TÉCNICAS DE CLASIFICACIÓN, ANÁLISIS CLÚSTER Y APLICACIÓN PRÁCTICA CON SPSS

Olga Blasco-Blasco
Área de Métodos Cuantitativos
Dpto. Economía Aplicada
Universitat de València

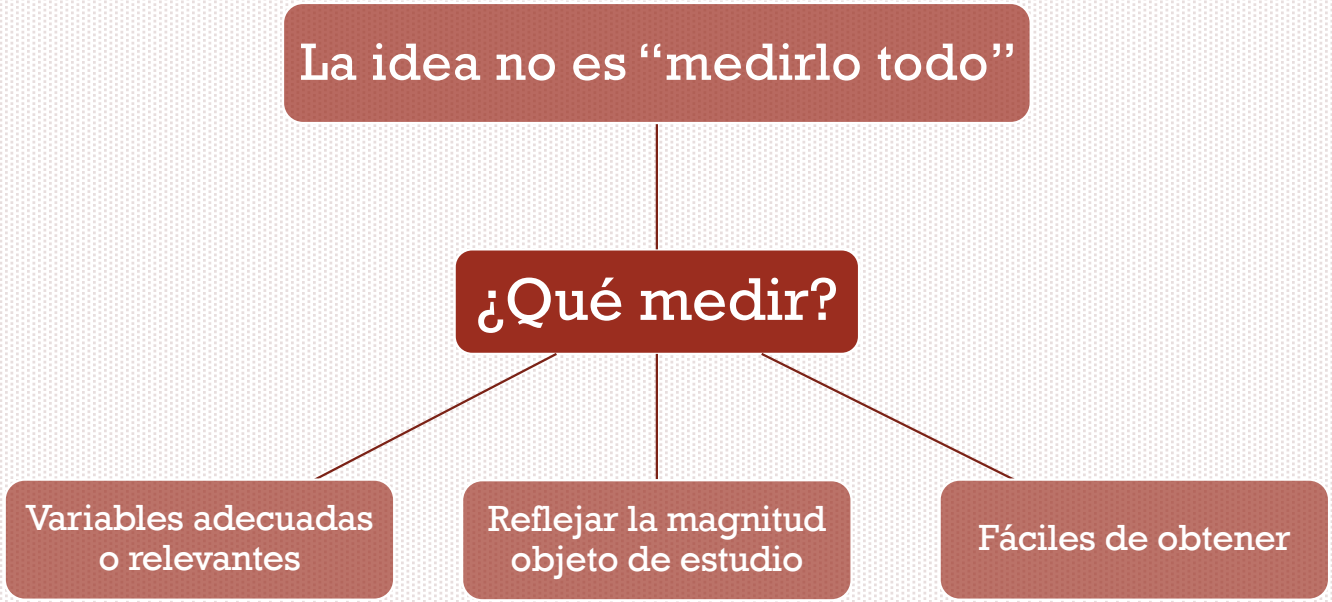
INTRODUCCIÓN

En comunicación, cada vez se hace más necesario recopilar datos, estadísticas, … que se producen, consumen y divulgan.

Existe, además un compromiso de llevar a cabo acciones que mejoren la situación inicial… Y MEDIR O CUANTIFICAR ESTAS ACCIONES.

INTRODUCCIÓN

En nuestro contexto, **medir** es asignar una cantidad a una actividad o a una acción, de manera que se pueda comparar con un patrón establecido previamente.



DATO ESTADÍSTICO: Par formado por una unidad observada y su correspondiente **característica**. A cada hecho le corresponde un n° real.
v. gr.: Un individuo y su edad.

VARIABLE ESTADÍSTICA: Se refiere a una característica que puede tomar cualquier modalidad de un conjunto determinado o "dominio de la variable".

EJEMPLO DE VARIABLES

Variable	Descripción de la variable
Edad	Edad en años
Marital	Estado civil
Direcc	Años en la dirección actual
Ingres	Ingresos del hogar
Ingcat	Categoría de ingresos en miles
Coche	Precio del coche principal
Cochecat	Categoría del precio del coche principal
Educ	Nivel educativo
Empleo	Años en la empresa actual

CARACTERÍSTICAS OBSERVADAS

- **Categóricas**: Las respuestas pertenecen a grupos o categorías.

- **Variables numéricas**: Son medibles.
  - ***Discretas***: Su dominio es un conjunto numerable de valores. Proviene de un proceso de recuento.
  - ***Continuas***: Su dominio es continuo, es decir, cuando entre dos valores la variable puede tomar, al menos teóricamente, cualquiera de los infinitos valores existentes entre ellos. Proviene de un proceso de medición.

CARACTERÍSTICAS OBSERVADAS

- **Cualitativas**: No se pueden medir.
  - ***Nominales o Atributos***: establecen diferentes categorías.
  - ***Ordinales***: es posible establecer cierta ordenación entre las diferentes categorías.
- **Cuantitativas**: Son medibles. Se refiere a los datos en una escala ordenada.
  - ***Escala de intervalos***: Indica el orden y la distancia con respecto a un cero arbitrario. (V. gr.: La temperatura o el año).
  - ***Razón***: Indican el orden y la distancia respecto al cero natural. (V. gr.: Peso, edad)

CLASIFICACIÓN SEGÚN LOS DATOS

- **Unidimensionales o Univariante:** Se consideran los elementos de UNA característica (variable).

- **Multidimensionales o Multivariante:** Se consideran conjuntamente VARIAS características.

- **Temporales o Cronológicos:** Evolución de la característica a lo largo del tiempo.

- **Atemporales o Cross-Section:** Corte transversal. Serie estudiada en un momento determinado del tiempo.

ESTADÍSTICA

Ciencia que estudia cómo debe emplearse la información y pretende dar una guía de acción en situaciones prácticas que entrañan incertidumbre, además de ofrecer métodos de investigación aplicables al resto de Ciencias.

Estadística Descriptiva

Recogida de información, descripción y análisis de un grupo de datos utilizando métodos numéricos y gráficos.

Inferencia Estadística

A partir de los datos de una muestra, basándose en el cálculo de probabilidades, se efectúan estimaciones, decisiones y predicciones.

Permite generalizar sobre un conjunto mayor de datos (población).

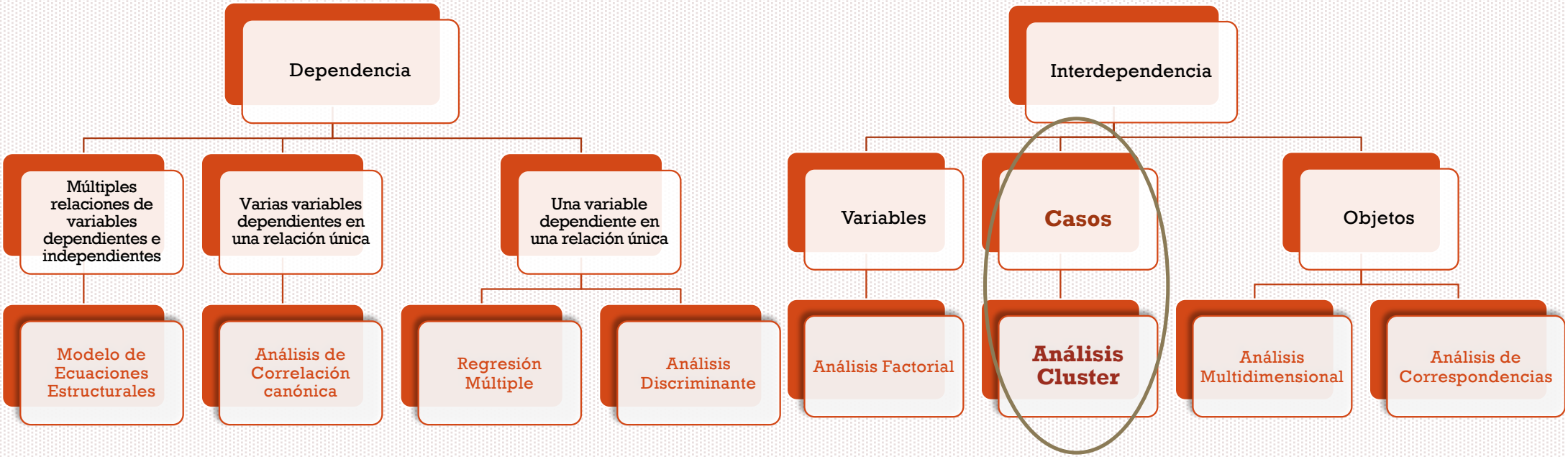
- Estimación
- Contraste de hipótesis
- **Análisis Multivariante**

Machine Learning

Rama de la Inteligencia Artificial que proporciona a las máquinas capacidad de aprendizaje, mediante algoritmos capaces de identificar patrones en grandes bases de datos.

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje semi-supervisado
- Aprendizaje de refuerzo

RELACIÓN ENTRE LAS VARIABLES



Fuente: Hair et al. (1999)

ANÁLISIS CLUSTER. INTRODUCCIÓN

Objetivo: Dividir un conjunto de observaciones en dos o más grupos, basándose en la semejanza entre los individuos.

- El análisis de conglomerados o análisis cluster es una técnica diseñada para clasificar distintas observaciones en grupos, mediante una agrupación de individuos, de forma que :

  1. Cada grupo sea homogéneo respecto a las variables utilizadas en el estudio.
  2. Que los grupos sean lo más diferente posible.

- Es importante señalar que la composición de los grupos es desconocida a priori y es necesario determinarlos a partir de las observaciones.

- ***Selección de variables***:

  1. Se toman aquellas que caractericen a los individuos que van a clasificarse.
  2. Deben estar vinculadas con los objetivos de la clasificación.

- En el análisis cluster, a diferencia de otras técnicas, no se dispone de herramientas que permitan saber qué variables son las relevantes, por lo que se forman los grupos atendiendo a todas las variables.

CÓMO PROCEDER EN EL ANÁLISIS DE CONGLOMERADOS:

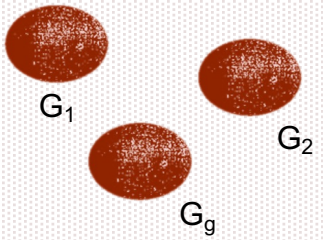
1. Se dispone de n observaciones de las que tiene información sobre k variables (edad, estado civil, nº de hijos,…)

	X_1	X_2	…	X_K
O_1				
O_2				
.				
O_n				

2. Se establece un indicador que nos dice en qué medida cada par de observaciones se parece entre sí. Medida de distancia o similaridad.

	O_1	O_2	…	O_n
O_1				
O_2				
.				
O_n				

3. Se crean grupos, mediante un análisis de conglomerados jerárquico o no jerárquico, de forma que cada grupo contenga las observaciones que más se parezcan entre sí, según la medida de similaridad calculada.



4. Se describen los grupos obtenidos y se comparan unos con otros. Para ello se observan los valores promedio que toman las *k* variables en cada uno de los *g* subgrupos obtenidos.

	X_1	X_2	…	X_K
G_1				
G_2				
.				
G_g				

ANÁLISIS PREVIO DE LOS DATOS

Valores perdidos

Existencia de outliers

Estandarización de los datos

VALORES PERDIDOS

- La existencia de valores perdidos es prácticamente inevitable en CCSS:
  - Los entrevistados no quieren facilitar la información sobre alguna variable.
  - El encuestador se equivoca al marcar la casilla.
  - La información no está disponible en el momento de recoger la información.
- Los valores perdidos se pueden generar de manera aleatoria o puede responder a un patrón de conducta de un determinado grupo.

- **Diagnóstico de la existencia o no de aleatoriedad:**
  - Se toma la misma variable con los datos perdidos y sin ellos y se comprueba si tienen la misma media mediante un contraste de medias. Si es la misma habrá que asumir aleatoriedad.
  - Prueba de las correlaciones dicotomizadas. Se transforman los valores de la variable en 1 (si no toma valor perdido) y 0 en caso contrario y se calcula la matriz de correlaciones.

VALORES PERDIDOS

Tratamiento de los valores perdidos:

- ***Eliminación de todos los casos que tengan un valor perdido***. (El más utilizado). Limitación: Pérdida de información por la disminución de la muestra.

- ***Eliminación de casos por parejas***. Solo se eliminan los casos si existe el valor perdido en la variable que se está estudiando. Inconveniente: Distintos tamaños muestrales según la variable.

- ***Imputación***, es decir, sustitución del valor perdido por alguna estimación de su valor. Se puede utilizar, entre otros, la media o la regresión

Problemas de la imputación:

- Las estimaciones son "demasiado" coherentes.
- La varianza se reduce porque las estimaciones están próximas a la media.
- Sólo se pueden utilizar si caen en el rango de la variable.

OUTLIERS: CASOS ATÍPICOS

- Casos que toman valores extremos para una, dos o múltiples variables, que los hacen diferir del resto de la muestra.

- **Importancia de detectar los valores atípicos**.
  - Distorsionan los resultados. Modifican el patrón de comportamiento del resto de los casos y se obtienen conclusiones que sin ellos, serían completamente distintas.
  - Pueden afectar a la condición de aplicabilidad más habitual en la mayoría de las técnicas de análisis multivariante: *La Normalidad*.

- **Causas que generan los valores atípicos**.
  - Errores en los datos: Tanto en la recogida como en la introducción.
  - Errores intencionados en la contestación.
  - Errores en el muestreo. Individuos de una población distinta a la población objetivo.
  - Outliers legítimos, es decir, casos atípicos pertenecientes a la población objetivo.

OUTLIERS: CASOS ATÍPICOS

- Identificación de los outliers.
  - Error en la introducción de los datos: Se consulta el cuestionario original y se corrige.
  - Si el error está en el registro y no es anónima la encuesta, se vuelve a entrevistar.
  - Si es anónima, una alternativa es el valor medio de la variable.
  - Si se detecta un error evidente, se elimina el valor atípico.
  - Si el valor atípico es legítimo:
    - Algunos autores consideran que lo mejor es eliminarlo para asegurar que las estimaciones son correctas para la mayoría de la población.
    - Otros autores, proponen intentar suavizar su influencia transformando las variables mediante raíces cuadradas o logaritmos, lo que reduce su rango.
    - Otra alternativa es la utilización de técnicas estadísticas robustas, por ejemplo contrastes no paramétricos.

COMPROBAR SI SE CUMPLEN LOS SUPUESTOS DEL ANÁLISIS DE CLASIFICACIÓN:

- Ausencia de correlación entre las variables.
- Número de variables no muy elevado.
- Que las variables no se encuentren medidas en unidades diferentes.

ESTANDARIZACIÓN DE LOS DATOS

- Para evitar la influencia en la variable de la unidad en la que viene medida, se puede corregir el efecto estandarizando los datos. La estandarización elimina la identidad de cada variable y las transforma en unidades adimensionales.
- El SPSS ofrece distintas posibilidades, entre la que se destaca:
  - ***Puntuaciones Z***: Los datos estandarizados se obtiene restando al valor de cada observación la media y dividiendo el resultado por su desviación típica. De esta forma, las variables estandarizadas tienen media 0 y desviación típica 1.

ANÁLISIS CLUSTER. MEDIDAS DE SIMILARIDAD

Medidas de similaridad para variables métricas.

Medidas de similaridad para datos binarios

EJEMPLO 1

- Suponemos que se dispone de una pequeña muestra de 10 individuos a los que hemos preguntado acerca de las siguientes dos características culturales de su viaje a una determinada ciudad:
  - X_1: Mi visita a la ciudad es una manera de divertirme y escapar de la rutina diaria.
  - X_2: Estoy especialmente interesado/a en saber más cosas de la cultura de la ciudad.

Individuo	X_1	X_2
1	5	3
2	4	2
3	5	3
4	3	4
5	5	2
6	4	3
7	2	5
8	3	5
9	2	5
10	3	4

- Una medida de proximidad, semejanza o distancia es un coeficiente que mide la semejanza global entre cada par de observaciones.
- En función del tipo de variables que se estén analizando para caracterizar a los objetos, las medidas más adecuadas serán diferentes.

MEDIDAS DE SIMILARIDAD PARA VARIABLES MÉTRICAS (CUANTITATIVAS)

- Las medidas distancia más comunes para determinar la similaridad entre dos individuos son:

  - Distancia euclídea: $$d(i,j)=\sqrt{\sum_{h=1}^{k}\left(X_{ih}-X_{jh}\right)^2}$$
  - Distancia euclídea al cuadrado: $$d(i,j)=\sum_{h=1}^{k}\left(X_{ih}-X_{jh}\right)^2$$
  - Distancia city block: $$d(i,j)=\sum_{h=1}^{k}\left|X_{ih}-X_{jh}\right|$$
  - Distancia de Minkowski: $$d(i,j)=\left[\sum_{h=1}^{k}\left(X_{ih}-X_{jh}\right)^2\right]^{1/n}$$

EJEMPLO 1

- Cálculo de dos distancias.

Distancia euclídea: $d(1,2)=\sqrt{\sum_{h=1}^{k=2}\left(X_{1h}-X_{2h}\right)^2}=\sqrt{(5-4)^2+(3-2)^2}=\sqrt{2}=1{,}414$

Distancia euclídea al cuadrado: $d(1,2)=\sum_{h=1}^{k=2}\left(X_{1h}-X_{2h}\right)^2=(5-4)^2+(3-2)^2=2$

Distancia City Block: $d(1,2)=\sum_{h=1}^{k=2}\left|X_{1h}-X_{2h}\right|=\left|5-4\right|+\left|3-2\right|=2$

EJEMPLO 1

- Utilizando el paquete estadístico SPSS, podemos obtener fácilmente las matrices de distancia. Para ello, seleccionamos *Analizar/Clasificar/Conglomerados jerárquicos…*
- A continuación, se seleccionan las variables X_1 y X_2 y hacemos clic en el botón *Método* para establecer la medida de distancias (disimilaridad). En función del tipo de variable considerada en el análisis: métrica, de recuento o binaria, las medidas de distancia o asociación son unas u otras.

Matriz de distancias

	distancia euclídea al cuadrado									
Caso	1	2	3	4	5	6	7	8	9	10
1	,000	2,000	,000	5,000	1,000	1,000	13,000	8,000	13,000	5,000
2	2,000	,000	2,000	5,000	1,000	1,000	13,000	10,000	13,000	5,000
3	,000	2,000	,000	5,000	1,000	1,000	13,000	8,000	13,000	5,000
4	5,000	5,000	5,000	,000	8,000	2,000	2,000	1,000	2,000	,000
5	1,000	1,000	1,000	8,000	,000	2,000	18,000	13,000	18,000	8,000
6	1,000	1,000	1,000	2,000	2,000	,000	8,000	5,000	8,000	2,000
7	13,000	13,000	13,000	2,000	18,000	8,000	,000	1,000	,000	2,000
8	8,000	10,000	8,000	1,000	13,000	5,000	1,000	,000	1,000	1,000
9	13,000	13,000	13,000	2,000	18,000	8,000	,000	1,000	,000	2,000
10	5,000	5,000	5,000	,000	8,000	2,000	2,000	1,000	2,000	,000

Esta es una matriz de disimilaridades

MEDIDAS DE SIMILARIDAD PARA DATOS BINARIOS

- Si se trabaja con variables que solo toman valores (1) presencia o (0) ausencia de una característica, la forma de actuar para calcular las medidas de similaridad es la siguiente:

  - Se construye una matriz 2x2 para cada par de observaciones que se están comparando, donde se recogen las coincidencias y divergencias entre las distintas variables correspondientes a las dos observaciones comparadas.

		E_j	
		1	0
E_i	1	a	b
	0	c	d

MEDIDAS DE DISTANCIA PARA DATOS BINARIOS

- Las medidas de similitud más utilizadas para dos observaciones *i* y *j* cualesquiera son:
  - Distancia euclídea al cuadrado: $d(i,j)=b+c$
  - Distancia euclídea: $d(i,j)=\sqrt{b+c}$
  - Diferencia de tamaño: $d(i,j)=\frac{(b-c)^2}{(a+b+c+d)^2}$
  - Diferencia de configuración: $d(i,j)=\frac{bc}{(a+b+c+d)^2}$

EJEMPLO 2

- Cálculo de dos distancias con datos binarios.

Ejemplo: Uriel, E. y Aldás, J. (2005): Análisis Multivariante Aplicado. Madrid. Thomson.

Observaciones	Variables			
	X1	X2	X3	X4
E1	1	1	0	0
E2	0	1	1	1
E3	1	1	0	1
E4	0	0	0	1
E5	1	1	1	0

		E_1	
		1	0
E_2	1	1	2
	0	1	0

$$d(1,2) = b + c = 2 + 1 = 3$$

Matriz de distancias

Caso	distancia euclídea al cuadrado binaria				
	1	2	3	4	5
1	0	3	1	3	1
2	3	0	2	2	2
3	1	2	0	2	2
4	3	2	2	0	4
5	1	2	2	4	0

Esta es una matriz de disimilaridades

Matriz de distancias

Caso	diferencia de tamaño				
	1	2	3	4	5
1	,000	,063	,063	,063	,063
2	,063	,000	,000	,250	,000
3	,063	,000	,000	,250	,000
4	,063	,250	,250	,000	,250
5	,063	,000	,000	,250	,000

Esta es una matriz de disimilaridades

FORMACIÓN DE LOS GRUPOS: ANÁLISIS JERÁRQUICO DE CONGLOMERADOS

Una vez obtenida la matriz de distancias (disimilaridades), el siguiente paso en el análisis cluster consiste en formar los grupos. Para ello, debemos tomar dos decisiones:

1. Aplicar un método que, partiendo de la matriz de distancias, establezca la composición de los grupos, es decir, seleccionar un algoritmo o método de agrupación:

   - **Métodos jerárquicos.** En un primer momento cada individuo forma un grupo y, sucesivamente, vamos fusionando grupos que se encuentran cercanos entre sí.
   - **Métodos no jerárquicos.** Se establece un número de grupos a priori y los individuos se van clasificando en cada uno de esos grupos.

2. Determinar un número de grupos que sea razonable.

MÉTODOS JERÁRQUICOS

- Vinculación intergrupos
- Vinculación intragrupos
- Método del vecino más cercano
- Método del vecino más lejano
- Agrupación de centroides
- Método de Ward

VINCULACIÓN ENTRE GRUPOS

- Se calcula la distancia entre dos grupos como distancia media entre todos los pares de observaciones que pueden formarse tomando un miembro de un grupo y otro miembro de otro grupo.

$$d\{(a,b);c\}=\frac{d(a,c)+d(b,c)}{2}$$

$$d[(a,b);(c,d)\}=\frac{d(a,c)+d(a,d)+d(b,c)+d(b,d)}{4}$$

- Este método tiene la ventaja de que utiliza información de todos lo pares de puntos en cada decisión.
- Es el método que por defecto utiliza SPSS.

Ejemplo 1

Vinculación promedio (Entre grupos)

Historial de conglomeración

Etapa	Conglomerado que se combina		Coeficientes	Etapa en la que el conglomerado aparece por primera vez		Próxima etapa
	Conglomerado 1	Conglomerado 2		Conglomerado 1	Conglomerado 2	
1	4	10	,000	0	0	4
2	7	9	,000	0	0	8
3	1	3	,000	0	0	5
4	4	8	1,000	1	0	8
5	1	6	1,000	3	0	7
6	2	5	1,000	0	0	7
7	1	2	1,500	5	6	9
8	4	7	1,667	4	2	9
9	1	4	8,960	7	8	0

VINCULACIÓN INTRAGRUPOS O DENTRO DE GRUPOS

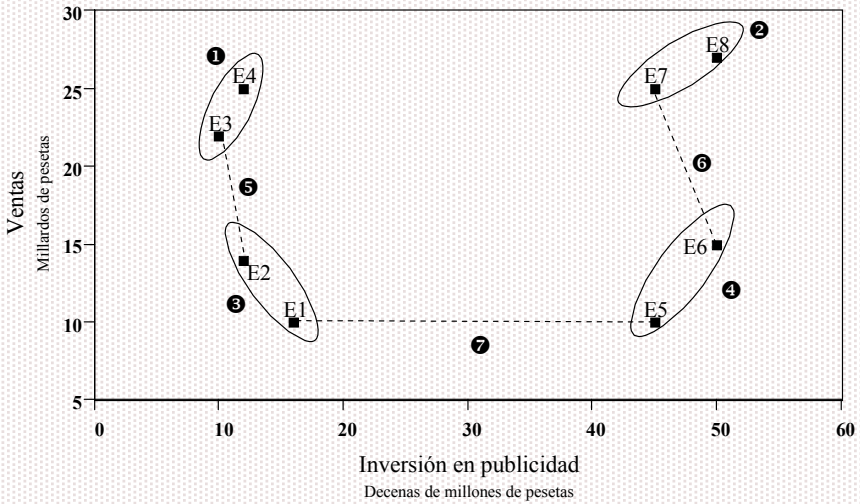
- Se basa en minimizar la distancia media entre todos los casos del cluster resultante.

$$d\{(a,b);c\}=\frac{d(a,b)+d(a,c)+d(b,c)}{3}$$

$$d\{(a,b);(c,d)\}=\frac{d(b,d)+d(c,d)+d(a,c)+d(a,d)+d(a,b)+d(b,c)}{6}$$

MÉTODO DEL VECINO MÁS CERCANO

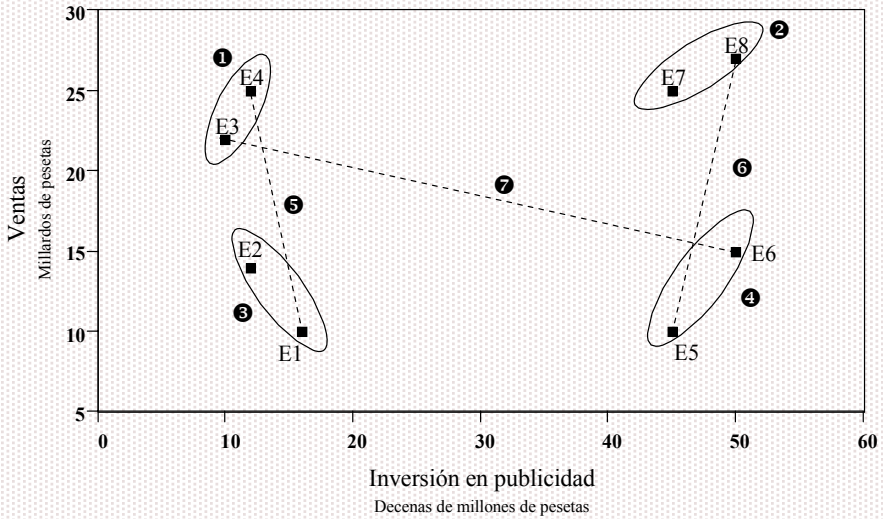
- El primer conglomerado está formado por los individuos más próximos.
- Volvemos a medir las distancias, pero cuando se mide la distancia entre el grupo formado y un individuo se toma la distancia mínima de los individuos del grupo al nuevo individuo.
- Al evaluar las distancias obtenidas y buscar la distancia mínima puede que al grupo formado en la etapa anterior se le añada un nuevo individuo o se forme un nuevo grupo de 2 observaciones.



Fte: Uriel, E. y Aldás, J. (2005): Análisis Multivariante Aplicado. Madrid. Thomson.

MÉTODO DEL VECINO MÁS LEJANO

- Este método es similar al anterior. Se parte de unir los dos casos con la distancia más próxima. En la segunda etapa se calculan nuevamente las distancias, pero se toma la distancia máxima de los individuos del grupo formado al nuevo individuo.



Fte: Uriel, E. y Aldás, J. (2005): Análisis Multivariante Aplicado. Madrid. Thomson.

AGRUPACIÓN DE CENTROIDES

- Se basa en el cálculo del centroide de cada grupo como la media de dicho grupo.
- A partir de las medias, la distancia entre dos grupos es la distancia entre sus centroides.
- En este método se consideran los dos individuos que se encuentran más cerca de manera que el grupo que forman se sustituye por la observación promedio (centroide).
- Se recalcula la matriz de distancias y se vuelven a unir los dos individuos que se encuentran más próximos. El proceso concluye cuando todos los individuos son asignados a un solo conglomerado.
- La principal ventaja del método de agrupación de centroides es que es menos sensible a los valores atípicos que otros métodos.

MÉTODO DE WARD

- También se le denomina *Método de la varianza mínima*.
- Lo que hace es buscar los dos conglomerados cuya unión conlleve el menor incremento de la varianza.
- En cada paso se debe probar con todas las posibles combinaciones de conglomerados.
- Para ello se calcula el valor del índice de la suma de los cuadrados y se selecciona aquel que tenga un menor valor.

OBSERVACIONES

- Los métodos centroide, de la mediana y el de Ward requieren el uso de la distancia euclídea al cuadrado como criterio. Los otros cuatro pueden utilizar cualquier medida de proximidad.

- De todos los métodos de agrupamiento, el más utilizado es el de vinculación entre grupos o intergrupos y el método de Ward.

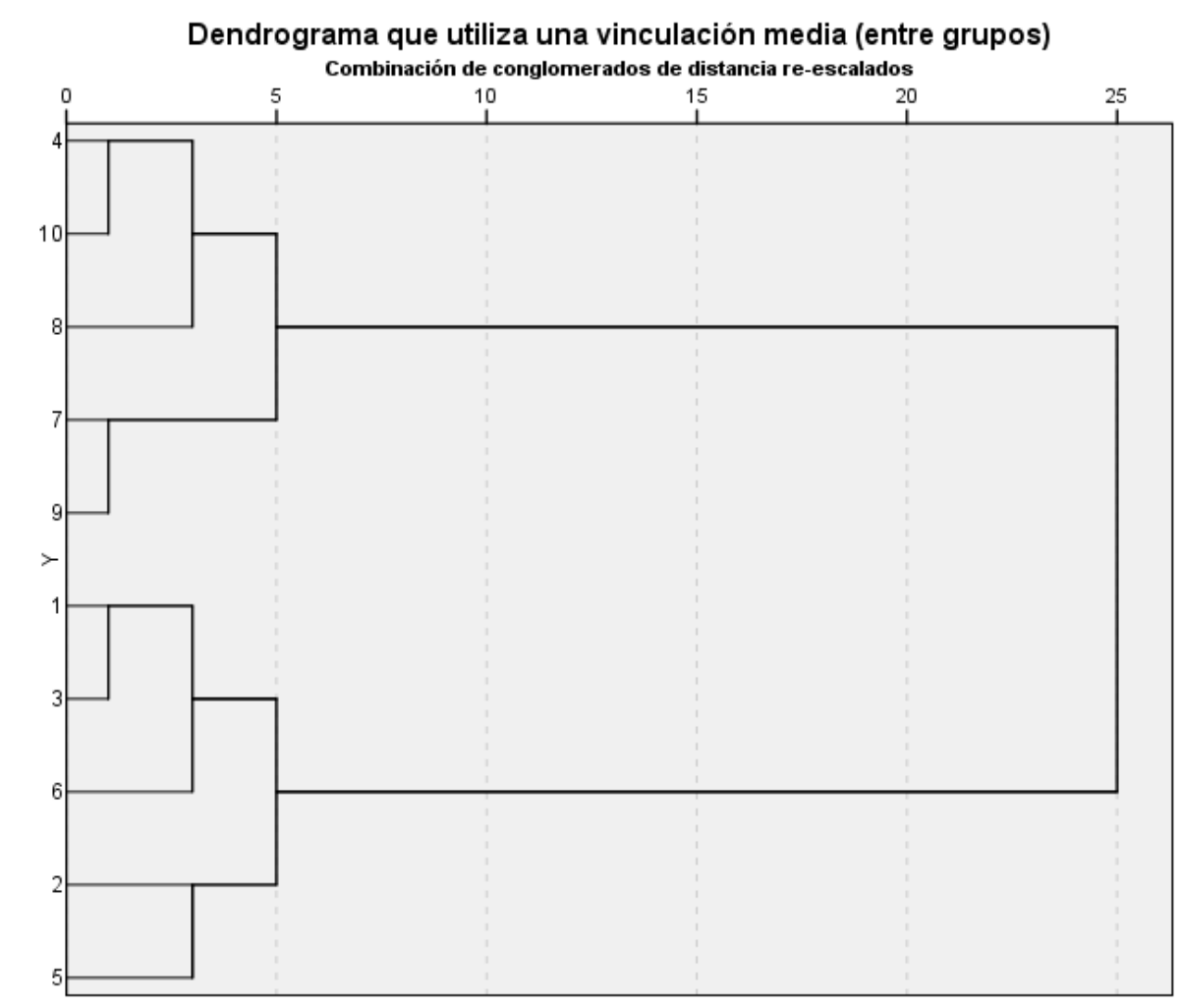
- Se aconseja utilizar distintos métodos para ver si la agrupación es natural. De lo contrario, en el método no jerárquico (*k*-medias) se prueba con distintos valores de k y en función de los resultados se toma la decisión.

SELECCIÓN DEL NÚMERO DE CONGLOMERADOS DE LA SOLUCIÓN

- Para decidir acerca del número grupos a considerar en nuestro análisis, podemos utilizar la representación gráfica que resume el historial de agrupación y que se conoce con el nombre de ***Dendrograma***.

- Para obtener el dendrograma en SPSS, en la ventana *Análisis de conglomerados jerárquico* debemos seleccionar la opción *Gráficos* y activar la casilla de verificación de Dendrograma y seleccionar la orientación del gráfico.

REPRESENTACIÓN DEL DENDROGRAMA



FORMACIÓN DE LOS GRUPOS: ANÁLISIS NO JERÁRQUICO DE CONGLOMERADOS

Método de K-medias

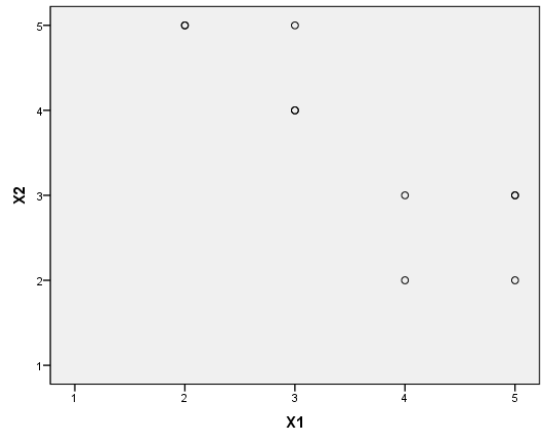
- En el análisis de conglomerados no jerárquico el investigador decide ***a priori*** **el número *k* de grupos** que desea, y las observaciones se asignan a cada uno de esos *k* conglomerados de forma que se maximiza la homogeneidad de los sujetos asignados a un mismo grupo y la heterogeneidad entre los distintos conglomerados.

MÉTODO K-MEDIAS

- Se basa en la asignación de los individuos a los conglomerados cuyo centroide se encuentre más próximo.
- Importante:
  - El centroide no tiene por qué coincidir con ningún individuo.
  - No se debe calcular este método cuando las variables no son métricas.
  - Siempre utiliza, para medir la distancia entre los casos, la distancia euclídea.

Ejemplo 1

- Con los datos del ejemplo 1, en el paquete estadístico SPSS, para aplicar el método k-medias, seleccionamos *Analizar/Clasificar/Conglomerado de k medias…*



Centros iniciales de los conglomerados

	Conglomerado	
	1	2
X1	5	2
X2	2	5

Centros de los conglomerados finales

	Conglomerado	
	1	2
X1	5	3
X2	3	5

Historial de iteraciones[a]

Iteración	Cambio en los …	
	1	2
1	,721	,721
2	,000	,000

a. Se ha logrado la convergencia debido a que los centros de los conglomerados no presentan ningún cambio o éste es pequeño. El cambio máximo de coordenadas absolutas para cualquier centro es de ,000. La iteración actual es 2. La distancia mínima entre los centros iniciales es de 4,243.

Número de casos en cada conglomerado

Conglomerado	1	5,000
	2	5,000
Válidos		10,000
Perdidos		,000

Individuo	X1	X2	QCL_1
1	5	3	1
2	4	2	1
3	5	3	1
4	3	4	2
5	5	2	1
6	4	3	1
7	2	5	2
8	3	5	2
9	2	5	2
10	3	4	2

ELECCIÓN ENTRE LOS DISTINTOS TIPOS DE ANÁLISIS DE CONGLOMERADOS

- Selección cuidadosa de las variables.
- Selección de una medida de similaridad o distancia adecuada:
  - Variables cualitativas
  - Variables cuantitativas: La más utilizada la distancia euclídea.
- Aplicación de un método jerárquico: Selección de un método de agrupamiento apropiado es muy importante.
  - Método de Ward (si no existen outliers).
  - Vinculación entre grupos.
- Solución final: Dendrograma.
- Aplicación de un método no jerárquico: K-medias.
- Caracterización de cada uno de los grupos formados.

ALGUNAS APLICACIONES EN MEDIA SYSTEMS

JOURNAL OF COMMUNICATION

Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

Hallin and Mancini Revisited: Four Empirical Types of Western Media Systems

Michael Brüggemann[1,*], Sven Engesser[1,*], Florin Büchel[1], Edda Humprecht[1], & Laia Castro[2]

1 IPMZ-Institute of Mass Communication and Media Research, University of Zurich, 8050 Zürich, Switzerland
2 Department of Communication and Media Research, University of Fribourg, 1700 Fribourg, Switzerland

The analysis of media systems has become a cornerstone in the field of comparative communication research. Ten years after its publication, we revisit the landmark study in the field, Hallin and Mancini's "Comparing Media Systems", and operationalize its framework for standardized measurement. The study at hand is, to the best of our knowledge, the first to comprehensively validate the original dimensions and models using aggregated data from the same sample of Western countries. Three out of four dimensions of media systems show relatively high levels of internal consistency, but "role of the state" should be disaggregated into 3 subdimensions. A cluster analysis reveals 4 empirical types of media systems that differentiate and extend the original typology.

Keywords: Media Systems, International Comparison, Journalism, Political Communication, Typology, Cluster Analysis.

doi:10.1111/jcom.12127

The study of media systems has become a cornerstone in the evolving field of comparative communication research: Typologies of media systems can serve as powerful heuristics that guide concept formation, hypotheses, and case selection. The unrivaled success of the two books, *Four Theories of the Press* (1956) and its successor *Comparing Media Systems* (2004), proves the power of media typologies. Hallin and Mancini's work has inspired and guided comparative research. It has received a lot of praise but also some criticism.

Ten years after its publication, it is time to revisit the theoretical framework and address a line of criticism that has been most prominently advanced by Norris (2011). She claims that Hallin and Mancini's three models cannot be replicated due to a lack of precise operationalization and standardized measurement. This study follows her call and contributes to the field by operationalizing, validating, and—in

Corresponding author: Michael Brüggemann; e-mail: michael@bruegge.net
*The first two authors have equally shared the lead on this paper.

International Journal of Communication 11(2017), 4797–4823 1932–8036/20170005

Rethinking Hallin and Mancini Beyond the West: An Analysis of Media Systems in Central and Eastern Europe

LAIA CASTRO HERRERO
University of Fribourg, Switzerland

EDDA HUMPRECHT
SVEN ENGESSER
University of Zurich, Switzerland

MICHAEL BRÜGGEMANN
University of Hamburg, Germany

FLORIN BÜCHEL
University of Zurich, Switzerland[1]

This study aimed to validate and extend Hallin and Mancini's framework of comparison to discriminate empirical types of media systems in Central and Eastern Europe. We tested and complemented their original dimensions by using aggregated data from 11 countries (Bulgaria, Croatia, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, and Slovenia). Our study shows the strength of political parallelism and public service broadcasting as variables for comparison. It also found that press freedom and foreign ownership point to significant differences between media systems in the region. Finally, a cluster analysis revealed the existence of three groups of media systems and provides empirical support for the assertion that there is no unique type of East-Central European media system.

Laia Castro Herrero: laia.castroherrero@unifr.ch
Edda Humprecht: e.humprecht@ipmz.uzh.ch
Sven Engesser: s.engesser@ipmz.uzh.ch
Michael Brüggemann: michael.brueggemann@uni-hamburg.de
Florin Büchel: f.buechel@ipmz.uzh.ch
Date submitted: 2016–07–07

[1] We thank Thomas Mathis for his valuable assistance. We are also grateful to Manuel Puppis, Frank Esser, and attendees at presentations at the International Communication Association and the European Communication Research and Education Association Political Communication conferences for their advice and suggestions. We also thank the editors and anonymous reviewers of the journal for their helpful comments.

Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

Media Systems in the Digital Age: An Empirical Comparison of 30 Countries

Edda Humprecht[1], Laia Castro Herrero[1,2], Sina Blassnig[1], Michael Brüggemann[3], & Sven Engesser[4]

[1]Department of Communication and Media Research, University of Zurich, Zurich, Switzerland
[2]Universitat Internacional de Catalunya, Spain
[3]Department of Journalism and Mass Communication, University of Hamburg, Hamburg, Germany
[4]Institute of Media and Communication, Technical University of Dresden, Dresden, Germany

Media systems have changed significantly as a result of the development of information technologies. However, typologies of media systems that incorporate aspects of digitalization are rare. This study fills this gap by identifying, operationalizing, and measuring indicators of media systems in the digital age. We build on previous work, extend it with new indicators that reflect changing conditions (such as online news use), and include media freedom indicators. We include 30 countries in our study and use cluster analysis to identify three clusters of media systems. Two of these clusters correspond to the media system models described by Hallin and Mancini, namely the democratic-corporatist and the polarized-pluralist model. However, the liberal model as described by Hallin and Mancini has vanished; instead, we find empirical evidence of a new cluster that we call "hybrid": it is positioned in between the poles of the media-supportive democratic-corporatist and the polarized-pluralist clusters.

Keywords: Media Systems, Typology, Cross-National Comparison, Information And Communication Technologies, Political Communication, Journalism

https://doi.org/10.1093/joc/jqab054

Research on media systems has played an important role in comparative communication research over the past two decades following the publication of Hallin and Mancini's (2004) groundbreaking book *Comparing Media Systems*. This work provided the scholarly community with a theoretical framework to compare Western media systems using a historical-institutional approach. The authors highlighted the differences and similarities between different political and media systems in North America and Western Europe. They further suggested four dimensions for comparing those media systems: the inclusiveness of the media market,

Corresponding author: Edda Humprecht; e-mail: edda.humprecht@uzh.ch

Referencias bibliográficas

- Hair, J.F.; Anderson, R.E.; Tatham, R.L. y Black, W.C. (1999): *Análisis Multivariante de Datos*. Madrid. Prentice Hall.
- Levy, J-P y Varela, J. (2003): Análisis multivariante para las Ciencias Sociales. Madrid. Prentice Hall.
- Uriel, E. y Aldás, J. (2005): Análisis Multivariante Aplicado. Madrid. Thomson.
- http://www.ucm.es/info/socivmyt/paginas/D_departamento/materiales/analisis_datosyMultivariable/21conglk_SPSS.pdf
- http://www.uoc.edu/in3/emath/docs/Cluster.pdf
- www.factum-marketing.com/download.php?file=Cluster.doc