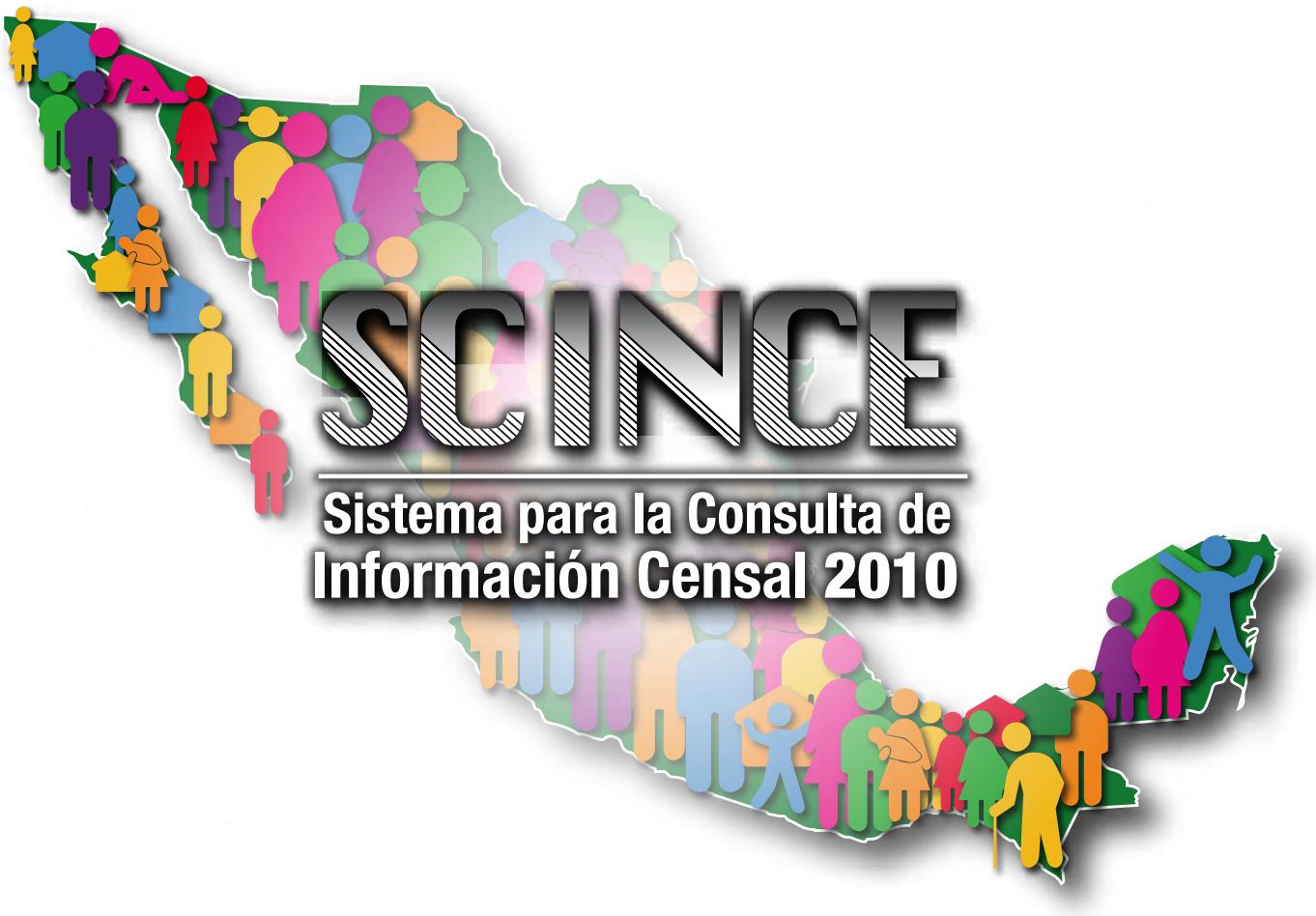
Instituto Nacional de Estadística y Geografía

Nota técnica

Estratificación multivariada



Censo de Población y Vivienda 2010

Censo de Población y Vivienda 2010

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

NOTA TÉCNICA

ESTRATIFICACIÓN MULTIVARIADA

Con la finalidad de que el usuario pueda realizar clasificaciones de las unidades geográficas del país considerando múltiples variables a la vez, se ha incorporado al Sistema para la Consulta de la Información Censal 2010 (SCINCE 2010) una herramienta de estratificación multivariada. Es importante que el usuario analice los resultados de la estratificación cuidadosamente antes de utilizar la clasificación obtenida.

El objetivo de la estratificación multivariada es resumir la información de todas las variables que se incluyen en el análisis, en una medida unidimensional que permita clasificar las observaciones en grupos homogéneos internamente y disímiles entre sí. El presente documento describe brevemente las técnicas empleadas para la estratificación; adicionalmente, se proporciona bibliografía para aquellos usuarios interesados en un estudio detallado de estas técnicas.

1. Método de Componentes principales y Dalenius-Hodges

Esta técnica de estratificación multivariada consiste en obtener una medida unidimensional en la que se resume la información de las variables consideradas para la estratificación, llamada primera componente principal, y aplicar a ésta el método de estratificación univariada de Dalenius-Hodges.

1.1 Componentes principales

Para realizar un análisis exploratorio de datos multivariados, se recomienda el uso de la técnica de componentes principales como primer paso. Esta técnica permite observar las estructuras de variación de los datos y, en algunos casos, identificar observaciones atípicas o variables cuya aportación es mínima o redundante para realizar la clasificación.

El método de componentes principales consiste básicamente en resumir la información de un conjunto de variables mediante la construcción de un conjunto con menor número de variables. El método de construcción de las componentes principales garantiza que la primera componente principal sea la que explique un mayor porcentaje de varianza de los datos, por ello, es esta primera componente principal la que se utiliza para realizar la estratificación. Es importante que el usuario evalúe la pertinencia de aplicar este método de estratificación considerando que el porcentaje de varianza explicada por la primera componente principal debe ser lo más cercano posible a 100 por ciento.

Los resultados que se proporcionan por medio del análisis de componentes principales permiten explorar la estructura y comportamiento de los datos que se incluyan en el modelo de estratificación. Al analizar los resultados numéricos y las gráficas que se presentan, el usuario podrá determinar si las variables que se incluyen en el estudio son pertinentes, o bien algunas son redundantes o aportan poca información para la estratificación.

De manera más formal, el método de componentes principales consiste en la descripción de la variación de un conjunto de p variables en términos de un conjunto de m ($m \leq p$) variables no correlacionadas, que en realidad son combinaciones lineales de las variables originales.

Así, si $\boldsymbol{x_1, x_2, \dots, x_p}$ son las variables originales, entonces las componentes principales tendrán la forma:

$$\begin{aligned} \boldsymbol{y_1} &= a_{11}\boldsymbol{x_1} + a_{12}\boldsymbol{x_2} + \cdots + a_{1p}\boldsymbol{x_p} \\ &\vdots \\ \boldsymbol{y_m} &= a_{m1}\boldsymbol{x_1} + a_{m2}\boldsymbol{x_2} + \cdots + a_{mp}\boldsymbol{x_p} \end{aligned}$$

Las componentes principales están construidas de tal modo que la varianza captada va decreciendo, es decir $Var(\boldsymbol{y_1}) \geq Var(\boldsymbol{y_2}) \geq \cdots \geq Var(\boldsymbol{y_m})$, por ello, en un contexto de reducción de dimensiones se seleccionan las m primeras componentes principales para representar a la población original.

Las componentes principales se obtienen mediante una técnica algebraica llamada descomposición espectral que se aplica a la matriz de covarianzas o correlación, según sea el caso. A los resultados de la descomposición espectral se les conoce como eigenvalores y eigenvectores. Los detalles sobre el cálculo de las componentes principales pueden consultarse en las referencias bibliográficas que se proporcionan.

1.2 Dalenius-Hodges

El método de Dalenius-Hodges (1959) consiste en la formación de estratos de manera que la varianza obtenida sea mínima para cada estrato. El procedimiento para la conformación de los estratos es el siguiente:

Sea n=número de observaciones y L=número de estratos.

1.- Ordenar las observaciones de manera ascendente.

2.- Agrupar las observaciones en J clases, donde J=min(L*10, n).

3.- Calcular los límites para cada clase de la siguiente manera:

$$\lim\inf C_k = \min\{x_{(i)}\} + (k-1) * \frac{\max\{x_{(i)}\} - \min\{x_{(i)}\}}{\mathrm{J}}$$

$$\lim\sup C_k = \min\{x_{(i)}\} + (k) * \frac{\max\{x_{(i)}\} - \min\{x_{(i)}\}}{\mathrm{J}}$$

Los intervalos se tomarán abiertos por la izquierda y cerrados por la derecha, a excepción del primero que será cerrado por ambos lados.

4.- A partir de estos límites, obtener la frecuencia de casos en cada clase f_i $(i = 1, \dots, J)$.

5.- Obtener la raíz cuadrada de la frecuencia de cada clase.

6.- Acumular la suma de la raíz cuadrada de las frecuencias.

$$C_i = \sum_{h=1}^{i} \sqrt{f_h} \quad (i = 1, \dots, J)$$

7.- Dividir el último valor acumulado entre el número de estratos.

$$Q = \frac{1}{L} C_J$$

8.- Los puntos de corte de cada estrato se tomarán sobre el acumulado de la raíz cuadrada de las frecuencias en cada clase de acuerdo a lo siguiente: $Q, 2Q, \dots, (h-1)Q$. Si el valor de Q queda entre dos clases, se tomará como punto de corte aquella clase que presente la mínima distancia a *Q*. Los límites de los *h* estratos conformados serán aquellos correspondientes a los límites inferior y superior de las clases comprendidas en cada estrato.

Resultados de la estratificación

A continuación se da una breve descripción de las salidas correspondientes a la estratificación con el método de Componentes principales y Dalenius–Hodges. Se recomienda al usuario consultar el estudio de caso contenido en esta aplicación así como la bibliografía proporcionada.

1.2.1 Resumen de resultados

En este apartado se muestran los resultados más importantes del análisis de componentes principales.

- Porcentaje de la varianza explicada por la primera componente principal

  Este es uno de los parámetros más importantes a considerar para una elección adecuada del modelo, ya que la estratificación se realiza considerando únicamente la primera componente principal. Es deseable que este porcentaje sea lo más cercano posible a 100% para que la estratificación arroje buenos resultados.

- Modelo

  Presenta un resumen del modelo planteado por el usuario, el cual consiste en las variables incluidas y sus descriptores, el número de observaciones, el tipo de análisis a realizar (covarianza o correlación) y la desviación estándar que se obtiene para cada componente principal.

Si el usuario elige realizar el análisis utilizando la matriz de correlaciones, las variables se estandarizan y la varianza total será igual al número de variables incluidas en el modelo; en cambio, si el análisis se hace utilizando la matriz de covarianza, las variables permanecerán en su métrica original. Sin embargo, el usuario deberá tener cuidado que las variables que incluya en el análisis tengan métricas similares. Para el caso de los indicadores incluidos en el SCINCE, el usuario deberá prestar atención de no incluir indicadores en absolutos y en porcentajes en un mismo modelo.

- Importancia de las componentes principales

Para cada una de las componentes se presentan los valores característicos, la desviación estándar y el porcentaje de varianza total explicada; adicionalmente se presenta el porcentaje de varianza total explicada de forma acumulada para las componentes. En este apartado el usuario deberá evaluar si la estratificación por medio de la primera componente principal y el método de Dalenius-Hodges es adecuada.

- Vectores de coeficientes para las componentes

Por medio de estos coeficientes, el usuario puede identificar la importancia de las variables consideradas para la estratificación. Las variables con coeficientes muy pequeños en la primera componente principal no contribuirán en realidad a la estratificación, sin embargo es posible que estas variables sí sean significativas en las demás componentes principales.

- Estratificación de la primera componente principal por medio del método de Dalenius-Hodges.

Se presentan los límites de cada estrato obtenido mediante el método de Dalenius-Hodges, es decir, el valor mínimo y máximo que se permiten para que una observación, al ser evaluada en la primera componente principal, quede incluida en un estrato dado. Se proporciona también el valor promedio de la primera componente principal en cada estrato; este dato permite observar qué tan distantes se encuentran los centroides de cada estrato, es decir, qué tan diferenciados están los estratos formados.

- Prueba Kaiser-Meyer-Olkin

La prueba Kaiser-Meyer-Olkin ayuda a determinar si los datos son adecuados para un análisis de componentes principales. El resultado de la prueba arroja un valor entre cero y uno, es deseable que el valor sea lo más cercano posible a uno y se sugiere 0.5 como valor mínimo aceptable. Los detalles técnicos de esta prueba pueden consultarse en la bibliografía proporcionada.

1.2.2 Gráfica de sedimentación

En el análisis de componentes principales, la gráfica de sedimentación ayuda a seleccionar el número de componentes principales que representan mejor a un determinado conjunto de datos. En este caso, es útil para verificar, como ya se mencionó anteriormente, que la primera componente explique la mayor cantidad posible de varianza, por lo que es deseable que en la gráfica se observe un desplome abrupto entre la primera y segunda componente.

1.2.3 Biplot

Los gráficos biplot representan al mismo tiempo las observaciones y las variables de un conjunto de datos, respecto a las dos primeras componentes principales. Las observaciones o valores de las variables están representados por puntos en el plano y las variables están representadas por vectores, con las siguientes características:

- La longitud de cada vector indica la importancia de cada variable en el modelo, de esta manera, vectores cortos indican que la variable es susceptible a eliminarse del modelo.
- El ángulo entre dos vectores representa el grado de correlación entre dos variables; cuanto menor sea el ángulo, mayor es el grado de correlación entre éstas variables. De esta manera, si el ángulo entre dos vectores es muy pequeño, puede optarse por eliminar una de las dos variables del modelo, de preferencia aquella cuyo vector sea más corto.
- La distancia entre los puntos son una medida de disimilitud de las observaciones reales, así dos puntos cercanos en el plano implican dos observaciones similares según las variables que se usan para clasificación. También permite identificar observaciones atípicas que se ubicarían muy alejadas del resto de las observaciones. En estos casos se puede considerar repetir el análisis eliminando estas observaciones atípicas y analizar los resultados obtenidos.

En el caso del SCINCE 2010, las observaciones se identifican en esta gráfica mediante su clave geográfica.

1.2.4 Gráfico de centroides

El gráfico de centroides muestra el valor promedio de cada variable en cada uno de los estratos conformados. Esta gráfica permite visualizar el comportamiento de las variables seleccionadas en los estratos, observando así las diferencias entre éstos.

2. Método de k-medias

2.1 k-medias

El método de k-medias es un algoritmo de formación de estratos que asigna cada elemento al estrato que tiene el centroide (punto medio) más cercano. El método se compone de los siguientes pasos:

1. Seleccionar al azar los *k* centroides iniciales de entre los datos.
2. Asignar cada elemento al estrato con el centroide más cercano.
3. Recalcular los centroides de los estratos resultantes en el paso 2.

Los pasos 2 y 3 se repiten hasta que los estratos conformados sean lo más homogéneos internamente y lo más disimiles entre sí.

Para medir las distancias entre los centroides y las observaciones, y entre cada uno de los centroides, se pueden utilizar varias medidas, aunque la más común es la distancia euclidiana que es la que se emplea en el caso de SCINCE 2010.

De manera más formal, el método de k-medias busca minimizar la suma de cuadrados del error intra-estrato. Así, si $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}$ son las variables originales, el método de k-medias tiene como objetivo determinar la conformación de un conjunto de *k* estratos $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ que minimice la suma de cuadrados del error intra-estrato, es decir

$$\underset{S}{\arg min} = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i}^{i} \left\| \boldsymbol{x}_j - \boldsymbol{\mu}_i \right\|^2$$

donde $\boldsymbol{\mu}_i$ es el vector de centroides del estrato S_i.

Un criterio de juicio para determinar si los conglomerados formados son adecuados, es comparar las sumas de cuadrados dentro de cada conglomerado con la suma de cuadrados entre los conglomerados; se espera que la suma de cuadrados dentro de los conglomerados sea menor. Los detalles técnicos sobre la conformación de estratos mediante el método k-medias pueden ser consultados en la bibliografía proporcionada.

2.2 Resultados de la estratificación

A continuación se da una breve descripción de las salidas correspondientes a la estratificación con el método de k-medias. Para el caso de la estratificación por el método de k-medias, adicionalmente a los resultados de la estratificación, se proporcionan las salidas correspondientes a un análisis de componentes principales con el fin que el usuario pueda realizar una análisis

exploratorio de las variables consideradas en el modelo. Este análisis de componentes principales es independiente de la estratificación por el método de k-medias, la descripción de las salidas de éste pueden consultarse en la sección correspondiente a la estratificación por medio del método de componentes principales y Dalenius-Hodges. Se recomienda al usuario consultar el estudio de caso contenido en esta aplicación así como la bibliografía presentada.

2.2.1 Estratificación por medio del método k-medias

Para cada una de las variables consideradas en el modelo, se presenta el valor promedio para cada estrato. De igual manera, se proporcionan las sumas de cuadrados del error al interior de cada estrato, la suma de cuadrados del error total y la suma de cuadrados entre estratos y la frecuencia de observaciones obtenida en cada estrato.

2.2.2 Dendograma

El dendograma tiene como objetivo facilitar la interpretación de los resultados obtenidos mediante la estratificación. Esencialmente, mediante esta gráfica se representa la formación de estratos así como la distancia entre ellos. Por medio de esta gráfica en forma de árbol invertido, el usuario puede identificar observaciones atípicas, y en algunos casos esta gráfica será de utilidad para determinar sí el número de estratos predefinidos es el más adecuado.

2.2.3 Biplot

Esta gráfica forma parte del análisis de componentes principales, en síntesis se representan al mismo tiempo las observaciones y las variables de un conjunto de datos, respecto a las dos primeras componentes principales. Las observaciones o valores de las variables están representados por puntos en el plano y las variables están representadas por vectores. En el apartado correspondiente a la estratificación por medio del método de componentes principales y Dalenius-Hodges se puede consultar una descripción un poco más detallada de esta gráfica.

2.2.4 Histograma

El histograma permite observar gráficamente la distribución de las observaciones en cada uno de los estratos. Con esto, el usuario podrá determinar si los estratos son homogéneos en cuanto al número de observaciones que contienen, o bien, si uno de los estratos resultantes contiene muy pocas observaciones, lo que pudiera indicar la presencia de observaciones atípicas.

Referencias Bibliográficas

Dalenius T. and Hodges J. (1959) *Minimum Variance Stratification*. Journal of the American Statistical Association Vol. 54, No. 285 p. 88-101

Everitt B. (2001) *Applied Multivariate Data Analysis*. Arnold

Everitt B. (2011) *Cluster Analysis 5th edition*. Wiley

Johnson D. (1998) *Métodos Multivariados Aplicados al Análisis de Datos*. Thomson Editores

Joliffe I. T. (2002) *Principal Component Analysis, Second Edition*. Springer Verlag

Levy J.P. (2005) *Análisis Multivariable para las Ciencias Sociales.* Pearson Educación

MacQueen J. B. (1967) *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Mardia K.V. (1979) *Multivariate Analysis*. Academic Press

Morrison D. (1967) *Multivariate Statistical Methods*. McGraw-Hill

Seber G. (1976) *Multivariate Observations*. Wiley