UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Imputación de datos faltantes del Censo de Población y Vivienda utilizando técnicas de estadística espacial

María Eugenia Riaño

Febrero, 2018

Serie Documentos de Trabajo

DT (18 / 1) - ISSN : 1688-6453

Forma de citación sugerida para este documento:

Riaño, María Eugenia. Imputación de datos faltantes del Censo de Población y Vivienda utilizando técnicas de estadística espacial. [en línea]. 2018. Serie Documentos de Trabajo, DT (18/1). Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República, Uruguay.

Imputación de datos faltantes del Censo de Población y Vivienda utilizando técnicas de estadística espacial

María Eugenia Riaño [1]

Departamento de Métodos Cuantitativos, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República

RESUMEN

En general, la calidad y cobertura de los Censos de Población y Vivienda del año 2011 fue calificada como positiva, cumpliendo con los estándares exigidos internacionalmente. Sin embargo, su implementación no estuvo exenta de inconvenientes. No se cuenta con información de determinados hogares cuyo domicilio fue relevado, y para algunos se cuenta con sólo información parcial relativa a la composición del hogar. La omisión censal se concentra en zonas socioeconómicamente vulnerables. Esto afectaría la construcción del mecanismo utilizado por el Ministerio de Desarrollo Social para seleccionar a la población beneficiaria de los programas de transferencia monetaria. Este mecanismo se basa en la Encuesta Continua de Hogares cuyo marco muestral es el del Censo, y refleja los problemas de omisión. El trabajo se desarrolla para la ciudad de Montevideo. El patrón espacial de la población objetivo y de la propia omisión hace necesaria una regionalización previa a la imputación, dado que la distribución espacial se muestra heterogénea en el mapa. La selección de los modelos a utilizar para la imputación es muy sensible a la escala del mapa, por lo que la definición de las regiones condiciona la selección del modelo final a utilizarse para realizar la imputación. Las regiones se construyen mediante el algoritmo de árboles oblicuos de decisión, implementado en el paquete SpODT de R. Se ajustan modelos autorregresivos espaciales en cada región (SAR) que son evaluados con métodos de validación cruzada, y se comparan los resultados con el de un modelo global para todo el mapa. Los modelos con menor error de validación cruzada dentro de cada región muestran un rezago similar medido en distancia, a excepción de un caso. El modelo global presenta un error de validación cruzada similar, pero muestra autocorrelación espacial en los residuos, por lo que las imputaciones se realizan con los modelos locales por región.

Palabras clave: árboles de decisión, imputación de datos faltantes, modelos SAR autorregresivos, validación cruzada.

CÓDIGOS JEL: C21,C31,C38,C88.
Clasificación MSC2010: 62H11,62H25,62P12,62M40.

[1] *email:*eugenia@iesta.edu.uy, ORCID:0000-0003-3451-8249

Missing data imputation using Spatial Statistics techniques applied to the Uruguay's National Census

María Eugenia Riaño [1]

Departamento de Métodos Cuantitativos, Instituto de Estadística, Facultad de Ciencias Económicas y de Administración, Universidad de la República

ABSTRACT

National Census's quality and coverage was positively evaluated in general, attaining international standard requirements. However, the data collecting process had some difficulties. For a number of households, there is no information about residents. In other cases, a brief version of the questionnaire was applied, remaining incomplete information. Also, the omission is focused in segments socioeconomically vulnerable. This could have an impact over the algorithm performed by the government to select the beneficiary population of some cash-transfer programs. The algorithm is based on the Continuous Household Survey data, being the Census the sample frame. The analysis is developed for the city of Montevideo. The heterogeneous spatial pattern of the target population and of the omission itself makes it necessary define regions for the imputation of the missing data. The model selection is sensitive to the choice of windowing, consequently, the region's definition conditions the model selection for data imputation. Regions are obtained by means of spatial oblique decision trees, performed with SpODT package. Spatial Autorregresive models (SAR) are adjusted for each region. The models are assessed using cross - validation methods. Results are compared with the performance of a global model for the whole map. Except by one region, models that minimize cross - validation's errors show a similar lag in each region. The cross - validation error for the global model is quite similar. Nevertheless, spatial autocorrelation is detected according to the Moran test for residuals. Hence, the data imputation is performed by regions, with local SAR models, selecting the lag according to the cross - validation error.

Key words: classification and regression trees, cross - validation, missing data imputation, SAR models.

Códigos JEL: C21,C31,C38,C88.
Clasificación MSC2010: 62H11,62H25,62P12,62M40 .

[1] *email:* eugenia@iesta.edu.uy, ORCID:0000-0003-3451-8249

1. Introducción

En general, la calidad y cobertura de los Censos de Población y Vivienda del año 2011 fue calificada como positiva, cumpliendo con los estándares exigidos internacionalmente. Sin embargo, su implementación no estuvo exenta de inconvenientes. No se cuenta con información de determinados hogares cuyo domicilio fue relevado, y para algunos se cuenta con sólo información parcial relativa a la composición del hogar.

La omisión censal se concentra en zonas socioeconómicamente vulnerables. Esto afectaría la construcción del mecanismo utilizado por el Ministerio de Desarrollo Social (de aquí en más MIDES) para seleccionar a la población beneficiaria de los programas de transferencia monetaria. Este mecanismo se basa en la Encuesta Continua de Hogares cuyo marco muestral es el del Censo, y refleja los problemas de omisión.

En el presente documento se presentan los resultados obtenidos para la imputación de población elegible del programa Tarjeta Uruguay Social en las zonas omisas del Censo. El marco geográfico es el del departamento de Montevideo. Se observa que el patrón espacial de la omisión es diferente por zonas, por lo que se hace necesario definir regiones de imputación.

Para la construcción de las regiones se utiliza el algoritmo SpODT (Spatial Oblique Decision Tree). El método se encuentra basado en los árboles de clasificación y regresión (CART). Las regiones de clasificación en los CART son rectangulares, siendo una limitante para captar patrones en el espacio. Los árboles oblicuos definen los límites entre regiones en base a rectas, lo que permite una mayor flexibilidad a la hora de definir la forma de la región.

En las siguientes secciones se presenta la descripción del problema, el marco teórico y la metodología a utilizar. Se presentan los resultados obtenidos y finalmente se realiza una discusión en donde se plantean las direcciones futuras de trabajo dentro del marco de esta investigación.

2. Descripción del problema

En el año 2008, en el marco de un convenio entre el MIDES y la Universidad de la República, se construye un índice denominado Índice de Carencias Críticas (de aquí en más ICC), con el fin de definir inicialmente la población elegible del programa de Asignaciones Familiares del Plan de Equidad (AFAM -PE). Luego también es utilizado para definir la población elegible de los programas de transferencias Tarjeta Uruguay Social

(TUS) (Ministerio de Desarrollo Social (2013)).

El ICC es un modelo probit que predice la probabilidad de que un hogar pertenezca al primer quintil de ingresos, en base a variables que reflejan su situación en términos de educación, vivienda, confort y composición del hogar. Junto con la estimación del ICC, se determina también un punto de corte por región (Montevideo - Interior) que permite distinguir entre los hogares que según la predicción del modelo pertenecen a la población elegible, y los que no. Es decir, se fija un valor del ICC a partir del cual un hogar sería población elegible para el programa por región.

Dichos umbrales se fijan de modo de permitir la elegibilidad de la cantidad de hogares que la Ley define deben recibir el beneficio, tomando como referencia la Encuesta Continua de Hogares (ECH) que se considera representativa del total poblacional.

Los hogares de la ECH cuentan con sus correspondientes pesos muestrales anuales, que permiten realizar las estimaciones pertinentes al caso. Para fijar los umbrales, se estima el ICC para los hogares de la ECH y se calcula la probabilidad de pertenecer al primer quintil de ingresos. Luego se ordenan en forma decreciente según la probabilidad calculada por el modelo, y dada la cantidad de hogares que pretende alcanzar el programa se estima la distribución de los mismos entre Montevideo - Interior. Se realiza una suma acumulada de los ponderadores muestrales que permite identificar cuáles son los hogares de la muestra que representarían a los beneficiarios del programa y luego tomando esos hogares se estima la proporción Montevideo - Interior.

Una vez determinada la cantidad de posibles beneficiarios del programa por región, el umbral del ICC se fija en el lugar donde (ordenado en forma decreciente) la estimación de la población beneficiaria se iguale a la cantidad de beneficiarios que pretende alcanzar el programa.

El procedimiento descrito en los párrafos anteriores corresponde al programa AFAM. Para los programas TUS se modifica levemente, tomando como criterio para ordenar los hogares la capacidad de compra de la canasta básica alimentaria (CBA) en lugar de los ingresos. Esto tiene como consecuencia que la cantidad de partidas de beneficios aumente para Montevideo, ya que la CBA es más cara.

Dado que el marco de la ECH es el del Censo 2011, cuya omisión se concentra en zonas socioeconómicamente vulnerables, se cuestiona cual sería el efecto de la omisión en los umbrales del ICC. Es decir, si los hogares que no se incluyen en la ECH por no estar en el marco son más vulnerables que los demás, los umbrales se verían modificados al alza. La omisión del censo tendría como consecuencia que la distribución de los programas entre

los hogares beneficiarios no fuera la correcta, así como también podría verse modificada la distribución de las partidas de beneficios entre Montevideo e Interior.

Surge así la necesidad de realizar una estimación de la cantidad de población elegible en las zonas omisas del Censo, con el fin de estimar el impacto en los umbrales principalmente de los programas TUS y en la distribución por región de las partidas de beneficios.

2.1. Objetivos

Objetivo general: Obtener una estimación de la cantidad de población elegible para el programa Tarjeta Uruguay Social en las zonas omisas del Censo.

Objetivos Específicos:

- Definir regiones de imputación.
- Comparar el desempeño de un modelo espacial global con el de modelos locales para las regiones de imputación.

3. Marco teórico

3.1. Introducción

Dentro de la Estadística espacial según Gaetan y Guyon (Gaetan y Guyon, 2010), existen tres grandes áreas de estudio:

- Análisis de Patrones de Puntos
- Geoestadística, y
- Datos de Área.

En Análisis de Patrones de Puntos (*Spatial Point Pattern*) el interés se centra en el lugar en donde ocurrirán los eventos. Se utiliza por ejemplo en epidemiología, y se concentra en la modelización de procesos estocásticos en tiempo y espacio.

En el segundo caso, los datos pueden medirse en un principio en cualquier punto del espacio (datos continuos). El interés no es el patrón de puntos observados en sí mismo, si no en la predicción sobre un espacio continuo de una variable de interés medida en los sitios observados.

Cuando los datos espaciales son observados en polígonos, se los definen como Datos de Área. En la mayoría de los casos los polígonos corresponden a unidades administrativas, como ser una zona censal, un departamento, país, etc. Los datos observados son frecuentemente agregados dentro de los límites del polígono, como por ejemplo totales o promedios dentro de las áreas administrativas. Los datos de área son datos discretos.

En el caso de Geoestadística, las distancias sobre una superficie continua son la base para determinar la estructura de la autocorrelación espacial. En el caso de Datos de Área no existe una noción de distancia euclidea, si no que se deben definir estructuras de vecindad - cercanía entre los polígonos. En una superficie continua, todos los puntos son vecinos uno a uno, aunque a algunos puede dársele un peso menor, por encontrarse muy apartados. En una superficie de polígonos, se define una partición del conjunto (excluyendo una observación i) separando los polígonos que son vecinos de la observación i del resto de las observaciones. Así se conforma un grafo dirigido que describe la dependencia espacial entre los datos. A cada polígono en la vecindad a su vez se le pueden asignar diferentes pesos, que reflejen la intensidad de la dependencia espacial. Esta diferencia entre Geoestadística y Datos de Área es fundamental, ya que en el segundo caso es el investigador que introduce la estructura de vecinos.

El presente trabajo corresponde a Datos de Área y las unidades de observación son las zonas censales del departamento de Montevideo. A continuación se presentan una serie de conceptos básicos para el análisis de este tipo de datos, comenzando por la definición de vecindad espacial.

3.2. Matriz de conectividad: definición de Vecindad - Cercanía Espacial

El primer paso es definir el criterio para elegir los vecinos. El segundo es asignar un peso a los vínculos entre observaciones definidos en el paso anterior.

Existen varias formas de definir los vecinos:

- **Por contigüidad**: se definen como vecinos a aquellos polígonos que compartan bordes o vértices con la observación. También se puede elegir el órden, en términos de series temporales, rezagos.
- Vecinos basados en **grafos**: triangulación de Delaunay, Esferas de Influencia, Gabriel graph, entre otros.
- Vecinos basados en **distancias**: vecinos más cercanos, radios de distancia.

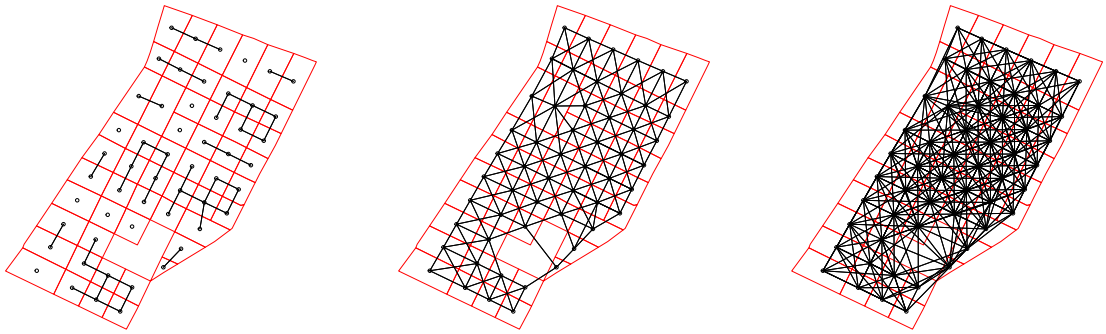
Los métodos mencionados anteriormente pueden ser implementados tanto en datos con polígonos irregulares (el caso de estudio por ejemplo) como regulares (una grilla regular como pueden ser los píxeles de una imagen). Existen otros métodos exclusivos para grillas regulares que no se consideran en este caso por no ser aplicables a los datos a analizar.

3.3. Definición de la matriz de Pesos Espaciales

Una vez establecida la matriz de conectividad que define a los vecinos se debe elegir la forma de ponderar las relaciones establecidas en la matriz de conectividad. A esta matriz de pesos se la denota como W y es la base del cálculo de la autocorrelación y de la estimación de modelos espaciales autorregresivos.

Existen varias formas de definir la matriz W. La primera forma es definirla como binaria, indicando con unos y ceros cuáles son vecinos y cuáles no. Otra forma es mantener la distancia euclidea entre los vecinos, asignando un cero en el resto de los vínculos. No se realizará una descripción exhaustiva de la forma de asignar los pesos, pero existe una variedad de métodos basados en distancia, que penalizan de distinta forma los vecinos más alejados.

Se dice que la matriz W está estandarizada cuando la suma de sus filas es uno. Cuando esto se cumple la interpretación de los parámetros de los modelos es más sencilla.



(a) 100 metros (b) 200 metros (c) 300 metros

Figura 1: Conectividad para 100, 200 y 300 metros

4. Autocorrelación espacial: Índice de Moran

Aunque no es el único, el Índice de Moran es el estadístico más utilizado para testear la autocorrelación espacial. Se define como:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

Donde N es el número de observaciones, X es la variable de interés y w_{ij} son los pesos de la matriz W. El valor esperado bajo la hipótesis nula de no autocorrelación es

$$E(I) = \frac{-1}{N-1}$$

Los valores de I se encuentran entre -1 y 1. Cuando se acerca a cero indica un patrón espacial aleatorio (no hay autocorrelación). Existen dos formas de interpretar la hipótesis nula de no existencia de autocorrelación espacial:

- Las observaciones son mutuamente independientes, o
- Cada permutación de las observaciones x_i es igualmente probable.

Se puede demostrar que la distribución asintótica del Índice de Moran escalado por una constante es $N(0,1)$ (Gaetan y Guyon, 2010).

4.1. Modelos Espaciales Simultáneos Autorregresivos (SAR)

Los modelos **SAR** son la extensión espacial de un modelo de regresión lineal, incluyendo un término correspondiente a la estructura espacial autorregresiva. Son una generalización de los procesos autorregresivos de series de tiempo.

El modelo espacial autorregresivo de primer orden es

$$y_i = \rho w_{i,n} Y + \varepsilon_i \qquad i = 1, \ldots, n$$

donde $Y = (y_1, \ldots, y_n)'$ es un vector columna de variables dependientes, $w_{i,n}$ es un vector fila n-dimensional de constantes y ε_i son i.i.d y $N\left(0, \sigma^2\right)$. En forma matricial

$$Y = \rho W Y + \varepsilon$$

El término WY se denomina "rezago espacial". Bajo el supuesto de que $(I_n - \rho W)$ es no singular se tiene que

$$y = (I_n - \rho W)^{-1} \varepsilon = S(\rho)^{-1} \varepsilon$$

Al modelo presentado anterioremente se le denomina espacial autorregresivo puro, ya que sólo considera como variable explicativa a la dependiente rezagada. Los modelos de este tipo son estimados por Máxima Verosimilitud.

Asumiendo que $\varepsilon \sim N(0, \sigma^2 I_n)$ el logaritmo de la verosimilitud es

$$lnL(\rho, \sigma^2) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) + ln|S(\rho)| - \frac{1}{2\sigma^2} Y' S'(\rho) S(\rho) Y$$

con $S(\rho) = I - \rho W$. La función de verosimilitud implica el cálculo del determinante de $S(\rho)$, que es función del parámetro desconocido ρ y dependiendo de n, de alta dimensión. Existen varios métodos numéricos para obtener una estimación de ρ. Entre ellos se encuentran los basados en la normalización de la matriz W por fila, utilizando la propiedad de que los valores propios de la matriz W son iguales a los de la matriz $D^{-1/2} W^* D^{-1/2}$, donde W^* es una matriz simétrica y $D = diag\left\{\sum_{j=1}^{n} w_{i,j}^*\right\}^{-1}$, y los basados en el polinomio característico de W.

4.2. Spatial Oblique Decision Tree (SpODT)

SpODT (Gaudart *et al.*, 2015) es un método no paramétrico para encontrar clusters espaciales basado en los árboles de clasificación y regresión (CART).

En CART, para cada variable se busca un umbral que particiona al espacio de la variable en dos clases, optimizando algún criterio definido a priori. Se realizan particiones binarias recursivas, hasta alcanzar una determinada regla de parada. Aplicado a datos espaciales, el CART busca entre las coordenadas planares $\{x_i, y_i\}$ (de cada ubicación M_i) un umbral o límite entre clases espaciales tal que la media de la variable de interés sea lo más diferente posible entre las dos clases. El algoritmo CART lleva a obtener clases rectangulares, proporcionando particiones perpendiculares de las longitudes y latitudes proyectadas. El algoritmo SpODT es una variante de CART que proporciona una partición oblicua al área de estudio, que es más apropiada a la forma de los datos espaciales. La forma funcional puede ser escrita como:

$$z_i = f(x_i, y_i) + \varepsilon_i$$

donde $\{x_i, y_i\}$ son las coordenadas planares para cada punto $M_i, i = 1, \ldots, N$ y $\varepsilon_i \in \mathcal{R}$. La función $f(x_i, y_i)$ se define como

$$f(x_i, y_i) = \sum_{j=1}^{P} \bar{z}_j \mathcal{I}_{\{M_i(x_i,y_i) \in class_j\}}$$

donde $class_j$ para $j = 1, \ldots, P$ son las P clases finales después de particionar el área de estudio, $\bar{z}_j = \frac{1}{N_j} \sum_{M_i \in class_j} z_i$ es el promedio de los valores observados de los N_j puntos $M_i \in class_j$.

El problema principal es determinar el conjunto de clases $\{class_j, j = 1, \ldots, P\}$. Los límites entre las regiones están definidos por rectas $s_j(x_i, y_i) = ax_i + by_i + c = 0$. Estos límites, o direcciones de partición, se determinan recursivamente por cada punto de la muestra, denotado como el nodo ξ, correspondiente al área de estudio al comienzo del algoritmo, o a una zona (clase geográfica) obtenida como resultado de una partición previa. El nodo ξ es particionado en dos clases, por la dirección $s_j(x_i, y_i)$. Si $s_j(x_i, y_i) \leq 0$, entonces el punto M_i pertenecerá al nodo "hijo" (jl) de la izquierda del árbol. Si no, el punto M_i pertenecerá al nodo "hijo" de la derecha (jr). Para cada nodo ξ constituido por el conjunto de $n(\xi)$ puntos, el algoritmo busca, entre el conjunto S de todas las funciones lineales de (x_i, y_i) la función $s_j(x_i, y_j)$ tal que:

$$SSE_{inter}(s_j, \xi) = max_{s \in S} \{SSE_{inter}\}$$

Gaudart et al (Gaudart *et al.*, 2005) demuestran que el conjunto S de funciones que particionan el área de estudio es de cardinal finito. Existe un número infinito de líneas que particionan el conjunto de puntos en dos subclases, sin embargo, muchas de ellas llevan a la misma clasificación, particionando el conjunto de puntos en forma idéntica. El algoritmo debe identificar las posibles líneas a analizar como candidatas a particionar el conjunto de puntos. Para este propósito el algoritmo usa propiedades relacionadas al orden de las abcisas de los puntos a ser particionados, después de una rotación del eje de las x. Luego, el algoritmo realiza una partición vertical de las imágenes de las x para cada rotación. Al igual que en CART, las particiones se hacen recursivamente hasta alcanzar una determinada regla de parada. Para más detalle ver (Gaudart *et al.*, 2005).

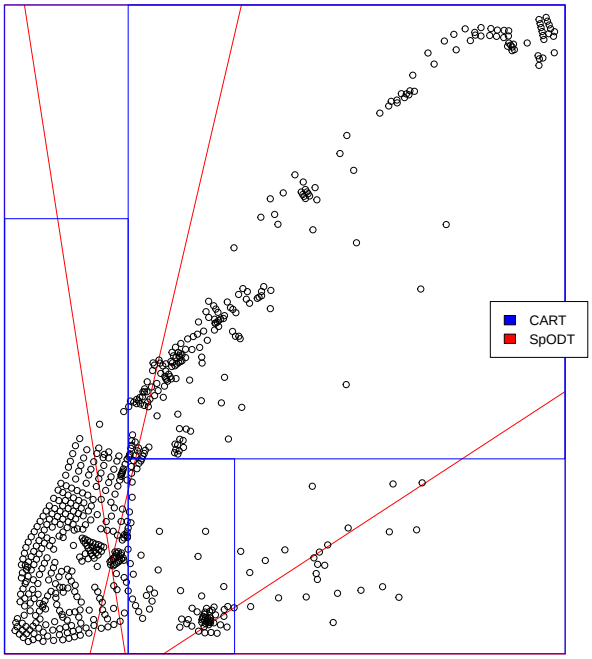


Figura 2: Comparación particiones CART y SpODT

5. Metodología

La metodología se desarrolla en las siguientes etapas:

1. Definición de las regiones de imputación.

   Utilizando el algoritmo SPOdT se construye una regionalización de la superficie, de acuerdo a la variable proporción de hogares elegibles del programa TUS en las zonas censales.

2. Ajuste de modelos SAR con diferentes rezagos para las distintas regiones y selección del modelo con menor error de validación cruzada.

   Para las regiones encontradas se obtienen los correlogramas por contigüidad y por clases de distancia basados en el Índice de Moran. Según la significación de los rezagos se elige un umbral de distancia para cada región y se ajustan modelos SAR cada 100 metros hasta alcanzar el umbral. Para cada modelo se estima el error de

validación cruzada, y se selecciona el modelo correspondiente al rezago de menor error.

3. Ajuste del modelo SAR para el mapa global. Se selecciona el rezago de distancia de menor error de validación cruzada.

   Se realiza el mismo procedimiento implementado en cada región para el mapa global, y se elige el rezago correspondiente al modelo de menor error de validación cruzada.

4. Comparación de resultados entre los modelos SAR locales y el global.

   Se comparan los errores de validación cruzada de los modelos locales con el modelo global y se realizan pruebas de diagnóstico.

5. Imputación de la cantidad de hogares elegibles para el programa TUS.

   Con las predicciones del modelo seleccionado en el paso anterior se obtiene una estimación de la población elegible para el programa TUS en las zonas omisas en Montevideo.

6. Descripción de la base de datos

Las bases cartográficas con las que se trabajan corresponden a las actualizadas por el Censo 2011. Las zonas censales son unidades de muestreo en la selección de la muestra de la ECH. En el país existen un total de 69.752 zonas de las cuales 16.486 corresponden a "zonas verdes". Ejemplos de zonas verdes son los parques, canteros, reservas naturales, etc. Estas zonas no se censan y no se toman en cuenta para la imputación.

Departamento	**Total de zonas**	**Total sin zonas verdes**	**Morador Ausente**	**Formato Papel**	**Zonas Mixtas**	**Omisas**
Montevideo	13621	10559	20	182	290	140

Tabla 1: Clasificación de Zonas sin información en Montevideo

Existen varios tipos de omisión. El primero es en donde el empadronador releva el domicilio, pero el censista no. Se tiene un registro de la existencia de la vivienda, pero no se tiene información de quienes residen en ella. A este caso se lo denomina Morador Ausente. Al estar registrada en el marco censal, la vivienda tiene posibilidades de ser seleccionada en la ECH. Si bien no se tienen datos al momento del censo sobre la cantidad de población

elegible residente en la vivienda, al no quedar excluida del sorteo de la ECH, no genera un problema de marco.

El segundo es consecuencia de una operativa de emergencia al final del relevamiento del Censo. Se decidió relevar solamente los datos referentes a la composición del hogar, es decir, sexo, edad y relación de parentesco, en un cuestionario papel. Para estos hogares no es posible calcular el ICC, pero sí se tiene algún tipo de información que puede ser de importancia para determinar si el hogar es población elegible o no, como ser la cantidad de menores de 15 años. A estos casos se los denomina Formato Papel. Como en el caso anterior, al no excluirse la vivienda del marco, no tiene consecuencias graves sobre la ECH, sí sobre la estimación de población elegible al momento del censo.

El tercer caso es en el que en donde la vivienda no fue relevada ni por el empadronador, ni por el censista. Para poder corroborar esta situación es necesaria una fuente externa de información, ya que en el marco del INE no hay registros de estos domicilios. El MIDES realiza un relevamiento permanente en zonas socioeconómicamente vulnerables que permitió identificar algunos de estos casos en zonas censales que se encontraban vacías según el censo. El trabajo que el MIDES realiza en campo es dirigido, y no es completo, es decir que no es posible identificar a todas las zonas que fueron omitidas completamente por el censo. Sí se obtiene una cota inferior para esta cantidad, no siendo posible identificar el resto por no contar con fuentes externas oficiales que lo permitan.

El tercer caso sí tiene consecuencias importantes sobre las estimaciones de la ECH ya que estas zonas no son incluidas en el marco. El error que implique esta omisión se arrastra a todas las ECH futuras hasta una nueva actualización del marco. Por este motivo es que este trabajo se enfoca en la imputación de los datos en estas zonas, en donde actualmente no se tiene ningún tipo de información.

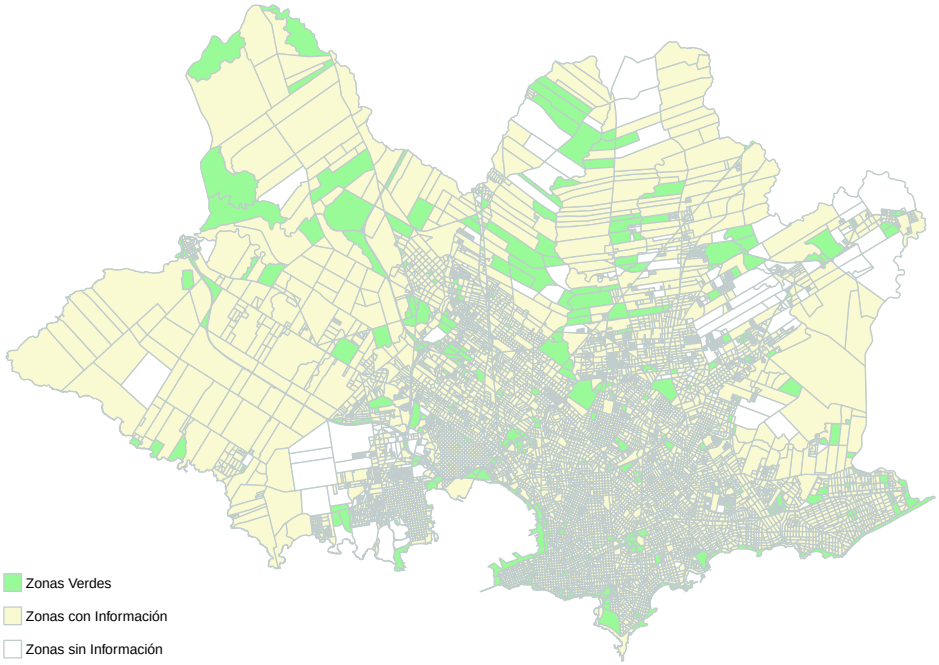


Figura 3: Distribución geográfica de las zonas sin información

La distribución geográfica de las zonas sin información se presenta en la Figura 3. Los problemas de omisión claramente se concentraron en zonas socioeconómicamente más vulnerables. Los comunales 9 y 17 concentran el 77.2 % de las zonas sin información. También con una participación menor se encuentran los comunales 10, 11 y 12 que concentran un 15 % de los casos.

Para la población elegible del Programa Tarjeta Uruguay Social, la distribución geográfica se presenta en la Figura 4. Si se compara con la distribución de las zonas sin información, se observa que estas zonas se encuentran cercanas a las zonas con mayor porcentaje de posibles beneficiarios de este programa. Al igual que en el caso anterior el patrón no es homogéneo, habiendo mayor concentración en algunos sectores del mapa.

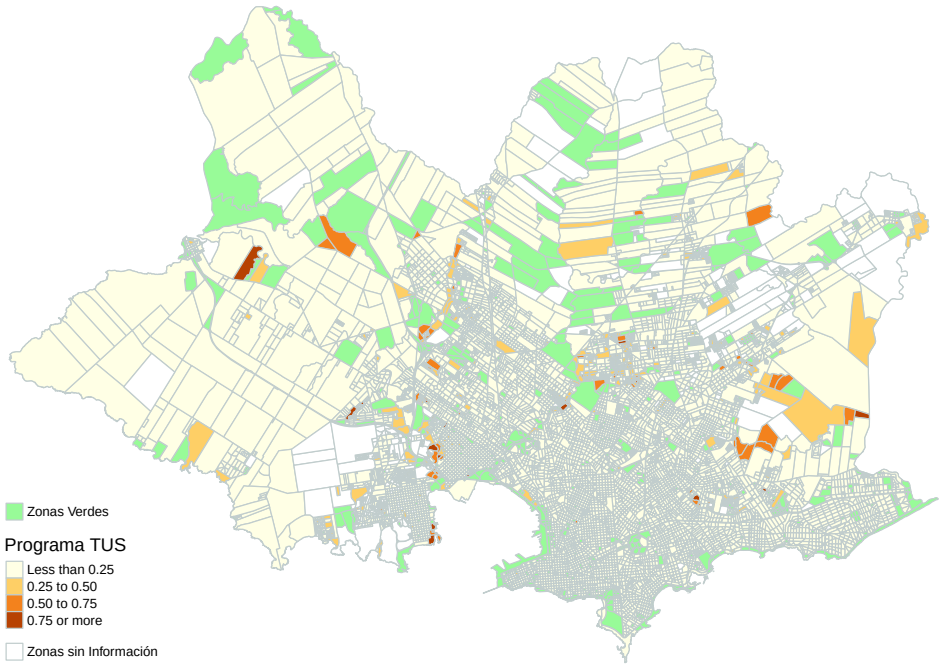


Figura 4: Distribución geográfica de la población elegible del programa TUS

7. Resultados

7.1. Definición de las regiones de imputación (SpODT)

La implementación se realiza con el paquete SpODT de R. Se imputarán aquellas zonas omisas de las regiones encontradas con un porcentaje de población elegible mayor a un 5 % (550 zonas).

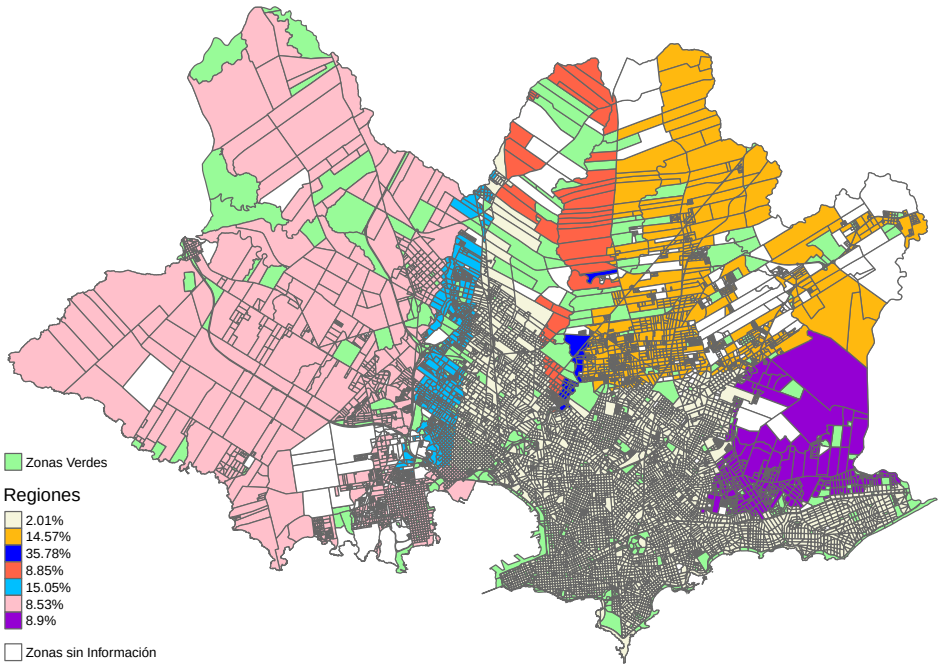


Figura 5: Regiones de imputación definidas por el algoritmo SpODT

El mapa se divide en 7 regiones en total. Del total de regiones sólo una tiene un porcentaje de población elegible menor al 5 % y corresponde a la zona del centro del mapa más la costa este del departamento. La región con más alta concentración es la coloreada en azul en el mapa. Es una región pequeña y presenta la particularidad de ser la única con una discontinuidad en el espacio. Se encuentra rodeada de zonas verdes y omisas.

La región del oeste de Montevideo es la más identificable en la partición obtenida, separada por una región "franja" con mayor porcentaje de población elegible. La zona noroeste del departamento es la que presenta zonas omisas de gran superficie, y dentro de ella se encuentran dos regiones. Por último, la región delimitada en la zona de Bañados de Carrasco incluye parte de Malvín Norte y de Carrasco Norte.

7.2. Ajuste y elección de modelos SAR para las regiones seleccionadas

En cada región se realiza un correlograma por contigüidad y por clase de distancia basados en el I de Moran, los resultados se presentan en el Anexo 1.

Basados en la significación de los índice de Moran para los rezagos de los correlogramas se elige una distancia máxima para la prueba de los modelos. Luego se ajusta un modelo SAR para los rezagos múltiplos de 100 metros hasta el rezago máximo encontrado, y se realiza una validación cruzada basada en K - *folds*, con $K = 20$ para cada modelo ajustado. Se selecciona el modelo con menor error de validación cruzada.

A continuación se presentan la estimación de los parámetros de los modelos con menor error de validación cruzada para las seis regiones encontradas, el rezago asociado, el p valor de la prueba de razón de verosimilitudes y el p valor asociado a la autocorrelación de los residuos realizada con el Índice de Moran.

	Error CV	**Rezago**	ρ	**LR test p- valor**	**Moran test p -valor**
Región 1	0.01825	200 m	0.5548	<2.22e-16	0.2747
Región 2	0.04387	300 m	0.61843	1.0078e-05	0.5715
Región 3	0.01408	200 m	0.46084	0.0002483	0.2719
Región 4	0.02502	200 m	0.60832	<2.22e-16	0.4663
Región 5	0.01538	300 m	0.49689	<2.22e-16	0.2646
Región 6	0.01884	600 m	0.66612	<2.22e-16	0.9257

Tabla 2: Resultados del ajuste de los modelos SAR locales con menor error de validación cruzada

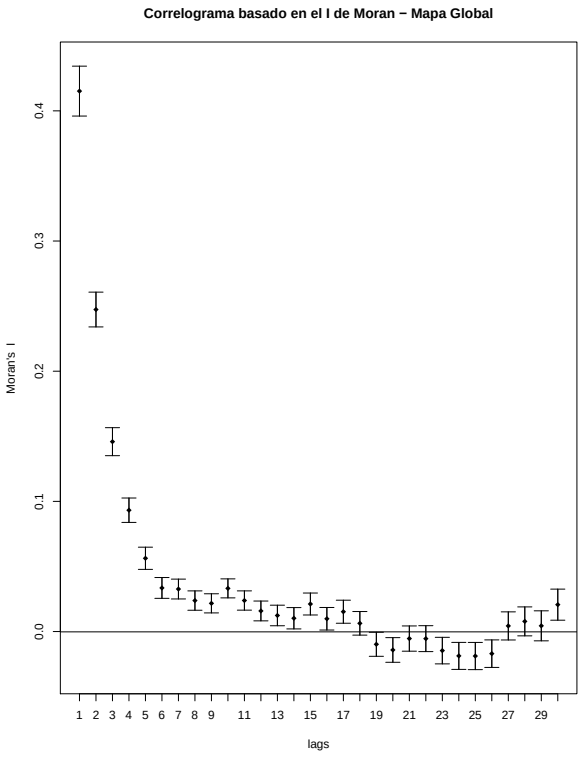
A excepción de la región 6, en donde el modelo con menor error de validación cruzada corresponde a un rezago de 600 metros, los rezagos encontrados varían entre 200 y 300 metros. El parámentro ρ es significativo en todos los casos, y los residuos no presentan autocorrelación espacial según el Índice de Moran. En las regiones 3 y 5 la "intensidad" de

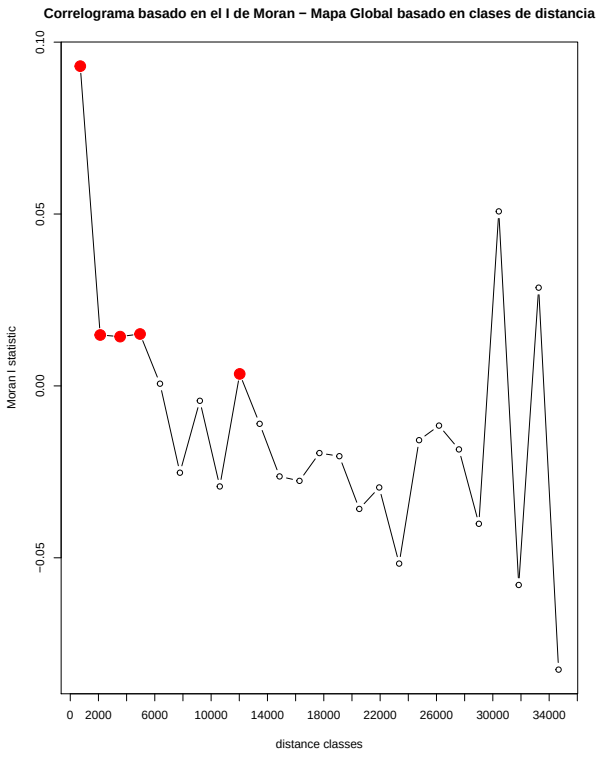
la dependencia espacial es menor a la de las otras regiones. El error global de validación cruzada se calcula como:

$$ErrorCV = \sum_{i=1}^{6} \frac{N_i}{N} ErrorCV_i = 0{,}01913$$

7.3. Modelo Global

A continuación se presentan los autocorrelogramas basados en contigüidad y en clases de distancias para el mapa de Montevideo.





Como se puede observar en el autocorrelograma basado en clases de distancias, aparece un rezago significativo hasta una distancia de aproximadamente 5000 metros. Luego en el rezago de 12000 metros vuelve a ser significativo. Según el basado en contigüidad, los rezagos son significativos hasta el número 17. Como son polígonos irregulares, no puede afirmarse que cada rezago basado en contigüidad tenga la misma distancia, por lo que se decide explorar los modelos hasta un rezago de 5000 metros. No se consideran los rezagos que vuelven a ser significativos a distancias mayores en los correlogramas, ya

que se presume que lo que capta es una discontinuidad en el mapa del fenómeno a estudiar.

Al igual que para los modelos locales, se prueban modelos SAR en rezagos múltiplos de 100, hasta el umbral de 5000 metros. Los resultados del modelo con menor error de validación cruzada se presentan a continuación:

Rezago	**Error CV**	ρ	**LR test p- valor**	**Moran test p -valor**
200 m	0.02000	0.57232	<2.22e-16	0.03747

Tabla 3: Resultados del ajuste del modelo global con menor error de validación cruzada

El error de validación cruzada para el modelo SAR global es apenas mayor que para el obtenido con los modelos locales. Sin embargo los errores del modelo global muestran correlación espacial según el Índice de Moran. Esto puede deberse a las características de la región 6, en donde el rezago de menor error de validación cruzada es de 600 metros, disímil con el del resto de las regiones. Esto puede hacer que el modelo global, con un rezago de 200 metros para toda la superficie, presente correlación espacial en los residuos. Se decide entonces imputar la población elegible en las zonas omisas con los modelos SAR locales obtenidos a partir de la regionalización.

7.4. Imputación

El total de hogares población elegible estimado en las zonas omisas es de 1058 hogares (5 % del total en zonas no omisas).

	0 %	**20 %**	**40 %**	**60 %**	**80 %**	**100 %**
p	0.0085	0.0687	0.0992	0.1261	0.1592	0.6242

Tabla 4: Quintiles de la proporción de hogares imputada

De los valores de los quintiles para las proporciones imputadas se puede decir que hasta un 80 % son proporciones bajas, de menos de un 16 % aroximadamente. Sólo un 20 % supera estos valores y hasta un 62 %. A continuación se presenta el mapa luego de realizar la imputación de población elegible para el programa Tarjeta Uruguay Social.

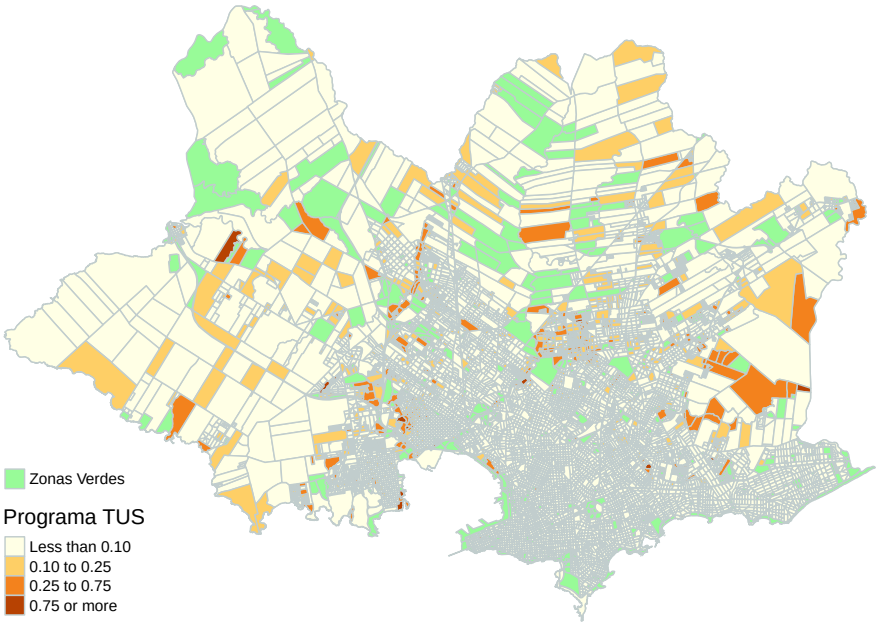


Figura 6: Distribución espacial de la población elegible luego de realizada la imputación

Acorde con los valores imputados, las zonas en blanco de la Figura 3 se completan con los colores más claros en la Figura 6.

8. Conclusiones y Líneas futuras de investigación

Luego de aplicar el algoritmo SpODT tomando como variable de respuesta a la proporción de hogares población elegible del Programa Uruguay Social, se obtiene una partición con siete regiones, de las cuales se excluye una del análisis por la baja proporción de

hogares población elegible. Para el resto de las zonas, se ajustan modelos SAR variando el rezago de a 100 metros, hasta un rezago límite que se define a partit de la observación de los correlogramas basados en el Índice de Moran (basados en clases de distancia y por contigüidad). Salvo en una región, los rezagos que presentan menor error de validación cruzada corresponden a los 200 o 300 metros.

Por otro lado, se estima un modelo global, considerando la totalidad del mapa. Con el mismo procedimiento utilizado en las regiones, se encuentra que el mejor modelo según el criterio de validación cruzada es el correspondiente a un rezago de 200 metros, presentando un error de validación cruzada apenas mayor que en los obtenidos con los modelos locales. Sin embargo, el modelo global presenta autocorrelación espacial en los residuos, según el Índice de Moran, probablemente debido a la región 6 que se muestra heterogénea respecto a las demás, con un rezago de 600 metros.

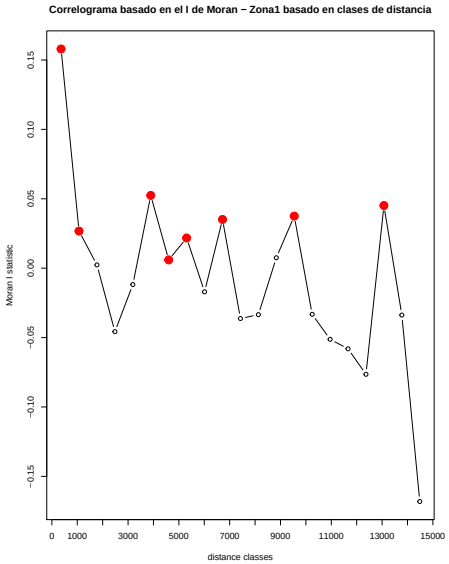
Por último, se obtiene una estimación del total de hogares población elegible en las zonas omisas, representando un 5 % de la población elegible estimada actual.

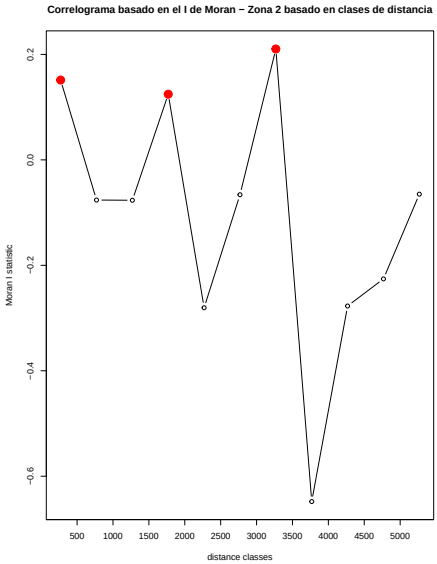
Como líneas futuras de investigación se detallan los siguientes problemas a analizar:

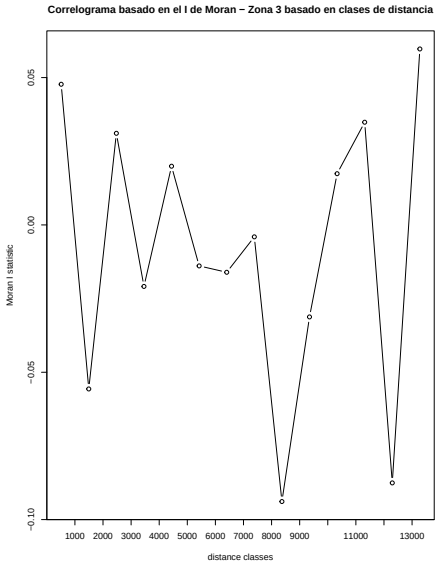
- Estimar la varianza de la predicción obtenida con los modelos locales: para este trabajo se obtuvo una estimación puntual, y en el caso de que el modelo global hubiese resultado más adecuado que los locales sería más sencillo. Se presenta el problema de realizar una estimación de varianza con los modelos locales agregados, que debería enfocarse en un principio con métodos de remuestreo.
- Probar el método con una diferencia más acentuada de los rezagos entre las regiones. El mapa de Montevideo en las seis regiones encontradas es homogéneo en cuanto a los rezagos que minimizan el error de validación cruzada, a excepción de un caso. Surge entonces la pregunta de cuáles serían los resultados si la variable de interés fuera totalmente heterogénea entre regiones. Mediante simulación podrían evaluarse distintos escenarios con diferentes grados de heterogeneidad entre regiones, de forma que permita arribar a una conclusión más general en cuanto a la metodología utilizada para la selección de los modelos locales y el global.
- Incorporar variables explicativas al modelo. En este trabajo no se incorporaron variables como la cantidad de menores por hogar, que pueden tener un poder explicativo fuerte cuando se quiere determinar si un hogar es población elegible. Tampoco se consideraron los datos faltantes dentro de zonas con datos completos, es decir, las zonas que se imputaron se encuentran completamente sin información y pueden haber zonas mixtas en donde hayan datos completos y faltantes. Se podría adaptar la metodología de forma que incorpore estos dos niveles de información.

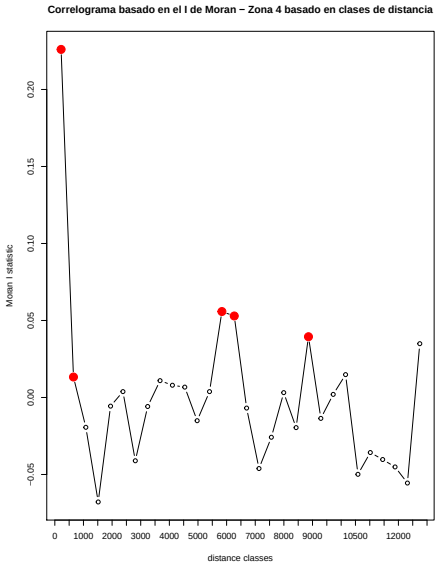
9. Anexo

9.1. Correlogramas basados en clases de distancia para las regiones

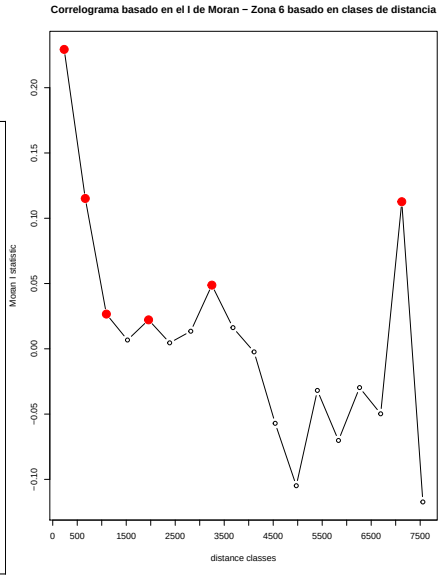




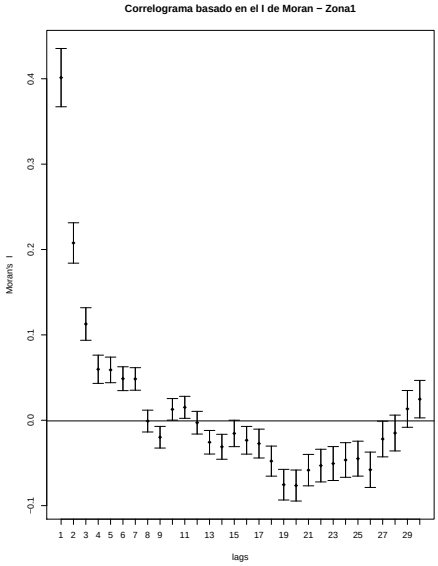


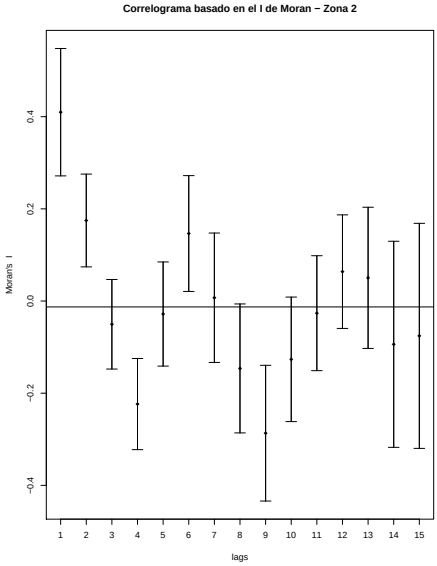


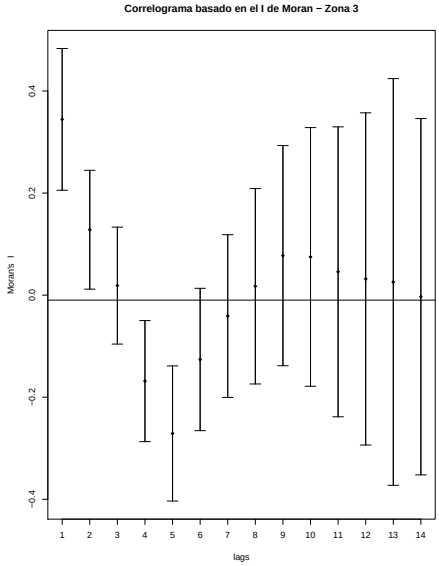
correlogDistZona5.pdf

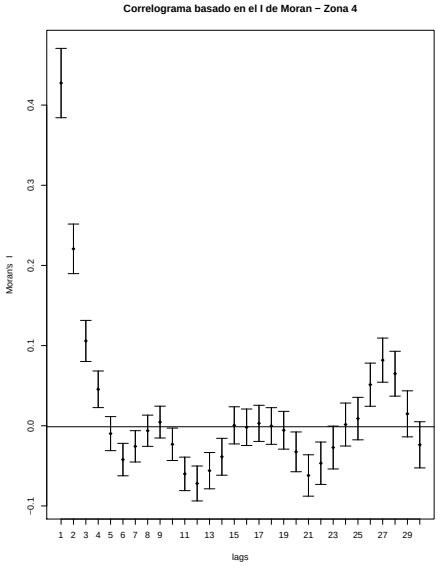


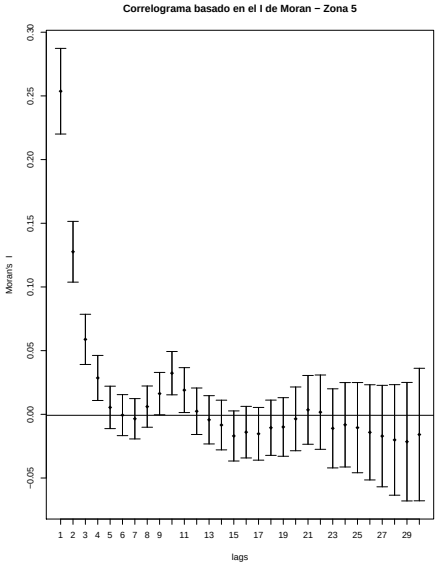
9.2. Correlogramas por contigüidad para las regiones

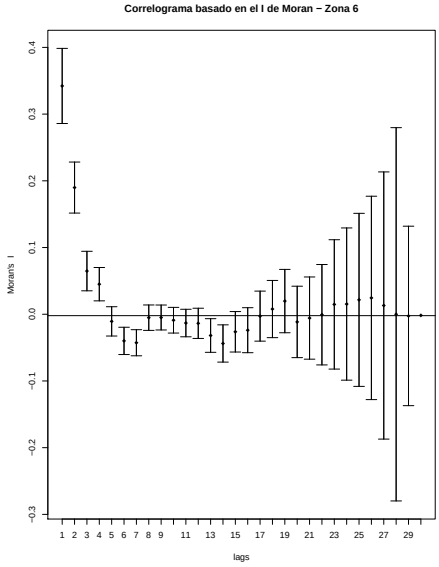












Referencias Bibliográficas

Gaetan, C. y Guyon, X. (2010). *Spatial Statistics and Modeling*. Springer, New York, NY.

Gaudart, J., Graffeo, N., Coulibaly, D., Barbet, G., Rebaudet, S., Dessay, N., Doumbo, O., y Giorgi, R. (2015). Spodt: An r package to perform spatial partitioning. *Journal of Statistical Software, Articles*, 63(16):1–23.

Gaudart, J., Poudiougou, B., Dicko, A., y Doumbo, O. (2005). Oblique decision trees for spatial pattern detection: optimal algorithm and application to malaria risk. *BMC Medical Research Methodology*, pp. 5–22.

Instituto de Estadística

Documentos de Trabajo

Eduardo Acevedo 1139. CP 11200 Montevideo, Uruguay
Teléfonos y fax: (598) 2410 2564 - 2418 7381
Correo: ddt@iesta.edu.uy
www.iesta.edu.uy
Área Publicaciones

Febrero, 2018
DT (18/1)