Análisis Multivariante

Teresa Villagarcía

1. Introducción

En estudios exploratorios en los que se tiene una gran cantidad de datos, los analistas suelen utilizar muchas variables y observaciones. El análisis subsiguiente es muy complejo, ya que las variables están muy relacionadas entre sí, y el número de observaciones es tan grande que, en ocasiones, es difícil utilizar técnicas descriptivas para reducir la dimensión de los datos.

Las técnicas de análisis multivariante son muy adecuadas para poder comprender la estructura de estas grandes masas de datos.

El análisis multivariante ha experimentado un fortisimo crecimiento en las últimas décadas debido, básicamente, al desarrollo de nuevas herramientas informáticas. Es importante indicar que estas técnicas, de uso intensivo de ordenador únicamente pueden utilizarse con apoyo informático. Consecuentemente esta documentación está elaborada para ser utilizada disponiendo de medios informáticos adecuados. Cualquier paquete estadístico tiene, hoy día, implementadas las más importantes herramientas de análisis multivariante Concretamente Statgraphics o SPSS -que son dos programas muy populares en el mundo empresarial- son perfectamente adecuados en este contexto.

Se van a estudiar tres técnicas diferentes:

1. Componentes Principales
2. Análisis Factorial
3. Análisis Cluster
   a. Clusters de Observaciones
   b. Clusters de Variables

Todas las técnicas van a ser introducidas siguiendo el mismo esquema. En primer lugar se van a explicar las características fundamentales de cada método. A continuación se presentarán algunos ejemplos sencillos y, finalmente se indicarán aplicaciones específicas en el campo de la Calidad.

2. Componentes Principales.

Uno de los problemas principales que afectan al estudio de grandes masas de datos, es que las variables explicativas suelen ser muy parecidas: contienen información equivalente. En efecto, cuando un investigador reúne información sobre cualquier fenómeno tiende a incorporar diversas variables que son semejantes pero no iguales, de modo que el análisis resulta complejo y surgen graves problemas de colinealidad entre las variables X. Así, por ejemplo en regresión múltiple, cuando existe multicolinealidad, no queda más remedio que eliminar algunas variables. Pero eso implica una pequeña pérdida de información.

Una forma alternativa de abordar el problema de la información recurrente es construir un conjunto de nuevas varibles Z menor que el inicial X y que contengan la máxima información posible. Componentes Principales parte de la idea de transformar las variables originales $X = X_1, X_2, \cdots, X_k$ en un conjunto de variables $Z = Z_1, Z_2, \cdots, Z_p$ tales que $p < k$. Es decir *vamos a sustituir las variables iniciales altamente correladas entre sí, por un nuevo conjunto menor de*

variables no correladas.

La idea es, por tanto, sustituir el conjunto de variable inicial X por otro conjunto más pequeño Z. Para ello vamos a definir las nuevas varibles Z como combinaciones lineales de las viejas variables X:

$$\begin{aligned} Z_1 &= v_{11}X_1 + v_{12}X_2 + \cdots + v_{1k}X_k \\ Z_2 &= v_{21}X_1 + v_{22}X_2 + \cdots + v_{2k}X_k \\ &\cdots \\ Z_p &= v_{p1}X_1 + v_{p2}X_2 + \cdots + v_{pk}X_k \end{aligned}$$

Es decir las nuevas variables serán una combinación de las variables existentes $X_1, X_2, \cdots, X_k$. A estas nuevas variables se les denomina Componentes Principales y se calculan como los autovectores de la matriz $X'X$ de observaciones. Estos cálculos -complicados- los realiza cualquier paquete estadístico.

Lo más importante en este tipo de análisis es que el número de variables Z sea menor que el número original de variables X. De hecho el objetivo de estas técnicas es, precisamente, reducir la dimensionalidad de los datos.

Surge pues el problema de determinar cuantas variables Z o componentes debemos incluir en el análisis. Para resolver este problema nos basaremos en las siguientes ideas:

1. La información total contenida en las k variables X va a ser explicada por un conjunto de variables Z que es menor.
2. Por tanto *NO TODA* la información contenida en las k variables X va estar contenida en las variables Z. Perderemos algo de información.
3. Debemos elegir el número de variables Z adecuado teniendo en cuenta que VAMOS a perder algo de información.

Para definir el número óptimo de Componentes Z utilizaremos la cantidad de varianza que explica cada componente. Así, si tenemos k variables X , el 100% de la varianza (Información) contenida por esas variables sólo puede ser explicada por k variables Z. El coste de reducir el número de variables es, precisamente, perder ”algo” de información.

Cuando en un programa (por ejemplo Statgraphics) accedemos a componentes principales, el programa pide que se le introduzcan las variables X que van a ser analizadas. Una vez introducidas se obtiene la siguiente información:

1. **Tabla de Autovalores** y sus varianzas asociadas.
2. **Valores de los Componentes**
3. **Diversos Gráficos**

La **Tabla de Autovalores** (**1**) proporciona la cantidad de información que contiene cada componente y las proporciones acumuladas.

Los **Valores de los Componentes** (**2**) son las ecuaciones que definen cada componente Z en función de las variables originales X. Es decir las ecuaciones:

$$\begin{aligned} Z_1 &= v_{11}X_1 + v_{12}X_2 + \cdots + v_{1k}X_k \\ Z_2 &= v_{21}X_1 + v_{22}X_2 + \cdots + v_{2k}X_k \\ &\cdots \\ Z_p &= v_{p1}X_1 + v_{p2}X_2 + \cdots + v_{pk}X_k \end{aligned}$$

Veamoslo con un ejemplo.

Ejemplo:

Los datos correspondientes al número de especies endémicas en cada una de las islas Galápagos son altamente colineales. Vamos a estudiar las regresiones del número de endémicas en función del resto de las variables:

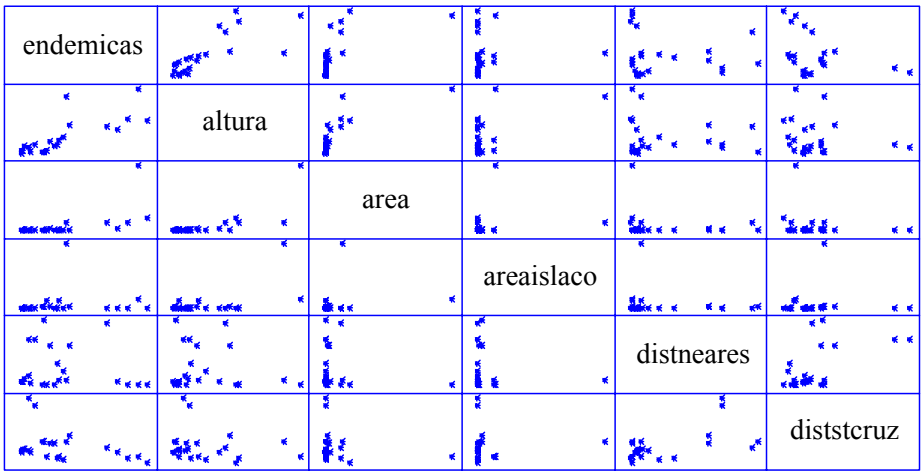


Figura 1: Gráficos de dispersión de los datos de las islas Galápagos.

Como puede observarse en el gráfico, la relación entre las variables es claramente no lineal. Para linealizar vamos a transformar todas las variables a Logaritmos.

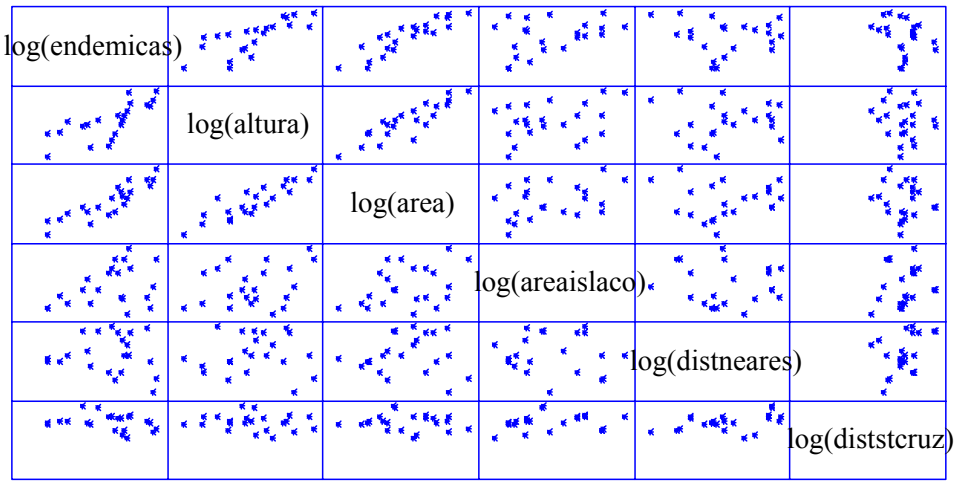


Figura 2: Datos de las islas Galápagos trnsformados.

Tras la transformación logarítmica la relación ha mejorado mucho. Veamos las regresiones con los estadísticos t. Inicialmente entre Area y Altura. Como puede observarse las variables son colineales, ya que la altura que es significativa en la regresión simple, deja de serlo en la múltiple. Esto es totalmente previsible a la vista del gráfico entre Area y Altura.

- 1. $\log(Endemicas) = 2.19 + 0.31 \log(Area)$
  $R^2 = 79.9 \quad (10.37)$
- 2. $\log(Endemicas) = -1.43 + 0.78 \log(Altura)$
  $R^2 = 57.5 \quad (5.33)$
- 3. $\log(Endemicas) = 3.27 + 0.38 \log(Area) - 0.23 \log(Altura)$
  $R^2 = 81.5 \quad (5.09) \quad (-1.01)$

Si introducimos variables adicionales:

$$\log(Endemicas) = 4.56 + 0.37 \log(Area) - 0.26 \log(Altura)$$
$$(4.77) \quad (-1.18)$$

$$+0.03 \log(areaislaco) - 0.01 \log(distnearest) - 0.27 \log(diststcruz)$$
$$(0.73) \quad (-0.15) \quad (-1.63)$$

Vemos que ni el Area de la Isla contigua (*areaislaco*), ni su distancia (*distnearest*) ni la distancia a la isla más biodiversa, Santa Cruz, (*diststcruz*) son significativas (ni en las regresiones simples ni en la múltiple) por lo que son variables NO SIGNIFICATIVAS.

Todas estas regresiones indican que existe una fuerte colinealidad entre las variables Area y Altura y que tenemos 3 variables no significativas. Tenemos 5 variables y vamos a construir un nuevo conjunto de variables de menor dimensión:

El Análisis de Componentes Principales lo hace cualquier programa estadístico (Statgraphics, SPSS...). El ejemplo se va a hacer con Statgraphics .

Las variables que queremos reducir son:

1. **log**(**Area**)
2. **log**(**Altura**)
3. **log**(**areaislaco**)

4. **log(distnearest)**
5. **log(diststcruz)**

Estas son las 5 variables *X*. Ahora vamos a construir las variables *Z*. Para ello hay que calcular los autovalores y autovectores de la Matriz $X'X$. Las componentes corresponden a los autovectores. El programa proporciona la **Tabla de Autovalores**:

Tabla de Autovalores

Componente	Eigenvalue	% de Varianza	% Acumulado
1	2,1187	42,374	42,374
2	1,4638	29,277	71,651
3	0,8768	17,537	89,187
4	0,4314	8,630	97,817
5	0,1091	2,183	100,000

Con la tabla de autovalores existen dos criterios para elegir el número de factores que se van a considerar:

1. Elegir aquellos factores con autovalor mayor de 1 (Autovalor=Eigenvalue)
2. Estudiar la tabla de varianzas explicadas por cada autovalor (tal como se hacía en Componentes Principales) y elegir un número razonable de factores.

En la tabla de autovalores se obtienen los porcentajes de varianza explicados por cada componente. En nuestro caso la primera componente explica el 42% de la varianza. La segunda explica el 29.2% y la tercera el 17.5%. Entre esas tres componentes explican 89.187% de la varianza de las 5 variables originales. Si seguimos el segundo criterio de elección del número de componentes, se elegirían como mucho las tres primeras. Y quizás únicamente 2, pues no hay que olvidar que queremos REDUCIR el número de variables.

Si utilizamos las tres primeras componentes para resumir la información, sólo dejamos de explicar un 11% de varianza de las variables originales.

Si utilizamos el criterio de Autovalor mayor que 1 elegiríamos las dos primeras componentes. Vamos a estudiar el significado de las componentes obtenidas.

Las nuevas variables *Z* tienen las siguientes expresiones:

- $C_1 = 0,64.\log(altura) \;\; +0,65.\log(area) \;\; +0,39.\log(areaislaco)$
  $+0,01.\log(distneares) \;\; -0,1\log(diststcruz)$

- $C_2 = -0,005.\log(altura) \;\; +0,029\log(area) \;\; +0,16.\log(areaislaco)$
  $+0,68.\log(distneares) \;\; +0,71\log(diststcruz)$

- $C_3 = 0,26.\log(altura) \quad +0,19\log(area) \quad . -0,79\log(areaislaco)$
$+0,44\log(distneares) \quad -0,25\log(diststcruz)$

Es decir que la primera componente C_1 es prácticamente una media de área, altura y ponderando menos, área de la isla contígua. Esta Componente está midiendo las características físicas de la isla.

La segunda no da peso a Area, Altura y Area de la isla contigua y si da peso a Distancia a la isla mas próxima y a Santa Cruz. Esta Componente mide claramente la distancia a las zonas "productoras de especies endémicas". Así, una isla con mucha distancia a Sta Cruz y a la isla mas próxima dará más valor en C_2.

Finalmente la tercera, C_3 da peso al Area de la isla contigua y su distancia.

Los programas estadísticos proporcionan también una serie de gráficos para entender mejos el significado de las componentes. Estos gráficos muestran los pesos de las componentes de dos en dos o de tres en tres.

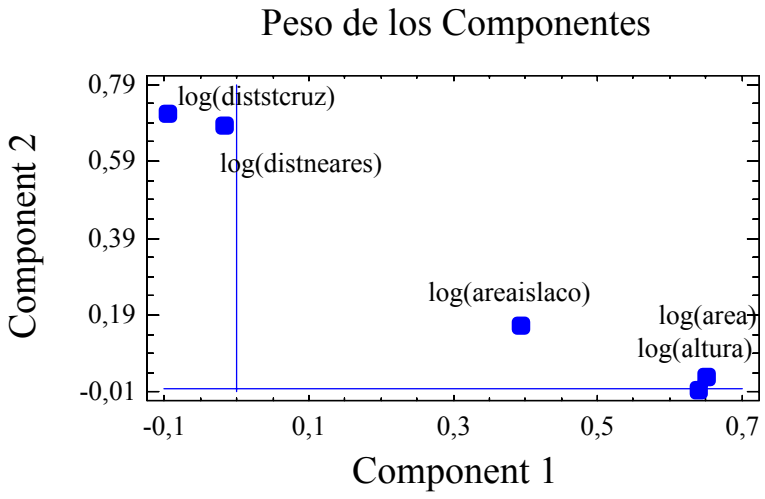


Figura 3: **Gráfico de los pesos de las Componentes 1 y 2. Se ve que Área y**

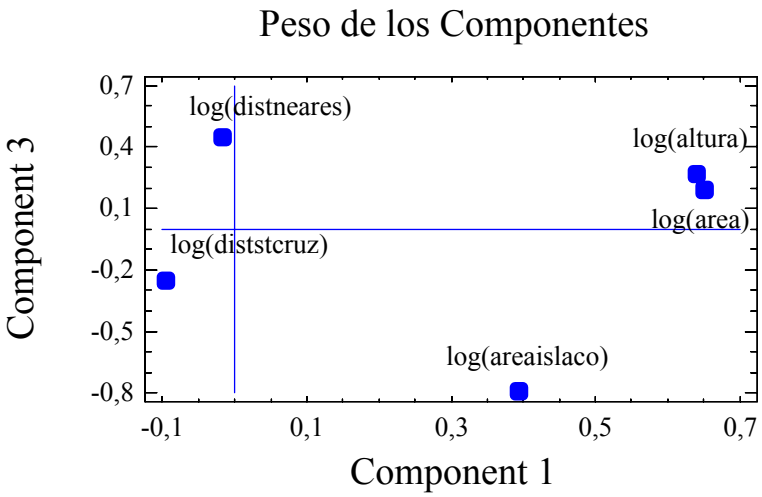


Figura 4: **Peso de las componentes 1 y 3**.

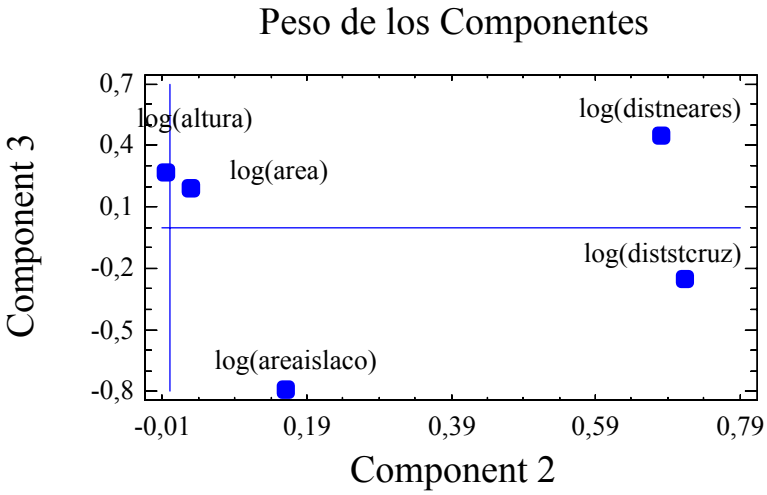


Figura 5: **Peso de las componentes 2 y 3**.

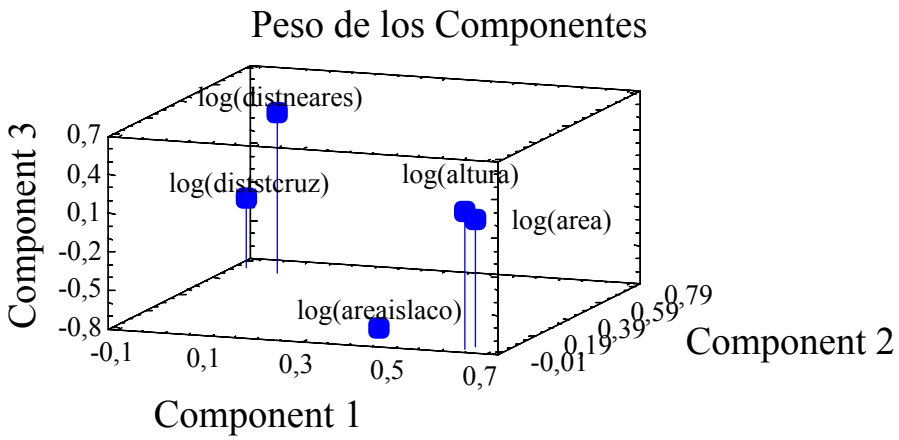


Figura 6: **Pesos de las tres componentes**.

Las conclusiones serían

- La primera componente mide las características físicas de la isla (Area y Altura) y en menor medida las de la isla vecina.
- La segunda mide la distancia a las islas "productoras de especies": la más cercana y la más biodiversa.
- La tercera (y de menor en importancia) mide básicamente el área y la distancia de la isla más cercana.

Si hacemos ahora la regresión entre el número de especies endémicas y la primera componente

$$\log(Endemicas) = 2,84 + 0,67.C_1$$
$$R^2 = 71.2 \qquad (7,1)$$

Aunque explique menos que el Area, en esta componente están integradas el Área y la Altura que era imposible integrar en un mismo modelo por los problemas derivados de la colinealidad. El analista deberá decidir si le interesa introducir las dos variables, o quedarse con la regresión explicada sólo por el área:

$$\log(Endemicas) = 2.19 + 0.31 \log(Area)$$
$$R^2 = 79.9 \qquad (10.37)$$

Si realizásemos la regresión entre Endémicas y las dos primera componentes obtendríamos:

$$\begin{array}{lll} \log(Endemicas) = 2.84 & +0,67C_1 & -.12C_2 \\ R^2 = 73.18 & (7.1) & (-1.06) \end{array}$$

Es decir que la segunda componente no es significativa. Pero es lógico pues la segunda componente mide la distancia a las islas "productoras de especies": la más cercana y la más biodiversa. Y en regresión ya habíamos visto que esas variables no eran significativas. La construcción de Componentes principales nos ha permitido "Entender" mejor el problema de las variables asociadas a las islas Galápagos. Ahora sabemos que esas variables están clasificadas en tres grupos definidos por las tres componentes: Las caracerísticas físicas de las islas y su vecina, la distancia a los centros productores de especies y las características de la isla vecina. La manera de obtener inforamción con estas componentes (Hacer regresiones con componentes o sin ellas) es decisión del Analista. Cualquier camino puede ser válido.

3. Análisis Factorial.

Como en Componentes Principales, el propósoto del Análisis Factorial (AF) es describir la información contenida por las variables originales X en función de un número menor de variables F. La diferencia es que en el AF, se especifica un número reducido de Factores Comunes. Y todas las correlaciones se van a explicar por esos factores comunes. Lo que quede sin explicar será debido a unos factores únicos o errores incorrelados entre ellos.

El origen del AF proviene del intento de los psicólogos de entender los factores determinantes de la inteligencia humana. Las variables observadas X eran los resultados de los diferentes tests de inteligencia. Se pensaba que estos resultados dependían muy directamente de unos Factores desconocidos que hacían que una persona fuera más hábil en determinadas pruebas.

Los factores resultaron ser lo que posteriormente se denominó Inteligencia Verbal, Matemática, Espacial etc. Así, los resultados que se obtenían en la pruebas de matemáticas estaban muy determinados por la Inteligencia Matemática.

3.1 El Modelo.

En el modelo de análisis factorial cada variable X será combinación de los factores F y de los errores U. El modelo quedará:

$$\begin{array}{lllll} X_1 = & a_{11}F_1 & +a_{12}F_2 & +\cdots + & a_{1k}F_k + U_1 \\ X_2 = & a_{21}F_1 & +a_{22}F_2 & +\cdots + & a_{2k}F_k + U_2 \\ & & \cdots & & \\ X_p = & a_{p1}F_1 & +a_{p2}F_2 & +\cdots + & a_{pk}F_k + U_P \end{array}$$

donde las variables originales (todas) se expresan en función de los factores comunes F y de unos errores U. En el caso de los test de inteligencia, las variables X representan los resultados

obtenidos en los tests. Los factores *F* serían los niveles de inteligencia Verbal...

Se puede demostrar que

1. La varianza de cada variable se expresa como la suma de la varianza debida a los factores *F* más la debida al error *U* (Inexplicada)
2. Cada variable *X* depende de los factores, que a su vez son componentes de las variables.

Ejemplo

Vamos a realizar un AF con los datos de conservación de las merluzas. Para ello vamos a utilizar el experimento CATAAM6 que son datos de cata de merluzas a las que se han aplicado 8 tratamientos aparte del testigo. Las variables recogen las valoraciones otorgadas por los catadores a las diferentes pescadillas que prueban. Cada catador evalúa diversos aspectos de la cata: Aspecto General, Color, Olor, Acuosidad, Jugosidad, Dureza, Firmeza y Flavor. Además se tienen los días transcurridos desde la captura del pescado y el catador que ha realizado la valoración.

Las correlaciones entre las variables son muy elevadas. La Figura 7 muestra un gráfico de dispersión múltiple de estas variables

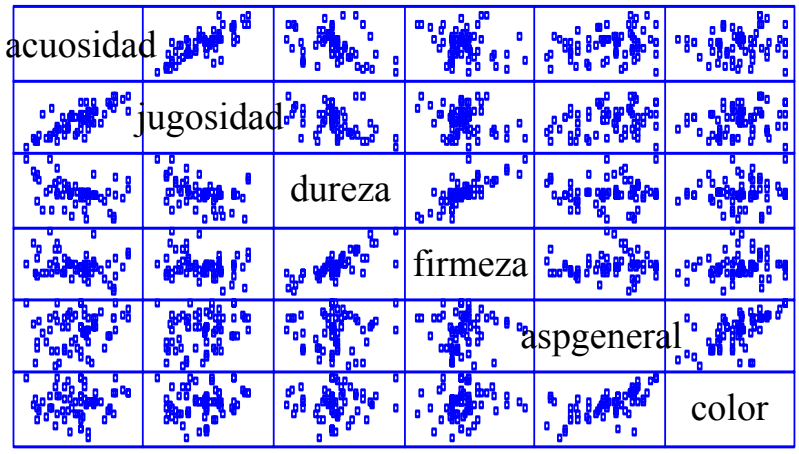


Figura 7: **Gráficos de dispersión de las variables de cata**.

Es evidente la relación entre Aspecto General-Color, Acuosidad-Jugosidad y Dureza-Firmeza.

Cuando realizamos el AF debemos introducir al ordenador las variables que queremos reducir footnote . En nuestro caso meteremos todas salvo Aceptabilidad General que consideramos una variable resumen. El AF parte del cálculo inicial de unos Componentes Principales que nos permiten decidir (mediante la cantidad de información que contienen las componentes) cuántos factores queremos tener en cuenta. Como ya se ha comentado existen dos criterios para elegir el número de factores que se van a considerar:

1. Elegir aquellos factores con autovalor mayor de 1 (Autovalor=Eigenvalue)
2. Estudiar la tabla de varianzas explicadas por cada autovalor (tal como se hacía en Componentes Principales) y elegir un número razonable de factores.

En el caso de los pescados, la Tabla de Autovalores es:

Tabla de Autovalores

Componente	Eigenvalue	% de Varianza	% Acumulado
1	3,13	31,34	31,34
2	1,95	19,57	50,9
3	1,09	10,95	61,87
4	1,04	10,4	72,28
5	0,85	8,52	80,80

La Tabla de hasta 10 factores muestra que hay 4 que tienen Eigenvalue mayor que uno. Estos cuatro factores cuentan por el 72.28 % de la varianza de los datos. Por tanto vamos a pensar en 4 Factores.

A continuación el programa ofrece la tabla de los pesos que los factores tienen para cada variable:

Pesos de factores para los cuatro primeros factores

Variable	F_1	F_2	F_3	F_4	Comunalidad
acuosidad	0,56	**0,65**	0,07	0,16	.77
aspgeneral	**0,75**	0,01	-0,01	-0,31	.67
catadores	-0,01	0,21	**0,64**	**0,54**	.76
color	**0,78**	-0,06	-0,01	-0,22	.67
dias	0,30	-0,35	**0,47**	**-0,49**	.68
dureza	0,33	**-0,81**	0,02	0,25	.83
firmeza	**0,51**	**-0,61**	0,16	0,37	.80
flavor	**0,74**	0,04	-0,06	0,01	.56
jugosidad	**0,66**	**0,55**	0,08	0,08	.77
olor	0,41	-0,07	**-0,648**	0,32	.69

Se ha marcado en negrita los pesos más importantes en cada factor. Esta tabla proporciona la siguiente información:

- F_1 está compuesto por la mayoría de las variables con la clara excepción de CATADORES.
- F_2 presenta un contraste entre **Acuosidad-Jugosidad** (positivos grandes) y **Dureza-Firmeza** (Negativos grandes).
- F_3 y F_4 tienen en cuenta **Catadores y Días**

El conjunto de los cuatro factores representa un elevado porcentaje de variablilidad de las variables.

Pero a su vez el AF ofrece una visión alternativa. Así, la variable Acuosidad queda explicada en un 77% por los cuatro factores. La variable Aspecto General en un 67%, Dureza en un 83% etc. Es decir las variables dependen de los factores. Por ejemplo Color es únicamente explicada por el primer factor ya que los pesos de los otros tres factores son muy pequeños. Lo mismo le ocurre a Flavor. Catadores es explicada sin embargo por los dos últimos factores. Lo mismo ocurre con días.

3.2 Rotaciones.

Una vez calculados los factores podemos rotarlos para conseguir unos pesos más sencillos. Para rotar los factores utilizaremos la denominada rotación Varimax que tiende a dar pesos mayores a las variables de mayor peso en el factor y menores a las de menor peso. De esta manera conseguimos mejores contrastes entre las variables. Las comunalidades se mantienen.

Tabla de Factores Rotados

Variable	F_1	F_2	F_3	F_4	**Comunalidad**
acuosidad	**0,74**	-0,25	0,34	-0,19	.77
aspgeneral	**0,73**	0,17	-0,24	0,21	.67
catadores	0,01	0,06	**0,86**	0,03	.76
color	**0,72**	0,27	-0,20	0,17	.67
dias	0,19	0,25	-0,11	**0,76**	.68
dureza	-0,02	**0,90**	-0,09	0,05	.83
firmeza	0,18	**0,87**	0,13	0,03	.80
flavor	**0,70**	0,25	-0,05	-0,04	.56
jugosidad	**0,81**	-0,16	0,25	-0,11	.77
olor	0,31	0,31	-0,28	**-0,65**	.69

Tras la rotación los factores son más claros:

- F_1 Media de **Acuosidad-Jugosidad**, **Color-Aspecto General**, **y Flavor**. Es decir un factor General: Representa la calidad general del pescado en la boca y al mirarlo en el plato.
- F_2 **Dureza-Firmeza**: Del mismo signo. Representa la textura tersa del pescado en la boca.
- F_3 **Catadores**
- F_4 Mezcla de **Días y Olor**: Días positivo y Olor negativo. Representa un factor de envejecimiento que se nota de forma rápida por el olor desagradable del pescado.

Así, nos ponen el pescado delante y hay tres factores que actúan:

1. Lo olemos: Factor 4
2. Lo miramos y lo probamos. Factor 1
3. Nos parece terso en la boca: Factor 2

La conclusión es que el proceso de cata se realiza en esos tres pasos. De nuevo el AF, como

Componentes Principales permite entender mejor la estructura de los datos.

4. Análisis Cluster.

El Análisis Cluster(AC) trata de encontar grupos de observaciones o variables semejates. Vamos a empezar por observaciones.

4.1 Clusters de observaciones (ACO).

El objetivo del ACO es encontrar grupos de observaciones semejantes. El conjunto de datos de la siguiente tabla es simulado. Se tienen tres variables X, Y, Z y los datos son los que se presentan en la Tabla:

TABLA DE DATOS

X	Y	Z	X	Y	Z
8,07	5,05	7,54	2,21	10,21	3,83
11,34	5,85	9,53	3,53	9,67	0,9
9,3	5,55	9,05	0,97	9,86	1,07
9,63	5,01	8,71	2,29	10,3	2,04
10,65	5,58	8,5	6,92	1,08	15,6
8,97	4,12	6,65	6,18	1,92	14,65
9,92	4,59	7,73	4,57	0,91	15,1
10,64	5,65	6,47	5,79	0,92	14,65
10,46	4,63	9,86	7,44	0,2	15,8
10,19	5,15	9,42	6,49	1,52	13,27
3,71	10,31	2,35	7,83	1,4	16,18
0,87	10,87	3,93	5,98	1,16	14,71
3,32	9,15	3,26	7,31	1,12	16,18
1,03	10,59	1,93	5,84	1,74	13,45
2,87	10,3	2,4	6,93	1,09	14,17
1,93	10,98	2,31	5,4	1,31	15,07
			4,85	0,93	14,93

Estos datos pueden representar las valoraciones dadas por una serie de individuos a ciertos atributos de calidad. Por ejemplo si **X**=”Limpieza” **Y**=”Rapidez” y **Z**=”Comida buena”, las variables podrían ser las valoraciones otorgadas por los clientes de una hamburguesería a tres atributos de calidad. En este caso, observaciones con **X**,**Y**,**Z** parecidas representan personas que valoran de forma semejante la cadena.

Estamos interesados en encontrar grupos de personas parecidas e identificarlos. Muchas veces estas identificaciones **No** serán las que se hayan pensado a priori. El Análisis Cluster permite segmentar el mercado de forma mas elástica y flexible que otros métodos tradicionales basados en características externas como Sexo, Edad o Nivel de Estudios.

Otra interesante utilización de este tipo de análisis es preguntar a los clientes por la importancia que conceden a diversos factores o atributos de los que depende la Calidad. Así, podremos presentar a nuestros clientes una lista de atributos y que ellos nos digan la importancia que les conceden. Los clientes que tienen una visión/expectativas parecidas presentarán valores semejantes de las distintas variables. Conociendo los diferentes grupos de clientes y sus expectativas, podremos saber en qué puntos se puede mejorar la calidad de un servicio con la máxima incidencia.

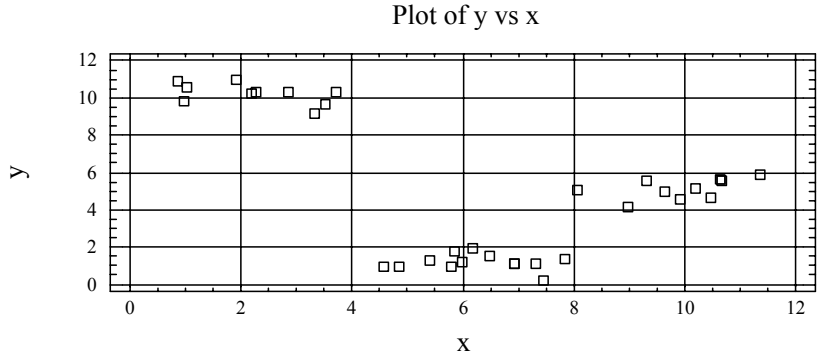
Si estudiamos las observaciones de la tabla, vemos que las observaciones 3 y 4 son muy semejantes:

Observación	X	Y	Z
3	9.3	5.55	9.05
4	9.63	5.01	8.71

y existen muchas otras observaciones semejantes. Evidentemente los indivíduos 3 y 4 tienen una valoración semejante de la Calidad de la hamburguesería. Para detectar indivíduos cercanos haremos un Análisis Cluster que se basa en:

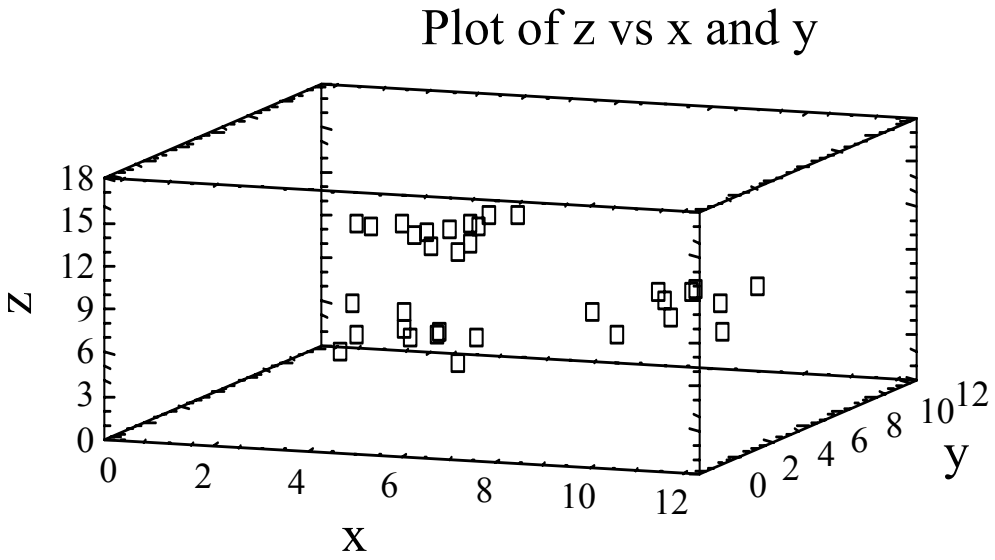
- Calcular las distancias entre observaciones
- Detectar aquellos grupos de observaciones con mínimas distancia y agruparlos en Clusters

En nuestro caso, las observaciones 3 y 4 están muy cercanas y pertenecerán al mismo grupo. El Gráfico que se presenta a continuación muestra los valores de **X e Y** para los datos de la Tabla:



Es evidente que hay tres grupos de observaciones claramente diferenciados. Cada uno de estos grupos tiene valores de **X e Y** semejantes.

Si observamos el gráfico tridimensional:



vemos que los tres grupos también están perfectamente claros. Estos tres grupos definen de forma natural tres Clusters

Cuando tenemos más de tres variables es mas difícil detectar los clusters pues no se pueden visualizar los puntos en más de tres dimensiones. Entonces es cuando vamos a utilizar el ACO con más provecho. El análisis cluster calcula la distancia entre los puntos y va clasificando las observaciones en uno u otro cluster según están las observaciones más cerca de uno u otro grupo.

Hay muchos métodos de clasificación que se pueden agrupar en :

- **Métodos jerárquicos**: La asignación de una observación a un cluster es irrevocable. Una vez que se incluye en un cluster **YA NO SE CAMBIA**
  - **Métodos Aglomerativos**: Hacen series de fusiones de las observaciones en los grupos fijados. El más característico es de el vecino más próximo que une las observaciones más cercanas en un primer cluster. Luego va realizando uniones de distancia mínima hasta que le queden tantos grupos como se le pide. Los métodos más utilizados son el de El Vecino más Próximo y más Lejano (Nearest Neighbor y Furthest Neighbor)
- **Métodos Partitivos**: El más conocido, que tiene bastantes variantes es de las K-medias (k-means) que consiste en encontrar grupos que minimizen la distancia al centro del cluster.

Todos los métodos se basan en ir agrupando las observaciones más cercanas. En cualquiera de ellos es preciso indicar al Programa en cuantos Clusters se quieren clasificar los datos y las variables que se van a utilizar en la clasificación (X_1,X_2,$\cdots X_k$).

El resultado es una tabla en la que se proporciona el número de elementos en cada grupo y los valores medios de cada una de las variables (**Centroides**). También se proporcionan diversos gráficos muy útiles.

Veamos un ejemplo con los datos simulados de la Tabla. Hemos elegido tres Clusters:

Cluster	Members	Percent
1	10	30.3
2	10	30.3
3	13	33.4

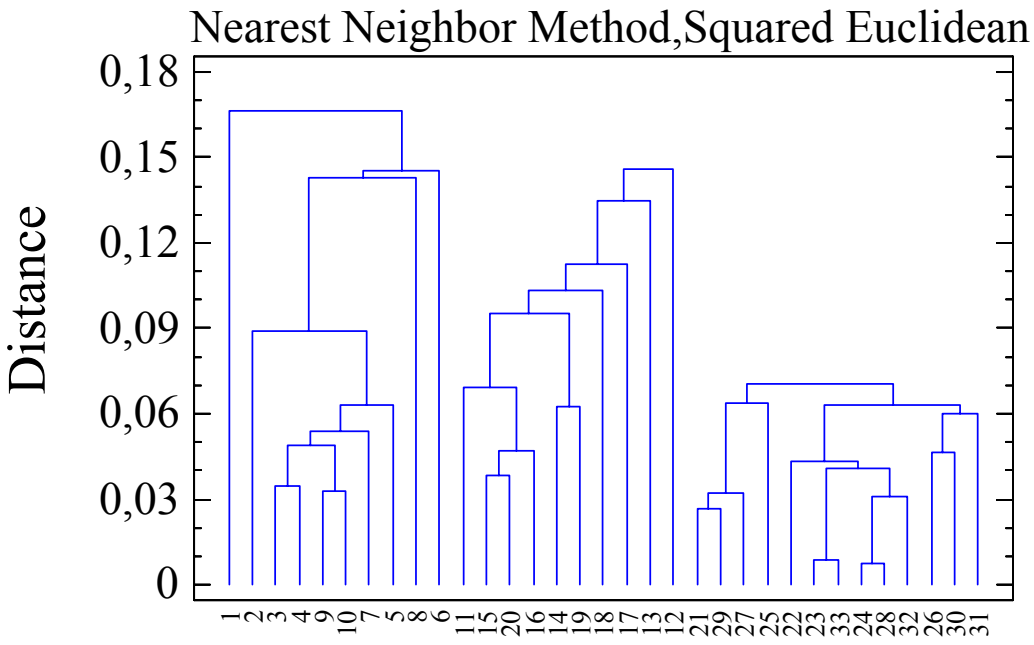
Centroides			
Cluster	**X**	**Y**	**Z**
1	9,91	5,11	8,34
2	2,27	10,2	2,40
3	6,27	1,17	14,9

Las tablas anteriores proporcionan los miembros de cada uno de los Clusters y los valores medios de las variables **X**,**Y y Z** por grupos . Así, si estos datos fueran de percepción de la calidad de la hamburguesería, el grupo 1 ofrecería unas medidas satisfactorias a las variables **X y Z**. Y el grupo 2 a la variable **Y**.

El ordenador proporciona además bastantes gráficos de interés.

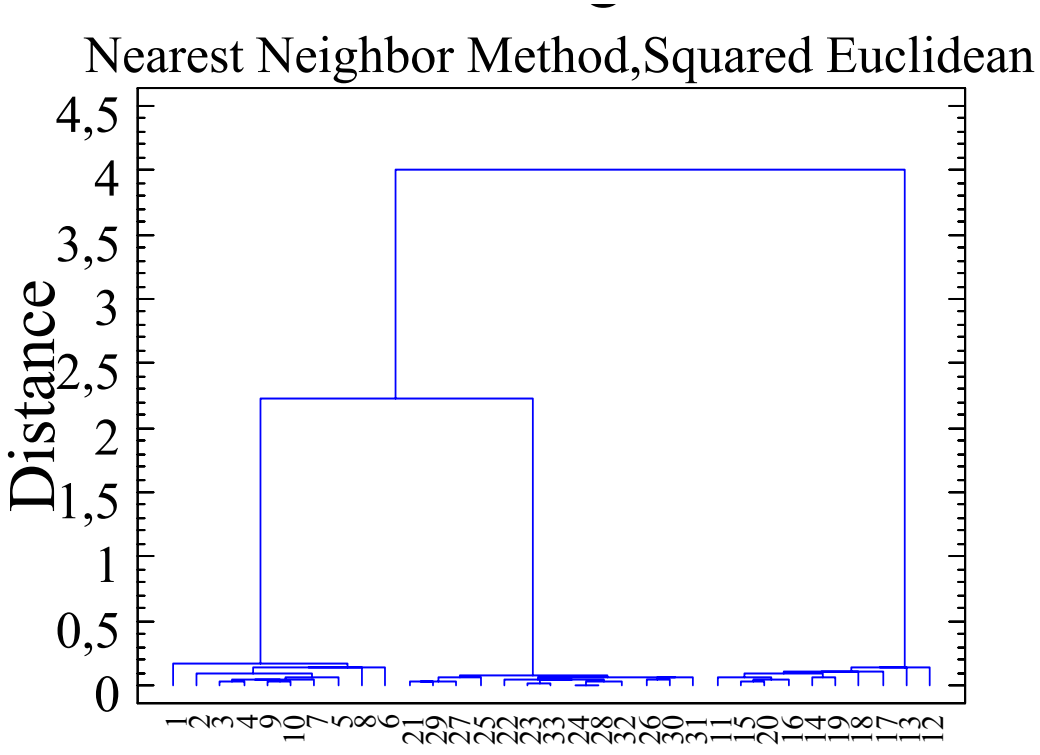
Dendograma

El dendograma es un gráfico que recoge la estructura de clasificación seguida. Se presenta en el siguiente Gráfico. En el eje vertical indica la distancia entre los grupos formados. En nuestro caso, la menor distancia se produce entre las observaciones 24 y 28 que forman el primer cluster. Luego entre las 23 y 33, 21 y 29, etc. Cuando empareja dos observaciones considera esas dos observaciones un grupo y calcula la distancia de todas las demás a ese grupo.



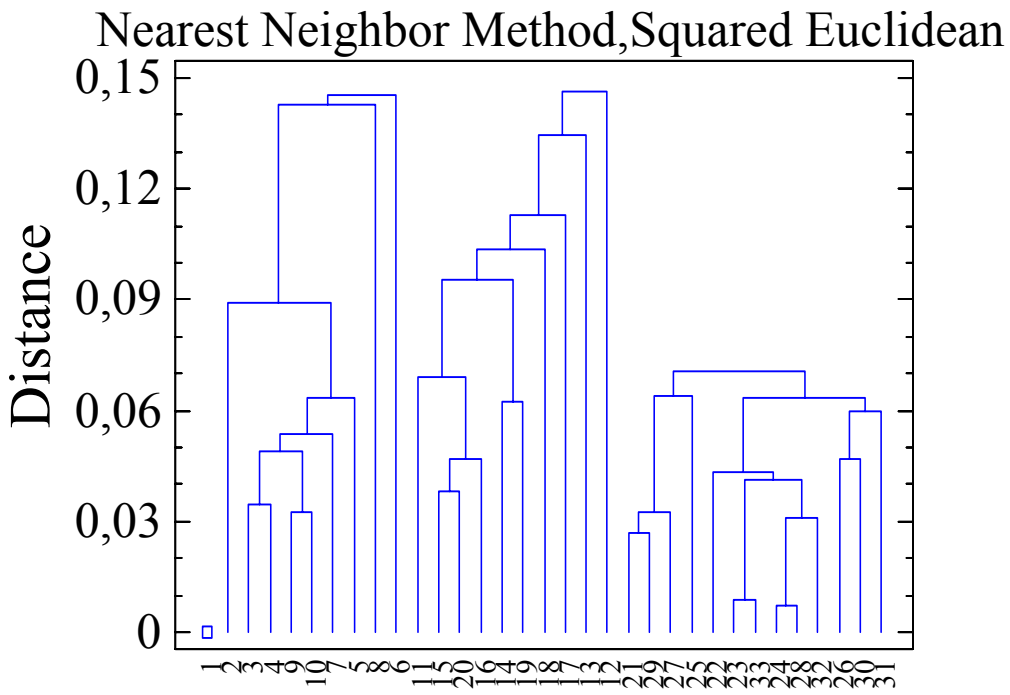
Dendograma para tres Clusters

En el gráfico se ven perfectamente los tres clusters. Si hubiéramos elegido un solo Cluster, el dendograma sería:



Dendograma para un solo Cluster. **La estructura de tres Clusters es muy evidente**.

y se ve perfectamente que realmente hay tres Clusters. Si hubiéramos clasificado en 4 grupos, el dendograma sería:



Dendograma para cuatro Clusters. **Obsérvese que la observación 1 sale sola en un**

En este caso la primera observación sale completamente aislada, formando un cluster ella sola. Esta observación estaba bastante bien integrada en uno de los tres clusters, así que no tiene mucho sentido aislarla.

Otros Gráficos

El ordenador proporciona diversos gráficos de dispersión para las variables clasificadas en Clusters:

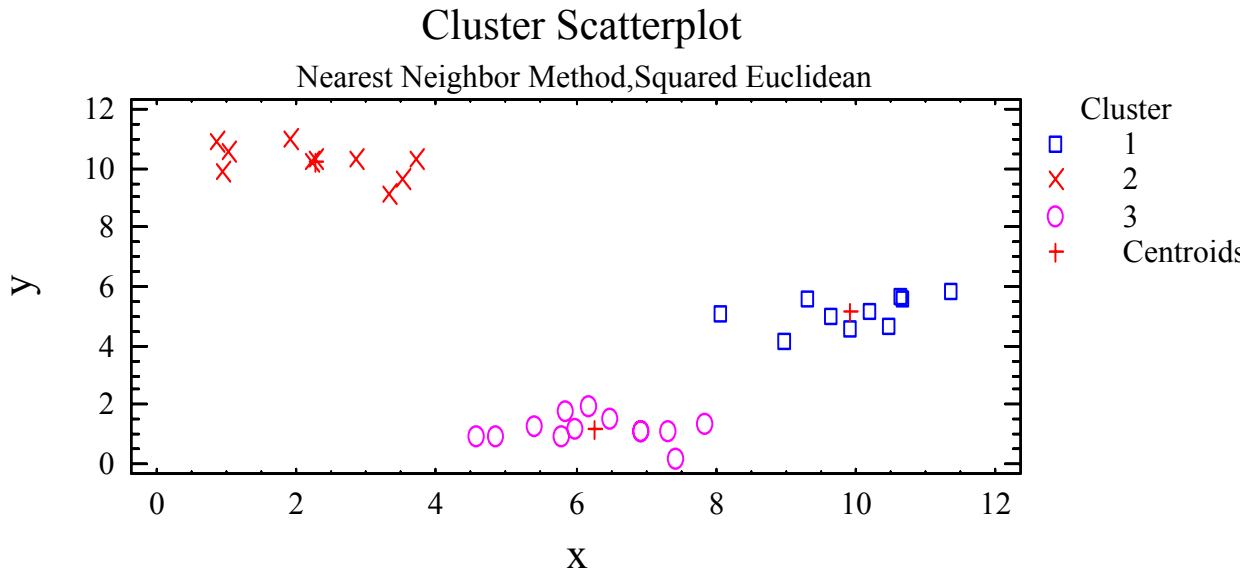


Gráfico de dispersión con tres clusters. **Centroides marcados con una cruz**.

en el gráfico se observan los valores de **X e Y** y los clusters perfectamente marcados. También aparecen los centroides.

Si hubiéramos elegido un solo cluster todos los datos están agrupados, pero el centroide está lejos de cualquier grupo.

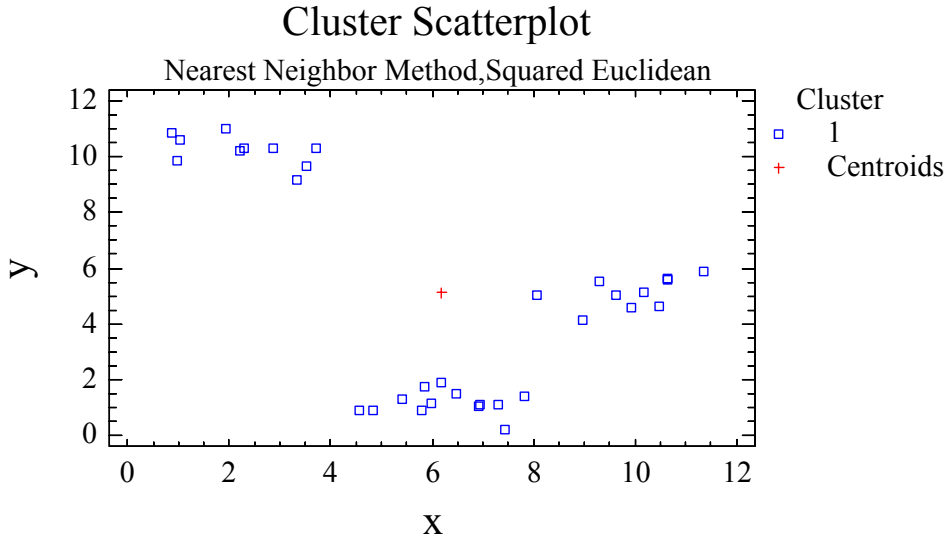


Gráfico de dispersión con un solo cluster. **Centroide no representativo**.

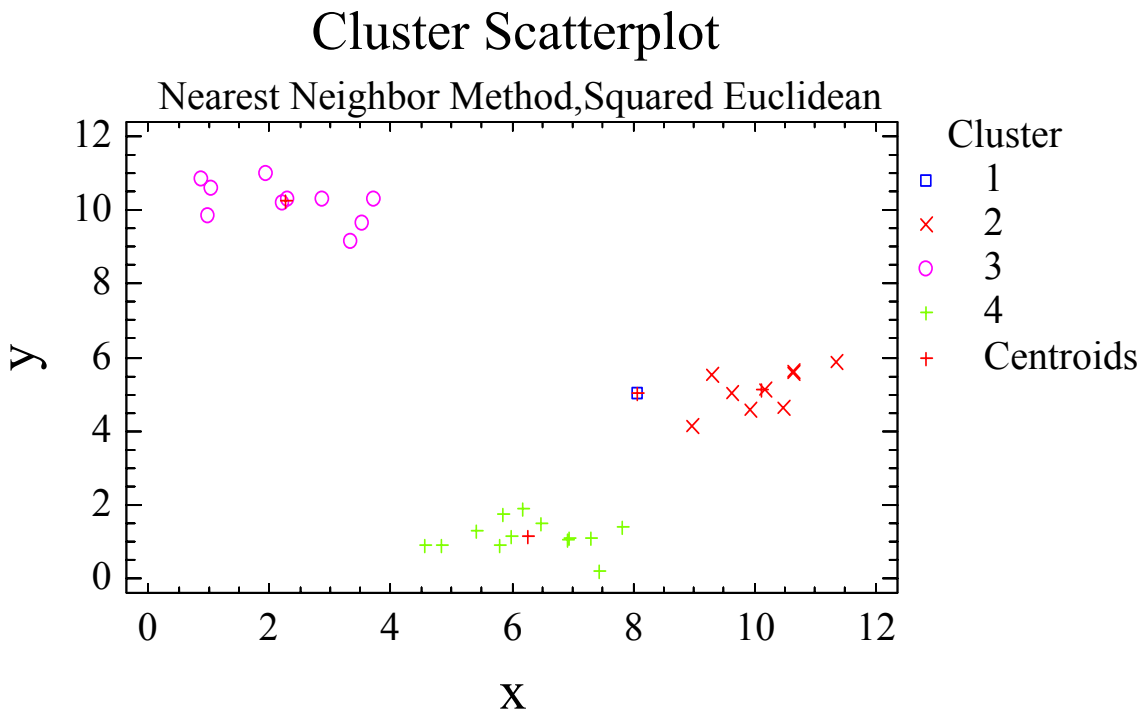


Gráfico de dispersión con 4 clusters. La primera obsevación forma un cluster aparte. Sería

Ejemplo 1:

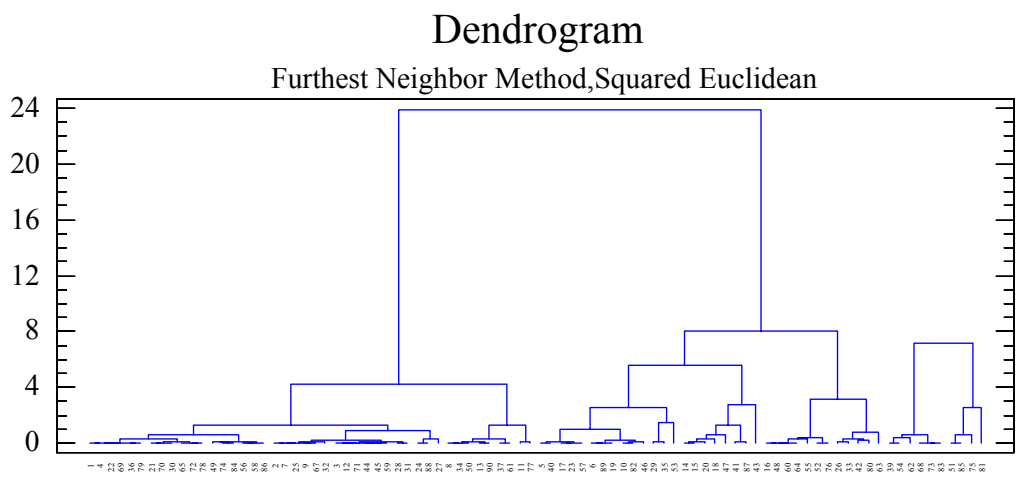
Vamos a estudiar los datos sobre características de automóviles del fichero CARDATA:SF de Statgraphics. Introducimos características técnicas de los vehículos: Número de Cilindros, Desplazamiento y Potencia.

Es importante probar divesos métodos y varios números de Cluster.

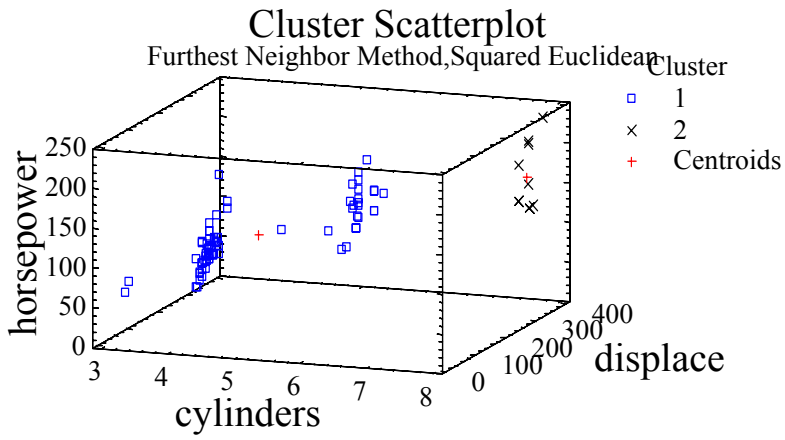
Empezamos con Furthest Neighbor:

Furthest Neighbor y 2 clusters:

El dendograma:



Y la representación de los grupos:



Cluster	Members	Percent
1	78	88,64
2	10	11,36

	Centroides		
Cluster	**cylinders**	**displace**	**horsepower**
1	4,65385	142,231	119,692
2	8,0	302,3	179,0

Evidentemente se ven más grupos. Repetimos el análisis para 3 y 4 grupos:

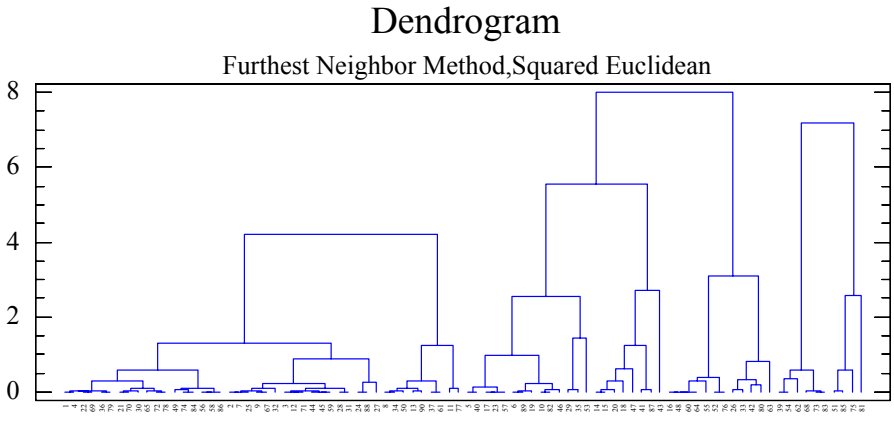
Furthest Neighbor para 3 grupos

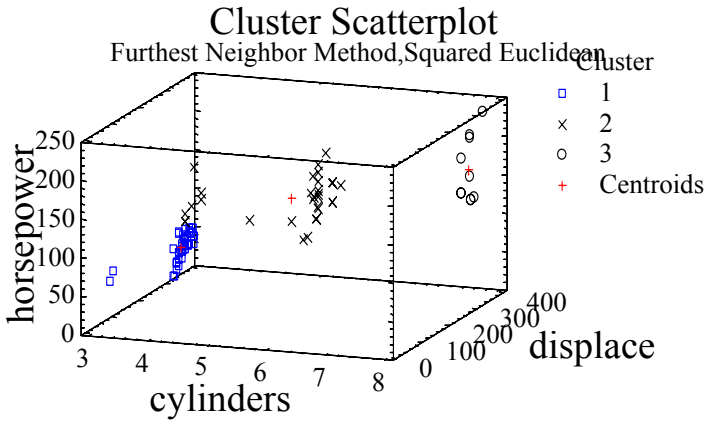
La tabla de Centroides:

```
Cluster    Members    Percent
----------------------------
1           44          50,00
2           34          38,64
3           10          11,36
----------------------------

Centroids

Cluster    cylinders     displace      hors
-----------------------------------------
1          3,95455       115,818       93,
2          5,55882       176,412       153,
```





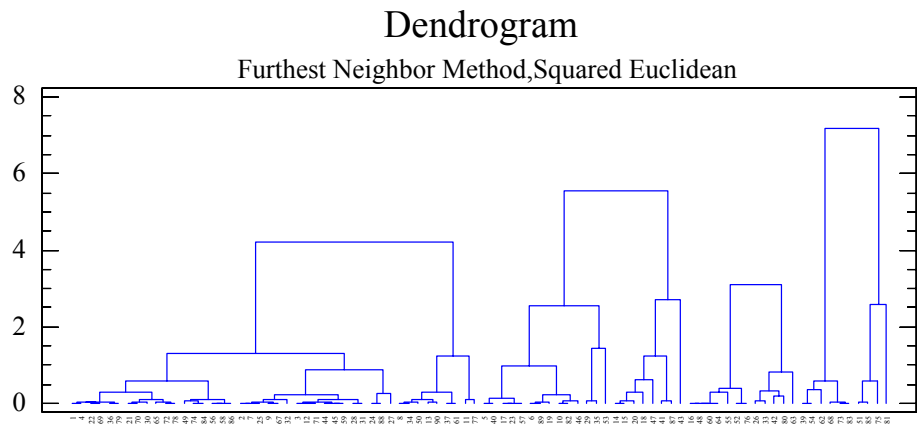
Furthest Neighbor con 4 grupos

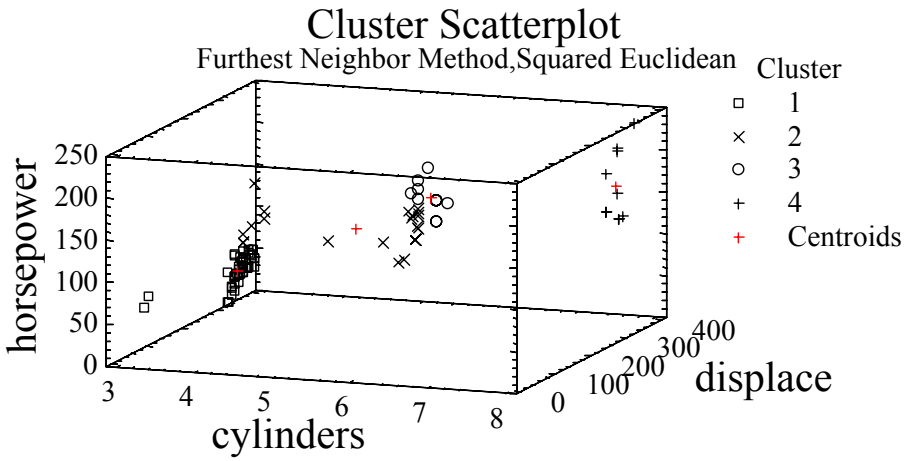
Tabla de Centroides:

Cluster	Members	Percent
1	44	50,00
2	22	25,00
3	12	13,64
4	10	11,36

Centroids

Cluster	cylinders	displace	horsepower
1	3,95455	115,818	93,7045
2	5,31818	155,773	143,409
3	6,0	214,25	171,5
4	8,0	302,3	179,0





En nuestro caso la clasificación en 4 grupos parece adecuada: Si se hace un análisis para identificar los modelos se verá que cada cluster agrupa tipos bastante claros:

Cluster 1:
44 coches de 3 y 4 cilindrros y en promedio 93CV. Es decir coches de la gama baja.

Cluster 2:
22 Coches de 4 y 6 cilindros de gama más alta.

Cluster 3:
12 coches de 6 cilindros de gama alta (Mercedes, BMW...)

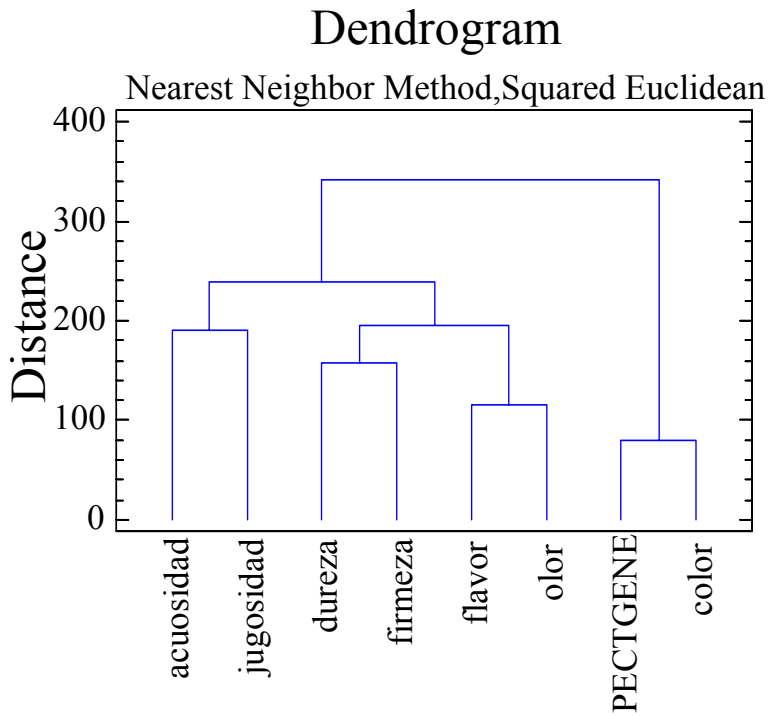
Cluster 4:

10 coches Americanos de gama alta.

Cluster de variables.

El Cluster de variables agrupa variables parecidas. Nos permite por tanto detectar grupos de variables semejantes. Es una herramienta más a la hora de detectar variables que miden lo mismo.

El dendograma siguiente representa la distancia y agrupación de las variables de cata:



En él se observan las mismas conclusiones que hemos obtenido de otros análisis estadísticos.