Metodologías de estratificación

Andrés Gutiérrez

Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Tabla de contenidos I

Metodologías de estratificación

Dimensiones Estructurales en el Marco de Muestreo

Información a nivel de UPM

Metodologías univariadas sobre medidas de resumen

Metodologías multivariadas sobre la matriz de información

Evaluación y escogencia de la mejor estratificación

Metodologías de estratificación

Introducción

- La importancias de la estratificación es que permite aumentar la eficiencia de la inferencia en las encuestas de hogares.
- La estratificación en el marco de muestreo permite clasificar a las UPM de acuerdo con su nivel socio-económico con el fin de poder realizar selecciones independientes en cada categoría de la clasificación.
- La estratificación garantiza homogeneidad dentro de grupos y reduce la incertidumbre de la estimación.

Procesos que Intervienen en la Estratificación

1. Ejecución de múltiples metodologías de estratificación de las UPM utilizando información agregada del censo.
2. Para cada método señalado anteriormente realizar particiones de 3, 4, o 5 grupos a nivel nacional y evaluar la pertinencia de realizarlo en las áreas rural y urbana de forma independiente.
3. Evaluar su efectividad mediante una única medida de calidad, definida como el DEFF generalizado y escoger el mejor escenario en términos de esta medida en conjunción con la viabilidad logística con respecto al número de particiones.

Dimensiones Estructurales en el Marco de Muestreo

Introducción

- Independientemente de las unidades de muestreo y jerarquías definidas, es crucial llevar a cabo un proceso de estratificación de las UPM. Esto implica agruparlas en categorías homogéneas en términos socioeconómicos y de bienestar, lo que define una partición del territorio nacional.
- Los INE utilizan las particiones geográficas y cartográficas generadas en el levantamiento del censo con el fin de seleccionar muestras de hogares, mediante la ejecución de diseños de muestreo probabilísticos, estratificados y en varias etapas

Homogeneidad dentro de los estratos y variables de estratificación

1. **Importancia de la homogeneidad**:
   - Reduce la incertidumbre de la estimación.
   - Minimiza los errores de muestreo en encuestas probabilísticas.
   - Garantiza homogeneidad dentro de grupos y reduce la incertidumbre de la estimación.
2. **Variables de estratificación en Latinoamérica**:
   - Información censal a nivel de personas, hogares y viviendas.
   - Dimensiones asociadas a la calidad de vida y bienestar: Demografía, Características de la vivienda, Tenencia de enseres y servicios públicos.

Ejemplos de Variables:

- Acceso a internet (Discrimina entre los hogares con mejores condiciones de bienestar)

  1: si el hogar tiene internet.

  0: si el hogar no tiene internet.

- Materialidad de la vivienda (Mejores materiales se asocian a una mayor capacidad económica y mejores condiciones habitacionales):

  1: si no hay materiales precarios.

  0: si hay materiales precarios.

- Educación del jefe del hogar:
- Ocupación del jefe del hogar:

Consideraciones:

- La estratificación se realiza a nivel de UPM.
- Todos los componentes de una UPM comparten la misma categoría, por consiguiente, las personas, los hogares y las viviendas de la UPM pertenecerán al estrato en el cual la UPM fue clasificada.
- Análisis exploratorio de variables candidatas.

Selección y Definición de Variables de Estratificación

- Relacionadas con los fenómenos a estudiar en las encuestas.
- Información disponible en el censo.
- Representativas de las dimensiones de calidad de vida y bienestar.

Dimensiones Relevantes

- **Demografía y estructura de la población:** sexo, edad, parentesco, origen extranjero, pertenencia a pueblos indígenas, número de hijos, etc.
- **Educación:** analfabetismo, asistencia escolar, años de estudios, grado de escolaridad, etc.
- **Mercado de trabajo:** población en edad de trabajar, pertenencia a la fuerza de trabajo por sexo, condición de ocupación por sexo, rama de actividad, etc.
- **Características de la vivienda:** tipo de vivienda, materiales de construcción, hacinamiento, equipamiento, etc.

Dimensiones Relevantes

- **Acceso a servicios:** fuente de agua, alcantarillado, internet, acceso a salud, seguridad social, etc.
- **Tenencia de bienes en el hogar:** televisor, microondas, aire acondicionado, autos, lavalozas automáticos, entre otros.
- **Necesidades básicas insatisfechas (NBI) o pobreza multidimensional:** viviendas con hacinamiento crítico, servicios inadecuados, alta dependencia económica, niños en edad escolar que no asisten a la escuela, precariedad en el aseguramiento en salud, entre otras.

Ejemplos de Estratificación por Dimensiones

- **Demografía:** UPM con mayor población indígena o afrodescendiente, mayor número de niños, hogares uniparentales con madres jefas de hogar.
- **Educación:** UPM con mayores tasas de analfabetismo, niños que no asisten a la escuela, mayor porcentaje de población con estudios de educación superior.
- **Mercado de trabajo:** UPM rurales con alta proporción de población ocupada, mayor incidencia de población desocupada y/o mayor proporción de personas dependientes.
- **Características de la vivienda:** UPM con alto porcentaje de viviendas con materiales de construcción precario, hacinamiento, acceso inadecuado a las fuentes de agua potable, o con servicios sanitarios y de eliminación de aguas grises deficientes.

Información a nivel de UPM

Clasificación de UPM

Tomando en cuenta la información recolectada en el censo, es posible también clasificar a las personas o a los hogares en una primera instancia para después agregarlos hasta llegar a una clasificación única de la UPM; sin embargo, en la práctica este proceso puede resultar un poco más complejo y no son claras sus ventajas. Por lo anteriormente mencionado, nos enfocamos en la **clasificación de las UPM a partir de una matriz de información a nivel de esta misma agregación**.

Dificultades en la estratificación

- Las UPM varían en tamaño, lo que afecta la clasificación.
- La matriz de información utilizada para la estratificación se basa en el número de personas con ciertas características dentro de cada UPM.
- Sin considerar el tamaño de la UPM, las metodologías de estratificación pueden no agruparlas de manera homogénea.

Ejemplo:

- Imagina dos UPM: una con 100 hogares y otra con 300 hogares, que agrupan a 200 y 400 persona en la fuerza laboral.
- Una variable clave es el número de personas ocupadas.
- Si no consideramos el tamaño de la UPM y el nivel socioeconómico, podrían clasificarse erróneamente en el mismo grupo.

Solución propuesta

- Definir la matriz de información en términos relativos (porcentaje de ocurrencia de cada variable).
- Esto controla el agrupamiento por el tamaño de la UPM y se basa en cambios estructurales en los constructos de medición del censo.

Matriz de Información

Una vez que se ha definido el conjunto de variables que entrarán en la matriz de información, es necesario verificar que todos los indicadores de esta matriz apunten hacia el mismo horizonte del constructo censal.

- La maatrix de información X debes estar compuesta por P variables de estratificación y N_I UPM.
- Cada fila representa una UPM con sus valores para las P variables.
- Realizar un proceso de refinamiento sobre esta matriz para eliminar aquellas variables que puedan estar altamente correlacionadas con el resto de las variables o que puedan expresarse como combinación lineal de otras variables

Estratificación óptima

- Minimiza los errores de muestreo de los estimadores.
- La variabilidad de los microdatos observados en el censo afecta las varianzas de los estimadores.
- Lo que podría ser una estratificación óptima para un indicador específico podría no ser eficiente para otros indicadores.
- La elección de la estratificación debe considerar múltiples escenarios y variables de interés.

Metodologías univariadas sobre medidas de resumen

Introducción

- La mejor estratificación para una variable de interés se basa en su propia variación.
- Las técnicas tradicionales se basan en una sola variable, ignorando el carácter multipropósito de las encuestas.
- Se propone partir de la matriz de información y resumir la variación y las correlaciones mediante técnicas multivariadas.
- Se pueden utilizar técnicas como componentes principales, análisis factorial o modelos no lineales.
- La medida de resumen debe ser interpretable y tener sentido en cuanto a la relación entre las variables y el bienestar de la UPM.

Medida de resumen

- Se define como $y = f(x_1, ..., x_P)$, donde x_i son las variables de la matriz de información.
- Se espera que esta variable tenga un comportamiento sesgado
- Se puede crear un estrato de inclusión forzosa con las unidades con mayor sesgo para asegurar un error de muestreo nulo en ese estrato.

Histograma de la medida de resumen (y) sobre las UPM

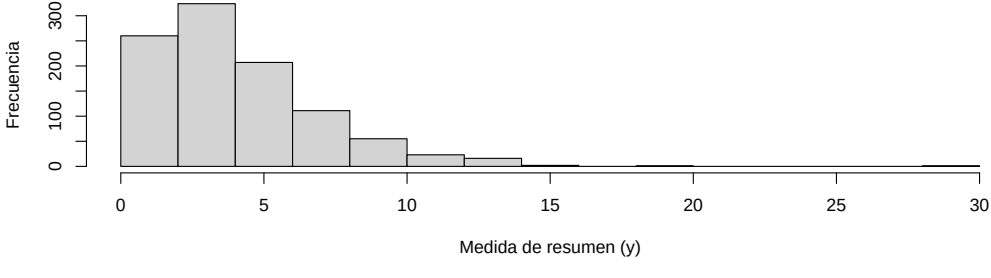

Figura 1: *Histograma de la medida de resumen (y) sobre las UPM*

Partición en cuantiles (Q)

Este método divide la población de UPM en grupos creados a partir de la división en intervalos regulares de la distribución de la medida de resumen. Los cuantiles más usados son los cuartiles, los quintiles y los deciles; sin embargo, con los propósitos de estratificación, también es útil considerar la partición en terciles.

Método de raíz de frecuencia acumulada (DH)

Dalenius y Hodges (1959) propusieron esta técnica de estratificación basada en la raíz cuadrada de las frecuencias acumuladas de la medida de resumen sobre las UPM. Esta técnica es exacta y no requiere de algún procedimiento iterativo. La idea principal de esta técnica es encontrar grupos que minimicen la siguiente función:

$$D = \sum_{h=1}^{H} W_h \sqrt{S_{y_h}^2}$$

En donde $W_h = N_h/N$ $(h = 1, \dots, H)$ es el tamaño relativo del estrato h y $S_{y_h}^2$ es la varianza de la medida de resumen en el estrato h.

Estratificación óptima (LH)

Lavallée y Hidiroglou (1988) propusieron por primera vez la construcción de una estratificación óptima para poblaciones de encuestas reales, basada en la minimización de la siguiente expresión ligada a la varianza de una estrategia de muestreo estratificada.

$$\sum_{h=1}^{H-1} \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{(n-N_H)a_h} - \frac{1}{N_h}\right) S_{x_h}^2$$

En donde N_h es el número de UPM en el estrato h, n es el tamaño de muestra de las UPM, N es el número de UPM en el marco de muestreo, $S_{x_h}^2$ es la varianza de la medida de resumen en el estrato h.

Regla de asignación

La regla de asignación a_h para el tamaño de muestra, dada por la siguiente relación:

$$a_h = \frac{\gamma_h}{\sum_h \gamma_h}$$

En donde, tomando en cuenta que $\bar{X}_h$ es la media de la medida de resumen en el estrato h, entonces, según (Baillargeon y Rivest 2011), γ_h es proporcional al tamaño de muestra n y está definida por:

$$\gamma_h = N_h^{2q_1} \times \bar{X}_h^{2q_2} \times S_{x_h}^{2q_3}$$

Tipos de asignación:

- Asignación proporcional:

$$\mathbf{q} = (0.5, 0, 0)'$$

- Asignación de Neyman:

$$\mathbf{q} = (0.5, 0, 0.5)'$$

- Asignación de potencia con exponente 0.7:

$$\mathbf{q} = (0.35, 0.35, 0)'$$

Los detalles técnicos de estos tipos de asignación pueden ser encontrados en Gutiérrez (2016).

Estratificación geométrica (GH)

Utilizando las técnicas de estratificación mencionadas anteriormente, algunos autores se percataron de que, para poblaciones de UPM con medidas de resumen sesgadas, las varianzas relativas (coeficientes de variación) de la medida de resumen en cada estrato eran similares; es decir:

$$\frac{S_{x_1}}{\bar{X}_1} \cong \frac{S_{x_2}}{\bar{X}_2} \cong ... \cong \frac{S_{x_H}}{\bar{X}_H}$$

Propuesta de Gunning y Horgan (2004)

- Los límites de los estratos se definen en progresión geométrica.
- Siendo X la variable que contiene la información de la medida de resumen para todas la UPM del marco de muestreo, entonces los límites de los estratos estarán dados por la siguiente expresión:

$$b_h = \min(X)\left(\frac{\max X}{\min X}\right)^{h/L}; \qquad h = 1, 2, \ldots, H-1.$$

Es posible encontrar que los coeficientes de variación de los estratos conformados por estos límites son equivalentes y por ende, este método es óptimo para encontrar mejores formas de estratificar teniendo en cuenta como función objetivo la variación relativa dentro los estratos.

Metodologías multivariadas sobre la matriz de información

Introducción

Teniendo en cuenta que en el periodo intercensal se realizarán encuestas que miden variables que están fuertemente ligadas a las observadas en el censo, entonces encontrar una estratificación que sea óptima para todo el conjunto de variables de la matriz de información asegurará una partición óptima para todas las encuestas realizadas en el periodo intercensal.

K-medias de Jarque (KmJ)

Jarque (1981) propuso utilizar una versión modificada del algoritmo de K-medias (Macqueen 1967), cuyo objetivo es la minimización de la siguiente función de distancia:

$$\sum_{h=1}^{H} \sum_{k \in U_h} (\mathbf{x}_k - \bar{\mathbf{x}}_h)' \Lambda^{-1} (\mathbf{x}_k - \bar{\mathbf{x}}_h)$$

En donde $\mathbf{x}_k$ corresponde a la medición de las P variables de la matriz de información en la k-ésima UPM, $\bar{\mathbf{x}}_h$ es el vector de medias de la matriz de información en el estrato h y Λ es una matriz diagonal de tamaño $P \times P$ cuyas entradas se definen como la varianza de las P variables de la matriz $\mathbf{X}$.

Partición genética Ballin y Barcaroli (2013)

- La partición genética (BB) es un método multivariado de estratificación que busca minimizar el costo de la muestra mientras se asegura la precisión de las estimaciones para un conjunto de variables.
- Función de costo:

$$c_0 + \sum_{h=1}^{H} c_h n_h$$

En donde c_0 define un costo fijo y c_h es el costo promedio de observar un hogar en el estrato h. En principio, es posible definir $c_0 = 0$ y $c_1 = c_2 = \cdots = c_H = 1$, lo cual da como resultado que el costo es el número de encuestas que deben realizarse en cada estrato.

Restricciones

El problema de optimización se complementa manteniendo las siguientes restricciones:

$$\sum_{h=1}^{H} \left(\frac{N_h^2}{n_h} \right) \left(1 - \frac{n_h}{N_h} \right) S_{x_h,p}^2 \leq V_{0p} \qquad p = 1, 2, \ldots, P.$$

En donde V_{0p} es un umbral predefinido por el usuario, que indica que la varianza de la estrategia estratificada está acotada; además, $S_{x_h,p}^2$ es la varianza poblacional de p-ésima variable de la matriz de información en el estrato h.

Algoritmo

- Se comienza con estratificaciones univariadas independientes para cada variable.
- Se crea el producto cartesiano de estas estratificaciones (estratos atómicos).
- Se utiliza un algoritmo genético evolutivo para unir grupos de forma jerárquica.
- El proceso se detiene cuando se alcanza el número de estratos deseado (H) y se cumplen las restricciones de precisión.

Evaluación y escogencia de la mejor estratificación

Evaluación de las estratificaciones

- Se evalúan las técnicas univariadas y multivariadas.
- Se calcula el DEFF para cada variable de la matriz de información.
- Se comparan los DEFF de las diferentes estratificaciones.

$$DEFF_p = \frac{Var_{ST}(\bar{x}_p)}{Var_{SI}(\bar{x}_p)} \qquad p = 1, \ldots, P.$$

Donde:

- $DEFF_p$: Efecto de diseño para la variable p.
- $Var_{ST}(\bar{x}_p)$: Varianza del diseño estratificado para la media poblacional de la variable p.
- $Var_{SI}(\bar{x}_p)$: Varianza de un muestreo aleatorio simple para la media poblacional de la variable p.

Evaluación de las estratificaciones

Por otro lado, Gutiérrez (2016, 184) demuestra que, cuando la asignación es proporcional, esta relación se puede escribir de la siguiente manera:

$$DEFF_p = \frac{\sum_{h=1}^{H} W_h S^2_{x_{hp}}}{S^2_{x_p}} \cong 1 - R^2_p \qquad p = 1, \dots, P.$$

En donde, para cada estrato $h = 1, \dots, H$, se tiene que $S^2_{x_p}$ es la varianza de la variable x_p en la población y $S^2_{x_{hp}}$ es la varianza de la variable x_p supeditada al estrato h.

Efecto de diseño generalizado, Jarque (1981)

El efecto de diseño generalizado (G(S)) es una medida de la calidad de una estratificación multivariante. Se define como:

$$G(S) = \sum_{p=1}^{P} DEFF_p = \sum_{p=1}^{P} \frac{1}{S^2_{x_p}} \sum_{h=1}^{H} W_h S^2_{x_{hp}}$$

Ante una estratificación pertinente, se esperaría que $Var_{ST}(\bar{x}_p) < Var_{SI}(\bar{x}_p)$, por lo tanto $0 < DEFF_p < 1$, lo que conlleva a que $0 < G(S) < P$. Luego, se debería escoger el escenario para el cual $G(S)$ fuera mínimo.

Ejemplo

En el se cuadro ejemplifica la evaluación de estas técnicas para dos escenarios de estratificación (tres y cuatro estratos) en una matriz de información que contiene 8 variables.

Tabla 1: *Efectos de diseño $DEFF_p$ y efecto de diseño generalizado $G(S)$ considerando tres ($H = 3$) y cuatro ($H = 4$) estratos para ocho variables.*

DEFF	Q (H=3)	DH (H=3)	LH (H=3)	GH (H=3)	KmJ (H=3)	BB (H=3)	Q (H=4)	DH (H=4)	LH (H=4)	GH (H=4)	KmJ (H=4)	BB (H=4)
$\bar{x}_1$	0.87	0.85	0.81	0.82	1	0.88	0.8	0.70	0.76	0.72	0.71	0.77
$\bar{x}_2$	0.89	0.82	0.95	0.97	0.94	0.88	0.79	0.74	0.75	0.77	0.75	0.71
$\bar{x}_3$	0.87	0.97	0.83	0.96	0.89	0.95	0.74	0.75	0.79	0.7	0.79	0.71
$\bar{x}_4$	0.92	0.89	0.81	0.94	0.96	1	0.77	0.73	0.73	0.7	0.71	0.74
$\bar{x}_5$	0.85	0.83	0.96	0.96	0.83	0.81	0.8	0.73	0.8	0.78	0.8	0.79
$\bar{x}_6$	0.87	0.88	0.9	0.88	0.86	0.81	0.8	0.72	0.76	0.7	0.74	0.73
$\bar{x}_7$	0.87	0.95	0.99	0.83	0.86	0.84	0.75	0.7	0.77	0.72	0.77	0.77
$\bar{x}_8$	0.93	0.82	0.91	0.99	0.93	0.88	0.77	0.74	0.72	0.78	0.76	0.75
G(S)	7.07	7.01	7.16	7.35	7.27	7.05	6.22	5.81	6.08	5.87	6.03	5.97

Comparabilidad y consistencia del proceso de estratificación

1. Evaluación independiente de áreas urbanas y rurales:

- Se aplican los algoritmos de evaluación a las UPM urbanas de forma independiente a las rurales.
- Se define si la estratificación independiente es más eficiente.

2. Evaluación conjunta de áreas urbanas y rurales:

- Se consideran las UPM de ambas áreas juntas.
- Se define si la comparabilidad entre estratos es imperante.
- La clasificación de las UPM urbanas se rige por las mismas condiciones que las rurales.

Viabilidad del número de estratos

- Se discute el número de estratos con las áreas involucradas en el INE.
- Se recomienda considerar $H = 3$ o $H = 4$ estratos.
- Se debe tener en cuenta un tamaño de muestra mínimo de dos UPM por estrato.

Consideraciones

El efecto diseño no es el único aspecto por evaluar para la elección del procedimiento de estratificación. Es necesario verificar la estabilidad del método con respecto a los otros procedimientos de estratificación. Por ejemplo, la siguiente tabla muestra la matriz de coincidencias entre las diferentes clasificaciones de los estratos.

Tabla 2: *Matriz de coincidencias, cuyas entradas están definidas como el porcentaje de UPM coincidentes en cada uno de los estratos creados por los métodos estudiados.*

Técnica	Jarque	K-means	DAL	GEO	LH-S	LH-K	Percentil
Q	1	0,64	0,92	0,84	0,89	0,89	0,82
DH	0,64	1	0,68	0,62	0,71	0,71	0,74
LH	0,92	0,68	1	0,82	0,96	0,96	0,90
GH	0,84	0,62	0,82	1	0,78	0,78	0,73
KmJ	0,89	0,71	0,96	0,78	1	1,00	0,93
BB	0,89	0,71	0,96	0,78	1,00	1	0,93

Puntos clave a considerar son:

1. **Comportamiento Esperado en los Estratos**:
   - Las proporciones de personas mayores de 15 años alfabetizadas deberían tener mayor incidencia en los estratos más altos.
   - Este patrón también debería observarse para otros indicadores, como la proporción de hogares con internet, tenencia de refrigerador, televisión por cable, automóvil, saneamiento adecuado, pisos adecuados y personas con educación superior.

Puntos clave a considerar son:

La **Figura** muestra el comportamiento esperado en los estratos de muestreo para algunas de estas variables.

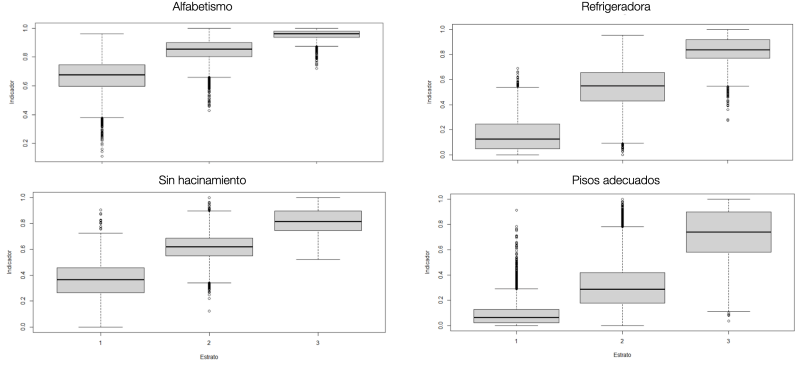

Figura 2: Comportamiento esperado en los estratos de muestreo para algunas variables de interés.

Observaciones Generales.

- El **estrato uno** debería representar condiciones económicas más adversas.
- El **estrato dos** debería tener mejores condiciones.
- El **tercer estrato** agrupa a las UPM con menores dificultades socioeconómicas.
- En áreas rurales, se espera una menor proporción de UPM en el **estrato 3**, debido a condiciones menos favorables.
- Si algunas unidades contribuyen de manera no significativa al total poblacional o son de difícil acceso, se puede crear un **estrato de exclusión forzosa**.
- En este estrato, no se realiza ninguna encuesta y las estimaciones no consideran a esta población excluida.

Estratificación implícita

La estratificación implícita es una herramienta valiosa en el proceso de muestreo, especialmente cuando se busca una asignación proporcional de hogares en los estratos sin necesariamente controlar el tamaño de la muestra final. A continuación, vemos los puntos clave sobre la estratificación implícita:

1. **Selección Ordenada y Proporcional**:
   - La estratificación implícita garantiza una asignación estrictamente proporcional de hogares en los estratos.
   - No se requiere control sobre el tamaño de la muestra final.
   - No se asume independencia en la selección.

Estratificación implícita

2. **Correlación con Indicadores de Interés**:
   - La estratificación implícita funciona bien cuando las variables de estratificación están correlacionadas con indicadores específicos (por ejemplo, tasas de desocupación, subocupación o informalidad).
3. **Enfoque en Temas Específicos**:
   - Es especialmente útil para encuestas enfocadas en temas particulares, como el mercado de trabajo.
   - Se utiliza con muestreo sistemático (probabilidades simples o desiguales) en la selección de Unidades Primarias de Muestreo (UPM).

Estratificación implícita

4. **Secuencia de Selección**:
   - En muchos países, la secuencia comienza con el área urbana, desagregada por departamento y municipio.
   - Luego se sigue con el área rural, desagregada por departamento y comuna o vereda.
   - La selección sistemática de UPM se basa en el ordenamiento de las UPM por el número de viviendas.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Baillargeon, Sophie, y Louis-Paul Rivest. 2011. «The construction of stratified designs in R with the package stratification». *Survey Methodology* 37 (1): 53-65.

Ballin, Marco, y Giulio Barcaroli. 2013. «Joint determination of optimal stratification and sample allocation using genetic algorithm». *Survey Methodology* 39 (2): 369-93.

Dalenius, Tore, y JosrEPH L Hodges. 1959. «Minimum Variance Stratification». *Journal of the American Statistical Association* 54 (285): 15.

Gunning, Patricia, y Jane M Horgan. 2004. «A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations». *Survey Methodology* 30 (2): 159-66.

Gutiérrez, H. A. 2016. *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Segunda edición. Ediciones de la U.

Jarque, Carlos 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): 163-69. https://doi.org/10.2307/2346387.

Lavallée, Pierre, y Michael A. Hidiroglou. 1988. «On the Stratification of Skewed Populations». *Survey Methodology* 14 (1): 33-43.

Macqueen, J. 1967. «Some methods for classification and analysis of multivariate observations». En *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-97. University of California Press.