REIRE
Revista d'Innovació i Recerca en Educació
www.ub.edu/ice/reire.htm

<Artículo>

El análisis de conglomerados bietápico o en dos fases con SPSS

María-José Rubio-Hurtado, Ruth Vilà-Baños

Enviado: 21/09/2016
Aceptado: 20/10/2016

//Resumen
El procedimiento de análisis de conglomerados en dos fases, también llamado *bietápico*, es una herramienta de exploración diseñada para descubrir las agrupaciones naturales de un conjunto de datos, permitiendo así la generación de criterios de información, frecuencias de los conglomerados y los estadísticos descriptivos por conglomerado, gráficos de barras, sectores y gráficos de importancia de las variables.
El método de análisis de conglomerados en dos fases tiene unas características únicas respecto a otros métodos de conglomeración tradicionales, que son las siguientes: un procedimiento automático del número óptimo de conglomerados, la posibilidad de crear modelos de conglomerados con variables tanto categóricas como continuas y la opción de trabajar con archivos de datos de gran tamaño.

//Palabras clave
Clasificación; Conglomerado; Clúster bietápico.

//Referencia recomendada
Rubio-Hurtado, M.-J., y Vilà-Baños, R. (2017). El análisis de conglomerados bietápico o en dos fases con SPSS. *REIRE. Revista d'Innovació i Recerca en Educació*, *10*(1), 118-126. doi: http://doi.org/10.1344/reire2017.10.11017

//Datos de las autoras
María-José Rubio-Hurtado. Universidad de Barcelona, mjrubio@ub.edu, orcid: http://orcid.org/0000-0003-2052-7611
Ruth Vilà-Baños. Universidad de Barcelona, ruth_vila@ub.edu, orcid: http://orcid.org/0000-0003-3768-1105

1. ¿Qué es un análisis de conglomerados o clúster?

El clúster es una técnica de clasificación que sirve para poder detectar y describir subgrupos de sujetos o variables homogéneas en función de los valores observados dentro de un conjunto aparentemente heterogéneo. Se basa en el estudio de las distancias entre ellos, lo cual permite cuantificar en el análisis el grado de similitud, en el caso de las proximidades, y el grado de diferencia, en el caso de las distancias. Como resultado aparecen agrupaciones o clústeres homogéneos (Vilà Baños, Rubio Hurtado, Berlanga Silvente y Torrado Fonseca, 2014).

2. ¿Qué es el análisis de conglomerados bietápico o en dos fases?

El análisis de conglomerados en dos fases, también llamado *bietápico*, es una herramienta de exploración diseñada para descubrir las agrupaciones naturales de un conjunto de datos (Pérez, 2011).

Con esta técnica se generan criterios de información, frecuencias de los conglomerados y los estadísticos descriptivos por conglomerado. También se pueden generar gráficos de barras de frecuencias de los conglomerados, gráficos de sectores de frecuencias y gráficos sobre la importancia de las variables. Finalmente, se pueden obtener medidas de la distancia para el cálculo de la similitud entre dos conglomerados.

3. ¿En qué se diferencia el análisis de conglomerados en dos fases de otros análisis clúster?

El método de análisis de conglomerados en dos fases tiene las siguientes características únicas respecto a otros métodos de clúster no jerárquico o jerárquico. Tal y como señala Pérez (2011), el algoritmo que emplea el análisis clúster en dos fases incluye varios rasgos que lo hacen diferente a las técnicas de conglomeración tradicionales:

- *Procedimiento automático del número óptimo de conglomerados.* Mediante la comparación de los valores de un criterio de selección del modelo para diferentes soluciones de conglomerados se determina automáticamente el número óptimo.
- *Posibilidad de crear modelos de conglomerados con variables categóricas y continuas.* Suponiendo una distribución normal multinomial conjunta y asumiendo que las variables son independientes.
- *Archivos de datos de gran tamaño.* Permite analizar grandes bases de datos mediante la construcción de un árbol de características de conglomerados que resume los registros.

4. Condiciones de aplicación del análisis de conglomerados en dos fases

Esta técnica se rige por las siguientes condiciones de aplicación:

A. *Las variables del modelo de conglomeración deben ser independientes.* La medida de la distancia de la verosimilitud implica que las variables del modelo de conglomerados sean independientes. Los procedimientos que pueden utilizarse para comprobar si se cumple este supuesto son los siguientes:

   - Correlaciones bivariadas para comprobar la independencia de dos variables continuas.
   - Tablas de contingencia para comprobar la independencia de dos variables categóricas.
   - El procedimiento de medias para comprobar la independencia entre una variable continua y una variable categórica.

B. *Las variables cuantitativas continuas siguen la ley normal.* El procedimiento de exploración para comprobar la normalidad de una variable continua puede ser la prueba de Kolmogorov-Smirnov.

C. *Las variables cualitativas categóricas tienen una distribución multinomial.* Se recomienda la prueba de chi-cuadrado para comprobar si una variable categórica sigue una distribución multinomial.

Las comprobaciones empíricas internas indican que este procedimiento es bastante robusto, incluso cuando no se cumplen estas condiciones. Aun así es preciso tener en cuenta hasta qué punto se cumplen estos supuestos.

Los resultados obtenidos pueden depender del orden de los casos. Para minimizar estos efectos se recomienda lo siguiente:

- Ordenar los casos aleatoriamente.
- Obtener varias soluciones distintas con los casos ordenados en distintos órdenes aleatorios para comprobar la estabilidad de una solución determinada.
- Cuando los tamaños de archivo son demasiado grandes, pueden sustituirse varias ejecuciones por una muestra de casos ordenados con distintos órdenes aleatorios.

5. Proceso para hacer un análisis de conglomerados en dos fases con SPSS

Para el análisis de conglomerados en dos fases con SPSS (versión 22) sugerimos realizar los siguientes pasos:

A. Verificar las condiciones de aplicación comentadas en el punto anterior.

B. Hacer una exploración inicial de los datos. Para una mejor solución se pueden explorar diversas selecciones de variables, aunque recomendamos tener presente el principio de parsimonia, que en este caso implica la selección de un número reducido de variables.

C. Seleccionar la medida de distancia más adecuada según las variables y otras opciones para el análisis. Se pueden obtener dos medidas de la distancia para el cálculo de la similaridad entre dos conglomerados:

- *Log-verosimilitud*. Realiza una distribución de probabilidad entre las variables: las continuas se suponen normales y las categóricas, multinomiales. Todas ellas se consideran independientes.

- *Euclídea.* Solo se puede utilizar cuando todas las variables son continuas.

D. Analizar los resultados obtenidos. En el análisis se sugiere seguir los siguientes pasos:

Si la solución es satisfactoria —ocurre cuando la medida de la silueta se sitúa en la franja "Buena" del gráfico "Calidad de conglomerados"—, se procede a confirmar dicha solución repitiendo el análisis, pero modificando previamente el orden en que aparecen los sujetos en la matriz (mediante la opción de aleatorización del orden de los sujetos que se encuentra en el menú "Datos/Ordenar casos").

Seguidamente se comprueba que la asignación de los sujetos a cada conglomerado es consistente mediante el cálculo del índice Kappa[1] como grado de acuerdo de dicha asignación[2] con la que se halle mediante otra técnica de conglomerados, como por ejemplo el K-medias —ver un ejemplo aplicado en Contreras Higuera, Martínez Olmo, Rubio Hurtado y Vilà Baños (2016).

Finalmente hay que caracterizar los conglomerados en función de las variables incluidas en el modelo. Para ello el SPSS ofrece la opción "Agrupaciones" en el menú desplegable (haciendo doble clic en el gráfico "Calidad de los conglomerados").

Si se ha marcado la opción "Crear variable de conglomerado de pertenencia", se pueden caracterizar también los grupos en función de cualquier otra variable no incluida en el modelo.

[1] En el SPSS el índice Kappa se encuentra en el menú «Análisis/Tablas de contingencia/Estadísticos».

[2] Hay que tener precaución a la hora de comparar las asignaciones hechas por diversos procesos de análisis de conglomerados, porque el SPSS nombra las etiquetas de los conglomerados arbitrariamente y quizás hay que recodificar dichas etiquetas.

6. Un caso práctico

Para ejemplificar lo expuesto, proponemos el siguiente caso práctico, en el que partimos de la siguiente pregunta de investigación:

¿Existen diferentes perfiles de profesores universitarios en relación con la docencia?

Para tal fin hemos recogido diferentes variables sociodemográficas y contextuales del profesorado universitario y las puntuaciones obtenidas por este en dos escalas relacionadas con el objeto de estudio: una escala de opinión sobre la capacidad de autonomía del alumnado y otra escala de opinión sobre la importancia del docente en el proceso de aprendizaje del alumnado.

PASO 1. Condiciones de aplicación

Para verificar las condiciones de aplicación realizamos los siguientes análisis:

- Correlaciones bivariadas, para comprobar la independencia de variables continuas.
- Tablas de contingencia, para comprobar la independencia de las variables categóricas.
- Comparación de medias, para comprobar la independencia existente entre la continua y la categórica.
- Exploración, para comprobar el ajuste a la ley normal.
- Prueba de chi-cuadrado, para las categóricas multinomiales.

PASO 2. Explorar los datos

Del total de las variables, y después de explorar la relevancia de las variables recogidas mediante pruebas de contraste, se consideran para el clúster bietápico las siguientes variables de entrada:

- Años de experiencia docente del profesorado (variable continua).
- Área de conocimiento del grado en el que imparte clases (variable nominal formada por cuatro categorías: ciencias, humanidades, técnicas, sociales).
- Puntuación en la escala de opinión sobre la capacidad de autonomía del alumnado (variable continua cuyo intervalo teórico oscila entre 1 y 20).
- Puntuación en la escala de opinión sobre la importancia del docente en el proceso de aprendizaje del alumnado (variable continua cuyo intervalo teórico oscila entre 1 y 20).
- Metodologías de enseñanza que utiliza habitualmente el docente (variable nominal formada por dos categorías: indagativas, expositivas).

- Tipología de recursos TIC que utiliza el docente (variable nominal formada por tres categorías: principalmente colaborativos, diversidad de recursos, solo programas de presentaciones).

PASO 3. Medida de distancia y otras opciones

La prueba clúster bietápico en el SPSS (versión 22) se encuentra en el menú "Analizar", submenú "Clasificar", opción "Conglomerado en dos fases".

Una vez seleccionado el proceso, en el cuadro de diálogo que aparece debemos colocar las variables continuas y las categóricas en sus respectivos campos, y seguidamente seleccionar la medida de distancia (ver Figura 1).

En nuestro caso, utilizaremos la medida de la distancia *log-verosimilitud*, dado que incorporamos variables continuas y categóricas.

También seleccionaremos *Determinar automáticamente* para obtener así el número óptimo de conglomerados. En nuestro caso no nos interesa especificar un número fijo de conglomerados. Dejamos, por defecto, el 15 como número máximo de conglomerados que tendrá en cuenta el procedimiento.

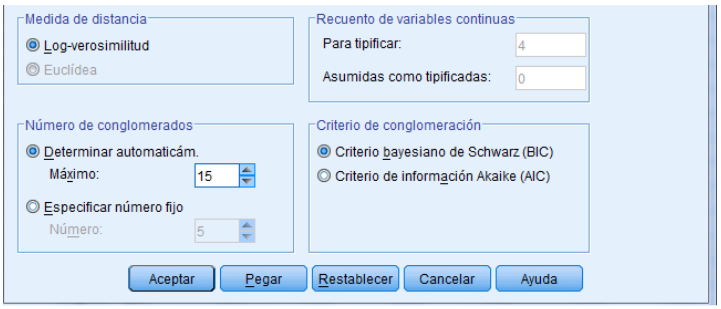
La sección de "Recuento de variables continuas" nos proporciona un resumen de las especificaciones acerca de la tipificación de variables continuas realizadas en las opciones. Por defecto, cuando se seleccionan variables continuas, estas se estandarizan automáticamente.

Finalmente, elegiremos el criterio de conglomeración que determina cómo el algoritmo halla el número de conglomerados: el criterio bayesiano de Schwarz (BIC) o el criterio de información Akaike (AIC). Ambos criterios valoran la calidad de los algoritmos teniendo en cuenta que una mayor cantidad de parámetros para mejorar el criterio de decisión sobre el número de conglomerados puede llevar a un sobreajuste. El criterio BIC penaliza el sobreajuste en mayor grado que el AIC, por lo que recomendamos dejar seleccionado el criterio BIC que el SPSS marca por defecto, aunque si la solución encontrada no es satisfactoria, se puede intentar una nueva solución con el criterio AIC.

En el botón "Resultados", marcaremos la opción "Crear variable del conglomerado de pertenencia". Eso nos permite poder caracterizar cada uno de los grupos en función de otras variables no consideradas en el modelo.

Figura 1

Cuadro de diálogo principal del clúster bietápico



Fuente: Elaboración propia

PASO 4. Resultados

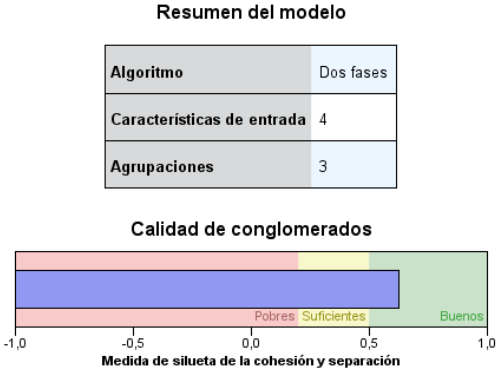
El programa muestra los resultados a través de las siguientes representaciones:

A. Tablas "Resumen del modelo" y "Calidad de los conglomerados" (Figura 2). Estas tablas nos informan de que del total de las variables introducidas se han considerado solo cuatro, que se ha generado un modelo formado por tres clústeres y que dicho modelo de conglomerado es bueno y, por lo tanto, podemos aceptarlo.[3]

B. Tabla "Agrupaciones". Se accede a ella clicando en la tabla "Calidad de los conglomerados". La tabla "Agrupaciones" muestra el tamaño de cada conglomerado, cada una de las variables consideradas en el modelo ordenadas de mayor a menor importancia y las respectivas puntuaciones de cada variable.

[3] En el *output* «Calidad de conglomerados», los resultados de «Pobres», «Suficientes» o «Buenos» se basan en el trabajo de Kaufman y Rousseeuw (1990) con respecto a la interpretación de los conglomerados. Un resultado en la zona «Buenos» significa que los datos evidencian de forma razonable o fuerte la estructura de los conglomerados; un resultado en la zona «Suficiente» significa que los datos evidencian de forma justa esta estructura de conglomerados, y un resultado en la zona «Pobres» refleja que los datos no aportan evidencias significativas de la estructura de conglomerados.
La medida de silueta se obtiene, sobre todos los casos, calculando (*B*-*A*) / max(*A*, *B*), donde *A* es la distancia del caso a su centro del conglomerado y *B* es la distancia del caso al centro del conglomerado más cercano al que no pertenece. Un coeficiente de silueta de 1 significaría que todos los casos se encuentran en sus centros de los conglomerados. Un valor de -1 significaría que todos los casos se encuentran en los centros de otros conglomerados a los que no pertenecen. Un valor de 0 significa que, en promedio, los casos son equidistantes entre el centro de su propio conglomerado y el centro de otro conglomerado cercano.

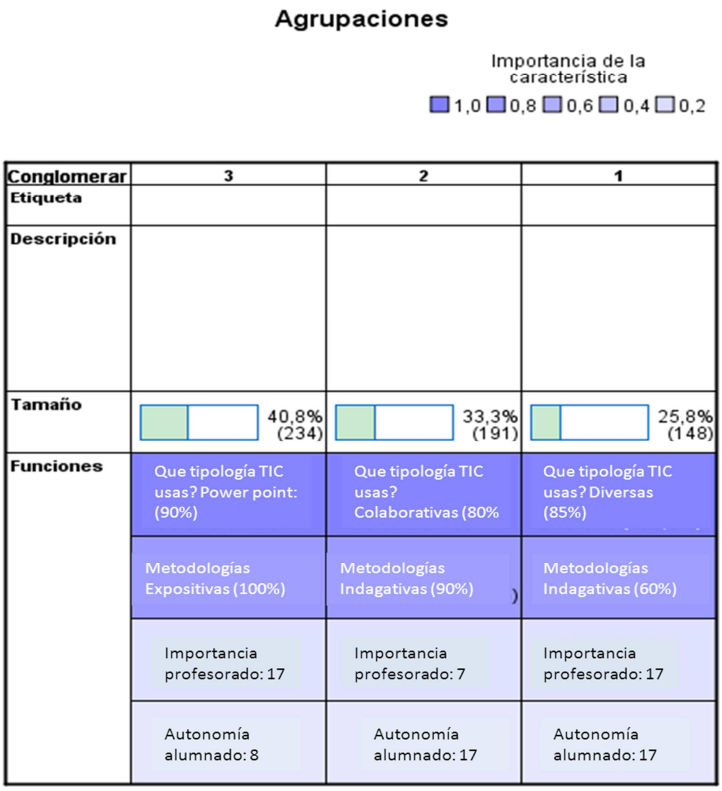
Figura 2
Output que muestra el resumen y la calidad del modelo de conglomerados



Fuente: Elaboración propia

En la tabla "Agrupaciones" (Figura 3) es donde observamos la caracterización de cada grupo, y donde podemos asignar un nombre a cada uno, en función de la representación teórica que suponga dicha caracterización. Este nombre se le puede añadir al campo "Etiqueta" clicando sobre él.

Figura 3
Output que muestra las agrupaciones propuestas por el modelo



Fuente: Elaboración propia

En nuestro caso han resultado tres agrupaciones, que incluyen cuatro variables ordenadas de mayor a menor peso en el modelo, siendo la de mayor peso la que aparece en primer lugar y de color más oscuro. Las características de estos tres grupos siguiendo la lectura de la tabla son las siguientes:

3. El grupo más numeroso, compuesto por el 40,8 % de la muestra, lo forma el profesorado con las siguientes características: usa sobre todo PowerPoint en su docencia (90 %), usa siempre metodologías expositivas (100 %), concibe un bajo nivel de autonomía en el alumnado (8), concibe una alta importancia del docente en el proceso de aprendizaje (17). A este grupo, por sus características, podríamos denominarlo *de enfoque tradicional*.

2. El segundo grupo está formado por el 33,3 % de la muestra y se caracteriza por los siguientes rasgos: usa principalmente TIC de tipo colaborativo (80 %), emplea sobre todo metodologías didácticas indagativas (90 %), concibe un alto nivel de autonomía en el alumnado (17) y poca importancia del docente en el proceso de aprendizaje del alumnado (7). A este grupo lo hemos denominado *de enfoque constructivista*.

1. El grupo menos numeroso, formado por el 25,8 % de la muestra, presenta las siguientes características: usa TIC diversas (85 %), usa metodologías indagativas, pero en menor grado que el grupo de enfoque constructivista (60 %), concibe un alto nivel de autonomía en el alumnado (17) y una alta importancia del docente en el proceso de aprendizaje (17). A este grupo lo hemos denominado *de enfoque innovador*.

En el campo "Etiqueta" de esta tabla podríamos escribir el nombre que le damos a cada agrupación y en el campo "Descripción", el resumen de sus características.

<Referencias bibliográficas>

Contreras Higuera, W. E., Martínez Olmo, F., Rubio Hurtado, M. J., y Vilà Baños, R. (2016). University students perceptions of e-portfolios and rubrics as combined assessment tools in education courses. *Journal of Educational Computing Research*, *54*(1), 85-107. doi: https://doi.org/10.1177/0735633115612784

Kaufman, L., y Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis*. New York: John Wiley & Sons. doi: https://doi.org/10.1002/9780470316801

Pérez, C. (2011). *Técnicas de segmentación. Conceptos, herramientas y aplicaciones.* Madrid: Gaceta Grupo Editorial.

Vilà Baños, R., Rubio Hurtado, M. J., Berlanga Silvente, V., y Torrado Fonseca, M. (2014). Cómo aplicar un clúster jerárquico en SPSS. *Revista d'Innovació i Recerca en Educació*, *7*(1), 113-127. doi: https://doi.org/10.1344/reire2014.7.1717

Copyright © 2017. Esta obra está sujeta a una licencia de Creative Commons mediante la cual, cualquier explotación de ésta, deberá reconocer a sus autores, citados en la referencia recomendada que aparece al inicio de este documento.