UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

CARRERA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

Estadística Espacial. Muestreo y modelización para la aplicación en estudios socioeconómicos

Autor: Virginia Laura Borra

Director: Dr. José Alberto Pagura

2015

TRIBUNAL EXAMINADOR:

Dr. Carlos Comari

Mg. Leticia Hachuel

M.Sc. Gonzalo Marí

Dedicado a mi hijo Nacho

Agradecimientos

Este trabajo de tesis realizado en la Escuela de Estadística es un esfuerzo en el cual, directa o indirectamente, participaron muchas personas opinando, corrigiendo, teniéndome paciencia, dando ánimo, acompañando en los momentos de crisis y en los momentos de felicidad.

En primer lugar quiero agradecer a mi director, Dr. José Alberto Pagura por su constante seguimiento y dedicación, pero sobre todo por la motivación y el apoyo recibido a lo largo de estos años.

Quiero agradecerles muy especialmente a mis amigas de la facultad, con las que compartí muchos momentos difíciles, pero también muchas charlas que hicieron que este duro camino sea más amigable y divertido.

No quiero olvidarme de mis compañeros de oficina que siempre me han apoyado para que siga adelante y estuvieron presentes para escucharme y aconsejarme, de mi compañera de Latex, de quienes leyeron este trabajo y aportaron su invalorable contribución y de mis amigas de siempre que me dieron fuerzas sin entender de qué les estaba hablando...

Y, por supuesto, el agradecimiento más profundo y sentido es para mi familia. Sin su apoyo, colaboración e inspiración habría sido imposible llevar a cabo esta dura tarea.

En especial, a mis padres, por haberme dado la oportunidad de estudiar en su momento y por su ejemplo de lucha y honestidad. A mis hermanos por estar siempre presentes.

A mi marido Rodrigo, por creer en mí, por su paciencia, su generosidad, apoyo incondicional y por compartir todos mis proyectos.

A mi hijo Nacho, por ser tan dulce y darme la fuerza necesaria para seguir adelante... Por aguantar desde la panza a esta mamá llena de dudas en este arduo camino.

A todos muchas gracias!!!

Resumen

Los enfoque de modelos y asistido por modelos, en planes de muestro para poblaciones finitas amplían las posibilidades de aprovechamiento de la información auxiliar.

En esta tesis se presentan propuestas que evidencian las mejoras obtenidas en la precisión de los estimadores en muestras de poblaciones en las que las unidades presentan correlación espacial. En particular se estudia la estimación del total de hogares con Necesidades Básicas Insatisfechas en Rosario en el año 2001, a partir de una muestra de radios censales, pudiendo generalizar la metodología para otros estudios socioeconómicos.

Se realiza una reseña de los procedimientos usuales para detectar, caracterizar y modelar la variabilidad espacial, pasos necesarios como fundamento de las siguientes etapas. Se presentan métodos de estimación basados en modelos con uso de información de variabilidad espacial, para obtener predicciones de totales y sus Errores Cuadráticos Medios.

Para apreciar las mejoras que pueden lograrse con los métodos planteados, se realizan estudios comparativos en los que se contrastan los resultados obtenidos con procedimientos de estimación usuales que no emplean información de la variabilidad espacial, con un método basado en modelos que incluye esa información mediante el semivariograma poblacional o muestral. La comparación se hace por medio del Error Cuadrático Medio obtenido en la distribución en el muestreo de la población finita.

Se concluye acerca de la conveniencia de utilizar los procedimientos que tienen en cuenta la variabilidad espacial y se plantea la necesidad de desarrollar métodos de estimación para las variancias, así como considerar modelos que sean específicos para variables de conteo y ampliar los procedimientos a los casos de muestreo multietápico.

Palabras Clave: muestreo de poblaciones finitas, enfoque basado en modelos, variabilidad espacial, modelos de semivariograma.

Índice general

Agradecimientos **III**

Resumen **IV**

Lista de figuras **VIII**

Lista de tablas **X**

1. Introducción **1**

2. Análisis exploratorio de datos espaciales **9**

2.1. Introducción 9

2.2. Conceptos generales 9

2.3. Métodos gráficos del Análisis Exploratorio de Datos Espaciales 16

2.3.1. Representación de distribuciones espaciales 17

2.3.2. Representación de la dependencia espacial 18

2.3.3. Representación de la heterogeneidad espacial 26

3. Modelos de variabilidad y correlación espacial **28**

3.1. Introducción 28

3.2. Definiciones 28

3.3. Modelos de semivariogramas y correlogramas 30

3.4. Métodos de estimación del semivariograma 37

3.4.1. Máxima verosimilitud restringida 38

3.4.2. Mínimos cuadrados ponderado 39

3.5. Criterios de elección del semivariograma 40

4. Inferencia basada y asistida por modelos 42

4.1. Introducción 42

4.2. Definiciones 43

4.3. Predicción del total bajo el enfoque de modelos 44

4.4. Algunos casos particulares de modelos de regresión 46

4.4.1. Modelo sin variable auxiliar: Homocedástico y sin autocorrelación 47

4.4.2. Modelo de regresión: Homocedástico y sin autocorrelación 48

4.4.3. Modelo de regresión sin ordenada al origen: Heterocedástico y sin autocorrelación 49

4.4.4. Modelo de regresión: Homocedástico y con autocorrelación 52

4.5. Enfoque asistido por modelos 53

5. Datos en látices 56

5.1. Introducción 56

5.2. Planes de muestreo para una población dispuesta en látices 56

5.2.1. Muestreo aleatorio simple 58

5.2.2. Muestreo estratificado 58

5.2.3. Muestreo sistemático 60

5.3. Estimación del total 61

5.3.1. Muestreo aleatorio simple 62

5.3.2. Muestreo estratificado 62

5.3.3. Muestreo sistemático 63

5.4. Eficiencias relativas 64

6. Resultados 66

6.1. Introducción 66

6.2. Estudio exploratorio 68

6.3. Predicción del total de hogares con NBI empleando diferentes modelos . . 74

6.4. Estudio comparativo: Predictores en muestreo sistemático 77

6.5. Estudio comparativo: Predictores en muestras aleatorias simples 81

6.6. Aplicación para datos en látices 86

7. Consideraciones finales **90**

Bibliografía **95**

A. Apendice **98**

Índice de figuras

3.1. Representación gráfica de los parámetros de un modelo teórico de semivariograma con efecto pepita 31
3.2. Representación gráfica del modelo teórico de semivariograma esférico . . 32
3.3. Representación gráfica del modelo teórico de semivariograma exponencial 32
3.4. Representación gráfica del modelo teórico de semivariograma gaussiano . 33
3.5. Representación gráfica del modelo teórico de semivariograma potencial . 33
3.6. Representación gráfica del modelo teórico de semivariograma efecto agujero 34
3.7. Representación gráfica del modelo teórico de correlograma esférico 35
3.8. Representación gráfica del modelo teórico de correlograma exponencial . 36
3.9. Representación gráfica del modelo teórico de correlograma gaussiano . . . 36

5.1. Esquema de la población dispuesta en látices 57
5.2. Esquema de una muestra aleatoria simple de la población dispuesta en látices 58
5.3. Esquema de una muestra estratificada de la población dispuesta en látices 59
5.4. Esquema de una muestra sistemática con tres arranques aleatorios de la población dispuesta en látices 61

6.1. "Box plot" y "box map" para el número de hogares con NBI 69
6.2. Diagrama de dispersión de Moran para el número de hogares con NBI . . 71
6.3. Mapas LISA para el número de hogares con NBI 72
6.4. Semivariograma empírico para el número de hogares con NBI, a partir de la totalidad de radios censales 73

6.5. Semivariograma empírico y ajustado para el número de hogares con NBI, a partir de la totalidad de radios censales 75
6.6. Semivariogramas muestrales para el número de hogares con NBI en cada muestra sistemática 79
6.7. Histograma de frecuencia y "box plot"para la predicción del total de hogares con NBI 84
6.8. Cantidad de centroides de los radios censales en la grilla definida 87
6.9. Modelo de semivariograma empírico y ajustado para el número de hogares con NBI, a partir de la totalidad de las 513 unidades de muestreo 88
6.10. Eficiencias relativas para diversos tamaños de bloque 89

Índice de cuadros

2.1. Técnicas gráficas del AEDE, según la perspectiva de látices 17

6.1. Predicción del total de hogares con NBI en la ciudad de Rosario, estimación de la raíz cuadrada del Error Cuadrático Medio de total predicho y de la eficiencia relativa, para cada propuesta 76

6.2. Predicción del total de hogares con NBI en cada muestra sistemática, raíz cuadrada del Error Cuadrático Medio según enfoque asistido por modelo y eficiencia relativa 80

6.3. Estimación de la raíz cuadrada del Error Cuadrático Medio basado en el enfoque de modelos para el número de hogares con NBI 81

6.4. Medidas descriptivas para la predicción del total de hogares con NBI . . 84

6.5. Raíz cuadrada del Error Cuadrático Medio según enfoque asistido por modelos y eficiencia relativa 85

6.6. Eficiencias relativas para diversos tamaños de bloque 89

Introducción

En muchos estudios de diferentes campos, se requiere obtener estimaciones de cantidades de interés de una población finita, para lo cual se hace necesaria la implementación de planes de muestreo probabilístico. El permanente esfuerzo de la estadística, se centra en la obtención de estimadores lo más precisos posible con la restricción de recursos disponibles limitados. La precisión hace referencia a la diferencia entre el estimador obtenido con la muestra y el valor poblacional objetivo (error de muestreo). Esta cantidad resulta siempre desconocida pero, teniendo en cuenta aproximaciones teóricas, puede cuantificarse a partir de la estimación del error cuadrático medio del estimador, que se define como el promedio de las diferencias al cuadrado que se pueden obtener entre el estimador para cada muestra posible y la cantidad poblacional de interés.

Un plan de muestreo, queda especificado cuando se decide un método aleatorio de selección de las unidades y un procedimiento que combine los valores observados de una variable de interés, para obtener un estimador del valor poblacional. El método de selección quedará determinado entre otras cosas, por el listado, por restricciones de costo, etc. Por ejemplo, es evidente que a la hora de realizar el trabajo de campo, resulta menos costoso emplear como unidad muestral (aquella que se selecciona) un conjunto de viviendas contiguas o conglomerado, que realizar una selección aleatoria de viviendas, aunque la variable en estudio posiblemente se mida sobre esta última unidad (unidad elemental). Por otra parte, una técnica muy utilizada es la estratificación, es decir agrupar unidades homogéneas formando estratos, para luego seleccionar muestras independientes de cada estrato. Combinando estas alternativas, se puede complejizar la selección aleatoria de unidades pero siempre con el objeto de mejorar los resultados del plan de muestreo.

Con respecto a los procedimientos de estimación, puede mencionarse desde el más

sencillo, de simple expansión, basado sólo en los valores de las variables observados en cada unidad, así como otros métodos que emplean valores de alguna otra variable, siendo los más conocidos, los estimadores de razón, regresión y diferencia. La bondad de estos estimadores ha sido evaluada tradicionalmente mediante la distribución que se obtendría observando los valores del estimador considerado, a través de todas las muestras posibles de extraer. Esto es imposible de llevar a cabo en la práctica, pero aproximaciones teóricas permiten realizar estos estudios. Este enfoque para el tratamiento del problema de las inferencias en el muestreo en poblaciones finitas se conoce como *enfoque de diseño*. En el mismo es muy importante el diseño muestral que se emplea, y no se requiere ninguna clase de supuestos acerca de un modelo que refleje la pauta de variabilidad de las unidades que se consideran.

Otro enfoque a considerar es el *basado en modelos* (Särndal *et al.* (1992), Ambrosio (2001), Ambrosio (2006)), en el cual se considera la población finita como una muestra de una población infinita o superpoblación, y se plantea un modelo teórico que refleje el comportamiento de la variable aleatoria que se estudia. En este caso, el problema de la inferencia se resuelve estimando los parámetros del modelo propuesto y obteniendo el valor de interés en la población finita a través de la predicción de los valores de las variables en las unidades no observadas, utilizando el modelo estimado. En este enfoque, la calidad de los resultados depende de la adecuación a la realidad del modelo que se proponga. Esta forma de abordar las inferencias, ha resultado muy útil en cuanto a la posibilidad de incorporación de información auxiliar.

El enfoque conocido como *asistido por modelos* integra a los anteriores, en cuanto a que se seleccionan los estimadores según el enfoque de modelos pero la inferencia se basa sólo en la distribución en el muestreo del estimador. Ambrosio *et al.* (2009) presentan, como un aporte a la mejora de las estadísticas agrarias y medioambientales, la ganancia en la precisión de las estimaciones que puede lograrse empleando muestreo de áreas y modelos que tengan en cuenta la información auxiliar.

En ciertos estudios, las unidades que conforman la población se encuentran distribui-

das en el espacio y por lo tanto, la posición de las mismas se especifica y es relevante. Los datos que así se obtienen se llaman datos espaciales y pueden encontrarse en estudios sociales, económicos, agrícolas, ambientales, geográficos, geológicos, climatológicos, etc. La particularidad de estos datos es la correlación positiva que presentan en el espacio, es decir, mientras más cercanía haya entre dos unidades, más parecidos son los valores de la variable que se estudia. Esta información puede aprovecharse, valiéndose del mencionado enfoque de modelos.

Para el análisis de datos espaciales, se han desarrollado métodos y técnicas específicas que se reúnen en la literatura estadística bajo el término Estadística Espacial o Geoestadística. La misma tiene como objetivo la descripción de la correlación espacial, la estimación de parámetros, la obtención de predicciones de la variable en puntos del espacio en los que no se han realizado observaciones y la provisión de métodos de selección de muestras para satisfacer adecuadamente los objetivos anteriores.

Para la descripción de la correlación espacial se cuenta con métodos e indicadores que son ampliamente conocidos. "El Análisis Exploratorio de Datos Espaciales se podría definir como el grupo de herramientas estadístico-gráficas que describen y visualizan las distribuciones espaciales, identificando localizaciones atípicas, descubriendo formas de asociación (autocorrelación espacial) que, a sus vez pueden ser de carácter global o local, y sugiriendo estructuras en el espacio geográfico (heterogeniedad espacial)" (Chasco Yrigoyen, 2003).

Entre las técnicas más utilizadas se encuentran los índices de asociación de Moran, "box plot", "box map", mapa de contigüidades espaciales, gráfico de retardo espacial, diagrama de dispersión de Moran y mapas LISA.

La modelización de la correlación espacial puede hacerse recurriendo a modelos de correlogramas o semivariograma (Cressie, 1993). Los primeros, expresan la correlación entre unidades en función de la distancia entre las mismas, mientras que los modelos de semivariograma reflejan la variabilidad entre unidades según la distancia que las separa. Entre los modelos teóricos de semivariograma más conocidos, se destacan los modelos exponencial, esférico, de potencia y gaussiano. La adecuada elección de este modelo pro-

porciona buenas consecuencias en las posteriores etapas de los estudios que se realicen.

La estimación del semivariograma se realiza especificando un modelo teórico y estimando sus parámetros por máxima verosimilitud restringida o mediante el cálculo del semivariograma empírico para luego ajustar por mínimos cuadrados ponderados un modelo teórico.

Como se mencionó, el enfoque de modelos basa la estimación de los valores poblacionales en la predicción de los valores de la variable en las unidades no incluidas en la muestra por medio de la utilización de un modelo teórico. Ese modelo puede contener variables auxiliares relacionadas con la variable en estudio, así como información de la ubicación geográfica de las unidades. La adecuada elección del modelo de semivariograma será crucial en esta fase, ya que el mismo contiene la información de la variabilidad espacial. Algunas propuestas y resultados pueden encontrarse en Iglesias (1998), Ambrosio (2001), Ambrosio (2006) y Ambrosio e Iglesias (2014), entre otros.

Surge aquí una cuestión que debe considerarse ya que no siempre se cuenta con información poblacional o de un estudio previo sobre la variabilidad espacial. ¿Cuál será el semivariograma adecuado para utilizar en las inferencias?.

En primer lugar, un planteo encontrado en la bibliografía analizada es obtener un modelo de semivariograma a partir de un estudio anterior. En este caso, cualquiera sea la muestra seleccionada siempre se utiliza ese semivariograma.

Teniendo en cuenta que se trabaja con sólo una muestra podría estimarse el semivariograma a partir de la misma, aunque esta elección agrega una fuente de variación. La estimación podría hacerse identificando un modelo aceptable para la población 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 o identificando el modelo de semivariograma y estimando sus parámetros con los datos proporcionados por la muestra.

Por otra parte, en la fase de selección en el enfoque de diseño, la utilización del se-

mivariograma puede resultar beneficiosa teniendo en cuenta que, en muchas situaciones, se sugiere la aplicación del muestreo sistemático. El semivariograma puede orientar en la elección del período y en la consideración de un muestreo uni o bidimensional, así como en la decisión del empleo de réplicas. Iglesias (1998) y Ambrosio e Iglesias (2000), exhiben aplicaciones a la estimación de usos del suelo con unidades de áreas regulares que presentan correlación espacial, mostrando los beneficios originados por su uso.

Cabe mencionar que muchos de los trabajos que tratan el tema de la variabilidad espacial de los datos, presentan la característica de tener las unidades distribuidas en el espacio en áreas regulares. Esto no ocurre cuando se realizan encuestas sociales o de hogares, en la que las áreas pueden ser, por ejemplo, los radios censales de una determinada ciudad. Cressie (1993) propone una posible solución para este tema particular, que permite el tratamiento de los datos de la misma manera que para látices regulares.

En esta tesis se considera un caso donde los métodos de la Estadística Espacial pueden aportar importantes mejoras, como lo es el estudio de la pobreza en la ciudad de Rosario. En particular, el interés se centra en la estimación del total de hogares con Necesidades Básicas Insatisfechas. Este indicador se calcula sólo en ocasión de los censos de población, pero podría interesar disponer de esta información en otros periodos mediante estudios por muestreo. Las técnicas muestrales que incorporan la información de la variabilidad espacial pueden proporcionar resultados más precisos. Los aportes de este trabajo pueden también resultar válidos para otras clases de estudios socioeconómicos en la ciudad.

Se propone entonces, la estimación del total de hogares con Necesidades Básicas Insatisfechas en la ciudad de Rosario en el año 2001, a partir de muestras de radios censales, considerados en la literatura de Estadística Espacial como látices irregulares.

Es de suponer que radios que se encuentren más cerca, presentan características similares en cuanto a las Necesidades Básicas Insatisfechas, presentando los datos variabilidad espacial, la que debe estudiarse y modelarse para luego aplicar esos resultados en la fase de estimación de características poblacionales de interés a partir de muestras. El uso de la metodología específica que tiene en cuenta la correlación espacial brindará estimadores

más precisos. Para verificar esta mejora, se realizan estudios comparativos a partir de los datos poblacionales correspondientes al Censo Nacional de Población, Hogares y Viviendas del año 2001.

La presente tesis está organizada en siete capítulos. En el Capítulo 2, se presenta en forma sucinta un conjunto de herramientas que pueden emplearse usualmente para detectar y caracterizar la variabilidad espacial en un conjunto de datos.

El Capítulo 3 está dedicado al desarrollo de los fundamentos de la construcción de modelos que describen la variabilidad espacial. Se enumeran los semivariogramas más usuales y se enuncian los procedimientos para la estimación de los parámetros de dichos modelos.

El Capítulo 4 se destina a la presentación de los enfoques para abordar el problema de la inferencia en el muestreo de poblaciones finitas, haciendo hincapié en aquellos procedimientos adecuados para la estimación de totales en poblaciones con variabilidad espacial y de los correspondientes Errores Cuadráticos Medios.

El Capítulo 5 analiza la aplicación de las propuestas de planes de muestreo del enfoque de diseño considerando látices regulares, de acuerdo a lo presentado en el trabajo de Iglesias (1998) ya mencionado. En particular se exhiben los estimadores del total por simple expansión y sus varianicas para el muestreo aleatorio simple, sistemático y estratificado.

El Capítulo 6 se compone de seis secciones destinadas a mostrar la utilidad del empleo de la metodología propuesta en el estudio de pobreza, presentando la estrategia de análisis y estudios comparativos que contrastan resultados con y sin el empleo la información espacial. Los análisis estadísticos utilizados en este capítulo se realizaron con los programas SAS 9.3 y GeoDa 0.9.5-i (Geodata Analysis Software).

La primera de ellas presenta una descripción del problema y la segunda muestra los resultados que se obtienen al aplicar las herramientas de Análisis Exploratorio de Datos Espaciales, las cuales conducen a concluir acerca de la existencia de variabilidad espacial en la población considerada y a la apreciación de aspectos particulares de dicho fenómeno. La misma finaliza con el cálculo del semivariograma empírico para la población.

La tercera sección, exhibe la forma de empleo de la información proporcionada por el semivariograma poblacional ajustado para la estimación, a partir de una muestra aleatoria simple, del total de hogares con Necesidades Básicas Insatisfechas, sus Errores Cuadráticos Medios y eficiencias relativas según el enfoque de modelos. Se compara con modelos que no tienen en cuenta dicha variabilidad.

La cuarta sección, consiste en un estudio comparativo del comportamiento de los estimadores de simple expansión, razón y los que se obtienen utilizando las tres alternativas de semivariograma antes mencionadas, para el muestreo sistemático.

La quinta sección contiene un estudio comparativo de los mismos métodos de estimación pero a partir de todas las muestras aleatorias simples posibles de extraer. Para el caso de los tres métodos de estimación que emplean información provista por el semivariograma, se utilizaron sólo 10000 muestras de la totalidad de muestras posibles, considerando que las esperanzas y los Errores Cuadráticos Medios así obtenidos constituían una aproximación apta para las comparaciones a realizar. Esta sección y la anterior, hacen referencia a una cuestión cuya discusión no se ha encontrado en la literatura y es la de la elección del semivariograma a emplear entre las tres alternativas propuestas.

La última sección del capítulo presenta el caso particular de la estimación del total de hogares con Necesidades Básicas Insatisfechas, pero considerando látices regulares artificiales. Para ello, dado que naturalmente las unidades de muestreo, radios censales, son látices irregulares, se agrupan los mismos en unidades cuadradas previamente definidas, en las que se ha dividido la ciudad de Rosario, y se muestra que el uso de la información proporcionada por los modelos de variabilidad espacial puede orientar en el período del muestreo sistemático.

Por último, el Capítulo 7, se dedica a la presentación de las consideraciones finales del estudio y las propuestas para la continuación del mismo, con diferentes líneas de investigación.

El trabajo de tesis tiene como objetivos:

- Revisión metodológica de las propuestas para emplear información geográfica en los

proceso de selección y estimación en muestreo de poblaciones finitas.

- Realización de propuestas para la estimación de características socioeconómicas de la ciudad de Rosario.
- Presentación de los resultados del empleo de las propuestas metodológicas.
- Realización de estudios comparativos para evaluar la eficiencia de los métodos de estimación.

Análisis exploratorio de datos espaciales

2.1. Introducción

El objetivo primordial de la presente investigación es el análisis de procedimientos de estimación en muestreo de poblaciones finitas que presentan variabilidad espacial, con la finalidad de mostrar que el uso de dicha información produce considerables ventajas. En instancias previas a la aplicación de estos procedimientos, se debe evaluar la existencia de la variabilidad espacial, para lo cual se realiza un primer estudio consistente en la aplicación de herramientas que permitan describir y caracterizar el fenómeno mencionado. En el contexto de la Estadística Espacial estas técnicas se agrupa bajo el título de Análisis Exploratorio de Datos Espaciales y se las refiere como AEDE.

En este capítulo se presenta una reseña de métodos descriptivos del AEDE, desde una perspectiva de látices, con sus fundamentos y particularidades.

2.2. Conceptos generales

La Estadística Espacial hace referencia a un conjunto de técnicas apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios de una región. Tiene por objeto la exploración, descripción, visualización y análisis de datos considerando sus características de distribución en el espacio.

Un grupo de observaciones tomadas en puntos cuya posición se especifica en el espacio constituyen un conjunto de datos espaciales. Para la descripción y el análisis es relevante conocer la posición o referencia geográfica asociada a cada unidad. Los datos espaciales

pueden ser representados en un mapa mediante puntos, líneas o polígonos. Los puntos están definidos por sus coordenadas de latitud y longitud, pudiendo estar asociados por ejemplo a individuos, establecimientos, delitos cometidos, accidentes, u otra clase de eventos. Las líneas son objetos que cubren una distancia dada y unen varios puntos. Éste sería el caso de líneas telefónicas, calles de una ciudad, ríos, etc. Los polígonos son áreas delimitadas por líneas, como pueden ser los países, departamentos, áreas municipales, radios censales, etc.

El problema que se plantea en el análisis de estos datos es que las unidades tomadas en una región específica no son independientes. Así, se supone que si se encuentra una determinada unidad en un punto de este área es más fácil (o inversamente más difícil) encontrar unidades semejantes en puntos próximos a éste que en puntos alejados. Es decir, la variable en estudio presenta mayor correlación cuanta más cercanía hay entre las unidades.

De manera formal, sea "Y" la variable aleatoria de interés, y_i el valor observado en la unidad i ubicada en el punto de coordenadas geográficas $\mathbf{s}_i$ de la región S y sea Y_i una realización espacial de la variable aleatoria. El comportamiento de la variable en el espacio se puede expresar de la siguiente forma:

$$\{Y_i; \mathbf{s}_i \in S \subset R^2\}.$$

Desde este punto de vista, el valor de la variable en las unidades espaciales que componen un mapa dado, constituye sólo una de las infinitas realizaciones posibles de un proceso general.

El conjunto no vacío S, denominado dominio, puede ser clasificado de dos maneras:

- Teniendo en cuenta las posibles localizaciones de los elementos que lo forman, S puede ser continuo o discreto. S es un conjunto continuo cuando la variable puede ser observada en cualquier sitio definido sobre la superficie en estudio (conjunto infinito no numerable) y es discreto cuando la variable corresponde a áreas definidas sobre la superficie estudiada (conjunto infinito numerable).
- Teniendo en cuenta la naturaleza de los sitios donde se observa la variable, S puede

ser fijo o aleatorio. S es fijo cuando los sitios son determinados por el investigador a priori y es aleatorio si la ubicación del evento no está definida e interesa estudiarla.

Dependiendo de estas características, se pueden diferenciar tres tipos de datos:

- Procesos puntuales (o mapas de puntos): Consisten en un número finito de localizaciones observadas en una región determinada, donde su selección no depende del investigador. Por lo cual el dominio S es discreto y aleatorio. En este caso, las localizaciones (y no las mediciones) son las variables de interés y por lo tanto sólo interesa la ubicación de ocurrencia del evento. En este tipo de datos se trata de valorar si los eventos tienden a exhibir algún patrón sistemático.
- Datos en látices: Se trata de observaciones de un proceso aleatorio sobre un conjunto numerable de regiones espaciales, regulares o irregulares. La selección de los sitios de medición depende del investigador, por lo que S es discreto y fijo. Las ubicaciones concretas suelen referirse al centroide de la región representado por sus coordenadas cartográficas. El objetivo del análisis en estos casos es descubrir y modelar el patrón espacial existente.
- Datos geoestadísticos (o datos espacialmente continuos): Los datos espacialmente continuos son mediciones tomadas en puntos fijos seleccionados por el investigador pero definidos en cualquier lugar del espacio, por lo que sus localizaciones definen una superficie espacialmente continua, siendo S continuo y fijo. La variable medida puede ser continua o discreta. El análisis de datos geoestadísticos contempla la modelización del patrón de variabilidad y la predicción en puntos donde no se ha muestreado.

Independientemente del tipo de datos, las unidades distribuidas en el espacio tienen características muy importantes. Una de ellas es la georreferenciación, en donde el espacio geográfico tiene una naturaleza georreferenciada que exige conocer la posición absoluta o relativa donde se producen los fenómenos analizados. La otra característica del espacio geográfico es la multidireccionalidad de las relaciones que sobre él se establecen, entendiendo que una región puede no sólo estar afectada por otra región contigua a ella sino

por otras muchas que la rodean.

Estas características producen los llamados efectos espaciales: dependencia o autocorrelación espacial y heterogeneidad espacial. Estos efectos son los que impiden que los métodos usuales de la estadística sean una buena herramienta para su modelación.

El efecto de **dependencia espacial** se define como la relación funcional existente entre los valores que adopta una variable en una unidad del espacio y en unidades vecinas. Es decir, el valor que toma una variable en una unidad, no sólo está explicado por condiciones internas, sino también por los valores que toma esa misma variable en otra unidad del espacio. Esta dependencia puede ser expresada según la primera ley de la geografía, en la cual todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes (Tobler, 1970).

Si existe dependencia espacial, entonces la variable se encuentra espacialmente autocorrelacionada cuando los valores observados en una unidad determinada dependen tanto de factores externos como de los valores de la misma variable en diferentes unidades.

Este efecto de autocorrelación espacial puede ser de signo positivo o negativo, así como nulo. Se entiende por autocorrelación espacial positiva el fenómeno de asociación entre valores similares de una variable en localizaciones cercanas; es decir, cuando la presencia de un fenómeno en una unidad hace que ese fenómeno se extienda hacia las unidades que lo rodean favoreciendo la concentración del mismo. Éste es el caso del llamado efecto contagio o desbordamiento. Por el contrario, existe autocorrelación espacial negativa en un espacio cuando las unidades con los valores altos de una variable se encuentran rodeadas por unidades con valores bajos de la misma, y viceversa. Por último, se produce ausencia de autocorrelación espacial en una variable geográfica cuando ésta se distribuye de forma aleatoria sobre el espacio, de manera que no se localizan valores de la variable parecidos (o disimiles) en un entorno geográfico próximo.

Para cuantificar la estructura de dependencia espacial en un conjunto de datos es necesario definir la relación espacial existente entre ellos, es decir, se debe definir quiénes se

consideran vecinos de una determinada unidad (Haining, 2003). Existen criterios basados en contigüidad y en distancia para definir esta vecindad:

- Distancia: Las áreas de un mapa pueden transformarse en puntos y viceversa, y como punto representativo de la misma se suele usar el punto central o centroide. En este caso, se utiliza un criterio de distancia para definir a los vecinos de la unidad i, tal que cada centroide que se encuentre dentro de una distancia máxima h_{max} al centroide de i se considera vecino. Teniendo en cuenta la distancia euclídea entre los centroides, las unidades i y i' son vecinos si $0 < h < h_{max}$, siendo h_{max} un valor prefijado de distancia máxima y h la distancia euclídea entre i e i'.

  Un problema con la elección de vecinos por medio de la distancia es la existencia de unidades aisladas que pueden no presentar algún vecino para un radio determinado. Para salvar este problema se suele determinar un radio h_{max} de amplitud tal que asegure que cada unidad tenga al menos un vecino.
- Contigüidad: Refleja la posición relativa de una unidad espacial hacia otras unidades del espacio. Las vecindades suelen establecerse a partir de un mapa; dos unidades se consideran vecinas si tienen una frontera en común. De aquí se desprenden varios tipos de contigüidad dependiendo de la frontera que compartan: contigüidad torre (aristas en común), alfil (vértices en común) o reina (vértices y aristas en común).
- D vecinos más cercanos: En este caso, considerando la distancia entre las unidades, se selecciona a los D vecinos más cercanos de cada punto. La ventaja de este criterio es que todas las unidades poseen la misma cantidad de vecinos evitando el problema de unidades aisladas.

Los tipos de vecindad presentados reflejan vecindad de primer orden, pero es posible considerar contigüidad de segundo orden (cuando se tiene en cuenta la influencia de vecinos de vecinos), de tercer orden, etc.

Una vez definidos los criterios de vecindad se crea la matriz de pesos espaciales W, también llamada matriz de conectividad. Ésta es una matriz cuadrada no estocástica de

orden nxn, siendo n el total observaciones en estudio, que recoge el efecto de la unidad i sobre la unidad i' a través de un peso o una ponderación $w_{ii'}$. Los pesos $w_{ii'}$ de la matriz deben ser positivos y finitos, pero no hay una definición única para ellos.

La formulación más simple es una matriz de contigüidad binaria, es decir:

$$w_{ii'} = \begin{cases} 1 & si\ i\ y\ i'\ son\ vecino \\ 0 & en\ otro\ caso \end{cases} .$$

Otras formas de definir las ponderaciones puede ser representando la cercanía entre las unidades espaciales: $w_{ii'} = h^{-2}$ o $w_{ii'} = \dfrac{1}{1+h}$; donde h es la distancia euclídea entre las unidades i e i' o bien, teniendo en cuenta algún dato referente a la naturaleza del problema que se pretende estudiar.

Una vez elegidos los pesos, se suele trabajar con alguna transformación de la matriz. La transformación más utilizada es la normalización por fila, donde los nuevos pesos obtenidos resultan:

$$w^*_{ii'} = \frac{w_{ii'}}{\sum_{i'=1}^{n} w_{ii'}}$$

y verifican que $\sum_{i'=1}^{n} w^*_{ii'} = 1$.

Esta matriz W^* pondera por igual la influencia total que recibe cada unidad de sus vecinos, con independencia del número total de vecinos de cada una de ellas. Al multiplicar esta nueva matriz por el vector de observaciones de una variable "Y" en las distintas unidades, la matriz producto "Y^*" = W^*"Y" representa una nueva variable igual a la media de las observaciones en las unidades vecinas.

La matriz de conectividad hace posible la conexión entre el valor de una variable en un punto del espacio geográfico y las observaciones de dicha variable en los puntos vecinos a través del llamado retardo espacial, que actúa como el retardo temporal en series de tiempo. En el contexto temporal, el operador de retardo B^j provoca un desplazamiento en el eje del tiempo tantos períodos como indique la potencia j. En el espacio no se puede considerar sólo una dirección de desplazamiento, es decir, se debe tener en cuenta las

diferentes vecindades de una unidad, según el criterio de contigüidad utilizado. Generalmente se puede encontrar una gran cantidad de direcciones posibles. Este problema se soluciona si se considera la suma de los pesos de todos los valores pertenecientes a una clase de contigüidad dada, en vez de tomar cada una de ellas individualmente.

Por lo tanto, el retardo espacial, consiste en un promedio ponderado de los valores que toma una variable en el subconjunto de observaciones vecinas a una dada. Los términos de esta suma se obtienen multiplicando las observaciones en cuestión, y_i, por sus correspondientes pesos de la matriz de conectividad W del modo siguiente:

$$B^c(y_i) = \sum_{i'=1}^{\#S_i} w_{ii'} y_{i'},$$

donde,

B^c es el operador retardo asociado con el criterio de contigüidad c,

S_i es el conjunto de unidades vecinas a i según el criterio de contigüidad c,

$w_{ii'}$ es el elemento de la matriz de conectividad que recoge la relación espacial entre las unidades i e i'.

En el caso de que la matriz W fuera normalizada por filas, el retardo espacial representa un suavizado de los valores vecinos, ya que la suma de todos los pesos de una determinada fila debe ser igual a la unidad. De esta forma, cada elemento del retardo espacial es igual a un promedio ponderado de los valores de la variable "Y" en el subgrupo de observaciones vecinas a ella, S_i, dado que $w_{ii'} = 0, \forall i' \notin S_i$.

El otro efecto espacial, la **heterogeneidad espacial**, se refiere a la ausencia de estabilidad en el espacio de la variable en estudio. Existen dos aspectos distintos a tener en cuenta aquí: inestabilidad estructural y heteroscedasticidad. La inestabilidad estructural se refiere a la falta de estabilidad en el espacio del comportamiento de la variable bajo estudio, mientras que la heteroscedasticidad se atribuye a la ausencia de estabilidad en la dispersión de un fenómeno, como sucede muchas veces con los residuos de una regresión.

A diferencia del caso de la dependencia espacial, los problemas causados por la heterogeneidad pueden ser, mayoritariamente, resueltos por las técnicas de estadística clásica. Sin embargo, en muchos casos, tener en cuenta la estructura espacial inherente en los

datos puede llevar a procedimientos más eficientes. Además, se puede hallar dependencia en combinación con heterogeneidad, en donde, el problema de distinguir entre ellas es altamente complejo.

A continuación se presenta un conjunto de métodos gráficos utilizados para detectar la variabilidad espacial de los datos.

2.3. Métodos gráficos del Análisis Exploratorio de Datos Espaciales

El Análisis Exploratorio de Datos Espaciales (AEDE) se define como el conjunto de técnicas que describen y visualizan las distribuciones espaciales, identifican ubicaciones atípicas ("outliers" espaciales), descubren formas de asociación espacial (autocorrelación espacial) de carácter global o local, y sugieren estructuras espaciales u otras formas de heterogeneidad espacial (Chasco Yrigoyen, 2003).

Según Cressie (1993), el AEDE se puede interpretar desde dos perspectivas diferentes, ya sea desde un análisis geoestadístico o bien por la econometría espacial. El análisis geoestadístico se utiliza generalmente en las ciencias medioambientales como la hidrología, la física, la teledetección o la geología, y se enfoca en una muestra de datos puntuales procedentes de distribuciones geográficas continuas (por ejemplo, humedad de la tierra, precipitaciones, procesamiento de imágenes satelitales, altura del mar, etc.). Por otra parte, la econometría espacial analiza localizaciones geográficas discretas de puntos o polígonos (provincias, municipios, radios censales, etc.), denominada perspectiva retícular o de látices. La misma se utiliza con mayor frecuencia en los estudios socioeconómicos (cantidad de lactantes con muerte súbita, votantes, etc.).

El AEDE emplea los métodos más sencillos de estadística descriptiva (mapas de cuartiles, percentiles, etc.) en combinación con herramientas más complejas para el estudio de los efectos espaciales de asociación y heterogeneidad.

Estos análisis se utilizan para identificar relaciones entre variables que se distribuyen

geográficamente cuando no se conoce la naturaleza de las mismas, considerando las características propias de estos datos: georreferenciación y multidireccionalidad.

A continuación se presentan los métodos gráficos más usuales del AEDE agrupados según si el objetivo es la representación de distribución espacial, dependencia espacial o heterogeneidad espacial, para una perspectiva de látices.

Cuadro 2.1: Técnicas gráficas del AEDE, según la perspectiva de látices

Dsitribución espacial	Mapa de percentiles	
	"Box plot"	
	"Box map"	
Dependencia espacial	Global	Índice de asociación espacial global de Moran
		Diagrama de dispersión de Moran
		Gráfico del retardo espacial
	Local	Índice de asociación espacial local de Moran
		Mapas LISA
		Diagrama de caja LISA
	Bivariante o multivariante	Índice de asociación espacial global y diagrama de dispersión de Moran
		Índice locales de Moran y gráficos LISA
Heterogeneidad espacial	Histograma de frecuencia	
	Diagrama de dispersión	

2.3.1. Representación de distribuciones espaciales

La distribución espacial de un fenómeno permite encontrar las condiciones y características de la ubicación de dicho fenómeno. En el AEDE, la visualización de la distribución espacial se realiza a través de métodos gráficos de la estadística clásica, destacándose el mapa de percentiles, el diagrama de caja ("box plot") y el correspondiente "box map". Estas herramientas proporcionan una idea preliminar de la distribución espacial de los datos y pueden ser la primera aproximación a la detección de cierto grado de asociación espacial entre los valores de la variable.

El **mapa de percentiles** permite la identificación de los valores extremos de una distribución espacial, para lo cual se agrupan los valores de la variable de forma tal que estos valores atípicos queden destacados. Una vez ordenados los valores de la variable de menor a mayor, usualmente se crean seis grupos correspondientes a los percentiles [0,1),

[1,10), [10,50), [50,90), [90,99) y [99,100). Estos percentiles se trasladan a un mapa, en el cual se diferencian los intervalos con colores en tonalidad "ascendente"de acuerdo a los valores de los intervalos definidos; así los colores más claros muestran los valores más bajos de la variable estudiada y los colores más oscuros representan los valores más altos.

El **"box map"** es la versión cartográfica del "box plot" y muestra la ubicación de los cuartiles representados en el diagrama de caja, sobre el mapa. Como criterio general agrupa bajo el mismo color aquellas áreas que se encuentran dentro del mismo intervalo intercuartílico y por lo tanto, hay seis categorías: cuatro correspondientes a cada cuartil y dos correspondientes a los valores extremos de las cotas inferior y superior. Este método permite identificar los valores atípicos en el mapa y sugiere criterios de agrupación en el espacio, características que en el "box plot"no se pueden detectar ya que sólo se descubren medidas de dispersión y simetría de la distribución de los datos y se individualizan valores atípicos.

2.3.2. Representación de la dependencia espacial

La dependencia o asociación espacial se define como la relación existente entre los valores que adopta una variable en una unidad o región del espacio y en unidades vecinas. Es decir, cuando los valores observados de una variable en una unidad dependen de los observados en unidades vecinas existe dependencia espacial.

Las técnicas para la caracterización de la asociación espacial están basadas en indicadores globales, locales y multivariantes. Los primeros analizan las unidades de forma conjunta y son incapaces de detectar aquellas regiones donde se concentran valores atípicos, pero permiten visualizar una superficie de tendencia o realizar un suavizado. Las técnicas del AEDE que analizan esta cualidad son herramientas de representación cartográfica para las cuales no es fundamental el mapa, sino las técnicas básicas de representación gráfica. Entre los procedimientos más utilizados se mencionan el índice de asociación global de Moran junto con su diagrama de dispersión y la comparación de los retardos espaciales. Estas medidas permiten una descripción general de la dependencia a través

de toda el área de estudio.

Los indicadores locales se utilizan para identificar patrones espaciales en áreas de menor tamaño: unidades muy distintas a sus vecinos en el mapa ("outliers" espaciales) o unidades agrupadas con valores altos o bajos de una variable (conglomerados). Estos métodos exploratorios analizan la existencia de concentraciones de observaciones cuyo valor se encuentra alejado de la tendencia general. Las técnicas de asociación espacial local son: los índices de asociación espacial local de Moran, los mapas LISA ("Local Indicator of Spatial Asociation") y el diagrama de caja LISA.

Por último, entre las técnicas de asociación espacial que consideran más de una variable, se utilizan los diagramas de dispersión multivariantes de Moran y gráficos LISA multivariante.

Dependencia espacial global

Los índices globales de autocorrelación espacial permiten verificar si se cumple la hipótesis de que una variable se encuentra distribuida en forma aleatoria en el espacio, o si por el contrario, existe asociación significativa entre unidades vecinas. Es decir, permiten probar la hipótesis nula de aleatoriedad espacial.

El **índice de asociación global de Moran** se utiliza para verificar si las unidades se distribuyen aleatoriamente en el espacio teniendo en cuenta la variable bajo estudio y proporciona una medida resumen de la intensidad de la autocorrelación de las unidades. Este índice se define como (Moran, 1950):

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{i'=1}^{n} w_{ii'}} \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} w_{ii'}(y_i - \bar{y})(y_{i'} - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \forall i \neq i',$$

donde,

n es el número de unidades espaciales en la región S,

y_i es el valor de la variable "Y" en la unidad i ubicada en el punto de coordenadas $\mathbf{s}_i$,

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ es la media de la variable,

$w_{ii'}$ es el elemento de la matriz de conectividad que recoge la relación de vecindad entre las unidades i e i' ($i = 1, 2, ..., n$ y $i' = 1, 2, ..., n$).

El índice de asociación global de Moran resume la intensidad y dirección de la dependencia entre los valores de una variable observados en distintas unidades o regiones del espacio, calculando los productos cruzados entre los valores de pares de unidades y ponderando por una medida de la relación de vecindad entre las unidades de cada par. Por lo tanto, I puede considerarse como una medida de correlación de cada y_i con el resto de las regiones con las que se encuentra vinculada. Al igual que el índice de correlación de Pearson varía entre -1 y 1 y $E[I] = -\dfrac{1}{n-1}$ bajo la hipótesis nula de aleatoriedad espacial.

Un coeficiente I mayor que su valor esperado indica autocorrelación espacial positiva, mientras que un valor inferior a la media pone de manifiesto la existencia de autocorrelación espacial negativa. Un valor cercano a $E[I]$ (la cual tiende a 0 cuando n crece) indica ausencia de autocorrelación.

Para probar la significación del índice I y así comprobar la hipótesis de no autocorrelación espacial se puede utilizar un test de hipótesis basado en supuestos de normalidad o en distribuciones experimentales.

Bajo la hipótesis nula de que no existe autocorrelación espacial y si $y_i \sim N(\mu, \sigma^2)$ o si n es suficientemente grande, la estadística $Z = \dfrac{I - E[I]}{\sqrt{Var[I]}}$ sigue una distribución normal estándar donde:

$$E[I] = -\frac{1}{n-1}$$

$$Var[I] = \frac{n^2 \sum_{i=1}^{n} \sum_{\substack{i'=1 \\ i' \neq i}}^{n} (w_{ii'} - w_{i'i})^2 - n \sum_{i=1}^{n} (\sum_{i'=1}^{n} w_{ii'} + \sum_{i'=1}^{n} w_{i'i})^2 + 3 \left[\sum_{i=1}^{n} \sum_{\substack{i'=1 \\ i' \neq i}}^{n} w_{ii'} \right]^2}{(n^2 - 1) \left[\sum_{i=1}^{n} \sum_{\substack{i'=1 \\ i' \neq i}}^{n} w_{ii'} \right]^2} - \frac{1}{(n-1)^2}.$$

Cuando no se cumple el supuesto de normalidad de la variable en estudio se recurre a los test permutacionales en los que se encuentran las $n!$ posibles configuraciones de las unidades asumiendo que sus valores son aleatorios y sobre cada una de ellas se calcula el valor de I, para luego calcular la probabilidad asociada a la hipótesis de aleatoriedad.

Comúnmente se puede utilizar un test basado en el Método de Montecarlo, que consiste en la realización de un test permutacional pero sólo considerando un subconjunto de configuraciones, y por lo tanto es útil cuando n es muy grande. Este tipo de técnicas obliga la utilización de métodos computacionales ya que para su aplicación intervienen gran cantidad de cálculos. Los "softwares" suelen utilizar 999 permutaciones.

Un instrumento gráfico habitual en el análisis de autocorrelación espacial es el **diagrama de dispersión de Moran**. Se trata de un diagrama de dispersión que representa en la abscisa la variable previamente estandarizada ($y_i^* = \frac{y_i - \bar{y}}{s}$ con $s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{(n-1)}$) y en la ordenada los retardos espaciales de dicha variable estandarizada ($B^c(y_i^*) = \sum_{i'=1}^{\#S_i} w_{ii'} y_{i'}^*$).

Este diagrama permite identificar los diferentes tipos de asociación: positiva, negativa o nula. Las categorías de asociación espacial positiva se corresponden a los cuadrantes I y III, mientras que la negativa se encuentra en los cuadrantes II y IV. Si la nube de puntos está dispersa en los cuatro cuadrantes es indicio de ausencia de autocorrelación espacial.

Otra forma de detectar la presencia de autocorrelación espacial es mediante la **comparación de los valores observados versus sus retardos espaciales**, ya sea mediante un gráfico o bien numéricamente. Como se mencionó, el retardo espacial se define como un promedio ponderado de los valores de una variable en las unidades vecinas a una dada. Este método compara el valor observado en una determinada unidad con el retardo espacial y por lo tanto la obtención de valores similares confirma que dicha unidad es similar a sus vecinas, indicando autocorrección espacial y sugiriendo posibles agrupaciones; mientras que valores muy disímiles se producen cuando la unidad es muy distinta a las unidades vecinas, la cual representa un "outlier" espacial.

Dependencia espacial local

El índice de autocorrelación global de Moran no es capaz de detectar ciertas estructuras locales de asociación, ya que supone que el grado de autocorrelación espacial es igual para todas las unidades espaciales. Existe la posibilidad de que no se detecte asociación

espacial global en la distribución de una variable, pero que existan pequeños conglomerados en los que dicha variable presente una concentración importante; o bien, habiéndose detectada dependencia a nivel global no todas las zonas contribuyan con el mismo peso en este indicador. Por este motivo se definen técnicas para detectar la asociación espacial local, la cual puede ser definida como una concentración, en un lugar del espacio global analizado, de valores especialmente altos o bajos de una variable en comparación con el valor medio esperado. Para estudiar este fenómeno se recurre al índice de asociación espacial local de Moran, el cual es capaz de detectar la contribución de cada región al indicador de asociación espacial global y posibles valores atípicos. También se presentan los mapas LISA de conglomerados y de significación y el diagrama de caja LISA para los índices locales.

El **índice de asociación espacial local de Moran**, para una unidad i, se define como (Anselin, 1995):

$$I_i = (y_i - \bar{y}) \sum_{i'=1}^{n} w_{ii'}(y_{i'} - \bar{y}),$$

donde,

n es el número de unidades espaciales en la región S,

y_i es el valor de la variable "Y" en la unidad i ubicada en el punto de coordenadas $\mathbf{s}_i$,

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ es la media de la variable,

$w_{ii'}$ es el elemento de la matriz de conectividad que recoge la relación de vecindad entre las unidades i e i' ($i = 1, 2, ..., n$ y $i' = 1, 2, ..., n$).

Al igual que el índice de asociación global de Moran, cada I_i varía entre -1 y 1. Un valor positivo de I_i indica que la unidad i tiene vecinos con valores similares (altos o bajos), entonces la unidad i forma parte de un conglomerado de unidades "parecidas". En cambio, un valor negativo para I_i es indicio de que la unidad i está rodeada de unidades diferentes y por lo tanto dicha unidad i es considerada un atípico espacial. Como resultado, el índice local de Moran identifica unidades donde valores altos o bajos se agrupan espacialmente, así como también unidades con valores muy distintos a los de las áreas vecinas.

En conclusión, este índice puede ayudar a reconocer cinco tipos de conglomerados espaciales:

- Alto-Alto (puntos calientes o "hot spots"): Unidades espaciales con valores altos de la variable en estudio rodeadas significativamente de unidades espaciales también con valores altos.
- Bajo-Bajo: Unidades espaciales con valores bajos rodeadas significativamente por unidades que también tienen valores bajos respecto a la variable de interés.
- Bajo-Alto: Presencia de unidades con valores bajos en la variable rodeadas significativamente por vecinos con valores altos.
- Alto-Bajo: Presencia de unidades con valores altos en la variable en estudio rodeadas significativamente por vecinos con valores bajos.
- Relación no significativa: Presencia de unidades donde el valor de la variable de interés no se relaciona significativamente con los valores que presentan sus vecinos.

Para cada valor de I_i es posible realizar un test para evaluar la significación estadística de la hipótesis nula de aleatoriedad de la distribución de valores en una región geográfica. Al igual que el test global de Moran, esta hipótesis se prueba con un test basado en la normalidad o bien con tests permutacionales considerando:

$$E[I_i] = -\frac{\sum_{i'=1}^{n} w_{ii'}}{n-1}$$

$$Var[I_i] = \frac{(n - \frac{m_4}{m_2^2})(n-2)\sum_{\substack{i'=1 \\ i' \neq i}}^{n} w_{ii'}^2 + 2(2\frac{m_4}{m_2^2} - n)\sum_{\substack{k=1 \\ k \neq i}}^{n}\sum_{\substack{h=1 \\ h \neq i}}^{n} w_{ik}w_{ih}}{(n-1)(n-2)} - \frac{\left[\sum_{i'=1}^{n} w_{ii'}\right]^2}{(n-1)^2}$$

donde,
$m_4 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4}{n}$, momento de orden 4,
$m_2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$, momento de orden 2.

En los **mapas LISA** se representan las unidades que tienen valores estadísticamente significativos en los indices de asociación local de Moran (Anselin, 1995). Hay dos clases

de mapas: los mapas de significación y los mapas de conglomerados. El mapa de significación LISA permite visualizar las unidades en las que el índice de asociación espacial local de Moran resulta significativo para diversos niveles de significación. Los mapas de conglomerados presentan los tipos de conglomerados espaciales detectados por los índices locales. En este gráfico se destacan con colores "calientes/fríos" aquellas observaciones en torno a las cuales se produce una concentración estadísticamente significativa de valores altos/bajos de una variable, poniendo así de manifiesto la presencia de atípicos espaciales.

Una característica de los índices locales I_i es que la suma de todos ellos es proporcional al estadístico global de Moran I, por lo cual, si se detectan los valores extremos en la distribución de los índices locales se puede determinar cuáles son las unidades que aportan más al índice global. Esto se puede observar tanto en el diagrama de dispersión de Moran, poniendo atención en los puntos muy alejados del resto, como en un **diagrama de caja LISA** para la distribución de los I_i. En este último, si los valores de I_i sobrepasan las cotas superior o inferior se considera a la unidad i como una unidad que muestra una valoración atípica de la variable en estudio. Los puntos atípicos en el diagrama de dispersión de Moran coincidirán con los valores más extremos en el "box plot".

Dependencia espacial bivariantes y multivariante

Las técnicas de representación de la asociación espacial multivariante se utilizan cuando en el estudio se consideran otras variables relacionada con la variable de interés.

El concepto de correlación espacial bivariante hace referencia al grado de semejanza entre el valor de una variable "X" observada en cierta unidad y los valores de la variable de interés "Y" observada en unidades vecinas. El análisis de asociación espacial se realiza, al igual que en el caso univariado, a través del índice de asociación de Moran y su diagrama de dispersión (análisis global) y de los índices locales de Moran y los mapas LISA (análisis local).

El **índice de asociación espacial global bivariante de Moran** está dado por:

$$I^{XY} = \frac{\sum_{i=1}^{n}\sum_{i'=1}^{n} w_{ii'}(x_i - \bar{x})(y_{i'} - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \forall i \neq i',$$

donde,

n es el número de unidades espaciales en la región S,

x_i es el valor de la variable "X" en la unidad i ubicada en el punto de coordenadas $\mathbf{s}_i$,

$y_{i'}$ es el valor de la variable "Y" en la unidad i' ubicada en el punto de coordenadas $\mathbf{s}_{i'}$,

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ son las medias de las variables,

$w_{ii'}$ es el elemento de la matriz de conectividad que recoge la relación de vecindad entre las unidades i e i' ($i = 1, 2, ..., n$ y $i' = 1, 2, ..., n$).

Este índice intenta mostrar la relación que existe en cada unidad, entre los valores de una variable y el valor medio de otra variable en el entorno de dicha unidad. Para probar esta relación se recurre a los test permutacionales, estandarizando las variables previamente.

De la misma manera que en el análisis de autocorrelación univariante se puede obtener el **diagrama de dispersión bivariante de Moran**. Se trata de un diagrama de dispersión en el que se representa en la ordenada el retardo espacial de la variable de interés y en la abscisa la otra variable. La pendiente de la recta de regresión muestra el grado de relación lineal existente entre estas variables, coincidiendo con el índice global de Moran bivariante.

El **índice local bivariante de Moran** tiene en cuenta, para cada unidad, los valores de la variable "X" y el retardo espacial de la variable de interés "Y", midiendo el grado de asociación de la variable "X" en una determinada unidad y la variable "Y" en los correspondientes vecinos. Está dado por:

$$I_i^{XY} = (x_i - \bar{x}) \sum_{i'=1}^{n} w_{ii'}(y_{i'} - \bar{y}) \qquad \forall i \neq i',$$

donde,

n es el número de unidades espaciales en la región S que se estudia,

x_i es el valor de la variable "X" en la unidad i ubicada en el punto de coordenadas $\mathbf{s}_i$,

$y_{i'}$ es el valor de la variable "Y" en la unidad i' ubicada en el punto de coordenadas $\mathbf{s}_{i'}$,

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ son las medias de las variables,

$w_{ii'}$ es el elemento de la matriz de conectividad que recoge la relación de vecindad entre las unidades i e i' ($i = 1, 2, ..., n$ y $i' = 1, 2, ..., n$).

Al igual que en el análisis univariado la significación de estos índices se prueba con test permutacionales y las regiones significativas se muestran a distintos niveles de significación en el **mapa de significación LISA bivariante**.

El **mapa de conglomerados LISA bivariante** hace posible determinar la naturaleza de la asociación espacial diferenciando cuatro grupos: dos categorías para asociación positiva o conglomerados espaciales (Alto-Alto, Bajo-Bajo) y dos categorías para asociación negativa u "outliers" espaciales (Alto-Bajo, Bajo-Alto). De esta manera, es posible identificar tanto agrupamientos como valores atípicos espaciales teniendo en cuenta simultáneamente dos variables.

El **diagrama de dispersión multivariante de Moran** es una técnica de exploración de asociación espacial multivariante derivada del clásico índice de asociación espacial global de Moran. Este diagrama multivariante, implantado en un entorno dinámico, permite comparar el comportamiento del fenómeno de asociación espacial en varios indicadores.

2.3.3. Representación de la heterogeneidad espacial

La heterogeneidad espacial se refiere a la ausencia de estabilidad en el espacio de la variable en estudio. Para identificarla, Anselin (1999) sugiere la utilización del histograma de frecuencias y el diagrama de dispersión para la variable en estudio y otras variables.

El **histograma de frecuencias** para una variable es la aproximación empírica a la

distribución teórica de dicha variable. En este tipo de gráfico también es posible obtener la aproximación a la distribución empírica de la variable en subregiones de la zona de estudio. Para ello, se seleccionan las barras del histograma que corresponden a valores similares de la variable en estudio, de forma tal que, al ubicar estas unidades en el mapa, permiten seleccionar un número adecuado de las mismas, identificando así las estructuras de comportamiento de la variable. Por ejemplo, si se eligen las barras correspondientes a los valores observados más chicos de la variable en cuestión, los mismos también son mostrados en el mapa. La existencia de heterogeneidad espacial implica que estos valores estén sobre una determinada región caracterizando así una estructura con bajos valores de la variable. De esta forma se identifican las distintas regiones donde la variable en cuestión se comporta diferente.

El **diagrama de dispersión** se basa en un sistema de dos ejes cartesianos, en los que se representan los valores de dos variables cuya relación estadística se quiera analizar. Además de los valores representados, en dicho diagrama se muestra la pendiente estimada por el método de los mínimos cuadrados ordinarios. En base a éste se puede detectar la presencia de heterogeneidad espacial comprobando la existencia de coeficientes de regresión diferentes entre la variable de interés versus otras variables en las estructuras detectadas mediante el método anterior. Se lleva a cabo construyendo gráficos de dispersión, uno para cada sector identificado y se analiza si hay variación entre los coeficientes de regresión estimados (Anselin, 1999).

Modelos de variabilidad y correlación espacial

3.1. Introducción

El uso de información auxiliar en planes de muestreo en poblaciones que presentan variabilidad espacial puede brindar importantes mejoras.

Para dicho aprovechamiento, es necesario expresar mediante algún modelo matemático, la forma de la relación existente entre los valores de la variable y la distancia que separa las correspondientes unidades. Con este fin, se proponen los modelos de correlograma y de semivariograma, cuya estimación resulta decisiva en la mejora del diseño muestral.

En este capítulo se presentan los aspectos fundamentales de los modelos de variabilidad y corelación espacial, los métodos propuestos para la estimación de los semivariogramas, así como los criterios para la selección del más adecuado.

3.2. Definiciones

Sea "Y" la variable de interés y sea y_i el valor observado de "Y" en el punto de coordenadas geográficas $\mathbf{s}_i$ de la región S. La estructura de la correlación espacial se modela considerando a Y_i como una realización espacial de una variable aleatoria. Si el comportamiento de "Y" responde a un proceso estocástico $\{Y_i; \mathbf{s}_i \in S \subset R^2\}$ estacionario de segundo orden, se verifica:

(i) $E[Y_i] = \mu \qquad \forall i \in S.$

(ii) $Cov[Y_i, Y_{i'}] = E[(Y_i - \mu)(Y_{i'} - \mu)] = C(i, i') \qquad \forall i, i' \in S.$

Es decir, el valor esperado de Y_i es constante para cualquier punto perteneciente a la región S y la covariancia entre los valores de la variable en dos unidades cualesquiera del área S (Y_i e $Y_{i'}$) depende del vector que separe a las unidades i y i' en módulo y orientación.

Cuando se cumple el supuesto (ii), se dice que el proceso es **anisotrópico**. En cambio, cuando la covariancia depende sólo de la distancia que une las unidades y no de la orientación, el proceso se llama **isotrópico** y en este caso la condición (ii) se la puede plantear como:

(ii) $Cov[Y_i, Y_{i'}] = E[(Y_i - \mu)(Y_{i'} - \mu)] = C(dist(i, i')) \qquad \forall i, i' \in S.$

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

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

$$\gamma(i, i') = \tfrac{1}{2} Var[Y_i - Y_{i'}] \qquad \forall i, i' \in S.$$

El gráfico de la semivariancia en función de la distancia que separa a las unidades se llama **semivariograma**. Comúnmente se utiliza el término de semivariograma en lugar de semivariancia.

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

$$\begin{aligned} 2\gamma(i, i') &= Var[Y_i - Y_{i'}] \\ &= E[Y_i - Y_{i'}]^2 \\ &= 2Var[Y_i] - 2Cov[Y_i, Y_{i'}]. \end{aligned}$$

Por lo tanto, el semivariograma $\gamma(i, i')$, en este caso, se puede escribir como:

$$\gamma(i, i') = Var[Y_i] - C[dist(i, i')].$$

Ahora, si se denota a la variancia de Y_i con $Var[Y_i] = C(i, i)$, se verifica:

$$\gamma(i, i') = C(i, i) - C(dist(i, i')),$$

y se denomina **función correlograma** a:

$$\rho(dist(i, i')) = \frac{C(dist(i, i'))}{C(i, i)} = 1 - \frac{\gamma(i, i')}{C(i, i)}.$$

Si la distancia que separa dos puntos se denota con $h = dist(i, i')$, se puede escribir al semivariograma y al correlograma como:

$$\gamma(h) = \frac{V[Y_{i+h} - Y_i]}{2} = \frac{E[Y_{i+h} - Y_i)]^2}{2} = C(0) - C(h),$$

$$\rho(h) = 1 - \frac{\gamma(h)}{C(0)}.$$

Una vez definidos los conceptos básicos, se describen, a continuación, los modelos de semivariogramas y correlogramas con sus características generales y particulares.

3.3. Modelos de semivariogramas y correlogramas

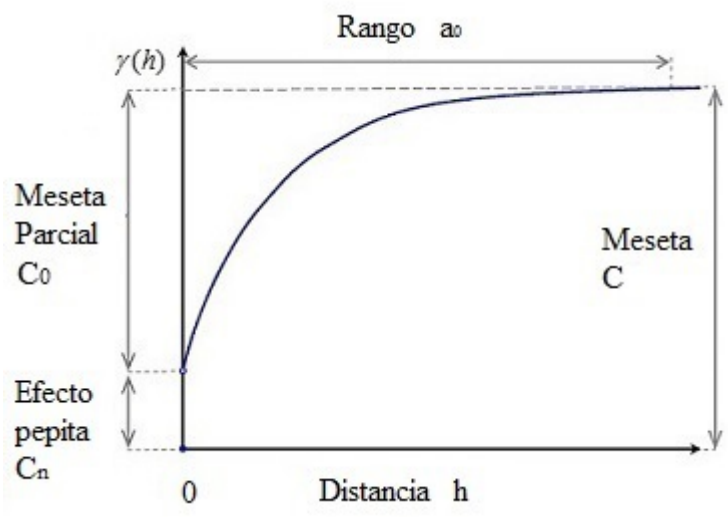
En un modelo de semivariograma para procesos estacionarios isotrópicos de segundo orden se distinguen los siguientes parámetros ((Cressie, 1993),(Ambrosio, 2000))

- **Meseta** (C): es el valor límite máximo constante que alcanza el semivariograma. Se corresponde con la variancia de la variable aleatoria.
- **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 unidades, mientras que cuando los valores de h son superiores al rango la correlación es nula.
- **Efecto pepita** (C_n): es la discontinuidad que puede presentar el semivariograma en el origen. El proceso estocástico se considera como suma de dos componentes, una correspondiente a distancias cortas, que no es posible modelar (componente caótica) y otra que es la que se modela en función de la distancia. El efecto pepita es debido a la componente caótica del proceso. La semivariancia es igual a 0 cuando la distancia entre dos unidades es $h = 0$ y por lo tanto el efecto pepita muestra por sí mismo un salto en la semivariancia en un entorno de $h > 0$.

La meseta C está compuesta por el efecto pepita, si existe, y la **meseta parcial** C_0; es decir $C = C_n + C_0$. Si el proceso Y_i es estacionario de segundo orden, la estimación de la meseta es una estimación de la variancia constante.

La Figura 3.1 presenta la representación gráfica de un modelo teórico de semivariograma con sus parámetros.

Figura 3.1: Representación gráfica de los parámetros de un modelo teórico de semivariograma con efecto pepita



Una forma de modelar la correlación espacial es mediante el modelo de semivariograma. Dependiendo de la forma del semivarigrama teórico, se distinguen algunos modelos, como ser el modelo esférico, exponencial, gausiano, potencial, efecto agujero o lineal.

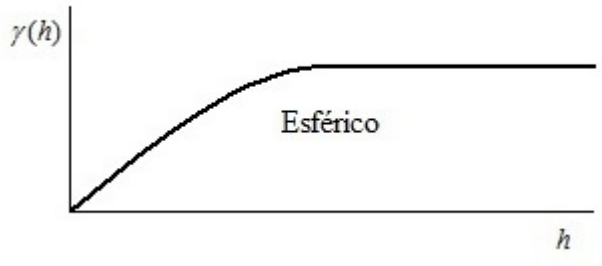
A continuación se presentan los modelos de semivariogramas más utilizados:

- Modelo esférico

$$\gamma_{sph}(h) = \begin{cases} C_n + C_0 \left(\dfrac{3}{2}\dfrac{h}{a_0} - \dfrac{1}{2}\dfrac{h^3}{a_0^3} \right) & 0 < h \leq a_0 \\ C_n + C_0 & h > a_0 \end{cases} .$$

  La meseta se alcanza para un valor de la distancia $h = a_0$, por lo tanto, el rango del semivariograma coincide con este parámetro. El modelo esférico es un modelo de transición. Un modelo de transición caracteriza un proceso aleatorio cuya variancia alcanza el valor de la meseta C dentro de un rango específico desde cualquier ubicación y por lo tanto $C(0) = C_n + C_0$.

Figura 3.2: Representación gráfica del modelo teórico de semivariograma esférico



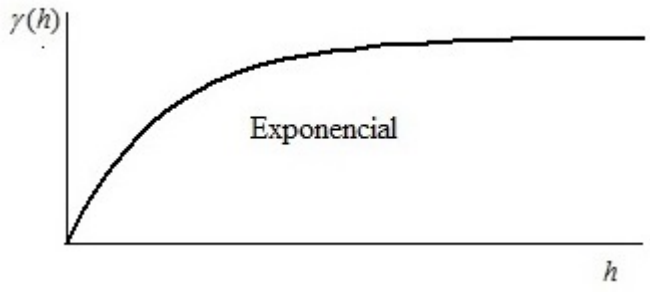
- Modelo exponencial

$$\gamma_{exp}(h) = C_n + C_0 \left(1 - e^{-\frac{h}{a_0}} \right) \qquad \forall h > 0.$$

La meseta se alcanza de forma asintótica, por lo que desde un punto de vista estricto no tiene rango. El llamado **rango efectivo** se define como la distancia para la cual el valor del semivariograma acumula el 95 % del valor de la meseta. En particular para este modelo, la relación entre el rango y el rango efectivo es $3a_0$.

Este modelo es de transición y por lo tanto, también se verifica que $C(0) = C_n + C_0$.

Figura 3.3: Representación gráfica del modelo teórico de semivariograma exponencial

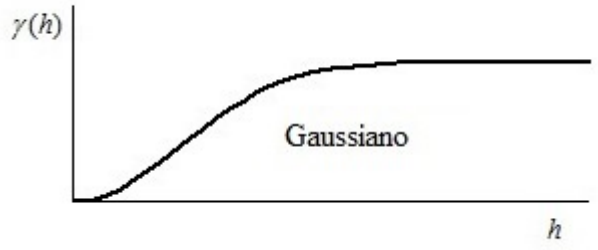


- Modelo guassiano

$$\gamma_{gau}(h) = C_n + C_0 \left(1 - e^{-\left(\frac{h}{a_0}\right)^2} \right) \qquad \forall h > 0.$$

Al igual que en el modelo anterior, la meseta se alcanza asintóticamente y el rango efectivo resulta igual a $\sqrt{3}a_0$. También es un modelo de transición o transitivo.

Figura 3.4: Representación gráfica del modelo teórico de semivariograma gaussiano



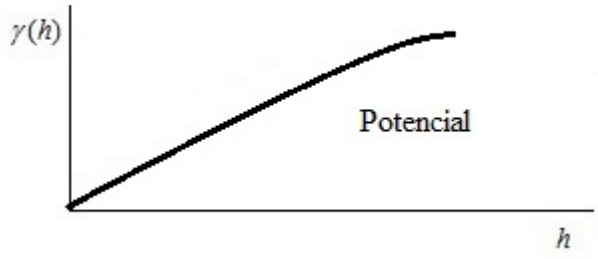
- Modelo potencial

$$\gamma_{pow}(h) = C_n + C_0 h^{a_0} \qquad \forall h > 0 \qquad 0 < a_0 < 2 \ .$$

El modelo de potencia es no transitivo y la variancia del proceso aumenta a medida que aumenta la distancia entre las unidades. No tiene meseta y rango; en su lugar, cuantifica la variación del proceso usando una pendiente positiva y una potencia que indica la rapidez con que dicha variancia aumenta.

El modelo lineal es un caso particular del modelo potencial cuando $a_0 = 1$.

Figura 3.5: Representación gráfica del modelo teórico de semivariograma potencial

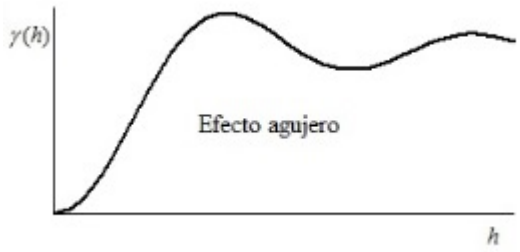


- Modelo efecto agujero

$$\gamma_{she}(h) = C_n + C_0 \left(1 - \frac{sen(\pi h/a_0)}{\pi h/a_0}\right) \qquad \forall h > 0.$$

La semivariancia para un modelo efecto agujero aumenta con la distancia. Tiene la característica distintiva que alcanza la meseta para una distancia igual al rango $h = a_0$ y luego oscila alrededor del valor de la meseta con una amplitud decreciente a medida que la distancia entre dos unidades es mayor.

Figura 3.6: Representación gráfica del modelo teórico de semivariograma efecto agujero



- Modelos anidados

  En ocasión, puede presentarse el caso en que ninguno de los modelos de semivariograma ajuste adecuadamente a un conjunto de datos pero aparecen características típicas de alguno de los modelos mezclados en los gráficos de semivariograma. En estos casos, se puede recurrir a **modelos anidados**, que consisten en la suma de dos o más estructuras de semivariancias.

  En general, una combinación lineal de modelos de semivariograma válidos produce un nuevo modelo de semivariograma válido. Por ejemplo, un semivariograma $\gamma(h)$ que contiene 2 estructuras, un modelo efecto agujero $\gamma_{she}(h)$ y otro exponencial $\gamma_{exp}(h)$, se puede expresar de la siguiente forma:

$$\gamma(h) = \gamma_{she}(h) + \gamma_{exp}(h).$$

Los modelos de semivariograma teóricos se utilizan para describir la estructura espacial de procesos aleatorios. Basados en su forma y características, pueden proporcionar una gran cantidad de información como se expresa en el capítulo "The VARIOGRAM Procedure"del libro SAS/STAT 9.3 User's Guide (2011b):

- El estudio del semivariograma en diferentes direcciones proporciona información acerca de la isotropía del proceso aleatorio.

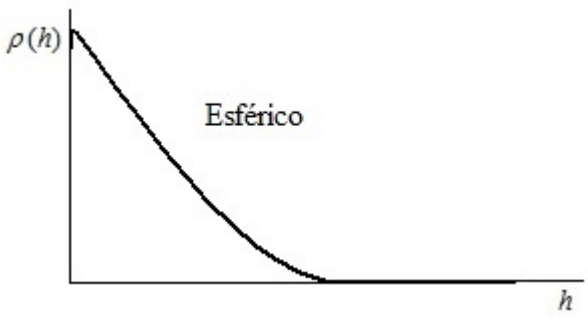
- El rango del semivariograma determina la extensión de la zona de influencia desde una región determinada. Una unidad específica está correlacionada con todas las unidades que se encuentran dentro de esta zona.
- El comportamiento del semivariograma a grandes distancias indica el grado de estacionariedad del proceso. En particular, un comportamiento asintótico sugiere un proceso estacionario, mientras que un incremento lineal o de otra forma es un indicador de no estacionariedad.
- El comportamiento del semivariograma cerca del origen indica el grado de regularidad de la variación del proceso.
- El comportamiento del semivariograma dentro del rango ofrece una descripción de periodicidades potenciales o anomalías en el proceso espacial.

Teniendo en cuenta la relación que existe entre el semivariograma y el correlograma para procesos estacionarios de segundo orden e isotrópicos, $\rho(h) = 1 - \dfrac{\gamma(h)}{C(0)}$, es posible definir un modelo de correlograma para cada modelo de semivariograma, teniendo en cuenta que en el correlograma se representa similitud (correlación) y el semivariograma se basa en diferencias (semivariancia). A continuación se presentan las expresiones matemáticas para los modelos de correlograma más clásicos con sus respectivos gráficos:

- Modelo esférico

$$\rho_{sph}(h) = \begin{cases} 1 - \dfrac{C_n + C_0\left(\dfrac{3}{2}\dfrac{h}{a_0} - \dfrac{1}{2}\dfrac{h^3}{a_0^3}\right)}{C(0)} & 0 < h \leq a_0 \\ 0 & h > a_0 \end{cases} .$$

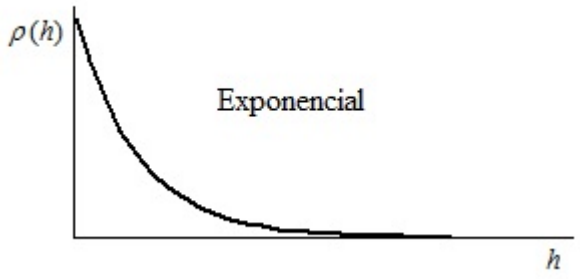
Figura 3.7: Representación gráfica del modelo teórico de correlograma esférico



- Modelo exponencial

$$\rho_{exp}(h) = \frac{C_0 e^{-\frac{h}{a_0}}}{C(0)} \qquad \forall h > 0.$$

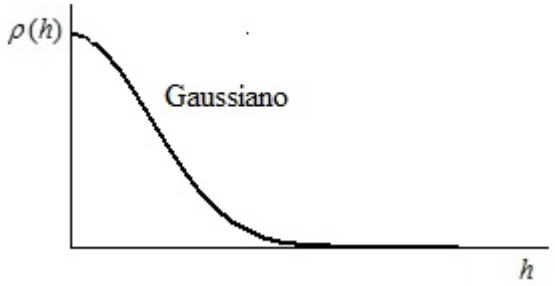
Figura 3.8: Representación gráfica del modelo teórico de correlograma exponencial



- Modelo guassiano

$$\rho_{gau}(h) = \frac{C_0 e^{-\left(\frac{h}{a_0}\right)^2}}{C(0)} \qquad \forall h > 0.$$

Figura 3.9: Representación gráfica del modelo teórico de correlograma gaussiano



Cuando el proceso es estacionario de segundo orden, el estimador clásico del semivariograma, que se presenta a continuación, es insesgado, mientras que el estimador del correlograma es sesgado.

También se verifica que bajo este supuesto, cualquiera de las tres funciones de dependencia espacial, semivariograma, covariograma o correlograma, se pueden utilizar en la

determinación de la relación espacial entre las unidades, pero la función de la semivariancias es la única que no necesita estimar previamente la media para su estimación. Por esta razón, fundamentalmente, en la práctica se emplea el semivariograma y no las otras dos funciones.

3.4. Métodos de estimación del semivariograma

Por lo general, como en todos los estudios por muestreo, los datos que se disponen pertenecen a una muestra de tamaño n, a partir de la cual se desea estimar una característica poblacional, como ser un total o una media. Si las unidades presentan correlación espacial, en una primera primera instancia se debe identificar el modelo de semivariograma y estimar sus parámetros.

La idea en la estimación del variograma es buscar uno válido que, como una medida de la dependencia espacial, sea lo más cercana posible a la presente en los datos $\mathbf{Y_n}^T = (y_1, ..., y_n)$. El espacio de todos los variogramas válidos es un gran conjunto, por lo que, usualmente, se eligen aquellos que pertenecen a una familia paramétrica de variogramas, como los descriptos en la sección anterior.

Una primera aproximación a la obtención del modelo de semivariograma es su estimación por el método de momentos o por una estimación robusta.

El estimador de momentos de un variograma $2\gamma(h)$, llamado **semivariograma empírico** o **experimental**, para un proceso estacionario isotrópico de segundo orden es (Matheron (1962), Cressie (1993)):

$$2\hat{\gamma}(h) = \frac{1}{|n(h)|} \sum_{n(h)} (y_i - y_{i'})^2$$

donde,

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

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

Como alternativa para suavizar el efecto que pueden ocasionar los outliers espaciales,

se define el semivariograma robusto (Cressie, 1993) como:

$$2\bar{\gamma}(h) = \frac{\left\{ \frac{1}{|n(h)|} \sum_{n(h)} |y_i - y_{i'}|^{1/2} \right\}^4}{0,457 + 0,494/|n(h)|}.$$

La representación gráfica de estas alternativas orientan a la elección del modelo teórico de semivariograma, cuyos parámetros deben estimarse. Los procedimientos para abordar esta estimación son: (i) especificación de un modelo teórico de variograma y estimación de sus parámetros por máxima verosimilitud restringida y (ii) cálculo del semivariograma empírico y ajuste de un modelo teórico de semivariograma al empírico mediante el método de mínimos cuadrados ponderados, cuyos fundamentos se presentan a continuación.

3.4.1. Máxima verosimilitud restringida

Para encontrar los estimadores máximo verosímiles debe asumirse un modelo subyacente para la distribución de la variable aleatoria. Se asume que el vector de observaciones muestrales $\mathbf{Y_n}$ tiene una distribución normal con vector de esperanza $\mu\mathbf{1}_n$, donde $\mathbf{1}_n$ es un vector de unos de dimension nx1 y matriz de variancias y covariancias $\mathbf{V_{n,n}}$. Un elemento genérico de $\mathbf{V_{n,n}}$ es $Cov[Y_i, Y_{i'}] = C(i, i')$, donde las unidades i, i' están ubicadas en los puntos de coordenadas $\mathbf{s}_i, \mathbf{s}_{i'} \in S$ o equivalentemente $i = 1, 2, ..., n$ e $i' = 1, 2, ..., n$, y depende del vector $\boldsymbol{\theta}_q$ a través de los q parámetros de los modelos de semivariogramas válidos.

El logaritmo de la función de verosimilitud de la muestra es:

$$L(\mu, \boldsymbol{\theta}_q) = -\frac{n}{2} ln 2\pi - \frac{1}{2} ln|\mathbf{V_{n,n}}| - \frac{1}{2}(\mathbf{Y_n} - \mu\mathbf{1}_n)^T \mathbf{V_{n,n}}^{-1}(\mathbf{Y_n} - \mu\mathbf{1}_n)$$

y los estimadores máximo verosímiles se obtienen igualando a cero las ecuaciones que resultan de derivar respecto a μ y cada uno de los parámetros del vector $\boldsymbol{\theta}_q$, de los que depende el modelo de semivariograma utilizado en la especificación de los términos de la matriz $\mathbf{V_{n,n}}$.

Debido a que el interés se centra sólo en la estimación máximo verosímil de los parámetros de la matriz de variancias y covariancias, se recurre a la estimación por máxima

verosimilitud restringida. El método consiste en eliminar la media de la función de verosimilitud de modo que quede definida sólo en términos de la matriz de variancias y covariancias. Una posible forma para obtener la verosimilitud restringida es transformar los datos a un conjunto de combinaciones lineales de las observaciones que tenga una distribución que no dependa de la media. Así, el logaritmo de la función de verosimilitud utilizando máxima verosimilitud restringida es (Cressie, 1993):

$$L(\mu, \boldsymbol{\theta}_q) = -\frac{n-1}{2} ln 2\pi - \frac{1}{2} ln|\mathbf{A}_{n,n}^T \mathbf{V}_{\mathbf{n},\mathbf{n}} \mathbf{A}_{n,n}| - \frac{1}{2} \mathbf{Y}_{\mathbf{n}}{}^T \mathbf{A}_{n,n} (\mathbf{A}_{n,n}^T \mathbf{V}_{\mathbf{n},\mathbf{n}} \mathbf{A}_{n,n})^{-1} \mathbf{A}_{n,n}^T \mathbf{Y}_{\mathbf{n}}$$

donde $\mathbf{A}_{n,n}$ es una matriz cuadrada de orden n y de rango $n-1$, tal que $\mathbf{A}_{n,n}^T \mathbf{1}_n = \mathbf{0}_n$. De esta forma resulta que $\mathbf{A}_{n,n}^T \mathbf{Y}_{\mathbf{n}}$ tiene una distribución normal de vector de esperanzas cero y matriz de variancia y covariancia $\mathbf{A}_{n,n}^T \mathbf{V}_{\mathbf{n},\mathbf{n}} \mathbf{A}_{n,n}$.

En consecuencia, los estimadores de máxima verosimilitud restringida, denominados también de máxima verosimilitud de los residuos, son estimadores de los parámetros desconocidos de $\mathbf{V}_{\mathbf{n},\mathbf{n}}$, que se obtienen maximizando la función de verosimilitud de una nueva variable definida a partir de la original. Dicha combinación lineal $\mathbf{A}_{n,n}^T \mathbf{Y}_{\mathbf{n}}$ debe verificar que $E(\mathbf{A}_{n,n}^T \mathbf{Y}_{\mathbf{n}}) = 0$.

En general, no es posible obtener expresiones analíticas para los estimadores obtenidos por máxima verosimilitud y máxima verosimilitud restringida, por ser las ecuaciones de verosimilitud no lineales en los parámetros. Para su solución se recurre a métodos iterativos, como por ejemplo el método de Newton-Raphson. También suelen aplicarse algunas variantes, como el método scoring, que reemplaza la inversa de la matriz de derivadas segundas por la inversa de la matriz de Fisher.

3.4.2. Mínimos cuadrados ponderado

A partir de la consideración del semivariograma experimental puede plantearse un modelo teórico que refleje adecuadamente sus características y utilizar el método de mínimos cuadrados para estimar los parámetros de dicho modelo.

El variograma es una función de la distancia h que se calcula clasificando todos los

pares de unidades en intervalos de acuerdo a la distancia de a pares, es decir, se discretiza la distancia $h_{(1)}, ..., h_{(L)}$ formando L intervalos.

En el ajuste basado en mínimos cuadrados, se desea estimar el vector de parámetros $\boldsymbol{\theta}_q$ del semivariograma teórico $\gamma(h)$, de modo tal que se minimice la suma de cuadrados de las diferencias ponderadas entre el semivariograma empírico y el teórico, $R(\boldsymbol{\theta}_q)$, dado por la expresión:

$$R(\boldsymbol{\theta}_q) = \sum_{l=1}^{L} p_l^2[\hat{\gamma}(h_{(l)}) - \gamma(h_{(l)}; \boldsymbol{\theta}_q)]^2$$

donde los pesos son $p_l^2 = \dfrac{1}{Var[\hat{\gamma}(h_{(l)})]}$ en el caso de mínimos cuadrados ponderados y $p_l^2 = 1$ en caso de mínimos cuadrados ordinarios.

Cressie (1985) mostró que la estimación de mínimos cuadrados ponderada aproximada del vector de parámetros $\boldsymbol{\theta}_q$ se obtiene minimizando:

$$R(\boldsymbol{\theta}_q)_{MCP} = \frac{1}{2}\sum_{l=1}^{L} n(h_{(l)})[\frac{\hat{\gamma}(h_{(l)})}{\gamma(h_{(l)}; \boldsymbol{\theta}_q)} - 1]^2,$$

donde $n(h_{(l)})$ es el número de pares de unidades en el l-ésimo intervalo de distancia.

3.5. Criterios de elección del semivariograma

En general, no existe un único camino para elegir e identificar entre varios modelos del semivariograma. La elección de los criterios para clasificar a los modelos puede depender de especificaciones conocidas o incluso de la valoración personal. Sin embargo, existen medidas de bondad de ajuste que permiten la comparación y elección del mejor modelo de semivariograma (SAS/STAT 9.3 User‘s Guide, 2011b).

Cuando se utiliza mínimos cuadrados ponderados, aquel modelo que tiene el menor valor de $R(\boldsymbol{\theta}_q)$, presentado en la sección anterior, tiene un mejor ajuste.

Otra medida de bondad de ajuste es el Criterio de Información de Akaike (AIC). En su definición estricta, AIC asume que los errores de los modelos se distribuyen de forma normal e independiente. Esta suposición no es correcta en el ajuste de la semivariancia, sin embargo, AIC también se puede definir de manera operativa sobre $R(\boldsymbol{\theta}_q)$ como:

$$AIC = L \ln\left(\frac{R(\boldsymbol{\theta}_q)}{L}\right) + 2q$$

donde L es la cantidad de intervalos que distan a una distancia h y q los parámetros del modelo de semivariograma.

Al igual que con el criterio que se basa en mínimo cuadrado ponderados, aquel modelo que presente el menor valor de AIC ofrece un mejor ajuste.

La expresión de AIC sugiere que cuando todos los modelos tienen el mismo número de parámetros, los mismos se ordenan de acuerdo a su ajuste en el mismo sentido por ambos criterios. Entre dos modelos con el mismo valor de $R(\boldsymbol{\theta}_q)$, AIC clasifica mejor al que tiene menor cantidad de parámetros.

Inferencia basada y asistida por modelos

4.1. Introducción

Un plan de muestreo probabilístico queda especificado cuando se define un método aleatorio de selección de la muestra y un procedimiento para estimar un valor poblacional. Este enfoque tradicional se conoce como basado en diseño. 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, convirtiendo a este enfoque en un procedimiento muy sólido y útil para llevarlo a cabo en la práctica.

Por otra parte, en las últimas décadas, el desarrollo del enfoque basado en modelos o en la predicción ha contribuido a la teoría del muestro en poblaciones finitas de varias formas. Se puede distinguir la incorporación de información auxiliar, con el fin de mejorar la precisión de las estimaciones, integrándola en un modelo estadístico. Además ofrece la posibilidad de reflejar en el modelo los diferentes métodos de selección, y permite la estimación en pequeñas áreas, que es un problema a destacar en los últimos años.

Bajo este enfoque, la población en estudio se considera como una muestra de una población infinita o superpoblación y las inferencias en la población finita 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.

Este enfoque tiene como inconveniente, frente al tradicional enfoque de diseño, la dependencia de la correcta especificación del modelo superpoblacional. Pero, como se mencionó, la mayor utilidad de esta aproximación es el conjunto de posibilidades que brinda en la incorporación de información auxiliar, permitiendo en algunos casos encon-

trar un Predictor Lineal Insesgado y Óptimo y estimar sus errores con las herramientas brindadas por el enfoque de diseño o de modelos.

Sin embargo, los enfoques basados en modelos y en diseño no se deben pensar como estrategias de muestreo contrapuestas, sino que pueden llegar a ser complementarias. Es en esta línea que surge el enfoque basado en el diseño, pero asistido por modelos, conocido con el nombre de enfoque asistido por modelos. El mismo integra los anteriores enfoques en el sentido que se seleccionan los estimadores según el enfoque de modelos pero la inferencia se basa sólo en la distribución en el muestreo del estimador.

"Una parte importante de los profesionales del muestreo, adoptan una actitud pragmática ante esta problemática, que consiste en seguir una aproximación basada en el diseño para la inferencia sobre características de poblaciones grandes a partir de muestras grandes y, en cambio, seguir el enfoque basado en modelos para la inferencia en pequeñas áreas o para el tratamiento del problema de la no respuesta"(Ambrosio, 2006).

En este capítulo se presenta el Predictor Lineal Insesgado y Óptimo del total, con su Error Cuadrático Medio según el enfoque de modelos, así como también las expresiones obtenidas para algunos modelos de regresión superpoblacionales específicos que se utilizan para dicha predicción. Además se presenta un resumen de las particularidades del muestreo según el enfoque asistido por modelos.

4.2. Definiciones

En el enfoque basado en modelos, se considera que la población finita, de la que se selecciona la muestra de tamaño n, es a su vez una muestra de una población infinita, llamada también superpoblación. El conjunto de valores $\{Y_i; i = 1, 2, ..., N\}$ se supone como un proceso estocástico, sobre el cual se especifican algunos supuestos que involucran a los momentos de primer y segundo orden:

(i) $E[Y_i] = \mu_i$,

(ii) $Cov[Y_i, Y_{i'}] = C(i, i')$.

Las inferencias se basan en la especificación de un modelo superpoblacional que puede o no tener en cuenta la autocorrelación de las unidades. En un primer lugar se estiman sus parámetros a partir de la información muestral y luego se obtienen las predicciones de los valores poblacionales de interés, cuyas propiedades se estudian considerando el modelo postulado.

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

- $\mathbf{Y}_N$: vector de datos poblacional de la variable en estudio, de dimensión Nx1.
- $\mathbf{Y}_n$: vector de datos observados de la variable en estudio para las unidades incluidas en la muestra, de dimensión nx1. Sin perder generalidad, se identifica a los valores de $\mathbf{Y}_n$ con los n primeros valores de $\mathbf{Y}_N$.
- $\mathbf{Y}_{N-n}$: vector de datos poblacional de la variable en estudio para las unidades no incluidas en la muestra, de dimensión $(N-n)$x1.
- $\mathbf{X}_{N,p}$: matriz de datos poblacional con una primera columna de unos, $\mathbf{1}_N$, y las restantes correspondientes a las $p-1$ variables auxiliares, de dimensión Nxp.
- $\mathbf{X}_{n,p}$: matriz de datos observados con una primera columna de unos, $\mathbf{1}_n$, y las restantes correspondientes a las $p-1$ variables auxiliares para las unidades incluidas en la muestra, de dimensión nxp, correspondientes a las n primeras filas de $\mathbf{X}_{N,p}$.
- $\mathbf{X}_{N-n,p}$: matriz de datos poblacional con una primera columna de unos, $\mathbf{1}_{N-n}$, y las restantes correspondientes a las $p-1$ variables auxiliares para las unidades no incluidas en la muestra, de dimensión $(N-n)$xp.
- $\mathbf{V}_{n,n}$: matriz de variancias y covariancias para las unidades incluidas en la muestra, de dimensión nxn.
- $\mathbf{V}_{N-n,n}$: matriz de variancias y covariancias entre las $N-n$ unidades no incluidas en la muestra y las n incluidas, de dimensión $(N-n)$xn.

4.3. Predicción del total bajo el enfoque de modelos

El total Y se puede expresar:

$$Y = \sum_{i=1}^{n} Y_i + \sum_{i=n+1}^{N} Y_i.$$

En forma matricial, dada la muestra de tamaño n, los elementos de la población $\{Y_i; i = 1, 2, ..., N\}$ se consideran ordenados de acuerdo a las definiciones dadas, representando a los n elementos de la muestra en los primeros lugares de vector $\mathbf{Y}_N$ y los restantes $(N - n)$ elementos no incluidos en la muestra en las siguientes posiciones. Es decir, $\mathbf{Y}_N$ se puede escribir como:

$$\mathbf{Y}_N = \begin{bmatrix} \mathbf{Y}_n \\ \mathbf{Y}_{N-n} \end{bmatrix}$$

y el total poblacional se especifica como:

$$Y = \mathbf{1}_n^T \mathbf{Y}_n + \mathbf{1}_{N-n}^T \mathbf{Y}_{N-n}$$

donde $\mathbf{1}_n$ y $\mathbf{1}_{N-n}$ son vectores de unos de dimensión nx1 y $(N - n)$x1 respectivamente.

La aproximación basada en modelos consiste en especificar un modelo superpoblacional para el proceso estocástico $\{Y_i; i = 1, 2, ..., N\}$ a partir del cual se predicen los valores del vector $\mathbf{Y}_{N-n}$.

Entonces, el predictor del total es:

$$\hat{Y} = \mathbf{1}_n^T \mathbf{Y}_n + \mathbf{1}_{N-n}^T \hat{\mathbf{Y}}_{N-n}.$$

El vector $\hat{\mathbf{Y}}_{N-n}$ se predice por medio del modelo de regresión lineal:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_{p-1} X_{p-1,i} + \varepsilon_i,$$

donde,

$E[\varepsilon_i] = 0$,

$$Cov[\varepsilon_i; \varepsilon_{i'}] = \begin{cases} \sigma_\varepsilon^2 = C(i, i') & i = i', \\ C(i; i') & i \neq i' \end{cases},$$

$X_{1i}, ..., X_{p-1,i}$ son las $p - 1$ variables auxiliares de la matriz $\mathbf{X}_{n,p}$.

En notación matricial $\mathbf{Y}_n = \mathbf{X}_{n,p}\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}_n$, con $E[\boldsymbol{\varepsilon}_n] = \mathbf{0}$ y $Cov[\boldsymbol{\varepsilon}_n] = \mathbf{V}_{n,n}$.

Por lo tanto, el Predictor Lineal Insesgado y Óptimo (PLIO) de $\mathbf{Y}_{N-n}$ viene dado por (Thompson, 1992):

$$\hat{\mathbf{Y}}_{N-n} = \mathbf{X}_{N-n,p}\hat{\boldsymbol{\beta}}_p + \mathbf{V}_{N-n,n}\mathbf{V}_{n,n}^{-1}(\mathbf{Y}_n - \mathbf{X}_{n,p}\hat{\boldsymbol{\beta}}_p)$$

donde,

$\mathbf{V}_{n,n} = Cov[\boldsymbol{\varepsilon}_n] = Cov[\mathbf{Y}_n]$,

$\mathbf{V}_{N-n,n} = Cov[\mathbf{Y}_{N-n}, \mathbf{Y}_n]$,

$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_{n,p}^T\mathbf{V}_{n,n}^{-1}\mathbf{X}_{n,p})^{-1}\mathbf{X}_{n,p}^T\mathbf{V}_{n,n}^{-1}\mathbf{Y}_n$,

$Var[\hat{\boldsymbol{\beta}}_p] = (\mathbf{X}_{n,p}^T\mathbf{V}_{n,n}^{-1}\mathbf{X}_{n,p})^{-1}$.

Entonces, el PLIO del total $\hat{Y}$ se puede escribir como:

$$\hat{Y} = \mathbf{1}_n^T\mathbf{Y}_n + \mathbf{1}_{N-n}^T\left[\mathbf{X}_{N-n,p}\hat{\boldsymbol{\beta}}_p + \mathbf{V}_{N-n,n}\mathbf{V}_{n,n}^{-1}(\mathbf{Y}_n - \mathbf{X}_{n,p}\hat{\boldsymbol{\beta}}_p)\right]. \tag{4.1}$$

El Error cuadrático medio de la predicción es (Thompson, 1992):

$$ECM[\hat{Y}] = E[\hat{Y} - Y]^2 =$$
$$= \mathbf{1}_{N-n}^T\left[(\mathbf{X}_{N-n,p} - \boldsymbol{\Omega}_{N-n,p})\boldsymbol{\Omega}_{p,p}^{-1}(\mathbf{X}_{N-n,p} - \boldsymbol{\Omega}_{N-n,p})^T + (\mathbf{V}_{N-n,N-n} - \mathbf{W}_{N-n,N-n})\right]\mathbf{1}_{N-n} \tag{4.2}$$

donde,

$\boldsymbol{\Omega}_{N-n,p} = \mathbf{V}_{N-n,n}\mathbf{V}_{n,n}^{-1}\mathbf{X}_{n,p}$,

$\boldsymbol{\Omega}_{p,p} = \mathbf{X}_{n,p}^T\mathbf{V}_{n,n}^{-1}\mathbf{X}_{n,p}$,

$\mathbf{V}_{N-n,N-n} = Cov[\mathbf{Y}_{N-n}, \mathbf{Y}_{N-n}]$,

$\mathbf{W}_{N-n,N-n} = \mathbf{V}_{N-n,n}\mathbf{V}_{n,n}^{-1}\mathbf{V}_{N-n,n}^T$.

4.4. Algunos casos particulares de modelos de regresión

A continuación se presentan casos particulares de modelos de regresión superpoblacionales en los que se trabaja con una única variable auxiliar. Algunos de estos modelos

se eligen debido a que se identifican con los estimadores obtenidos bajo el enfoque de diseño, como ser el estimador del total de simple expansión, razón o regresión, pero no tienen en cuenta la correlación espacial. También se presenta un modelo que sí utiliza la información de dicha variabilidad espacial a través de los parámetros del semivariograma.

4.4.1. Modelo sin variable auxiliar: Homocedástico y sin autocorrelación

El predictor del total que se obtiene a partir del modelo sin variable auxiliar homocedástico y sin autocorrelación se identifica con el estimador de simple expansión conocido del enfoque de diseño.

Modelo: $Y_i = \beta_0 + \varepsilon_i$

donde,

$E[\varepsilon_i] = 0$,

$$Cov[\varepsilon_i; \varepsilon_{i'}] = \begin{cases} \sigma_\varepsilon^2 & i = i' \\ 0 & i \neq i' \end{cases}.$$

En notación matricial el modelo se puede explicitar como $\mathbf{Y}_n = \mathbf{1}_n\beta + \boldsymbol{\varepsilon}_n$, donde

$E[\boldsymbol{\varepsilon}_n] = \mathbf{0}$,

$$Cov[\boldsymbol{\varepsilon}_n] = \mathbf{V}_{n,n} = \sigma_\varepsilon^2 \mathbf{I}_{n,n} = \begin{pmatrix} \sigma_\varepsilon^2 & 0 & \dots & 0 \\ 0 & \sigma_\varepsilon^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_\varepsilon^2 \end{pmatrix}.$$

El Estimador Lineal Insesgado y Óptimo (ELIO) del parámetro β (igual a la media muestral de la variable) y su variancia son:

$$\hat{\beta} = (\mathbf{1}_n^T \mathbf{1}_n)^{-1} \mathbf{1}_n^T \mathbf{Y}_n,$$

$$Var[\hat{\beta}] = \sigma_\varepsilon^2 (\mathbf{1}_n^T \mathbf{1}_n)^{-1} = \sigma_\varepsilon^2 / n.$$

Un estimador insesgado de la variancia de este estimador es $\hat{Var}[\hat{\beta}] = \hat{\sigma}_\varepsilon^2 (\mathbf{1}_n^T \mathbf{1}_n)^{-1}$,

donde $\hat{\sigma}^2_\varepsilon = \dfrac{(\mathbf{Y}_n - \mathbf{1}_n\beta)^T(\mathbf{Y}_n - \mathbf{1}_n\beta)}{n-1}$.

A continuación se presentan la predicción del total y su Error Cuadrático Medio para este caso, que se obtienen reemplazando por las matrices correspondientes en las formulas (4.1) y (4.2). La predicción del total resulta:

$$\hat{Y} = \mathbf{1}_n^T\mathbf{Y}_n + \mathbf{1}_{N-n}^T[\mathbf{1}_{N-n}\hat{\beta}].$$

Debido a que $\mathbf{V}_{N-n,n} = \mathbf{0}_{N-n,n}$, $\mathbf{\Omega}_{N-n,1} = \mathbf{0}_{N-n,1}$, $\mathbf{\Omega}^{-1}_{1,1} = \dfrac{\sigma^2_\varepsilon}{n}$,$\mathbf{V}_{N-n,N-n} = \sigma^2_\varepsilon\mathbf{I}_{N-n,N-n}$ y $\mathbf{W}_{N-n,N-n} = \mathbf{0}_{N-n,N-n}$ se obtiene que:

$$ECM[\hat{Y}] = N^2(1 - \frac{n}{N})\frac{\sigma^2_\varepsilon}{n}.$$

4.4.2. Modelo de regresión: Homocedástico y sin autocorrelación

El predictor del total que se obtiene a partir del modelo de regresión homocedástico y sin autocorrelación se identifica con el estimador de regresión conocido en el enfoque de diseño.

Modelo: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

donde,

$E[\varepsilon_i] = 0$,

$$Cov[\varepsilon_i;\varepsilon_{i'}] = \begin{cases} \sigma^2_\varepsilon & i = i' \\ 0 & i \neq i' \end{cases}.$$

En notación matricial el modelo se puede escribir como $\mathbf{Y}_n = \mathbf{X}_{n,2}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_n$, donde

$E[\boldsymbol{\varepsilon}_n] = \mathbf{0}$

$$Cov[\boldsymbol{\varepsilon}_n] = \mathbf{V}_{n,n} = \sigma^2_\varepsilon\mathbf{I}_{n,n} = \begin{pmatrix} \sigma^2_\varepsilon & 0 & \dots & 0 \\ 0 & \sigma^2_\varepsilon & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2_\varepsilon \end{pmatrix}.$$

El ELIO del parámetro $\boldsymbol{\beta}_2$ y la variancia del estimador resultan:

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_{n,2}^T \mathbf{X}_{n,2})^{-1} \mathbf{X}_{n,2}^T \mathbf{Y}_n,$$

$$Var[\hat{\boldsymbol{\beta}}_2] = \sigma_\varepsilon^2 (\mathbf{X}_{n,2}^T \mathbf{X}_{n,2})^{-1}.$$

El estimador insesgado de la variancia de este estimador es $\hat{Var}(\hat{\boldsymbol{\beta}}_2) = \hat{\sigma}_\varepsilon^2 (\mathbf{X}_{n,2}^T \mathbf{X}_{n,2})^{-1}$, donde $\hat{\sigma}_\varepsilon^2 = \dfrac{(\mathbf{Y}_n - \mathbf{X}_{n,2}\hat{\boldsymbol{\beta}})^T (\mathbf{Y}_n - \mathbf{X}_{n,2}\hat{\boldsymbol{\beta}})}{n-2}$.

El predictor del total resulta igual a:

$$\hat{Y} = \mathbf{1}_n^T \mathbf{Y}_n + \mathbf{1}_{N-n}^T [\mathbf{X}_{N-n} \hat{\beta}].$$

Debido a que $\mathbf{V}_{N-n,n} = \mathbf{0}_{N-n,n}$, $\boldsymbol{\Omega}_{N-n,2} = \mathbf{0}_{N-n,2}$, $\boldsymbol{\Omega}_{2,2}^{-1} = \sigma_\varepsilon^2 (\mathbf{X}_{n,2}^T \mathbf{X}_{n,2})^{-1}$,$\mathbf{V}_{N-n,N-n} = \sigma_\varepsilon^2 \mathbf{I}_{N-n,N-n}$ y $\mathbf{W}_{N-n,N-n} = \mathbf{0}_{N-n,N-n}$ se deduce que:

$$ECM[\hat{Y}] = N(1 - \frac{n}{N})\sigma_\varepsilon^2 + N^2 \left(1 - \frac{n}{N}\right)^2 \bar{\mathbf{X}}_{N-n,2} Var[\hat{\boldsymbol{\beta}}_2] \bar{\mathbf{X}}_{N-n,2}^T$$

donde $\bar{\mathbf{X}}_{N-n,2} = \dfrac{\mathbf{X}_{N-n,2}}{N-2}$.

4.4.3. Modelo de regresión sin ordenada al origen: Heterocedástico y sin autocorrelación

El predictor del total que se obtiene a partir del modelo de regresión heterocedástico, sin autocorrelación y sin ordenada al origen se identifica con el estimador de razón del enfoque de diseño.

Modelo: $Y_i = \beta_1 X_i + \varepsilon_i$

donde,

$E[\varepsilon_i] = 0$

$$Cov[\varepsilon_i;\varepsilon_{i'}] = \begin{cases} \sigma^2_{\varepsilon i} & i = i' \\ 0 & i \neq i' \end{cases}.$$

En notación matricial el modelo se explicita como $\mathbf{Y}_n = \mathbf{X}_n\beta + \boldsymbol{\varepsilon}_n$, donde

$E[\boldsymbol{\varepsilon}_n] = \mathbf{0}$,

$$Cov[\boldsymbol{\varepsilon}_n] = \mathbf{V}_{n,n} = \begin{pmatrix} \sigma^2_{\varepsilon i} & 0 & \dots & 0 \\ 0 & \sigma^2_{\varepsilon i} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2_{\varepsilon i} \end{pmatrix},$$

$\mathbf{X}_n$ es el vector que contiene sólo los datos observados para la variable auxiliar para las n unidades incluidas en la muestra.

El Estimador Lineal Insesgado y Óptimo del parámetro β supuesta $\mathbf{V}_{n,n}$ (coincide con la razón muestral) y su variancia son:

$$\hat{\beta} = (\mathbf{X}_n^T\mathbf{V}_{n,n}^{-1}\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{V}_{n,n}^{-1}\mathbf{Y}_n,$$

$$Var[\hat{\beta}] = (\mathbf{X}_n^T\mathbf{V}_{n,n}^{-1}\mathbf{X}_n)^{-1}.$$

Supuestos relativos a la estructura de la variancia de los errores.

Si $\mathbf{V}_{n,n}$ es desconocida, entonces $\hat{\beta}$ no es un estimador porque depende de los parámetros desconocidos de $\mathbf{V}_{n,n}$: en el caso más general, $\mathbf{V}_{n,n}$ consta de n elementos. En la práctica, $\mathbf{V}_{n,n}$ es desconocida por lo que el número de parámetros a estimar se eleva a $n + p$ ($n + 1$ para este caso particular) y crece al aumentar el tamaño de la muestra, de modo que se hace necesario introducir supuestos acerca de $\mathbf{V}_{n,n}$, que permitan reducir el número de parámetros a estimar.

El supuesto más simple consiste en asumir que la matriz de variancias y covariancias es de la forma $\mathbf{V}_{n,n} = \sigma^2_\varepsilon\Psi_{n,n}$, donde $\Psi_{n,n}$ es semidefinida positiva y conocida, de modo que $\mathbf{V}_{n,n}$ depende sólo de σ^2_ε.

Este supuesto es suficientemente útil en varios casos prácticos, ya que el modelo anterior se reduce a $\mathbf{Y}_n = \mathbf{X}_n\beta + \boldsymbol{\omega}_n$, con las siguientes características, en las que $\Psi_{n,n}$ se

supone conocido: $E[\boldsymbol{\omega}_n] = \mathbf{0}$ y $Cov[\boldsymbol{\omega}_n] = \mathbf{V}_{n,n} = \sigma_\varepsilon^2 \Psi_{n,n}$.

Por lo tanto el estimador de β y su variancia resultan:

$$\hat{\beta} = (\mathbf{X}_n^T \boldsymbol{\Psi}_{n,n}^{-1} \mathbf{X}_n)^{-1} \mathbf{X}_n^T \boldsymbol{\Psi}_{n,n}^{-1} \mathbf{Y}_n,$$

$$Var[\hat{\beta}] = \sigma_\varepsilon^2 (\mathbf{X}_n^T \boldsymbol{\Psi}_{n,n}^{-1} \mathbf{X}_n)^{-1}.$$

Un estimador insesgado de σ_ε^2 es $\hat{\sigma}_\varepsilon^2 = \dfrac{(\mathbf{Y}_n - \mathbf{X}_n\hat{\beta})^T \boldsymbol{\Psi}_{n,n}^{-1} (\mathbf{Y}_n - \mathbf{X}_n\hat{\beta})}{n-2}$, de modo que un estimador insesgado de la matriz de varianzas y covarianzas del estimador es $\hat{Var}[\hat{\beta}] = \hat{\sigma}_\varepsilon^2 (\mathbf{X}_n^T \boldsymbol{\Psi}_{n,n}^{-1} \mathbf{X}_n)^{-1}$.

A continuación se presenta el predictor del total para este caso particular:

$$\hat{Y} = \mathbf{1}_n^T \mathbf{Y}_n + \mathbf{1}_{N-n}^T [\mathbf{X}_{N-n} \hat{\beta}].$$

Debido a que $\mathbf{V}_{N-n,n} = \mathbf{0}_{N-n,n}$, $\boldsymbol{\Omega}_{N-n,2} = \mathbf{0}_{N-n,2}$, $\boldsymbol{\Omega}_{2,2} = \sigma_\varepsilon^2 (\mathbf{X}_{n,2}^T \boldsymbol{\Psi}_{n,n}^{-1} \mathbf{X}_{n,2})$, $\mathbf{V}_{N-n,N-n} = \sigma_\varepsilon^2 \boldsymbol{\Psi}_{N-n,N-n}$ y $\mathbf{W}_{N-n,N-n} = \mathbf{0}_{N-n,N-n}$ se obtiene que:

$$ECM[\hat{Y}] = \sigma_\varepsilon^2 \mathbf{1}_{N-n}^T \boldsymbol{\Psi}_{N-n,N-n}^{-1} \mathbf{1}_{N-n} + +N^2 \left(1 - \frac{n}{N}\right)^2 \bar{\mathbf{X}}_{N-n,2} Var[\hat{\beta}] \bar{\mathbf{X}}_{N-n,2}^T.$$

En la práctica resulta común utilizar a la matriz $\boldsymbol{\Psi}_{n,n}$ igual a una matriz diagonal con valores de la variable auxiliar, es decir:

$$\mathbf{V}_{n,n} = \sigma_\varepsilon^2 \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_n \end{pmatrix}.$$

4.4.4. Modelo de regresión: Homocedástico y con autocorrelación

El modelo de regresión homocedástico y con autocorrelación espacial incorpora no sólo la información de variable auxiliar relacionada con la variable bajo estudio, sino que también tiene en cuenta la correlación entre las unidades.

Modelo: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

donde,

$E[\varepsilon_i] = 0$,

$$Cov[\varepsilon_i; \varepsilon_{i'}] = \begin{cases} \sigma_\varepsilon^2 & i = i' \\ \sigma_\varepsilon^2 - \gamma(i; i') & i \neq i' \end{cases}.$$

Las estimaciones de σ_ε^2 y de $\gamma(i; i')$ se obtienen a partir del modelo de semivariograma ajustado a los datos.

En notación matricial el modelo se define como $\mathbf{Y}_n = \mathbf{X}_{n,2}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_n$ donde,

$E[\boldsymbol{\varepsilon}_n] = \mathbf{0}$

$$Cov[\boldsymbol{\varepsilon}_n] = \mathbf{V}_{n,n} = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_\varepsilon^2 - \gamma(1;2) & \dots & \sigma_\varepsilon^2 - \gamma(1;n) \\ \sigma_\varepsilon^2 - \gamma(1;2) & \sigma_\varepsilon^2 & \dots & \sigma_\varepsilon^2 - \gamma(2;n) \\ \dots & \dots & \dots & \dots \\ \sigma_\varepsilon^2 - \gamma(1;n) & \sigma_\varepsilon^2 - \gamma(2;n) & \dots & \sigma_\varepsilon^2 \end{pmatrix}.$$

El ELIO del vector de parámetros $\boldsymbol{\beta}$ en el modelo antes explicitado, supuesto conocido el semivariograma $\gamma(i; i')$, es:

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_{n,2}^T \mathbf{V}_{n,n}^{-1} \mathbf{X}_{n,2})^{-1} \mathbf{X}_{n,2}^T \mathbf{V}_{n,n}^{-1} \mathbf{Y}_n,$$

$$Var[\hat{\boldsymbol{\beta}}_2] = (\mathbf{X}_{n,2}^T \mathbf{V}_{n,n}^{-1} \mathbf{X}_{n,2})^{-1}.$$

Aquí, el PLIO del total y su variancia se obtienen a partir de las formulas (4.1) y (4.2), debido a que la existencia de correlación espacial, no permite simplificar ninguna de las fórmulas.

Supuestos relativos a la estructura de la variancia de los errores.

Como se mencionó en la sección anterior, $\hat{\boldsymbol{\beta}}_2$ no es un estimador si $\mathbf{V}_{n,n}$ es desconocida, ya que depende de sus parámetros desconocidos. Un estimador del estimador $\hat{\boldsymbol{\beta}}_2$ se obtiene sustituyendo en la expresión del estimador $\mathbf{V}_{n,n}$ por su estimador $\hat{\mathbf{V}}_{n,n}$: $\hat{\hat{\boldsymbol{\beta}}}_2 = (\mathbf{X}_{n,2}^T \hat{\mathbf{V}}_{n,n}^{-1} \mathbf{X}_{n,2})^{-1} \mathbf{X}_{n,2}^T \hat{\mathbf{V}}_{n,n}^{-1} \mathbf{Y}_n$.

Como estimador $\hat{\mathbf{V}}_{n,n}$ se utiliza el que resulte de sustituir en $\mathbf{V}_{n,n}$ los valores del variograma desconocidos, por sus estimaciones obtenidas a partir del variograma empírico.

El estimador $\hat{\hat{\boldsymbol{\beta}}}_2$ del estimador $\hat{\boldsymbol{\beta}}_2$ deja de ser lineal insesgado y óptimo, porque $\mathbf{V}_{n,n}$ depende también de $\mathbf{Y}_n$. Sin embargo, bajo ciertas condiciones, muy generales, puede gozar de buenas propiedades asintóticas (Ambrosio, 1999).

En la práctica, surge el problema de decidir cuál es el semivariograma más adecuado para utilizar. Se pueden destacar estas posibilidades:

- Utilizar un semivariograma obtenido a partir de una muestra piloto o estudio poblacional, en cuyo caso, los valores de $\gamma(i;i')$ serán cantidades constantes.
- Utilizar un modelo de semivariograma específico de la población y estimar sus parámetros a partir de la muestra, en donde las cantidades $\hat{\sigma_\varepsilon}^2$ y $\hat{\gamma}(i;i')$ son aleatorias.
- Identificar la forma del modelo de semivariograma y estimar sus parámetros con la información muestral, en donde también las cantidades $\hat{\sigma_\varepsilon}^2$ y $\hat{\gamma}(i;i')$ son aleatorias.

4.5. Enfoque asistido por modelos

En el enfoque de modelos, la inferencia, y en particular las propiedades de los estimadores depende estrechamente de los supuestos en los que se basa el modelo: los predictores tienen buenas propiedades estadísticas sólo si el modelo utilizado es el correcto. Ambrosio (2001) presenta una reseña de las consecuencias sobre el sesgo del predictor y de su Error Cuadrático Medio en casos de mala especificación del modelo.

Por este motivo se vuelve necesario un planteo en la forma de realizar las inferencias

por medio de un enfoque que permita que los predictores no se vean afectados por los problemas que puedan surgir en la elección del modelo adecuado. De esta manera surge el enfoque asistido por modelos, consistente en la elección de los estimadores más idóneos teniendo en cuenta los modelos, pero basando la inferencia en el diseño.

Bajo este enfoque, los estimadores que se analizan y sus Errores Cuadráticos Medios son:

- El estimador del total por simple expansión, $\hat{Y}_{se} = N \sum_{i=1}^{n} y_i$, cuyo Error Cuadrático Medio es $ECM[\hat{Y}_{se}] = N^2(1-\frac{n}{N})\frac{S_y^2}{n}$ y que se corresponde con el modelo sin variable auxiliar, homocedástico y sin autocorrelación presentado en la sección anterior.
- El estimador del total por regresión, $\hat{Y}_{rl} = \bar{y} + \hat{\beta}_1(\bar{X} - \bar{x})$, cuyo Error Cuadrático Medio resulta $ECM[\hat{Y}_{rl}] = N^2(1 - \frac{n}{N})\frac{S_y^2 + \beta_1^2 S_x^2 - 2\beta_1 S_{xy}}{n}$ y se corresponde con un modelo de regresión homocedástico y sin autocorrelación.
- El estimador del total por razón, $\hat{Y}_r = rX = \frac{\bar{y}}{\bar{x}}X$ (donde X es el total poblacional de la variable auxiliar), cuyo Error Cuadrático Medio del total estimado es $ECM[\hat{Y}_r] = N^2(1 - \frac{n}{N})\frac{S_y^2 + R^2 S_x^2 - 2RS_{xy}}{n}$. Este estimador se corresponde con el modelo de regresión sin ordenada al origen, heterocedástico y sin autocorrelación.
- El estimador del total que tiene en cuenta la variabilidad espacial. Éste no presenta una fórmula cerrada bajo el enfoque de diseño ni tampoco su Error Cuadrático Medio.

Debido a este último hecho, en los estudios comparativos a realizar, se calcula el Error Cuadrático Medio a partir de todas las muestras simples al azar de tamaño n de un población de tamaño N para construir la distribución de frecuencia del estimador. También se realizan estudios comparativos utilizando como método de selección el muestreo sistemático, el cuál es frecuentemente utilizado en poblaciones con variabilidad espacial.

Para ambos métodos de selección se calcula el Error Cuadrático Medio del total de la

siguiente manera:

$$ECM(\hat{Y}) = Var(\hat{Y}) = \frac{1}{k}\sum_{i=1}^{k}(\hat{Y}_i - Y)^2, \tag{4.3}$$

donde k representa la cantidad de muestras posibles para el muestreo sistemático y la cantidad de muestras consideradas en el muestreo aleatorio simple.

Datos en látices

5.1. Introducción

Como se mencionó en el Capítulo 2, las unidades de la población finita pueden estar dispuestas en látices regulares o irregulares. Cuando las áreas que se desean seleccionar son regulares existen artículos que presentan los procedimientos útiles para modelar la variabilidad espacial y utilizar esta información en las fases de selección de muestras y de estimación. Sin embargo, cuando las unidades de muestreo son irregulares existe escasa bibliografía. Para este caso, Cressie (1993) propone dividir al área bajo estudio en cuadrículas regulares, resultando algunas retículas sin elementos, otras con uno sólo y las restantes pueden tener más de un elemento. Luego se pueden aplicar las técnicas de muestreo usuales.

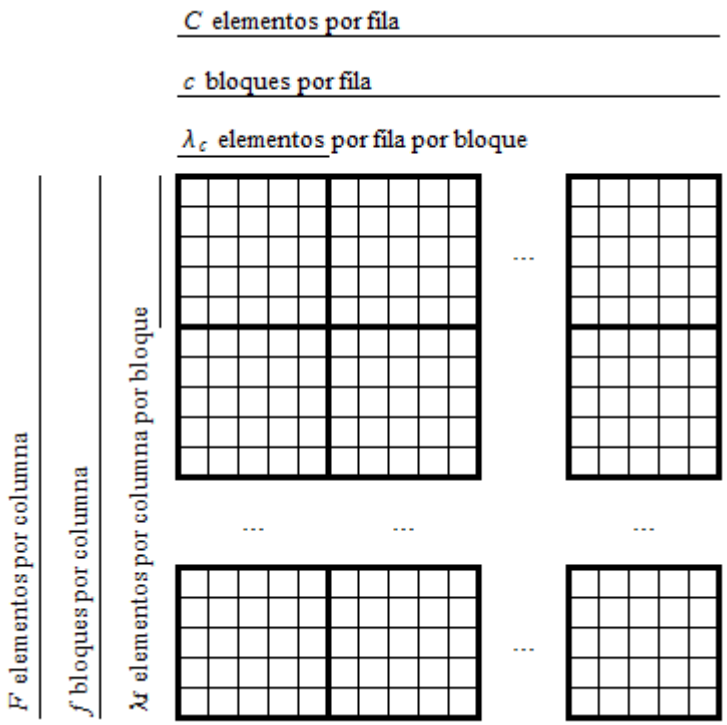
En este capítulo se recoge la propuesta metodológica presentada por Iglesias (1998) para su aplicación en estudios de estimaciones de usos del suelo a partir de látices regulares, empleando diversos planes de muestreo ampliamente conocidos.

5.2. Planes de muestreo para una población dispuesta en látices

La disposición en látices es frecuentemente considerada en el muestreo en poblaciones con variabilidad espacial y por lo tanto, el marco más idóneo para la selección de muestras es el mapa, en el cual es posible considerar divisiones del territorio cuadradas o rectangulares.

Se considera que la población está dispuesta en látices de F filas y C columnas y que los elementos se encuentran agrupados en bloques de $\lambda_f \lambda_c$ elementos en cada uno (λ_f elementos por columna por bloque y λ_c elementos por fila por bloque). Sea f el número de bloques por columna y c el número de bloques por fila, por lo que se verifica que $F = \lambda_f f$, $C = \lambda_c c$ y $FC = \lambda_f \lambda_c fc$. Es decir, las unidades están ordenadas de la siguiente manera:

Figura 5.1: Esquema de la población dispuesta en látices



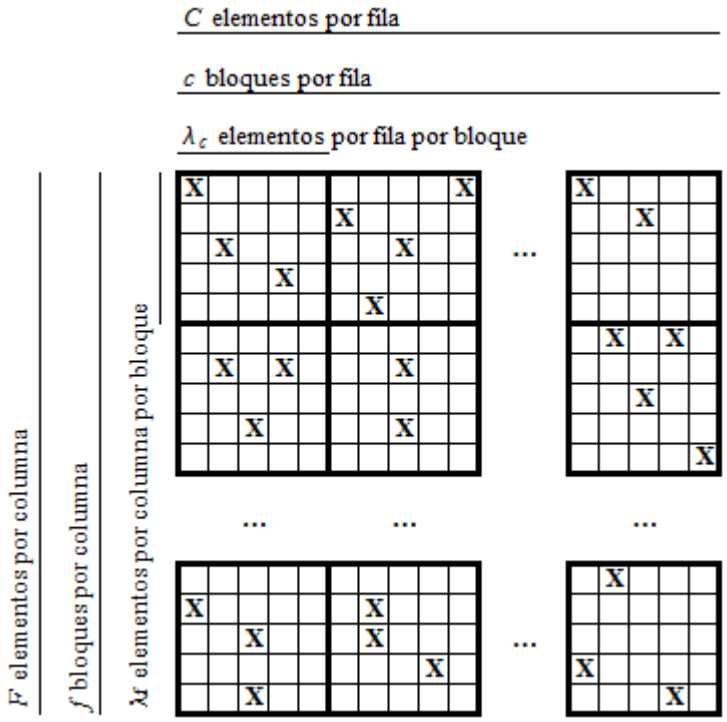
Un plan de muestreo queda especificado cuando se determina un método aleatorio de selección de la muestra y un estimador de la característica de interés.

A continuación se presentan tres métodos de selección de la muestra y en la próxima sección se describe el estimador con su variancia para un mismo tamaño de muestra n correspondiente a cada uno de estos métodos: muestreo aleatorio simple, muestreo estratificado y muestreo sistemático.

5.2.1. Muestreo aleatorio simple

El muestreo aleatorio simple consiste en extraer una muestra de n elementos con igual probabilidad y sin reposición entre los FC elementos que conforman la población. A continuación se presenta un posible esquema de una muestra aleatoria simple.

Figura 5.2: Esquema de una muestra aleatoria simple de la población dispuesta en látices



El número de unidades en cada uno de los bloques puede diferir, así como su ubicación dentro del bloque y por lo tanto, la distancia entre las unidades de la muestra es aleatoria.

La probabilidad de selección de un elemento es $p_{ij} = \frac{1}{FC}\ \forall i = 1, 2, ..., F;\ \forall j = 1, 2, ..., C$, mientras que la probabilidad de inclusión en la muestra es $\pi_{ij} = \frac{n}{FC}\ \forall i = 1, 2, ..., F;\ \forall j = 1, 2, ..., C$, llamando fracción de muestreo a $\frac{n}{FC}$.

5.2.2. Muestreo estratificado

En el muestreo estratificado, se considera a cada uno de los bloques como un estrato, de forma tal que el número de estratos resulta igual a fc, cada uno con $\lambda_f \lambda_c$ elementos. Dentro de cada estrato se extrae una muestra sin reposición independiente de tamaño n_{kl},

y de esta forma cada elemento del estrato tiene la misma probabilidad de ser seleccionado. El tamaño de muestra total queda conformado por la suma de los tamaños de muestra en cada estrato.

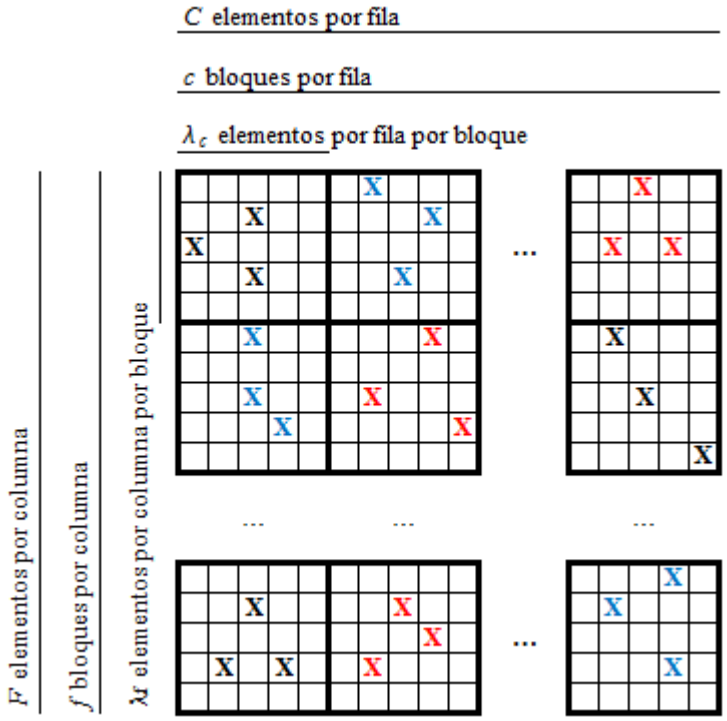
Con respecto al tipo de adjudicación de la muestra en cada estrato, una de las posibilidades es que el tamaño de la muestra se reparta proporcionalmente al tamaño del estrato, es decir:

$$n_{kl} = n\frac{\lambda_f \lambda_c}{FC} \tag{5.1}$$

en donde la fracción de muestreo en cada estrato $\frac{n_{kl}}{\lambda_f \lambda_c}$ es la misma y además igual a la del muestreo aleatorio simple $\frac{n}{FC}$.

En la Figura5.3 se presenta un posible esquema de un muestreo estratificado.

Figura 5.3: Esquema de una muestra estratificada de la población dispuesta en látices



Debido al tipo de adjudicación utilizada, el tamaño de muestra dentro de cada estrato es el mismo, $n_{kl} = \frac{n}{fc}$. La ubicación de los elementos dentro de cada uno es aleatoria y, por lo tanto, la distancia también.

La probabilidad de selección de un elemento es $p_{klgh} = \frac{1}{\lambda_f \lambda_c}$ $\forall k = 1, 2, ..., f$; $\forall l = 1, 2, ..., c$ $\forall g = 1, 2, ..., \lambda_f$; $\forall h = 1, 2, ..., \lambda_c$, mientras que la probabilidad de inclusión

es $\pi_{klgh} = \dfrac{n_{kl}}{\lambda_f \lambda_c}$ $\forall k = 1, 2, ..., f$; $\forall l = 1, 2, ..., c$ $\forall g = 1, 2, ..., \lambda_f$; $\forall h = 1, 2, ..., \lambda_c$, donde sustituyendo por la formula (5.1), resulta $\pi_{klgh} = \dfrac{n}{FC}$, la cual coincide con la probabilidad de inclusión en muestreo aleatorio simple.

5.2.3. Muestreo sistemático

Una muestra sistemática de tamaño n se puede extraer de diversas maneras, en particular, se considera extraer n_t unidades con igual probabilidad de entre las $\lambda_f \lambda_c$ unidades de un bloque cualquiera, a las que se denominan arranques aleatorios. Luego, en cada uno de los $(fc - 1)$ bloques restantes se seleccionan aquellas unidades que se encuentran en la misma posición relativa de estos arranques aleatorios. En este caso, se verifica que $n = n_t fc$.

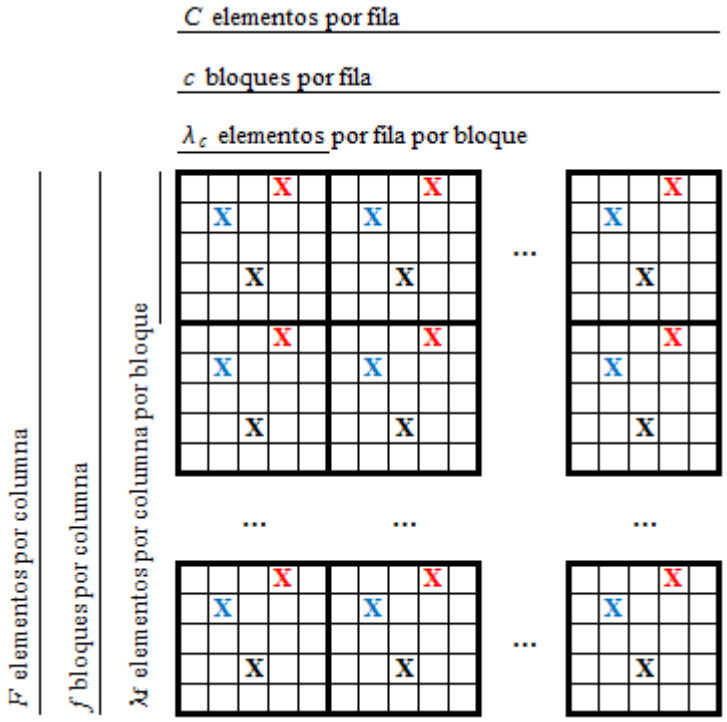
Este método de selección es equivalente a considerar un muestreo por conglomerado, donde las $FC = \lambda_f \lambda_c fc$ unidades de la población están agrupados en $\lambda_f \lambda_c$ conglomerados de fc unidades cada uno, de forma tal que las unidades de cada uno de estos conglomerados se encuentran espaciadas regularmente a una distancia λ_f en las filas y λ_c en las columnas. De los $\lambda_f \lambda_c$ conglomerados se seleccionan sin reposición y con igual probabilidad, una muestra de n_t conglomerados.

La ubicación de los n_t arranques es aleatoria pero, los restantes elementos se disponen regularmente a una distancia λ_c en la dirección de las filas y λ_f en las columnas, presentándose en la Figura 5.4 un posible esquema de un muestreo sistemático con tres arranques aleatorios.

La probabilidad de selección de un elemento es $p_{ij} = \dfrac{1}{\lambda_f \lambda_c}$ $\forall i = 1, 2, ..., F$; $\forall j = 1, 2, ..., C$, mientras que la probabilidad de inclusión en la muestra es $\pi_{ij} = \dfrac{n_t}{\lambda_f \lambda_c} = \dfrac{n_t fc}{\lambda_f \lambda_c fc} = \dfrac{n}{FC}$ $\forall i = 1, 2, ..., F$; $\forall j = 1, 2, ..., C$.

De este modo se evidencia que la probabilidad de inclusión para cualquiera de los tres métodos de selección presentados resulta ser la misma.

Figura 5.4: Esquema de una muestra sistemática con tres arranques aleatorios de la población dispuesta en látices



5.3. Estimación del total

Sea Y_{ij} el valor de la variable en estudio ubicado en la fila i y columna j del esquema de látices regulares o grilla y se define el total poblacional como:

$$Y = \sum_{i=1}^{F} \sum_{j=1}^{C} Y_{ij}.$$

Como estimador de Y se considera la media de los valores observados en los n elementos de la muestra por la cantidad de unidades de la población, FC:

$$\hat{Y} = FC \frac{1}{n} \sum_{r=1}^{n} y_r.$$

Este estimador junto con los métodos de selección antes descriptos, definen tres planes de muestreo distintos, que utilizan estimadores insesgados del total, pero con variancias diferentes según el caso. El plan de muestreo más eficiente es aquel que proponga el estimador con menor variancia. El análisis de los métodos de estimación se aborda en las siguientes secciones.

5.3.1. Muestreo aleatorio simple

El estimador de simple expansión del total para un muestreo aleatorio simple, $\hat{Y}_{mas}$, con su correspondiente variancia se definen como:

$$\hat{Y}_{mas} = FC\frac{1}{n}\sum_{r=1}^{n} y_r,$$

$$Var(\hat{Y}_{mas}) = (FC)^2(1-\frac{n}{FC})\frac{1}{n}S^2, \tag{5.2}$$

donde,
$S^2 = \frac{1}{FC-1}\sum_{i=1}^{F}\sum_{j=1}^{C}(y_{ij}-\bar{Y})^2$ es la variancia poblacional.

Un estimador insesgado de la variancia del total es $\hat{Var}(\hat{Y}_{mas}) = (FC)^2(1-\frac{n}{FC})\frac{1}{n}s^2$, donde $s^2 = \frac{1}{n-1}\sum_{r=1}^{n}(y_r-\bar{y})^2$ es un estimador insesgado de la variancia poblacional.

5.3.2. Muestreo estratificado

El estimador del total poblacional de simple expansión en un muestreo estratificado con asignación proporcional, $\hat{Y}_{est}$, se puede escribir como:

$$\hat{Y}_{est} = FC\sum_{k=1}^{f}\sum_{l=1}^{c} w_{kl}\bar{y}_{kl},$$

donde,
$w_{kl} = \frac{\lambda_f\lambda_c}{FC} = \frac{1}{fc}$, es el tamaño relativo del estrato,
$\bar{y}_{kl} = \frac{1}{n_{kl}}\sum_{s=1}^{n_{kl}} y_{kls}$, media muestral dentro del estrato kl-ésimo.

La variancia del total estimado es:

$$Var(\hat{Y}_{est}) = \frac{(FC)^2}{fc}(1-\frac{n}{FC})\frac{1}{n}\sum_{k=1}^{f}\sum_{l=1}^{c} S_{kl}^2, \tag{5.3}$$

donde,
$S_{kl}^2 = \frac{1}{\lambda_f\lambda_c-1}\sum_{g=1}^{\lambda_f}\sum_{h=1}^{\lambda_c}(Y_{klgh}-\bar{Y}_{kl})^2$, es la variancia poblacional dentro del estrato kl,
$\bar{Y}_{kl} = \frac{1}{\lambda_f\lambda_c}\sum_{g=1}^{\lambda_f}\sum_{h=1}^{\lambda_c} Y_{klgh}$,

Y_{klgh} es el valor de la variable en estudio asociado a la unidad que se encuentra en la posición gh-ésima del estrato kl.

Un estimador de la variancia del estimador del total se obtiene reemplazando en la expresión anterior S_{kl}^2 por s_{kl}^2:

$$\hat{Var}(\hat{Y}_{est}) = \frac{(FC)^2}{fc}(1 - \frac{n}{FC})\frac{1}{n}\sum_{k=1}^{f}\sum_{l=1}^{c} s_{kl}^2,$$

donde,

$s_{kl}^2 = \frac{1}{n_{kl}-1}\sum_{s=1}^{n_{kl}}(y_{kls} - \bar{y}_{kl})^2$.

5.3.3. Muestreo sistemático

El estimador del total por simple expansión para una muestra sistemática, $\hat{Y}_{sist}$, se define como:

$$\hat{Y}_{sist} = \frac{FC}{fcn_t}\sum_{t=1}^{n_t} y_t = \frac{1}{fcn_t}\sum_{t=1}^{n_t}\sum_{k=1}^{f}\sum_{l=1}^{c} y_{klt},$$

siendo y_t el total dentro de la muestra sistemática t-ésima.

De esta forma, los tres estimadores del total se definen como el total estimado de los valores observados en las unidades incluidas en la muestra.

Con este plan de muestreo, la variancia del estimador del total resulta:

$$Var(\hat{Y}_{sist}) = (FC)^2\frac{1}{(fc)^2}(1 - \frac{n}{FC})\frac{S_{gh}^2}{n_t}, \tag{5.4}$$

donde,

$S_{gh}^2 = \frac{1}{\lambda_f\lambda_c - 1}\sum_{g=1}^{\lambda_f}\sum_{h=1}^{\lambda_c}(Y_{gh} - \bar{Y}_{...})^2$ es la variancia poblacional entre los totales de las muestras sistemáticas,

Y_{gh} es el total de la muestra sistemática gh-ésima de las $\lambda_f\lambda_c$ posibles,

$\bar{Y}_{...}$ es la media de los totales de las muestras sistemáticas.

Un estimador insesgado de la variancia del total es:

$$\hat{Var}(\hat{Y}_{sist}) = (FC)^2\frac{1}{(fc)^2}(1 - \frac{n}{FC})\frac{s_{gh}^2}{n_t},$$

siendo,
$s^2_{gh} = \frac{1}{n_t}\sum_{t=1}^{n_t}(y_t - \bar{y}_{\cdot}),$
y_t el total dentro de la muestra sistemática t-ésima,
$\bar{y}_{\cdot} = \frac{1}{n_t}\sum_{t=1}^{n_t} y_t.$

5.4. Eficiencias relativas

Cabe mencionar que el estimador de simple expansión utilizado en las tres estrategias de muestreo resulta el mismo, por lo que sólo difieren en el método de selección. Cada plan de muestreo genera un espacio muestral distinto y una distribución del estimador distinta. El estimador del total es insesgado en todos los planes de muestreo pero la variancia difiere de un plan a otro, por lo que la eficiencia relativa es una buena opción para comparar estos planes.

La eficiencia relativa ($ER_{A/B}$) entre dos planes de muestreo A y B, es el cociente entre las variancias de los estimadores cuando el tamaño de muestra es el mismo en ambos procedimientos. Es decir:

$$ER_{A/B} = \frac{Var_B(\hat{Z})}{Var_A(\hat{Z})} \text{ suponiendo que } n_A = n_B.$$

Si $ER_{A/B} > 1$ la estrategia A es más eficiente que la B, mientras que si $ER_{A/B} < 1$, la estrategia B es más eficiente que la A.

A continuación se presentan las eficiencias relativas para las diversas estrategias de muestreo consideradas, que surgen como el cociente de las variancias del total presentadas en las formulas (5.2), (5.3) y (5.4) (Iglesias, 1998):

- Eficiencia relativa del muestreo estratificado respecto al muestreo aleatorio simple:

$$ER_{est/mas} = \frac{fcS^2}{\sum_{k=1}^{f}\sum_{l=1}^{c} S^2_{kl}}. \tag{5.5}$$

- Eficiencia relativa del muestreo sistemático respecto al muestreo aleatorio simple:

$$ER_{sist/mas} = \frac{fcS^2}{S_{gh}^2}. \tag{5.6}$$

- Eficiencia relativa del muestreo sistemático respecto al muestreo estratificado:

$$ER_{sist/est} = \frac{\sum_{k=1}^{f}\sum_{l=1}^{c} S_{kl}^2}{S_{gh}^2}. \tag{5.7}$$

Resultados

6.1. Introducción

Este capítulo se dedica a la presentación de un conjunto de resultados que muestran la forma de llevar a cabo la propuesta metodológica contenida en esta tesis, así como estudios comparativos que ponen en evidencia las mejoras que pueden lograrse utilizando información de la variabilidad espacial.

Se ha elegido como caso de estudio y con la intención de realizar un aporte a la práctica de encuestas por muestreo en el área socioeconómica, la estimación de un parámetro de sumo interés en estudios de pobreza de la ciudad de Rosario, como lo es el total de hogares con Necesidades Básicas Insatisfechas.

El empleo de herramientas que permiten observar la existencia de variabilidad espacial de la variable número de hogares con Necesidades Básicas Insatisfechas medida a los radios censales de la ciudad de Rosario se muestra en la segunda sección, así como también se mencionan aquellas características del conjunto de datos espaciales que surgen a partir de la aplicación de las técnicas mencionadas.

La obtención de predictores del total de hogares con Necesidades Básicas Insatisfechas en la ciudad de Rosario a partir de una muestra, utilizando los modelos propuestos en el Capítulo 4, se presentan en la tercera sección. Además se estiman los Errores Cuadráticos Medios y las eficiencias relativas de los métodos. Uno de los modelos empleados utiliza información de la variabilidad espacial provista por el semivariograma, debiendo previamente identificar el modelo y realizar el ajuste correspondiente.

El primer estudio comparativo consiste en la evaluación, en el muestreo sistemático, del comportamiento de los estimadores elegidos a partir de los predictores propuestos por

los modelos de regresión (Sección 6.4). Para el caso de los modelos que emplean información espacial, se tuvieron en cuenta diversos modelos de semivariograma: se utilizó el semivariograma poblacional, el modelo poblacional de semivariograma pero estimado con los datos de la muestra y el semivariograma identificado y estimado con dichos datos. Se obtuvieron los Errores Cuadráticos Medios empleando los estimadores obtenidos con cada una de las muestras sistemáticas y además se observó el comportamiento de la estimación del Error Cuadrático Medio del total bajo el enfoque de modelos.

En la quinta sección se presenta el segundo estudio comparativo. En él, se evalúa el comportamiento de los estimadores mencionados en la sección anterior utilizando el Error Cuadrático Medio, pero en el muestreo aleatorio simple. Para aquellos predictores que incluyen información de la variabilidad espacial, se obtuvieron aproximaciones al Error Cuadrático Medio a partir de 10000 muestras.

En la última sección se presenta un estudio comparativo, partiendo de la propuesta de Iglesias (1998), quien presenta diversos planes de muestreo para datos con correlación espacial dispuestos en látices regulares. Se adapta la información de los radios censales organizada en látices irregulares a la disposición mencionada, con el fin de estimar el total de hogares con Necesidades Básicas Insatisfechas en la ciudad de Rosario. A partir de estimadores de simple expansión, se muestra de qué forma el semivariograma puede orientar la selección de muestras sistemáticas, para lograr resultados más eficientes que los otros métodos de selección considerados.

Previo al desarrollo de las mencionadas secciones, debe decirse que para dichos análisis se cuenta con información poblacional de la variable, en base a datos del Censo Nacional de Población, Hogares y Viviendas de 2001.

El indicador de Necesidades Básicas Insatisfechas (NBI) tiene como principal finalidad medir el nivel y la intensidad de la pobreza. Los hogares con Necesidades Básicas Insatisfechas son aquellos que presentan condiciones de privación en al menos uno de los siguientes aspectos (publicación de Indec (2003)):

- Hacinamiento: Un hogar en el que habitan más de tres personas por cuarto (habi-

tación de uso exclusivo).

- Vivienda: Un hogar que habita en una vivienda de tipo inconveniente como ser pieza de inquilinato, vivienda precaria, ect. excluyendo casa, departamento y rancho.
- Condiciones sanitarias: Un hogar que habita en una vivienda sin retrete.
- Asistencia escolar: Un hogar que tienen al menos un niño en edad escolar (6 a 12 años) que no asiste a la escuela.
- Capacidad de subsistencia: Un hogar que tiene cuatro o más personas por miembro ocupado, cuyo jefe no completó el tercer grado de escolaridad primaria.

El número de hogares con NBI se obtiene actualmente, a partir de relevamientos censales, momento en el que se miden estas variables en forma simultánea. En dichos operativos, la ciudad se divide en áreas pequeñas, denominadas fracciones y las mismas se subdividen en radios censales, definidos a partir de la cantidad de viviendas que contienen.

Para el censo 2001 la ciudad de Rosario se dividió en 56 fracciones y en 896 radios censales y el número de hogares con Necesidades Básicas Insatisfechas se dispone totalizado por radio censal. En esa ocasión el total de hogares con NBI en la ciudad de Rosario resultó igual a 29622.

También se tiene en cuenta la información de la variable número de hogares por radio censal, la cual se encuentra relacionada positivamente con la variable en estudio.

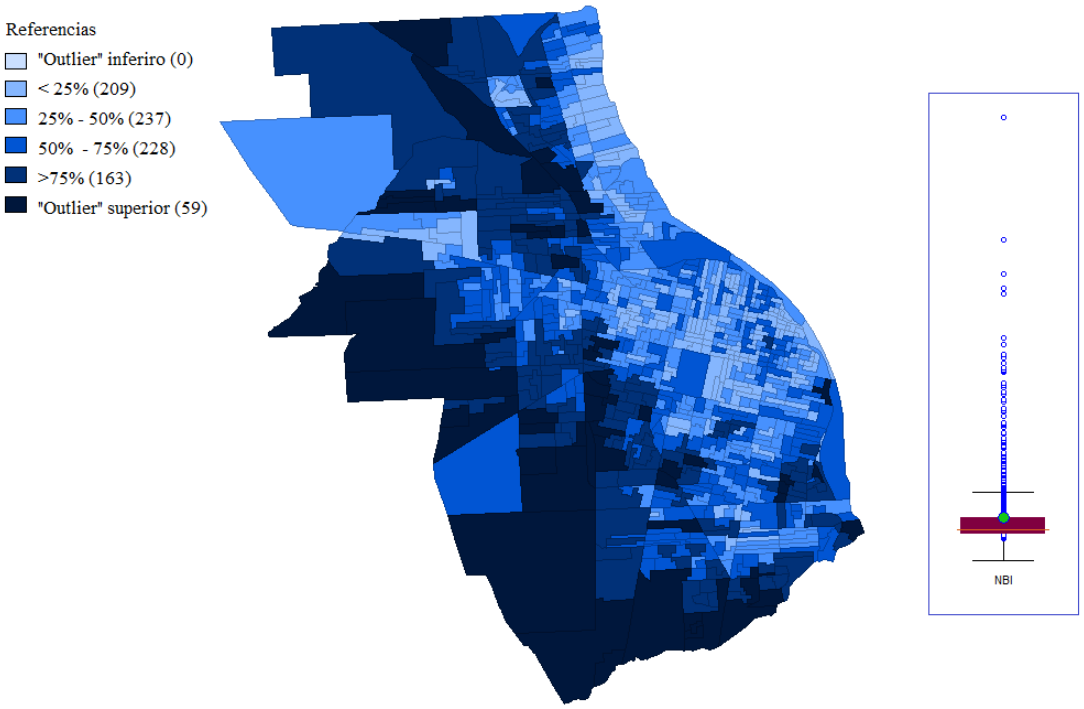
6.2. Estudio exploratorio

A continuación se presentan los resultados del análisis descriptivo con la finalidad de evaluar la existencia de variabilidad espacial para la variable número de hogares con NBI, utilizando como unidad de análisis el radio censal, según lo desarrollado en el Capítulo 2. Como se mencionó, la herramienta informática empleada para el tratamiento de los datos en esta sección es el programa GeoDa 0.9.5-i (Geodata Analysis Software).

Un primer paso en el análisis consiste en observar la representación gráfica de la distribución espacial del número de hogares con NBI de acuerdo a los radios censales de la ciudad de Rosario, recurriendo al “box plot” y el “box map” asociado, los que se presentan en la Figura 6.1.

En el “box map”, los radios aparecen sombreados en seis colores diferentes según el valor que tome la variable número de hogares con NBI en cada uno de ellos, correspondientes a 6 intervalos del valor de la variable en estudio definidos a partir del “box plot”: el primer cuartil, la mediana, el tercer cuartil y los usuales límites para considerar que una observación es anómala.

Figura 6.1: “Box plot” y “box map” para el número de hogares con NBI



Como puede apreciarse, los radios censales con menores cantidades de hogares con NBI se encuentran distribuidos mayoritariamente en la zona centro de la ciudad y en un sector de la zona norte, están rodeados de radios con la misma característica y aparecen en colores claros. En la zona noroeste se distingue un grupo de radios censales con bajas cantidades de hogares con NBI pero rodeados de radios con altas cantidades de hogares con esta característica. Se observa también que radios censales con mayores cantidades de hogares con NBI se concentran mayoritariamente en las zonas sur y oeste de Rosario, pintados con colores más oscuros.

El “box plot” muestra que la distribución del número de hogares con NBI es asimétrica a la derecha. El número de hogares con Necesidades Básicas Insatisfechas varía de 0 a

671. La mitad de los radios censales de la ciudad tienen menos de 13 hogares con NBI y el 75 % de los radios tiene menos de 33 de hogares con esta característica, valor muy alejado del máximo (671).

Considerando el corte de 3 veces el rango intercuartílico ($Q_3 + 3(Q_3 - Q_1)$ y $Q_1 - 3(Q_3 - Q_1)$), hay 59 radios censales que se destacan por sus altos valores de hogares con Necesidades Básicas Insatisfechas, considerados "outliers" y coloreados en azul más oscuro en la Figura 6.1. La mayoría de estos radios censales están rodeados de otros que también tienen altas cantidades de hogares con NBI aunque en menor medida. Estos radios, en su mayoría, son de gran superficie y están ubicados lejos del centro de la ciudad, más específicamente se encuentran en la periferia rosarina.

El índice de asociación global de Moran mide la tendencia de valores similares a agruparse en el espacio, es decir, hasta que punto áreas con altos niveles de pobreza están cerca de otras áreas de alta pobreza mientras que las zonas de poca pobreza están rodeadas de otras similares. Según el criterio de conectividad tipo reina, este índice resulta igual a 0,42 (p-value=0,00) mostrando existencia de autocorrelación espacial positiva significativa, lo que concuerda con lo observado en el "box map".

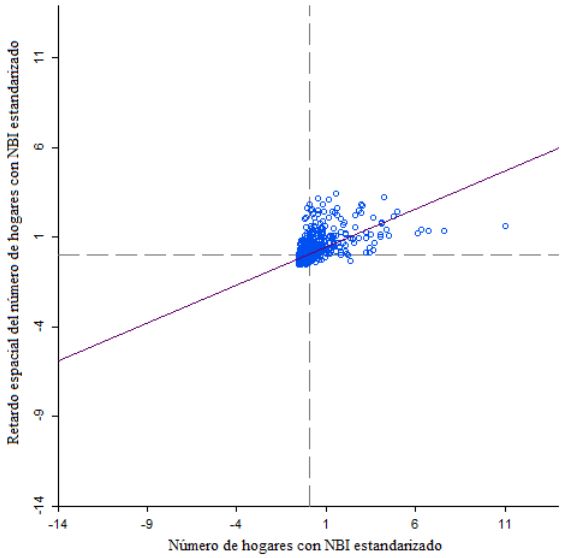
En la Figura 6.2 se presenta el gráfico de dispersión de Moran bajo el criterio reina. En él se observa que la nube de puntos aparece concentrada mayoritariamente en los cuadrantes I y III. En el cuadrante I se ubican los radios censales con alta cantidad de hogares con NBI, que están rodeados de radios censales que también tiene una alta cantidad de hogares con NBI. También se destacan varias observaciones alejadas del resto, las que corresponden a los radios censales considerados como "outliers" en el "box plot".

En el cuadrante III se ubican radios con un bajo número de hogares con NBI rodeados de radios también con baja cantidad de hogares con las mismas características y presentan mayor concentración que los radios ubicados en el cuadrante I.

Los puntos que en el diagrama de dispersión de Moran aparecen en los cuadrantes II y IV corresponden a radios censales con baja cantidad de hogares con NBI rodeados de otros con alta cantidad de hogares con NBI (cuadrante II) y viceversa (cuadrante IV).

La cantidad de unidades en estos cuadrantes resulta mucho menor.

Figura 6.2: Diagrama de dispersión de Moran para el número de hogares con NBI



Para identificar la correlación espacial local se calculan los índices de asociación espacial local de Moran para cada radio censal bajo el criterio de conectividad tipo reina. Los mismos se representan en los mapas LISA (Figura 6.3), donde se observan aquellas localizaciones con valores significativos en cuanto a asociación espacial local, permitiendo identificar conglomerados y "outliers" espaciales.

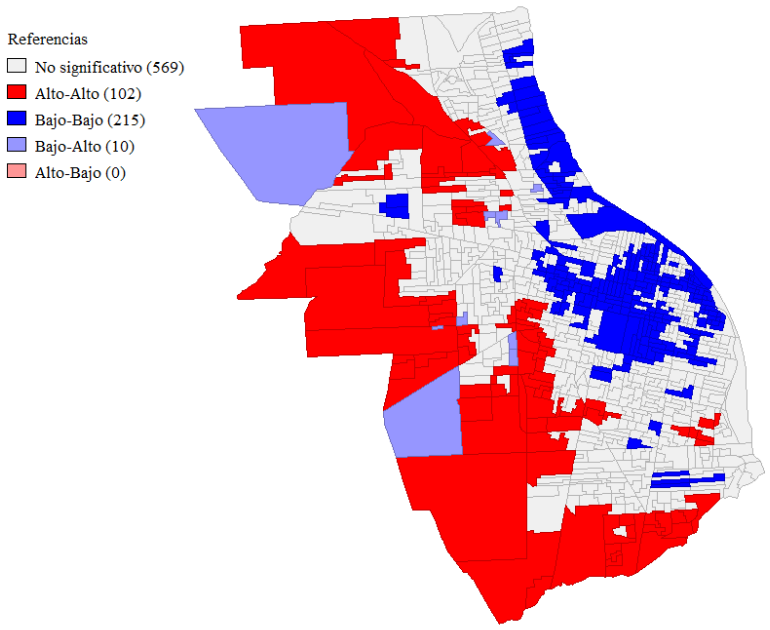
La Figura 6.3a muestra el mapa de conglomerados y la Figura 6.3b el mapa de significación para el número de hogares con NBI. De los 896 índices locales analizados 327 resultan significativos (36,5 %).

El color rojo identifica aquellos radios censales que tienen alta cantidad de hogares con NBI y se encuentran rodeados de otros radios en la misma situación, los que se ubican mayormente en las zonas sur, oeste, noroeste y suroeste. Los índices obtenidos para estos radios resultan todos significativos al 5 %, lo que se observa en el mapa de significación.

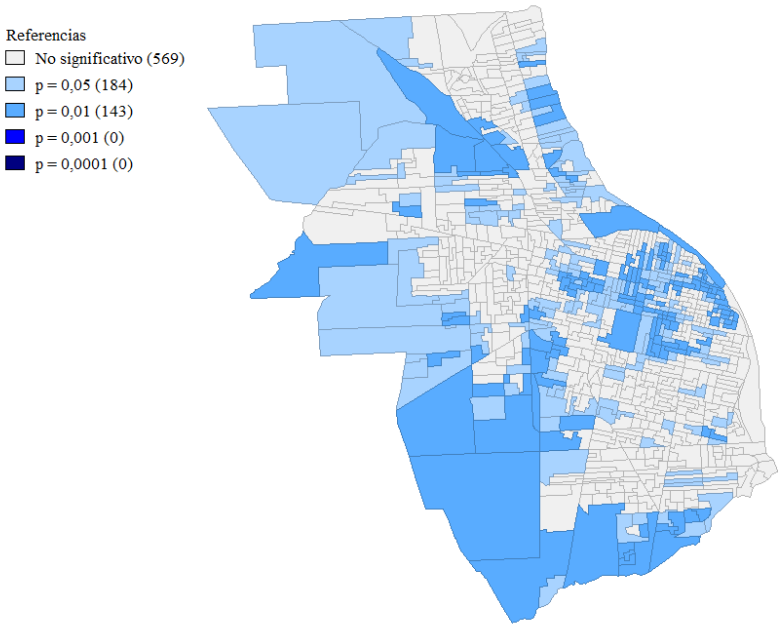
En color azul oscuro se encuentran radios con valores bajos de NBI rodeados de radios también con valores bajos de dicha variable, la mayoría en la zona centro y norte de la ciudad, los cuales también resultan significativos al 5 %. Estos conglomerados están menos dispersos que los correspondientes a los de valores altos, lo cual se puede observar también en el diagrama de dispersión de Moran.

El color celeste identifica a los radios con bajas cantidades de hogares con NBI rodeados de radios con valores altos y en color rosa se encuentra el caso inverso: radios con valores altos rodeados de vecinos con valores bajos del número de hogares con NBI, todos significativos al 5 %. El resto de los radios censales no presentan correlaciones locales significativas.

Figura 6.3: Mapas LISA para el número de hogares con NBI



(a) Mapa de conglomerado LISA



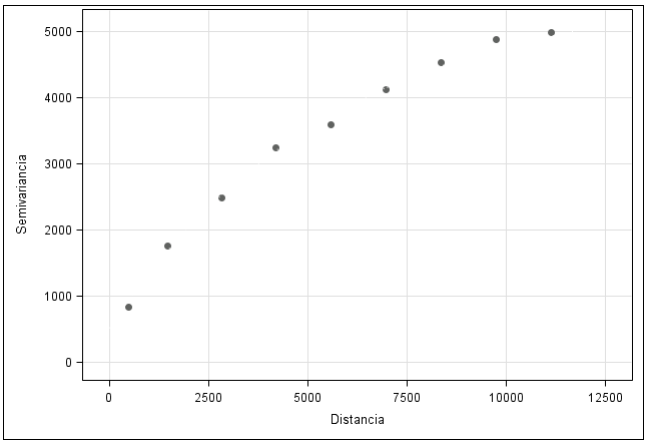
(b) Mapa de significacion LISA

De lo analizado, en la ciudad de Rosario en el año 2001 se detecta que:

- Existe correlación espacial positiva para la variable número de hogares con Necesidades Básicas Insatisfechas.
- Existen conglomerados de radios que presentan valores bajos de la variable rodeados de otros con las mismas características en las zonas centro y norte de la ciudad.
- Existen conglomerados de radios que presentan altas cantidades de hogares con Necesidades Básicas Insatisfechas rodeados de otros con las mismas características en las zonas sur, oeste, noroeste y suroeste de la ciudad.
- Existen outliers espaciales ubicados mayormente en la periferia de Rosario.

Dichas conclusiones fundamentan la búsqueda de un modelo para su explicación. La variabilidad espacial puede modelarse recurriendo a los semivariogramas. En esta fase exploratoria, se presenta el semivariograma empírico de la variable número de hogares con NBI calculado a partir de la población de radios censales de la ciudad de Rosario.

Figura 6.4: Semivariograma empírico para el número de hogares con NBI, a partir de la totalidad de radios censales



Como se verá más adelante, el aspecto de este semivariograma parece reflejar el comportamiento de un modelo exponencial o esférico.

Previo al ajuste del semivariograma empírico a un modelo teórico, se calculan los semivariogramas empíricos en función de la orientación de los vectores que unen los puntos, con el fin de evaluar una posible anisotropía. Si bien los gráficos de los semivariogramas en las direcciones de 0º, 45º, 90º y 135º no son exactamente los mismos, las diferencias

encontradas podrían hacer suponer que no existe anisotropía para la variable en cuestión. Debido a esto, para los siguientes análisis, se considera que los modelos de semivariograma sólo dependen de la distancia entre unidades y no de la orientación del vector que las separa.

6.3. Predicción del total de hogares con NBI empleando diferentes modelos

En esta sección se presentan los resultados, según el enfoque basado en modelos, de la predicción del total de hogares con Necesidades Básicas Insatisfechas en Rosario a partir de una muestra aleatoria simple de 148 radios censales de una población finita conocida, utilizando diferentes modelos de regresión, donde uno de ellos incluye la información provista por el modelo de semivariograma.

La existencia de variabilidad espacial observada en el estudio exploratorio y luego representada en el semivariograma empírico, sugiere la identificación de un modelo teórico y la estimación de sus parámetros. Se emplea el método de mínimos cuadrados ponderados utilizando el procedimiento PROC VARIOGRAM del programa SAS 9.3 para dicha estimación.

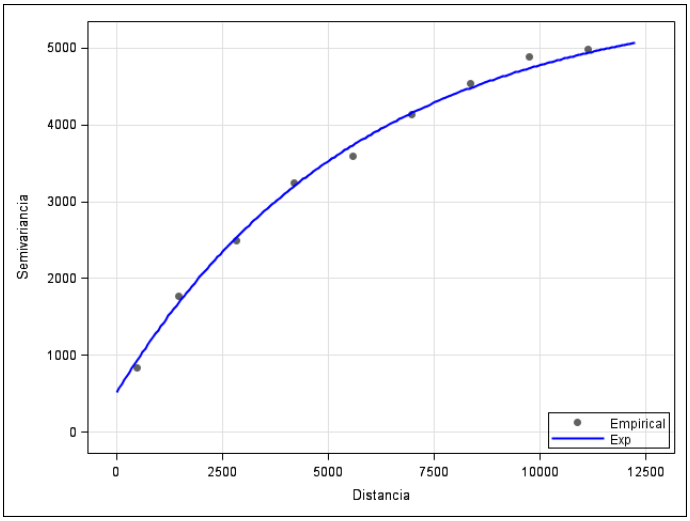
En una primera aproximación se identifican los modelos exponencial y esférico como adecuados para representar el comportamiento de la semivariancia en función de la distancia para la variable número de hogares con NBI. Se procede al ajuste de ambos modelos y se opta por el semivariograma exponencial a partir del análisis de las estadísticas $R(\boldsymbol{\theta}_q)$ y AIC que resultan iguales a 150,63 y 31,36 para el modelo exponencial y 422,23 y 40,63 para el esférico, respectivamente.

La expresión analítica del modelo de semivariograma poblacional exponencial es:

$$\hat{\gamma}_{exp}(h) = 515,78 + 5150,60 \left(1 - e^{-\frac{h}{5684,24}} \right) \qquad \forall h > 0, \tag{6.1}$$

y su representación gráfica se presenta en la Figura 6.5.

Figura 6.5: Semivariograma empírico y ajustado para el número de hogares con NBI, a partir de la totalidad de radios censales



Con la finalidad de realizar una primera evaluación de los métodos, se calcula la predicción del total de hogares con NBI para diversos modelos, incluyendo algunos que tienen en cuenta una variable auxiliar y/o la variabilidad espacial (según el semivariograma poblacional exponencial). Además se estiman los Errores Cuadráticos Medios del predictor y las eficiencias relativas, tomando como referencia el modelo más simple.

Se aplicaron los modelos de regresión, descriptos en la Sección 4.4, para predecir los valores de la variable en las unidades de la población que no fueron incluidas en la muestra. A partir de ellas se obtiene la predicción del total de hogares con NBI. Las alternativas consideradas son:

1. PLIO teniendo en cuenta un modelo sin variable auxiliar, homocedástico y sin autocorrelación.
2. PLIO teniendo en cuenta un modelo de regresión, homocedástico y sin autocorrelación.
3. PLIO teniendo en cuenta un modelo de regresión sin ordenada al origen, heterocedástico y sin autocorrelación.
4. PLIO teniendo en cuenta un modelo de regresión, homocedástico y con autocorrelación espacial.

La variable auxiliar utilizada en los modelos de regresión es el total de hogares por radio censal.

Para el cálculo del predictor del total de hogares con NBI y la estimación de su Error Cuadrático Medio se utilizan las formulas (4.1) y (4.2) respectivamente aplicadas a los casos particulares. Los resultados obtenidos se presentan en el Cuadro 6.1 conjuntamente con la estimación de la eficiencia relativa de cada propuesta con respecto al modelo más sencillo.

Cuadro 6.1: Predicción del total de hogares con NBI en la ciudad de Rosario, estimación de la raíz cuadrada del Error Cuadrático Medio de total predicho y de la eficiencia relativa, para cada propuesta

Propuesta	$\hat{Y}$	$\sqrt{E\hat{C}M[\hat{Y}]}$	$E\hat{R}_{*/se}$
Modelo sin variable auxiliar Homocedástico y sin autocorrelación	29883	3897	1,0
Modelo de regresión Homocedástico y sin autocorrelación	30122	3118	1,56
Modelo de regresión sin ordenada al origen Heterocedástico y sin autocorrelación	29980	2550	2,34
Modelo de regresión Homocedástico y con autocorrelación espacial	29679	1957	3,96

Una observación respecto de la bondad de las estimaciones obtenidas en esta muestra particular es que todas ellas fueron cercanas al valor poblacional conocido del total de hogares con NBI en la ciudad de Rosario en el año 2001 (29622). Además se aprecia que el uso de la información brindada por la variable auxiliar correlacionada con la variable en estudio ha proporcionado una mejora importante en la precisión de las estimaciones, de acuerdo a la estimación de las eficiencias relativas.

Además la incorporación del semivariograma en el modelo superpoblacional muestra una importante reducción en el Error Cuadrático Medio estimado, presentando para este caso particular una eficiencia estimada 4 veces mayor con respecto al modelo más sencillo. Esto era esperable debido a la existencia de autocorrelación espacial.

Estos resultados que provienen de la observación de una sóla muestra, se completan por medio de un estudio comparativo en el que se evalúa la calidad de los estimadores derivados del enfoque de modelos, pero teniendo en cuenta la distribución de los mismos

obtenida con las muestras posibles de la población finita (o un subconjunto de ellas).

Otro aspecto que completa la comparación es el ajuste del semivariograma. En esta sección se utilizó el semivariograma poblacional para explicar la variabilidad presente en los datos, sin embargo cabe destacar que en la mayoría de los problemas aplicados la población es desconocida y resulta necesario estimar el semivariograma con los datos de la muestra.

6.4. Estudio comparativo: Predictores en muestreo sistemático

Los enfoques basado en modelos y asistido por modelos brindan 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.

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 (Sección 4.4.4).

En esta sección, como un aporte original, se presenta un primer estudio comparativo utilizando muestras sistemáticas con período 6. Para cada una de las 6 muestras posibles de tamaño 149 se obtienen los predictores del total de hogares con NBI utilizando diferentes modelos presentados en la sección anterior, excepto el modelo de regresión homocedástico y sin autocorrelación y agregando los modelos de regresión que presentan

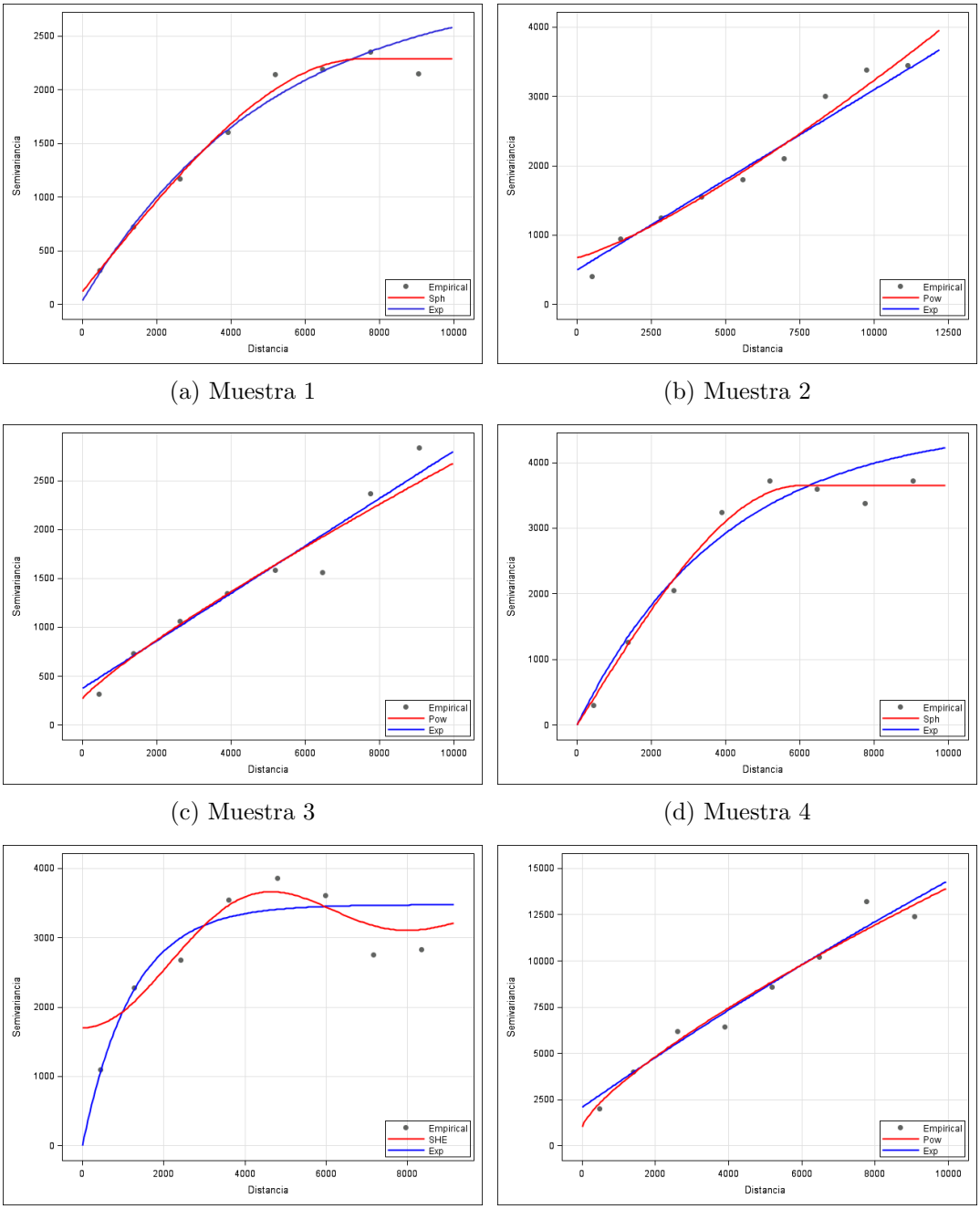
autocorrelación para distintas variantes de la estimación del semivariograma. Luego, se analiza el comportamiento de estos estimadores por medio del Error Cuadrático Medio. Las alternativas planteadas son:

1. PLIO teniendo en cuenta un modelo sin variable auxiliar, homocedástico y sin autocorrelación (asociado al estimador de simple expansión).
2. PLIO teniendo en cuenta un modelo de regresión sin ordenada al origen, heterocedástico y sin autocorrelación (asociado al estimador de razón).
3. PLIO teniendo en cuenta un modelo de regresión, homocedástico y con autocorrelación. Modelo poblacional de semivariograma exponencial (llamado *modelo poblacional exponencial*).
4. Expresión matemática del PLIO teniendo en cuenta un modelo de regresión, homocedástico y con autocorrelación. Modelo poblacional exponencial y parámetros estimados con la muestra (llamado *modelo muestral exponencial*).
5. Expresión matemática del PLIO teniendo en cuenta un modelo de regresión, homocedástico y con autocorrelación. Identificación del modelo de semivariograma y estimación de sus parámetros con la información muestral (llamado *modelo muestral*).

Para calcular el predictor del total en la 3º propuesta se utiliza el modelo se semivariograma poblacional exponencial, presentado en la sección anterior. Su representación y expresión matemática se encuentran en el Gráfico 6.5 y formula (6.1) respectivamente. En la 4º opción el cálculo de los predictores utiliza estimadores de los parámetros del modelo exponencial basándose en las observaciones muestrales. En la 5º propuesta primero debe identificarse con los datos de la muestra el modelo de semivariograma a adoptar y luego estimar sus parámetros.

En las figuras 6.6a a 6.6f se han reunido, para cada una de las 6 muestras sistemáticas, el semivariograma empírico (Empirical), el semivariograma exponencial (Exp) estimado con los datos de la muestra y usado en la 4º propuesta y el semivariograma identificado y estimado a partir de los datos de la muestra, que se utiliza en la 5º propuesta.

Figura 6.6: Semivariogramas muestrales para el número de hogares con NBI en cada muestra sistemática



(a) Muestra 1

(b) Muestra 2

(c) Muestra 3

(d) Muestra 4

(e) Muestra 5

(f) Muestra 6

Puede apreciarse que la identificación del modelo de semivariograma a partir de la muestra conduce a diferentes modelos: en la muestra 1 y 4 se identifica un modelo esférico (Sph), en las muestras 2, 3 y 6 modelos de potencia (Pow) y en la 5 se considera un modelo de efecto agujero (She).

En el Cuadro 6.2 se presentan las predicciones del total de hogares con Necesidades Básicas Inssatisfechas obtenidas para cada muestra en cada propuesta, fórmula (4.1), la raíz cuadrada del Error Cuadrático Medio a través de todas las muestras posibles, formula (4.3), y las eficiencias relativas respecto a dos estimadores distintos, el de simple expansión y el de razón.

Cuadro 6.2: Predicción del total de hogares con NBI en cada muestra sistemática, raíz cuadrada del Error Cuadrático Medio según enfoque asistido por modelo y eficiencia relativa

Propuesta	Predicción del total en las muestras sistemáticas						$\sqrt{ECM[\hat{Y}]}$	$ER_{*/se}$	$ER_{*/r}$
	1	2	3	4	5	6			
Simple expansión	25344	26862	24798	30582	29520	40626	5342	1	0,64
Razón	26254	27598	25699	30129	28798	38461	4280	1,56	1
Modelo poblacional Exponencial	26352	27669	26516	29523	27352	32890	2582	4,28	2,75
Modelo muestral Exponencial	26148	27972	26555	29045	27203	33095	2659	4,03	2,59
Modelo muestral	26231	28163	26536	29830	27248	32952	2579	4,29	2,75

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, inclusive con respecto al estimador de razón que sólo tiene en cuenta la variable auxiliar total de hogares del radio censal relacionada con la variable en estudio.

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 las predicciones del total resultaron similares.

Debe mencionarse que en 3 muestras, se encuentran semivariogramas que se identifican como el modelo de potencia. Este modelo, que no coincide con el poblacional, indica falta de estacionariedad en el proceso y pondría en tela de juicio la aplicación debido a los supuestos teóricos necesarios. Este es un motivo por el cual se sostiene la sugerencia

de utilizar un modelo poblacional de un estudio anterior o un modelo razonable para el problema estimando sus parámetros con los datos de la muestra.

Una vez evaluados los Errores Cuadráticos Medios para los estimadores a través de todas las muestras posibles (Cuadro 6.2), interesa considerar el comportamiento de las estimaciones del Error Cuadrático Medio de acuerdo al enfoque basados en modelos (formula (4.2)). Se detecta que las mismas son más parecidas entre sí para todas las muestras cuando se utiliza el modelo de semivariograma poblacional, conclusión que también lleva a sugerir que se utilice un modelo conocido.

Cuadro 6.3: Estimación de la raíz cuadrada del Error Cuadrático Medio basado en el enfoque de modelos para el número de hogares con NBI

	Estimación del $\sqrt{ECM[\hat{Y}]}$ basado en el enfoque de modelos					
Propuesta	1	2	3	4	5	6
Simple expansión	2863	2932	2687	3751	3705	6238
Razón	2404	2209	2154	2504	2554	3259
Modelo poblacional exponencial	2038	1973	2010	2038	1981	2046
Modelo muestral exponencial	1028	1659	1481	1454	2064	3561
Modelo muestral	1183	1816	1394	1335	2780	3222

El estudio realizado constituye un aporte en cuanto a mostrar la utilidad del uso de la información de la variabilidad espacial de las unidades aunque está limitado a una población determinada y a que las unidades de las muestras sistemáticas dependen del orden en que hayan estado en la población.

6.5. Estudio comparativo: Predictores en muestras aleatorias simples

En esta sección se presenta un segundo estudio comparativo empleando planes de muestreo que tienen como método de selección el muestreo simple al azar ($n = 150$) y como métodos de estimación los mismos de la sección anterior. Las comparaciones se

realizan a partir del los Errores Cuadráticos Medios según el enfoque asistido por modelos, es decir considerando la distribución de los estimadores a través de todas las muestras posibles, de la población finita.

Las formulas para los Errores Cuadráticos Medios de los estimadores del total por simple expansión y razón son expresiones cerradas que pueden calcularse sin inconvenientes, como se presentaron en la Sección 4.5. Para los estimadores que emplean información de la variabilidad espacial, deberían tomarse todas las muestras posibles, calcular cada estimación y luego el Error Cuadrático Medio. En este caso esta tarea es imposible de realizar ya que existen $\binom{894}{150}$ muestras posibles. Como alternativa, se proponen extraer 10000 muestras aleatorias simples independientes y calcular para cada una la estimación del total de hogares con NBI, y obtener el Error Cuadrático Medio según el enfoque asistido por modelos.

Previo a la presentación de los resultados, cabe hacer las algunas consideraciones. Cuando se estima la característica de interés en cada una de las muestras, asumiendo un *modelo poblacional exponencial*, no hay inconvenientes ya que las especificaciones de los parámetros del modelo se dan a priori y el único cálculo es el necesario para obtener las estimaciones del total de hogares con NBI según formula (4.1), y a partir de ellas obtener el Error Cuadrático Medio (formula (4.3)).

Cuando la estimación se realiza asumiendo un *modelo exponencial muestral* se agrega la dificultad de estimar sus parámetros con los datos de la muestra. Es decir, primero con cada muestra se estiman los parámetros del modelo de semivariograma exponencial, luego se predice el total de hogares con NBI y por último se calcula el Error Cuadrático Medio según la formula (4.3).

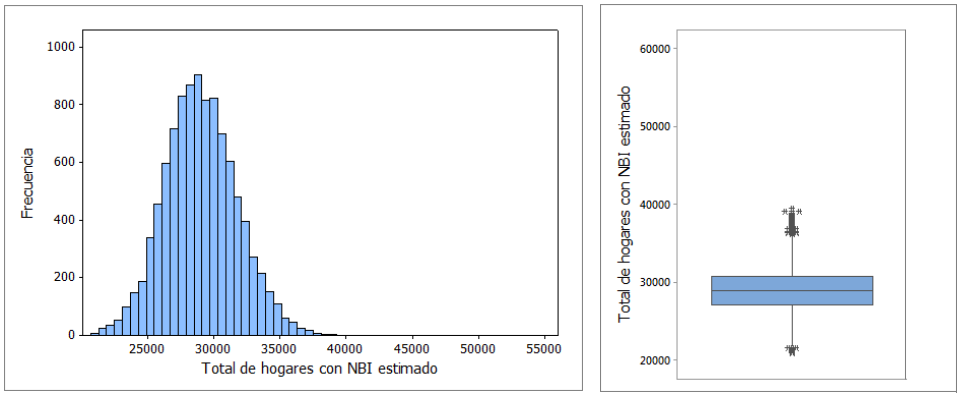
Cuando la elección del mejor modelo de semivariograma y la estimación de sus parámetros se realiza con los datos de la muestra, la tarea es más compleja. En un primer paso, para cada muestra se debe seleccionar el mejor modelo de semivariograma que ajusta los datos para luego estimarlo. Cuando se trabajó con las 6 muestras sistemáticas este paso pudo realizarse manualmente, observando distintas alternativas de semivariograma

y eligiendo aquella que presentara menor valor de $R(\boldsymbol{\theta}_q)$ o menor AIC y que además se aproxime mejor al semivariograma empírico, más allá de que el PROC VARIOGRAM del programa SAS 9.3 puede seleccionar el mejor modelo automáticamente según el criterio del mínimos cuadrados ponderados. Por lo tanto, al trabajar con una cantidad tan grande de muestras este cálculo manualmente resulta casi imposible de llevar a cabo. Debido a esto se decidió que el programa, seleccione automáticamente el mejor modelo de semivariograma. Luego se predice el total de hogares con NBI para cada muestra y se calcula el Error Cuadrático Medio en la población finita (asistido por modelos). Esta tarea incrementa el tiempo de procesamiento en un 40 %. Los programas utilizados se incluyen en el Anexo.

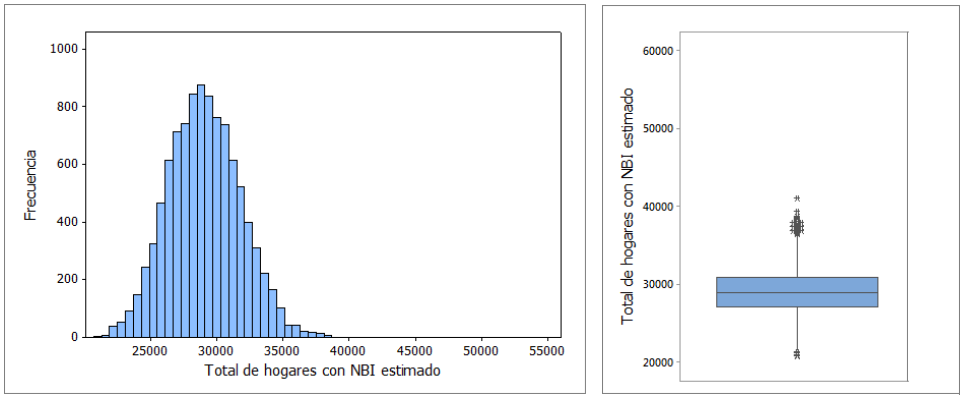
Una vez calculados las predicciones del total de hogares con NBI para cada muestra, se presenta en primer lugar un resumen descriptivo de los valores de las predicciones. La Figura 6.7 contiene los histogramas de frecuencia y "box plot"para la variable total de hogares con NBI estimado, para cada uno de los posibles modelos de semivariograma presentados y en el Cuadro 6.4 se exhiben medidas descriptivas para esta varibale. De la observación de estos resúmenes se encuentra que el *modelo poblacional exponencial* es el que presenta menor dispersión o en otras palabras, mayor precisión.

Los promedios del total de hogares con NBI estimado resultan muy similares para cualquiera de los modelos considerados, al igual que los desvíos estándar. El 50 % de las estimaciones del total de hogares con NBI presentan valores inferiores a 28916 para el *modelo poblacional exponencial*, 28977 para el *modelo muestral exponencial* y 29213 para el *modelo muestral*, recordando que el total de hogares con NBI para la ciudad de Rosario es 29622 según datos del censo.

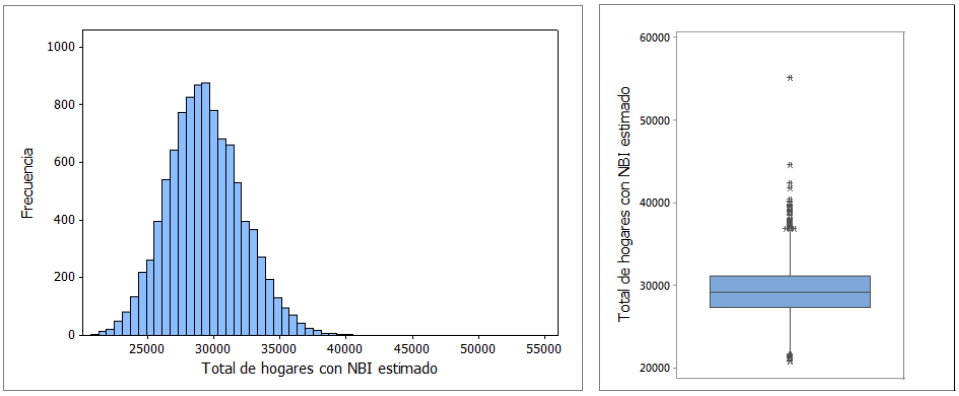
Figura 6.7: Histograma de frecuencia y "box plot"para la predicción del total de hogares con NBI



(a) *Modelo poblacional Exponencial*



(b) *Modelo muestral Exponencial*



(c) *Modelo muestral*

Cuadro 6.4: Medidas descriptivas para la predicción del total de hogares con NBI

Propuesta	Media	Desvío estándar	Mínimo	Q_1	Q_2	Q_3	Máximo
Modelo poblacional Exponencial	29032	2719	20995	27169	28916	30805	39594
Modelo muestral Exponencial	29069	2728	20803	27148	28977	30897	41154
Modelo muestral	29352	2836	20854	27417	29213	31195	55189

El Cuadro 6.5 presenta el Error Cuadrático Medio en la población finita para el estimador de simple expansión del total de hogares con NBI (asociado al modelo sin variable auxiliar, homocedástico y sin autocorrelación), de razón (asociado al modelo de regresión sin ordenada al origen, heterocedástico y sin autocorrelación) y estimadores que tienen en cuenta la variablidad espacial según los diferentes modelos de semivariograma. El Error Cuadrático Medio para el estimador de simple expansión y de razón se calculan con la tradicional forma de la variancia poblacional del total. Para los restantes estimadores se utilizan valores aproximados obtenidos con la formula del Error Cuadrático Medio en la población finita con los datos de las 10000 muestras.

Cuadro 6.5: Raíz cuadrada del Error Cuadrático Medio según enfoque asistido por modelos y eficiencia relativa

Propuesta	$\sqrt{ECM[\hat{Y}]}$	$ER_{*/se}$	$ER_{*/r}$
Simple expansión	3898	1	0,72
Razón	3318	1,38	1
Modelo poblacional Exponencial	2782	1,96	1,42
Modelo muestral Exponencial	2783	1,96	1,42
Modelo muestral	2849	1,87	1,36

Se observa que siempre que se tenga en cuenta la variabilidad espacial, con cualquier de los 3 modelos de semivariograma propuestos, la eficiencia relativa resulta de aproximadamente 2 veces en comparación con el estimador de simple expansión. En cambio, al considerar el estimador de razón que tiene como variable auxiliar al total de hogares de cada radio censal, los estimadores que utilizan la variabilidad espacial son más eficientes pero esa eficiencia resulta alrededor del 40 %.

Al igual que con el muestro sistemático cualquiera de los 3 modelos planteados para estimar el semivariograma presentan valores muy similares de Error Cuadrático Medio, mostrando una pequeña diferencia cuando se trabaja con el *modelo poblacional exponencial* o *modelo muestral exponencial*.

6.6. Aplicación para datos en látices

En un análisis estadístico de datos que están dispuestos en látices, es primordial determinar (1) si las unidades son regulares o irregulares, (2) si representan puntos o regiones, y (3) si en ellas se miden variables continuas o discretas. Los datos de conteo procedentes de regiones geopolíticas vecinas ofrecen un desafío particular debido a que las unidades se encuentran en látices irregulares, son regiones y la variable es discreta.

Esta sección se refiere a la utilización de los métodos de selección tradicionales para información que se encuentra dispuesta en látices regulares. Los datos del censo de Población, Hogares y Vivienda que han sido considerados en esta tesis podrían ser tratados como un caso particular de látices regulares cambiando la unidad de análisis.

A continuación se presenta una comparación de los diferentes planes de muestreo del enfoque de diseño, descriptos en el Capítulo 5, para estimar el total de hogares con Necesidades Básicas Insatisfechas en la ciudad de Rosario, a partir de una muestra aleatoria, suponiendo que esta variable presenta variablidad espacial.

Para llevar a cabo este estudio se propone "cubrir" a la ciudad de Rosario con una grilla cuadrangular de 40*40 unidades, siguiendo la metodología utilizada por Cressie (1993). Luego se determina la ubicación de los centroides de los radios censales en estas unidades y se encuentra que existen unidades sin radios censales, otras con un sólo un radio censal y las restantes unidades pueden tener más de un radio. Dentro de la grilla se forman bloques de diversos tamaños (20*20, 10*10, 8*8, 5*5) para poder comparar la precisión de los métodos de selección.

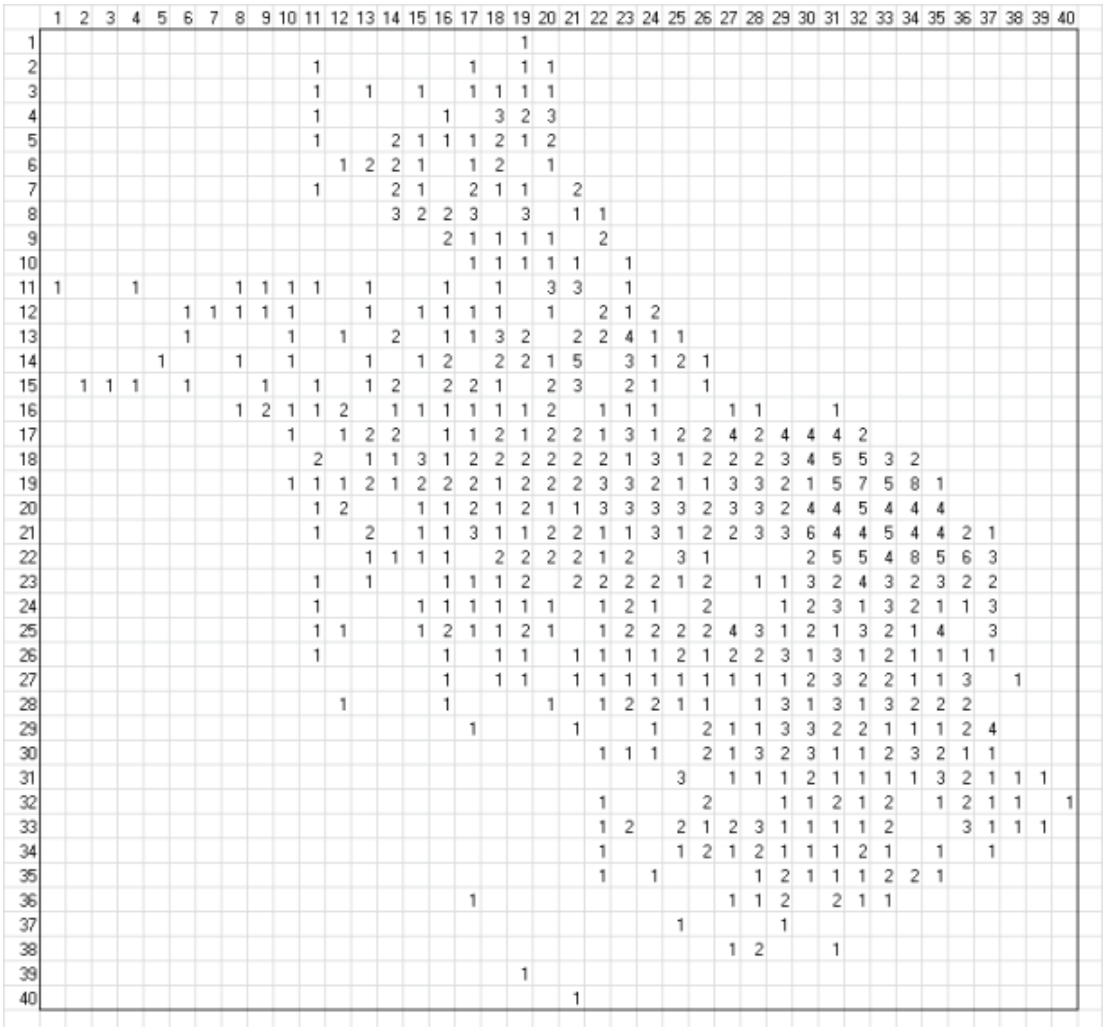
Los métodos de selección empleados son muestreo aleatorio simple, muestreo estratificado (donde cada bloque es un estrato) y muestreo sistemático y empleando como método de estimación el de simple expansión bajo el enfoque basado en diseño.

La Figura 6.4 presenta la grilla para la ciudad de Rosario con la cantidad de centroides de radios censales que se encuentran en cada unidad. Como puede observarse la forma geográfica de la ciudad se mantiene, y se detecta que hay muchas unidades que no tienen

radios censales, lo cual constituye un inconveniente a la hora de estimar el total de hogares con Necesidades Básicas Insatisfechas.

La solución adoptada para tratar este problema es no tener en cuenta para el análisis aquellos bloques que no tienen ningún elemento, es por este motivo que el valor del total del unidades, FC, no es igual a 1600 para todos los tamaños de bloque. En el cálculo de las variancias del total estimado para cada uno de los métodos de selección analizados, se tuvieron en cuenta todas las celdas cuyo bloque tenía alguna unidad, considerando aquellas unidades vacías con un valor de total de hogares con NBI igual a 0.

Figura 6.8: Cantidad de centroides de los radios censales en la grilla definida



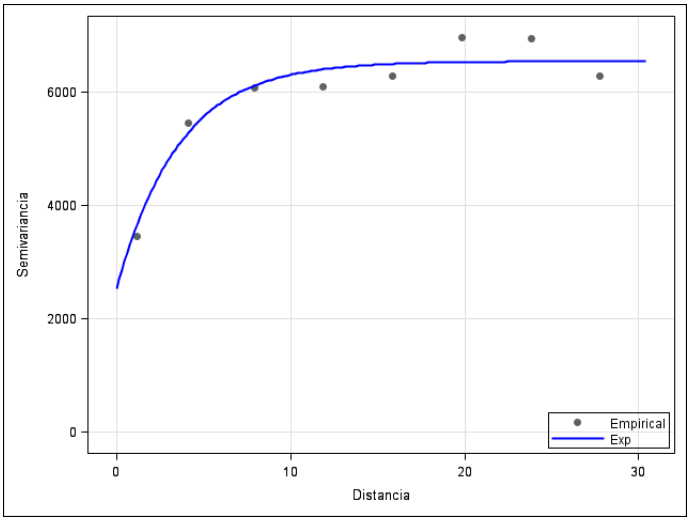
En primer lugar se presenta una descripción de la variabilidad espacial existente en la población de 513 unidades con datos, información que puede orientar en la elección del período a emplear en la selección de las muestras sistemáticas, a partir del semivariograma

poblacional. Para determinar el mejor modelo, se eligió aquel que fuese lo más parecido al semivariograma empírico y que presentara el menor valor de $R(\boldsymbol{\theta}_q)$ y de Akaike. El modelo de semivariograma seleccionado, corresponde al modelo exponencial, cuya especificación es:

$$\hat{\gamma}_{exp}(h) = 2531,24 + 4014,65 \left(1 - e^{-\frac{h}{3,54}} \right) \qquad \forall h > 0.$$

A continuación se presenta el semivariograma empírico conjuntamente con el modelo de semivariograma exponencial estimado.

Figura 6.9: Modelo de semivariograma empírico y ajustado para el número de hogares con NBI, a partir de la totalidad de las 513 unidades de muestreo



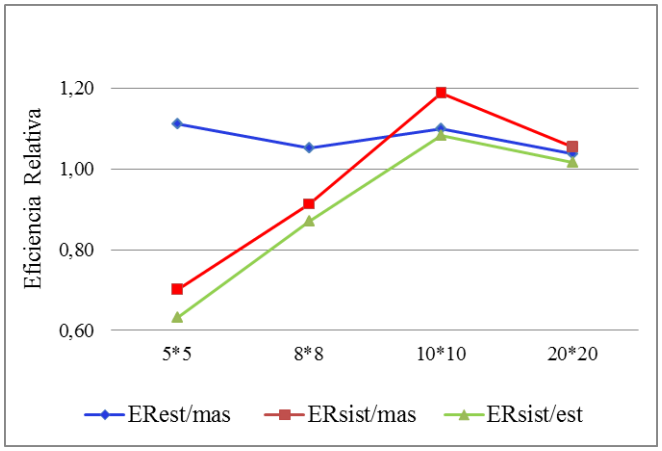
Se detecta que la meseta se alcanza de forma asintótica y el rango efectivo resulta aproximandamente igual a 10 $(3 * 3,54)$, distancia para el cual el semivariograma toma un valor igual al 95 % del valor de la meseta. Se concluye que para distancias inferiores a 10 existe correlación espacial entre las unidades, mientras que para distancias entre las unidades mayores a 10, la correlación es nula.

A continuación, se presentan los resultados en términos de eficiencia para los tres métodos de selección utilizados teniendo en cuenta los diferentes tamaños del bloque y también un gráfico comparativo para su mayor compresión.

Cuadro 6.6: Eficiencias relativas para diversos tamaños de bloque

Tamaño del bloque	$ER_{est/mas}$	$ER_{sist/mas}$	$ER_{sist/est}$	FC	n
5*5	1,11	0,70	0,63	925	37
8*8	1,05	0,91	0,87	1088	17
10*10	1,10	1,19	1,08	1200	12
20*20	1,04	1,06	1,02	1600	4

Figura 6.10: Eficiencias relativas para diversos tamaños de bloque



Los diferentes planes de muestreo usuales para datos espaciales dispuestos en látices regulares, muestran que:

- El muestreo sistemático con período 10, rango efectivo del semivariograma, y 20 resulta ser el más eficiente frente al muestreo aleatorio simple y al muestreo estratificado.
- El muestreo estratificado resulta más eficiente que el muestreo aleatorio simple en todas las situaciones, encontrándose la máxima eficiencia cuando el tamaño del estrato es el menor (5*5).
- El muestreo estratificado resulta más eficiente que el muestreo sistemático excepto en la situación señalada anteriormente, es decir para bloques de 10*10 y de 20*20.

Consideraciones finales

Los enfoques conocidos como de modelos o de predicción y asistido por modelos, en el muestreo de poblaciones finitas, han ampliado las posibilidades de utilización de información auxiliar, muchas veces disponible en los marcos muestrales, para la mejora de los planes de muestreo.

En esta tesis se han presentado propuestas para el uso de información auxiliar relativa a la variabilidad espacial que puede existir en unidades de muestreo asociadas a una localización geográfica, comparándose con métodos tradicionalmente aplicados, en los que no se tiene en cuenta dicha información.

Para mostrar la utilidad de la metodología planteada se ha considerado un problema con características particulares, como lo es la estimación de cantidades poblacionales en estudios socioeconómicos. En particular se estudia la predicción del total de hogares con Necesidades Básicas Insatisfechas en la ciudad de Rosario en el año 2001, a partir de una muestra de radios censales.

En forma sintética, se han expuesto herramientas usuales para el análisis exploratorio de datos espaciales, pensando que la descripción y caracterización de este fenómeno es un paso fundamental para sustentar los análisis posteriores. Estos métodos han sido aplicados satisfactoriamente en el problema considerado en esta tesis, encontrando que existe variabilidad espacial en el número de hogares con Necesidades Básicas Insatisfechas medido a cada radio censal. Este hecho da lugar a la posibilidad de construcción de un modelo de semivariograma a emplear en la fase de estimación.

Luego se han desarrollado los conceptos fundamentales de los modelos de semivariograma, útiles para expresar la variabilidad como función de la distancia que separa a

las unidades, y aquellos procedimientos necesarios para la identificación y selección del modelo y la estimación de sus parámetros. La variabilidad espacial del número de hogares con Necesidades Básicas Insatisfechas en la ciudad de Rosario pudo ser expresada adecuadamente mediante un modelo de semivariograma exponencial.

Como paso siguiente, se ha planteado la estimación de un total, de acuerdo al enfoque de modelos utilizando cuatro diferentes modelos de regresión: tres de ellos no emplean información de la variabilidad espacial (se identifican con los estimadores de simple expansión, regresión y razón) y el restante tiene en cuenta esta información, la cual es incorporada al proceso de predicción mediante un modelo de semivariograma. Se ha planteado en cada caso el predictor lineal insesgado y óptimo del total, su Error Cuadrático Medio y la estimación del mismo.

Esta metodología se ha implementado, obteniendo las predicciones del total de hogares con Necesidades Básicas Insatisfechas en Rosario a partir de una muestra aleatoria de acuerdo a los siguientes modelos considerados:

1. Modelo sin variable auxiliar, homocedástico y sin autocorrelación, coincidente con el clásico estimador de simple expansión en el enfoque de diseño.
2. Modelo de regresión, considerando como variable auxiliar el total de hogares en el radio censal, homocedástico y sin autocorrelación, identificado con el estimador de regresión en el enfoque de diseño.
3. Modelo de regresión sin ordenada al origen, utilizando como variable auxiliar el total de hogares en el radio censal, heterocedástico y sin autocorrelación, coincidente con el estimador de razón en el enfoque de diseño.
4. Modelo de regresión considerando como variable auxiliar el total de hogares en el radio censal, homocedástico y con autocorrelación espacial, utilizando el semivariograma encontrado con los datos de la población.

Se ha encontrado, para esta muestra, que la estimación del Error Cuadrático Medio del predictor obtenida con el primer modelo, identificado con el estimador de simple expansión, es cuatro veces la obtenida empleando un modelo que tiene en cuenta la variabilidad espacial. También, se ha observado que el Error Cuadrático Medio del tercer

modelo correspondiente al estimador de razón, es casi dos veces mayor que el que tiene en cuenta la correlación espacial.

Como un aporte original, se han realizado dos estudios comparativos del comportamiento de los predictores planteados, considerando su distribución en el muestreo de la población finita para los casos de muestreo sistemático y muestreo aleatorio simple. En ambos estudios se excluyó el estimador de regresión y se consideraron tres alternativas para el cálculo del semivariograma:

- Semivariograma a partir de los datos de la población. Modelo exponencial.
- Modelo exponencial pero estimado con los datos de la muestra.
- Identificación del modelo de semivariograma y estimación con los datos de la muestra.

En el primer estudio, se han seleccionado las 6 muestras sistemáticas con período 6, y se han calculado las eficiencias relativas tomando como referencia el estimador de simple expansión y de razón. Los estimadores que utilizan el modelo de semivariograma han presentado un mejor desempeño, con eficiencias relativas de aproximadamente 4 con respecto al de simple expansión, y casi 2 veces con respecto al de razón.

En el segundo estudio comparativo, se ha evaluado el comportamiento de los estimadores en el muestreo aleatorio simple. Los Errores Cuadráticos Medios de los estimadores de simple expansión y de razón pueden calcularse mediante sus expresiones matemáticas exactas, en cambio, para los estimadores que emplean el modelo de semivariograma no hay fórmulas para el Error Cuadrático Medio en el muestreo de la población finita. Para estos casos, se requeriría considerar todas las muestras posibles y este número resulta muy elevado, motivo por el cual se han extraído 10000 muestras aleatorias simples y se han obtenido aproximaciones al valor del Error Cuadrático Medio.

Las eficiencias relativas de los estimadores que utilizan información de la variabilidad espacial son cercanas a dos (es decir, sus Errores Cuadráticos Medios son cercanos a la mitad del correspondiente al estimador de simple expansión). Con respecto al estimador de razón, su comportamiento ha sido superior al de simple expansión, pero ha resultado

superado en precisión, por los que emplean el modelo de semivariograma.

Cabe mencionar también que los estimadores que utilizan el modelo de semivariograma poblacional (exponencial), han logrado ser un poco más eficientes que aquel en el que se decide el modelo con los datos de la muestra.

Por último, se ha realizado un estudio atendiendo a una forma usual de obtener muestras en poblaciones cuyas unidades se encuentran distribuidas en un área, a través del muestreo aleatorio simple, muestreo estratificado y muestreo sistemático. Estos procedimientos se han propuesto por ejemplo, en el campo de la agricultura (Iglesias, 1998) cuando se divide el área total en látices regulares. Para adaptar los radios censales, que son látices irregulares, a la propuesta mencionada, se superpuso una grilla al mapa de Rosario, y se asignaron los centroides de cada radio censal a la correspondiente retícula de pertenencia, como lo sugiere Cressie (1993).

Se han utilizado estimadores del total de simple expansión, y los métodos de selección fueron: muestreo sistemático con períodos 5, 8, 10 y 20, muestreo aleatorio simple con los tamaños de muestra dados por el muestreo sistemático, y muestreo estratificado con los mismos tamaños muestrales. El muestreo sistemático con períodos 10 y 20 ha resultado más eficiente que los demás, siendo 10 el rango efectivo del semivariograma, y por lo tanto demostrando que esta información puede ser útil para la selección adecuada del período.

El trabajo realizado ha permitido verificar importantes beneficios que puede brindar el uso de la información espacial en la mejora de planes de muestreo, que hacen recomendable el uso de esta metodología. Sin embargo, a lo largo de la investigación se han ido planteando interrogantes que han dado lugar a las siguientes líneas de investigación:

- Estudio de la eficiencia de los estimadores que emplean variabilidad espacial de acuerdo al tamaño de la muestra, en diferentes esquemas de selección.
- Elaboración de una propuesta metodológica para el muestreo bietápico, ya que podría plantearse la selección de radios censales y luego de segmentos. En este caso se debería estudiar la forma de tratar el fenómeno de la variabilidad espacial de las

unidades primarias y también la que podría existir para las unidades secundarias en cada unidad primaria.

- Desarrollo de propuestas, que agreguen, la consideración de modelos que sean específicos para variables de conteo, ya sea para el caso de una etapa o de dos etapas.
- Planteamiento de estimadores de la variancia basados en métodos de remuestreo o bootstrap.
- Profundización del estudio del impacto sobre la precisión de los estimadores, que puede tener la estimación del semivariograma a partir de los datos de la muestra.

Bibliografía

Ambrosio, L. (1999). *Muestreo*. E.T.S.I.A. Madrid.

Ambrosio, L. (2000). *Estadística Espacial*. E.T.S.I.A. Madrid.

Ambrosio, L. (2001). Modelos para el diseño y análisis de encuestas. *Escuela Técnica Superior de Ingenieros Agrónomos de la Universidad Politécnica de Madrid (España)y Escuela de Estadística de la Universidad Nacional de Rosario (Argentina)*.

Ambrosio, L. (2006). Estimación del total con datos de conteo sobredispersos y espacio-temporalmente correlacionados: una aproximación basada en la predicción. *Curso de las Jornadas Internacionales de Estadística*.

Ambrosio, L. e Iglesias, L. (2000). Small area estimation by ground survey and remote sensing. *Remote sensing of environment, Elsevier*.

Ambrosio, L. e Iglesias, L. (2014). Technical reports on identifying the most appropriate sampling frame for specific landscape types. *Improving Agricultural and Rural Statistics, FAO*.

Ambrosio, L., Iglesias, L., y Marín, C. (2003). Systematic simple design for the estimation of spatial means. *Environmetrics*, 14:45–61.

Ambrosio, L., Iglesias, L., Marín, C., y Del Monte, J. P. (2004). Evaluation of sampling methods and assessment of the sample size to estimate the weed seedbank in soil, taking into account spatial variability. *Weed Research*, 44:224–236.

Ambrosio, L., Iglesias, L., Marín, C., Pascual, V., y Serrano Bermejo, A. (2008). A

spatial high-resolution model of the dynamics of agricultural land use. *Agr Econ*, 38:233–245.

Ambrosio, L., Marín, C., Iglesias, L., Pascual, V., Fuertes, A., y Mena, M. A. (2009). Agricultural and environmental information systems: the integrating role of area samples. *Spanish Journal of Agricultural Research*, 7:957–973.

Anselin, L. (1995). Local indicators of spatial association-lisa. *Geographical Analysis*, 27(2):93–115.

Anselin, L. (1999). *Interactive techniques and exploratory spatial data analysis*. John Wiley, New York, NY.

Anselin, L., S. I. y O., S. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. *New Tools in Spatial Data Analysis, Proceedings of a Workshop. Center for Spatially Integrated Social Science, University of California, Santa Barbara*, (CD-ROM).

Chambers, R. L. y Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press.

Chasco Yrigoyen, C. (2003). Métodos gráficos del análisis exploratorio de datos espaciales. *Asociación Española de Economía Aplicada, ASEPELT*.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17(5):563–570.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, N. Y. 990p.

Haining, R. (2003). *Spatial Data Analysis, Theory and Practice*. Cambridge University Press.

Iglesias, L. (1998). Tesis doctoral: Muestreo de áreas: Diseño de muestras y estimación en pequeñas áreas. *Escuela Técnica Superior de Ingenieros Agrónomos. Universidad Politécnica de Madrid. España*.

Indec (2003). Aquí se cuenta. *Revista Informativa del Censo 2001*, 7.

Matheron, G. (1962). *Traité de géostatistique appliquée*, volumen 14. Editions Technip, Paris.

Moran, P. A. (1950). Notes on continuos stochastic phenomena. *Biometrika*, 37:17–33.

SAS/STAT 9.3 User's Guide, . (2011a). *The KRIGE2D Procedure (Charter)*. Institute Inc., Cart. N.C, USA.

SAS/STAT 9.3 User's Guide, . (2011b). *The VARIOGRAM Procedure (Charter)*. Institute Inc., Cart. N.C, USA.

Särndal, C. E., Swensson, B., y Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag New York.

Thompson, S. K. (1992). *Sampling*. John Wiley y Sons Inc.

Thompson, S. K. (2012). *Sampling, 3rd Edition*. John Wiley y Sons Inc.

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:34–240.

Wang, J. F., Stein, A., Gao, B. B., y Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics. ELSEVIER*, pp. 1–14.

Apendice

Generador de 10000 muestras aleatorias simples suponiendo un modelo de semivariograma exponencial poblacional.

```
libname a "D:\generadores finales";
%macro Generar_yest_y_ECM;
%do i=1 %to 10000;

/* REALIZA UNA MUESTRA ALEATORIA SIMPLE */
proc surveyselect data=a.base
   method=srs n=150 out=MAS&i. outseed;
run;

/*CREA UN CONJUNTO DATOS PARA SE UTILIZADO EN EL PROCEDIMIENTO IML */
PROC SQL;
CREATE TABLE no_MAS&i. AS
SELECT DISTINCT a.*
FROM a.base a WHERE i NOT IN (SELECT i FROM MAS&i.);
QUIT;

proc iml;

/* usa la muestra i para crear la matriz de datos    */
use MAS&i.;
read all var{x y thogares sinbi InitialSeed} into datos;

/* creación de matriz de datos no incluidos en la muestra   */
use no_MAS&i.;
read all var{x y thogares} into datosnm;
d1=nrow(datos);
uno1=j(d1,1,1);
v=j(d1,d1,0);
/* diagonal de v   */
  do j2=1 to d1;
  v[j2,j2]=3399.4444486;
  end;
/*                */
/* cálculo de los elementos de fuera de la diagonal de v */
```

```
  do j3=1 to d1;
    do j4=j3+1 to d1;
/* cálculo de la distancia */
    h=sqrt((datos[j3,1]-datos[j4,1])**2+(datos[j3,2]-datos[j4,2])**2);
/* cálculo del semivariograma estimado */
g=515.78+5150.60*(1-exp((-1)*h/5684.24));/* modelo exponencial poblacional*/
v[j3,j4]=3399.4444486-g;
    v[j4,j3]=v[j3,j4];
    end;
  end;
/* fin de cálculo de lementos de v   */

x=j(d1,1,1)||datos[,3];

/* estimación de los beta  */
bestim= inv(x'*inv(v)*x)* x' * inv(v) * datos[,4];

/* cálculo de V_N-n,n que se llamará cova              */
  d2=nrow(datosnm);
  cova=j(d2,d1,0);
  do j21=1 to d2;
    do j22=1 to d1;
    h=sqrt((datosnm[j21,1]-datos[j22,1])**2+(datosnm[j21,2]-datos[j22,2])**2);

/* cálculo del semivariograma estimado */
g=515.78+5150.60*(1-exp((-1)*h/5684.24));/* modelo exponencial poblacional*/
    cova[j21,j22]=3399.4444486-g;
    end;
  end;

  /* predicción del vector de unidades no incluidas en la muestra   */
  xnomuest=j(d2,1,1)||datosnm[,3];
  yipred=xnomuest*bestim;
  d3=nrow(yipred);
  ytotalest= sum(datos[,4])+sum(yipred+cova*inv(v)*(datos[,4]-x*bestim));

  /*calculo de ecm*/
  omega1=cova*inv(v)*x;
  omega2=x'*inv(v)*x;
  w=cova*inv(v)*cova';
  uno2=j(d2,1,1);

  /*matriz V_N-n,N-n se llama cova2*/
  cova2=j(d2,d2,0);

  /* diagonal de V_N-n,N-n   */
  do j5=1 to d2;
  cova2[j5,j5]=3399.4444486;
```

```
  end;

  /* cálculo de los elementos de fuera de la diagonal de v */
  do j6=1 to d2;
    do j7=j6+1 to d2;

 /* cálculo de la distancia */
    h=sqrt((datosnm[j6,1]-datosnm[j7,1])**2+(datosnm[j6,2]-datosnm[j7,2])**2);

 /* cálculo del semivariograma estimado */
g=515.78+5150.60*(1-exp((-1)*h/5684.24));/* modelo exponencial poblacional*/
    cova2[j6,j7]=3399.4444486-g;
    cova2[j7,j6]=cova2[j6,j7];
    end;
  end;
 /* fin de cálculo de lementos de cova2  */

  y=894*((d1/894)/d1*uno1‘*datos[,4]+(1-d1/894)/(894-d1)*uno2‘*
  (xnomuest*bestim+cova*inv(v)*(datos[,4]-x*bestim)));
  ECM=uno2‘*((xnomuest-omega1)*inv(omega2)*(xnomuest-omega1)‘+(cova2-w))*uno2;
  Sem=datos[1,5];
  print y ECM sem;

  *ALMACENA LOS RESULTADOOS EN UN DATASET EXTERNO;
  resul=shape(0,1,3);
  resul[1,1]= y;
  resul[1,2]= ECM;
  resul[1,3]= datos[1,5];
  create a.resultad from resul[colname={"yest" "ECM" "Semilla"}];
  append from resul;
  data aleat&i.;
set a.resultad;
  run;
  proc delete data=MAS&i.;run;
  proc delete data=no_MAS&i.;run;
  dm "log;clear;out;clear";
%end;
quit;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT1;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_1 separated by " "
from sashelp.vtable
where memname like '%ALEAT1%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 1;
```

```
data a.aleat_1;
set &lista_tablas_1.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT2;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_2 separated by " "
from sashelp.vtable
where memname like '%ALEAT2%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 2;
data a.aleat_2;
set &lista_tablas_2.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT3;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_3 separated by " "
from sashelp.vtable
where memname like '%ALEAT3%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 3;
data a.aleat_3;
set &lista_tablas_3.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT4;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_4 separated by " "
from sashelp.vtable
where memname like '%ALEAT4%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 4;
data a.aleat_4;
set &lista_tablas_4.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT5;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_5 separated by " "
from sashelp.vtable
```

```
where memname like '%ALEAT5%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 5;
data a.aleat_5;
set &lista_tablas_5.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT6;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_6 separated by " "
from sashelp.vtable
where memname like '%ALEAT6%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 6;
data a.aleat_6;
set &lista_tablas_6.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT7;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_7 separated by " "
from sashelp.vtable
where memname like '%ALEAT7%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 7;
data a.aleat_7;
set &lista_tablas_7.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT8;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_8 separated by " "
from sashelp.vtable
where memname like '%ALEAT8%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 8;
data a.aleat_8;
set &lista_tablas_8.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT9;
```

```
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_9 separated by " "
from sashelp.vtable
where memname like '%ALEAT9%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 9;
data a.aleat_9;
set &lista_tablas_9.;
run;

*UNE TODOS LOS DATASET;
data a.aleat;
set a.aleat_1 a.aleat_2 a.aleat_3 a.aleat_4 a.aleat_5
a.aleat_6 a.aleat_7 a.aleat_8 a.aleat_9;
run;

*BUSCAR LAS OBSERVACIONES QUE ESTÁN REPETIDAS EN LA SEMILLA;
proc sql;
create table a.duplicados (where=(frec>1)) as select
semilla,
count(*) as frec
from a.aleat
group by 1;
quit;

/* CREA UN DATASET SIN LAS OBSERVACIONES DUPLICADAS EN EL NÚMERO DE SEMILLA*/
PROC SQL;
CREATE TABLE a.aleat_sin_dupli AS
SELECT DISTINCT a.*
FROM a.aleat a WHERE semilla NOT IN (SELECT semilla FROM a.duplicados);
QUIT;

/* CREA UN DATASET SIN LAS OBSERVACIONES QUE CERO EN "ECM" e "Y" */
DATA a.aleat_sin_dupli;
SET a.aleat_sin_dupli;
IF (yest=0) and (ECM=0) then delete;
run;

run;
%mend;
%Generar_yest_y_ECM;
```

Generador de 10000 muestras aleatorias simples suponiendo un modelo de semivariograma exponencial poblacional y estimando los parámetros con los datos de la muestra.

```
%macro Generar_yest_y_ECM;
%do i=1 %to 10;
```

```
*realiza una MSA de n=150 del conjunto de Datos;
proc surveyselect data=a.base
   method=srs n=150 out=MAS&i. outseed;
run;

*Crea el complemento de la muestra en el conjunto de datos;
PROC SQL;
CREATE TABLE no_MAS&i. AS
SELECT DISTINCT a.*
FROM a.base a WHERE i NOT IN (SELECT i FROM MAS&i.);
QUIT;

*Crear una salida a un archivo de los resultados del
PROC VARIOGRAM que se corre para la muestra;
ODS OUTPUT Variogram.SINBI.Model.Angle1.SemivModel1.ParameterEstimates=SAL&i.;

proc variogram data=MAS&i. outv=a.semivar_1 ;
store out=semi / label='VIR';
compute lagd=1300 maxlag=7 cl robust autocorr(assum=random);
coordinates xc=x yc=y;
model form=( exp );
var sinbi;
run;

ODS OUTPUT CLOSE;

*Coloca en un DataSet solo los parámetos nugget, Scale y Range;
data a.salida ;
set SAL&i. (KEEP=Estimate);
if (Estimate=0) then Estimate=.;
run;

proc iml;

*Toma los tres parámetros nugget, Scale y Range;
use a.salida;
read all into par;
close a.salida;
print par;
nugget=par[1];
scale=par[2];
range=par[3];

print nugget scale range;

*toma el valor de la covariancia;
use a.semivar_1;
```

```
read all into semivar;
close a.semivar_1;
covar=semivar[1,7];
print covar;

* Crear una tablas de n filas y dos columnas donde se almacenaran
el "y" estimado con ECM respectivos;

/* usa la muestra para la creación de matriz de datos    */
use MAS&i.;
read all var{x y thogares sinbi InitialSeed} into datos;

/* creación de matriz de datos no incluidos en la muestra   */
use no_MAS&i.;
read all var{x y thogares} into datosnm;
d1=nrow(datos);
uno1=j(d1,1,1);
v=j(d1,d1,0);

/* diagonal de v   */
  do j2=1 to d1;
  v[j2,j2]=covar;
  end;

/* cálculo de los elementos de fuera de la diagonal de v */
  do j3=1 to d1;
    do j4=j3+1 to d1;

/* cálculo de la distancia */
    h=sqrt((datos[j3,1]-datos[j4,1])**2+(datos[j3,2]-datos[j4,2])**2);

/* cálculo del semivariograma estimado */
g=nugget+scale*(1-exp((-1)*h/range));/* modelo exponencial poblacional*/
v[j3,j4]=covar-g;
    v[j4,j3]=v[j3,j4];
    end;
  end;
/* fin de cálculo de lementos de v   */

x=j(d1,1,1)||datos[,3];

/* estimación de los beta  */
  bestim= inv(x'*inv(v)*x)* x' * inv(v) * datos[,4];

/* cálculo de V_N-n,n que se llamará cova               */
  d2=nrow(datosnm);
  cova=j(d2,d1,0);
  do j21=1 to d2;
```

```
    do j22=1 to d1;
    h=sqrt((datosnm[j21,1]-datos[j22,1])**2+(datosnm[j21,2]-datos[j22,2])**2);

/* cálculo del semivariograma estimado */
g=nugget+scale*(1-exp((-1)*h/range));/* modelo exponencial poblacional*/
    cova[j21,j22]=covar-g;
    end;
  end;

/* predicción del vector de unidades no incluidas en la muestra   */
  xnomuest=j(d2,1,1)||datosnm[,3];
  yipred=xnomuest*bestim;
  d3=nrow(yipred);
  ytotalest= sum(datos[,4])+sum(yipred+cova*inv(v)*(datos[,4]-x*bestim));

/*calculo de ecm*/
  omega1=cova*inv(v)*x;
  omega2=x`*inv(v)*x;
  w=cova*inv(v)*cova`;
  uno2=j(d2,1,1);

/*matriz V_N-n,N-n se llama cova2*/
  cova2=j(d2,d2,0);

/* diagonal de V_N-n,N-n   */
  do j5=1 to d2;
  cova2[j5,j5]=covar;
  end;

/* cálculo de los elementos de fuera de la diagonal de v */
  do j6=1 to d2;
    do j7=j6+1 to d2;

/* cálculo de la distancia */
    h=sqrt((datosnm[j6,1]-datosnm[j7,1])**2+(datosnm[j6,2]-datosnm[j7,2])**2);

/* cálculo del semivariograma estimado */
g=nugget+scale*(1-exp((-1)*h/range));/* modelo exponencial poblacional*/
    cova2[j6,j7]=covar-g;
    cova2[j7,j6]=cova2[j6,j7];
    end;
  end;

/* fin de cálculo de lementos de cova2  */
  y=894*((d1/894)/d1*uno1`*datos[,4]+(1-d1/894)/(894-d1)*uno2`*
  (xnomuest*bestim+cova*inv(v)*(datos[,4]-x*bestim)));
  ECM=uno2`*((xnomuest-omega1)*inv(omega2)*(xnomuest-omega1)`+(cova2-w))*uno2;
  Sem=datos[1,5];
```

```
  print y ECM sem;

*ALMACENA LOS RESULTADOOS EN UN DATASET EXTERNO;
  resul=shape(0,1,3);
  resul[1,1]= y;
  resul[1,2]= ECM;
  resul[1,3]= datos[1,5];
  create a.resultad from resul[colname={"yest" "ECM" "Semilla"}];
  append from resul;
  data aleat&i.;
set a.resultad;
  run;
  proc delete data=MAS&i.;run;
  proc delete data=no_MAS&i.;run;
  proc delete data=SAL&i.;run;

 dm 'clear log; clear output';
%end;
quit;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT1;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_1 separated by " "
from sashelp.vtable
where memname like '%ALEAT1%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 1;
data a.aleat_1;
set &lista_tablas_1.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT2;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_2 separated by " "
from sashelp.vtable
where memname like '%ALEAT2%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 2;
data a.aleat_2;
set &lista_tablas_2.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT3;
proc sql noprint ;
```

```
select compress(libname||"."||memname)
into: lista_tablas_3 separated by " "
from sashelp.vtable
where memname like '%ALEAT3%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 3;
data a.aleat_3;
set &lista_tablas_3.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT4;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_4 separated by " "
from sashelp.vtable
where memname like '%ALEAT4%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 4;
data a.aleat_4;
set &lista_tablas_4.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT5;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_5 separated by " "
from sashelp.vtable
where memname like '%ALEAT5%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 5;
data a.aleat_5;
set &lista_tablas_5.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT6;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_6 separated by " "
from sashelp.vtable
where memname like '%ALEAT6%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 6;
data a.aleat_6;
set &lista_tablas_6.;
```

```
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT7;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_7 separated by " "
from sashelp.vtable
where memname like '%ALEAT7%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 7;
data a.aleat_7;
set &lista_tablas_7.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT8;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_8 separated by " "
from sashelp.vtable
where memname like '%ALEAT8%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 8;
data a.aleat_8;
set &lista_tablas_8.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT9;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_9 separated by " "
from sashelp.vtable
where memname like '%ALEAT9%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 9;
data a.aleat_9;
set &lista_tablas_9.;
run;

*UNE TODOS LOS DATASET;
data a.aleat;
set a.aleat_1 a.aleat_2 a.aleat_3 a.aleat_4 a.aleat_5
a.aleat_6 a.aleat_7 a.aleat_8 a.aleat_9;
run;

*BUSCAR LAS OBSERVACIONES QUE ESTÁN REPETIDAS EN LA SEMILLA;
```

```
proc sql;
create table a.duplicados (where=(frec>1)) as select
semilla,
count(*) as frec
from a.aleat
group by 1;
quit;

/* CREA UN DATASET SIN LAS OBSERVACIONES DUPLICADAS EN EL NÚMERO DE SEMILLA*/
PROC SQL;
CREATE TABLE a.aleat_sin_dupli AS
SELECT DISTINCT a.*
FROM a.aleat a WHERE semilla NOT IN (SELECT semilla FROM a.duplicados);
QUIT;

/* CREA UN DATASET SIN LAS OBSERVACIONES QUE TENGAN A "ECM" e "Y" DISTINTOO DE CERO*/
DATA a.aleat_sin_dupli;
SET a.aleat_sin_dupli;
IF (yest=0) and (ECM=0) then delete;
run;

run;
%mend;
%Generar_yest_y_ECM;
```

Generador de 10000 muestras aleatorias simples suponiendo un modelo de semivariograma muestral y estimando los parámetros con los datos de la muestra.

```
libname a "D:\generadores finales";
%macro Generar_yest_y_ECM;
%do i=1 %to 10;

*Realiza una MSA de n=150 del conjunto de Datos;
proc surveyselect data=a.base
   method=srs n=150 out=MAS&i. outseed;
run;

*Crea el complemento de la muestra en el conjunto de datos;
PROC SQL;
CREATE TABLE no_MAS&i. AS
SELECT DISTINCT a.*
FROM a.base a WHERE i NOT IN (SELECT i FROM MAS&i.);
QUIT;

*Crear una salida a un archivo de los resultados del
PROC VARIOGRAM que se corre para la muestra;
ODS OUTPUT Variogram.SINBI.Model.Angle1.SelectedModel.ParameterEstimates=SAL&i.;
ODS OUTPUT Variogram.SINBI.Model.Angle1.SelectedModel.FitGenInfo=MOD&i.;
```

```
proc variogram data=MAS&i. outv=a.semivar_1;
store out=semi / label='VIR';
compute lagd=1300 maxlag=7 cl robust autocorr(assum=random);
coordinates xc=x yc=y;
model form=auto( mlist =(exp sph pow gau she ) nest=1 to 1);
var sinbi;
run;

ODS OUTPUT CLOSE;
ODS OUTPUT CLOSE;

*Colocar en un DataSet solo los parámetos nugget, Scale y Range;
data a.salida;
set SAL&i. (KEEP=Estimate);
if (Estimate=0) then Estimate=.;
run;

*colocar en un dataset el método o modelo seleccionado por SAS;
data a.salmod;
set MOD&i. (KEEP=Value FIRSTOBS = 2);
run;

proc iml;
*Definición la variable pi con el valor constante del número pi;
pi = constant("pi");

*Toma solo los parámetros nugget, Scale y Range;
use a.salida;
read all into par;
close a.salida;
print par;
nugget=par[1];
scale=par[2];
range=par[3];

*Toma el método o el modelo generado por SAS;
use a.salmod;
do data;
read next;
modelo= Value;
end;
close a.salmod;
print modelo;

print nugget scale range modelo;

*toma el valor de la covariancia;
use a.semivar_1;
```

```
read all into semivar;
close a.semivar_1;
covar=semivar[1,7];
print covar;

* Crear una tablas de n filas y dos columnas donde se almacenaran
el "y" estimado con ECM respectivos;
/* muestra i, creación de matriz de datos    */
use MAS&i.;
read all var{x y thogares sinbi InitialSeed} into datos;
/* creación de matriz de datos no incluidos en la muestra   */
use no_MAS&i.;
read all var{x y thogares} into datosnm;
d1=nrow(datos);
uno1=j(d1,1,1);
v=j(d1,d1,0);
/* diagonal de v   */
  do j2=1 to d1;
  v[j2,j2]=covar;
  end;
/*                */
/* cálculo de los elementos de fuera de la diagonal de v */
  do j3=1 to d1;
    do j4=j3+1 to d1;
/* cálculo de la distancia */
    h=sqrt((datos[j3,1]-datos[j4,1])**2+(datos[j3,2]-datos[j4,2])**2);
/* cálculo del semivariograma estimado */
if (modelo="Exp") then g=Nugget+Scale*(1-exp((-1)*h/Range));
/*modelo exponencial (exp)*/
else if (modelo="Sph") then g=Nugget+Scale*((3/2)*(h/Range)-(1/2)*(h/Range)**3);
/*modelo esferico (SPH)*/
 else if (modelo="Pow") then g=Nugget+Scale*h**Range;
 /*modelo potencia (pow)*/
  else if (modelo="Gau") then g=Nugget+Scale*(1-exp((-1)*(h/Range)**2));
  /*modelo gaussiano (gau)*/
   else  g=Nugget+Scale*(1-(sin(pi*h/Range)/(pi*h/Range)));
   /*efecto agujero (She)*/
v[j3,j4]=covar-g;
    v[j4,j3]=v[j3,j4];
    end;
  end;
/* fin de cálculo de lementos de v   */

  x=j(d1,1,1)||datos[,3];

/* estimación de los beta  */
  bestim= inv(x'*inv(v)*x)* x' * inv(v) * datos[,4];
```

```
/* cálculo de V_N-n,n que se llamará cova                */
  d2=nrow(datosnm);
  cova=j(d2,d1,0);
  do j21=1 to d2;
    do j22=1 to d1;
    h=sqrt((datosnm[j21,1]-datos[j22,1])**2+(datosnm[j21,2]-datos[j22,2])**2);

/* cálculo del semivariograma estimado */
if (modelo="Exp") then g=Nugget+Scale*(1-exp((-1)*h/Range));
/*modelo exponencial (exp)*/
else if (modelo="Sph") then g=Nugget+Scale*((3/2)*(h/Range)-(1/2)*(h/Range)**3);
/*modelo esferico (SPH)*/
 else if (modelo="Pow") then g=Nugget+Scale*h**Range;
 /*modelo potencia (pow)*/
  else if (modelo="Gau") then g=Nugget+Scale*(1-exp((-1)*(h/Range)**2));
  /*modelo gaussiano (gau)*/
   else  g=Nugget+Scale*(1-(sin(pi*h/Range)/(pi*h/Range)));
   /*efecto agujero (She)*/
    cova[j21,j22]=covar-g;
    end;
  end;

/* predicción del vector de unidades no incluidas en la muestra   */
  xnomuest=j(d2,1,1)||datosnm[,3];
  yipred=xnomuest*bestim;
  d3=nrow(yipred);
  ytotalest= sum(datos[,4])+sum(yipred+cova*inv(v)*(datos[,4]-x*bestim));

  /*calculo de ecm*/
  omega1=cova*inv(v)*x;
  omega2=x`*inv(v)*x;
  w=cova*inv(v)*cova`;
  uno2=j(d2,1,1);

  /*matriz V_N-n,N-n se llama cova2*/
  cova2=j(d2,d2,0);
  /* diagonal de V_N-n,N-n   */
  do j5=1 to d2;
  cova2[j5,j5]=covar;
  end;

/* cálculo de los elementos de fuera de la diagonal de v */
  do j6=1 to d2;
    do j7=j6+1 to d2;

/* cálculo de la distancia */
    h=sqrt((datosnm[j6,1]-datosnm[j7,1])**2+(datosnm[j6,2]-datosnm[j7,2])**2);
```

```
/* cálculo del semivariograma estimado */
if (modelo="Exp") then g=Nugget+Scale*(1-exp((-1)*h/Range));
/*modelo exponencial (exp)*/
else if (modelo="Sph") then g=Nugget+Scale*((3/2)*(h/Range)-(1/2)*(h/Range)**3);
/*modelo esferico (SPH)*/
 else if (modelo="Pow") then g=Nugget+Scale*h**Range;
 /*modelo potencia (pow)*/
  else if (modelo="Gau") then g=Nugget+Scale*(1-exp((-1)*(h/Range)**2));
  /*modelo gaussiano (gau)*/
   else  g=Nugget+Scale*(1-(sin(pi*h/Range)/(pi*h/Range)));
   /*efecto agujero (She)*/
    cova2[j6,j7]=covar-g;
    cova2[j7,j6]=cova2[j6,j7];
    end;
  end;
/* fin de cálculo de lementos de cova2  */

  y=894*((d1/894)/d1*uno1`*datos[,4]+(1-d1/894)/(894-d1)*uno2`*
  (xnomuest*bestim+cova*inv(v)*(datos[,4]-x*bestim)));

  ECM=uno2`*((xnomuest-omega1)*inv(omega2)*(xnomuest-omega1)`+(cova2-w))*uno2;
  Sem=datos[1,5];

  print y ECM sem modelo;

*ALMACENA LOS RESULTADOOS EN UN DATASET EXTERNO;
  if (modelo="Exp") then m=1;
  if (modelo="Sph") then m=2;
  if (modelo="Pow") then m=3;
  if (modelo="Gau") then m=4;
  if (modelo="SHE") then m=5;
  resul=shape(0,1,4);
  resul[1,1]= y;
  resul[1,2]= ECM;
  resul[1,3]= datos[1,5];
  resul[1,4]= m;
  create a.resultad from resul[colname={"yest" "ECM" "Semilla" "Modelo"}];
  append from resul;
  data aleat&i. (drop=modelo);
set a.resultad;
   if (modelo=1) then mod="Exp";
   if (modelo=2) then mod="Sph";
   if (modelo=3) then mod="Pow";
   if (modelo=4) then mod="Gau";
   if (modelo=5) then mod="SHE";
  run;

  proc delete data=MAS&i.;
```

```
  proc delete data=no_MAS&i.;
  proc delete data=SAL&i.;
  proc delete data=MOD&i.;
  run;

 dm 'clear log; clear output; clear result';
%end;
quit;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT1;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_1 separated by " "
from sashelp.vtable
where memname like '%ALEAT1%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 1;
data a.aleat_1;
set &lista_tablas_1.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT2;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_2 separated by " "
from sashelp.vtable
where memname like '%ALEAT2%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 2;
data a.aleat_2;
set &lista_tablas_2.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT3;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_3 separated by " "
from sashelp.vtable
where memname like '%ALEAT3%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 3;
data a.aleat_3;
set &lista_tablas_3.;
run;
```

```
*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT4;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_4 separated by " "
from sashelp.vtable
where memname like '%ALEAT4%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 4;
data a.aleat_4;
set &lista_tablas_4.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT5;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_5 separated by " "
from sashelp.vtable
where memname like '%ALEAT5%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 5;
data a.aleat_5;
set &lista_tablas_5.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT6;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_6 separated by " "
from sashelp.vtable
where memname like '%ALEAT6%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 6;
data a.aleat_6;
set &lista_tablas_6.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT7;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_7 separated by " "
from sashelp.vtable
where memname like '%ALEAT7%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 7;
```

```
data a.aleat_7;
set &lista_tablas_7.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT8;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_8 separated by " "
from sashelp.vtable
where memname like '%ALEAT8%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 8;
data a.aleat_8;
set &lista_tablas_8.;
run;

*SELECCIONA LOS DATASET QUE EMPIECEN POR ALEAT9;
proc sql noprint ;
select compress(libname||"."||memname)
into: lista_tablas_9 separated by " "
from sashelp.vtable
where memname like '%ALEAT9%' and libname = "WORK";
quit;

*UNE TODOS LOS DATASET QUE COMIENZAN CON 9;
data a.aleat_9;
set &lista_tablas_9.;
run;

*UNE TODOS LOS DATASET;
data a.aleat;
set a.aleat_1 a.aleat_2 a.aleat_3 a.aleat_4 a.aleat_5
a.aleat_6 a.aleat_7 a.aleat_8 a.aleat_9;
run;

*BUSCAR LAS OBSERVACIONES QUE ESTÁN REPETIDAS EN LA SEMILLA;
proc sql;
create table a.duplicados (where=(frec>1)) as select
semilla,
count(*) as frec
from a.aleat
group by 1;
quit;

/* CREA UN DATASET SIN LAS OBSERVACIONES DUPLICADAS EN EL NÚMERO DE SEMILLA*/
PROC SQL;
CREATE TABLE a.aleat_sin_dupli AS
```

```
SELECT DISTINCT a.*
FROM a.aleat a WHERE semilla NOT IN (SELECT semilla FROM a.duplicados);
QUIT;

/* CREA UN DATASET SIN LAS OBSERVACIONES QUE CERO EN "ECM" e "Y" */
DATA a.aleat_sin_dupli;
SET a.aleat_sin_dupli;
IF (yest=0) and (ECM=0) then delete;
run;

run;
%mend;
%Generar_yest_y_ECM;
```