Publicación Electrónica **ISSN 2718- 6636**

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARIA DE CIENCIA Y TECNOLOGIA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales
"Investigaciones en la Facultad"
Ciencias Económicas y Estadística

Pagura, José Alberto
Borra, Virginia Laura
Marfetán Molina, Diego
Mignoni, César|
Riaño, María Eugenia
López, Elisabet
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.

PLANES DE MUESTREO PARA DATOS ESPACIALES: SU COMPORTAMIENTO EN UN ESTUDIO SOCIOECONÓMICO[1]

Resumen

En muchos estudios por muestreo, las unidades que constituyen la población se encuentran situadas en el espacio. Es frecuente observar en esta clase de poblaciones una característica conocida como *correlación espacial positiva*: los valores de la variable de interés son muy parecidos en unidades cercanas entre sí y menos parecidos a medida que la distancia entre ellas es mayor. Este comportamiento se puede representar mediante los modelos estadísticos conocidos como correlograma y semivariograma. La información obtenida de los mismos podrá resultar de utilidad al momento de diseñar la muestra, en cuanto a mejorar la precisión de las estimaciones. Esta información se puede aprovechar utilizándola en el proceso de estimación recurriendo al enfoque de predicción, el cual permite la incorporación del modelo de semivariograma para lograr las mejoras buscadas o, en el procedimiento de selección de la muestra, tópico en el cual se centra el presente trabajo.

Existen varias propuestas para la selección de la muestra atendiendo a la correlación espacial presente. En este trabajo, se han considerado algunas de ellas, que tienen su origen en estudios sobre recursos naturales o problemas ambientales. Se plantea su aplicación en un estudio socioeconómico como lo es la estimación del número de hogares con NBI en la ciudad de Rosario a partir de una muestra de radios censales y se evalúa la eficiencia de los mismos.

Se ha encontrado que varios de los métodos analizados proporcionan una mayor eficiencia que el muestreo aleatorio simple alentando la ampliación de los escenarios para los estudios de eficiencia, así como su aplicación en encuestas de este campo.

Palabras clave: Muestreo de datos espaciales. Estimador de Horvitz-Thompson. Encuestas sociales.

Abstract

In many sampling applications the units composing the population are spatially distributed. When working with this type of populations it is common to observe a characteristic known as *positive spatial correlation*: values of the response variable are similar for units close to each other, and more diverse as the distance between units becomes larger. This situation is taken into account by two models: correlogram and semivariogram. Information obtained from these

[1] Trabajo elaborado en el marco del Proyecto 80020180300081UR, titulado: “Planes de Muestreo para Poblaciones con Variabilidad Espacial. Propuestas y Estudio de Propiedades”, dirigido por José Pagura.

two models is of great value when designing a sample, as it improves the precision of the estimates. This information is capitalized by using it during the estimation process, choosing a prediction approach that allows the inclusion of the semivariogram model in order to achieve the desired improvement, or else during the sample selection process, which is this articles' topic of interest.

There are several proposals pertaining the selection of a sample while taking into account the spatial correlation present in the data. Some of the techniques considered in this study were developed from studies regarding natural or environmental resources. We postulate their application in a socioeconomic study concerning the number of households with at least one Unsatisfied Basic Needs (NBI, Necesidades Básicas Insatisfechas) in the city of Rosario via a sample of census units, and their efficiency is assessed.

We find that many of the analyzed methods provide better efficiency than simple random sampling, motivating their application in this type of studies and also broadening the scenarios for efficiency studies.

Introducción

En muchos estudios por muestreo las unidades se encuentran ubicadas en el espacio; por ejemplo, encuestas agropecuarias, socioeconómicas, estudios ambientales, estudios sobre recursos naturales. Estas poblaciones presentan con frecuencia el fenómeno conocido como correlación espacial positiva: los valores de las variables que se estudian serán más parecidos en unidades cercanas y ese parecido es menor a mayor distancia entre las mismas.

Este fenómeno se puede modelar mediante un semivariograma o un correlograma, los cuales permiten detectar y cuantificar dichos efectos. La información dada por estos modelos puede emplearse en la fase de estimación mediante el empleo del enfoque conocido como "de predicción".

Por otra parte, admitiendo la existencia de correlación espacial positiva, incluir en la muestra unidades vecinas cercanas es incurrir en redundancias, lo que se traduce en perdida de precisión de las estimaciones; en consecuencia, parece conveniente que la muestra esté compuesta por unidades que se encuentren separadas y repartidas por toda la región en la que se encuentra la población en estudio. En general, los planes de muestreo tradicionales basados en el diseño no tienen en cuenta esta necesidad, a excepción del muestreo sistemático. Esto lleva a plantear la necesidad de considerar otros procedimientos que respondan a la definición de muestreo probabilístico y que a la vez tengan en cuenta esta característica de la población, procurando brindar así estimadores más eficientes.

En este trabajo se analizan algunas de las alternativas existentes originadas en estudios ambientales y de estimación de recursos naturales para su aplicación en estudios socioeconómicos.

Planes de muestreo espacial

Un plan de muestreo queda especificado cuando se define un método de estimación y un método de selección. En el presente estudio se considerará selección sin reposición y con probabilidades desiguales, empleando el estimador de Horvitz-Thompson.

El estimador del total es $\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ y su variancia resulta igual a $V(\hat{Y}_{HT}) = \sum_{i<j}^{N} \sum_{j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$. Si las probabilidades de inclusión de segundo orden π_{ij} son proporcionales a la

distancia que separa a las unidades i y j se obtendrá un plan más eficiente. Sobre esta base, se construyen los métodos de selección que se han considerado en el presente trabajo.

Atendiendo a los requerimientos enunciados, se presenta una descripción sintética de algunas propuestas que dan satisfacción a los mismos.

Diseño estratificado en teselas aleatorizadas (GRTS)

Fue desarrollado por Stevens & Olsen (1999, 2004) para tratar el problema de muestreo en el contexto de estudio de recursos naturales. Divide en forma recursiva el territorio en "teselas" y traslada las unidades del espacio de dos dimensiones a una dimensión, manteniendo la estructura espacial, para luego extraer una muestra sistemática en una dimensión. Para su implementación se puede utilizar el paquete spsurvey (Kincaid et al., 2019) de R.

Se ha estudiado que este método resulta de suma utilidad en los estudios ambientales o de recursos naturales en los que las unidades de muestreo son puntos en la región, pero no hay suficiente evidencia sobre la ganancia de eficiencia en poblaciones finitas.

Técnica de selección secuencial dependiente de las unidades de área (DUST)

Esta técnica fue propuesta por Arbia (1993). Consiste en seleccionar una unidad k, para luego actualizar las probabilidades de selección mediante una fórmula que tiene en cuenta las distancias entre la unidad k y las restantes unidades hasta completar el tamaño de la muestra establecido.

Su implementación se realiza utilizando un programa R presentado en Benedetti et al. (2015). Se muestran resultados con dos fórmulas diferentes de actualización de las probabilidades de selección (DUST1 y DUST2).

Muestreo de Poisson espacialmente correlacionado (SCPS)

Grafström (2012) propuso este método secuencial. La muestra se selecciona ordenando el listado en forma aleatoria y decidiendo para cada unidad de la población, si ingresa o no a la muestra a partir de las probabilidades de inclusión, las que van cambiando a medida que se incluye una nueva unidad en la muestra.

Este método puede implementarse mediante el uso de la función scps del paquete BalancedSampling (Grafström & Lisic, 2019) de R.

Método pivotal local (LPM)

Grafstrom et al. (2012) presentaron este método para seleccionar muestras con probabilidades de inclusión π_k fijas y correlacionadas. Se basa en un criterio de actualización para las probabilidades π_k y π_l. En cada paso, las reglas establecen que la suma de las probabilidades actualizadas es tan localmente constante como sea posible, y que difieren entre sí en la forma en que se eligen las dos unidades cercanas k y *l*.

Existen 2 alternativas: LPM 1, que produce resultados mejor equilibrados espacialmente y LPM 2 que es más simple y rápido. Las funciones LPM1 y LPM2 del paquete BalancedSampling de R permiten el empleo de este método.

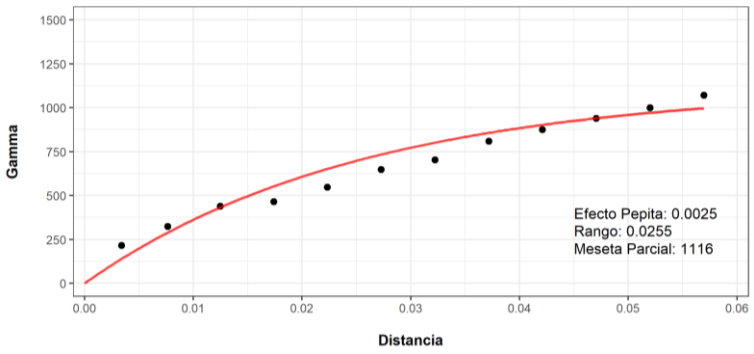
Comportamiento de los planes de muestreo considerados

Benedetti et. al. (2015) utilizó una población de datos simulados con atributos específicos para examinar el comportamiento de las propuestas antes descriptas. En dicha población planteó 9 escenarios distintos y en cada uno de ellos tomó 10000 muestras de tamaño 10, 50 y 100. Como resultado observó que los métodos DUST con parámetros de ajuste 1 y 2 presentaron los peores resultados en la mayoría los casos, y los métodos SCPS, LPM1 y LPM2, en general, tuvieron buenos resultados.

El interés de este trabajo fue evaluar el comportamiento de los planes considerados en estudios socioeconómicos eligiendo como problema, la estimación del total de hogares con Necesidades Básicas Insatisfechas (NBI) en la ciudad de Rosario a partir de una muestra de radios censales.

Para verificar la existencia de *correlación espacial positiva* en la población estudiada se construyó el semivariograma muestral y se ajustó un modelo exponencial.

Gráfico 1: Semivariograma muestral y ajuste modelo exponencial

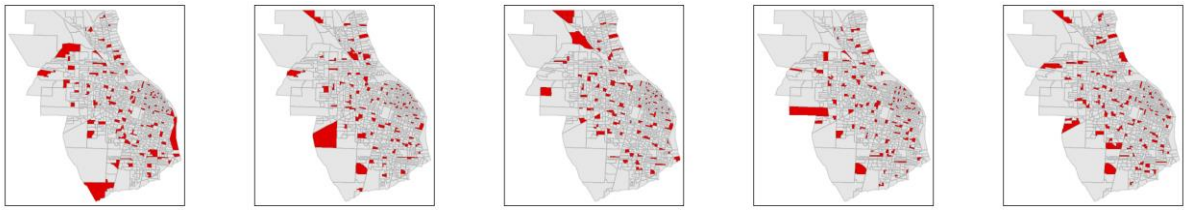


Con la finalidad de evaluar la eficiencia de los métodos, se obtuvieron 10000 muestras de tamaño 10, 50 y 100 para cada uno de ellos. Se estimó en cada caso y con los resultados arrojados por las muestras, el ECM del estimador para luego obtener el cociente de ellos con respecto al correspondiente al muestreo aleatorio simple (Tabla 1).

Para ilustrar la cobertura del territorio que se logra con las muestras que se obtienen con cada método, se grafica el mapa de Rosario con las unidades seleccionadas en 5 muestras.

Gráfico 2: Cobertura espacial para muestras generadas por los seis métodos considerados

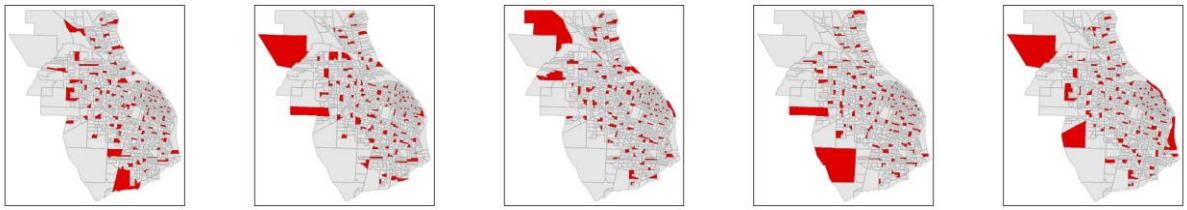
MÉTODO GRTS



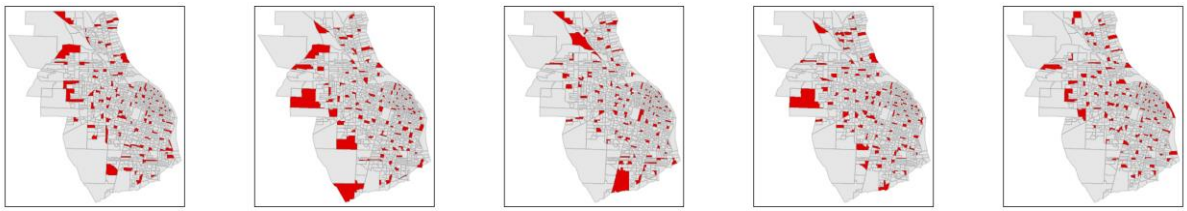
MÉTODO DUST 1

MÉTODO DUST 2

MÉTODO SCPS



MÉTODO LPM 1



MÉTODO LPM 2

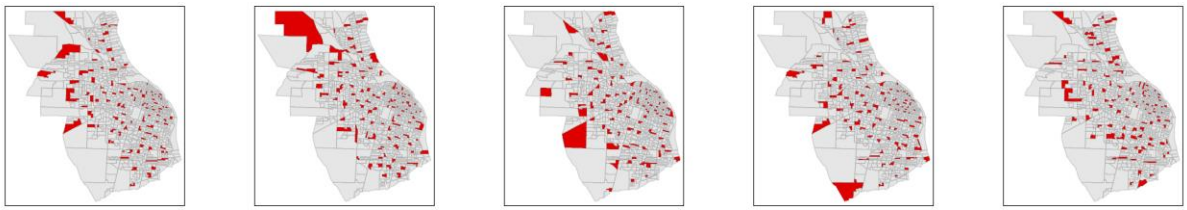
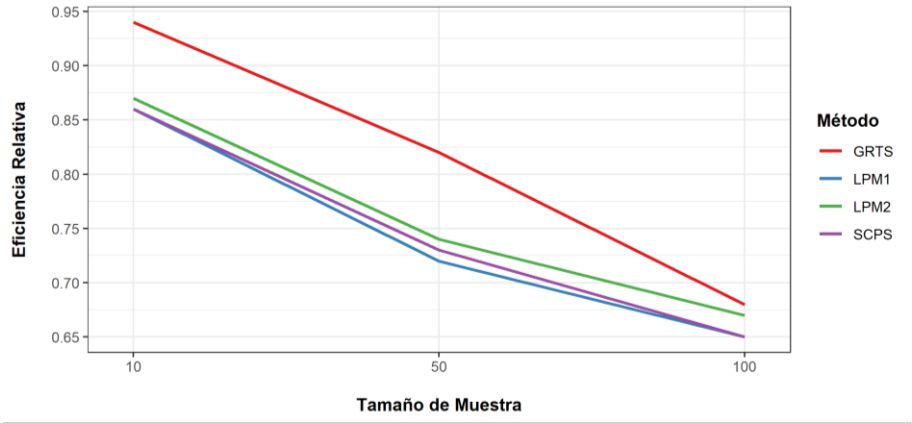


Tabla 1: Eficiencia relativa con respecto al muestreo aleatorio simple

MÉTODO	N = 10	N = 50	N = 100
GRTS	0,94	0,82	0,68
SCPS	0,86	0,73	0,65
DUST 1	1,47	13,20	27,72
DUST 2	1,49	12,98	27,44
LPM 1	0,86	0,72	0,65
LPM 2	0,87	0,74	0,67

Gráfico 3: Eficiencia relativa con respecto al muestreo aleatorio simple



Consideraciones Finales

Los resultados obtenidos son coherentes con los estudios por simulaciones encontrados en Benedetti (2015). El método DUST es el de peor desempeño, siendo menos eficiente que el muestreo aleatorio simple para los tres tamaños de muestra. Los restantes métodos fueron más eficientes que el muestreo aleatorio simple y con eficiencias parecidas entre ellos.

Futuros avances

Los resultados obtenidos impulsan la profundización del estudio de los métodos de muestreo espacial para su aplicación en estudios económicos y sociales. Se planea la inclusión de otros métodos disponibles en la bibliografía y la realización de estudios comparativos en diferentes escenarios con las características de las poblaciones habituales en las encuestas socioeconómicas.

Referencias Bibliográficas

- Arbia G. (1993) The use of GIS in spatial statistical surveys. Int Stat Rev 61:339–359.
- Benedetti R.; Piersimoni F.; Postiglioni P. (2015) Sampling Spatial Units for Agricultural Surveys. Springer.
- Bondesson L., Grafström A. (2011) An extension of Sampford's method for unequal probability sampling. Scand J Stat 38:377–392

- Grafström A., Lundström N., Schelin L. (2012) Spatially balanced sampling through the pivotal method. Biometrics 68:514–520

- Grafström A., Lisic J. (2019). BalancedSampling: Balanced and Spatially Balanced Sampling. R package version 1.5.5. https://CRAN.R-project.org/package=BalancedSampling

- Kincaid, T. M., Olsen, A. R., and Weber, M. H. (2019). spsurvey: Spatial Survey Design and Analysis. R package version 4.1.0.

- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

- Stevens D.L., Olsen A.R. (1999) Spatially Restricted Surveys over Time for Aquatic Resources. Journal of Agricultural, Biological and Environmental Statistics, 4: 4, 415-428.

- Stevens D.L.; Olsen A.R. (2004) Spatially Balanced Sampling of Natural Resources. JASA, 99: 465, 262-278.

- Thompson S. (2012) Sampling. Wiley.