5 39.5 21.0  8.0 53.0 42.0 23.0  8.0

[[5]]
       m0       m25       m50       m75        w0       w25       w50       w75
66.428571 48.000000 22.857143  7.857143 72.714286 50.714286 27.714286  9.714286
```

 # Se dibujan los cruces de variables con el cluster de pertenencia identificado
pairs(life,panel= function(x,y){text(x,y,cuantos)})



```
# Para que las escalas de las variables sean iguales, se divide cada valor entre
# el rango de las variables: (max-min)
rge <- apply(cacharros,2,max)-apply(cacharros,2,min)
cacharros <- sweep(cacharros,2,rge,FUN="/")
n <- length(cacharros[,1])

# Calculo las sumas de cuadrados dentro de grupos para todos los datos
# Calculo la suma de cuadrados dentro de grupos con 1 solo grupo
scd1 <- (n-1)*sum(apply(cacharros,2,var))

# Calculo la suma de cuadrados dentro de grupos con 2 a 6  grupos
scd <- numeric(0)
for(i in 2:6) {
      W <- sum(kmeans(cacharros,i)$withinss)
      scd <- c(scd,W)
}

# Junto los resultados de 1 grupo con los de 2:6 grupos
scd <- c(scd1,scd)

# Dibujo las sumas de cuadrados dentro de grupos frente al numero de grupos
plot(1:6,scd,type="l",xlab="Numero de grupos",ylab="Suma de cuadrados dentro de
grupos",lwd=2)
```




```
# El resultado mejor es con 2 o 3 grupos
cacharros.kmedia <- kmeans(cacharros,3)
cacharros.kmedia

K-means clustering with 3 clusters of sizes 14, 10, 21
```

```
Cluster means:
     AL2O3     FE2O3        MGO        CAO      NA2O       K2O      TIO2
1 1.162216 0.7218439 0.71311301 0.12458472 0.2821429 1.3337125 0.8754579
2 1.658879 0.1874419 0.09552239 0.02267442 0.0637500 0.6436306 1.3076923
3 1.581219 0.8637874 0.27498223 0.54595792 0.4321429 0.9881711 1.2020757
         MNO      BAO
1 0.72619048 1.137755
2 0.01975309 1.142857
3 0.43915344 1.224490

Clustering vector:
 [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
[39] 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 2.874794 1.466713 3.164386

Available components:
[1] "cluster"  "centers"  "withinss" "size"


# Los resultados anteriores son para los datos estandarizados.
# Para calcular los resultados sobre los resultados reales:

lapply(1:3,
function(eso){apply(cacharros[cacharros.kmedia$cluster==eso,],2,mean)} )

[[1]]
    AL2O3     FE2O3       MGO       CAO      NA2O       K2O      TIO2       MNO
1.1622163 0.7218439 0.7131130 0.1245847 0.2821429 1.3337125 0.8754579 0.7261905
      BAO
1.1377551

[[2]]
     AL2O3      FE2O3        MGO        CAO       NA2O        K2O       TIO2
1.65887850 0.18744186 0.09552239 0.02267442 0.06375000 0.64363057 1.30769231
       MNO        BAO
0.01975309 1.14285714

[[3]]
    AL2O3     FE2O3       MGO       CAO      NA2O       K2O      TIO2       MNO
1.5812194 0.8637874 0.2749822 0.5459579 0.4321429 0.9881711 1.2020757 0.4391534
      BAO
1.2244898
```

Se observa que el cluster 3 se caracteriza por tener un valor alto en óxido de aluminio, un valor bajo en óxido de hierro y un valor bajo en óxido de calcio.
El cluster 2 tiene un valor alto en óxido de manganeso y también en óxido de potasio.
El cluster 1 tiene un valor alto en óxido de calcio.

Análisis MDS (con R)

Se considera la matriz de distancias entre 10 ciudades de USA, estas distancias no son euclídeas dado que se consideran sobre una esfera.

```
aire.dist <- read.table("c:\\cursoCIII\\AnMultiv\\practicas\\aireMDS.txt")
dimnames(aire.dist)[[1]] <- c("Atlanta", "Chicago", "Denver", "Houston", "Los
Angeles", "Miami", "New York", "San Francisco", "Seattle", "Washington DC")
dimnames(aire.dist)[[2]] <- dimnames(aire.dist)[[1]]
```

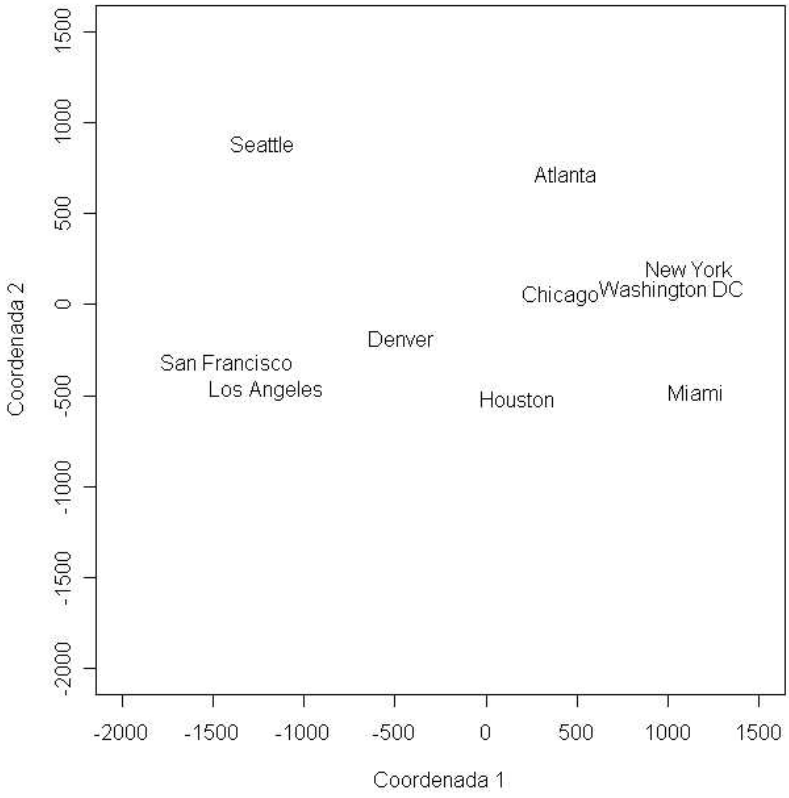
```
# Se efectua un analisis clasico MDS metrico
aire.mds <- cmdscale(as.matrix(aire.dist),k=9,eig=T)

Warning messages:
1: some of the first 9 eigenvalues are < 0 in: cmdscale(as.matrix(aire.dist), k
= 9, eig = T)
2: NaNs produced in: sqrt(ev)
```

```
# Calculamos los autovalores
aire.mds$eig
[1]  9.213705e+06  2.199924e+06  1.082863e+06  3.322361e+03  3.858824e+02
[6]  6.984919e-10 -9.323115e+01 -2.168535e+03 -9.090644e+03

# Normalizo los dos primeros autovalores
sum(abs(aire.mds$eig[1:2]))/sum(abs(aire.mds$eig))
[1] 0.9122472

sum(aire.mds$eig[1:2]^2)/sum(aire.mds$eig^2)
[1] 0.9870998
```

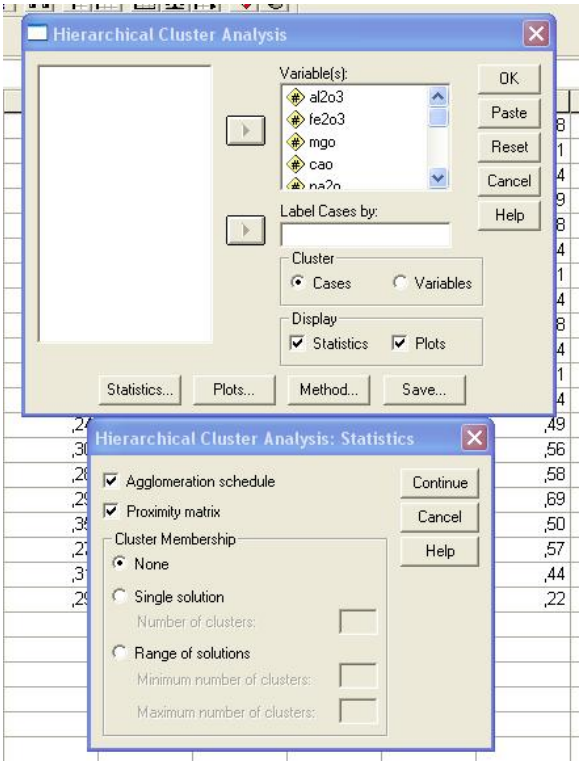
La solución con dos dimensiones es adecuada

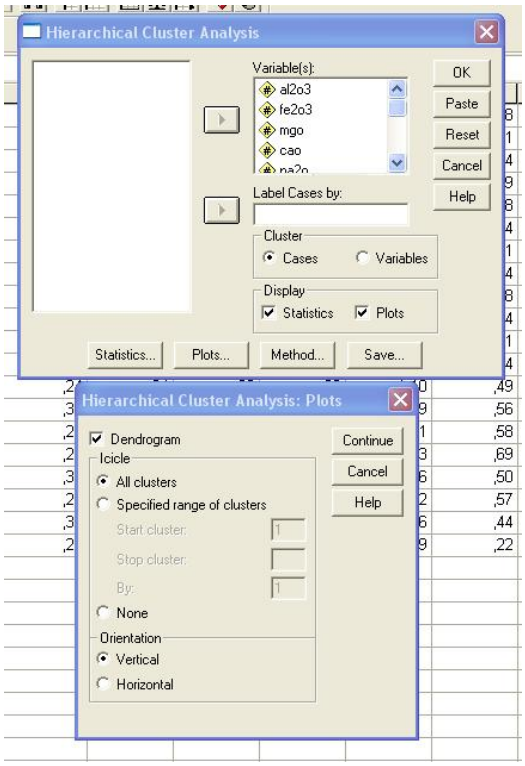
```
# Se muestran las coordenadas de las ciudades en las dos dimensiones
aire.mds$points[,1:2]

                   [,1]       [,2]
Atlanta        -434.7588  724.22221
Chicago        -412.6102   55.04016
Denver          468.1952 -180.65789
Houston        -175.5816 -515.22265
Los Angeles    1206.6772 -465.63705
Miami         -1161.6875 -477.98261
New York      -1115.5609  199.79247
San Francisco  1422.6887 -308.65595
Seattle        1221.5351  887.20174
Washington DC -1018.8972   81.89956
```

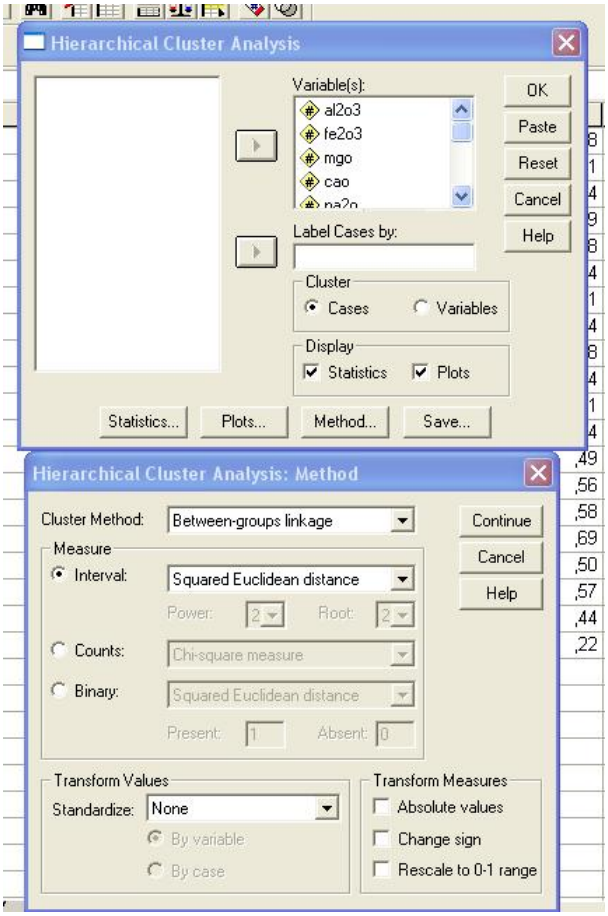
```
# Se dibujan las coordenadas de las ciudades en las dos dimensiones
par(pty="s")
plot(-aire.mds$points[,1],aire.mds$points[,2],type="n",xlab="Coordenada
1",ylab="Coordenada 2", xlim = c(-2000,1500),ylim=c(-2000,1500))
text(-aire.mds$points[,1],aire.mds$points[,2],labels=row.names(aire.dist))
```

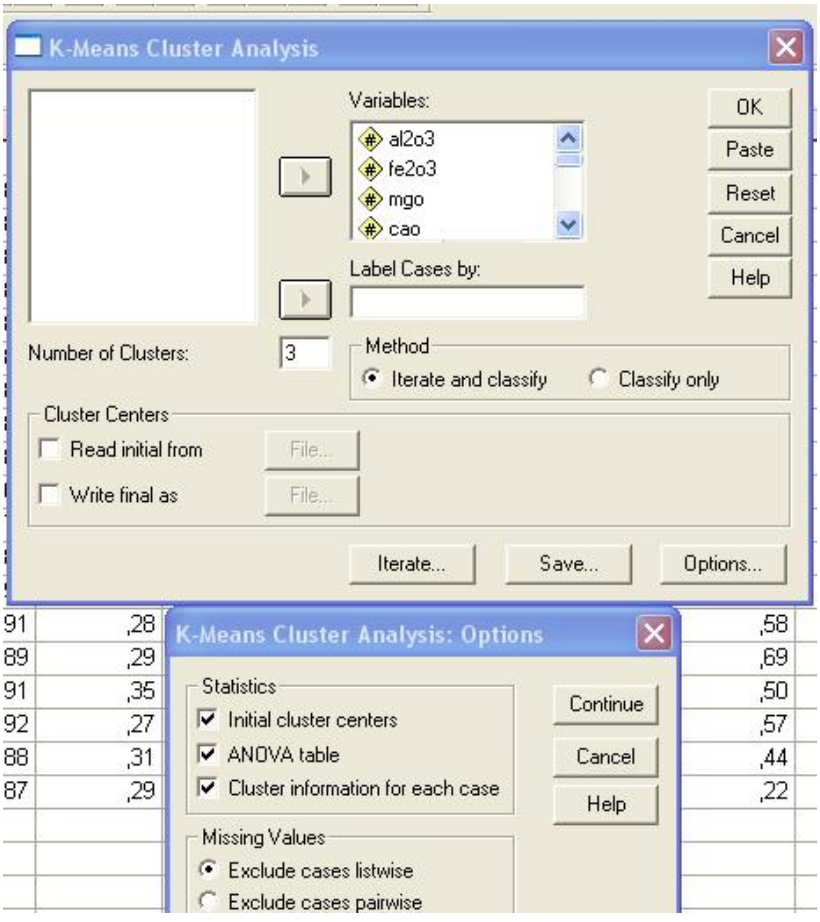

Seattle
Atlanta
New York
Chicago
Washington DC
Denver
San Francisco
Los Angeles
Houston
Miami
Coordenada 2
Coordenada 1
1500
1000
500
0
-500
-1000
-1500
-2000
-2000
-1500
-1000
-500
0
500
1000
1500

Análisis de Cluster (con SPSS)






Hierarchical Cluster Analysis
Variable(s):
al2o3
fe2o3
mgo
cao
Label Cases by:
Cluster
Cases
Variables
Display
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Method...
Save...
Hierarchical Cluster Analysis: Method
Cluster Method:
Between-groups linkage
Measure
Interval:
Squared Euclidean distance
Power:
Root:
Counts:
Chi-square measure
Binary:
Squared Euclidean distance
Present:
Absent:
Transform Values
Standardize:
None
By variable
By case
Transform Measures
Absolute values
Change sign
Rescale to 0-1 range
Continue
Cancel
Help


K-Means Cluster Analysis
Variables:
al2o3
fe2o3
mgo
cao
Label Cases by:
Number of Clusters:
3
Method
Iterate and classify
Classify only
Cluster Centers
Read initial from
Write final as
File...
File...
Iterate...
Save...
Options...
OK
Paste
Reset
Cancel
Help
K-Means Cluster Analysis: Options
Statistics
Initial cluster centers
ANOVA table
Cluster information for each case
Missing Values
Exclude cases listwise
Exclude cases pairwise
Continue
Cancel
Help

Cluster Jerárquico

Proximity Matrix																				
Case	**Squared Euclidean Distance**																			
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	,000	,153	,064	,195	,189	,119	,528	,131	,200	,342	,855	,543	,452	,217	,233	,549	,091	,529	,169	,285
2	,153	,000	,115	,008	,024	,082	,288	,063	,033	,259	,630	,301	,256	,059	,064	,395	,049	,260	,206	,194
3	,064	,115	,000	,168	,170	,102	,497	,107	,133	,197	,624	,437	,344	,207	,215	,499	,096	,556	,139	,279
4	,195	,008	,168	,000	,027	,105	,336	,095	,045	,322	,685	,364	,325	,057	,063	,407	,067	,253	,248	,198
5	,189	,024	,170	,027	,000	,093	,255	,072	,077	,369	,747	,336	,296	,050	,068	,397	,046	,198	,250	,216
6	,119	,082	,102	,105	,093	,000	,333	,015	,127	,327	,680	,336	,293	,160	,185	,689	,040	,426	,044	,101
7	,528	,288	,497	,336	,255	,333	,000	,232	,383	,610	,987	,145	,137	,340	,346	,769	,322	,241	,477	,515
8	,131	,063	,107	,095	,072	,015	,232	,000	,098	,284	,659	,241	,199	,135	,160	,610	,042	,352	,074	,132
9	,200	,033	,133	,045	,077	,127	,383	,098	,000	,152	,473	,284	,250	,115	,128	,443	,119	,397	,222	,181
10	,342	,259	,197	,322	,369	,327	,610	,284	,152	,000	,214	,286	,280	,471	,473	,799	,389	,897	,306	,323
11	,855	,630	,624	,685	,747	,680	,987	,659	,473	,214	,000	,436	,500	,916	,894	1,533	,833	1,424	,592	,465
12	,543	,301	,437	,364	,336	,336	,145	,241	,284	,286	,436	,000	,029	,456	,452	,990	,411	,557	,368	,359
13	,452	,256	,344	,325	,296	,293	,137	,199	,250	,280	,500	,029	,000	,352	,343	,804	,332	,454	,337	,398
14	,217	,059	,207	,057	,050	,160	,340	,135	,115	,471	,916	,456	,352	,000	,006	,258	,061	,133	,346	,357
15	,233	,064	,215	,063	,068	,185	,346	,160	,128	,473	,894	,452	,343	,006	,000	,258	,079	,131	,372	,386
16	,549	,395	,499	,407	,397	,689	,769	,610	,443	,799	1,533	,990	,804	,258	,258	,000	,435	,340	,993	1,087
17	,091	,049	,096	,067	,046	,040	,322	,042	,119	,389	,833	,411	,332	,061	,079	,435	,000	,266	,152	,215
18	,529	,260	,556	,253	,198	,426	,241	,352	,397	,897	1,424	,557	,454	,133	,131	,340	,266	,000	,712	,714
19	,169	,206	,139	,248	,250	,044	,477	,074	,222	,306	,592	,368	,337	,346	,372	,993	,152	,712	,000	,093
20	,285	,194	,279	,198	,216	,101	,515	,132	,181	,323	,465	,359	,398	,357	,386	1,087	,215	,714	,093	,000
This is a dissimilarity matrix																				

Average Linkage (Between Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	14	15	,006	0	0	9
2	2	4	,008	0	0	4
3	6	8	,015	0	0	6
4	2	5	,026	2	0	7
5	12	13	,029	0	0	12
6	6	17	,041	3	0	10
7	2	9	,052	4	0	9
8	1	3	,064	0	0	11
9	2	14	,076	7	1	13
10	6	19	,090	6	0	11
11	1	6	,119	8	10	13
12	7	12	,141	0	5	17
13	1	2	,157	11	9	15
14	10	11	,214	0	0	19
15	1	20	,220	13	0	17
16	16	18	,340	0	0	18
17	1	7	,357	15	12	18
18	1	16	,493	17	16	19
19	1	10	,584	18	14	0

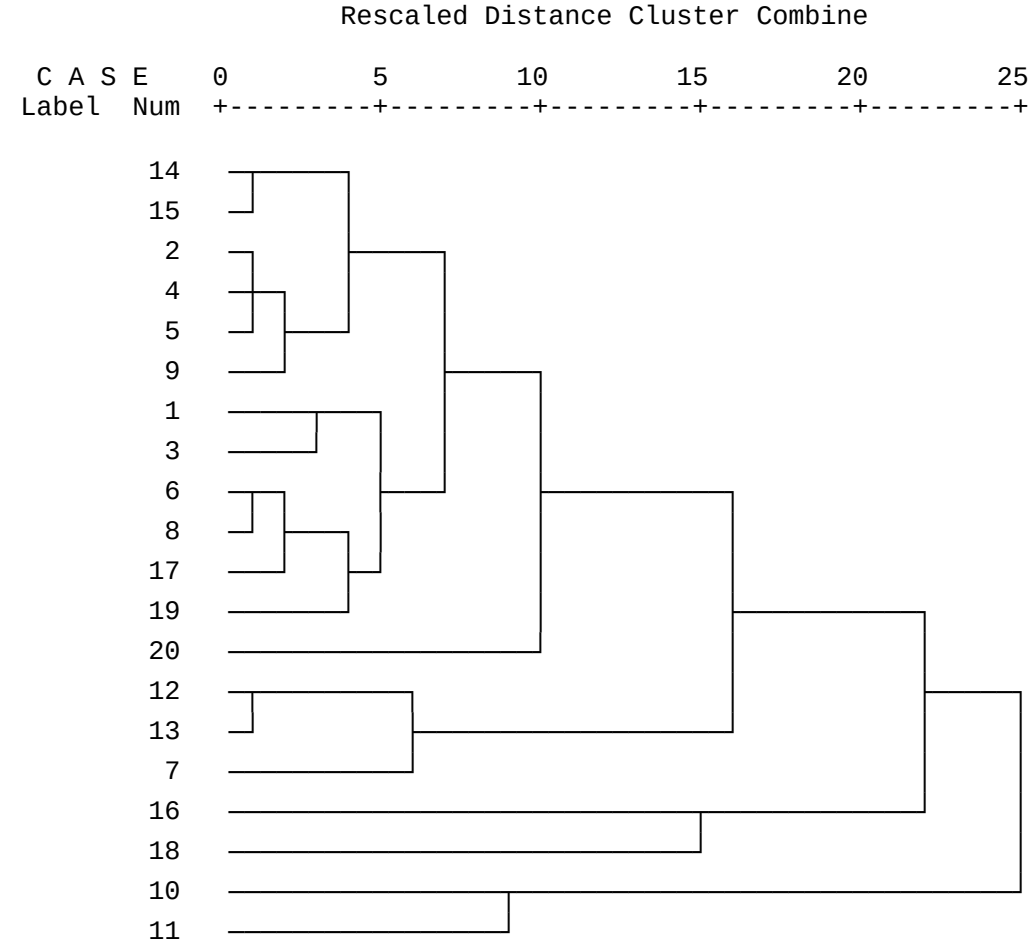
Vertical Icicle

N of clust	Case																																					
	11		**10**		**18**		**16**		**13**		**12**		**7**		**20**		**15**		**14**		**9**		**5**		**4**		**2**		**19**		**17**		**8**		**6**		**3**		**1**
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3	X	X	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4	X	X	X		X	X	X		X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
5	X	X	X		X		X		X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6	X	X	X		X		X		X	X	X	X	X		X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
7	X		X		X		X		X	X	X	X	X		X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
8	X		X		X		X		X	X	X	X	X		X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
9	X		X		X		X		X	X	X		X		X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
10	X		X		X		X		X	X	X		X		X		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X		X	X	X
11	X		X		X		X		X	X	X		X		X		X	X	X	X	X	X	X	X	X	X	X		X		X	X	X	X	X		X	X	X
12	X		X		X		X		X	X	X		X		X		X	X	X		X	X	X	X	X	X	X		X		X	X	X	X	X		X	X	X
13	X		X		X		X		X	X	X		X		X		X	X	X		X	X	X	X	X	X	X		X		X	X	X	X	X		X		X
14	X		X		X		X		X	X	X		X		X		X	X	X		X		X	X	X	X	X		X		X	X	X	X	X		X		X
15	X		X		X		X		X	X	X		X		X		X	X	X		X		X	X	X	X	X		X		X		X	X	X		X		X
16	X		X		X		X		X		X		X		X		X	X	X		X		X	X	X	X	X		X		X		X	X	X		X		X
17	X		X		X		X		X		X		X		X		X	X	X		X		X		X	X	X		X		X		X	X	X		X		X
18	X		X		X		X		X		X		X		X		X	X	X		X		X		X	X	X		X		X		X		X		X		X
19	X		X		X		X		X		X		X		X		X	X	X		X		X		X		X		X		X		X		X		X		X

Dendrogram

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

Dendrogram using Average Linkage (Between Groups)



K Medias

Initial Cluster Centers			
	Cluster		
	1	**2**	**3**
al2o3	1,48	1,54	1,28
fe2o3	,89	,82	,68
mgo	,29	,27	,22
cao	,47	1,01	,38
na2o	1,04	,41	,16
k2o	1,06	1,02	,72
tio2	1,23	1,22	,96
mno	,69	,41	,21
bao	1,36	1,36	,86

Iteration History(a)			
Iteration	**Change in Cluster Centers**		
	1	**2**	**3**
1	,338	,439	,339
2	,095	,040	,000
3	,000	,000	,000

a Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 3. The minimum distance between initial centers is ,877.

Cluster Membership		
Case Number	**Cluster**	**Distance**
1	2	,355
2	2	,172
3	2	,301
4	2	,246
5	2	,222
6	2	,195
7	2	,459
8	2	,116
9	2	,244
10	3	,260
11	3	,339
12	2	,447
13	2	,398
14	1	,170
15	1	,169
16	1	,379
17	2	,210
18	1	,284
19	2	,339
20	3	,389

Final Cluster Centers			
	Cluster		
	1	**2**	**3**
al2o3	1,56	1,61	1,44
fe2o3	,91	,87	,78
mgo	,28	,28	,25
cao	,55	,58	,41
na2o	,71	,41	,24
k2o	,99	1,01	,90
tio2	1,21	1,22	1,11
mno	,60	,43	,26
bao	1,46	1,21	,98

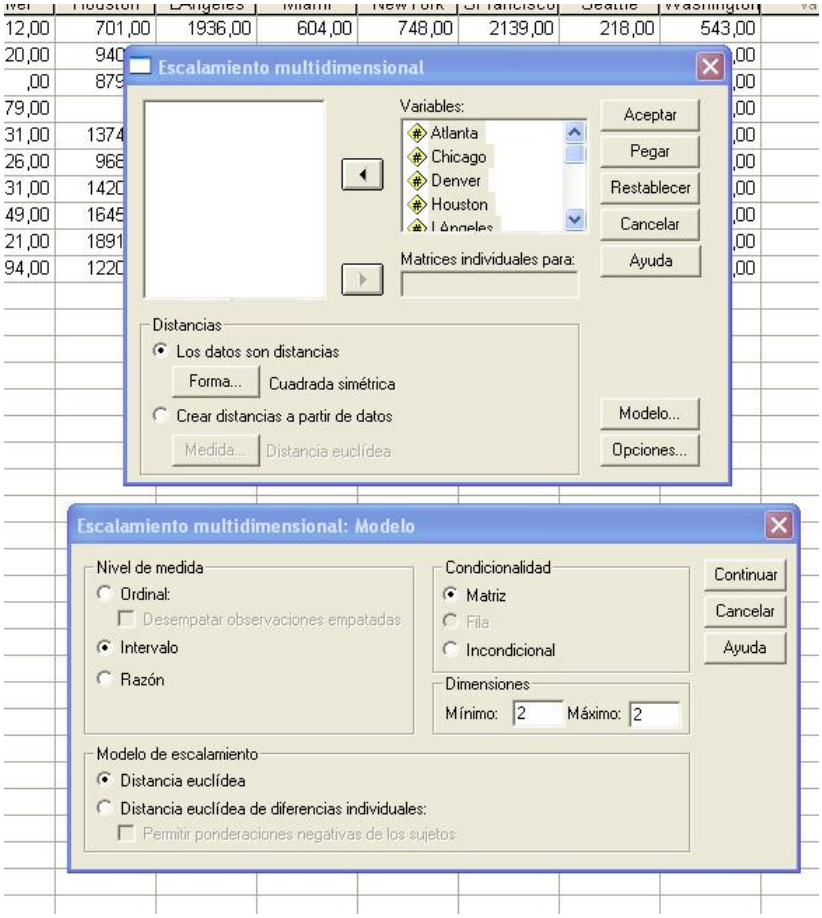
Distances between Final Cluster Centers			
Cluster	**1**	**2**	**3**
1		,439	,805
2	,439		,455
3	,805	,455	

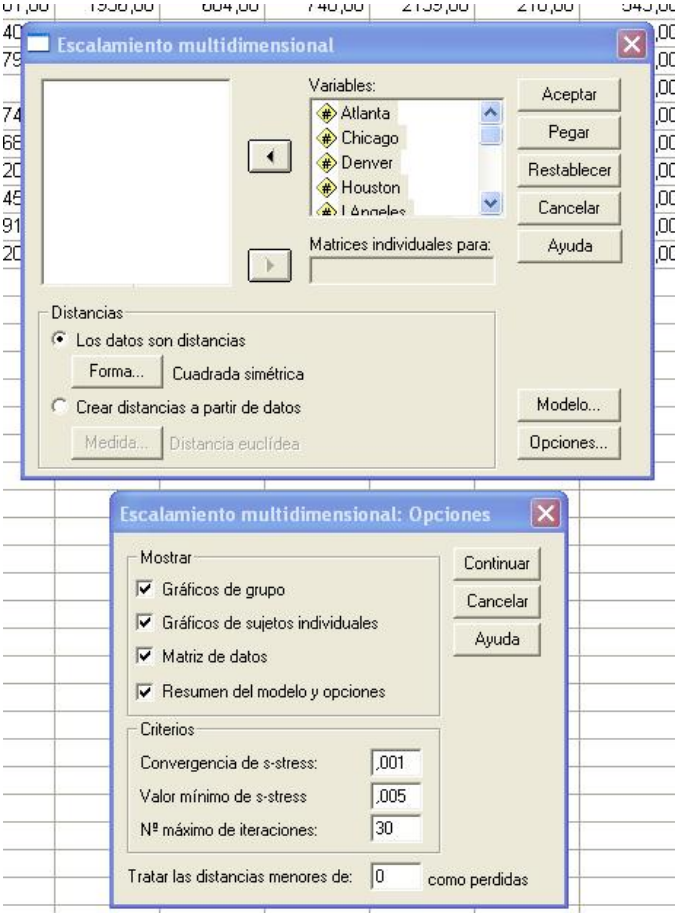
ANOVA						
	Cluster		**Error**		**F**	**Sig.**
	Mean Square	**df**	**Mean Square**	**df**		
al2o3	,037	2	,020	17	1,893	,181
fe2o3	,014	2	,005	17	2,500	,112
mgo	,001	2	,001	17	,810	,461
cao	,036	2	,029	17	1,246	,313
na2o	,215	2	,021	17	10,192	,001
k2o	,015	2	,004	17	3,420	,056
tio2	,017	2	,004	17	3,820	,043
mno	,102	2	,004	17	28,374	,000
bao	,209	2	,018	17	11,890	,001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Number of Cases in each Cluster		
Cluster	**1**	4,000
	2	13,000
	3	3,000
Valid		20,000
Missing		,000

Multidimensional Scaling (MDS) con SPSS




Escalamiento multidimensional
Variables:
Atlanta
Chicago
Denver
Houston
Matrices individuales para:
Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Distancias
Los datos son distancias
Forma...
Cuadrada simétrica
Crear distancias a partir de datos
Medida...
Distancia euclídea
Modelo...
Opciones...
Escalamiento multidimensional: Opciones
Mostrar
Gráficos de grupo
Gráficos de sujetos individuales
Matriz de datos
Resumen del modelo y opciones
Continuar
Cancelar
Ayuda
Criterios
Convergencia de s-stress:
,001
Valor mínimo de s-stress
,005
Nº máximo de iteraciones:
30
Tratar las distancias menores de:
0
como perdidas

Escalamiento multidimensional

Raw (unscaled) Data for Subject 1

	1	2	3	4	5
1	,000				
2	587,000	,000			
3	1212,000	920,000	,000		
4	701,000	940,000	879,000	,000	
5	1936,000	1745,000	831,000	1374,000	,000
6	604,000	1188,000	1726,000	968,000	2339,000
7	748,000	713,000	1631,000	1420,000	2451,000
8	2139,000	1858,000	949,000	1645,000	347,000
9	218,000	1737,000	1021,000	1891,000	959,000
10	543,000	597,000	1494,000	1220,000	2300,000

	6	7	8	9	10
6	,000				
7	1092,000	,000			
8	2594,000	2571,000	,000		
9	2734,000	2408,000	678,000	,000	
10	923,000	205,000	2442,000	2329,000	,000

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	,22655	
2	,15756	,06899
3	,15012	,00745
4	,14926	,00085

Iterations stopped because
S-stress improvement is less than ,001000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances. Stress values are Kruskal's stress formula 1.

For matrix

Stress = ,15369 RSQ = ,87390

Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	2
1	Atlanta	,5756	-,5818
2	Chicago	,6231	-,3291
3	Denver	-,7066	,1110
4	Houston	,1622	,9063
5	LAngeles	-1,6769	,4319
6	Miami	1,4468	,9603
7	NewYork	1,5567	-,3628
8	SFrancis	-1,9170	,0961
9	Seattle	-1,4507	-1,0821
10	Washingt	1,3869	-,1499

Optimally scaled data (disparities) for subject 1

	1	2	3	4	5
1	,000				
2	,911	,000			
3	1,692	1,327	,000		
4	1,054	1,352	1,276	,000	
5	2,596	2,357	1,216	1,894	,000
6	,933	1,662	2,334	1,387	3,099
7	1,112	1,069	2,215	1,952	3,239
8	2,850	2,499	1,363	2,233	,612
9	,451	2,348	1,453	2,540	1,376
10	,856	,924	2,044	1,702	3,051

	6	7	8	9	10
6	,000				
7	1,542	,000			
8	3,418	3,389	,000		
9	3,593	3,185	1,025	,000	
10	1,331	,434	3,228	3,087	,000

Abbreviated Name	Extended Name
SFrancis	SFrancisco
Washingt	Washington

Configuración de estímulos derivada

Modelo de distancia euclídea

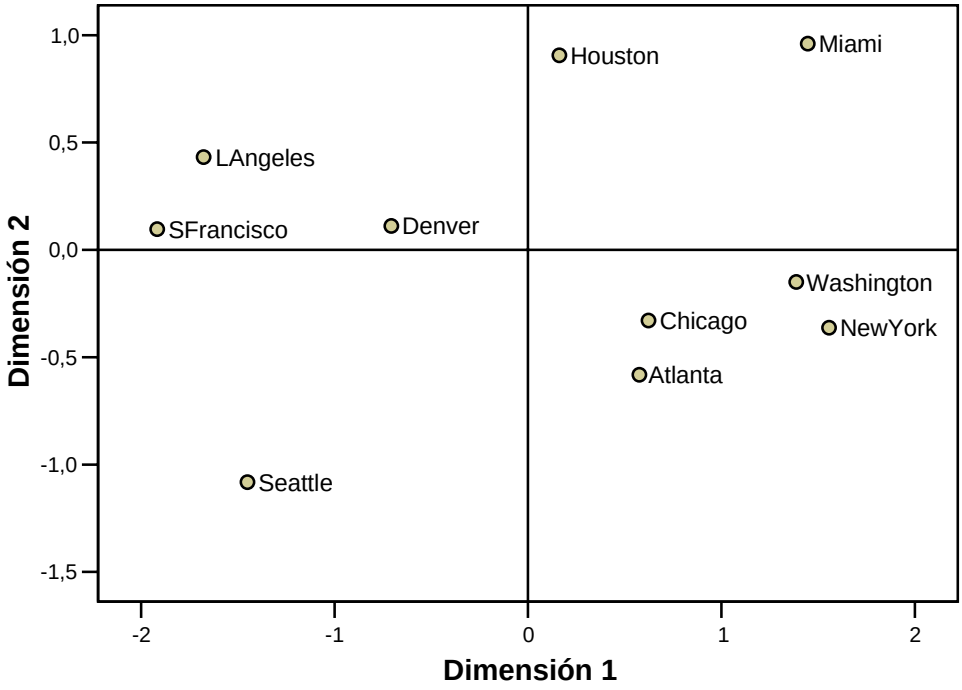
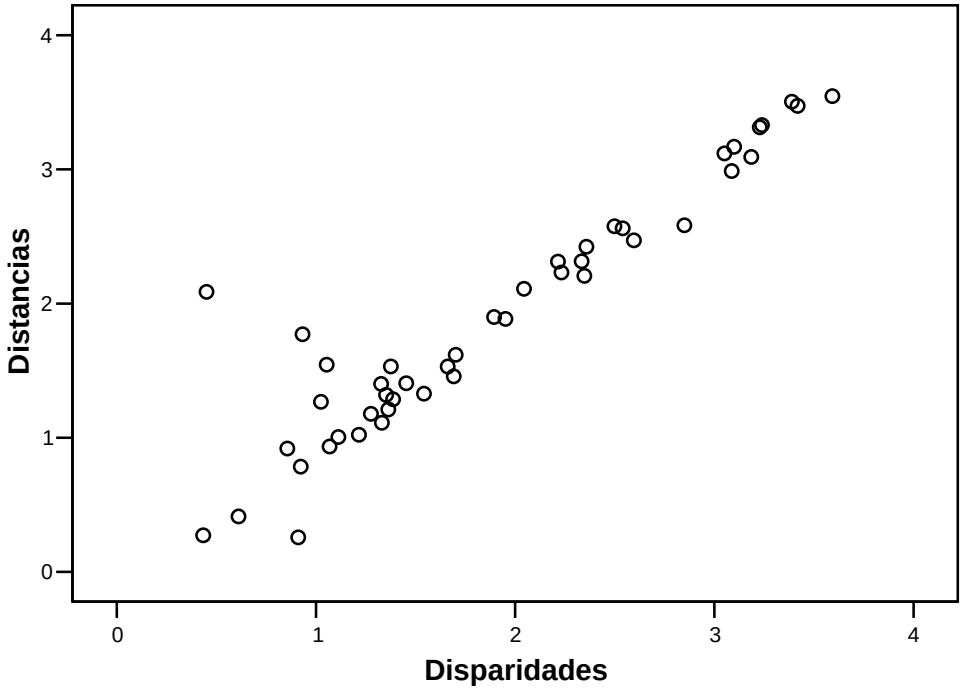


Gráfico de ajuste lineal

Modelo de distancia euclídea



Análisis de Cluster (con *SAS*)

```
/* Analisis de Cluster */
options ls=80 nodate nonumber;
title 'Análisis de Cluster de datos de cerámica';
data ceramica;
infile 'c:\...\CachaSAS.txt';
/* Hay 9 variables: */
input al2o3 fe2o3 mgo cao na2o k2o tio2 mno bao;
run;

proc cluster data=ceramica method=single simple ccc std outtree=
single;
var al2o3 fe2o3 mgo cao na2o k2o tio2 mno bao;
proc tree horizontal;
run;
```

Análisis de Cluster de datos de cerámica

The CLUSTER Procedure
Single Linkage Cluster Analysis

Variable	Mean	Std Dev	Skewness	Kurtosis	Bimodality
al2o3	1.4673	0.2533	-0.4026	-0.8763	0.4967
fe2o3	0.6693	0.2794	-1.0097	-0.3827	0.7128
mgo	0.3713	0.2601	0.9876	0.0891	0.5977
cao	0.2989	0.2638	0.7834	-0.0871	0.5158
na2o	0.3049	0.2232	0.9509	1.0509	0.4463
k2o	1.0189	0.2720	0.1279	-0.6614	0.3979
tio2	1.1240	0.2306	0.4038	0.3180	0.3291
mno	0.4356	0.2887	0.0916	-0.5279	0.3751
bao	1.1789	0.2135	-0.2723	0.1408	0.3200

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.19785459	1.67273872	0.4664	0.4664
2	2.52511587	1.64701600	0.2806	0.7470
3	0.87809987	0.42173477	0.0976	0.8446
4	0.45636510	0.07443216	0.0507	0.8953
5	0.38193294	0.11083473	0.0424	0.9377
6	0.27109821	0.15433632	0.0301	0.9678
7	0.11676189	0.02468356	0.0130	0.9808
8	0.09207833	0.01138513	0.0102	0.9910
9	0.08069320		0.0090	1.0000

The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Mean Distance Between Observations = 3.966628

Cluster History

NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	ERSQ	CCC	Norm Min Dist	Tie
44	OB14	OB15	2	0.0001	1.00	.	.	0.0797	
43	OB2	OB4	2	0.0002	1.00	.	.	0.0996	
42	OB6	OB8	2	0.0003	.999	.	.	0.1216	
41	OB12	OB13	2	0.0004	.999	.	.	0.1494	
40	CL43	OB5	3	0.0006	.998	.	.	0.1606	
39	OB37	OB44	2	0.0005	.998	.	.	0.1655	
38	OB23	OB24	2	0.0007	.997	.	.	0.1913	
37	CL42	OB21	3	0.0010	.996	.	.	0.1917	
36	CL40	OB9	4	0.0016	.995	.	.	0.1953	
35	CL37	OB20	4	0.0023	.992	.	.	0.2069	
34	CL36	CL44	6	0.0033	.989	.	.	0.2106	
33	CL34	OB17	7	0.0016	.987	.	.	0.2128	
32	OB41	OB43	2	0.0010	.986	.	.	0.2202	

Cluster History

NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	ERSQ	CCC	Norm Min Dist	Tie
31	CL33	CL35	11	0.0129	.973	.	.	0.2213	
30	OB34	OB35	2	0.0010	.972	.	.	0.2292	
29	OB36	CL32	3	0.0027	.970	.	.	0.2309	
28	OB1	OB3	2	0.0011	.969	.	.	0.2365	
27	CL31	OB19	12	0.0066	.962	.	.	0.239	
26	OB42	OB45	2	0.0014	.961	.	.	0.2666	
25	OB38	CL26	3	0.0016	.959	.	.	0.2704	
24	CL29	CL25	6	0.0118	.947	.	.	0.3002	
23	OB29	OB30	2	0.0021	.945	.	.	0.3252	
22	CL28	CL27	14	0.0082	.937	.	.	0.3271	
21	CL38	OB26	3	0.0034	.934	.	.	0.3418	
20	CL22	OB18	15	0.0145	.919	.	.	0.3726	
19	CL24	OB39	7	0.0142	.905	.	.	0.3737	
18	OB25	OB27	2	0.0031	.902	.	.	0.3929	
17	OB7	CL41	3	0.0043	.897	.	.	0.4008	
16	CL21	OB33	4	0.0056	.892	.	.	0.4041	
15	CL16	CL23	6	0.0144	.877	.	.	0.4124	
14	CL15	CL30	8	0.0155	.862	.	.	0.4403	
13	CL19	OB40	8	0.0121	.850	.	.	0.4439	
12	CL20	OB10	16	0.0127	.837	.	.	0.4459	
11	CL12	CL17	19	0.0228	.814	.	.	0.4464	
10	CL11	OB11	20	0.0237	.791	.	.	0.4703	
9	CL14	OB31	9	0.0098	.781	.786	-.26	0.4711	
8	OB22	OB32	2	0.0047	.776	.764	0.59	0.4849	
7	CL8	CL9	11	0.0219	.754	.739	0.72	0.4868	
6	CL7	OB28	12	0.0098	.744	.708	1.65	0.5	
5	CL13	CL39	10	0.0247	.720	.670	2.18	0.5196	
4	CL6	CL18	14	0.0327	.687	.619	2.87	0.5203	
3	CL10	OB16	21	0.0236	.663	.543	4.25	0.5562	
2	CL3	CL4	35	0.2890	.374	.372	0.07	0.7492	
1	CL2	CL5	45	0.3744	.000	.000	0.00	0.7507	

Variable	Mean	Std Dev	Skewness	Kurtosis	Bimodality
al2o3	1.4673	0.2533	-0.4026	-0.8763	0.4967
fe2o3	0.6693	0.2794	-1.0097	-0.3827	0.7128
mgo	0.3713	0.2601	0.9876	0.0891	0.5977
cao	0.2989	0.2638	0.7834	-0.0871	0.5158
na2o	0.3049	0.2232	0.9509	1.0509	0.4463
k2o	1.0189	0.2720	0.1279	-0.6614	0.3979
tio2	1.1240	0.2306	0.4038	0.3180	0.3291
mno	0.4356	0.2887	0.0916	-0.5279	0.3751
bao	1.1789	0.2135	-0.2723	0.1408	0.3200

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.19785459	1.67273872	0.4664	0.4664
2	2.52511587	1.64701600	0.2806	0.7470
3	0.87809987	0.42173477	0.0976	0.8446
4	0.45636510	0.07443216	0.0507	0.8953
5	0.38193294	0.11083473	0.0424	0.9377
6	0.27109821	0.15433632	0.0301	0.9678
7	0.11676189	0.02468356	0.0130	0.9808
8	0.09207833	0.01138513	0.0102	0.9910
9	0.08069320		0.0090	1.0000

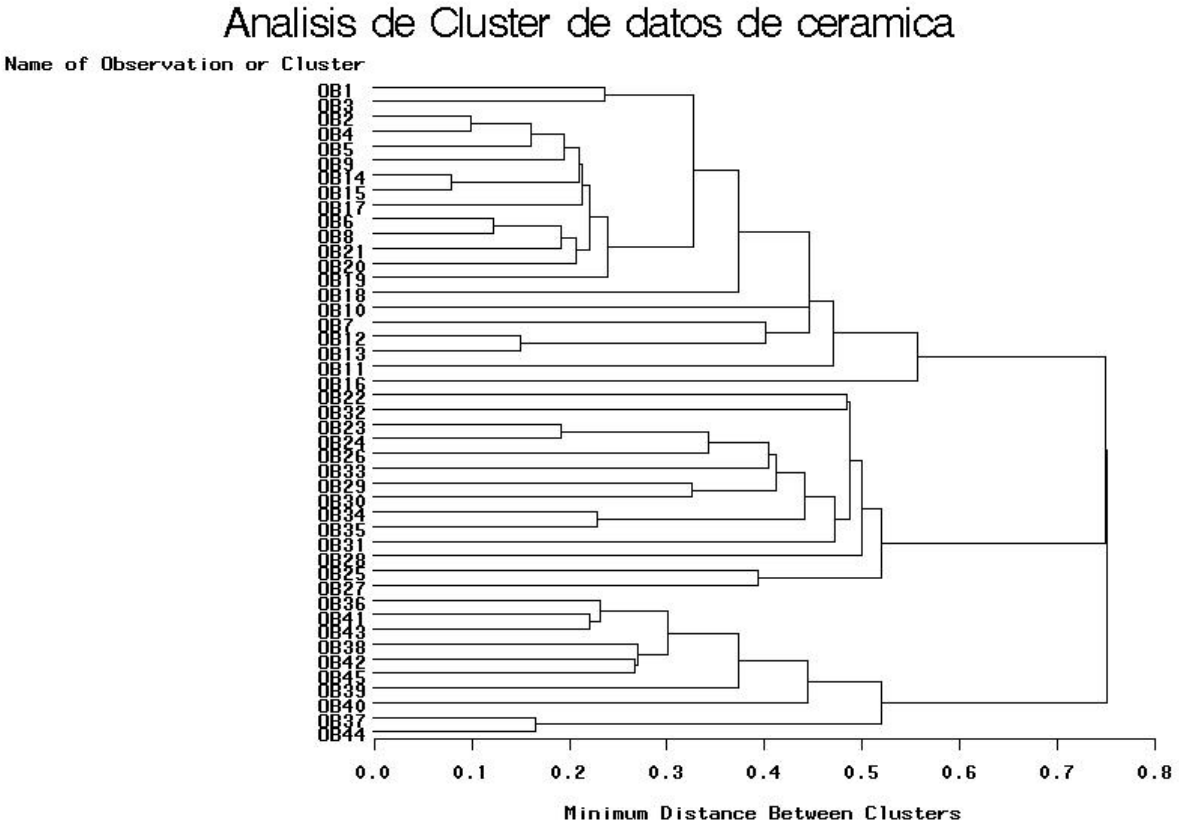
The data have been standardized to mean 0 and variance 1
Root-Mean-Square Total-Sample Standard Deviation = 1
Mean Distance Between Observations = 3.966628

Cluster History

NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	ERSQ	CCC	Norm Min Dist	Tie
44	OB14	OB15	2	0.0001	1.00	.	.	0.0797	
43	OB2	OB4	2	0.0002	1.00	.	.	0.0996	
42	OB6	OB8	2	0.0003	.999	.	.	0.1216	
41	OB12	OB13	2	0.0004	.999	.	.	0.1494	
40	CL43	OB5	3	0.0006	.998	.	.	0.1606	
39	OB37	OB44	2	0.0005	.998	.	.	0.1655	
38	OB23	OB24	2	0.0007	.997	.	.	0.1913	
37	CL42	OB21	3	0.0010	.996	.	.	0.1917	
36	CL40	OB9	4	0.0016	.995	.	.	0.1953	
35	CL37	OB20	4	0.0023	.992	.	.	0.2069	
34	CL36	CL44	6	0.0033	.989	.	.	0.2106	
33	CL34	OB17	7	0.0016	.987	.	.	0.2128	
32	OB41	OB43	2	0.0010	.986	.	.	0.2202	

Cluster History

NCL	--Clusters Joined--		FREQ	SPRSQ	RSQ	ERSQ	CCC	Norm Min Dist	Tie
31	CL33	CL35	11	0.0129	.973	.	.	0.2213	
30	OB34	OB35	2	0.0010	.972	.	.	0.2292	
29	OB36	CL32	3	0.0027	.970	.	.	0.2309	
28	OB1	OB3	2	0.0011	.969	.	.	0.2365	
27	CL31	OB19	12	0.0066	.962	.	.	0.239	
26	OB42	OB45	2	0.0014	.961	.	.	0.2666	
25	OB38	CL26	3	0.0016	.959	.	.	0.2704	
24	CL29	CL25	6	0.0118	.947	.	.	0.3002	
23	OB29	OB30	2	0.0021	.945	.	.	0.3252	
22	CL28	CL27	14	0.0082	.937	.	.	0.3271	
21	CL38	OB26	3	0.0034	.934	.	.	0.3418	
20	CL22	OB18	15	0.0145	.919	.	.	0.3726	
19	CL24	OB39	7	0.0142	.905	.	.	0.3737	
18	OB25	OB27	2	0.0031	.902	.	.	0.3929	
17	OB7	CL41	3	0.0043	.897	.	.	0.4008	
16	CL21	OB33	4	0.0056	.892	.	.	0.4041	
15	CL16	CL23	6	0.0144	.877	.	.	0.4124	
14	CL15	CL30	8	0.0155	.862	.	.	0.4403	
13	CL19	OB40	8	0.0121	.850	.	.	0.4439	
12	CL20	OB10	16	0.0127	.837	.	.	0.4459	
11	CL12	CL17	19	0.0228	.814	.	.	0.4464	
10	CL11	OB11	20	0.0237	.791	.	.	0.4703	
9	CL14	OB31	9	0.0098	.781	.786	-.26	0.4711	
8	OB22	OB32	2	0.0047	.776	.764	0.59	0.4849	
7	CL8	CL9	11	0.0219	.754	.739	0.72	0.4868	
6	CL7	OB28	12	0.0098	.744	.708	1.65	0.5	
5	CL13	CL39	10	0.0247	.720	.670	2.18	0.5196	
4	CL6	CL18	14	0.0327	.687	.619	2.87	0.5203	
3	CL10	OB16	21	0.0236	.663	.543	4.25	0.5562	
2	CL3	CL4	35	0.2890	.374	.372	0.07	0.7492	
1	CL2	CL5	45	0.3744	.000	.000	0.00	0.7507	


Analisis de Cluster de datos de ceramica
Name of Observation or Cluster
OB1
OB3
OB2
OB4
OB5
OB9
OB14
OB15
OB17
OB6
OB8
OB21
OB20
OB19
OB18
OB10
OB7
OB12
OB13
OB11
OB16
OB22
OB32
OB23
OB24
OB26
OB33
OB29
OB30
OB34
OB35
OB31
OB28
OB25
OB27
OB36
OB41
OB43
OB38
OB42
OB45
OB39
OB40
OB37
OB44
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Minimum Distance Between Clusters