Borra, Virginia Laura
Pagura, José Alberto
Mignoni, César Antonio
López, Elisabet
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

ALTERNATIVAS EN LA ELECCION DEL SEMIVARIOGRAMA A EMPLEAR EN LA ESTIMACION DE TOTALES A PARTIR DE MUESTRAS CON DATOS ESPACIALMENTE CORRELACIONADOS

Resumen

Los enfoques basados en modelos y asistidos por modelos han resultado de utilidad en el muestreo en poblaciones finitas, brindando un soporte metodológico para la incorporación de información auxiliar en la fase de estimación, con el fin de obtener mejoras en la precisión de las estimaciones.

Cuando las unidades se encuentran ubicadas en el espacio, la característica principal es que pueden presentar autocorrelación espacial positiva. La Estadística Espacial ha desarrollado métodos para la recolección, descripción y modelamiento de la variabilidad espacial de las variables de interés, los que pueden ser empleados en las fases de selección y de estimación de valores de la población finita.

Las propuestas para la estimación bajo el enfoque basado en modelos, plantean la incorporación de la correlación espacial utilizando modelos de semivariograma, por lo que será clave la identificación y la estimación de sus parámetros. Una opción es el uso de un modelo de semivariograma obtenido de una muestra piloto, o de un estudio anterior. También, podría utilizarse para este fin la información muestral, ya sea para establecer y estimar el modelo o dado un modelo poblacional, estimar sus parámetros.

En este trabajo, se presenta una aplicación con el fin de comparar los resultados obtenidos por medio de los tres procedimientos y su correspondiente discusión.

Abstract

Model-Based Approach and Model-Assisted Approach have been useful in finite population sampling, providing a methodological support for incorporating auxiliary information in the estimating stage, in order to obtain improvements in accuracy estimates.

When units are located in space, it`s main feature is that they may have positive spatial autocorrelation. Spatial Statistics has developed methods for collection, description and modeling of spatial variability of the interest variables, which can be used in the stages of selection and estimation of values of the finite population.

Proposals for estimation under the Model-Based Approach, suggest the incorporation of spatial correlation using semivariogram models, so it will be important the identification and estimation of its parameters. One option is the use of a semivariogram model obtained from a pilot sample, or from a previous study. Also, the sample information could be used for this purpose, either to establish and estimate the model or given a population model, estimate its parameters.
In this paper, an application is presented in order to compare the results obtained by the three procedures and it`s corresponding discussion.

1-INTRODUCCION

Los enfoque basado en modelos y asistido por modelos han resultado de gran utilidad en el muestreo en poblaciones finitas, brindando un soporte metodológico para la incorporación de información auxiliar en las fases de selección y de estimación, con la finalidad de obtener mejoras en la precisión de las estimaciones.

En los casos en los que las unidades de muestreo se encuentran ubicadas en el espacio, la característica principal de estos datos es que pueden presentar autocorrelación espacial positiva, es decir, mientras más cercanía haya entre dos unidades, más parecidas son en cuanto a la variable que se estudia. La Estadística Espacial ha desarrollado métodos para la recolección, descripción y modelamiento de la variabilidad o correlación espacial de las variables que caracterizan el fenómeno, los que pueden ser empleados en mejoras de los diseños muestrales.

Los procedimientos de estimación que tienen en cuenta la variabilidad espacial plantean incorporar esta información mediante modelos de semivariograma, siendo entonces una cuestión primordial, la identificación y estimación del mismo.

Para ello, algunos autores sugieren el uso de un modelo de semivariograma obtenido de una muestra piloto, o proveniente de un estudio anterior. Naturalmente surge también la idea de identificar el modelo y estimar sus parámetros con los datos de la muestra o definir un modelo poblacional de acuerdo al conocimiento a priori que se tenga del comportamiento de la variable en el espacio y luego emplear los datos de la muestra para la estimación de sus parámetros. Los dos últimos planteos, introducen variabilidad adicional en la estimación de la característica de interés en la población finita.

En trabajos anteriores, los autores han presentado resultados de estudios comparativos mostrando mejoras en planes de muestreo que utilizan información espacial para estimación de características socioeconómicas en la ciudad de Rosario. En todos ellos se ha procedido con semivariogramas poblacionales.

En este trabajo se presenta la aplicación de las propuestas enunciadas utilizando muestreo sistemático de radios censales, para la estimación del total de hogares con necesidades básicas insatisfechas en la ciudad de Rosario, contando para ello con datos poblacionales del Censo Nacional de Población Hogares y Viviendas 2001. Finalmente, se presenta la correspondiente discusión destinada a destacar aspectos relevantes del empleo de la información auxiliar mencionada.

2-METODOLOGIA

Como ya se ha expresado, los enfoques más recientes para la estimación han propiciado el aprovechamiento de la información auxiliar de la que se suele disponer frecuentemente. El enfoque de modelos parte de asumir que la población finita es una realización de un determinado proceso aleatorio para luego, estimar los parámetros del proceso con la información provista por la muestra. Luego se predice la característica de interés de la población finita. Frecuentemente, la información auxiliar de referencia consiste en un conjunto de variables observadas a cada unidad de la población finita y su uso se traduce en la estratificación, la selección con probabilidades desiguales o las estimaciones de regresión o razón.

El enfoque asistido por modelos, atiende al enfoque de modelos para la elección del estimador, pero su error se considera de acuerdo a la distribución del mismo a través de todas las muestras posibles.

Un caso particular es aquel en el que las unidades se encuentran distribuidas en el espacio y presentan correlación espacial positiva. Dicha correlación puede describirse mediante un correlograma o un semivariograma, prefiriendo este último por sus propiedades estadísticas. El modelo puede incorporarse al proceso de estimación como se muestra más adelante o utilizarse en la fase de selección de la muestra, lo cual hace que sea clave la estimación del mismo.

En esta sección se presentan aspectos generales del modelo de semivariograma y la forma de empleo del mismo en la fase de estimación.

2.a. Modelado de la variabilidad espacial.

Si Y(s) es la variable de interés, observada en el punto de coordenadas s perteneciente a una región A, la estructura de la correlación espacial se modela considerando al valor observado Y(s) como una realización espacial de una variable aleatoria. En general, se asume que el comportamiento de Y responde a un proceso estocástico es estacionario de segundo orden, es decir:

(i) $E[Y(s)] = \mu$

(ii) $Cov[Y(s), Y(s')] = E[(Y(s) - \mu)(Y(s') - \mu)] = C(s, s') \quad \forall s, s' \in A$.

De esta forma, el valor esperado de Y(s) es constante para cualquier punto perteneciente a la región A y la covariancia entre los valores de la variable en dos puntos cualesquiera del área A (Y(s) y Y(s')) dependen del vector que separe a los puntos s y s' en módulo y orientación.

Cuando la covariancia entre los valores de la variable en dos puntos depende del módulo y de la orientación del vector que une dichos puntos, se dice que el proceso es anisotrópico.

En cambio, cuando la covariancia depende sólo de la distancia que separa los puntos el proceso se llama isotrópico y en ese caso la condición (ii) puede escribirse como:

(ii) $Cov[Y(s), Y(s')] = E[(Y(s) - \mu)(Y(s') - \mu)] = C(dist(s, s')) \quad \forall s, s' \in A$.

A la función $C(dist(s, s'))$ se la denomina covariograma.

Esta condición se puede expresar en términos del variograma o correlograma, en lugar del covariograma. El variograma se define en forma general como:

$$2\gamma(s, s') = Var(Y(s) - Y(s')) \quad \forall s, s' \in A .$$

Cuando el proceso es estacionario de segundo orden e isotrópico, el valor esperado de Y(s) es constante para todas las unidades y la expresión del variograma se simplifica a:

$$2\gamma(s, s') = 2Var(Y(s)) - 2Cov(Y(s), Y(s')) .$$

A $\gamma(s, s')$ se lo denomina semivariograma y si se denota con C(s, s) a la variancia de Y(s) y sea d la distancia entre los puntos s y s', se lo puede escribir como:

$$\gamma(s, s') = C(s, s) - C(dist(s, s')) = C(0) - C(h) = \gamma(h) .$$

Modelos de semivariograma

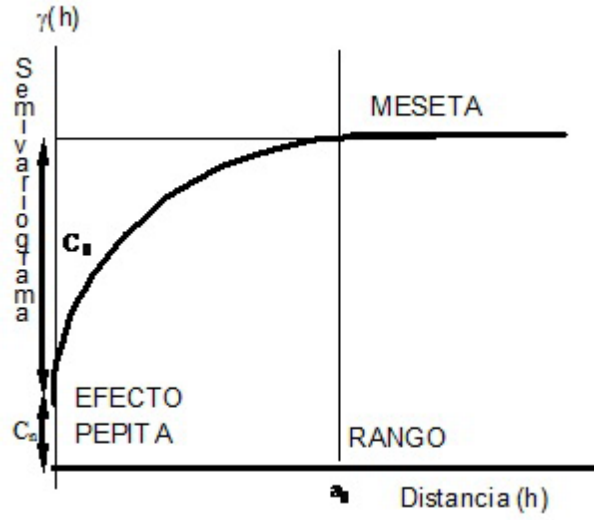
En un modelo de semivariograma para procesos estacionarios isotrópicos de segundo orden se distinguen los parámetros:

Meseta (C_0): es el valor límite máximo constante que alcanza el semivariograma.

Rango (a_0): es el valor de la distancia a partir del cual se alcanza la meseta. Para valores de h inferiores al rango existe correlación espacial entre las observaciones.

Efecto pepita (c_n): es la discontinuidad que presenta el semivariograma en el origen.

Figura1. El semivariograma y sus parámetros.



Dependiendo de la forma del semivarigrama teórico, se distinguen algunos modelos, como los modelos exponencial, esférico, gausiano, efecto agujero (Ambrosio, 2000). Por ejemplo el modelo de semivariograma exponencial responde a la siguiente expresión:

$$\gamma_{exp}(h) = c_n + c_0\left[1 - \exp\left(\frac{-h}{a_0}\right)\right] \quad h > 0$$

y su representación gráfica es similar a la presentada en la Figura 1.

Métodos de estimación del semivariograma

Un problema de importancia lo constituye la identificación del modelo y la estimación de sus parámetros. Los procedimientos que se plantean para abordar esta cuestión, se presentan en forma sintética, a continuación.

- Semivariograma empiríco o experimental

El estimador de momentos de $\gamma(h)$ para un proceso estacionario isotrópico de segundo orden es (Matheron (1962), Cressie (1993)):

$$\hat{\gamma}(h) = \frac{1}{2|n(h)|}\sum_{n(h)}(Y(s) - Y(s'))^2$$

donde:

$$n(h) = \{(s,s') / dist(s,s') = h \quad \forall s,s' \in muestra\}$$

$|n(h)|$ es el número de pares distintos en n(h).

- Mínimos Cuadrados Ponderado

Dado un modelo de semivariograma teórico, se estiman sus parámetros dependiendo de la forma del semivariograma experimental.

En el ajuste basado en mínimos cuadrados, se desea estimar el vector de parámetros $\boldsymbol{\theta} = (c_n, c_0, a_0)$ del semivariograma teórico $\gamma(h)$ que minimice la suma de cuadrados de la diferencia $R(\boldsymbol{\theta})$ dado por la expresión:

$$R(\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i^2 \left[\hat{\gamma}(h_i) - \gamma(h_i;\boldsymbol{\theta})\right]^2 .$$

Para i=1,...,k, las ponderaciones son $w_i^2 = 1/\operatorname{var}\left[\hat{\gamma}(h_i)\right]$ en el caso de Mínimos Cuadrados Ponderados y $w_i^2 = 1$ en caso de Mínimos Cuadrados Ordinarios.

Para Mínimos Cuadrados Ponderados, Cressie (1985) mostró que bajo los supuestos de que las observaciones se distribuyen normalmente y de diferencias cuadradas no correlacionadas en la semivariancia empírica, la estimación de mínimos cuadrados ponderada aproximada del vector de parámetros $\boldsymbol{\theta}$ se obtiene minimizando

$$R(\boldsymbol{\theta})_{MCP} = \frac{1}{2}\sum_{i=1}^{k} n(h_i)\left[\frac{\hat{\gamma}(h_i)}{\gamma(h_i;\boldsymbol{\theta})} - 1\right]^2$$

donde $n(h_i)$ es el número de pares de puntos en la i-ésima lag distancia.

2.b. Inclusión de la variabilidad espacial en el proceso de estimación: Aproximación basada en modelos

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.

El más reciente enfoque de modelos, ha contribuido a la teoría del muestro en poblaciones finitas de varias formas. 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

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

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

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

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

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

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

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

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

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

El Predictor lineal insesgado y óptimo (BLUP) 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 BLUP de $\hat{Y}$ es $\hat{Y} = 1_n' Y_n + 1_{N-n}' \left[X_{N-n,p} \hat{\beta} + V_{N-n,n} V_{n,n}^{-1} (Y_n - X_{n,p} \hat{\beta}) \right]$ (1).

El error cuadrático medio de la predicción es:

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

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,

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

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

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

En el presente trabajo se plantean dos modelos, donde para los radios censales se considera la variable en estudio Y, número Hogares con necesidades básicas insatisfechas (NBI) y una variable auxiliar X, total de hogares.

Un modelo es: $Y_i = \beta_1 X_i + \varepsilon_i$ con ε_i no correlacionados pero con $\sigma_{\varepsilon_i}^2$ proporcional a X_i, el cual conduce al empleo del conocido estimador de razón; y otro modelo: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ con:

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

La covariancia depende del modelo de semivariograma elegido.

2.c. Aproximación asistida por modelos

La inferencia basada en el diseño realiza las estimaciones de los parámetros dependiendo del diseño muestral elegido para seleccionar la muestra sin tener en cuenta las propiedades de la población finita.

Como se mencionó en la sección anterior, la inferencia basada en el modelo utiliza información auxiliar relacionando a la variable de interés con dicha información mediante un modelo superpoblacional. Bajo este enfoque se necesita que los datos provengan de una muestra probabilística pero la forma en la que se selecciona dicha muestra no se tiene en cuenta para la estimación de los parámetros de interés y a la variable en estudio se la considera como una variable aleatoria

Los enfoques basados en modelos y en diseño no se deben pensar como puntos de vistas contrapuestos, sino que pueden llegar a ser complementarios.

En el presente trabajo, se utilizan los estimadores del total basados en modelos y las comparaciones para evaluar el comportamiento de los mismos, se hará teniendo en cuenta los valores que ellos asumen en las muestras sistemáticas que se seleccionan.

El error cuadrático medio del total queda especificado de la siguiente manera:

$$ECM(\hat{Y}) = Var(\hat{Y}) = \frac{1}{k}\sum_{i=1}^{k}(\hat{Y}_i - Y)^2 \quad (3)$$

Las comparaciones se llevan a cabo mediante la eficiencia relativa, la cual se calculará con el error cuadrático medio del estimador de razón, en el numerador.

3. RESULTADOS

Para la implementación de las propuestas enunciadas, se partió de la población de radios censales de la ciudad de Rosario extrayendo 6 muestras sistemáticas de 149 unidades, considerando como variable en estudio el número de hogares con necesidades básicas insatisfechas. El total poblacional de esta variable fue 29622.

Los datos de cada muestra sistemática se utilizaron para estimar el total de hogares con NBI en Rosario usando:

- Estimador de razón del total
- BLUP con modelo poblacional de semivariograma
- Expresión matemática del BLUP pero con modelo poblacional exponencial y parámetros estimados con la muestra.
- Expresión matemática del BLUP pero identificando el modelo de semivariograma y estimando sus parámetros con la información muestral.

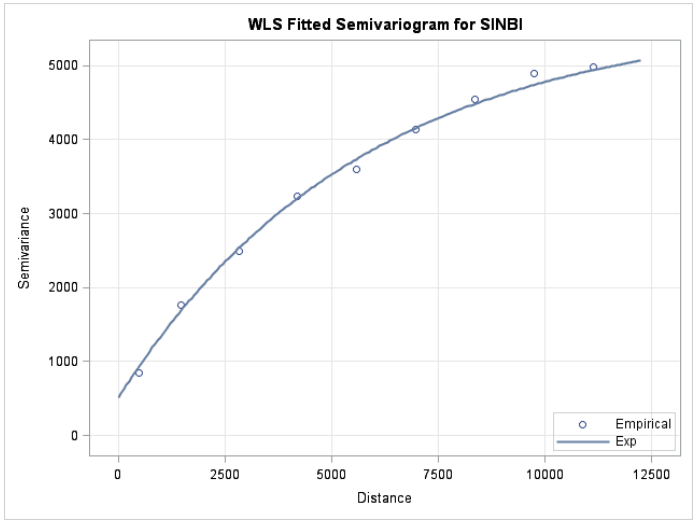
Encontradas las seis estimaciones correspondientes a cada caso, se procedió a calcular el error cuadrático medio.

Modelos de semivariograma utilizados

El modelo se semivariograma poblacional utilizado es exponencial, su representación se encuentra en el Gráfico 1 y la expresión analítica es:

$$\hat{\gamma}_{exp}(h) = 515,78 + 5150,60\left[1 - \exp\left(\frac{-h}{5684,24}\right)\right] \quad h > 0$$

Gráfico 1. Semivariograma empírico y ajustado (Resultado de SAS).



Los gráficos 2 a 7 contienen, para cada muestra, el semivariograma identificado con los datos de la muestra y la estimación de sus parámetros de acuerdo a dicho modelo. Se agrega, superpuesto, el semivariograma exponencial, que es el modelo poblacional, estimado con los datos de la muestra.

En la muestra 1 se identificó un modelo esférico, en las muestras 2, 3 y 6 modelos de potencia, en la 4 esférico y en la 5 se considera un efecto agujero.

Gráfico 2. Semivariogramas muestrales para la muestra 1

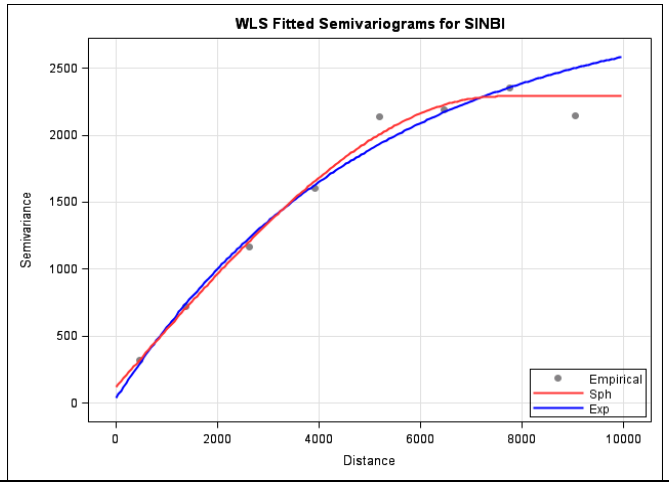


Gráfico 3. Semivariogramas muestrales para la muestra 2

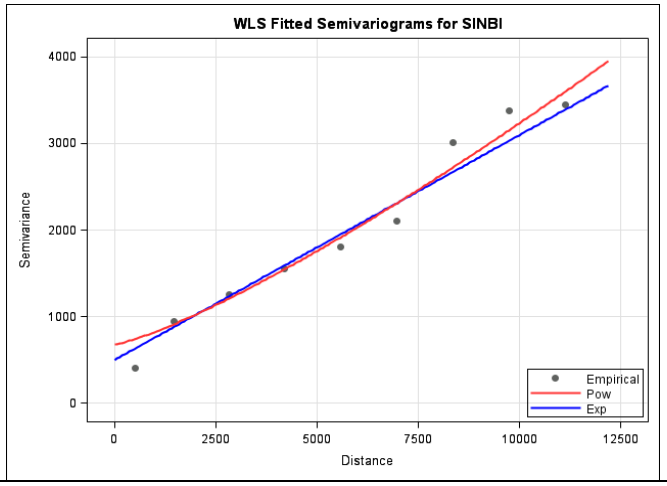


Gráfico 4. Semivariogramas muestrales para la muestra 3

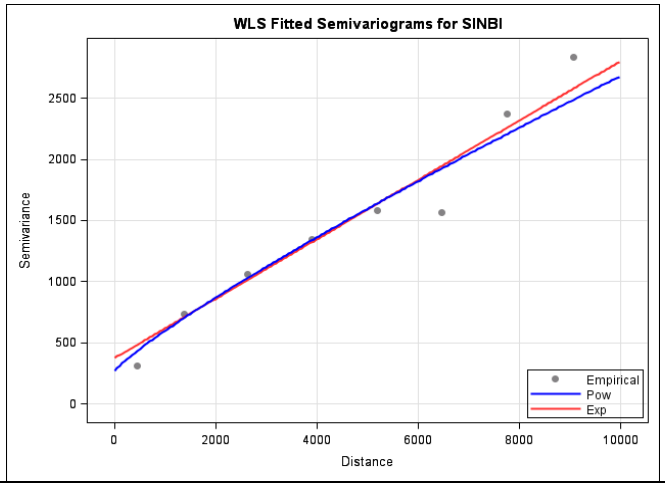


Gráfico 5. Semivariogramas muestrales para la muestra 4

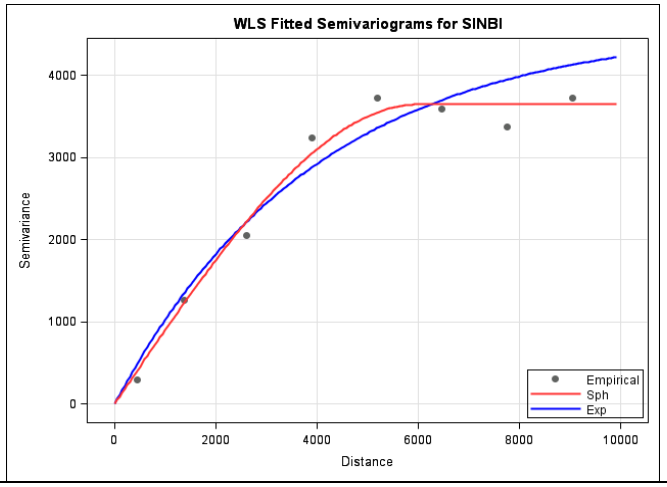


Gráfico 6. Semivariogramas muestrales para la muestra 5

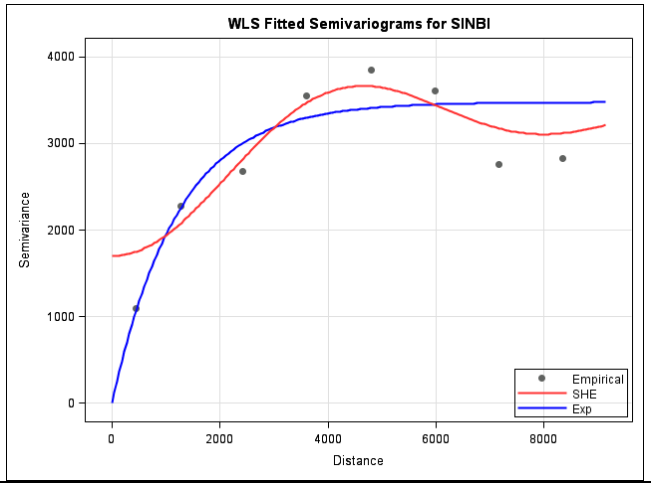
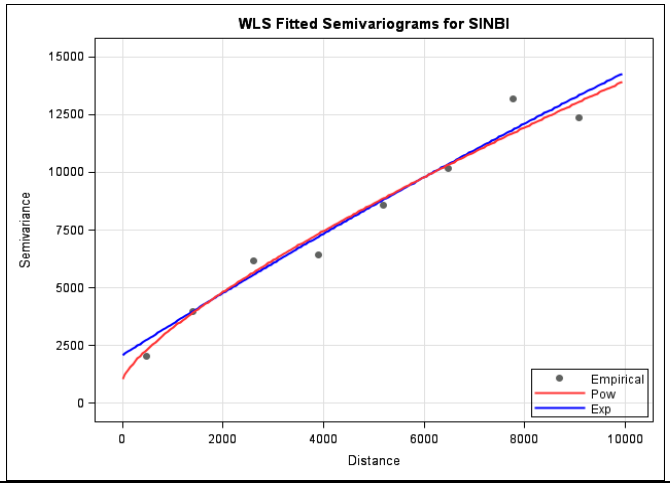


Gráfico 7. Semivariogramas muestrales para la muestra 6



En el Cuadro 1 se presentan las estimaciones obtenidas para cada muestra en cada propuesta, fórmula (1), el error cuadrático medio calculado según (3) y la eficiencia relativa obtenida como el cociente del error cuadrático medio del estimador de razón sobre el de cada una de las propuestas.

Cuadro 1. Estimadores del total en cada muestra sistemática, $\sqrt{ECM(\hat{Y})}$ y eficiencia relativa, para cada estimador propuesto.

Estimador según:	Estimador del total en las muestras sistemáticas:						√ECM	Eficiencia relativa
	1	2	3	4	5	6		
Razón	26254	27598	25699	30129	28798	38461	4280	1
Modelo poblacional Exponencial	26352	27669	26516	29523	27352	32890	2582	2,75
Modelo muestral Exponencial	26148	27972	26561	29045	27203	33095	2659	2,59
Modelo muestral	26231	28163	26536	29830	27248	32952	2579	2,75

4. COMENTARIOS FINALES

Puede decirse que, para el caso en estudio, con cualquiera de las alternativas consideradas, el uso de la información que caracteriza la variabilidad espacial, ha redundado en una mejora importante de la precisión de las estimaciones.

Se esperaba que al plantear estimaciones del modelo de semivariograma a partir de la muestra, el error cuadrático medio aumentara con respecto al que se obtiene al utilizar el semivariograma poblacional, debido a la variabilidad introducida por el uso de diferentes modelos, dependiendo de lo observado en cada muestra y las estimaciones de los parámetros del mismo. En el presente caso los errores cuadráticos medios de los estimadores del total resultaron similares.

Cabe notar que se obtuvieron estimaciones del error cuadrático medio expresado en (3), para cada propuesta y las mismas son más parecidas entre sí para las estimaciones usando el modelo de semivariograma poblacional.

Si bien el estudio realizado se limita a una población determinada y con cierto orden, lo que debe aclararse por haber obtenido resultados para muestras sistemáticas, constituye un aporte en cuanto a mostrar la utilidad del uso de la información de la variabilidad espacial de las unidades.

5. 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 L. (2006). “Estimación del total con datos de conteo sobredispersos y espaciotemporalmente correlacionados: una aproximación basada en la predicción”. Curso de las Jornadas Internacionales de Estadística.

- 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.

- Borra V. L.; Pagura, J. A. (2013). Estimación del total de hogares con necesidades básicas insatisfechas en la ciudad de Rosario utilizando modelos de semivariograma.Decimoctavas Jornadas “Investigaciones en la Facultad de Ciencias Económicas y Estadística”, Noviembre de 2013. Rosario.

- 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.

- SAS/STAT 9.3 User`s Guide: The VARIOGRAM Procedure (Charter) (2011). SAS Institute Inc., Cart. N.C, USA.

- Thompson S. K. (2012) Sampling. John Wiley & Sons Inc.

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