Publicación Electrónica ISSN 2718-6636

UNIVERSIDAD NACIONAL DE ROSARIO
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA
SECRETARÍA DE CIENCIA Y TECNOLOGÍA E INSTITUTOS DE INVESTIGACIONES

Resumen Ampliado

Jornadas Anuales

“Investigaciones en la Facultad” de Ciencias Económicas y Estadística

Virginia Laura Borra
Valentina Celeste Castellanos
José Alberto Pagura
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística.

PROPUESTA DE PLANES DE MUESTREO PARA ESTUDIOS EPIDEMIOLÓGICOS EN LA CIUDAD DE ROSARIO[1]

Resumen

En muchos estudios por muestreo las unidades de interés se encuentren distribuidas geográficamente. En esta situación, suele ocurrir que observaciones cercanas en el espacio tengan valores de la variable de interés similares, lo que se conoce como correlación espacial positiva. Si esta característica se tiene en cuenta al momento del definir un diseño muestral, éste puede mejorar notablemente. Algunos de los diseños muestrales con esta particularidad son el diseño estratificado en teselas aleatorizadas (GRTS), método pivotal local (LPM) y el muestreo Poisson espacialmente correlacionado (SCPS). En un principio, estas técnicas surgieron debido al interés en realizar estudios sobre recursos naturales o problemas ambientales. En este trabajo, se hace uso de ellas en un estudio epidemiológico, como lo es la estimación del total de casos de dengue en la ciudad de Rosario entre el 30/07/2023 y 27/07/2024. Dicha aplicación, servirá como base para la utilización de estas técnicas para otras enfermedades de transmisión virales. En este trabajo se compararon las técnicas mencionadas, contrastándolas entre sí y a su vez con el muestreo simple al azar (MSA), el cual no contempla la ubicación geográfica de la población. Se ha concluido que los métodos que consideran dicha información presentan una mayor eficiencia respecto al MSA.

Palabras clave: Muestreo espacial. Muestreo balanceado. Estudio epidemiológico.

Abstract

In many sampling studies, the units of interest are geographically distributed. In such cases, it is often observed that nearby units have similar values of the variable of interest, which is known as positive spatial correlation. If this characteristic is considered when defining a sampling design, it can significantly improve. Some of the sampling designs with this particularity are the stratified design with randomized tessellation (GRTS), the local pivotal method (LPM), and spatially correlated Poisson sampling (SCPS). Initially, these techniques emerged due to the interest in conducting studies on natural resources or environmental issues. In this work, they are applied in an epidemiological study, specifically the estimation of the total number of dengue cases in the city of Rosario between July 30, 2023, and July 27, 2024. This application will serve as a foundation for using these techniques for other viral transmission diseases. In this work, the mentioned techniques were compared and contrasted with each other, and simple random sampling (MSA), which does not consider the

[1] Trabajo elaborado en el marco del Proyecto 80020220700061UR titulado: “Muestreo espacial y su aplicación en estudios económicos y sociales”, dirigido por José Alberto Pagura.

geographical location of the population. It has been concluded that the methods that take this information into account present greater efficiency compared to MSA.

Keywords: Spatial Sampling – Balanced Sampling – Epidemiology Studies

Introducción

Los planteos usuales para la selección de muestras en encuestas sociales se basan en métodos tradicionales, como por ejemplo el muestreo aleatorio simple, estratificado, sistemático o por conglomerados. Cuando las unidades que conforman la población se encuentran distribuidas en el espacio geográfico, frecuentemente las unidades cercanas son más parecidas entre ellas respecto a aquellas que están a mayor distancia, lo que se conoce como correlación espacial positiva. Una muestra brindará estimaciones precisas, si contiene unidades que representan la variabilidad existente en la población, por lo que en poblaciones con estas características, convendrá evitar la inclusión en la muestra de unidades vecinas.

El muestreo estratificado a partir de una división de la región en la que se encuentran las unidades, así como el muestreo sistemático, fueron inicialmente las estrategias preferentemente utilizadas para lograr mejor precisión en casos como el mencionado.

Posteriormente surgieron métodos de selección que obtienen muestras con unidades "espacialmente dispuestas" en forma similar a su disposición en la población (*muestras espacialmente balanceadas*), constituyendo una alternativa beneficiosa para la reducción del error de muestreo. Estas técnicas de muestreo agrupadas bajo la denominación de "muestreo espacial" surgieron como una opción para este tipo de poblaciones; las mismas aprovechan la información sobre la ubicación de la unidad para lograr extraer una muestra representativa de la población, captando la variabilidad de la variable en estudio. El origen de este tipo de muestreo se encuentra en la estimación de características relacionadas con recursos naturales o problemas ambientales, pero su aplicación en estudios sociales, socioeconómicos o epidemiológicos es escasa.

Entre los métodos de muestreo espacial más conocidos se destacan:

-Diseño estratificado en teselas aleatorizadas (GRTS).

-Muestreo Poisson espacialmente correlacionado (SCPS).

-Método pivotal local (LPM).

El objetivo principal de este trabajo es la propuesta de aplicación de los métodos mencionados en estudios epidemiológicos, como lo es la estimación del total de casos de dengue en la ciudad de Rosario y la comparación del desempeño de estas técnicas aprovechando la información disponible para toda la población en el período comprendido entre las semanas epidemiológicas 31/2023 y 30/2024 en la ciudad de Rosario.

Presentación del problema

De acuerdo con la Organización Mundial de la Salud (OMS), el dengue es una infección vírica, la cual se transmite a partir de los mosquitos hembra infectados (principalmente los mosquitos Aedes aegypti y Aedes albopictus) a las personas.

Entre las semanas epidemiológicas 31/2023 y 30/2024 (período comprendido entre el 30/07/2023 al 27/07/2024) han sido confirmados 24.480 casos en la ciudad de Rosario,

informados por organismos municipales y privados. De estos casos, 19.225 fueron reportados a la red de salud pública municipal (Gobierno de Rosario, 2024).

Esta enfermedad es de declaración obligatoria y, por lo tanto, para este trabajo se cuenta con el total de casos de dengue registrados a nivel municipal para la ciudad de Rosario en el período de referencia. La aplicación del muestreo espacial a este conjunto de datos y su análisis detallado servirá como base para la utilización de estas técnicas en otras enfermedades de transmisión viral que no sean de declaración obligatoria y que sea de interés estimar el total de casos para la ciudad.

Específicamente, se tiene la cantidad de casos de dengue por radio censal de la ciudad de Rosario (definidos según el Censo de Población, Hogares y Vivienda 2010) y a partir de muestras espacialmente balanceadas, se propone estimar el total de casos de dengue con su error estándar, para luego comparar la precisión de los métodos que tienen en cuenta la ubicación geográfica de las unidades (GRTS, SCPS, LPM) en contraste con muestreo simple al azar (MSA).

El estudio se realiza mediante la extracción de 10.000 muestras para cada uno de los métodos de selección propuestos, compuestas por 100 radios censales cada una. El comportamiento de cada plan se realiza observando la eficiencia relativa con respecto al muestreo aleatorio simple, el valor medio de la estimación usual de la variancia, el valor medio de la estimación de la variancia por el método del vecindario local (propuesta específica para estos métodos de selección), y teniendo en cuenta el índice de balance espacial (PEI).

Planes de muestreo espacial

A continuación, se describen brevemente los métodos espaciales empleados:

Diseño estratificado en teselas aleatorizadas (GRTS):

El método GRTS fue presentado por Stevens & Olsen (2004) y se basa en la creación de una función que mapea el espacio bidimensional en un espacio unidimensional, definiendo una dirección espacial ordenada. Dicha técnica fue diseñada principalmente para poblaciones en el espacio con el objetivo de estudiar recursos naturales.

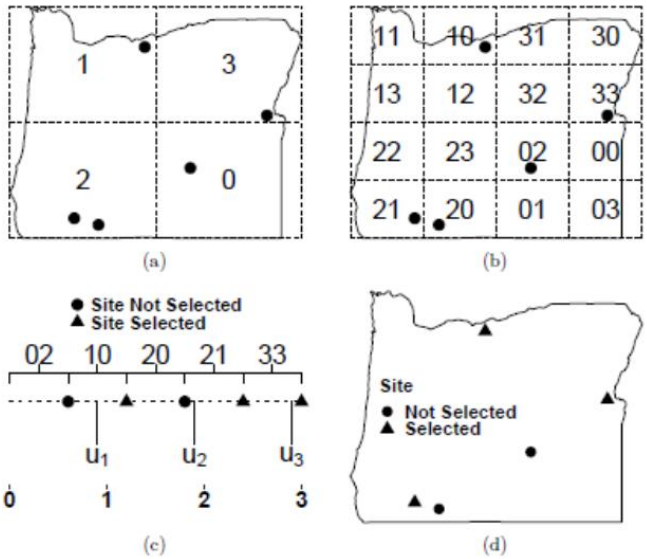
El método consiste en dividir la región en cuatro cuadrados distintos de igual tamaño llamados celdas de nivel uno. A cada celda de nivel uno se le asigna aleatoriamente una dirección o referencia de nivel uno que puede ser 0, 1, 2 o 3 (Figura 1 a).

Cada celda de nivel uno se divide en cuatro cuadrados distintos de igual tamaño llamados celdas de nivel dos. A cada celda de nivel dos, se le asigna aleatoriamente una dirección de nivel dos que puede ser 0, 1, 2 o 3. La división continúa durante k pasos (Figura 1 b).

Se llevan las unidades, siguiendo el orden de las direcciones, a una recta. Cada unidad queda representada por un segmento de longitud proporcional a la probabilidad de inclusión de la misma y luego se selecciona una muestra sistemática (Figura 1 c).

Por último, se vuelven a ubicar las unidades seleccionadas en el territorio, pudiendo apreciar cómo se distribuye la muestra en el mismo (Figura 1 d).

Figura 1: Una descripción visual del método GRTS utilizando sitios de una muestra ilustrativa en Oregon, EEUU.



Fuente: "spsurvey: Spatial Sampling Design and Analysis in R", Journal of Statistical Software, 105 (3). https://doi.org/10.18637/jss.v105.i03

Para la implementación de esta técnica, se desarrolló el paquete de RStudio "spsurvey" (Kincaid et al., 2019) que fue utilizado en el presente trabajo.

Muestreo Poisson espacialmente correlacionado (SCPS):

Este método fue propuesto por Grafström (2012). En una primera instancia, se elige un orden de las unidades que componen la población. Luego, se toma la primera unidad y se decide mediante un determinado criterio, si esta pertenecerá o no a la muestra. La decisión tomada se plasma en la probabilidad de inclusión actualizada que será $I_1 = 1$ si la unidad se incluye en la muestra, o $I_1 = 0$ si no forma parte de la misma. Una vez tomada la decisión, se actualizan las probabilidades de inclusión de las unidades restantes, según la siguiente regla de actualización:

$$\pi_i^{(j)} = \pi_i^{(j-1)} - \left(I_j - \pi_j^{(j-1)}\right) w_j^{(i)}$$

donde $\pi_i^{(j)}$ es la probabilidad de inclusión de orden j, I_j es la variable indicadora de inclusión de la muestra y $w_j^{(i)}, j < i$, es el peso dado por la unidad *j* a las unidades $i = j+1, j+2, \ldots N$. Además, π_i es la probabilidad de inclusión original de la unidad i, igual a $\pi_i^{(0)}$.

La actualización se puede ilustrar de la siguiente manera:

$$\begin{array}{lccccccc}
\pi^{(0)}: & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \dots & \pi_N \\
\pi^{(1)}: & I_1 & \pi_2^{(1)} & \pi_3^{(1)} & \pi_4^{(1)} & \dots & \pi_N^{(1)} \\
\pi^{(2)}: & I_1 & I_2 & \pi_3^{(2)} & \pi_4^{(2)} & \dots & \pi_N^{(2)} \\
\pi^{(3)}: & I_1 & I_2 & I_3 & \pi_4^{(3)} & \dots & \pi_N^{(3)} \\
 & & & \dots & & & \\
\pi^{(N)}: & I_1 & I_2 & I_3 & I_4 & \dots & I_N
\end{array}$$

De este modo, se actualizan gradualmente las probabilidades de inclusión en N pasos.

El peso $w_j^{(i)}, j < i$, determina cómo el resultado del muestreo en la unidad j, se ve afectado por la probabilidad de inclusión de la unidad i. Los pesos deben cumplir la siguiente restricción:

$$-min\left(\frac{1-\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \leq w_j^{(i)} \leq min\left(\frac{\pi_i^{(j-1)}}{1-\pi_j^{(j-1)}}, \frac{1-\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right)$$

para que se cumpla que $0 \leq \pi_i^{(j)} \leq 1, i = j+1, j+2, \dots, N$.

Un enfoque para elegir los pesos es a través de la estrategia de peso máximo que consiste en que primero, la unidad j asigne tanto peso como sea posible a la unidad más cercana (en distancia) entre las unidades $i = j+1, j+2, \dots, N$, luego tanto peso a la segunda unidad más cercana, etc., con la restricción de que los pesos sumen 1.

Para utilizar este método en RStudio, se hizo uso de la función *scps* del paquete "BalancedSampling" (Grafström & Lisic, 2019).

Método pivotal local (LPM):

El método pivotal, propuesto por Deville y Tillé (1998), es un método de muestreo que permite la selección de unidades con probabilidades de inclusión desiguales. En cada paso, se seleccionan aleatoriamente dos unidades y se actualizan sus probabilidades de inclusión de tal manera que se decide el resultado de muestreo para al menos una de las dos unidades. Por lo tanto, se obtiene la muestra en un máximo de N pasos.

Es decir, en un paso *t* se eligen aleatoriamente dos unidades (i y j). Una de estas unidades no será elegida en este paso para formar parte de la muestra y continuará siendo parte de la población con una probabilidad de inclusión actualiza. La otra unidad estará "terminada"; su probabilidad de inclusión será actualizada igualándola a 1 o 0, es decir: será incluida o no en la muestra y no será más parte del conjunto de unidades a seleccionar para repetir los pasos.

Las probabilidades de inclusión se actualizan según la siguiente regla:

- Si $\pi_i + \pi_j < 1$:

$$(\pi_i', \pi_j') = \begin{cases} (0, \pi_i + \pi_j) \text{ con probabilidad } \dfrac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) \text{ con probabilidad } \dfrac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

- Si $\pi_i + \pi_j \geq 1$

$$(\pi_i', \pi_j') = \begin{cases} (1, \pi_i + \pi_j - 1) \text{ con probabilidad } \dfrac{1 - \pi_j}{2 - \pi_i - \pi_j} \\ (\pi_i + \pi_j - 1, 1) \text{ con probabilidad } \dfrac{1 - \pi_i}{2 - \pi_i - \pi_j} \end{cases}$$

De dicho método, surgen dos alternativas: método pivotal local 1 (LPM1) y método pivotal 2 (LPM2). Estos dos tienen en cuenta la ubicación geográfica de las unidades, ya que, al momento de seleccionar las unidades, lo hacen según la cercanía entre ellas. A continuación, se describen los pasos a seguir en cada uno de ellos.

LPM1:

1.Se elige aleatoriamente una unidad i.

2.Se elige la unidad j, que sea el vecino más cercano de i.

3.Si j tiene a i como su vecino más cercano, se actualizan las probabilidades de inclusión de acuerdo con la regla de actualización. De lo contrario, se vuelve al paso número 1.

4.Si todas las unidades están "terminadas", se detiene el proceso. De lo contrario, se vuelve al paso 1.

LPM2:

Los pasos 1, 2 y 4 son iguales al método LPM1. En cuanto al paso 3, directamente se actualizan las probabilidades de inclusión para las unidades i y j de acuerdo a la regla de actualización.

Respecto a LPM1, este produce resultados más equilibrados espacialmente que LPM2, aunque este otro es más simple y rápido.

Para hacer uso de estos métodos en RStudio, se utilizaron las funciones *lpm1* y *lpm2* del paquete BalancedSampling.

Comparación de los métodos

Teniendo en cuenta que el plan de muestreo se completa utilizando como estimador del total al conocido estimador de Horvitz-Thompson, las comparaciones se realizan mediante el estimador de la variancia del total de Horvitz-Thompson, el estimador de la variancia del vecindario local del total, el error cuadrático medio, la eficiencia relativa y el valor PEI.

Sea *Y* la variable en estudio, y_i el valor de la *i-ésima* unidad, π_i la probabilidad de inclusión de la *i-ésima* unidad y n es el tamaño de la muestra s.

El estimador del total de Horvitz-Thompson es:

$$\hat{\tau}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i},$$

y el estimador de la variancia del total es:

$$\hat{V}_{HT}(\hat{\tau}_{HT}) = \sum_{i \in s} \sum_{j \in s} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

El estimador de la variancia del vecindario local es:

$$\hat{V}_{NBH}(\hat{\tau}) = \sum_{i \in s} \sum_{j \in s} w_{ij} \left(\frac{y_j}{\pi_j} - \bar{y}_{D_i} \right)^2,$$

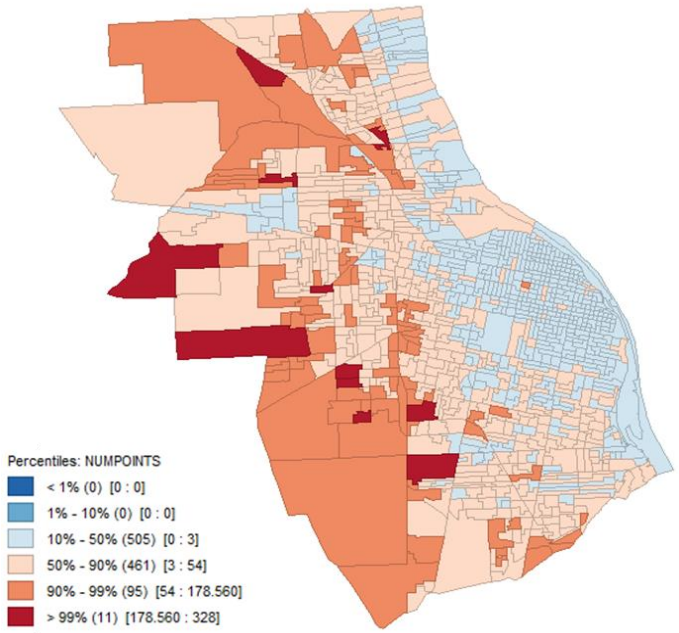
donde w_{ij} son pesos que decrecen a medida que aumenta la distancia entre i y j y cuya suma es igual a 1. D_i es un vecindario de la unidad *i*, $\bar{y}_{D_i}$ es el promedio en el vecindario de la unidad *i*.

Descripción de la población

Según el reporte municipal la distribución geográfica de los casos de dengue, representada en la Figura 2, muestra que los radios censales con menores cantidades de casos de dengue se encuentran localizados mayoritariamente en la zona centro de la ciudad y en un sector de la zona norte y están rodeados de radios con la misma característica (color celeste). En la zona noroeste se distingue un grupo de radios censales con bajas cantidades de casos de dengue, pero rodeados de radios con altas cantidades de casos de dengue. Se observa también que radios censales con mayores cantidades casos de dengue se concentran mayormente en la periferia de la ciudad (colores naranja y rojo).

El índice de asociación global de Moran mide la tendencia de valores similares a agruparse en el espacio, es decir, hasta qué punto áreas con altos niveles de casos están cerca de otras áreas con iguales características mientras que las zonas de pocos casos están rodeadas de otras similares. Según el criterio de conectividad tipo reina, este índice resulta igual a 0,46 (p-value=0,00) mostrando existencia de autocorrelación espacial positiva significativa, lo que concuerda con lo observado en el mapa.

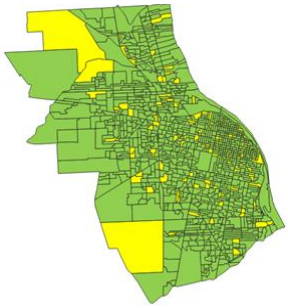
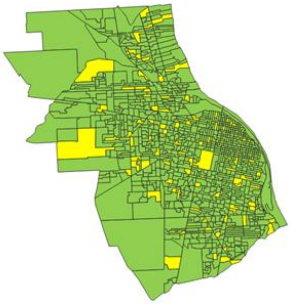
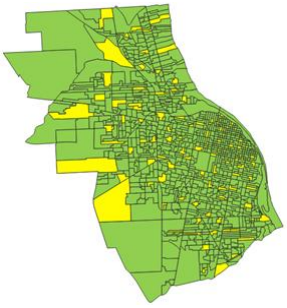
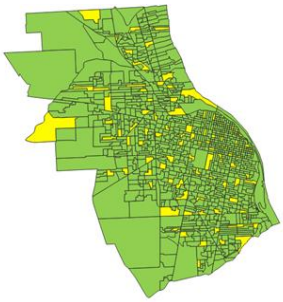
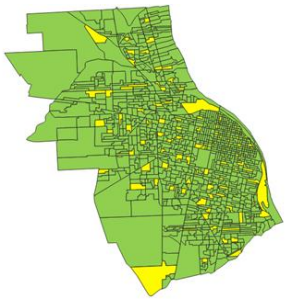
Figura 2: Mapa de percentiles para la cantidad de casos de dengue en la ciudad de Rosario



Para ilustrar la cobertura del territorio que se logra con las muestras obtenidas con cada método, se grafica el mapa de Rosario con las unidades seleccionadas en una muestra para cada una de las técnicas consideradas (Figura 3).

Para todos los planes de muestreo estudiados, las unidades que forman parte de la muestra (marcadas en color amarillo) se encuentran distribuidas en todo el territorio, incluido el MSA.

Figura 3: Ejemplo ilustrativo de la localización de las unidades de una muestra extraída por cada método

Muestra MSA	Muestra GRTS	Muestra SCPS
		
Muestra LPM1	Muestra LPM2	
		

Estudio comparativo

Los resultados de las 10000 muestras de tamaño 100 extraídas por cada método estudiado se estimó el error cuadrático medio (ECM) del estimador para luego obtener el cociente de ellos con respecto al correspondiente al MSA (ER), el valor medio de las variancias estimadas, el valor medio de las variancias del vecindario local estimadas y el valor medio del PEI (Tabla 1).

Tabla 1: Medidas resúmenes para los diferentes métodos de muestreo para un tamaño muestral de 100 unidades

Método	ECM	ER	Valor medio de las variancias estimadas	Valor medio de las variancias del vecindario local estimadas	Valor medio de PEI
MSA	14.150.220	1,00	13.901.424	8.016.813	0,033
GRTS	10.105.837	1,40	14.047.873	8.205.642	0,020
SCPS	8.678.337	1,63	13.996.578	8.424.287	0,009
LPM1	8.612.287	1,64	14.033.867	8.436.766	0,010
LPM2	8.793.744	1,61	13.981.704	8.374.266	0,010

De los resultados, se desprende:

- El método GRTS es el que presenta la menor ganancia, lográndose un 40% de disminución en el error de muestreo con su utilización. Los métodos SCPS y LPMs presentan ganancias mayores y cercanas al 60%.
- Los promedios a través de las 10.000 muestras, de las variancias estimadas por HT, son mayores que los ECM para aquellos planes que tienen en cuenta la variabilidad espacial.
- En cambio, los promedios de las variancias estimadas por el método del vecindario local resultaron similares a los ECM en aquellos planes que tienen en cuenta la variabilidad espacial.
- Los métodos SCPS, LPM1 y LPM2 presentan los mejores balances espaciales de acuerdo a los valores PEI, seguido por el método GRTS y por último, como era de esperar, el MSA. Si bien todos los valores son pequeños, en las 10.000 muestras se ha encontrado que para el MSA este índice varía entre 0,016 y 0,059, mientras que en los otros métodos varían entre 0,004 y 0,016.
- Todos los métodos que emplean información de la correlación espacial son más eficientes que el MSA.

Las conclusiones anteriores se obtienen para tamaños de muestra igual a 100. Sin embargo, se ha probado con otros tamaños de muestra menores y, para ellos, las variancias del vecindario local son mayores que el ECM.

Futuros avances

Los resultados obtenidos motivan el estudio del comportamiento de estos métodos para diferentes tamaños de muestra. A su vez, se considera interesante evaluar el comportamiento del sesgo del estimador de la variancia del vecindario local tanto como para los diferentes métodos de selección como para diferentes tamaños muestrales. También, se propone utilizar métodos de selección y estimación en muestras complejas.

Por otro lado, se plantea el estudio del comportamiento de estimadores basados en modelos.

Referencias Bibliográficas

- Benedetti R.; Piersimoni F.; Postiglioni P. (2015) Sampling Spatial Units for Agricultural Surveys. Springer.
- GeoDa (versión 1.22.0.4). *Software de análisis espacial*. (2023).
- Gobierno de Rosario. *Situación Epidemiológica Semana 29 2024.* 2024. https://datos.rosario.gob.ar/sites/default/files/2024-07/Situaci%C3%B3n%20epidemiol%C3%B3gica%20semana%2029%202024.pdf
- Grafström A. (2012). *Spacially correlated Poisson sampling.* Journal of Statistical Planning and Inference, 142(1), 139-147.
- 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.
- QGIS Development Team (2024). *QGIS Geographic Information System* (versión 3.38).
- 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. (2004) Spatially Balanced Sampling of Natural Resources. JASA, 99: 465, 262-278.