# Segmentación de mercados: proceso

*Jordi López Sintas*

*9 de enero de 2015*

## Modelos de segmentación: proceso

1. El procedimiento estándar

Fase exploratoria: la clasificación jerárquica

La clasificación jerárquica con el procedimiento de Ward minimiza la suma de las distancias euclidianas al cuadrado entre el individuo i y el centro del segmento al que se asigna,$\sum_k \sum_{i \in k} \sum_j = (x(i,j) - x(j',j))^2$. El primer sumatorio calcula el error de la observación $i$ en el segmento $k$ para todas las variables, el segundo realiza el cálculo para todos los individuos en el segmento $k$ y, finalmente, el último sumatorio calcula la suma del error para todos los segmentos (Wishart 1998).

Funciona de la siguiente manera

1. Computa la matriz de las distancias euclidianas al cuadrado para todo par de observaciones, *i*, *i'*, $d^2(i,i')$=$\sum_k (x(i,j) - x(j',j))^2$.
2. Agrupa las dos observaciones o grupos $i$ y $i'$ más cercanos, es decir, cuya agrupación minimiza el incremento en el error, *E*. Inicialmente serán aquellos casos cuya distancia, $d^2(i,i')$, sea mínima.
3. Transforma la matriz inicial de distancias euclidianas al cuadrado $D^2$ en $E^2$, la cual contendrá el error al cuadrado de la unión del nuevo segmento $i \cup i$.
4. Repite los pasos 2 y 3 y cada vez forma un nuevo grupo con las observaciones o grupos cuya unión resulte en un incremento mínimo en el error $E^2$.
5. Finaliza cuando todos los casos se hayan agrupado en un solo segmento

El procedimiento minimiza el cuadrado de la suma de las desviaciones entre los individuos y el centro del grupo al que se le ha asignado. Realiza un proceso de aglomeración con tantas fases o etapas como individuos ha de clasificar, $n$, de manera que una vez clasificados en su totalidad se minimiza una medida de la heterogeneidad, $min \sum_{SQD-I}$. En la etapa inicial existen tantos grupos como individuos debe agrupar. A partir de ese momento en cada etapa se formará un nuevo segmento agrupando a dos de los segmentos ya formados en etapas anteriores o un grupo y un individuo aun no clasificado, aquéllos más parecidos entre sí, de forma que se minimice en cada agrupación el incremento en la suma de las diferencias entre el individuo agrupado y la media del grupo al que se asigna, $min \left\{ \sum_i x^2(i,j) - \frac{\sum_j x^2(i,j)}{n} \right\}$. En la etapa inicial la suma del cuadrado de las distancias es cero y se va incrementando a medida que se realizan las agrupaciones.

Suponogamos que tenemos cinco individuos medidos en una única variable $\{A = 2, B = 5, C = 9, D = 10, E = 15\}$.

d(A, B)=22+52-(1/2)(2+5)2=29-24,5=4,5

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 4,5 | 24,5 | 32 | 98 |
| B | | 0 | 8 | 12,5 | 60,5 |
| C | | | 0 | **0,5** 24 | ,5 |
| D | | | | 0 | 18 |
| E | | | | | 0 |

La agrupación C y D es la que minimiza el SQD.

| | A | B | CD | E |
|---|---|---|---|---|
| A | 0 | **4,5** 38 | 98 | |
| B | | 0 | 14 | 60,5 |
| CD | | | 0 | 28,66 |
| E | | | | 0 |

La agrupación A y B es la que minimiza la SQD

| | AB | CD | E |
|---|---|---|---|
| AB | 0 | 41 | 108,66 |
| CD | | 0 | **28,66** |
| E | | | 0 |

La agrupación *CDE* es la que minimiza la *SQD*. Finalmente $d(AB,\ CDE)$= 113,2.

La agrupación se muestra visualmente con un gráfico denominado dendrograma, donde las letras identifican a los individuos.

```
ejemplo<-c(2,5,9,10,15)
ejemplo
```

```
## [1]  2  5  9 10 15
```

```
ejemplo.dist<-dist(ejemplo)
ejemplo.dist
```

```
##    1  2  3  4
## 2  3
## 3  7  4
## 4  8  5  1
## 5 13 10  6  5
```

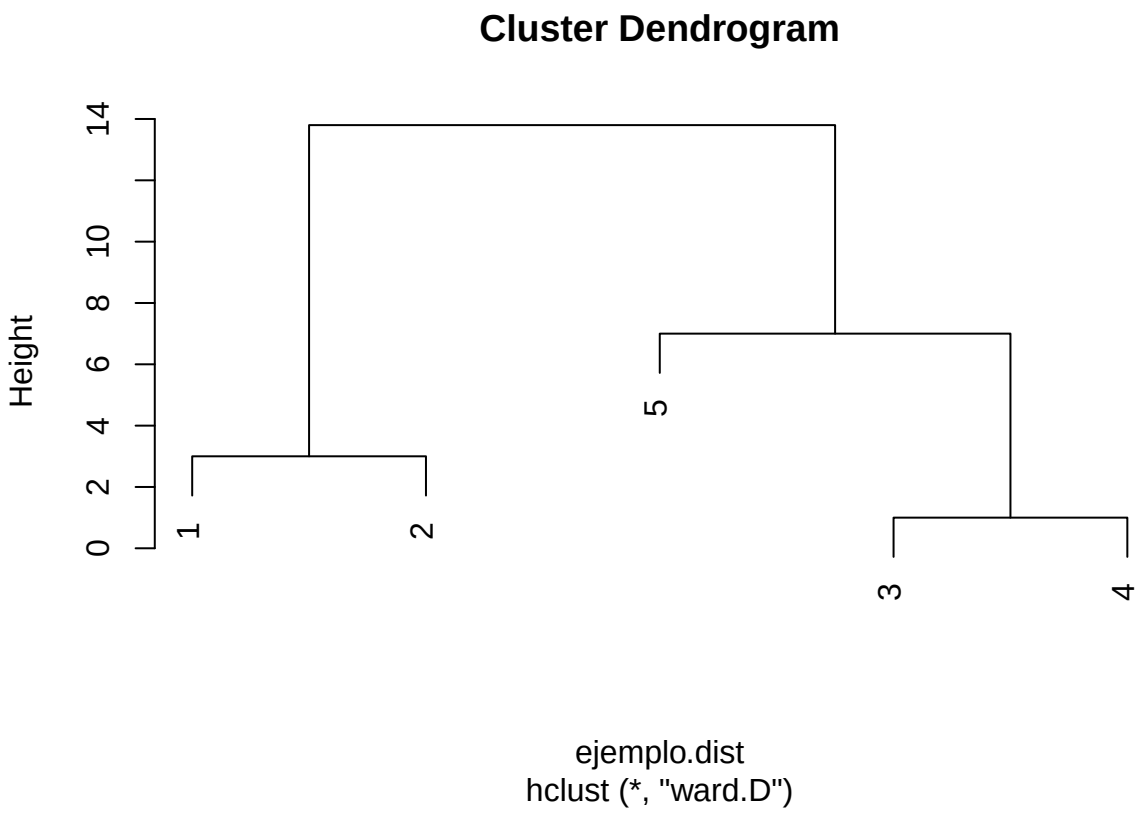
```
ejemplo.hclust<-hclust(ejemplo.dist, method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

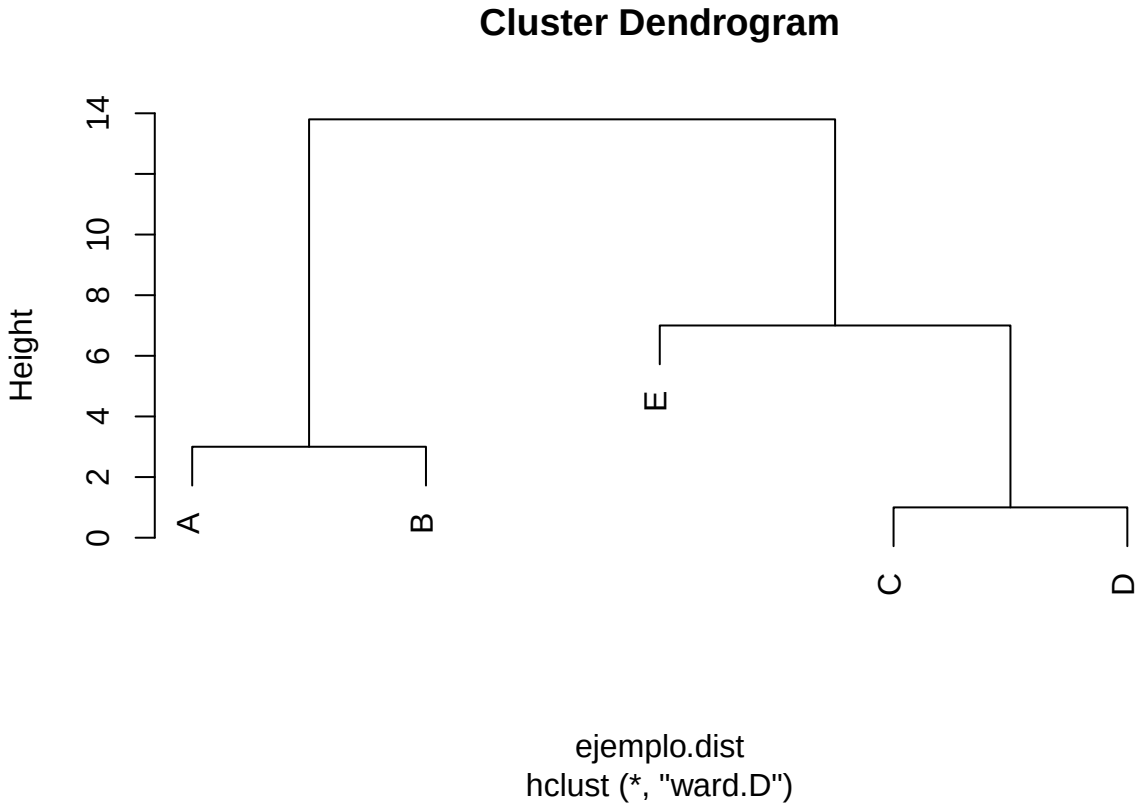
```
ejemplo.hclust
```

```
##
## Call:
## hclust(d = ejemplo.dist, method = "ward")
##
## Cluster method   : ward.D
## Distance         : euclidean
## Number of objects: 5
```

```r
plot(ejemplo.hclust)
```




```r
labels<-c("A", "B", "C", "D", "E")
ejemplo.hclust$labels<-labels
plot(ejemplo.hclust)
```




Veamos cómo lo haríamos en el entorno R de análisis de datos.

## Partición de la muestra en k segmentos: el procedimiento k-means

Los algoritmos de clasificación del tipo k-means inician el proceso de selección especificando el número de clases que desea el analista, $k$, así como los centros iniciales para cada uno de los $k$ clusters a formar. Después asigna cada individuo al grupo cuyo centro tenga más próximo. La solución inicial se mejora de manera iterativa hasta que se alcanza alguna forma de estabilidad, por ejemplo, hasta que las reasignaciones no reducen la suma del cuadrado de las distancias de los consumidores a los centros de los segmentos a los que han sido asignados (Helsen y Green 1991). Concretamente:

1. Empieza el proceso con un conjunto de centros iniciales o coordenadas de los $k$ segmentos. Estos centros pueden obtenerse de manera aleatoria o por algún procedimiento como el propuesto por Hartigan y Wong (1979).
2. Después asignamos al consumidor $i$ al cluster cuyo centro está más próximo. Los centros permanecen inalterados durante todo el proceso de asignación.
3. Seguidamente calculamos un nuevo conjunto de centros de los segmentos, como la media de los consumidores asignados a cada uno. Estos nuevos centros serán la base de un nuevo ciclo de reasignaciones.
4. Repetimos los procesos 2 y 3 hasta que ningún individuo cambie de grupo en el proceso de asignación especificado en el paso 2.

En este algoritmo el centro de cada segmento se define como la media de todos los individuos asignados al segmento en cada una de las variables en las que se han medido sus propiedades. Por ello necesita una matriz de datos $nxp$ en lugar de una matriz de distancias entre los individuos (a diferencia de lo que ocurre en los

algoritmos de clasificación jerárquica). El propósito del algoritmo es minimizar la suma de las distancias euclidianas al cuadrado de la partición realizada en $k$ segmentos. Implícitamente asume que los grupos muestran una distribución normal esférica. Vemos un ejemplo.

$n$= número de individuos a clasificar

$p$= número de variables en las que se han medido las propiedades de los individuos

$x(i,j)$= valor que muestra el individuo $i$ en la variable $j$

$x(l,j)$= valor medio de la variable $j$ en el segmento $l$.

$n(l)$=número de individuos en el grupo $l$.

$d(i,l)$= distancia entre el individuo $i$ y el centro del segmento $l$.

$d(i,l)$=raíz cuadra de la suma para toda $p$ de las diferencias al cuadrado

$e(p(n,\ k))$=error de la partición

$p(n,\ k)$= resultado de partir la muestra en $k$ segmentos y asignar a los $n$ individuos a cada uno de los $k$ segmentos.

$min\ e(p(n,\ k))=\sum_{i=1,\ldots,n} d^2(i,l(i))=min\ SQD_I$

| Tipo de pescado | Energía | Calorías | Calcio | Sum(i) |
|---|---|---|---|---|
| Caballa | 5 | 9 | 20 | 34 |
| Perca | 6 | 11 | 2 | 19 |
| Salmón | 4 | 5 | 20 | 29 |
| Sardina | 6 | 9 | 46 | 61 |
| Atún | 5 | 7 | 1 | 13 |
| Camarones | 3 | 1 | 12 | 16 |

Supongamos que queremos formar tres segmentos. El procedimiento nos daría como resultado los siguientes segmentos:

Segmento 1: perca, atún y caballa

Segmento 2: caballa y salmón

Segmento 3: sardina

Seguidamente calcularemos la media de las propiedades de los objetos clasificados en cada uno de los segmentos.

| Segmento | Energía | Calorías | Calcio |
|---|---|---|---|
| 1 | 14/3 | 19/3 | 5 |
| 2 | 9/2 | 7 | 20 |
| 3 | 6 | 9 | 46 |

Y calculamos la distancia euclidiana entre los individuos y las medias de cada grupo:

E(p(n=6,k=3))=SQDI=$d^2$ˆ(1,1)+$d^2$(2,1)+...+$d^2$(6,1)+...+$d^2$(1,3)+...+$d^2$(6,3)=137,805

Seguidamente probamos si cualquier cambio en la asignación de individuos a los segmentos reduce el error de la clasificación o suma del cuadrado de las distancias entre individuos y centros. Siendo $n(l)$ el número de individuos asignados al segmento $l$, y $l(i)$ el segmento que contiene al individuo $i$, primero calculamos las distancias al cuadrado entre el primer individuos y los centros de cada uno de los grupos:

$d^2(1,1)=(5-14/3)^2+(9-19/3)^2+(20-5)^2=232{,}22$

$d^2(1,2)=4,25$

$d^2(1,2)=677$

A la hora de decidir donde clasificar al individuo i, calcularemos la variación en la $SQD_I$. En este caso, cambiar la caballa de segmento incrementaría el error de la partición realizada.

Realizamos el proceso para todos los objetos y encontramos, en este caso, que el objeto 6, el camarón, puede ser clasificado en el segmento 2, en lugar del *1* inicial, y reducir el error de la clasificación. Quedando así la clasificación siguiente:

Segmento 1: perca y atún

Segmento 2: caballa, salmón y camarón

Segmento 3: sardinas.

Código (modificar)

Energia Calorias Calcio

Caballa 5 9 20

Perca 6 11 2

Salmon 4 5 20

Sardina 6 9 46

Atun 5 7 1

Camarones 3 1 12

```
peces<-read.csv("peces.csv", row.names=1, header=T)
peces
```

```
##           Energia Calorias Calcio
## Caballa         5        9     20
## Perca           6       11      2
## Salmon          4        5     20
## Sardina         6        9     46
## Atun            5        7      1
## Camarones       3        1     12
```

```
peces.dist<-dist(peces)

peces.dist
```

```
##             Caballa     Perca    Salmon   Sardina      Atun
## Perca     18.138357
## Salmon     4.123106 19.078784
## Sardina   26.019224 44.045431 26.381812
## Atun      19.104973  4.242641 19.131126 45.055521
## Camarones 11.489125 14.456832  9.000000 35.057096 12.688578
```

Caballa Perca Salmon Sardina Atun

Perca 18.138357

Salmon 4.123106 19.078784

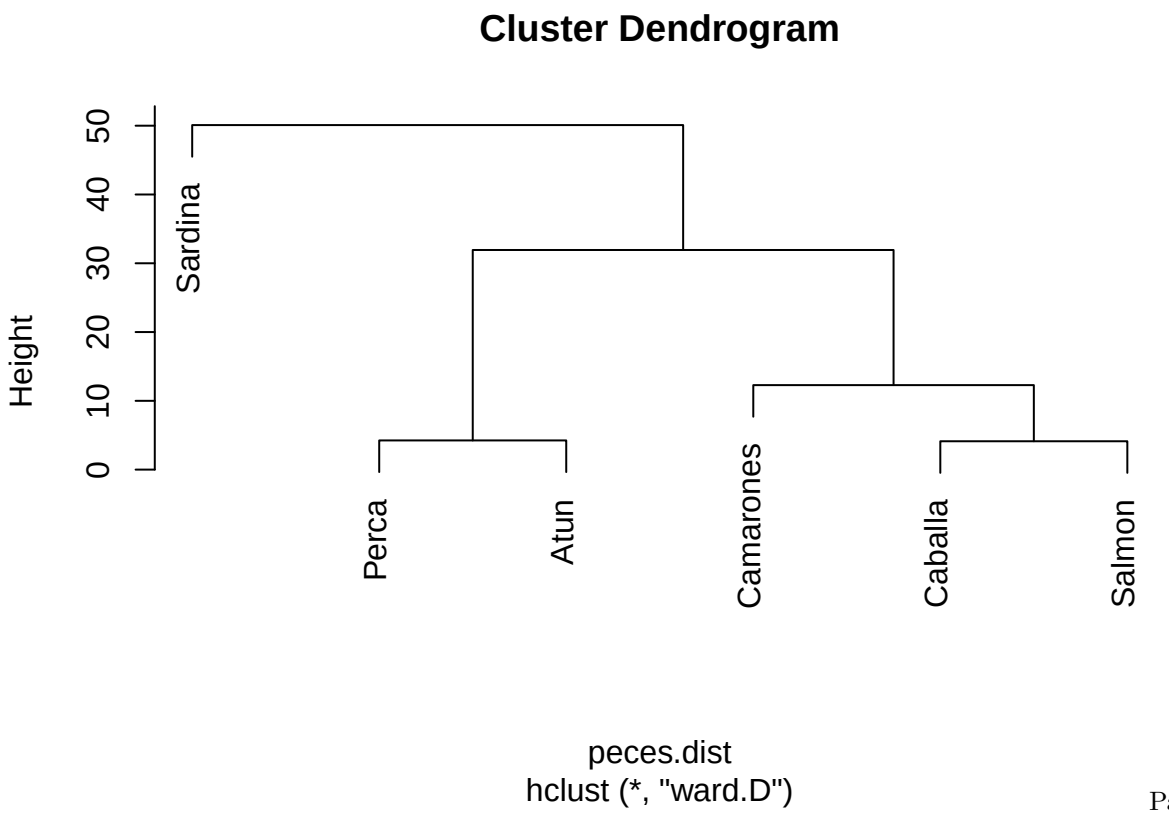
Sardina 26.019224 44.045431 26.381812

Atun 19.104973 4.242641 19.131126 45.055521

Camarones 11.489125 14.456832 9.000000 35.057096 12.688578

You can also embed plots, for example:

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```




Partimos la muestra en tres segmentos

```
## KMNS(*, k=3): iter=  1, indx=0
##  QTRAN(): istep=6, icoun=1
## KMNS(*, k=3): iter=  2, indx=3
## KMNS(*, k=3): iter=  3, indx=6
```

Si queremos saber las opciones de lal función k-means, podemos utilizar la función ? seguida del nombre de la función

```
?kmeans
```

Para saber los objetos incluidos en el resultado de clasificar la muestra, podemos uitlizar la función names()

```
names(peces.kmeans)
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Los centros de los segmentos se encuentran en el objeto `centers`. La classificación se encuentra en el objeto `cluster`, y en el objeto `iter`el número de iteraciones realizadas.

```
peces.kmeans$centers
```

```
##   Energia Calorias   Calcio
## 1     5.5        9  1.50000
## 2     6.0        9 46.00000
## 3     4.0        5 17.33333
```

```
peces.kmeans$cluster
```

```
##   Caballa     Perca    Salmon   Sardina      Atun Camarones
##         3         1         3         2         1         3
```

```
peces.kmeans$iter
```

```
## [1] 3
```

El resultado lo podemos visualizar en el espacio de las bases de segmentación

```
plot(peces[1:2], col=peces.kmeans$cluster)
points(peces[1:2])
points(peces.kmeans$centers, col=1:2, pch=8,cex=2)
```