Clase 12
Análisis de conglomerados II

Análisis Avanzado de Datos

Gabriel Sotomayor

Recordatorio de la clase anterior

Tipos de técncias multivariadas

En el análisis multivariable, las técnicas se pueden clasificar en **técnicas de dependencia** y **técnicas de interdependencia**. La selección de la técnica adecuada depende de la naturaleza de la pregunta de investigación y la relación entre las variables.

Las **técnicas de dependencia** se utilizan cuando existe una clara distinción entre variables dependientes (o respuesta) y variables independientes (o predictoras). Es decir, hay una relación de dependencia que se quiere modelar y analizar. El objetivo principal de estas técnicas es estimar el efecto de las variables independientes sobre las dependientes.

Las **técnicas de interdependencia** se aplican cuando no existe una distinción clara entre variables dependientes e independientes, o no es necesario establecer una relación de dependencia. En lugar de eso, todas las variables se tratan de igual forma, y el objetivo es descubrir patrones o estructuras ocultas en los datos.

Introducción al Análisis de Conglomerados

El análisis de conglomerados (o “cluster analysis”) es una técnica que se ubica dentro de las técnicas multivariables de clasificación. Su objetivo es agrupar datos en grupos llamados conglomerados, donde los integrantes de un conglomerado son lo más similares posibles entre sí y diferentes de los otros grupos.

El análisis de conglomerados se utiliza ampliamente en diferentes disciplinas debido a su capacidad para identificar estructuras ocultas dentro de los datos. Esta técnica se enfoca principalmente en la exploración de datos, sin tener hipótesis a priori sobre la estructura de los mismos.

Objetivos del Análisis de Conglomerados

1. **Desarrollar tipologías o clasificaciones de datos.**
    - Permite identificar patrones dentro de los datos y agruparlos según características compartidas.
2. **Buscar esquemas conceptuales útiles para agrupar entidades.**
    - Estos esquemas pueden utilizarse para comprender mejor el comportamiento de los datos y realizar predicciones.
3. **Generalizar hipótesis a través de la exploración de datos.**
    - Explorar datos sin hipótesis previas y generar nuevas hipótesis basadas en los conglomerados formados.
4. **Validar tipos definidos a través de otros procedimientos.**
    - Se pueden utilizar técnicas complementarias para validar la calidad de los conglomerados, como el análisis discriminante o pruebas estadísticas.

Etapas del Análisis de conglomerados

Pasos Principales en el Análisis de Conglomerados

En la ejecución de un análisis de conglomerados se siguen una serie de fases que pueden resumirse en los siguientes pasos:

1. **Selección de Variables**:
    - Selección de las variables que favorezcan la agrupación de los datos. Esta es una decisión clave y previa a cualquier análisis de conglomerados, ya que las variables elegidas determinan las características de clasificación que identifican a cada conglomerado.
2. **Selección del Procedimiento de Conglomeración**:
    - Decidir qué tipo de procedimiento de conglomeración se seguirá, ya sea jerárquico y/o no jerárquico, junto con el algoritmo específico de clasificación para la creación de los conglomerados.

Pasos Principales en el Análisis de Conglomerados

3. **Elección de Medidas de Distancia y Proximidad**:
    - Elegir las medidas de distancia y proximidad para proceder a la formación de los conglomerados. Esta elección depende en gran medida de la naturaleza de las variables incluidas en el análisis. Para variables métricas, hay más opciones disponibles, mientras que para variables no métricas (nominales u ordinales), se utilizan medidas de co-ocurrencia.
4. **Decisión sobre el Número de Conglomerados**:
    - Determinar el número de conglomerados que se constituirán, basado en criterios estadísticos, visualizaciones (como dendogramas) o criterios teóricos.

Pasos Principales en el Análisis de Conglomerados

5. **Presentación e Interpretación de Resultados**:
    - Presentar e interpretar los resultados obtenidos, tanto en su forma numérica (tablas de aglomeración) como gráfica (dendogramas y gráficos de témpanos).
6. **Validación de Resultados**:
    - Validar los resultados del análisis. Si los resultados no alcanzan la calificación de válidos, se deben introducir modificaciones para mejorar la solución, lo cual implica repetir el proceso desde el principio, revisando las decisiones previas.

Estas fases se pueden agrupar en cuatro bloques principales: selección de variables, procedimiento de conglomeración, interpretación de resultados, y validación de los hallazgos, que se repiten iterativamente hasta alcanzar una solución adecuada.

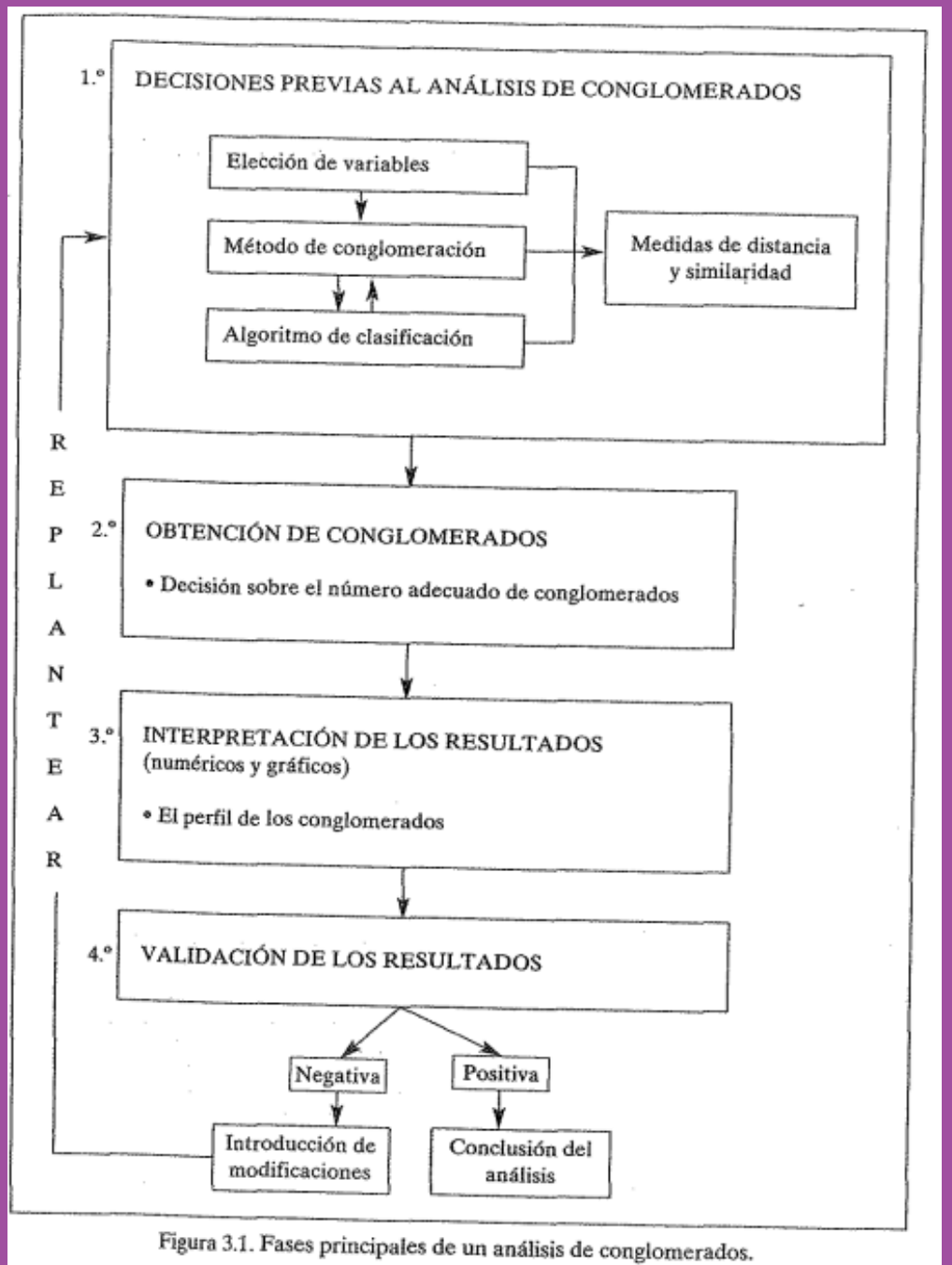


Figura 3.1. Fases principales de un análisis de conglomerados.

Elección de Variables

La elección de las variables es crucial ya que determina la calidad de la agrupación. Las decisiones deben basarse en la naturaleza de los datos y el objetivo de la investigación:

- **Relevancia**: Se deben elegir variables que sean relevantes para los objetivos del estudio.
- **Colinealidad**: Evitar incluir variables muy correlacionadas, ya que pueden influir demasiado en la formación de los conglomerados.
- **Escalado**: Variables con rangos muy diferentes deberían ser estandarizadas para evitar sesgos, ya que el análisis de conglomerados es muy sensible a las escalas de las variables.
    - **Estandarización**: Convertir las variables a puntuaciones Z para que tengan media cero y desviación estándar uno.

Métodos Jerárquicos vs No Jerárquicos

Métodos Jerárquicos

Los **métodos jerárquicos** son aquellos en los que los conglomerados se forman de manera jerárquica. Esto significa que el proceso de agrupamiento se realiza en una serie de pasos sucesivos, donde cada observación se va uniendo progresivamente a conglomerados más grandes, o conglomerados se van dividiendo en grupos más pequeños. Los métodos jerárquicos pueden ser **aglomerativos** o **divisivos**.

Métodos Jerárquicos

- **Aglomerativos**: Forman conglomerados empezando desde objetos individuales y uniéndolos en conglomerados mayores hasta que todos los objetos se encuentran en un solo conglomerado. Se utiliza un criterio de distancia para decidir qué objetos agrupar.
  - Algoritmos comunes: Método de Ward, Método del centroide, Distancias mínimas/máximas.
  - **Ventaja**: Permite observar la estructura completa de los conglomerados.
  - **Desventaja**: Puede ser computacionalmente costoso en grandes volúmenes de datos.
- **Divisivos**: Comienzan con un solo conglomerado que incluye todos los objetos, y sucesivamente dividen los conglomerados hasta que cada objeto pertenece a su propio conglomerado.

Métodos No Jerárquicos

Los **métodos no jerárquicos** no siguen un proceso jerárquico de agrupamiento, sino que intentan dividir el conjunto de datos en un número predefinido de conglomerados. Estos métodos asignan iterativamente los datos a conglomerados con base en una medida de similitud, buscando optimizar la homogeneidad dentro de los conglomerados y la heterogeneidad entre ellos.

- **K-Means**: Agrupa datos en un número predefinido de grupos. El algoritmo selecciona aleatoriamente los centroides iniciales y, luego, cada dato se asigna al centroide más cercano. Los centroides se recalculan iterativamente hasta que no se producen más cambios significativos.
  - Se eligen los centroides aleatoriamente y los datos se reasignan iterativamente hasta minimizar la variación intragrupal.
  - **Ventaja**: Es eficiente en grandes conjuntos de datos.
  - **Desventaja**: Es sensible a los valores iniciales y puede converger a soluciones subóptimas si no se eligen buenos centroides.

Distancia en el Análisis de Conglomerados

- Las decisiones sobre cuáles objetos combinar para formar un conglomerado se basan en estas matrices.
- Dependiendo del tipo de variables, se utilizan diferentes medidas para calcular las distancias o similaridades (por ejemplo, medidas euclidianas para variables continuas o coeficientes de Jaccard para datos binarios).
- **Tipos de Medidas**:
  - **Medidas de Distancia**: Se enfocan en cuán diferentes son los objetos en términos de magnitudes. Ejemplo: Distancia euclidiana.
  - **Medidas de Similaridad**: Enfocadas en patrones comunes entre objetos. Ejemplo: Correlación de Pearson.

Distancia Euclidiana

Distancia Euclidiana

- **Distancia Euclidiana al Cuadrado**: Frecuentemente usada en algoritmos de conglomerados como el método del centroide y el método de Ward. Se define como la suma de los cuadrados de las diferencias entre los valores de las variables de los objetos.
  - **Problema de Escalabilidad**: Si las variables están en diferentes unidades de medida, las variables con valores mayores pueden influir más en el cálculo de la distancia. La solución común es la estandarización de las variables.

Estandarización

- **¿Cuándo Estandarizar?**: La estandarización de las variables es recomendable cuando las variables tienen diferentes unidades de medida o rangos significativamente distintos. Por ejemplo, si una variable está medida en ingresos anuales (en miles) y otra en edad (en años), la variable con valores más altos dominará el cálculo de la distancia.
- **Efecto de la Estandarización**:
  - **Reducción de Influencia Desigual**: La estandarización reduce la influencia de variables con grandes valores numéricos, asegurando que todas las variables contribuyan de manera equitativa al cálculo de la distancia.
  - **Posibles Desventajas**: La estandarización puede minimizar diferencias que podrían ser relevantes en el contexto del análisis. Esto es especialmente cierto cuando la variabilidad en una variable tiene un significado sustancial importante.

Obtrención de conglomerados

A la elección de la medida de similaridad o de distancia le sigue la obtención de la solución de conglomerados, en conformidad con las diversas decisiones adoptadas. Es decir, el método de conglomeración, el algoritmo de clasificación y la medida de similaridad o distancia. Antes de proceder a la interpretación de los resultados, se debe dirimir una cuestión crucial: la referente al número de conglomerados a retener entre las distintas alternativas posibles de clasificación de los objetos de interés.

En la conglomeración no jerárquica, la decisión sobre el número de conglomerados a retener es previa a la ejecución de cualquier análisis. En la conglomeración jerárquica, sin embargo, esta decisión se toma al final del análisis, una vez que todos los conglomerados han sido formados. De ahí que se incluya esta discusión en este apartado posterior a la exposición de decisiones clave previas al análisis de conglomerados.

Elección del Número de Conglomerados

La finalidad de todo análisis de conglomerados es la clasificación de una serie de objetos en conglomerados (o grupos) homogéneos. Pero, ¿cuántos conglomerados se requieren para describir de forma precisa la similitud y la diversidad en una población?

No existe una respuesta única para esta cuestión, pero existen varios procedimientos alternativos que se pueden aplicar para determinar el número idóneo de conglomerados:

- **Criterios Teóricos**: Seguir algún criterio teórico que fundamente la elección de un número de conglomerados específico. Es recomendable probar diferentes soluciones de clasificación y luego elegir aquella que tenga mayor significado teórico y estadístico.
- **Coeficientes de Conglomeración**: En la conglomeración jerárquica, se pueden observar los coeficientes de conglomeración, también llamados “coeficientes de fusión” o “coeficientes de aglomeración”. Estos coeficientes indican el valor numérico (medida de distancia o similitud) que propicia la unión de objetos para formar conglomerados. La solución idónea corresponde al número de conglomerados previo al “salto” apreciable en el valor del coeficiente de conglomeración.
- **Dendograma**: Además, la información proporcionada por el dendograma es uno de los gráficos más característicos en la conglomeración y puede ser útil para la determinación del número de conglomerados

Cuando se observan varias variaciones o “saltos” en los coeficientes, puede resultar difícil decidir cuál es relevante. Esta subjetividad es una crítica frecuente a los métodos de conglomeración jerárquica.

Presentación de resultados de conglomerados

Análisis de conglomerados Jerárquico

Historial de Conglomeración

- El historial de conglomeración es una tabla que resume el proceso de constitución de los conglomerados.
- Incluye las siguientes columnas:
  - **Etapa**: número de la etapa del análisis.
  - **Objetos combinados**: los dos objetos o conglomerados que se combinan en cada etapa.
  - **Coeficiente**: valor de la medida de distancia o similitud utilizada.
- **Interpretación del Historial**:
  - Ayuda a determinar el número óptimo de conglomerados observando los saltos grandes en los coeficientes.
  - Útil para identificar conglomerados atípicos que se unen en etapas tardías.

Análisis de conglomerados Jerárquico

Historial de Conglomeración

	Paso	Cluster1	Cluster2	Altura
1	1	-3	-14	0.4412289
2	2	-8	-9	0.4419112
3	3	-10	-17	0.4962598
4	4	-16	3	0.7013850
5	5	-11	-12	0.7960683
6	6	-2	-22	0.7978042
7	7	-1	-5	0.8750157
8	8	-20	2	0.9021514
9	9	-18	4	0.9512569
10	10	1	8	1.0842550
11	11	-19	7	1.3545367
12	12	-15	9	1.5607937
13	13	-6	6	1.5690875
14	14	10	12	1.8440775
15	15	-13	5	2.1954467
16	16	-7	11	2.6043262
17	17	14	15	2.9425617
18	18	13	16	2.9468935
19	19	-21	18	4.3467653
20	20	-4	17	4.9913026

Análisis de conglomerados Jerárquico

Conglomerado de Pertenencia

- Tabla que muestra la asignación de objetos a conglomerados.
- Puede obtenerse para una única solución o un rango de soluciones definido previamente.
- **Utilidad**: Permite conocer a qué conglomerado pertenece cada objeto, facilitando la interpretación del análisis.

Análisis de conglomerados Jerárquico

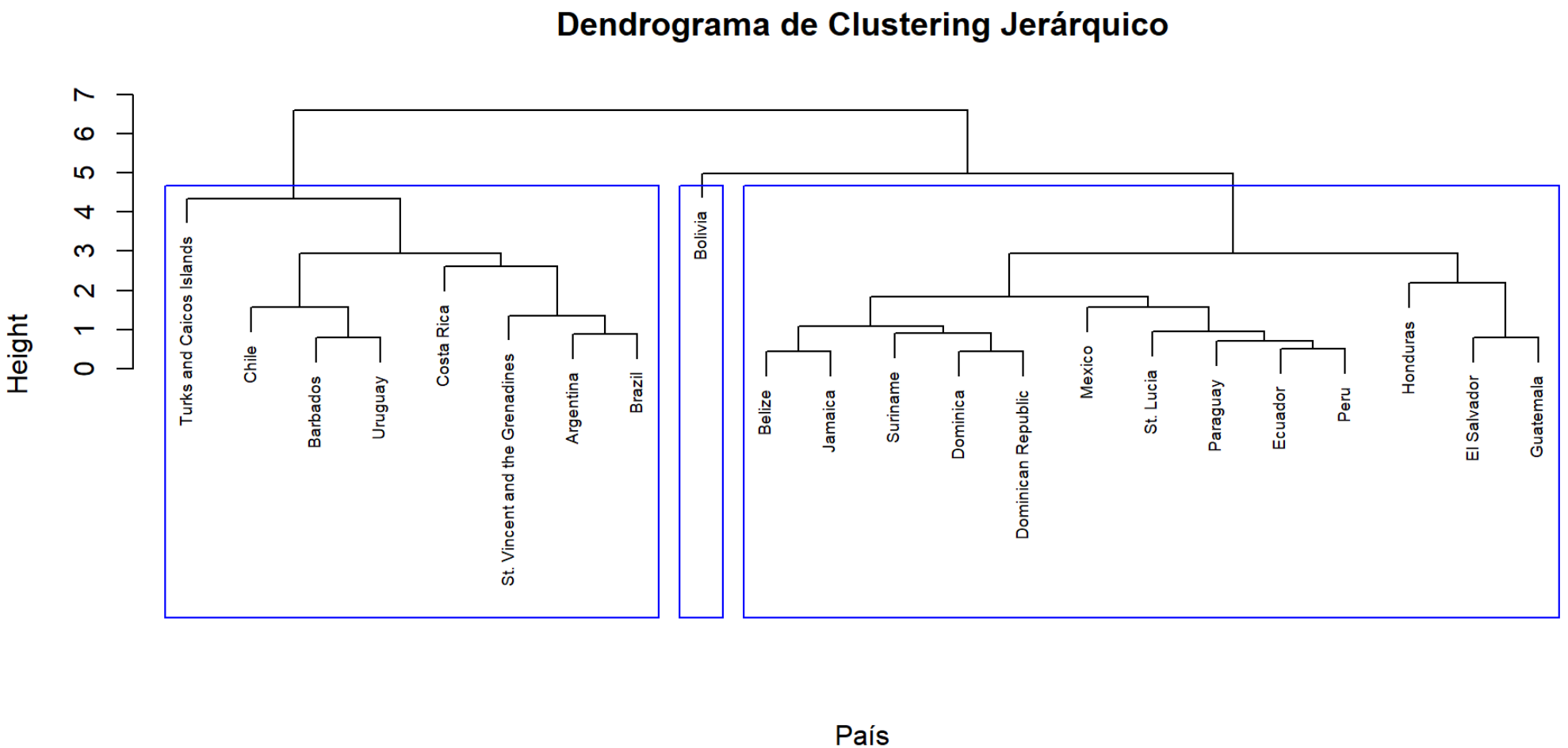
Conglomerado de Pertenencia

	country	conglomerado
2	Argentina	1
5	Barbados	1
6	Belize	2
7	Bolivia	3
8	Brazil	1
11	Chile	1
13	Costa Rica	1
16	Dominica	2
17	Dominican Republic	2
18	Ecuador	2
19	El Salvador	2
21	Guatemala	2
24	Honduras	2
25	Jamaica	2
26	Mexico	2
29	Paraguay	2
30	Peru	2
34	St. Lucia	2
36	St. Vincent and the Grenadines	1
37	Suriname	2

Dendograma

- Muestra el proceso de aglomeración en un formato jerárquico.
- Ayuda a identificar el número de conglomerados adecuados.
- **Interpretación**:
  - Las ramas más largas indican conglomerados que se formaron más tarde en el proceso, lo que puede ser un indicador de conglomerados menos homogéneos.

Dendograma


Dendrograma de Clustering Jerárquico
Height
0 1 2 3 4 5 6 7
Turks and Caicos Islands
Chile
Barbados
Uruguay
Costa Rica
St. Vincent and the Grenadines
Argentina
Brazil
Bolivia
Belize
Jamaica
Suriname
Dominica
Dominican Republic
Mexico
St. Lucia
Paraguay
Ecuador
Peru
Honduras
El Salvador
Guatemala
País

Análisis de Conglomerados No Jerárquico

Tablas de Resultados No Jerárquico

Centros de los Conglomerados

- Los **centroides** de cada conglomerado se calculan iterativamente.
- Se presenta información de los centros **iniciales** y **finales**.
- La medida de distancia utilizada para reasignar los objetos es la **distancia euclídea**.
- **Iteraciones**:
  - Los centroides iniciales y finales normalmente no coinciden, reflejando la evolución de la composición de los conglomerados durante el proceso.
  - Cada iteración busca minimizar la distancia total dentro de los conglomerados.

Tablas de Resultados No Jerárquico

Centros de los Conglomerados

Caracterización de los Conglomerados I

Conglomerado	Adopción Homoparental	Aborto	Rol del Estado en Educación	Capitalización Individual Pensiones	Restriccione Migratoria
1	3.365796	2.000000	4.004751	2.546318	4.23753
2	3.893424	2.909297	4.074830	3.682540	4.28117
3	3.920981	4.152589	4.305177	2.629428	3.90190
4	1.714286	1.864169	3.946136	3.252927	4.28337

Tablas de Resultados No Jerárquico

Tabla ANOVA

- Evalúa la contribución de cada variable a la diferenciación entre los conglomerados.
- La tabla ANOVA incluye:
  - **Medias Cuadráticas entre conglomerados** y **dentro de conglomerado**.
  - **Estadístico F**: indica qué variables diferencian mejor entre los conglomerados.
- **Interpretación**:
  - Un **valor F** alto sugiere que la variable tiene un papel importante en la diferenciación de los conglomerados.
  - **Significancia estadística**: Se utiliza para identificar qué variables deben ser consideradas al interpretar las características distintivas de los conglomerados.

Tablas de Resultados No Jerárquico

Tabla ANOVA

```r
# ANOVA para la variable Aborto
anova_aborto <- aov(c37_02 ~ conglomerado, data = elsoc)
summary(anova_aborto)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
conglomerado    1    7.6   7.605    5.43 0.0199 *
Residuals    1654 2316.2   1.400
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# ANOVA para la variable Rol del Estado en Educación
anova_educacion <- aov(c37_03 ~ conglomerado, data = elsoc)
summary(anova_educacion)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
conglomerado    1      0  0.0080   0.012  0.913
Residuals    1654   1109  0.6703
```

Gráficos No Jerárquicos

Ubicación de Conglomerados y Centros

- Representación de la ubicación de cada conglomerado y sus centros.
- Muestra claramente la distribución espacial de los conglomerados.
- **Utilidad**:
  - Facilita la identificación de conglomerados bien separados, indicando una buena partición de los datos.
  - Permite identificar conglomerados solapados que podrían requerir un ajuste del número de conglomerados.

Gráficos No Jerárquicos

Ubicación de Conglomerados y Centros