Borra, Virginia Laura
Pagura, José Alberto
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

ESTIMACIÓN DEL TOTAL DE HOGARES CON NECESIDADES BÁSICAS INSATISFECHAS EN LA CIUDAD DE ROSARIO UTILIZANDO MODELO DE SEMIVARIOGRAMA

1-INTRODUCCIÓN

El enfoque basado en la predicción o en modelos ha resultado de suma utilidad en el muestreo en poblaciones finitas, proporcionando un soporte metodológico que permite la incorporación de información auxiliar con el fin de mejorar la precisión de las estimaciones.

Bajo este enfoque, la población finita se considera una muestra de una población infinita o superpoblación que sigue un determinado modelo estadístico. Las inferencias en la población finita bajo este enfoque se realizan estimando los parámetros del modelo, con la información proporcionada por la muestra, para luego, empleando dicho modelo estimado, obtener una predicción del valor poblacional de interés en la población finita, considerado una cantidad aleatoria. Este enfoque tiene como inconveniente, frente al tradicional enfoque de diseño, la dependencia de la correcta especificación del modelo superpoblacional, pero, a la vez, es muy útil a la hora de incorporar información auxiliar. La mayor utilidad de esta aproximación es el conjunto de posibilidades que brinda en la incorporación de información auxiliar permitiendo en algunos casos encontrar un predictor lineal insesgado y óptimo (PLIO), para luego utilizarlo con el enfoque de diseño o bajo el enfoque de modelos y estimar sus errores con las herramientas brindadas por dicha aproximación.

Cuando las unidades de una población a muestrear se encuentran ubicadas en el espacio, podría suceder que existiese un agrupamiento de las mismas en cuanto a los valores de una variable de interés, hecho que puede ser aprovechado con el uso de las herramientas dadas por la aproximación basada en la predicción.

En este trabajo se presenta una aplicación de dicho enfoque teniendo en cuenta la variabilidad espacial reflejada en un modelo de semivariograma con la finalidad de mostrar la mejora en la precisión de las estimaciones del total de hogares con necesidades básicas insatisfechas (NBI) en Rosario.

El estudio tiene como objetivo mostrar la factibilidad y buenos resultados que puede brindar esta metodología y se realiza con una muestra de radios de una población finita conocida como lo es la de número de hogares con NBI en cada radio censal de la ciudad de Rosario según el Censo Nacional de Población, Hogares y Viviendas de 2001.

En la sección 2, se presentan los fundamentos de la metodología propuesta. Luego, en el siguiente apartado se muestra la existencia de variabilidad espacial a través de un breve análisis exploratorio utilizando los datos de toda la población.

Los resultados empleados para la comparación de los métodos de estimación provienen de una muestra aleatoria simple. La sección 4 contiene la estimación del semivariograma utilizado para modelar la variabilidad espacial, a partir de la muestra seleccionada. Por último, se compara la precisión estimada con las obtenidas usando otros modelos correspondientes a algunos planes de muestreo usuales.

La sección 6 se ha destinado a la presentación de los comentarios finales.

2-APROXIMACIÓN BASADA EN LA PREDICCIÓN

El enfoque que tradicionalmente se emplea en el muestreo de poblaciones finitas se conoce como basado en el diseño y consiste en definir un método probabilístico de selección de la muestra y un procedimiento para estimar un valor poblacional. El análisis del comportamiento del estimador se realiza en base a la distribución del mismo obtenida a través de todas las muestras posibles y tiene en cuenta muy pocos supuestos, lo que lo convierte en un procedimiento muy sólido y útil para llevarlo a la práctica.

Por otra parte, el desarrollo del enfoque de modelos ha contribuido a la teoría del muestro en poblaciones finitas de varias formas. Entre ellas pueden distinguirse la incorporación de información auxiliar integrándola en un modelo estadístico, la posibilidad de reflejar en el modelo los diferentes métodos de selección, y la estimación en pequeñas áreas. Bajo este enfoque, la población finita es una muestra de una población infinita llamada también superpoblación. La muestra es a su vez una submuestra de la población finita

Las inferencias se basan en el planteamiento de un modelo superpoblacional que tenga en cuenta o no la autocorrelación de las unidades para luego estimar sus parámetros con los datos de la muestra y obtener las predicciones de los valores poblacionales de interés. Las propiedades de los predictores se estudian considerando el modelo postulado.

Para formalizar el enfoque de modelos, a continuación se presentan las siguientes definiciones:

Sean

- $\mathrm{Y_n}$: vector de datos observados en la muestra
- $\mathrm{Y_{N-n}}$: vector de datos de la población para unidades no incluidas en la muestra
- $\mathrm{X_{N,p}}$: matriz de datos para la población de p-1 variables auxiliares

El Predictor Lineal Insesgado y Óptimo (PLIO) del total es:

$$\hat{\mathrm{Y}} = 1_{\mathrm{n}}' \mathrm{Y_n} + 1_{\mathrm{N-n}}' \hat{\mathrm{Y}}_{\mathrm{N-n}}$$

donde $\hat{\mathrm{Y}}_{\mathrm{N-n}}$ se predice por medio del modelo lineal $\mathrm{y_i} = \beta_0 + \beta_1 \mathrm{x_{1,i}} ... \beta_{\mathrm{p-1}} \mathrm{x_{p-1,i}} + \varepsilon_{\mathrm{i}}$

$$\mathrm{E}(\varepsilon_{\mathrm{i}}) = 0 \qquad \mathrm{Cov}(\varepsilon_{\mathrm{i}}; \varepsilon_{\mathrm{i`}}) = \begin{cases} \sigma_\varepsilon^2 & \text{si } \mathrm{i} = \mathrm{i}' \\ \mathrm{C}(\mathrm{dist}(\mathrm{s_i}; \mathrm{s_{i`}})) = \mathrm{C_{i;i`}} & \text{si } \mathrm{i} \neq \mathrm{i}' \end{cases}.$$

El estimador lineal de los parámetros del modelo $\beta' = (\beta_0; \beta_1; ...; \beta_{\mathrm{p-1}})$ es:

$$\hat{\beta} = (\mathrm{X}_{\mathrm{n,p}}' \mathrm{V}_{\mathrm{n,n}}^{-1} \mathrm{X}_{\mathrm{n,p}})^{-1} (\mathrm{X}_{\mathrm{n,p}}' \mathrm{V}_{\mathrm{n,n}}^{-1} \mathrm{Y_n})$$

$$\mathrm{Var}(\hat{\beta}) = (\mathrm{X}_{\mathrm{n,p}}' \mathrm{V}_{\mathrm{n,n}}^{-1} \mathrm{X}_{\mathrm{n,p}})^{-1}$$

donde $\mathrm{V_{n,n}}$ es la matriz de covariancias para las n unidades en la muestra.

El Predictor lineal insesgado y óptimo (PLIO) es:

$$\hat{Y}_{N-n} = X_{N-n,p}\hat{\beta} + V_{N-n,n}V_{n,n}^{-1}(Y_n - X_{n,p}\hat{\beta})$$

Por lo tanto el PLIO de $\hat{Y}$ es $\hat{Y} = 1_n' Y_n + 1_{N-n}'\left[X_{N-n,p}\hat{\beta} + V_{N-n,n}V_{n,n}^{-1}(Y_n - X_{n,p}\hat{\beta})\right]$.

El error cuadrático medio de la predicción es:

$$ECM(\hat{Y}) = E\left(\hat{Y} - Y\right)^2 = 1_{N-n}'\left[\left(X_{N-n,p} - \Omega_{N-n,p}\right)\Omega_{p,p}^{-1}\left(X_{N-n,p} - \Omega_{N-n,p}\right)' + \left(V_{N-n,N-n} - W_{N-n,N-n}\right)\right]1_{N-n}$$

donde $V_{N-n,n}$ es la matriz de covariancias entre los "n" elementos incluidos en la muestra y los "N-n" elementos no incluidos en la muestra,

$\Omega_{N-n,p} = V_{N-n,n}V_{n,n}^{-1}X_{n,p}$,

$\Omega_{p,p} = X'_{n,p}V_{n,n}^{-1}X_{n,p}$,

$W_{N-n,N-n} = V_{N-n,n}V_{n,n}^{-1}V'_{N-n,n}$.

A continuación se presentan algunos casos particulares en los que se trabaja con una única variable aleatoria.

a) Modelo sin variable auxiliar: Homocedástico y sin autocorrelación

Modelo: $y_i = \beta_0 + \varepsilon_i$

$$E(\varepsilon_i) = 0 \qquad Cov(\varepsilon_i;\varepsilon_{i`}) = \begin{Bmatrix} \sigma_\varepsilon^2 & \text{si } i = i' \\ 0 & \text{si } i \neq i' \end{Bmatrix}$$

b) Modelo de regresión: Homocedástico y sin autocorrelación

Modelo: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$E(\varepsilon_i) = 0 \qquad Cov(\varepsilon_i;\varepsilon_{i`}) = \begin{Bmatrix} \sigma_\varepsilon^2 & \text{si } i = i' \\ 0 & \text{si } i \neq i' \end{Bmatrix}$$

c) Modelo de regresión: Heterocedástico y sin autocorrelación

Modelo: $y_i = \beta_1 x_i + \varepsilon_i$

$$E(\varepsilon_i) = 0 \qquad Cov(\varepsilon_i;\varepsilon_{i`}) = \begin{Bmatrix} \sigma_{\varepsilon i}^2 & \text{si } i = i' \\ 0 & \text{si } i \neq i' \end{Bmatrix}$$

Suponiendo que la matriz de variancias y covariancias es de la forma $V_{n,n} = \sigma_\varepsilon^2 \Psi_{n,n}$, donde $\Psi_{n,n}$ es una matriz definida positiva y conocida.

d) Modelo de regresión con autocorrelación: Variograma y correlograma

Modelo: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$$E(\varepsilon_i) = 0 \qquad Cov(\varepsilon_i ; \varepsilon_{i`}) = \begin{cases} \sigma_\varepsilon^2 & \text{si } i = i' \\ \sigma_\varepsilon^2 - \gamma_{i,i'} & \text{si } i \neq i' \end{cases}$$

donde $\gamma_{i,i'}$ es el valor del semivariograma para las unidades i e i`.

Las estimaciones de σ_ε^2 y de $\gamma_{i,i'}$ se obtienen a partir del modelo de semivariograma ajustado a los datos.

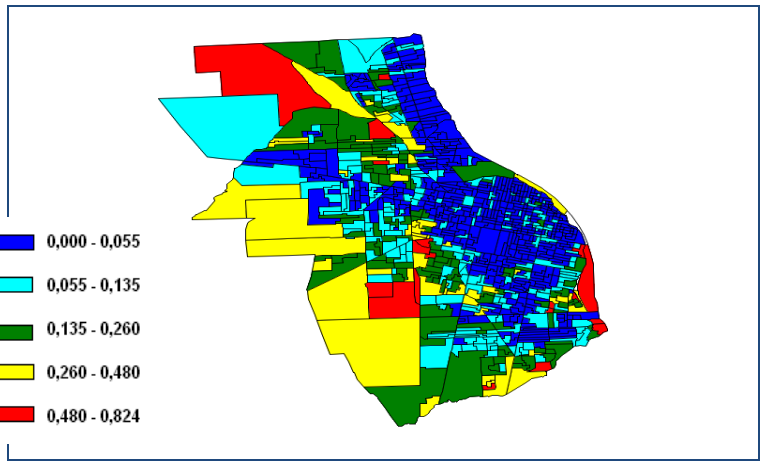
3-ANÁLISIS EXPLORATORIO

A continuación, se presenta un mapa de coropletas construido a partir de los datos poblacionales, que permite apreciar la distribución espacial de la variable proporción de hogares con NBI de acuerdo a radios censales en la ciudad de Rosario.

En el Gráfico 1 se observa que las proporciones más bajas de hogares con NBI se encuentran en la zona centro de la ciudad mientras que en la periferia se detectan los valores más altos de hogares con dicha característica.

El índice de Moran global resulta igual a 0,47 (p<0,001), mostrando la existencia de autocorrelación espacial moderada y positiva

Gráfico 1: Proporción de hogares con NBI según radio censal



Por lo tanto se observar la existencia de correlación espacial, conclusión que fundamenta la búsqueda de un modelo para su explicación. Para ello, se recurre a los modelos de semivariograma.

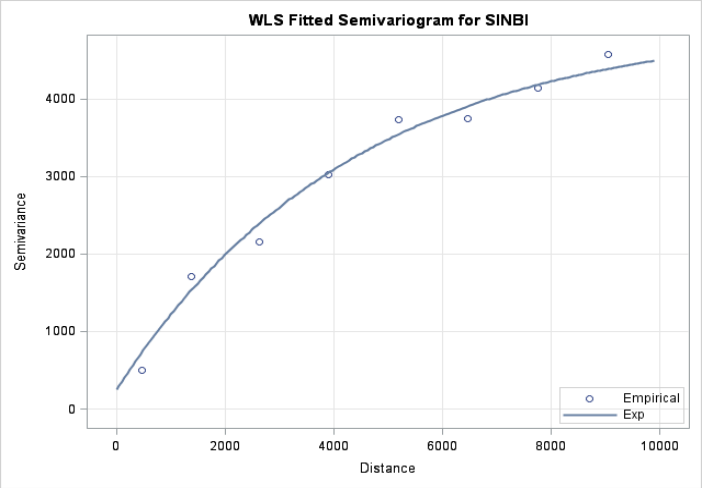
4-MODELO DE SEMIVARIOGRAMAESTIMADO PARA LA MUESTRA

Para la estimación del semivariograma se consideran dos aproximaciones: (i) la aproximación de un modelo teórico de variograma y estimación de los parámetros del modelo por máxima varosimilitud restringida y (ii) el cálculo del variograma empírico y ajuste por mínimos cuadrados ponderados de un modelo teórico de variograma al empírico, que fue el procedimiento utilizado recurriendo a SAS 9.3.

Para el proceso estacionario del 2º orden, el modelo elegido es el exponencial y los parámetros estimados son iguales a c_n=254,89, c_0=4726,59, a_0=4367,17, resultando:

$$\hat{\gamma}_{\exp}(h) = 254{,}89 + 4726{,}59\left[1 - \exp\left(\frac{-h}{4367{,}17}\right)\right] \quad h > 0$$

Gráfico 2: Semivariograma estimado para el número de hogares con NBI



5-ANÁLSIS COMPARATIVO

Con la finalidad de comparar los resultados que se obtienen empleando un modelo que tiene en cuenta la variabilidad espacial, con otros que pueden identificarse con estimadores empleados habitualmente, se calculan las predicciones del total y de los errores cuadráticos medios del predictor en una muestra de 148 radios censales. Se obtienen además, las estimaciones de las eficiencias relativas de los métodos, tomando como referencia el modelo más simple. Los resultados se presentan en la Tabla 1 y los modelos considerados fueron:

a) modelo sin variable auxiliar homocedástico y sin autocorrelación que se identifica con el estimador de simple expansión,
b) modelo de regresión homocedástico y sin autocorrelación identificado con el estimador de regresión
c) modelo de regresión heterocedástico, sin autocorrelación y sin ordenada al origen que se identifica con el estimador de razón
d) modelo de regresión homocedástico y con autocorrelación espacial.

Cabe mencionar que el total de hogares con NBI en la ciudad, según el Censo considerado es 29622 y que la variable auxiliar a considerar es el total de hogares por radio censal.

Tabla1: Estimaciones de total de NBI en la ciudad de Rosario, del Error cuadrático medio y de la eficiencia, para cada modelo empleado.

Método	$\hat{Y}$	$\sqrt{\hat{ECM}(\hat{Y})}$	$\hat{ER}$
Modelo sin variable auxiliar $y_i = \beta_0 + \varepsilon_i$ Homocedástico y sin autocorrelación	29.883	3.897	1,0
Modelo de regresión $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ Homocedástico y sin autocorrelación	30.122	3.118	1,6
Modelo de regresión $y_i = \beta_1 x_i + \varepsilon_i$ Heterocedástico y sin autocorrelación	29.980	2.550	2,3
Modelo de regresión $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ Homocedástico y con autocorrelación	29.483	1.277	9,3

6-CONSIDERACIONES FINALES

De acuerdo a los resultados obtenidos con una sola muestra y presentados en la Tabla 1, se aprecia que:

- El uso de la información brindada por variables auxiliares correlacionadas con la variable en estudio ha mostrado una mejora importante en la precisión de las estimaciones.
- La incorporación del semivariograma en el modelo superpoblacional provocó una importante reducción en el error cuadrático medio estimado presentando para este caso particular una eficiencia estimada de aproximadamente 9 con respecto al modelo más sencillo. Esto era esperable debido a la existencia de autocorrelación espacial.

Por otra parte, dado que los resultados obtenidos para las comparaciones corresponden solo a una muestra, convendrá llevar a cabo un estudio comparativo en el que se evalúe la calidad de los estimadores propuestos por los modelos, pero teniendo en cuenta la distribución obtenida con las muestras posibles de la población finita. Esto es factible debido a que se dispone de los datos de toda la población, pero seguramente requerirá una importante cantidad de tiempo de computación para la obtención de resultados.

7-REFERENCIAS BIBLIOGRÁFICAS

- Ambrosio Flores, L. (1999). Muestreo. Monografías de Escuela Técnica Superior de Ingenieros Agrónomo, 156, Universidad Politécnica de Madrid. España.

- Ambrosio Flores, L. (2000). Estadística Espacial. Monografías de Escuela Técnica Superior de Ingenieros Agrónomo, 157, Universidad Politécnica de Madrid. España.

- Ambrosio Flores, L.; Marín, C.; Iglesias, L.; Pascual, V.; Fuertes, A.; Mena, M.A. (2009). Agricultural and environmental information systems: the integrating role of área samples. Spanish Journal of Agricultural Research, pp. 957-973.

- Cochran, W. G. (1998). Técnicas de Muestreo. Wiley. México.

- Cressie, N. A. C. (1993). Statistics for Spatial Data. Wiley. New York.

- Iglesias Martínez, L. (2000). Tesis Doctoral: Muestreo de áreas: Diseño de muestras y estimación en pequeñas áreas. Escuela Técnica Superior de Ingenieros Agrónomos. Universidad Politécnica de Madrid. España.

- Wang, J-F.; Stein, A.; Gao, B-B; Ge, Y. (2012). A review of Spatial Sampling. Spatial Statistics