Muestreo y Planificación de Encuestas 1 - 2023

Matías Bajac, Lucas Pescetto, Andrés Vidal.

Trabajo 2 - Parte 1 - Cerro Largo

Este informe corresponde a la primera parte del segundo trabajo de la UC "Muestreo y Planificación de Encuestas 1" de la Facultad de Ciencias Económicas y Administración (FCEA) de la Universidad de la República (UdelaR), dictada en el primer semestre de 2023.

Relevamiento de Padrones en Cerro Largo

La Oficina de Planeamiento y Presupuesto (OPP) busca estimar la proporción de padrones urbanos no regularizados en el departamento de Cerro Largo, con el fin de estimar el costo de la actualización del catastro nacional y poder actualizar el valor de la contribución inmobiliaria. Se define que un padrón no está regularizado cuando su superficie construída es mayor a la declarada en la oficina de catastro.

En este contexto, se requiere una propuesta de diseño muestral ajustada a las restricciones presupuestales de la OPP, enumeradas a continuación. Debido a que los datos catastrales presentes en el Catálogo de Datos Abiertos no se actualiza en tiempo real, la OPP aún debe iniciar gestiones con la Dirección General de Catastro para obtener los datos actualizados de los padrones registrados. Por lo tanto, en esta etapa se debe presentar únicamente la muestra de zonas censales a partir de la cual se seleccionarán los padrones a relevar. No obstante, sí es necesario específicar el diseño de muestreo para los padrones.

Restricciones Presupuestales

La encuesta debe adaptarse a una serie de restricciones presupuestales, a saber:

1. se relevarán hasta 1000 padrones;
2. se visitarán hasta 3 localidades con más de 5000 habitantes;
3. se visitarán hasta 4 localidades con menos de 5000 habitantes; y
4. se debe incluir la capital departamental en conjunto con su conurbano, que no contribuye a las restricciones anteriores

Para optimizar el desempeño de las estimaciones realizadas a partir de la encuesta, es ideal utilizar todo el presupuesto disponible.

Definiciones previas

En la tabla a seguir presentan definiciones generales que serán tomadas en cuenta en este reporte.

Localidad **chica**	Localidad con menos de 5000 habitantes
Localidad **grande**	Localidad con más de 5000 habitantes
Localidad con **pocas viviendas**	Localidad con menos de 150 viviendas

Localidad con **muchas viviendas**	Localidad con más de 150 viviendas
Padrón **regular**	Padrón cuya superficie construída coincide con la declarada en la oficina de catastro
Padrón **irregular**	Padrón cuya superficie construida es superior a la declarada en la oficina de catastro

Especificación de la Encuesta

En la tabla a seguir se presentan los aspectos clave de la encuesta.

Población Objetivo	Padrones urbanos del departamento de Cerro Largo, Uruguay.
Variable de Interés	Estado de regularización catastral del padrón. (`REGULAR` o `IRREGULAR`)
Parámetro a Estimar	Proporción de padrones irregulares en la población

Marco de Muestreo

A partir del marco censal se construyó un marco de muestreo de zonas censales. La información adicional considerada compreende la localidad, la cantidad de habitantes y la cantidad de viviendas de cada zona censal.

Diseño de Muestreo

Se propone realizar muestreo en tres etapas, para seleccionar jerárquicamente localidades, zonas cenales y padrones. En la tabla a seguir se definen aspectos centrales del muestreo en etapas:

Unidad Primaria de Muestreo	Localidad
Unidad Secundaria de Muestreo	Zona Censal
Elemento de Muestreo	Padrón

Esto es, en la primera etapa se obtendrá una muestra S_L de localidades:

$$S_L = \{L_1, \dots L_{n(S_L)}\}$$

En la segunda etapa, se obtendrá una muestra S_{Z_i} de zonas censales para cada localidad L_i, con $i \in \{1, \dots, n(S_L)\}$:

$$S_{Z_i} = \{Z_{i,1}, \dots, Z_{i,n(S_{Z_i})}\}$$

Finalmente, se obtendrá una muestra de padrones $S_{P_{i,j}}$ para cada zona censal $Z_{i,j}$, con $i \in \{1, \dots, n(S_Z)\}$ y $j \in \{1, \dots, n(S_{Z_i})\}$:

$$S_{P_{i,j}} = \{P_{i,j,1}, \dots, n(S_{P_{i,j}})\}$$

Recordando que la ejecución del muestreo, en esta estadio, incluye únicamente hasta la obtención de cada muestra S_{Z_i} de zonas censales por localidad. Las muestras de padrones por zona censal $S_{P_{i,j}}$ podrán ser obtenidas cuando se cuente con la información catastral actualizada sobre los padrones de Cerro Largo.

Primera Etapa: Muestreo de Localidades

Para obtener la muestra S_L de localidades se utilizó el marco de muestreo de zonas censales para construir un marco de localidades. Se agruparon las zonas censales según su localidad, y se sumaron los datos de número de habitantes y número de viviendas. Por lo tanto, se cuenta con una lista de localidades e información adicional referente al número de habitantes y el número de viviendas de cada localidad.

Se observó que las únicas localidades grandes del departamento son Melo (capital departamental) y Rio Branco. Además, Melo no posee conurbano, por lo que el restante de las localidades deben incluirse en la muestra individualmente.

Las restricciones presupuestales indican que, además de la capital departamental, podrán visitarse 3 localidades grandes y 4 localidades chicas. Debido a que Rio Branco es la única localidad grande exceptuando a Melo, con el fin de optimizar el uso del presupuesto se visitarán 5 localidades chicas. El intercambio de dos localidades grandes por una chica se debe a que éstas últimas suponen más costos al realizar la encuesta. Además, se concluyó que Rio Branco deberá ser visitada obligatoriamente. Por lo tanto, el tamaño de la muestra S_L de localidades será

$$n(S_L) = 7$$

Considerando este contexto, se realizará muestreo estratificado con el fin de controlar la cantidad de localidades de cada tipo en la muestra. Sea U_L el conjunto con todas las localidades de Cerro Largo, se definen tres estratos:

- $L^1 = \{\text{Rio Branco}\}$
- $L^2 = \{\text{Melo}\}$
- $L^3 = \{l \in U_L : \text{población}(l) < 5000\}$

En los primeros estratos, L^1 y L^2, se seleccionarán sus elementos con probabilidad 1.

En el tercer estrato L^3 se observó gran variabilidad respecto al número de viviendas. Las localidades de "Lago Merin" y "Fraile Muerto" cuentan con 1384 y 1296 viviendas, respectivamente, mientras que otras 6 localidades tienen entre 150 y 886 viviendas y las otras 18 cuentan con entre 2 y 100 viviendas.

Estas características de L^3 hacen que en el caso de usar un diseño simple sin reposición haya gran probabilidad de sortear localidades con muy pocas viviendas y que en el caso de utilizar un diseño proporcional al número de viviendas la muestra se vea demasiado sesgada a las localidades con más viviendas.

Es deseable mitigar el riesgo de seleccionar demasiadas localidades con pocas viviendas, ya que pueden ser de difícil acesso o no tener padrones suficientes para el muestreo en la tercera etapa.

Por otro lado, no se desea incorporar un sesgo fuerte hacia las localidades con más viviendas, debido a que no es evidente la proporcionalidad entre la cantidad de viviendas y la cantidad de padrones irregulares.

Entonces, se optó por realizar nuevamente un muestreo estratificado según la cantidad de viviendas:

- $L^3_1 = \{l \in L^3 : \text{viviendas}(l) < 150\}$
- $L^3_2 = \{l \in L^3 : \text{viviendas}(l) \geq 150\}$

En cada estrato se realizará muestreo simple sin reposición, con tamaños de muestra 2 y 3 para L^3_1 y L^3_2. De esta forma, se introduce un sesgo controlado hacia las poblaciones con localidades con más viviendas garantizando que también se incluirán localidades del otro grupo en la muestra.

Segunda Etapa: Muestreo de Localidades

Para obtener las muestra S_{Z_i} de zonas censales para cada localidad L_i perteneciente a la muestra obtenida en la etapa anterior, se utiliza el marco de muestro de zonas censales construido inicialmente.

Dentro de cada localidad, se utilizará un diseño de muestreo simple sin reposición, cuyo tamaño se calcula en base a la cantidad de viviendas de cada estrato, y a la cantidad de viviendas de las localidades de la muestra, de la siguiente forma:

$$n(S_{Z_i}) = \frac{\rho_i}{r_i}$$

dónde r_i es el número de padrones que se relevarán por zona censal en la localidad L_i y ρ_i es la cantidad de padrones que se relevarán en la localidad L_i:

$$\rho_i = 1000\alpha_k^j \frac{\text{viviendas}(L_i)}{\sum_{S_k^j} \text{viviendas}(L_k)}$$

donde α_k^j es la proporción de viviendas del estrato j y el subestrato k (esto último solo aplica para el estrato 3) sobre el total de viviendas de la población. S_k^j es el conjunto de las localidades del estrato j y el subestrato k que salieron en la muestra.

Calculamos los ρ_i de esta forma para que la cantidad de padrones seleccionados en cada estrato sea aproximadamente proporcional a la cantidad total de padrones del mismo. Esto permite atenuar el impacto que tiene sobre las localidades chicas el hecho de haber seleccionado en ese subestrato una proporción menor de localidades que en los otros.

Luego, para distribuir los padrones de cada estrato entre las localidades que salieron en la muestra, se realiza de forma proporcional a la cantidad de viviendas de cada una de esas localidades (en los casos de Melo y Rio Branco, todos los padrones seleccionados en sus respectivos estratos quedan en esas localidades ya que son únicas en sus estratos).

Tercera Etapa: Muestreo de Padrones

Para obtener las muestras $S_{P_{i,j}}$ de padrones para la zona censal Z_j y la localidad L_i, se propone utilizar muestreo simple sin reemplazo, con tamaño de muestra r.

Como el parámetro a estimar es una proporción y el diseño es simple sin reemplazo, el criterio para escoger el tamaño de muestra es:

$$r_i = \frac{N_i(z_{1-\frac{\alpha}{2}})^2 S_y^2}{N_i\varepsilon^2 + (z_{1-\frac{\alpha}{2}})^2 S_y^2}$$

dónde N_i es la cantidad de padrones en cada zona censal de la localidad, $z_{1-\frac{\alpha}{2}}$ es el percentil $1-\frac{\alpha}{2}$ de una Normal estándar, $1-\alpha$ es el nivel de confianza, ε es el error de muestreo y S_y^2 es la varianza de la variable de interés. Como la variable de interés es una Bernoulli se sabe que

$$S_y^2 \leq \frac{1}{4}$$

por lo que se usa este valor para estimar el tamaño de muestra. Se decidió también utilizar la información del número de viviendas y del número de zonas censales de cada localidad para estimar el número de padrones por zona censal en la localidad L_i:

$$N_i = \left\lceil \frac{\text{viviendas}(L_i)}{\text{zonas\_censales}(L_i)} \right\rceil$$

El valor de r queda entonces determinado por el error de muestreo objetivo ε y el nivel de confianza $1-\alpha$:

$$r_i(\varepsilon, \alpha) = \frac{N_i \frac{1}{4}(z_{1-\frac{\alpha}{2}})^2}{N_i \varepsilon^2 + \frac{1}{4}(z_{1-\frac{\alpha}{2}})^2}$$

Resúmen del Diseño de Muestreo

En síntesis, se realizará muestreo en 3 etapas:

1. Localidades con muestreo estratificado
2. Zonas censales con muestreo simple sin reposición
3. Padrones con muestreo simple sin reposición

Las localidades se seleccionarán usando muestreo estratificado en tres estratos:

1. Melo por inclusión forzosa (probabilidad 1)
2. Rio Branco por inclusión forzosa (probabilidad 1)
3. Localidades chicas por muestreo estratificado con tamaño de muestra global igual a 5

Las localidades chicas también se seleccionarán usando muestreo estratificado

1. Localidades con pocas viviendas por muestreo simple sin reposición con tamaño de muestra igual a 2
2. Localidades con muchas viviendas por muestreo simple sin reposición con tamaño de muestra igual a 3

De esta forma, se pretende introducir un sesgo controlado en la muestra hacia las localidades con muchas viviendas, garantizando que se incluirán localidades con pocas viviendas. Esto se debe a las complicaciones operativas de visitar localidades muy chicas y a la falta de evidencia que soporte la proporcionalidad entre la cantidad de viviendas y la cantidad de padrones irregulares en la localidad.

Las zonas censales se seleccionarán con muestreo simple sin reposición de tamaño $n(S_{Z_i})$ y los padrones se seleccionarán con muestreo simple sin reposición de tamaño $n(S_{P_{i,j}})$, dónde:

$$n(S_{Z_i}) = \frac{\rho_i}{r_i}$$

$$\rho_i = 1000 \frac{\text{viviendas}(L_i)}{\sum_{k=1}^{n(S_L)} \text{viviendas}(L_k)}$$

$$n(S_{P_{i,j}}) = r_i = \frac{N_i \frac{1}{4}(z_{1-\frac{\alpha}{2}})^2}{N_i \varepsilon^2 + \frac{1}{4}(z_{1-\frac{\alpha}{2}})^2}$$

$$N_i = \left\lceil \frac{\text{viviendas}(L_i)}{\text{zonas\_censales}(L_i)} \right\rceil$$

Los tamaños de muestra $n(S_{Z_i})$ y $n(S_{P_{i,j}})$ quedan determinados por el nivel de confianza $1-\alpha$ y el error de muestreo ε en la etapa de muestreo de padrones, ambos valores predefinidos.