Practice 4 SPSS and RCommander

Cluster Analysis

It is a class of techniques used to classify *cases* (or *variables*) into ***groups*** that are relatively **homogeneous** *within* themselves, and **heterogeneous** *between* each other, on the basis of a defined set of variables. These groups are called **clusters**.

Example: clustering of consumers according to their attribute preferences or similar behaviours and characteristics. Then, clusters of similar brands or products can help identifying market opportunities.

General Steps to conduct a Cluster Analysis

i. Select a distance measure.
ii. Select a clustering algorithm.
iii. Determine the number of clusters.
iv. Validate the analysis.

Clustering procedures

Hierarchical procedures

Agglomerative (start from *n* clusters, to get to 1 cluster)

Divisive (start from 1 cluster, to get to *n* cluster)

Non hierarchical procedures

K-means clustering

Agglomerative clustering

Linkage methods

Single linkage (minimum distance)

Complete linkage (maximum distance)

Average linkage

The distance between two clusters is defined as the difference between the centroids (cluster averages)

K-means clustering

The number k of cluster is fixed

An initial set of *k seeds* (aggregation centres) is provided

First *k* elements or

Other seeds

Given a certain treshold, all units are assigned to the nearest cluster seed

New seeds are computed

Go back to step 3 until no reclassification is necessary

Units can be reassigned in successive steps (*optimising partioning*)

Hierarchical vs Non hierarchical methods

Hierarchical clustering

No decision about the number of clusters.

Problems when data contain a high level of error.

Can be very slow...

Initial decision are more influential (one-step only).

Non hierarchical clustering

Faster, more reliable.

Need to specify the number of clusters (arbitrary).

Need to set the initial seeds (*arbitrary*).

Hierarchical cluster analysis

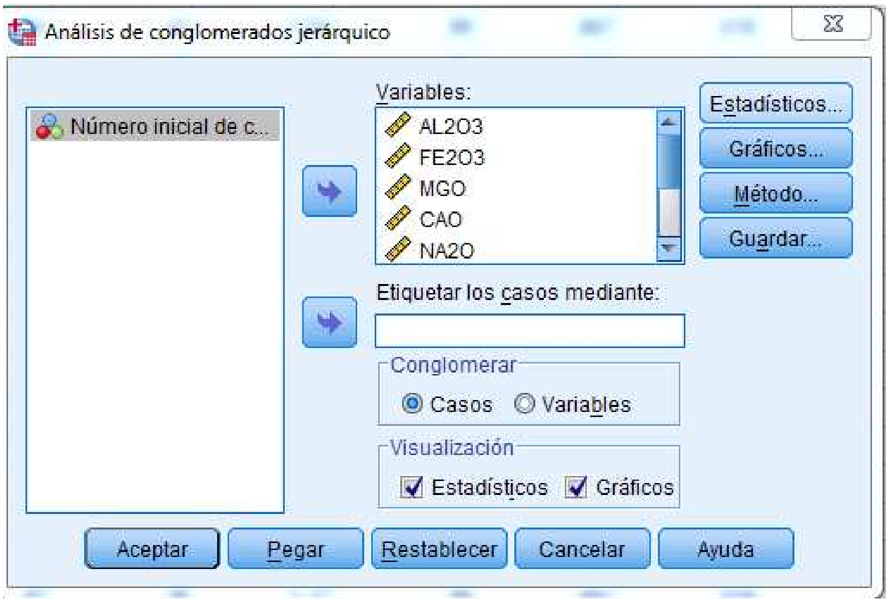
Here, we will be use the file **pottery.txt**

The data give the chemical composition of 48 specimens of *Romano-British* pottery, determined by atomic absorption spectrophotometry, for nine oxides. In addition to the chemical composition of the pots, the kiln site at which the pottery was found is known for these data.

For these data, interest centres on whether, on the basis of their chemical compositions, the pots can be divided into distinct groups, and how these groups relate to the *kiln* site.

Go to Analyze, followed by Classify, and then Hierarchical Cluster. Drag drop all variables except *kiln*.

Analizar → Clasificar → Conglomerados Jerarquicos



Análisis de conglomerados jerárquico

Número inicial de c...

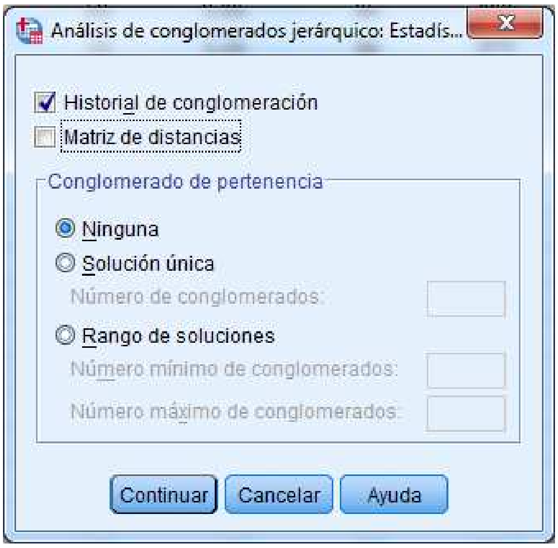
Variables:
AL2O3
FE2O3
MGO
CAO
NA2O

Etiquetar los casos mediante:

Conglomerar: ● Casos ○ Variables

Visualización: ☑ Estadísticos ☑ Gráficos

Press in **Estadisticos**


Análisis de conglomerados jerárquico: Estadís...

☑ Historial de conglomeración

□ Matriz de distancias

Conglomerado de pertenencia

● Ninguna

○ Solución única

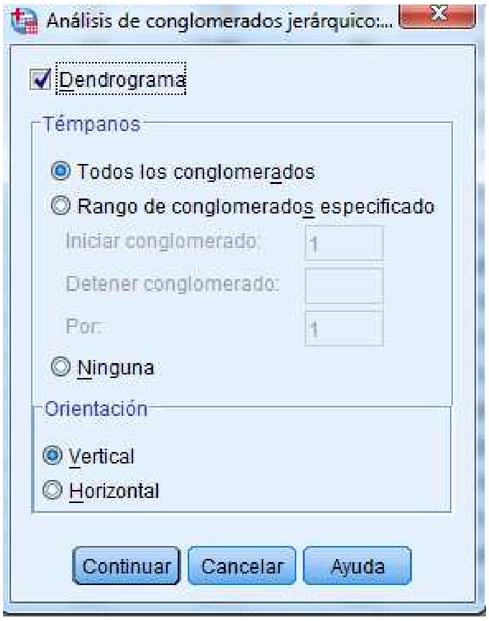
Número de conglomerados:

○ Rango de soluciones

Número mínimo de conglomerados:

Número máximo de conglomerados:

Press in **Graficos**


Análisis de conglomerados jerárquico:...

☑ Dendrograma

Témpanos

● Todos los conglomerados

○ Rango de conglomerados especificado

Iniciar conglomerado: 1

Detener conglomerado:

Por: 1

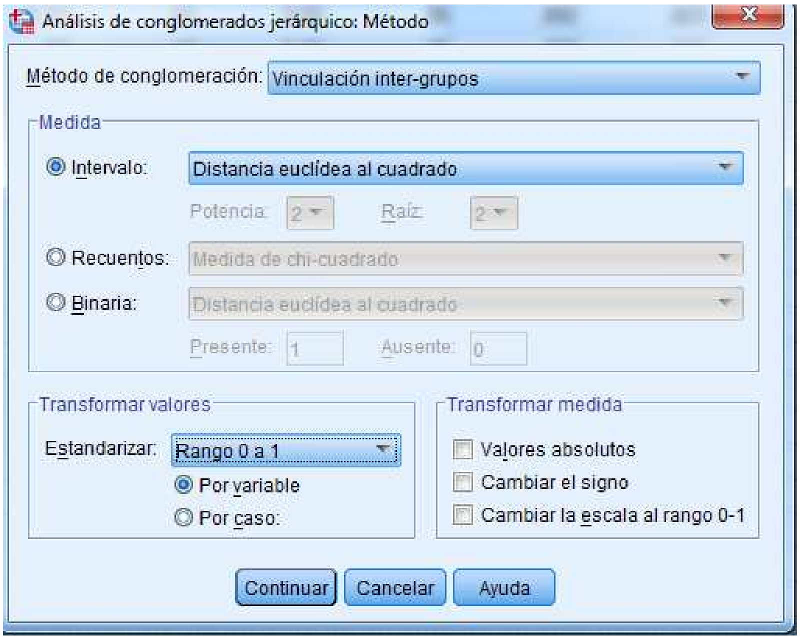
○ Ninguna

Orientación

● Vertical

○ Horizontal

Press in **Metodo**



Análisis de conglomerados jerárquico: Método

Método de conglomeración: Vinculación inter-grupos

Medida

● Intervalo: Distancia euclídea al cuadrado

Potencia: 2 Raíz: 2

○ Recuentos: Medida de chi-cuadrado

○ Binaria: Distancia euclídea al cuadrado

Presente: 1 Ausente: 0

Transformar valores

Estandarizar: Rango 0 a 1

● Por variable

○ Por caso:

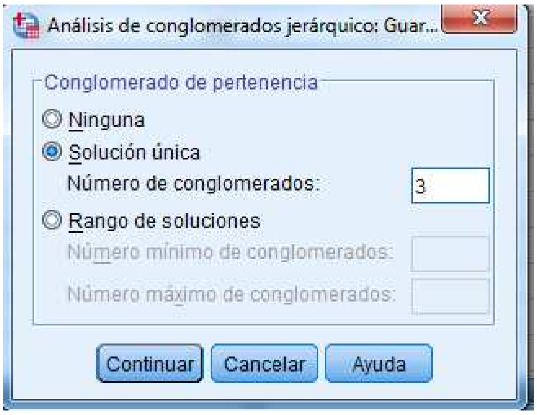
Transformar medida

□ Valores absolutos

□ Cambiar el signo

□ Cambiar la escala al rango 0-1

Press in **Guardar**



Análisis de conglomerados jerárquico: Guar...

Conglomerado de pertenencia

○ Ninguna

● Solución única

Número de conglomerados: 3

○ Rango de soluciones

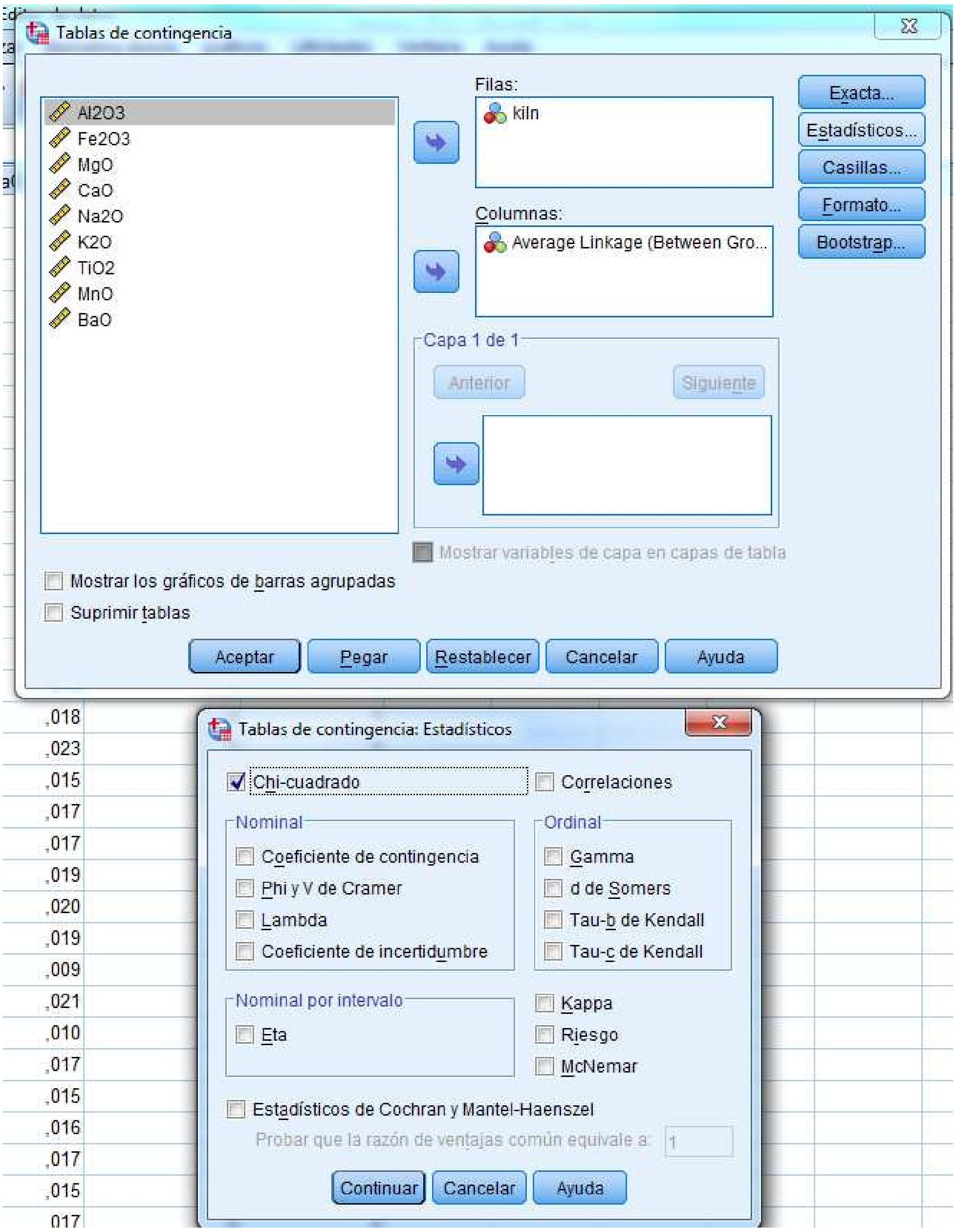
Número mínimo de conglomerados:

Número máximo de conglomerados:

Hierarchic classifications may be represented by a two-dimensional diagram known as a **dendrogram**, which illustrates the fusions made at each stage of the analysis.

We derive the three-cluster solution by cutting the dendrogram at a *height* of 10. Our interest is now a comparison with the *kiln* sites at which the pottery was found.

Analizar → Estadisticos Descriptivos → Tablas de Contingencia



Al2O3
Fe2O3
MgO
CaO
Na2O
K2O
TiO2
MnO
BaO

Filas: kiln

Columnas: Average Linkage (Between Gro...

Capa 1 de 1

☐ Mostrar variables de capa en capas de tabla

☐ Mostrar los gráficos de barras agrupadas

☐ Suprimir tablas

,018
,023
,015
,017
,017
,019
,020
,019
,009
,021
,010
,017
,015
,016
,017
,015
,017

☑ Chi-cuadrado ☐ Correlaciones

Nominal
☐ Coeficiente de contingencia
☐ Phi y V de Cramer
☐ Lambda
☐ Coeficiente de incertidumbre

Ordinal
☐ Gamma
☐ d de Somers
☐ Tau-b de Kendall
☐ Tau-c de Kendall

Nominal por intervalo
☐ Eta

☐ Kappa
☐ Riesgo
☐ McNemar

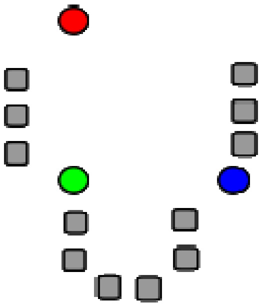
☐ Estadísticos de Cochran y Mantel-Haenszel

Probar que la razón de ventajas común equivale a: 1

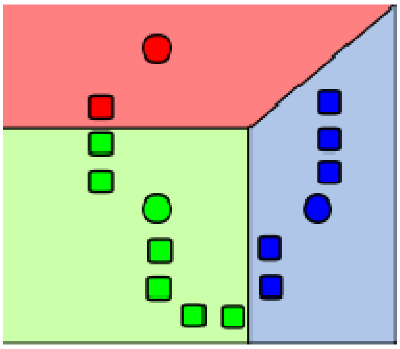
The contingency table shows that cluster 1 contains all pots found at *kiln* site number one, cluster 2 contains all pots from *kiln* sites number two, and cluster three collects pots from *kiln* sites four and five. In fact, so the clusters actually correspond to pots from three different regions.

K-means clustering

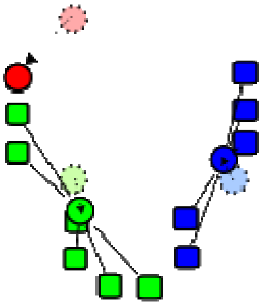
Basic ideas and scheme (from *Wikipedia*):



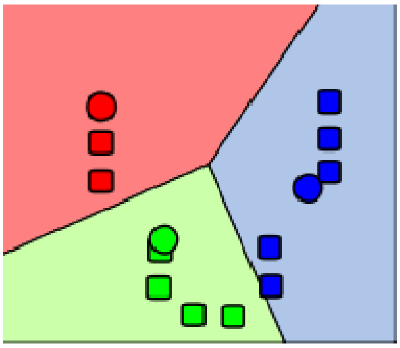
1) *k* initial *means* (in this case *k*=3) are randomly selected from the data set (shown in color).



2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the *Voronoi diagram* generated by the means.



3) The *centroid* of each of the *k* clusters becomes the new means.

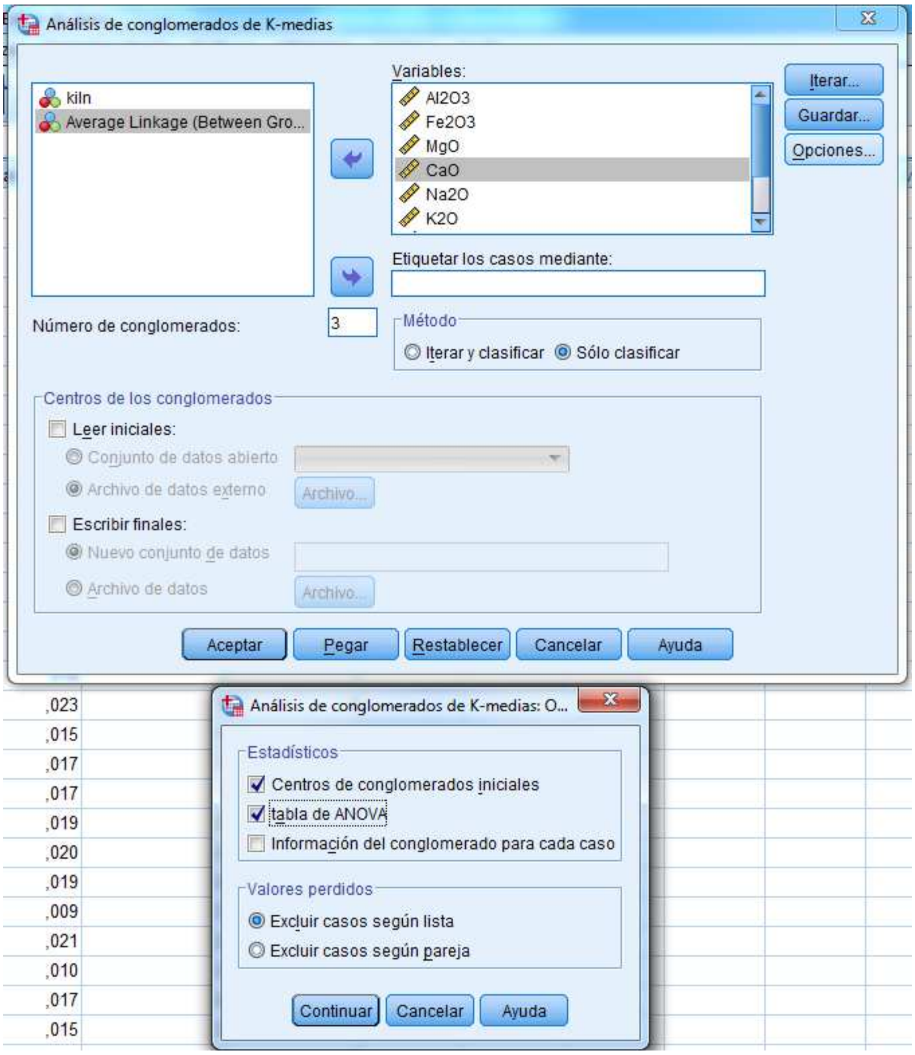


4) Steps 2 and 3 are repeated until convergence has been reached.

We use the same data **pottery.txt** and set 3 as the number of clusters.

Go to Analyze, followed by Classify, and then K means Cluster. Drag drop all variables except *kiln*.

Analizar → Clasificar → Conglomerados K medias



Análisis de conglomerados de K-medias

kiln
Average Linkage (Between Gro...

Variables:
Al_2O_3
Fe_2O_3
MgO
CaO
Na_2O
K_2O

Etiquetar los casos mediante:

Número de conglomerados: 3

Método
○ Iterar y clasificar ● Sólo clasificar

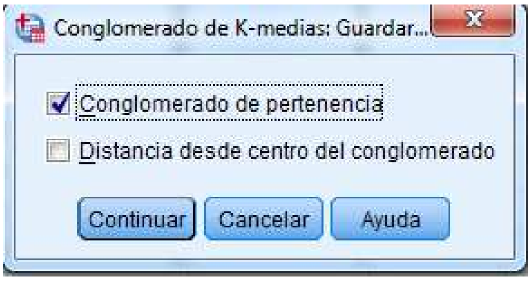
Centros de los conglomerados
□ Leer iniciales:
○ Conjunto de datos abierto
● Archivo de datos externo
□ Escribir finales:
● Nuevo conjunto de datos
○ Archivo de datos

Análisis de conglomerados de K-medias: O...

Estadísticos
☑ Centros de conglomerados iniciales
☑ tabla de ANOVA
□ Información del conglomerado para cada caso

Valores perdidos
● Excluir casos según lista
○ Excluir casos según pareja

Press in **Guardar**



Conglomerado de K-medias: Guardar...

☑ Conglomerado de pertenencia
□ Distancia desde centro del conglomerado

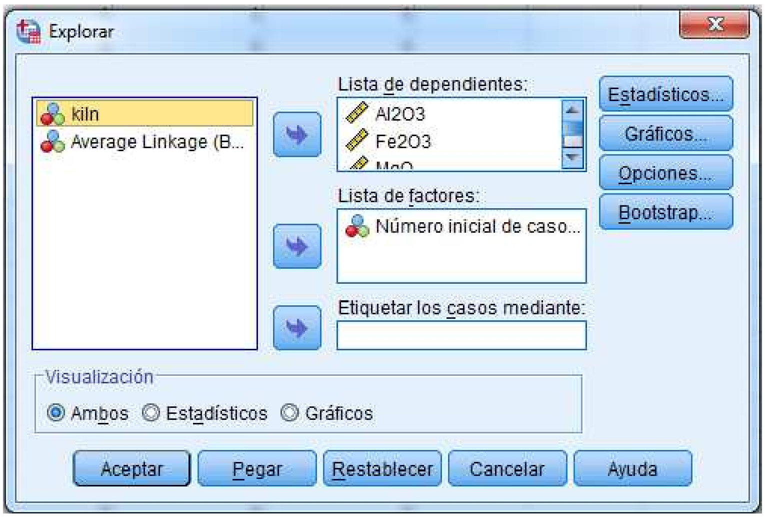
We derive the three-cluster solution and we make now a comparison with the *kiln* sites at which the pottery was found.

Analizar → Estadisticos Descriptivos → Tablas de Contingencia

We obtain similar results as in the hierarchical clustering case.

It may be very informative to compute the descriptive statistics on the original variables for each cluster in order to interpret them.

Analizar → Estadisticos Descriptivos → Explorar


Explorar

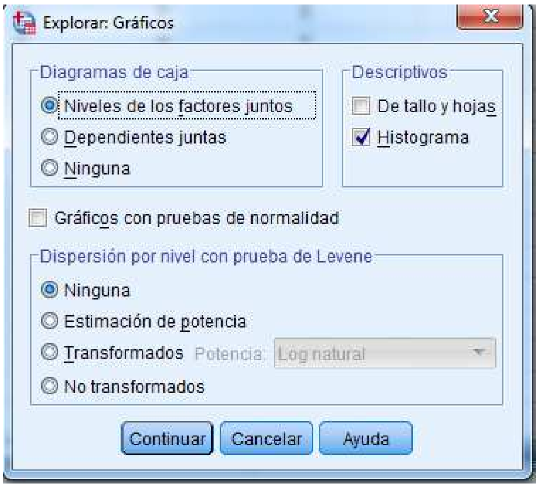
kiln
Average Linkage (B...

Lista de dependientes:
Al2O3
Fe2O3
MgO

Lista de factores:
Número inicial de caso...

Etiquetar los casos mediante:

Visualización
● Ambos ○ Estadísticos ○ Gráficos


Explorar: Gráficos

Diagramas de caja
● Niveles de los factores juntos
○ Dependientes juntas
○ Ninguna

Descriptivos
□ De tallo y hojas
☑ Histograma

□ Gráficos con pruebas de normalidad

Dispersión por nivel con prueba de Levene
● Ninguna
○ Estimación de potencia
○ Transformados Potencia: Log natural
○ No transformados

Cluster Analysis with RCommander

Hierarchical cluster analysis

Here, we will be use the file **pottery.txt**

The data give the chemical composition of 48 specimens of *Romano-British* pottery, determined by atomic absorption spectrophotometry, for nine oxides. In addition to the chemical composition of the pots, the kiln site at which the pottery was found is known for these data.

For these data, interest centres on whether, on the basis of their chemical compositions, the pots can be divided into distinct groups, and how these groups relate to the *kiln* site.

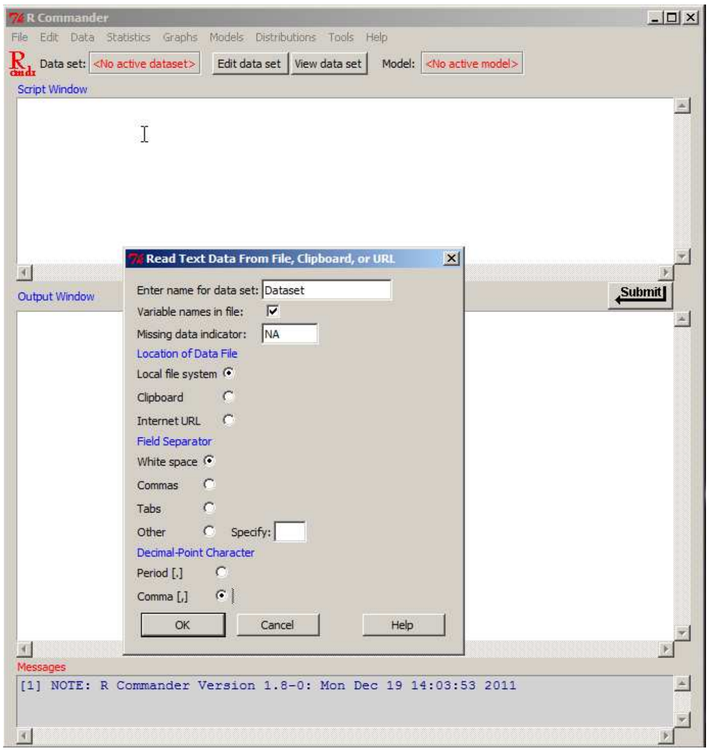
In the working panel of R, type

library(Rcmdr)

Go to

Data → Import Data → From text File…

Take the file **pottery.txt**



R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Edit data set
View data set
Script Window
Output Window
Submit
Messages
Read Text Data From File, Clipboard, or URL
OK
Cancel
Help

Data set: <No active dataset>

Model: <No active model>

Enter name for data set: Dataset

Variable names in file: ☑

Missing data indicator: NA

Location of Data File

Local file system ●

Clipboard ○

Internet URL ○

Field Separator

White space ●

Commas ○

Tabs ○

Other ○ Specify:

Decimal-Point Character

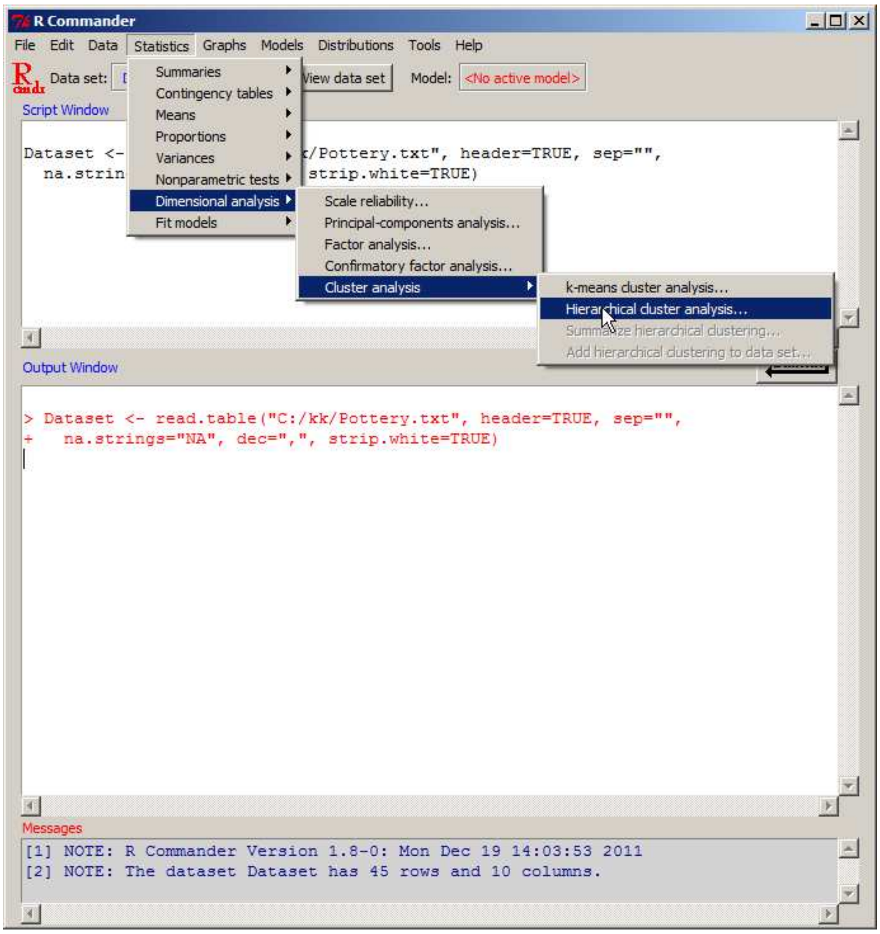
Period [.] ○

Comma [,] ●

```
[1] NOTE: R Commander Version 1.8-0: Mon Dec 19 14:03:53 2011
```

Go to

Statistics → Dimensional analysis → Cluster analysis → Hierarchical cluster analysis

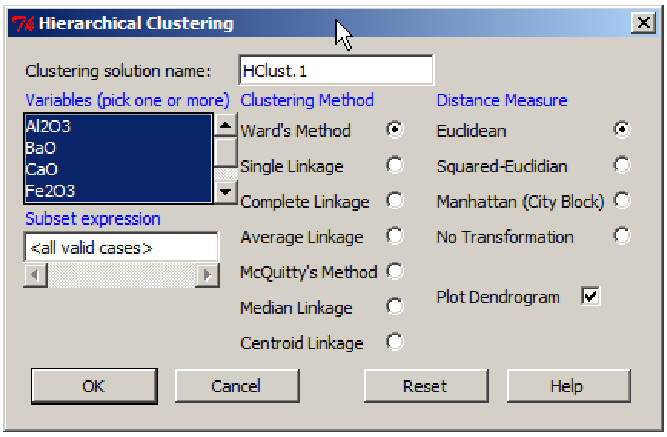

R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set:
View data set
Model: <No active model>
Script Window
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Scale reliability...
Principal-components analysis...
Factor analysis...
Confirmatory factor analysis...
Cluster analysis
k-means cluster analysis...
Hierarchical cluster analysis...
Summarize hierarchical clustering...
Add hierarchical clustering to data set...
Output Window
Messages

```
Dataset <-                    /Pottery.txt", header=TRUE, sep="",
  na.strin                    strip.white=TRUE)
```

```
> Dataset <- read.table("C:/kk/Pottery.txt", header=TRUE, sep="",
+   na.strings="NA", dec=",", strip.white=TRUE)
```

```
[1] NOTE: R Commander Version 1.8-0: Mon Dec 19 14:03:53 2011
[2] NOTE: The dataset Dataset has 45 rows and 10 columns.
```

Include all variables **except** *kiln*



Clustering solution name: HClust.1

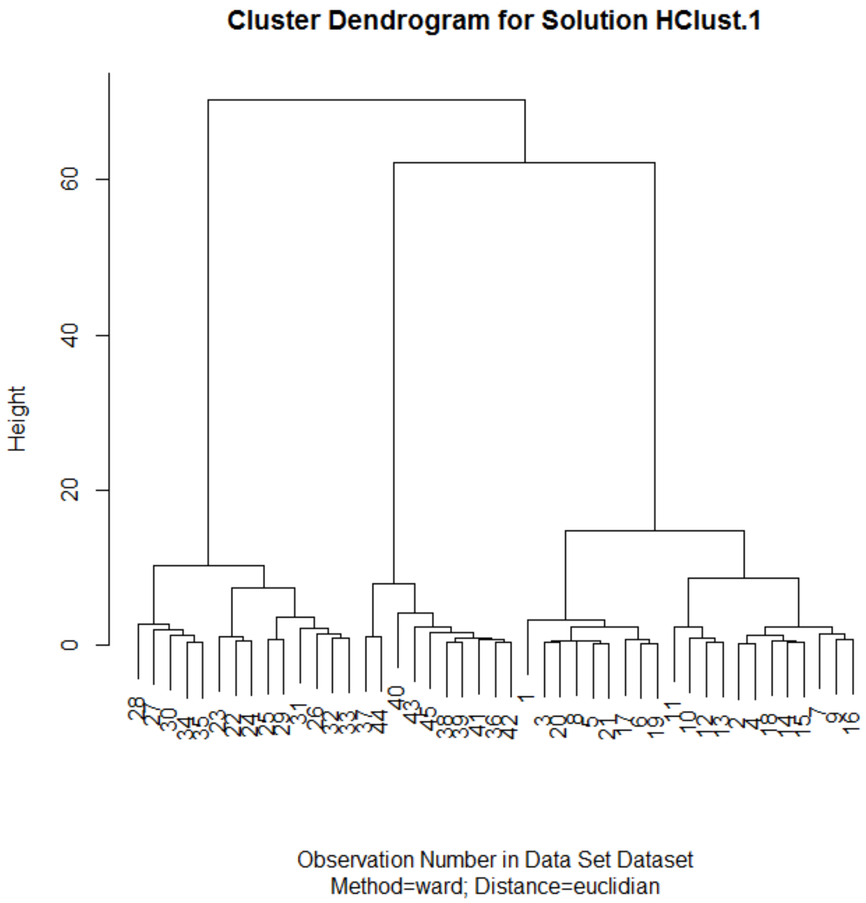
Variables: Al2O3, BaO, CaO, Fe2O3

Subset expression: <all valid cases>

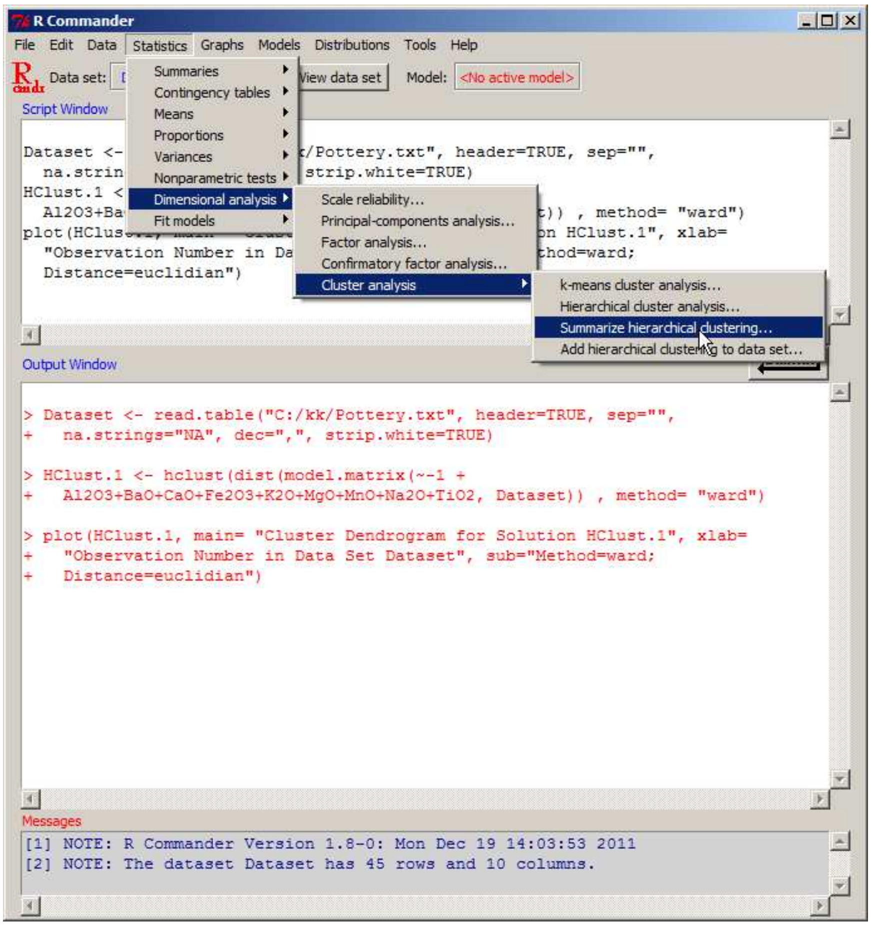
Ward's Method ●
Single Linkage ○
Complete Linkage ○
Average Linkage ○
McQuitty's Method ○
Median Linkage ○
Centroid Linkage ○

Euclidean ●
Squared-Euclidian ○
Manhattan (City Block) ○
No Transformation ○

Plot Dendrogram ☑



Summarize results:

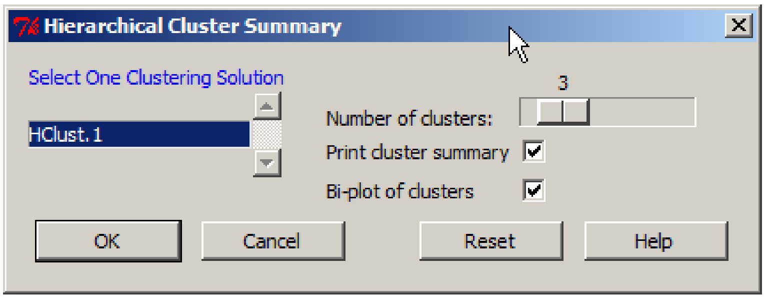

R Commander
File Edit Data Statistics Graphs Models Distributions Tools Help
Data set:
View data set
Model: <No active model>
Script Window
Summaries
Contingency tables
Means
Proportions
Variances
Nonparametric tests
Dimensional analysis
Fit models
Scale reliability...
Principal-components analysis...
Factor analysis...
Confirmatory factor analysis...
Cluster analysis
k-means cluster analysis...
Hierarchical cluster analysis...
Summarize hierarchical clustering...
Add hierarchical clustering to data set...
Output Window
Messages

```
> Dataset <- read.table("C:/kk/Pottery.txt", header=TRUE, sep="",
+   na.strings="NA", dec=",", strip.white=TRUE)

> HClust.1 <- hclust(dist(model.matrix(~-1 +
+   Al2O3+BaO+CaO+Fe2O3+K2O+MgO+MnO+Na2O+TiO2, Dataset)) , method= "ward")

> plot(HClust.1, main= "Cluster Dendrogram for Solution HClust.1", xlab=
+   "Observation Number in Data Set Dataset", sub="Method=ward;
+   Distance=euclidian")
```

```
[1] NOTE: R Commander Version 1.8-0: Mon Dec 19 14:03:53 2011
[2] NOTE: The dataset Dataset has 45 rows and 10 columns.
```



Select One Clustering Solution: HClust.1

Number of clusters: 3

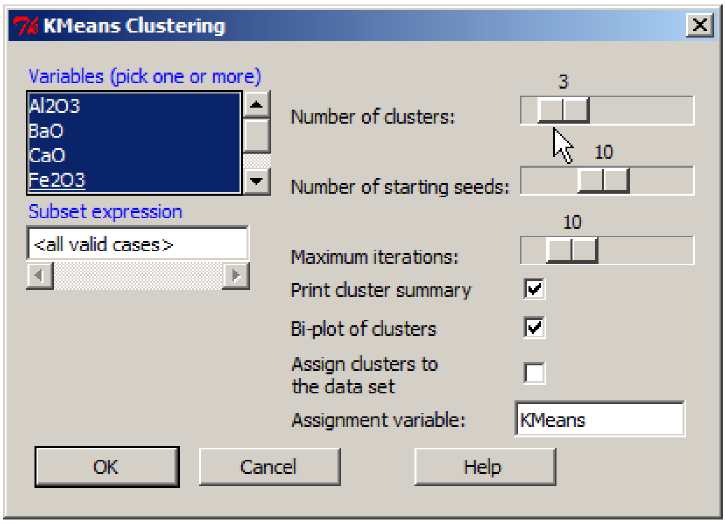
Print cluster summary ☑

Bi-plot of clusters ☑

K-means clustering

Statistics → Dimensional analysis → Cluster analysis → K-means cluster analysis

Include all variables **except** *kiln*



KMeans Clustering

Variables (pick one or more)

Al2O3
BaO
CaO
Fe2O3

Subset expression

<all valid cases>

Number of clusters: 3

Number of starting seeds: 10

Maximum iterations: 10

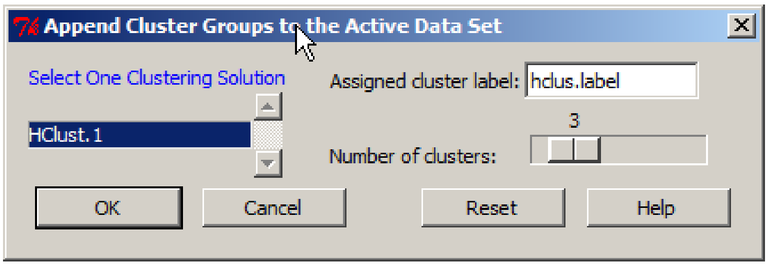
Print cluster summary ☑

Bi-plot of clusters ☑

Assign clusters to the data set □

Assignment variable: KMeans

It may be very informative to compute the descriptive statistics on the original variables for each cluster in order to interpret them.



Append Cluster Groups to the Active Data Set

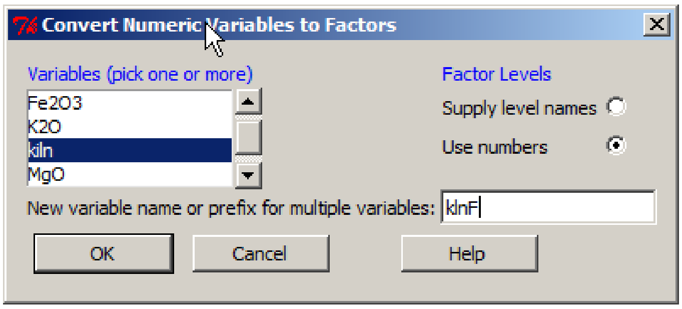
Select One Clustering Solution

HClust.1

Assigned cluster label: hclus.label

Number of clusters: 3

Convert variable *kiln* to type **factor**



Convert Numeric Variables to Factors

Variables (pick one or more)

Fe2O3
K2O
kiln
MgO

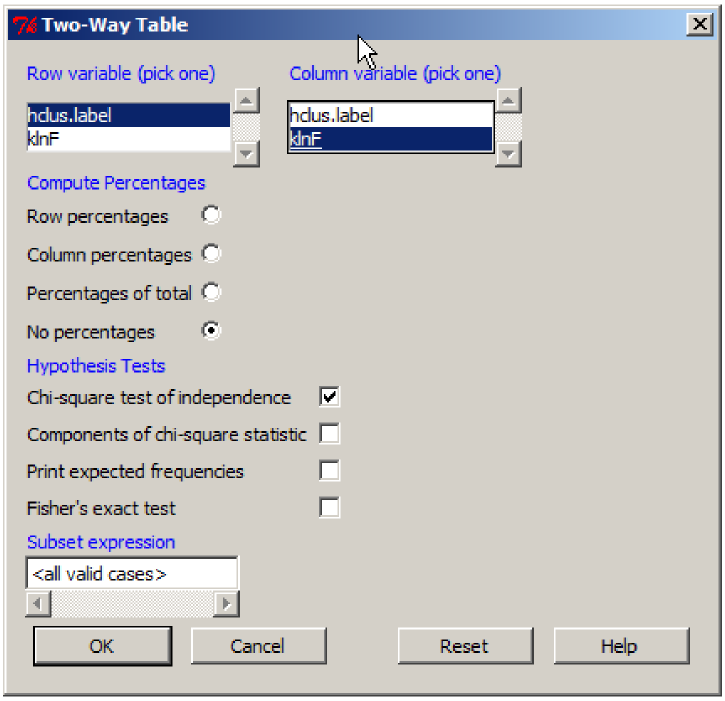
Factor Levels

Supply level names ○

Use numbers ●

New variable name or prefix for multiple variables: klnF

Compute a table to check groups



Two-Way Table

Row variable (pick one)

hclus.label
klnF

Column variable (pick one)

hclus.label
klnF

Compute Percentages

Row percentages ○

Column percentages ○

Percentages of total ○

No percentages ●

Hypothesis Tests

Chi-square test of independence ☑

Components of chi-square statistic □

Print expected frequencies □

Fisher's exact test □

Subset expression

<all valid cases>