Documento metodológico del Estratificador INEGI 2.0

Octubre, 2024

Contenido

Introducción

Hoy en día es frecuente escuchar que la implementación de una política pública, que busca atender un problema de naturaleza multidimensional, se basa en algún indicador compuesto o en alguna aplicación de éste. Sobra decir que un índice desarrollado de forma inadecuada puede limitar el efecto esperado de la aplicación de recursos fiscales en la población a la que son dirigidos. Es por esta razón que se ha vuelto muy importante mejorar nuestra capacidad tanto para realizar los cálculos requeridos como para evaluar los resultados. En general, los métodos para obtener indicadores compuestos siguen siendo objeto de análisis y mejora; de cualquier modo, aún es frecuente encontrar indicadores de este tipo construidos mediante diversos procedimientos y métodos.

Por su parte, en relación con nuestra capacidad para evaluar los resultados, hay en general ausencia de respuestas a la pregunta: ¿se puede medir de mejor manera lo que se quiere medir? Por ejemplo, en presencia de alguno de los indicadores de mayor uso actual para la ordenación de países o regiones (p. ej., el índice de desarrollo humano de la Organización de las Naciones Unidas), resulta difícil determinar si alguna modificación al valor del coeficiente de uno de los indicadores componentes mejora o empeora la descripción que el índice compuesto resultante hace del fenómeno en cuestión. Es claro que se requiere desarrollar la capacidad para medir, además, de qué tan bien se mide lo que se desea medir.

La naturaleza multidimensional de los aspectos que se busca conocer trae consigo la dificultad adicional de comunicar los resultados alcanzados. Por esta razón se ha recurrido a procedimientos simples y fáciles de explicar. Llama por ello la atención que el esfuerzo pionero del CONAPO (iniciado durante la primera mitad de la década de los 90 y basado en metodologías no tradicionales) haya encontrado buena aceptación y amplia aplicación. En efecto, el uso del análisis de componentes principales (ACP) como parte de la metodología permitió incorporar al análisis la covariabilidad de los indicadores utilizados, como lo exige un tratamiento formal del análisis de fenómenos multidimensionales o multivariados.

En relación con el análisis de componentes principales, la Organización para la Cooperación y el Desarrollo Económicos (OCDE) publicó en el 2005 un manual[1] para apoyar la elaboración de indicadores compuestos en el que se refiere al uso de esta técnica. Entre sus fortalezas, destaca su capacidad de resumir un conjunto de indicadores básicos en tanto se preserva la proporción máxima posible de la variación total en el archivo de datos original. Indica que las mayores ponderaciones son asignadas a los indicadores básicos que muestran la mayor variación entre países y destaca que ésta es una propiedad deseable para realizar comparaciones entre naciones, ya que los indicadores básicos que son parecidos entre ellas carecen de interés pues no pueden

[1]Nardo, M.; Saisana, M.; Saltelli, A.; Tarantola, S.; Hoffman, A.; Giovannini, E. (2005). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OCDE Statistics Working Paper, recuperado en www.olis.oecd.org/olis/2005doc.nsf/LinkTo/NT00002E4E/$FILE/JT00188147.PDF

explicar las diferencias en desempeño. En contrapartida, los autores del manual señalan entre las debilidades del método que las correlaciones no representan necesariamente la influencia real de los indicadores básicos en el fenómeno que está siendo medido; del mismo modo, indican que es sensible a modificaciones en los datos, así como a la presencia de observaciones aberrantes que pueden introducir variabilidad espuria.

Debido a que a lo largo del texto se hará referencia a técnicas multivariadas de conglomeración (o estratificación), conviene recordar otros esfuerzos encaminados a establecer clasificaciones socioeconómicas de áreas geográficas en México, desde el nivel de área geoestadística básica (AGEB hasta el de entidad federativa. El Instituto Nacional de Estadística y Geografía (INEGI) desarrolló un ejercicio basado en métodos de estratificación que, en sus orígenes, fue denominado *niveles de bienestar*. Este enfoque reconoce la naturaleza multidimensional del bienestar al incorporar un número importante de indicadores, aunque el procedimiento estadístico de clasificación al que recurre (conocido como k-medias) basa la clasificación de una unidad en la distancia euclidiana entre esta y los centroides de los conglomerados[2], en otras palabras, no aprovecha la información relativa a las correlaciones entre los indicadores utilizados.

Al respecto, el ya citado manual de la OCDE señala que el análisis de conglomerados es otra herramienta para clasificar grandes cantidades de información en conjuntos más tratables, y ha sido, también, usado en el desarrollo de indicadores compuestos para agrupar información sobre países basada en su semejanza con base en diferentes indicadores básicos; además, sirve como a) un método meramente estadístico de agregación de los indicadores; b) una herramienta de diagnóstico para explorar el impacto del uso de diversas metodologías durante la fase de construcción del indicador compuesto; c) un método para la diseminación de información sobre el indicador compuesto, sin perder la que se refiere a las dimensiones de los indicadores básicos y d) un método para seleccionar grupos de países para imputar datos faltantes con el propósito de reducir la varianza de los valores imputados.

[2] Hartigan, J. A. (1975). Clustering algorithms. *John Wiley google schola*, *2*, 25-47.

1. Objetivo de la estratificación

La estratificación es un procedimiento estadístico relacionado con los denominados de clasificación o agrupación. En estos casos se busca que los grupos sean integrados por unidades similares en su interior, pero heterogéneos entre sí.

En el caso de la estratificación es deseable que además el resultado esté dado por clases ordenadas, de mayor a menor o de mejor a peor, cuando las variables en las que se basa la agrupación lo permitan.

En adelante se hará uso de la siguiente notación. Sea $\boldsymbol{X}_{n\times k}$ la matriz de datos. Esta queda expresada como sigue, cuando se dispone de k variables, que corresponden a las columnas, para cada una de las n unidades muestrales, a las cuales corresponden los renglones.

$$\boldsymbol{X}_{n\times k} = \begin{bmatrix} \boldsymbol{x}_{1\times k} \\ \boldsymbol{x}_{2\times k} \\ \boldsymbol{x}_{3\times k} \\ \vdots \\ \boldsymbol{x}_{n\times k} \end{bmatrix}_{n\times k}$$

donde $n = n_1 + n_2 + \cdots + n_G$ es el tamaño de la población y n_j el de las observaciones que corresponden al j-ésimo estrato, y $\boldsymbol{x}_{i\times k}, i = 1, \ldots, n$ es el i-ésimo vector de observaciones, de dimensión $1 \times k$.

Con el propósito de simplificar algunas de las expresiones a las que se recurrirá para evaluar la calidad del resultado de un ejercicio de estratificación, es conveniente reordenar los renglones de la matriz de datos de modo que los primeros n_1 de ellos se refieran a las unidades agrupadas en el primer estrato; los siguientes n_2 a las correspondientes al segundo estrato; y así sucesivamente, hasta llegar a las agrupadas en el estrato G. Bajo estas condiciones la matriz $\boldsymbol{X}_{n\times k}$ quedará expresada como sigue:

$$\boldsymbol{X}_{n\times k} = \begin{bmatrix} \boldsymbol{X}_{n_1\times k} \\ \boldsymbol{X}_{n_2\times k} \\ \boldsymbol{X}_{n_3\times k} \\ \vdots \\ \boldsymbol{X}_{n_G\times k} \end{bmatrix}_{n\times k}$$

2. Métodos de estratificación incluidos

2.1 k-medias

El primero de los tres procedimientos es uno de los mejor conocidos y de mayor uso en la práctica. Ello se debe a que el procedimiento puede ser expresado de forma sencilla: con base en los valores de los *k*

indicadores seleccionados, asigne cada unidad geográfica a aquel de los K grupos[3] cuyo punto central le sea más cercano. En otras palabras, a aquel grupo cuya distancia (Euclidiana) con el punto que representa a la unidad en un espacio de k dimensiones, sea mínima. Por supuesto, la distancia entre un grupo y un punto puede definirse de diversas formas; las más usuales son las identificadas como "vecino más próximo", "vecino más distante" y "al centroide". En particular, k-medias hace uso del último en esa lista. Cuando las distancias se minimizan, debe tenerse que además la suma (de cuadrados) de las distancias entre las unidades y sus centroides alcanza su valor mínimo. Es decir, la estratificación óptima es la que minimiza ese criterio[4] cuando los centroides han sido adecuadamente elegidos.

2.2 mulvar

Por su parte, el segundo de los tres métodos incluidos, el que ha sido denominado mulvar en el Estratificador, es el usado en los ejercicios denominados "Niveles de Bienestar" que fueron elaborados por el INEGI a partir de información recolectada por los censos de 1990 y de 2000. El procedimiento fue propuesto en Jarque (1981)[5] como un intento por extender la estratificación univariada óptima de Dalenius-Hodges a un contexto multivariado. La esencia del método puede resumirse como la aplicación del procedimiento de k-medias a una versión estandarizada de los indicadores seleccionados; la mencionada estandarización se realiza usando las desviaciones típicas de los estimadores muestrales del promedio poblacional de cada indicador. Ha sido habitual suponer un tamaño de muestra equivalente a 10 % del tamaño de la población y así ha sido instrumentado en el Estratificador INEGI.

2.3 mclust[6]

El tercero de procedimientos se incluye buscando corregir una limitación de los primeros dos, que se refiere a aquellas circunstancias en que es necesario tomar en cuenta a las correlaciones exhibidas entre indicadores para evitar que redundancias entre ellos sesguen los resultados en

---

[3] A lo largo de la discusión se ha hecho uso de la letra G para designar el número de grupos. La costumbre es usar K; de ahí el nombre del procedimiento. Se espera que este cambio momentáneo no cause confusión.

[4] Este procedimiento ilustra también la complejidad que enfrentan los procedimientos de estratificación multivariada. Antes de llevar a cabo cualquier estratificación, los valores de los centroides son desconocidos; en consecuencia, no es posible calcular a priori las distancias entre las unidades y ellos. Alternativamente, puede pensarse en proponer una asignación arbitraria de las unidades a G grupos, para después calcular los centroides correspondientes y finalmente las distancias entre éstos y los puntos del grupo. Dos asignaciones tales pueden ser comparadas con base en las sumas de distancias para determinar cuál es "mejor". Procediendo de este modo, después de hacer una enumeración completa, cabe esperar que sería sencillo identificar la óptima. Sin embargo, cuando el número de unidades es relativamente grande, resulta materialmente imposible enumerar todas las posibles estratificaciones para encontrar la que sería considerada óptima; por ejemplo, para el caso de los más de 2450 municipios mexicanos se tiene que el número de todas sus estratificaciones en cinco grupos rebasa el valor 5^{2450} o, lo que es casi lo mismo, a un 10 seguido de 1711 ceros. Si el tiempo que toma la asignación de unidades a grupos, más el que toma el cálculo de centroides, más el que toma calcular las distancias de éstos a las unidades, consumiera en total 1 segundo, todavía tomaría más de 10^{1700} segundos, o más de 3 millones de siglos, hacer una enumeración completa para tener la certeza de que se encontró la solución óptima. Por lo anterior se han desarrollado estrategias que permiten encontrar soluciones aproximadas en tiempos razonables. Una de ellas, instrumentada en el Estratificador, consiste en seleccionar aleatoriamente G unidades para que hagan las veces de centroides iniciales. Cada vez que una unidad es asignada a un grupo, el centroide correspondiente es recalculado hasta que todas las unidades han sido asignadas. El proceso se repite usando ahora como centroides iniciales los que resultan de la iteración anterior hasta que ninguna unidad cambia de estrato. Es claro que aun cuando se use el mismo conjunto de indicadores y el mismo número de grupos, así como el mismo procedimiento, los resultados pueden variar dependiendo de las selecciones iniciales en cada aplicación; sin embargo, se espera que no difieran mucho.

[5] Jarque, C. M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *30*(2), 163-169.

[6] Ver A quick tour of mclust (r-project.org) o Mclust function - RDocumentation

alguna dirección. Para fijar ideas, se recurrirá a la situación descrita enseguida, en la cual se supone que el vector aleatorio $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{k}}$, del cual los n datos de la muestra son realizaciones independientes, proviene de una mezcla de un número desconocido G de distribuciones normales; tantas como grupos a formar. Es decir:

$$\boldsymbol{X} \sim \sum_{i=1}^{G} \frac{n_i}{n} N(\boldsymbol{\mu_i}, \Sigma_i)$$

Es decir, que la fracción de la muestra que pertenece al estrato $i = 1, \dots, G$, a su vez una muestra de tamaño desconocido n_i, pero tal que $n = \sum_{i=1}^{G} n_i$, proviene de una distribución normal multivariada con vector de medias (centroide) $\boldsymbol{\mu_i}$ y matriz de covarianzas Σ_i, con $i = 1, \dots, G$ (con fines de ejemplificación, en el cuadro siguiente el valor de G se supone igual a 4).

TABLA 1. ILUSTRACIÓN DE PROBLEMA A RESOLVER

Estrato 1: $n_1, \frac{\lvert\Sigma_1\rvert^{-1}}{2\pi^{k/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})^t \Sigma_1^{-1} (\boldsymbol{x}-\boldsymbol{\mu_1})}$	Estrato 2: $n_2, \frac{\lvert\Sigma_2\rvert^{-1}}{2\pi^{k/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_2})^t \Sigma_2^{-1} (\boldsymbol{x}-\boldsymbol{\mu_2})}$
Estrato 3: $n_3, \frac{\lvert\Sigma_3\rvert^{-1}}{2\pi^{k/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_3})^t \Sigma_3^{-1} (\boldsymbol{x}-\boldsymbol{\mu_3})}$	Estrato 4: $n_4, \frac{\lvert\Sigma_4\rvert^{-1}}{2\pi^{k/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_4})^t \Sigma_4^{-1} (\boldsymbol{x}-\boldsymbol{\mu_4})}$

Elaboración propia

El problema de clasificación en este contexto se refiere a determinar cuántas y cuáles unidades observadas pertenecen a cada uno de un número G desconocido de grupos. Si se conociera la respuesta de antemano, el método de máxima verosimilitud ayudaría para encontrar estimadores de los anteriores parámetros. En ausencia de la información anterior se hace necesario desarrollar estrategias eficientes de cálculo con el fin de dar respuesta satisfactoria al problema planteado. Con este fin, la función mclust del paquete R desarrolla una variedad de modelos para el comportamiento de las estructuras de los segundos momentos. Por ejemplo, es posible asumir de entrada que todas las matrices son iguales y que lo que distingue a los grupos es su centroide. Entonces, se impondría la restricción $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4$. También podría asumirse que todas las matrices son diagonales. Las anteriores restricciones podrían imponerse una, o la otra, o ambas. En el otro extremo estaría el no suponer nada y dejar que los datos hablen sugiriendo el tipo de modelo a usar en el ejercicio de estratificación.

Los mencionados modelos son caracterizados por un vector de 3 posiciones cada una de las cuales representa una característica del conjunto de matrices: volumen, forma y orientación. Por ello, para la aplicación de este método se incluye un procedimiento, que se activa al elegir mclust por primera vez, momento en el que aparece la leyenda "GENERANDO GRÁFICA BIC" en la esquina inferior derecha de la pantalla, y que evalúa (ver una descripción detallada del método, más abajo) diversas combinaciones de número de estratos a usar, entre 1 y 10, y de modelos locales para la estructura de covarianza, dentro de cada estrato, denotados por hasta 10 combinaciones de 3 letras[7]. La primera posición de la combinación se refiere al tamaño de las nubes de puntos (o al volumen de los elipsoides de concentración; E cuando son iguales o V cuando pueden variar); la segunda a su forma (I para todas esferas, E para todas elipsoides

[7] Esta parte del procedimiento puede consumir varios minutos dependiendo del número de variables y de unidades consideradas. Por ejemplo, para la ejemplificación numérica (más de 9300 manzanas y 19 variables) fueron requeridos casi dos minutos. Un botón animado en la esquina inferior derecha de la pantalla indicará que la elaboración de la gráfica está en proceso. El servicio puede seguir siendo usado para realizar otras funciones mientras esto ocurre.

y V para mezclas); y la tercera a la orientación de los ejes principales (I para esferas o sin orientación definida, E para igual orientación de todos los elipsoides y V para orientaciones diversas). De este modo, el modelo EII resulta ser el más restrictivo y el VVV, el más libre y, por ello, aquel al que se asocia una mayor cantidad de parámetros a estimar.

3. Transformaciones

Las transformaciones a los datos ayudan a prepararlos para que los modelos usados en esta aplicación puedan ser más eficaces y den resultados más congruentes. En el caso de la estandarización, se sugiere usarla cuando las variables tienen diferentes escalas. Esto evitará que las variables con mayor rango tengan mayor influencia sobre los resultados y se eviten así sesgos en los mismos.

Para el caso de Componentes Principales se sugiere usar cuando se sospeche que las variables elegidas están correlacionadas. La persona usuaria confirmará o descartará su sospecha cuando analice la gráfica de componentes principales y note la alta variabilidad recogida en las primeras dos componentes principales. Al usar esta herramienta estadística se reducirá dimensionalidad y evitará redundancia, ruido que se presentaría si se usaran todas las variables correlacionadas.

3.1 Estandarización

Con el fin de evitar que la separación entre los grupos se deba solamente a las unidades en las cuales son medidas algunas de las variables, es conveniente expresarlas en dimensiones similares. Por ejemplo, se puede eliminar la influencia de las unidades en el resultado mediante la estandarización de cada una de las variables. Esta es una operación simple cuando se conocen los valores de la media, μ_i, y de la desviación estándar, σ_i de cada variable $x_i, i = 1, \dots, k$. Su expresión algebraica queda dada por:

$$y_i = \frac{(x_i - \mu_i)}{\sigma_i}, i = 1, \dots, k.$$

Todas las nuevas variables tienen ahora media 0 y varianza igual a 1. Es de esperarse que ninguna de ellas domine a las restantes en la definición de los estratos, preservando de esta manera la intención de realizar un ejercicio multivariado. Cuando los valores de los parámetros anteriores sean desconocidos, se les remplazará por sus contrapartes muestrales.

3.2 Componentes principales

El propósito de esta técnica es el de construir variables latentes, que son sumas ponderadas de las variables originales, pero con características deseables. En este caso se pide que la denominada primera componente principal sea aquella combinación lineal, o suma ponderada, cuya varianza es máxima entre todas las combinaciones lineales posibles. A su vez, la segunda componente principal es una nueva combinación lineal cuya varianza es máxima entre todas las combinaciones lineales cuya correlación con la primera componente es igual a cero. Procediendo del mismo modo es posible obtener tantas componentes principales como variables iniciales,

siempre maximizando la varianza, pero imponiendo la condición de que su correlación sea nula con las anteriores componentes.

Sean $\boldsymbol{X}$, el vector aleatorio correspondiente a las mediciones, y Σ su matriz de covarianzas. El primer problema de maximización es planteado como sigue:

$$\underset{\underline{\alpha}_1}{Max} Var(\underline{\alpha}_1^t \boldsymbol{X}) = \underset{\underline{\alpha}_1}{Max}\ \underline{\alpha}_1^t \Sigma \underline{\alpha}_1$$
$$s.a.\ \underline{\alpha}_1^t \underline{\alpha}_1 = 1.$$

Y los subsecuentes de esta manera:

$$\underset{\underline{\alpha}_j}{Max} Var(\underline{\alpha}_j^t \boldsymbol{X}) = \underset{\underline{\alpha}_j}{Max}\ \underline{\alpha}_j^t \Sigma \underline{\alpha}_j$$
$$s.a.\ \underline{\alpha}_j^t \underline{\alpha}_j = 1, y$$
$$Cov(\underline{\alpha}_j^t \boldsymbol{X}, \underline{\alpha}_i^t \boldsymbol{X}) = \underline{\alpha}_j^t \Sigma \underline{\alpha}_i = 0, \forall\ i < j = 2, \dots, k.$$

Al proceder de esta forma las varianzas de las últimas componentes pueden ser relativamente insignificantes, llegando hasta a anularse; por ejemplo, si existe alguna redundancia entre las variables originales. Cuando este es el caso, es usual ignorar a aquellas componentes en todo análisis posterior. Se dice entonces que la dimensión del problema ha sido reducida, pues se ha obtenido un conjunto con un número menor de indicadores, los cuales, por construcción, no están correlacionados entre sí.

Una interpretación geométrica de las componentes principales es que se asocian con los ejes principales de las nubes de puntos de los datos. El eje más largo correspondería a la primera componente, la de mayor varianza, y así sucesivamente. Nuevamente se puede dar el caso de que una mayor dispersión se deba solamente a las unidades diferentes en que se midieron algunos indicadores originales. En estos casos se ha sugerido trabajar con las versiones estandarizadas de los mismos. Cabe señalar que esto equivale a convertir la nube de puntos en una hiperesfera con lo que se puede ocultar la dirección de mayor dispersión, como era el objetivo. Se recomienda evitar esta forma de proceder. La opción se ha incluido en el Estratificador para permitir la producción de resultados en Grados de Marginación como lo hacía CONAPO tiempo atrás.

4. Criterios para evaluar estratificaciones

Se tiene entonces que la matriz de datos una vez realizada una estratificación está dada por

$$\boldsymbol{X}_{n\times k} = \begin{bmatrix} \boldsymbol{x}_{1\times k} \\ \boldsymbol{x}_{2\times k} \\ \boldsymbol{x}_{3\times k} \\ \vdots \\ \boldsymbol{x}_{n\times k} \end{bmatrix}_{n\times k} = \begin{bmatrix} \boldsymbol{X}_{n_1\times k} \\ \boldsymbol{X}_{n_2\times k} \\ \boldsymbol{X}_{n_3\times k} \\ \vdots \\ \boldsymbol{X}_{n_G\times k} \end{bmatrix}_{n\times k} ;$$

Es necesario recordar que $n = n_1 + n_2 + \cdots + n_G$ es el tamaño de la muestra y n_j el de las observaciones que corresponden al j-ésimo de los G estratos.

Sean las matrices J_n, de dimensión $n \times n$, cuyas entradas son todas iguales a $\frac{1}{n}$, por lo que cuando premultiplica a un vector da como resultado un nuevo vector cuyas componentes son todas iguales al promedio de los valores del vector inicial; J_{n_j}, de dimensión $n_j \times n_j$, cuyas entradas son todas iguales a $\frac{1}{n_j}$, su efecto es similar al de J_n al multiplicar vectores de menor dimensión; y H, de dimensión $n \times n$, que es diagonal en bloques con las G matrices J_{n_j} en la diagonal. Es decir, sea $\underline{1}$ un vector de dimensión n todas cuyas componentes son iguales a 1, entonces

$$J_n = \underline{1}\left(\underline{1}^t\underline{1}\right)^{-1}\underline{1}^t = \frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n\times n}; J_{n_j} = \frac{1}{n_j}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n_j\times n_j};$$

$$H = \begin{bmatrix} J_{n_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & J_{n_G} \end{bmatrix}_{n\times n};$$

Note que, si I_n es una matriz identidad de orden n, se cumplen las siguientes igualdades

$$J_{n_j}J_{n_j} = J_{n_j} \Rightarrow HH = H \Rightarrow (I_n - H)(I_n - H) = (I_n - H);$$

Es decir, que las matrices J_{n_j}, H y $(I_n - H)$ son todas idempotentes. Asimismo, que

$$HJ_n = J_n$$

Por lo que

$$(I_n - H)(H - J_n) = H - H - J_n + HJ_n = 0.$$

Note adicionalmente que

$$Rango(I_n - H) = n - G; Rango(H - J_n) = G - 1.$$

Bajo las anteriores condiciones se tiene que la matriz $(I_n - H)\boldsymbol{X}$, de dimensión n$\times k$, tiene como sus entradas a las desviaciones de los valores de cada variable observada con respecto a su media muestral en el estrato correspondiente. Es decir,

$$(I_n - H)\boldsymbol{X} = \begin{bmatrix} (I_{n_1} - J_{n_1})\boldsymbol{X}_{n_1 \times k} \\ (I_{n_2} - J_{n_2})\boldsymbol{X}_{n_2 \times k} \\ (I_{n_3} - J_{n_3})\boldsymbol{X}_{n_3 \times k} \\ \vdots \\ (I_{n_G} - J_{n_G})\boldsymbol{X}_{n_G \times k} \end{bmatrix}_{n \times k} \quad ;$$

De este modo, la matriz W, de dimensión $k \times k$, igual a la suma de matrices de sumas de cuadrados y de productos cruzados dentro de los estratos puede ser expresada de la siguiente manera:

$$W = \boldsymbol{X}^t(I_n - H)\boldsymbol{X} = \sum_{i=1}^{G} \boldsymbol{X}_{n_1 \times k}^t (I_{n_1} - J_{n_1})\boldsymbol{X}_{n_1 \times k} = \sum_{i=1}^{G} W_i.$$

Por su parte, se define $\Sigma_i = 1/(n_i - 1)\, W_i$. Del mismo modo, se define a la matriz de sumas de cuadrados y de productos cruzados como:

$$(n-1)\Sigma = \boldsymbol{X}^t(I_n - J_n)\boldsymbol{X}.$$

A partir de lo anterior se hace posible interpretar y calcular los criterios de comparación entre estratificaciones alternativas en la siguiente tabla.

TABLA 2. CRITERIOS

Determinante de la suma de las matrices locales de sumas de cuadrados y de productos cruzados (Ward)	$\lvert W \rvert$
Suma de cuadrados de las distancias Euclidianas de cada punto al centroide de su grupo (SC)	$tr(W) = \sum_{i=1}^{G} tr(W_i)$
Suma ponderada del logaritmo de trazas de matrices locales de covarianzas (SLT)	$\sum_{i=1}^{G} n_i log\big(tr(\Sigma_i)\big)$
Suma ponderada del logaritmo de determinantes de matrices locales de covarianzas (SLD)	$\sum_{i=1}^{G} n_i log(\lvert\Sigma_i\rvert)$
Promedio ponderado de las relaciones entre determinantes locales y globales (DEff)	$\sum_{i=1}^{G} \frac{n_i}{n} \frac{\lvert\Sigma_i\rvert}{\lvert\Sigma\rvert}$

Elaboración propia

Además, la matriz $\boldsymbol{\Sigma}$ de sumas de cuadrados y productos cruzados de desviaciones de los puntos muestrales con la media (centroide) global puede ser descompuesto según medidas de homogeneidad tanto al interior de los estratos, $\boldsymbol{W}$, como entre estratos, $\boldsymbol{B}$.

$$\begin{aligned} \boldsymbol{\Sigma} = \boldsymbol{X}^t(I_n - J_n)(I_n - J_n)\boldsymbol{X} &= \boldsymbol{X}^t(I_n - H + H - J_n)(I_n - H + H - J_n)\boldsymbol{X} \\ &= \boldsymbol{X}^t[(I_n - H) - (I_n - H)J_n - J_n(I_n - H) + (H - J_n)(H - J_n)]\boldsymbol{X} \\ &= \boldsymbol{X}^t[(I_n - H) + (H - J_n)]\boldsymbol{X} = \boldsymbol{X}^t[I_n - H]\boldsymbol{X} + \boldsymbol{X}^t[H - J_n]\boldsymbol{X} = \boldsymbol{W} + \boldsymbol{B}. \end{aligned}$$

Como ya quedó demostrado, la matriz $\boldsymbol{X}^t[I_n - H]\boldsymbol{X}$ representa la suma de las G matrices de sumas de cuadrados y productos cruzados de desviaciones dentro de los estratos. Por su parte, ya que

$$[H - J_n]\boldsymbol{X} = \begin{bmatrix} 1_{n_1 \times 1} \otimes \left(\overline{\boldsymbol{x}}_{1\times k}^{(1)} - \overline{\boldsymbol{x}}_{1\times k}\right) \\ 1_{n_2 \times 1} \otimes \left(\overline{\boldsymbol{x}}_{1\times k}^{(2)} - \overline{\boldsymbol{x}}_{1\times k}\right) \\ 1_{n_3 \times 1} \otimes \left(\overline{\boldsymbol{x}}_{1\times k}^{(3)} - \overline{\boldsymbol{x}}_{1\times k}\right) \\ \vdots \\ 1_{n_G \times 1} \otimes \left(\overline{\boldsymbol{x}}_{1\times k}^{(G)} - \overline{\boldsymbol{x}}_{1\times k}\right) \end{bmatrix},$$

donde $\overline{\boldsymbol{x}}_{1\times k}^{(i)}$ y $\overline{\boldsymbol{x}}_{1\times k}$ son los vectores renglón cuyas componentes son los valores promedios de las variables para las unidades en el i-ésimo estrato y en la muestra completa, respectivamente. Entonces, la matriz $\boldsymbol{X}^t[H - J_n]\boldsymbol{X}$ representa las sumas, ponderadas por el tamaño del estrato, de cuadrados y productos cruzados de las desviaciones de los centroides de cada estrato con el centroide global. En otras palabras,

$$\boldsymbol{B} = \boldsymbol{X}^t[H - J_n]\boldsymbol{X} = \sum_{j=1}^{G} n_j \left(\overline{\boldsymbol{x}}_{1\times k}^{(j)} - \overline{\boldsymbol{x}}_{1\times k}\right)^t \left(\overline{\boldsymbol{x}}_{1\times k}^{(j)} - \overline{\boldsymbol{x}}_{1\times k}\right)$$

Es decir, se ha descompuesto a la suma de cuadrados total como la suma de las influencias al interior de los estratos (Within o dentro), que se desea minimizar, y la de los estratos entre si (Between o entre), que se busca maximizar.

Consecuentemente, se tiene que:

$$tr(\boldsymbol{\Sigma}) = tr(\boldsymbol{X}^t(I_n - J_n)\boldsymbol{X}) = tr(\boldsymbol{X}^t[I_n - H]\boldsymbol{X}) + tr(\boldsymbol{X}^t[H - J_n]\boldsymbol{X}) = tr\boldsymbol{W} + tr\boldsymbol{B}$$

Por lo que se considerará la inclusión del siguiente criterio para comparar entre estratificaciones alternativas:

$$C_6 = \frac{tr(\boldsymbol{X}^t[I_n - H]\boldsymbol{X})}{tr(\boldsymbol{X}^t[H - J_n]\boldsymbol{X})} = \frac{tr(\boldsymbol{W})}{tr(\boldsymbol{B})}$$

En el numerador se tiene una medida de lo que se quiere minimizar y en el denominador de lo que se quiere maximizar. Valores menores del cociente indican un mejor cumplimiento del objetivo general del ejercicio de agrupación, según se estableció en la sección correspondiente.