UNIVERSIDAD DE JAÉN

# Material del curso "Análisis de datos procedentes de investigaciones mediante programas informáticos"

Manuel Miguel Ramos Álvarez

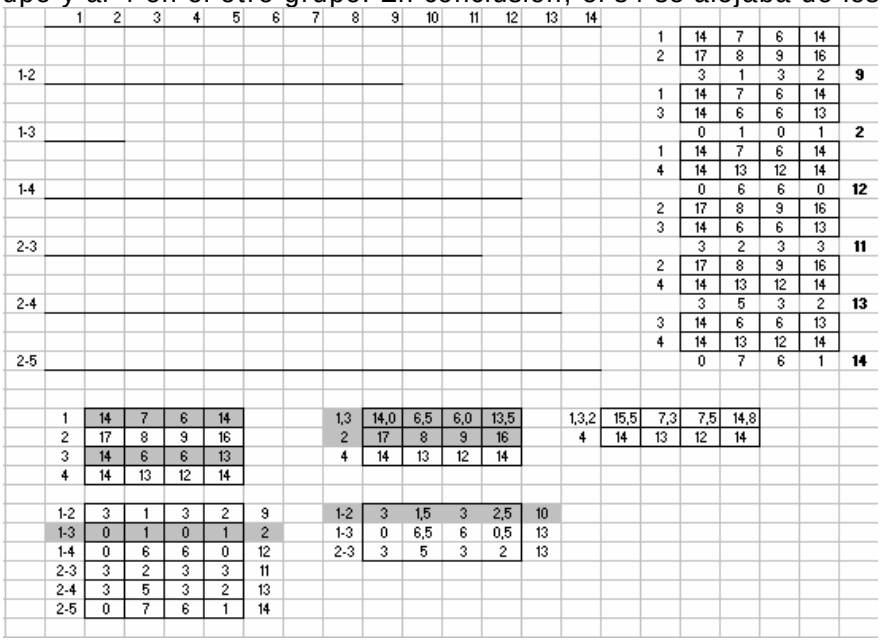
**Índice**

## MATERIAL XVA "DESCRIPCIÓN CON ANÁLISIS DE CLUSTER"

1. Conceptos fundamentales del análisis de cluster (o conglomerados) ..... 2
2. Planteamiento computacional del análisis de cluster ..... 2
3. El Análisis de Cluster o Conglomerados en profundidad ..... 4
4. Secuencia de investigación en el Análisis de Cluster o Conglomerados..... 9

# 1. Conceptos fundamentales del análisis de cluster (o conglomerados)

- Objetivo: Comenzamos con un conjunto de n-individuos a los que se ha medido en una serie de p-variables **métricas** y entonces se pretende reducir las dimensiones necesarias para interpretar un conjunto de datos. Típicamente para identificar grupos de individuos sin criterio de agrupamiento a priori. De especial interés para establecer taxonomías de forma empírica (i.e. grupos de enfermedades o trastornos).
- Lógica. Opera pues asignando individuos a grupos, cluster o conglomerados, de manera que cada grupo sea lo más homogéneo posible y los grupos sean heterogéneos unos respecto de otros.
- Lógica estadística. Para agrupar calculamos alguna distancia entre los individuos dos a dos, como por ejemplo la suma de las diferencias en todas las variables (parte superior de la figura). A continuación decidimos agruparlos a partir de tales diferencias. Los que están más cerca entran en un primer paso (i.e. el s1 y el s3, ver los cálculos en la parte inferior). Así, ahora tenemos 3 cluster para nueva agrupación en un 2º paso. Entonces volvemos a recalcular las distancias y juntamos a los dos más cercanos (i.e. el s1+3 y el s2). Así, hemos progresado desde 4 cluster hasta 2, que aglutinaría a los sujetos 1,3 y 2 en un grupo y al 4 en el otro grupo. En conclusión, el s4 se alejaba de los otros tres.




- También puede emplearse para formar grupos de variables. Semejante al AFE y al AD. Pero en el AD, el criterio de formación de los grupos es conocido y permite determinar la importancia que cada variable medida tiene para discriminar a los grupos entre sí. En contraposición, en el AC, se concede a todas las variables el mismo peso. En el AFE se proporcionan estadísticos relativamente claros acerca de si la solución final es satisfactoria, lo que no ocurre en el AC.

# 2. Planteamiento computacional del análisis de cluster

- La distancia entre los individuos se puede computar con la métrica de Minkowski:

$$\underbrace{d_{ij}}_{\substack{distancia \\ sujetos\ i-j}} = \underbrace{\sqrt[r]{\sum \left| x_{ik} - x_{jk} \right|^r}}_{\substack{suma\ a\ traves\ k \\ variables\ medidas}}$$

- Si la potencia r=1, medida de Manhattan (o city-block) y si r=2, medida euclídea.
- A partir de estas distancias hay diferentes algoritmos para realizar el agrupamiento, jerárquicamente o basado en las medias.

# 3. El Análisis de Cluster o Conglomerados en profundidad

Analizar -> Clasificar -> Conglomerados jerárquicos -> Variables: stai_est, ..., evaneg_t -> Conglomerar: Casos (sujetos) -> Estadísticos -> Mostrar: Matriz de distancias -> Historial de conglomeración -> Continuar -> Gráficos -> Dendrograma -> Témpanos: todos los conglomerados -> Continuar -> Método -> Método de conglomeración: Vinculación Inter-grupo -> Medida: Intervalo: Distancia euclídea al cuadrado -> Continuar -> Aceptar (análisis)

1. Identificar el **dominio** en el cual se inserta la investigación de cara a una adecuada interpretación. La importancia que las variables tengan para separar después grupos de sujetos es algo que sólo tiene una solución teórica y que depende, por tanto, del conocimiento que el investigador tenga del campo en el que trabaja.
2. **Planificación y estructuración del diseño:**
   a. **Seleccionar** un conjunto de p-variables que se medirán a una muestra de n- sujetos.
   b. En principio no hay limitaciones al número de variables o sujetos y de hecho apenas impone supuestos. Pero es recomendable una replicación de las soluciones, lo que exige un tamaño muestral adecuado pues solo una mitad servirá para las estimaciones.
   c. Las variables tienen que ser métricas, aunque algunos programas como Statistica incluyen alternativas computacionales para escalas de medidas más débiles que básicamente trabajan sobre la Moda en lugar de la Media Aritmética y además el método jerárquico sí se puede aplicar con otras escalas de medida pero en principio el de K-Medias no se puede.
3. Se realiza la investigación y se obtienen los datos.
4. Se elige una medida de semejanza. Medidas de distancia vs. igualación. Todas se calculan por pares (entre dos sujetos o entre un sujeto y el centroide de un grupo). Ver la tabla.

| Escala Variable | Semejanza-Igualación | Diferenciación-distancia |
|---|---|---|
| **Intervalo** | • Coseno (semejante Pearson)<br>• Pearson | • D. euclídea (al cuadrado)<br>• Manhattan (Bloques o city-block)<br>• **Métrica Minkowsky** (Potencia)<br>• Chevycheb<br>• D. generalizada de Mahalonobis (Statistica).<br>• Personalizada |
| **Categóricas-Frecuencias** | | • Chi-cuadrado<br>• Phi-cuadrado (Chi-normalizada) |
| **Dicotómicas-Binarias** | • Dispersión<br>• Concordancia simple<br>• Coeficiente Phi (Correl.)<br>• Porcentaje Desacuerdo<br>• Lambda Goodman & Kruskal<br>• D de Andemberg<br>• Dice<br>• Hamann<br>• Jaccard<br>• Kulczynski 1, 2<br>• Ochiai<br>• Rogers y Tanimoto<br>• Russel y Rao<br>• Sokal y Sneath 1, ..., 5<br>• Y de Yule<br>• Q de Yule | • D. euclídea (al cuadrado)<br>• Diferencia de tamaño<br>• Diferencia de configuración<br>• Disepersión<br>• Varianza<br>• Forma<br>• Lance y Williams |

- Algunas de las medias (i.e. basadas en **Minkowsky)** dependen de la escala, lo que minimiza la influencia relativa de las escalas más restringidas. La solución es estandarizar o emplear otras medidas como la de Mahalanobis, que son invariantes a la escala.
- Las de igualación-binarias son medidas de asociación basadas en matrices de confusión.
- La decisión sobre la medida es importante, porque puede determinar los agrupamientos que se realicen. La mayoría difieren en el peso relativo que se concede a las diferencias de mayor magnitud respecto a las de menor magnitud o bien por el peso concedido a las diferentes variables o dimensiones consideradas en los cálculos.

5. Se realiza el agrupamiento, para lo cual hay que decidir una técnica concreta: jerárquica vs. no jerárquica.
    a. Variantes de agrupamiento:
        i. **Jerárquicos**. Proceso en pasos sucesivos que progresa desde un cluster por sujeto hasta el cluster con todos los sujetos (Ascendente o por aglomeración) o a la inversa (Descendente, por disociación o por división). Entre ambos extremos se va progresando en orden de complejidad. Una vez que el individuo ha sido asignado a un cluster no puede ser reasignado a otro. Básicamente el siguiente sujeto-cluster que se agrupa con alguno de los restantes es el de mayor puntuación en la matriz de semejanzas (según las medidas de distancia del paso previo).
            - M. del **centroide**. Lo característico es que cada cluster es reducido al centroide del grupo.
            - M. **vinculación promedio** (o *vinculación inter-grupo)*. La distancia entre dos clusters corresponde al promedio de las distancias entre todos los individuos de un cluster respecto de todos los individuos del otro. **De los que más información aprovecha**.
                a. vinculación *intra-grupo* es una variante que minimiza la distancia entre todos los casos dentro de un cluster
            - M. **vecino más próximo** (o distancia mínima -single linkage-). La regla para medir la distancia entre los clústeres es la mínima de entre todas las que pueden definirse entre los miembros de uno y otro cluster en lugar de ser el promedio como en la variante anterior.
            - M. **vecino más lejano** (o distancia máxima -complete linkage-). La distancia entre dos clústeres es la máxima de las distancias entre todas las parejas posibles de miembros de un cluster con respecto a miembros del otro.
            - M. de **Ward**. La agrupación no se basa en las distancias entre clústeres, sino minimizando la **variabilidad intra-cluster**, buscando que cada cluster sea lo más homogéneo posible. La regla se traduce en asignar a un cluster el sujeto que produzca el menor incremento posible en la suma de cuadrados error. Una vez que un sujeto ha sido asignado a un cluster ya no puede ser excluido del mismo.
            - Los resultados obtenidos se representan en un gráfico tipo **dendrograma:** distancia de agrupamiento vs. sujetos. Un árbol lógico que indica visualmente la secuencia en que se han ido formando los clústeres. Permite tomar una decisión clara respecto del número

... -> Guardar -> Solución única: 2 grupos -> Continuar -> ...

de cluster que es posible formar. Alternativamente se puede usar un diagrama de témpanos. Se puede complementar con la opción de forzar una solución única con un número determinado de cluster de cara a ayudar en la decisión del nº de cluster y con el método de dos fases. En este último, el algoritmo nos informa del número óptimo de cluster (para decidir si puede o no haber divisiones razonables).

Analizar -> Clasificar -> Conglomerados en dos fases -> Variables: stai_est, ..., evaneg_t -> Graficos -> Gráfico del porcentaje intra-conglomerado, Gráfico de sectores de los conglomerados > Continuar→ Resultados → Descriptivos por conglomerado, Crear variable del Conglomerado de pertenencia → Continuar -> Aceptar (análisis)

- **De división, tipo descendente.** Se asume que todos los individuos son miembros de un mismo cluster para luego asignarlos a diferentes grupos. Técnicas con cómputo excesivo que están en desuso.
  a. Variantes.
     i. **Distancia astilla-promedio** (Splinter-Average Distance). Se separan los sujetos más distantes para formar dos grupos (Grupo Propio y Grupo Otro) y en 2º paso se calculan dos distancias, sujeto a su Gp y Sujeto a Go y si dGp> dGo, se segrega de su grupo y se añade a Go. Así se procede hasta llegar a un único cluster con todos los sujetos.
     ii. **Detección automática interacción** (Automatic Interaction Detection, AID. Se puede realizar en el módulo "Data Mining" de Statistica). Dividir el grupo inicial en dos subgrupos según Máximo SCEG/SCT de entre todas las variables medidas.

ii. **De partición.** Es también un método iterativo**.** No suponen que la asignación de un sujeto a un cluster sea irrevocable sino que ésta puede ir cambiando. Es necesario establecer de antemano los grupos que se quieren establecer.

Analizar -> Clasificar -> Conglomerados de K-Medias -> Variables: stai_est, ..., evaneg_t -> Iterar y Clasificar -> Número máximo de iteraciones: 10 -> Criterio de convergencia: 0 -> Usar medias actualizadas -> Continuar -> Centros -> Escribir finales en archivo (indicar archivo) -> Opciones -> Información sobre el conglomerado de cada caso -> Continuar -> Aceptar (análisis)

- **K-Medias.** Obtiene los grupos especificados por el investigador, asignando el sujeto a aquél grupo del que está más cerca según la distancia **euclídea** entre sujeto y centroides de grupo. Los centroides iniciales se corresponden con los individuos más **distantes**. Una vez seleccionados los centroides iniciales, se hace la 1ª asignación con el principio general y se recalculan los centroides. Así se procede en **varias iteraciones** (en torno a 20) reasignando sujetos y recalculando centroides. De esta forma con cada iteración los centroides experimentarán un **desplazamiento** y el error en cada cluster irá disminuyendo, cuestiones que conviene analizar con detenimiento.
- **K-Medoides**. El centroide lo constituye uno de los sujetos del cluster y en cada iteración se sustituye el centroide cambiando de sujeto si se reduce el error dentro del cluster. Es de uso minoritario e inadecuado con grandes muestras de sujetos.

iii. El tipo de técnica vendrá marcado por los objetivos de investigación. Los **jerárquicos** son preferibles cuando no se tiene una idea precisa de qué patrones de respuesta se espera encontrar en los sujetos y cuando pudiera haber puntos extremos. Cuando se opta por los de partición, emplear como punto de partida del algoritmo las soluciones

aportadas por los jerárquicos y evitar iniciarlo a partir de extracciones aleatorias o según los primeros de la lista.

iv. Entre los jerárquicos, los métodos más frecuentes son los del promedio y los del centroide pero se ven muy afectados por el problema de **encadenamiento**. Para minimizarlo, el M. vecino más lejano cuando se pretende construir clústeres en los que los sujetos sean muy semejantes entre sí. El M. Ward es preferible cuando se pretende que sean homogéneos y de aproximadamente el mismo número de sujetos.

v. Mejor complementarlas: Con las jerárquicas se produce una solución exploratoria y se emplean las de partición para confirmar dicha solución.

- La decisión sobre la técnica de agrupamiento, al igual que sucede con el tipo de medida, puede determinar los agrupamientos que se realicen.

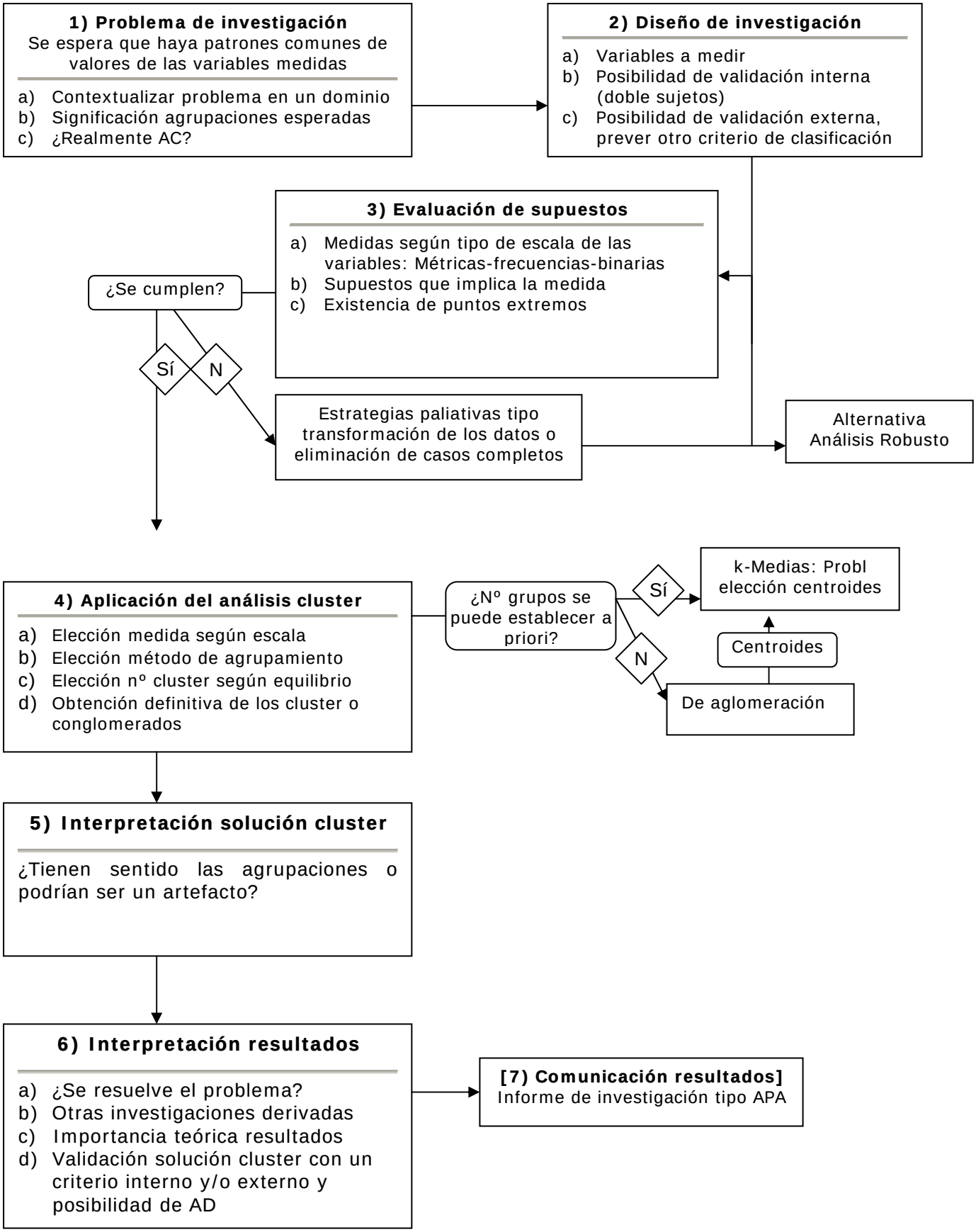
6. Evaluar la **bondad de ajuste** del agrupamiento efectuado. Fundamentalmente decidir el número de agrupaciones adecuado y la contribución de variables al agrupamiento.
   a. Para evaluar la solución de agrupamiento alcanzada, así como si el número de clústeres seleccionado es adecuado o no. Esta es una decisión difícil pues el número de grupos no está determinado a priori al no haber criterio objetivo de clasificación. En las jerárquicas este número puede definirse a posteriori pero en las no jerárquicas este número debe ser especificado antes de aplicar la técnica. El problema es que las técnicas de agrupamiento no proporcionan ningún criterio objetivo al respecto, aunque las jerárquicas sí lo facilitan.
   b. Se obtienen para cada etapa del análisis de manera que podamos decidir el punto en el que quedaría la agrupación. No se implementan en la mayoría de los programas del tipo SPSS (ver Catena, Ramos y Trujillo, 2003).
   c. Variantes.
      i. **Raíz Cuadrática Media** de las Desviaciones Típicas (RMSSTD). Desviación típica promedio de todas las variables que forman un cluster. Sólo se puede interpretar de forma relativa, buscando **el menor** RMSSTD, es decir un punto de equilibrio entre n y 1 cluster.
      ii. **$R^2$**. Cociente de la suma de cuadrados entre los cluster y suma de cuadrados total. Proporción de variabilidad total explicada por los clusters. Es semejante a la medida **RPE** del Modelo Lineal General. Lo deseable es un valor elevado.
      iii. **$R^2$ Semiparcial**. Pérdida de homogeneidad que se produce cuando se añade un nuevo sujeto en la etapa actual, respecto de la etapa precedente. Mejor valores de menor magnitud (estamos incluyendo en el cluster un sujeto que es bastante semejante a los que ya estaban).
      iv. La **distancia entre dos cluster**. Se computa alguna de las medidas indicadas pero en este caso para determinar la distancia entre cluster (por Pares) en lugar de hacerlo entre sujetos. Esta distancia no debe ser grande para que los dos clústeres puedan mezclarse.

7. **Interpretación** de la solución obtenida. Siempre que se aplique la técnica es posible obtener una solución de agrupamiento de sujetos (o de variables), pero no siempre ese agrupamiento tiene sentido. Es posible que los grupos obtenidos no existan en realidad, a pesar de lo cual, se obtendrá una solución de tantos grupos como se desee.

8. La solución de agrupación puede ser contrastada, validada, con respecto a un criterio interno o bien externo.
    a. Validación **cruzada o replicación**, siempre que el número de sujetos sea suficientemente grande
        i. La muestra se divide en dos mitades de forma aleatoria.
        ii. Se hace un AC sobre la 1ª mitad y se obtienen los centroides de cada cluster.
        iii. Se calculan las distancias de cada sujeto de la 2ª muestra a cada uno de los centroides de los clústeres de la 1ª y así agrupar los sujetos de la segunda muestra según las características de la primera.
        iv. AC de la segunda muestra.
        v. Se comparan las 2 medidas de agrupamiento de la 2ª para evaluar el grado de **acuerdo** de las dos soluciones. A mayor acuerdo, más evidencia a favor de la estabilidad de la solución.
    b. Alternativamente aplicar un **AD** para establecer si los grupos obtenidos son realmente diferentes y cuáles son las variables que contribuyen a esa diferencia.
    c. Validación **externa**. Valorar la solución del AC en comparación a un criterio externo a la investigación (i.e. correlación entre la clasificación de los jueces sobre los sujetos y la obtenida mediante el AC).

- El análisis cluster es una técnica que corre un riesgo importante, su mal uso, desembocando en soluciones “vacías de contenido psicológico”.

# 4. Secuencia de investigación en el Análisis de Cluster o Conglomerados



Volver Principio