Publicación Electrónica ISSN 1668-5008

UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

ACTAS

Jornadas Anuales

"Investigaciones en la Facultad" de Ciencias Económicas y Estadística

Maldonado, Leonor
Marí, Gonzalo
Instituto de investigación, Escuela de Estadística

INTRODUCCIÓN AL DISEÑO DE UNA MUESTRA MAESTRA PARA LA PROVINCIA DE SANTA FE[1]

Resumen:
Las oficinas de estadística utilizan entre sus operativos a las encuestas a hogares para realizar mediciones de indicadores específicos de las problemáticas públicas, con el fin de favorecer la toma de decisiones políticas para mejorar la calidad de vida humana en general. Para ello se valen de operativos que, en su mayoría, se apoyan en muestras que son seleccionadas a través de procedimientos probabilísticos, los cuales aseguran la representatividad de la muestra respecto a la población. Para la selección de una muestra por procedimientos aleatorios, se requiere un marco de muestreo que está compuesto por el conjunto de materiales que sirven para identificar, localizar y acceder a cada uno de los elementos de la población. Una opción para la construcción de un marco es la de seleccionar una muestra grande en una primera fase inicial que posteriormente sirva de marco para seleccionar las muestras. El IPEC tiene como objetivo la realización de una Muestra Maestra en la provincia de Santa Fe, la cual estará basada en el diseño de la MMUVRA desarrollada por el INDEC. El objetivo del presente trabajo es comenzar con el estudio de los distintos componentes que forman parte del diseño muestral a partir del cual se construirá la Muestra Maestra para ir determinando aquellos aspectos que brinden estimaciones más precisas. Se consideró el diseño muestral que se emplea en la localidad más grande de la provincia, Rosario, teniendo como objetivo evaluar la precisión que se obtiene en la estimación de parámetros relacionados con el ámbito laboral que forman parte de la Encuesta Permanente de Hogares. Se consideraron diseños con distintos métodos de selección y distintos tamaños de unidades de muestreo de primera etapa basados en las desagregaciones geográficas censales que considera el INDEC en los Censos de Población. Las unidades de mayor tamaño formadas por la unión de 2 o 3 radios censales tuvieron los mejores comportamientos en términos de precisión, mientras que las desagregaciones menores, radios y segmentos, tuvieron un desempeño pobre. El diseño que considera probabilidades de inclusión distintas para las unidades de primera etapa brinda mejores resultados respecto a la precisión comparado a un diseño con igual probabilidad como el muestreo simple al azar.

Palabras claves: MUESTRA MAESTRA, DISEÑO MUESTRAL COMPLEJO, EFECTO DE DISEÑO.

Abstract:

National Statistical Offices use household surveys among their statistical tools to measure specific indicators of public problems, to take political decision-making to improve the quality of human life in general. To do this, they use operations that rely on samples that are selected through probabilistic procedures, which ensure the representativeness of the sample with respect to the population. For the selection of a sample by random procedures, a sampling frame is required that is made up of the set of materials that serve to identify, locate, and access each of the elements of the population. One option for the construction of a sampling frame is to select a large

[1] Trabajo elaborado en el marco del proyecto titulado: “Métodos muestrales y series de tiempo en las estadísticas oficiales”, dirigido por Gonzalo Marí

sample in a first initial phase that later serves as a frame to select the samples. The objective of IPEC is to carry out a Master Sample in the province of Santa Fe, which will be based on the design of the MMUVRA developed by INDEC. The objective of this work is to begin with the study of the different components that are part of the sample design from which the Master Sample will be built to determine those aspects that provide more precise estimates. The sample design used in the largest town in the province, Rosario, was considered, with the objective of evaluating the precision obtained in the estimation of parameters related to the work environment, which are part of the Permanent Household Survey. Designs with diverse selection methods and different sizes of first-stage sampling units were considered based on the geographical census disaggregation considered by INDEC in the Population Censuses. The largest units formed by the union of 2 or 3 census radios had the best behaviour in terms of precision, while the smaller disaggregation, radios and segmentos, had an inferior performance. The design which considers different inclusion probabilities for the primary stage units provides better results regarding precision compared to a design with equal probability such as simple random sampling.

Keywords: MASTER SAMPLE, COMPLEX SAMPLING DESIGN, DESIGN EFFECT.

Introducción

Las oficinas de estadística utilizan entre sus operativos a las encuestas a hogares para realizar mediciones de indicadores específicos de las problemáticas públicas, con el fin de favorecer la toma de decisiones políticas para mejorar la calidad de vida humana en general. Para ello se valen de operativos que, en su mayoría, se apoyan en muestras que son seleccionadas a través de procedimientos probabilísticos, los cuales aseguran la representatividad de la muestra respecto a la población.

Para la selección de una muestra por procedimientos aleatorios, se requiere un marco de muestreo que está compuesto por el conjunto de materiales que sirven para identificar, localizar y acceder a cada uno de los elementos de la población. El marco debe ser actual y completo, esto es, debe incluir sin omisión ni repetición a todos los elementos de la población. Una opción para la construcción de un marco es la de seleccionar una muestra grande en una primera fase inicial que posteriormente sirva de marco para seleccionar las muestras. Kish denomina a este instrumento marco maestro (*master frame*) (Kish, 1978). El mismo incluye listados de elementos físicos y procedimientos de muestreo que, a su vez, permiten obtener muestras de unidades elementales, sin el esfuerzo de haberlas listado a todas.

Dicho concepto se aplica por primera vez en la Muestra Maestra de Agricultura en Estados Unidos (King y Jesen, 1945), una muestra de escala extremadamente grande; luego se comenzó a aplicar en otros ámbitos. El primer antecedente en Argentina de elaboración de un marco de muestreo nacional para encuestas a hogares data de fines de la década del 60 (INDEC, 1999). Dicho marco fue elaborado con el propósito de extraer periódicamente muestras de viviendas para la Encuesta Nacional de Salud, un proyecto conjunto del Gobierno Nacional, la Organización Panamericana de la Salud y la Asociación de Facultades de Medicina. La encuesta se realizó con cobertura nacional urbana y rural desde 1970 a 1972. Luego de esa experiencia, el Instituto Nacional de Estadística y Censos (INDEC) crea el Marco de Muestreo Nacional Urbano (MMNU) para encuestas a hogares en 1992 (INDEC, 1999) con el propósito de extraer muestras de viviendas para encuestas con alcance nacional urbano, para aglomerados de 5.000 o más habitantes según el Censo Nacional de Población, Hogares y Viviendas de 1991 (Censo 1991). A partir del Censo Nacional de Población, Hogares y Viviendas del año 2001 (Censo 2001), se realiza una ampliación del MMNU para posibilitar la extracción de muestras para encuestas con mayor cobertura, incluyendo a la población de 2000 y más habitantes, según la información del Censo 1991 y algunas actualizaciones realizadas en el Censo 2001. En 2011, y en base a los datos del Censo Nacional de Población, Hogares y Viviendas 2010 (Censo

2010), se desarrolla la Muestra Maestra de Viviendas Urbanas de la República Argentina (MMUVRA), la cual es usada actualmente para seleccionar muestras para todas las encuestas a hogares que lleva a cabo el INDEC.

El diseño muestral de la MMUVRA considera una muestra probabilística, polietápica y estratificada de viviendas. La provincia es la primera variable de estratificación. La selección de la muestra se hizo en todas las provincias con un plan bietápico. En la primera etapa se muestrea, considerando un diseño con probabilidad proporcional al tamaño (PPT) y estratificado, unidades primarias (UPM) definidas como aglomeraciones[2] urbanas con 2.000 y más habitantes, residiendo en viviendas particulares al momento del Censo 2010. Luego, dentro de las UPM muestreadas, se seleccionan unidades secundarias de muestreo (USM) definidas a partir de radios o conjuntos de radios censales[3]. Los procedimientos de selección aplicados en esta segunda etapa fueron los de un muestreo PPT y estratificado. Las probabilidades de selección asignadas a las unidades de muestreo en cada etapa fueron proporcionales al total de población registrada en el Censo 2010. Por último se listan todas las viviendas de las USM muestreadas de cada UPM seleccionadas.

Basado en el diseño de la MMUVRA, el IPEC se propone realizar pruebas de diseños para concretar una Muestra Maestra en la provincia de Santa Fe (MUMSAFE), con el objetivo de seleccionar muestras para encuestas a hogares en el período intercensal. En primer lugar y como objetivo del presente trabajo, se estudia la posibilidad de considerar distintos tamaños de las unidades de muestreo y diferentes métodos de selección de estas.

En la siguiente sección se presenta la metodología sobre las medidas de precisión de las estimaciones de totales para muestreo en dos etapas. Luego, se presentan los resultados en la aplicación de dichas medidas basados en los datos del Censo 2010, y sus respectivas conclusiones.

Metodología

Diseño de muestra

El diseño de la muestra tiene impacto directo en las estimaciones, que se construyen en función a los valores de las variables obtenidos una vez aplicada la encuesta a las unidades seleccionadas y los elementos que componen el diseño utilizado. Ya que su precisión varía conforme se apliquen diferentes diseños, se realizan pruebas de estimaciones planteando diseños muestrales variando las distintas componentes del mismo para evaluar como impactan en la precisión de las estimaciones y, teniendo en cuenta su factibilidad, se busca obtener conclusiones que permitan seleccionar el diseño más eficaz. Para ello a continuación se exponen los diseños propuestos con sus respectivas medidas de precisión.

Diseño de muestras en varias etapas

Es frecuente utilizar en encuestas a hogares un muestreo utilizando más de una etapa. El uso de muestras de varias etapas a menudo puede ser una solución práctica y económica en situaciones en las que no se dispone de una lista de unidades elementales (o analíticas) para el muestreo directo de las mismas, o aun existiendo dicha lista, la selección y concreción de una muestra traería aparejado un aumento de los costos operativos por la gran dispersión geográfica de la misma.

Para adjudicar una muestra entre las diferentes etapas de muestreo, se deben considerar las contribuciones de las mismas a la variancia de un estimador. Estos

[2]Se entiende por aglomeración a una localidad o conjunto de localidades que por continuidad de edificaciones y calles configuran una misma unidad urbana

[3]Los radios censales son unidades menores definidas para la asignación de las cargas de trabajo de censistas; en las zonas urbanas generalmente se definen como un conjunto de manzanas o sectores con alrededor de 300 viviendas

componentes de la variancia generalmente dependen de la variable de análisis y también de la forma del estimador. A continuación, se presentan algunos resultados básicos para estimadores lineales en un muestreo de dos etapas.

Muestreo en dos etapas

Se considera un diseño de muestra de dos etapas en el que las unidades de la primera etapa se seleccionan utilizando un muestreo πps, es decir, con probabilidades proporcionales al tamaño y sin reemplazo. Los elementos se seleccionan en la segunda etapa mediante muestreo simple al azar sin reemplazo (MSA). Sea:

U = universo de UPM

M = número de UPM en el universo

U_i = universo de elementos en UPM i

N_i = número de elementos en la población para UPM i

$N = \sum_{i \in U} N_i$ es el número total de elementos en la población

π_i = probabilidad de inclusión de UPM i

π_{ij} = probabilidad de inclusión conjunta de las UPMs i y j

m = tamaño de muestra de UPMs

n_i = tamaño de muestra de elementos en UPM i

s = conjunto muestra de UPMs

s_i = conjunto de muestra de elementos en UPM i

y_{ik} = variable de análisis para el elemento k en UPM i

$\bar{y}_U$ = media por elemento de la población

$\bar{y}_{U_i}$ = media por elemento de la población en UPM i

El π estimador del total de la población, $t_U = \sum_{i \in U} \sum_{k \in U_i} y_k$, de una variable *Y* viene dado por

$$\hat{t}_\pi = \sum_{i \epsilon s} \frac{\hat{t}_i}{\pi_i}$$

donde $\hat{t}_i = (N_i/n_i) \sum_{k \in s_i} y_k$, es el estimador del total de la UPM *i* bajo MSA. La variancia de diseño del total estimado se puede escribir como la suma de dos componentes:

$$V(\hat{t}_\pi) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{t_i}{\pi_i} \frac{t_j}{\pi_j} + \sum_{i \in U} \frac{N_i^2}{\pi_i n_i} \left(1 - \frac{n_i}{N_i}\right) S_{U2i}^2 \qquad (1)$$

dónde $S_{U2i}^2 = \sum_{k \in U_i} \left(y_k - \bar{y}_{U_i}\right)^2 / (N_i - 1)$ es la variancia de *Y* entre los elementos de la UPM *i*.

Uno de los inconvenientes que trae aparejada la fórmula (1) para el cálculo del tamaño de la muestra de las UPM es que el mismo no se presenta de manera explícita resultando inconveniente a la hora de determinar los tamaños de muestra correspondientes a cada una de las etapas. Es común basar el tamaño de la muestra

en un diseño que sea menos complicado que el que realmente se utiliza. Una alternativa es considerar un MSA de UPMs y USMs como en el caso que se plantea a continuación.

MSA en la primera y segunda etapa.

Se considera una MSA en la primera etapa de *m* de *M* UPMs y una MSA en la segunda etapa de n_i elementos seleccionados de la población de N_i. El π estimador para este diseño viene dado por

$$\hat{t}_\pi = \frac{M}{m}\sum_{i\epsilon s}\frac{N_i}{n_i}\sum_{k\epsilon s_i} y_{ik}.$$

y la variancia es igual a

$$V(\hat{t}_\pi) = \frac{M^2}{m}\frac{M-m}{M}S_{U1}^2 + \frac{M}{m}\sum_{i\epsilon U}\frac{N_i^2}{n_i}\frac{N_i - n_i}{N_i}S_{U2i}^2$$

donde $S_{U1}^2 = \frac{\sum_{i\epsilon U}(t_i - \bar{t}_U)^2}{M-1}$, siendo t_i el total poblacional de *Y* en la UPM *i* y $\bar{t}_U = \sum_{i\epsilon U} t_i / M$ es la media de los totales de las UPMs. La variancia relativa de $\hat{t}_\pi$, $V(\hat{t}_\pi)/t_U^2$ viene dada por

$$V(\hat{t}_\pi)\Big/ t_U^2 = \frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{t_U^2}\frac{M}{m}\sum_{i\epsilon U}\frac{N_i^2}{n_i}\frac{N_i - n_i}{N_i}S_{U2i}^2 \qquad (2)$$

donde $B^2 = S_{U1}^2/\bar{t}_U^2 = M^2 S_{U1}^2/\, t_U^2$ es la variancia relativa unitaria entre los totales de las UPMs.

Si se seleccionan $\bar{n}$ elementos en cada UPM y las fracciones de muestreo de las UPMs y de los elementos dentro de las UPMs son pequeñas, la variancia relativa puede expresarse de manera simplificada como

$$V(\hat{t}_\pi)\Big/ t_U^2 = \frac{B^2}{m} + \frac{W^2}{m\bar{n}} \qquad (3)$$

donde $W^2 = M\sum_{i\epsilon U} N_i^2 S_{U2i}^2/\, t_U^2$ es la componente de variancia dentro de las UPMs. La función *BW2stageSRS* del paquete *PracTools* (Valliant et al, 2022) del programa *R* considera la ecuación (3) como versión aproximada de la variancia relativa del estimador del total. Si todas las UPMs tienen el mismo tamaño, $\bar{N}$, y se seleccionan $\bar{n}$ elementos en cada una, la fracción de muestreo de la segunda etapa es $\bar{n}$ / $\bar{N}$. Esto implica que la muestra es autoponderada, o sea, $\pi_i\pi_{k|i} = \mathrm{m}\,\bar{n}\,/\,\mathrm{M}\,\bar{N}$. Bajo esta situación, la variancia relativa expresada en la ecuación (2) se simplifica a

$$V(\hat{t}_\pi)\Big/ t_U^2 = \frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{m\bar{n}}\frac{\bar{N}-\bar{n}}{\bar{N}}W^2 \qquad (4)$$

Suponiendo que se seleccionan $\bar{n}$ elementos en cada muestra de UPM, y que *m/M* y $\bar{n}/N_i$ son pequeños, la forma más general de la variancia relativa $V(\hat{t}_\pi)/t_U^2$ en la ecuación (2), también se puede escribir, de manera aproximada, en términos de una medida de homogeneidad δ:

$$V(\hat{t}_\pi)\Big/ t_U^2 \cong \frac{\tilde{V}}{m\bar{n}}k[1 + \delta(\bar{n}-1)] \qquad (5)$$

donde $\tilde{V} = S_U^2/\bar{y}_U^2$, $k = (B^2 + W^2)/\tilde{V}$, y

$$\delta = \frac{B^2}{B^2+W^2} \tag{6}$$

En el caso que $N_i = \bar{N}$ y tanto M como $\bar{N}$ sean grandes, se puede expresar la variancia relativa de la variable en la población como la suma de las variancias relativas entre y dentro, o sea,

$$\frac{S_U^2}{\bar{y}_U^2} = \frac{1}{\bar{y}_U^2}\frac{\sum_{i\in U}(y_{ik}-\bar{y}_U)^2}{(N-1)} \cong B^2 + W^2 \tag{7}$$

Si *k* = 1, la ecuación (5) queda expresada de la manera usual que aparece en la bibliografía sobre muestreo. Sin embargo, k depende del tamaño poblacional de elementos por conglomerado, y cuando los mismos presentan una marcada heterogeneidad, el valor de k puede ser muy superior a 1. Tener una variación grande en los tamaños de las UPMs conduce a grandes diferencias en los tamaños de los grupos, N_i, y de los totales, t_i. De esta forma, B^2, la componente de variancia entre UPMs, toma un valor grande. En esos casos, es conveniente utilizar la ecuación (5) con una estimación de k para determinar los tamaños de muestra y calcular estimaciones anticipadas de los coeficientes de variación. La expresión (5) es útil para calcular el tamaño de la muestra, ya que el número de UPMs de la muestra y las unidades de muestra en cada UPM están explícitamente indicados en la fórmula.

También relaciona la variancia del total estimado con la variancia relativa que se obtendría de una muestra aleatoria simple, ya que $\tilde{V}/m\bar{n}$ es la variancia relativa del total estimado en una *MSA* de tamaño $m\bar{n}$ cuando la fracción de muestreo es pequeña. El producto $k[1+\delta(\bar{n}-1)]$ es un tipo de efecto de diseño. La expresión (5) con $k = 1$ generalmente se utiliza sin considerar el método de selección de las muestras de UPMs y de los elementos dentro de las UPM, ni tampoco tiene en cuenta el tipo de estimador que se utilice. Cabe notar que es esta fórmula no es la que corresponde cuando se utilizan métodos de muestreo distintos del MSA en diferentes etapas. Por este motivo, a continuación, se presenta un diseño de muestreo diferente a MSA en la primera etapa.

PPT en la primera y MSA en la segunda etapa.

En este caso, las UPM se seleccionan con reemplazo con distinta probabilidad, y la muestra dentro de cada UPM es seleccionada por MSA. Si bien es cierto que los diseños con reemplazo no se utilizan con frecuencia en la práctica para evitar que las mismas unidades sean seleccionadas más de una vez en la muestra, también lo es que dichos diseños poseen fórmulas de variancia simples. Bajo ciertas condiciones, estas variancias pueden servir de aproximaciones para los diseños sin reemplazo. Bajo este diseño, el estimador pwr del total viene dado por

$$\hat{t}_{pwr} = \frac{1}{m}\sum_{i\in s}\hat{t}_i/p_i$$

donde $\hat{t}_i = \frac{N_i}{n_i}\sum_{k\in s_i} y_{ik}$ es el total estimado de la UPM *i* de una muestra aleatoria simple y p_i es la probabilidad de selección de la UPM *i*. La variancia de $\hat{t}_{pwr}$ viene dada por

$$V(\hat{t}_{pwr}) = \frac{1}{m}\sum_{i\in U} p_i\left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i\in U}\frac{N_i^2}{mp_in_i}(1-\frac{n_i}{N_i})S_{U2i}^2 \tag{8}$$

Haciendo la misma suposición anterior que $\bar{n}$ elementos se seleccionan en cada UPM, la varianza se reduce a

$$V(\hat{t}_{pwr}) = \frac{S_{U1(pwr)}^2}{m} + \frac{1}{m\bar{n}}\sum_{i\in U}(1-\frac{\bar{n}}{N_i})\frac{N_i^2S_{U2i}^2}{p_i}$$

donde $S^2_{U1(pwr)} = \sum_{i\in U} p_i(\frac{t_i}{p_i} - t_U)^2$. Dividiendo por t_U^2 y asumiendo que la fracción de muestreo dentro de la UPM *i*, $\bar{n}$/ Ni, es despreciable, se obtiene la variancia relativa aproximada de $\hat{t}_{pwr}$

$$\frac{V(\hat{t}_{pwr})}{t_U^2} \cong \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}} k[1 + \delta(\bar{n} - 1)] \quad (9)$$

con $V(\hat{t}_{pwr}) = \frac{S_U^2}{\bar{y}_U^2}$, $k = (B^2 + W^2)/\tilde{V}$, y

$$B^2 = \frac{S^2_{U1(pwr)}}{t_U^2}, \quad (10)$$

$$W^2 = \frac{1}{t_U^2}\sum_{i\in U} N_i^2 \frac{S^2_{U2i}}{p_i} \quad (11)$$

$$\delta = {B^2}/{(B^2 + W^2)} \quad (12)$$

La expresión (9) tiene la misma forma que la ecuación (5) pero con diferentes definiciones de B^2 y W^2. La expresión (9) también tiene la interpretación de una variancia relativa MSA, $\tilde{V}/m\bar{n}$, multiplicada por un efecto de diseño, $k[1 + \delta(\bar{n} - 1)]$ como en la ecuación (5). Por lo tanto, se utiliza la ecuación (5) y la (9) para comparar la precisión de ambos diseños respectivamente, mediante el efecto de diseño $k[1 + \delta(\bar{n} - 1)]$.

Resultados

En muchas de las encuestas a hogares que desarrolla el INDEC, se consideran a los aglomerados como dominios de estimación. En estos casos, en cada uno de los aglomerados considerados se seleccionan muestras en dos etapas, considerando en la primera de ellas la selección de radios o conjuntos de radios denominados áreas. En la segunda etapa, se seleccionan viviendas dentro de cada una de las áreas seleccionadas. En las viviendas seleccionadas se recolectan datos de las unidades de análisis que pueden estar constituidas por personas u hogares. En algunas encuestas existe una tercera etapa de selección entre las personas que forman parte de la población objetivo dentro de cada uno de los hogares. De todas formas, para este ejercicio, esta última posibilidad no va a estar cubierta en el mismo.

En este trabajo se analizan resultados obtenidos a partir de las bases del Censo Nacional de Población, Hogares y Viviendas 2010, tomando la localidad de Rosario como dominio de estimación, considerando a la misma de inclusión forzosa. Se selecciona como UPM a radios o conjunto de radios y como USM a las viviendas.

El problema que se plantea en esta instancia del diseño de MUMSAFE es: ¿qué unidad usar en la primera etapa de muestreo dentro de una localidad o un aglomerado? Radios, conjunto de radios o segmentos, y ¿Cómo seleccionar esas unidades? Muestreo simple al azar sin reemplazo (MSA) o con probabilidad proporcional al tamaño (PPT).

Se establece como objetivo probar qué tipo de UPM conviene utilizar en términos de precisión. Las UPM que se plantean son:

- Tramos, unión de 3 radios[4] consecutivos (radiox3)

[4] Los radios con 500 viviendas o más se consideran en forma independiente

- Áreas, unión de 2 radios[5] consecutivos (radiox2)
- Radios, división censal: conjunto aproximado de 300 viviendas
- Segmentos, subdivisión censal: carga del censista alrededor de 30 viviendas

Se considera la estimación de los totales: población económicamente activa (PEA), personas desocupadas, mujeres, varones, mujer entre 14-29 años, mujer entre 30-64 años, varón de 14-29 años y varón de 30-64 años. Estos totales están asociados a una de las principales encuestas socioeconómicas planteadas, la Encuesta Permanente de Hogares de ciudades Nodos (EPH Nodo), dentro del programa de encuestas que se pretende seleccionar de la MUMSAFE.

Se trabaja con la base de datos de personas y la de viviendas habitables. Se define como viviendas habitables[6], a todas aquellas viviendas que al momento del censo están habitadas o son potencialmente habitables. Luego de una depuración de las bases con unidades que operativamente resultaría inconveniente trabajar, se obtienen 919.545 personas en 322.493 viviendas habitables, divididas geográficamente en 62 fracciones, 380 tramos, 551 áreas, 1.066 radios y 13.317 segmentos.

Selección de UPM y USM con muestreo simple al azar

Fijando entonces la muestra en dos etapas dentro de la ciudad de Rosario, se toma la primera etapa como una muestra simple al azar sin reemplazo (MSA) de unidades primarias de muestreo (UPM) y la segunda etapa también como una muestra MSA de viviendas o unidades secundarias de muestreo (USM).

A partir de la función *BW2stageSRS* del paquete *PracTools* del programa *R* se obtiene $S_U^2/\bar{y}_U^2$ la relación unitaria de la población, $B^2 + W^2$ para la comparación, la relación $k = (B^2 + W^2)/(S_U^2/\bar{y}_U^2)$ y la versión completa de δ en la ecuación (6). A continuación, se presentan los resultados correspondientes a los totales de las variables consideradas para cada uno de los tamaños de UPM considerados:

Tabla 1: Componentes entre y dentro de la variancia en el diseño MSA/MSA, UPM tramos y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: tramo	B^2	W^2	$S_U^2/\bar{y}_U^2$	$B^2 + W^2$	k	δ
PEA	0,107	0,941	0,825	1,048	1,271	0,102
Personas desocupadas	0,316	32,711	27,483	33,027	1,202	0,010
mujer	0,129	1,039	0,904	1,168	1,292	0,111
varón	0,175	1,271	1,106	1,446	1,307	0,121
mujer 14-29 años	0,195	7,259	6,285	7,454	1,186	0,026
mujer 30-64 años	0,102	4,069	3,629	4,172	1,150	0,025
varón 14-29 años	0,219	7,473	6,407	7,692	1,200	0,028
varón 30-64 años	0,133	4,679	4,108	4,812	1,171	0,028

Tabla 2: Componentes entre y dentro de la variancia en el diseño MSA/MSA, UPM áreas y USM viviendas. Ciudad de Rosario. Censo 2010

[5] Los radios con 500 viviendas o más se consideran en forma independiente

[6] Incluye las categorías viviendas con personas presentes y ausentes, como también las viviendas en venta/alquiler y en construcción.

UPM= área	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,111	0,947	0,825	1,058	1,283	0,105
Personas desocupadas	0,337	32,934	27,483	33,271	1,211	0,010
Jefe	0,067	2,138	1,929	2,206	1,143	0,030
mujer	0,136	1,047	0,904	1,183	1,308	0,115
varón	0,186	1,280	1,106	1,466	1,326	0,127
mujer 14-29 años	0,216	7,333	6,285	7,549	1,201	0,029
mujer 30-64 años	0,105	4,085	3,629	4,190	1,155	0,025
varón 14-29 años	0,239	7,545	6,407	7,784	1,215	0,031
varón 30-64 años	0,137	4,703	4,108	4,840	1,178	0,028

Tabla 3: Componentes entre y dentro de la variancia en el diseño MSA/MSA, UPM radios y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: radio	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,180	1,011	0,825	1,191	1,445	0,151
Personas desocupadas	0,505	35,566	27,483	36,072	1,313	0,014
mujer	0,211	1,120	0,904	1,331	1,472	0,159
varón	0,274	1,369	1,106	1,644	1,486	0,167
mujer 14-29 años	0,331	7,887	6,285	8,218	1,308	0,040
mujer 30-64 años	0,172	4,341	3,629	4,513	1,244	0,038
varón 14-29 años	0,355	8,113	6,407	8,468	1,322	0,042
varón 30-64 años	0,209	5,010	4,108	5,218	1,270	0,040

Tabla 4: Componentes entre y dentro de la variancia en el diseño MSA/MSA, UPM segmentos y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: segmento	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,382	1,197	0,825	1,579	1,915	0,242
Personas desocupadas	1,773	43,241	27,483	45,013	1,638	0,039
mujer	0,438	1,349	0,904	1,786	1,976	0,245
varón	0,567	1,649	1,106	2,216	2,004	0,256
mujer 14-29 años	0,777	9,669	6,285	10,447	1,662	0,074
mujer 30-64 años	0,345	5,067	3,629	5,412	1,491	0,064
varón 14-29 años	0,837	9,996	6,407	10,832	1,691	0,077
varón 30-64 años	0,429	5,922	4,108	6,350	1,546	0,068

Los valores de δ oscilan entre 0,010 y 0,121 cuando las UPM son tramos (Tabla 1). Cuando las UPM son áreas, los valores de δ oscilan entre 0,010 y 0,127 (Tabla 2). Mientras que las δ van desde 0,014 a 0,167 cuando las UPM se toman como radios (Tabla 3) y de 0,039 a 0,256 para segmentos como UPM (Tabla 4). Las medidas de homogeneidad aumentan sustancialmente cuando se toman las agrupaciones como radios o segmentos. Por otra parte, se observa un aumento de los valores de *k* a medida que disminuyen los tamaños de los conglomerados propuestos, siendo esta diferencia mínima cuando se consideran tramos y áreas como UPM. Cabe recordar que el aumento de los valores de *k* es un indicativo de la heterogeneidad de las unidades de muestreo de primera etapa consideradas respecto al tamaño.

Selección de UPM y USM con muestreo proporcional al tamaño y simple al azar

Se considera la selección de viviendas habitables a través de un diseño en dos etapas donde en la primera, las UPMs son seleccionadas con probabilidad proporcional al tamaño (PPT) con reemplazo, y en la segunda etapa a través de una MSA.

Se calcula, B^2+W^2 según las ecuaciones (10) y (11), la relación $k=(B^2+W^2)/(S_U^2/\bar{y}_U^2)$ y la versión de δ de la ecuación (12). A continuación, se presentan los resultados correspondientes a los distintos tamaños de UPMs consideradas:

Tabla 5: Componentes entre y dentro de la variancia en el diseño PPT/MSA, UPM tramos y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: tramo	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,065	0,911	0,825	0,976	1,184	0,066
Personas desocupadas	0,279	31,693	27,483	31,972	1,163	0,009
mujer	0,092	1,006	0,904	1,098	1,214	0,084
varón	0,139	1,230	1,106	1,369	1,238	0,101
mujer 14-29 años	0,160	7,032	6,285	7,192	1,144	0,022
mujer 30-64 años	0,062	3,931	3,629	3,994	1,101	0,016
varón 14-29 años	0,185	7,244	6,407	7,428	1,159	0,025
varón 30-64 años	0,092	4,521	4,108	4,614	1,123	0,020

Tabla 6: Componentes entre y dentro de la variancia en el diseño PPT/MSA, UPM áreas y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: área	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,071	0,917	0,825	0,988	1,199	0,072
Personas desocupadas	0,302	31,978	27,483	32,280	1,175	0,009
mujer	0,099	1,014	0,904	1,113	1,230	0,089
varón	0,150	1,240	1,106	1,389	1,256	0,108
mujer 14-29 años	0,176	7,096	6,285	7,272	1,157	0,024
mujer 30-64 años	0,099	4,551	4,108	4,650	1,132	0,021
varón 14-29 años	0,203	7,314	6,407	7,517	1,173	0,027
varón 30-64 años	0,068	3,953	3,629	4,021	1,108	0,017

Tabla 7: Componentes entre y dentro de la variancia en el diseño PPT/MSA, UPM radios y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: radios	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,087	0,936	0,825	1,024	1,241	0,085
Desocupadas/os	0,386	32,794	27,483	33,180	1,207	0,012
mujer	0,122	1,037	0,904	1,158	1,281	0,105
varón	0,180	1,268	1,106	1,448	1,309	0,124
mujer 14-29 años	0,230	7,297	6,285	7,527	1,198	0,031
mujer 30-64 años	0,083	4,016	3,629	4,099	1,129	0,020
varón 14-29 años	0,257	7,518	6,407	7,775	1,213	0,033
varón 30-64 años	0,117	4,636	4,108	4,753	1,157	0,025

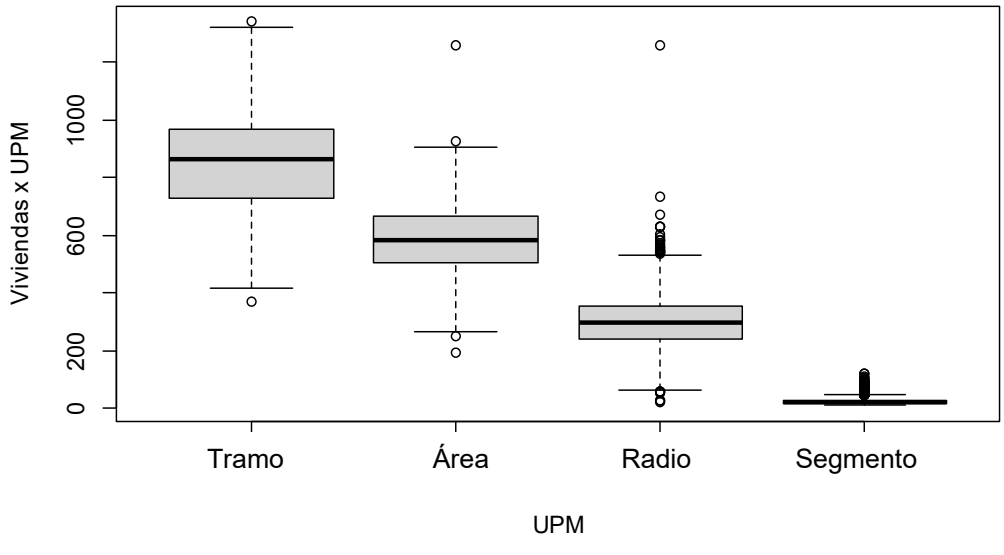
Tabla 8: Componentes entre y dentro de la variancia en el diseño PPT/MSA, UPM segmentos y USM viviendas. Ciudad de Rosario. Censo 2010

UPM: segmentos	B^2	W^2	$S_U^2/\bar{y}_U^2$	B^2+W^2	k	δ
PEA	0,183	1,009	0,825	1,192	1,446	0,154
Desocupadas/os	1,325	35,730	27,483	37,055	1,348	0,036
Jefa/jefe	0,102	2,280	1,929	2,382	1,235	0,043
mujer	0,225	1,138	0,904	1,363	1,507	0,165
varón	0,311	1,392	1,106	1,703	1,540	0,183
mujer 14-29 años	0,480	8,062	6,285	8,542	1,359	0,056
mujer 30-64 años	0,174	4,332	3,629	4,505	1,242	0,039
varón 14-29 años	0,518	8,315	6,407	8,833	1,379	0,059
varón 30-64 años	0,227	5,026	4,108	5,252	1,279	0,043

Con éste nuevo diseño, los valores de δ oscilan entre 0,009 y 0,101 cuando las UPM son tramos (Tabla 5). Cuando las UPM son áreas, los valores de δ varían entre 0,009 y 0,108 (Tabla 6). Mientras que las δ van desde 0,012 a 0,124 cuando las UPM se toman como radios (Tabla 7) y de 0,036 a 0,183 para segmentos como UPM (Tabla 8). Las medidas de homogeneidad siguen aumentando cuando se toman los conglomerados más pequeños como radios o segmentos. Las mismas conclusiones respecto a los *k* se observa en este caso, si bien las diferencias no son tan grandes como en el caso anterior.

Para observar que está sucediendo con los tamaños de las UPM estudiadas a continuación se presenta la distribución de viviendas por UPM en Rosario según el Censo 2010.

Gráfico 1: Distribución de viviendas por UPM. Rosario. Censo 2010



Como se aprecia en el Gráfico 1, la variación de los tamaños de las divisiones geográficas en la población de Rosario es considerablemente mayor de las que se preferirían al definir las UPM. El rango del número de viviendas por tramo es de 372 a 1.339 viviendas, con una media de 865 viviendas. El rango de las áreas varía de 194 a 1.260 viviendas, con una media de 584. El de los radios va desde 23 a 1.260; su media de 295 viviendas. Los segmentos poseen desde 10 hasta 120 viviendas, con una media de 23 viviendas. Tener una variación tan grande en los tamaños de las UPM conduce a grandes diferencias en los tamaños de los grupos (N_i) y los totales (t_i). Esto hace que la componente de variancia entre, B^2, sea grande, lo que a su vez conduce a las altas medidas de homogeneidad, vistas anteriormente. Esta es también la razón por la que la aproximación $(S_U^2/\bar{y}_U^2) = B^2 + W^2$ es pobre mayormente para las variables de UPM tomados como radios o como segmentos. Se observa que la distribución más dispersa la presentan las UPM con tamaños más pequeñas, los radios y los segmentos, ya que se distinguen muchos valores extremos, formando una marcada asimetría en su distribución. Mientras que las UPM más grandes, presentan una distribución de tamaños de viviendas más simétrica. La variación en el tamaño de los conglomerados tiene un efecto en los factores, como δ, necesarios para diseñar una muestra, esto parece ser raramente enfatizado en los textos de muestreo, como también el factor k, el cual se tendrá en cuenta en el cálculo de la eficiencia.

Para evaluar la eficiencia de los diseños planteados se compara la precisión de las estimaciones mediante el efecto de diseño. Bajo el supuesto que se toman *m* de las *M* UPMs, y $\bar{n}$ =10 viviendas en cada UPM seleccionada, y que las fracciones de muestreo de las UPM y los elementos dentro de las UPM son pequeñas, se calcula el efecto de diseño de la ecuación (5) y (9) respectivamente según el diseño considerado, para cada uno de los tamaños de UPM, correspondiente a cada variable:

$$Deff = k[1 + \delta(\bar{n} - 1)]$$

Según se observa en la Tabla 9 se mejora la precisión para todas las estimaciones cuando se extrae el muestreo de cualquier UPM considerando un diseño PPT en vez de un MSA. Con respecto a los conglomerados, la peor precisión la presentan los segmentos, luego los radios, mientras que los conglomerados de mayor tamaño presentan la mejor precisión, tramos o áreas, en forma indistinta.

Tabla 9: Efecto de diseño variando la selección MSA o PPT en la primera etapa de muestreo y la UPM. USM viviendas. Ciudad de Rosario. Censo 2010

Deff MSA/MSA	UPM			
	Tramos (Rx3)	Áreas(Rx2)	Radios(R)	Segmentos
PEA	2,44	2,49	3,41	6,09
Personas desocupadas	1,31	1,32	1,48	2,22
mujer	2,58	2,66	3,57	6,33
varón	2,73	2,84	3,72	6,62
mujer 14-29 años	1,47	1,51	1,78	2,78
mujer 30-64 años	1,40	1,41	1,67	2,35
varón 14-29 años	1,51	1,55	1,82	2,87
varón 30-64 años	1,46	1,48	1,73	2,49

Deff PPT/MSA	UPM			
	Tramos (Rx3)	Áreas(Rx2)	Radios(R)	Segmentos
PEA	1,89	1,97	2,20	3,44
Personas desocupadas	1,25	1,27	1,33	1,78
mujer	2,13	2,22	2,49	3,75
varón	2,37	2,47	2,77	4,07
mujer 14-29 años	1,37	1,41	1,53	2,05
mujer 30-64 años	1,26	1,35	1,33	1,67
varón 14-29 años	1,42	1,46	1,57	2,11
varón 30-64 años	1,33	1,28	1,41	1,78

Conclusiones

El IPEC tiene como objetivo la realización de una Muestra Maestra en la provincia de Santa Fe, la cual estará basada en el diseño de la MMUVRA desarrollada por el INDEC. El objetivo de contar con dicho instrumento es que permita seleccionar muestras que servirán para recolectar datos correspondientes a encuestas a hogares durante el período intercensal. El objetivo del presente trabajo es comenzar con el estudio de los distintos componentes que forman parte del diseño muestral a partir del cual se construirá la Muestra Maestra para ir determinando aquellos aspectos que brinden estimaciones más precisas. Por ello, se consideró el diseño muestral que se emplea en la localidad más grande de la provincia, Rosario, teniendo como objetivo evaluar la precisión que se obtiene en la estimación de parámetros relacionados con el ámbito laboral, que forman parte de la Encuesta Permanente de Hogares.

Se consideraron diseños con distintos métodos de selección y distintos tamaños de unidades de muestreo de primera etapa basados en las desagregaciones geográficas censales que considera el INDEC en los Censos de Población. Se evaluaron distintos aspectos como la heterogeneidad de las unidades respecto al tamaño, las componentes de variancias que surgen de la estimación de un conjunto de parámetros de interés, y los efectos de diseño que se obtienen para los distintos escenarios.

Como conclusión cabe destacar que las unidades de mayor tamaño formadas por la unión de 2 o 3 radios censales tuvieron los mejores comportamientos en términos de

precisión, mientras que las desagregaciones menores, radios y segmentos, tuvieron un pobre desempeño. Las causas pueden venir de la mano de la heterogeneidad en los tamaños de las unidades más pequeñas sumado a la distribución marcadamente asimétrica de su distribución.

Por otra parte, el diseño con considera probabilidades de inclusión distintas para las unidades de primera etapa brinda mejores resultados respecto a la precisión comparado a un diseño con igual probabilidad como el muestreo simple al azar. La justificación puede deberse a la mencionada en el párrafo anterior respecto a la heterogeneidad entre los tamaños de las unidades, medida que determina la probabilidad de inclusión de las unidades.

En estudios futuros se estudiará la extensión de esta evaluación a otras localidades de la provincia, como también considerar otros diseños para evaluar el desempeño de estos en la estimación de parámetros para toda la provincia.

Referencias Bibliográficas

Cochran, W. (1977). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

INDEC (1999). *Marco de Muestreo Nacional Urbano para Encuestas a Hogares*. Buenos Aires: Publicaciones del INDEC. Recuperado de https://biblioteca.indec.gob.ar/bases/minde/4si9_12.pdf.

King, A.J., Jesen, R.J. (1945). The Master Sample of Agriculture. *Journal of the American Statistical Association,* 40, 38-56.

Kish, L. (1978). *Muestreo de Encuestas*. México: Editorial Trillas.

Valliant, R., Dever, J.A., Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples. Second Edition.* Cham, Switzerland: Springer International Publishing AG.

Valliant R., Dever J.A., Kreuter F. (2022). PracTools: Tools for Designing and Weighting Survey Samples. R package version 1.2.6.

Fuente

INDEC. Censo Nacional de Población, Hogares y Viviendas 2010. Localidad Rosario.