Tesina

“Indicadores globales de autocorrelación espacial para unidades de diferente tamaño”

Alumno: Ferraro, Sebastián Mario

Director: Dr. Pagura, José A.

Codirector: AUS. Mignoni, César

Carrera: Licenciatura en Estadística

Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística

AGRADECIMIENTOS

En este día se consuma una etapa muy importante de mi vida, y es grande la felicidad y emoción que siento. Este logro no es solo fruto de mi esfuerzo, sino que muchas otras personas estuvieron involucradas, a las cuales les debo mi agradecimiento.

En primer lugar, a mis padres, por su amor, apoyo y entrega incondicional, todos mis logros también son suyos.

A mis hermanos, Ariel y Pri, por aguantarme y acompañarme en cada momento de mi vida.

A Mica, por su amor invaluable y aliento para finalizar este trabajo.

A los profesores de la carrera, por formarme como profesional con mucha devoción y paciencia.

A Beto y César, por el compromiso y aporte fundamental en este trabajo, sin ellos no hubiese sido posible.

Y, principalmente, a Dios, de quien proviene toda fuente de conocimiento, sabiduría e inteligencia.

RESUMEN

Uno de los primeros pasos en el análisis de datos espaciales consiste en la verificación de existencia de autocorrelación espacial mediante el cálculo de algún indicador. El más divulgado y utilizado es el índice de Moran.

Cuando la variable de interés se observa sobre áreas irregulares de una región y sus valores son proporcionales al tamaño de cada área, suelen utilizarse razones o proporciones que tienen en cuenta el tamaño de dicha área utilizándolas en los análisis posteriores. En estos casos y frente a desigualdades importantes en los tamaños, se observan incrementos en la probabilidad de error de tipo I (Walter, 1992) y disminuciones en la potencia (Assunção y Reis, 1999) en las pruebas de significación para el índice de Moran.

Por este motivo, en la literatura se proponen índices alternativos que podrían corregir estas deficiencias. Entre estas propuestas, pueden distinguirse el índice de Oden (Oden, 1995) y el Índice Empírico de Bayes (Assunção y Reis, 1999).

En esta tesina se analizan aspectos teóricos de los tres índices mencionados, y se aplican en dos problemas: el estudio de la distribución espacial de los hogares con NBI en la ciudad de Rosario en el año 2010 y de los heridos por delitos con armas de fuego en la ciudad de Rosario durante un determinado año.

En términos generales, el índice Empírico de Bayes es el que posee las mejores propiedades estadísticas en cuanto a potencia, probabilidad de error de tipo I y robustez, principalmente en escenarios en los que las diferencias de tamaño entre las unidades son importantes.

ÍNDICE

1. INTRODUCCIÓN

En muchos problemas estadísticos, las variables que se consideran corresponden a unidades que se encuentran ubicadas en el espacio ocurriendo que aquellas unidades más cercanas tienen valores parecidos y a medida que la distancia es mayor las diferencias en los valores de las variables son también mayores (autocorrelación espacial). A diferencia de muchos métodos de la estadística clásica que suponen la independencia entre los datos, aquí no se cumple y en consecuencia se requieren métodos especiales para su análisis, los que se encuentran comprendidos en lo que se denomina Estadística Espacial. En estos estudios, se puede identificar una fase exploratoria destinada a comprender y describir las características relevantes del fenómeno observado y una fase destinada a modelar el comportamiento de las variables para la posterior explotación de dichos modelos.

En la primera fase, los recursos usuales consisten en herramientas gráficas e indicadores que pongan en evidencia la existencia de autocorrelación espacial y permitan detectar la naturaleza de la misma. Un indicador muy utilizado es el índice de Moran (Moran, 1950), el cual permite verificar si las unidades se distribuyen o no aleatoriamente en el espacio y proporciona una medida resumen de la intensidad de la autocorrelación.

En muchos de estos estudios ocurre que las unidades son de diferente tamaño, como los casos en que se observan variables para cada radio censal o departamentos de una región, lo que lleva a considerar la necesidad de tener en cuenta el tamaño de las distintas áreas contempladas a la hora de construir indicadores que expresen la autocorrelación.

Por ejemplo, si se desea estudiar la autocorrelación espacial para la variable número de hogares con necesidades básicas insatisfechas (NBI) observado en los radios censales de la ciudad de Rosario, puede ocurrir que en un radio censal con 100 hogares haya 10 con NBI y en otro radio censal con 10000 hogares haya 1000 con NBI. En ambos casos, la proporción de hogares con NBI es 0,10 pero evidentemente la situación es diferente.

Al calcular el índice de Moran para evaluar la autocorrelación espacial de las proporciones no hay distinción alguna entre situaciones como la

ejemplificada. El no tener en cuenta el desequilibrio en el tamaño de las correspondientes áreas trae como consecuencia reducciones en la potencia del test de hipótesis que se aplica para estudiar la significación estadística de índice mencionado.

Por este motivo, se han planteado diferentes índices alternativos como el presentado por Oden (Oden, 1995) y el Índice Empírico de Bayes o EBI (siglas de Empirical Bayes Index) desarrollado por Assunção y Reis en 1999. El primero de ellos brinda resultados satisfactorios para los problemas planteados, pero posee otro tipo de desventajas que se mencionarán en el cuerpo de la tesina. Mientras que el EBI posee mejores propiedades estadísticas y se adapta de una manera más razonable a este tipo de situaciones.

En el presente trabajo se exponen los conceptos fundamentales de cada uno de los tres índices mencionados y se aplican a los siguientes dos problemas: el estudio de la distribución espacial de los hogares con NBI en la ciudad de Rosario en el año 2010 y de los heridos por delitos con armas de fuego en la ciudad de Rosario durante un determinado año, el cual no se especifica por motivos de confidencialidad de la información.

Los objetivos de la tesina son:

I) Estudiar los fundamentos teóricos de los índices de Moran, Oden y el Índice Empírico de Bayes.
II) Comparar las propiedades de los índices mencionados, con la finalidad de determinar en qué situaciones es pertinente su aplicación.
III) Aplicar estos índices en problemas reales con la finalidad de observar e interpretar los resultados que proporcionan.

Tras este primer capítulo introductorio este escrito se estructura de la siguiente forma.

El capítulo 2, Material y Método, contiene la exposición de los fundamentos de los tres índices estudiados, la descripción de los problemas de aplicación y menciones necesarias acerca de la programación. El capítulo 3 se ha dedicado a presentar los resultados obtenidos en cada uno de los dos problemas mencionados, describiendo en primer lugar el comportamiento

espacial de las variables estudiadas, presentando luego los valores obtenidos para cada uno de los índices y finalmente se enuncian comentarios sobre estas aplicaciones.

El último capítulo del presente trabajo está destinado a los Comentarios Finales que se realizan a partir de cuestiones metodológicas y de aplicación.

2. MATERIAL Y MÉTODO

Este capítulo está dedicado a presentar los fundamentos metodológicos de diferentes alternativas para el cálculo de indicadores de autocorrelación espacial, así como a describir los problemas de aplicación en los que se utilizarán dichos índices.

En primer lugar, se introducen los conceptos sobre Estadística Espacial necesarios para la construcción de indicadores de autocorrelación espacial.

Luego, se realizan consideraciones a tener en cuenta a la hora de trabajar con índices de autocorrelación espacial, presentando los fundamentos teóricos de los índices de Moran, Oden y el EBI.

En tercer lugar, se presentan dos problemas en los que se utilizan los tres índices mencionados, con las particularidades que condujeron a su elección y se agrega la descripción de los conjuntos de datos empleados para las mencionadas aplicaciones.

2.1 Estadística Espacial. Conceptos básicos

En muchas ocasiones los datos con los que se debe tratar se corresponden con unidades que se encuentran situadas en el espacio y cuya localización puede especificarse (datos georreferenciados). Si la ubicación de las unidades es relevante para la descripción y análisis del fenómeno que se estudia, estos datos se denominan “datos espaciales”. Una particularidad que puede presentar esta clase de datos es la dependencia espacial, fenómeno conocido como autocorrelación espacial.

Para el tratamiento de esta clase de datos se ha desarrollado un conjunto de métodos, conocidos bajo el título Estadística Espacial, que tienen por objetivo la exploración, descripción, visualización, análisis, detección de la estructura espacial y su modelización y predicción espacial. Estos métodos atienden a las características particulares debidas a su distribución en el espacio. Los mismos se basan en considerar a los datos espaciales como la realización de un proceso estocástico en la región que se estudia.

En un análisis de datos espaciales pueden distinguirse tres etapas:

- Análisis exploratorio: se aplican métodos estadísticos convencionales para realizar un reconocimiento del comportamiento de las variables que se estudian y evidenciar la existencia de autocorrelación espacial. Sin la realización de esta etapa no se puede avanzar sobre las siguientes. Los métodos descriptivos se complementan con el cálculo de índices de autocorrelación espacial, el más utilizado es el índice de Moran cuyos fundamentos se presentan más adelante.
- Análisis estructural: constituye el objetivo principal de la Estadística Espacial y consiste en la construcción de modelos que describen la autocorrelación espacial.
- Predicción: a partir del modelo construido se predicen valores de la variable aleatoria que se estudia en sitios donde no se han realizado observaciones.

El interés del presente trabajo está en el estudio de los índices disponibles para la evaluación de la autocorrelación espacial. El ya mencionado índice de Moran, que es el más ampliamente difundido, puede encontrarse en muchos estudios publicados y los procedimientos de cálculo se encuentran programados en cualquier software de Estadística Espacial.

Este índice puede presentar inconvenientes cuando las unidades son de tamaños diferentes y por ese motivo se han propuesto otras alternativas que consideran dichas diferencias, las cuales se tratan en la presente tesina, como lo son el índice de Oden y el EBI.

Tipos de datos espaciales

En función de la manera en que se considere el espacio donde se observa la variable en estudio, los datos espaciales pueden dividirse en tres tipos:

- Geoestadísticos o espacialmente continuos: estos datos se pueden observar en cualquier posición y corresponden a un fenómeno que se desarrolla en forma continua en la región que se considera. Se puede

seleccionar cualquier localización del espacio en estudio para realizar una observación de la variable de interés.

- Reticulares o Lattice (látices): en esta situación cada observación se corresponde con agregaciones espaciales, es decir se observa una variable aleatoria sobre cada una de diferentes áreas en las que se divide la región que se estudia. Estas áreas son polígonos definidos por vértices y lados (fronteras). Según la forma que presenten estas superficies serán regulares o irregulares, las primeras dividen al espacio total de estudio en áreas idénticas, mientras que las segundas presentan distintas formas y tamaños. La definición de las áreas no resulta trivial, ya que el resultado final del estudio podría estar afectado por la delimitación elegida. La naturaleza del problema determinará la mejor manera de definir la división territorial en distintas áreas.
- Puntuales: la población en estudio es un conjunto de objetos distribuidos en el espacio y de tamaño pequeño en relación a la distancia que los separa. Cada objeto se encuentra en una posición aleatoria en el espacio.

Las unidades de análisis en los conjuntos de datos considerados en la presente tesina se corresponden con la segunda clase de los tipos de datos presentados anteriormente (reticulares), ya que se trata de los radios censales de la ciudad de Rosario.

2.2 Indicadores de Autocorrelación Espacial

Los métodos estadísticos tradicionales asumen que las observaciones de una variable se toman bajo condiciones idénticas y de manera independiente. Ellos consideran que los datos son una muestra aleatoria simple, es decir, son independientes e idénticamente distribuidos. Bajo esta suposición se construye la mayoría de la teoría estadística.

La consideración de la dependencia en los datos es un gran inconveniente a la hora de trabajar con los modelos usuales. Sin embargo, en muchos casos los modelos que incluyen dependencia son más realistas que los que no lo hacen. La idea de que datos cercanos, en el espacio o en el tiempo, son más parecidos que aquellos que se encuentran más alejados es natural.

En el contexto espacial, esta falta de independencia recibe el nombre de dependencia o autocorrelación espacial, la cual se define mediante una relación funcional entre lo que ocurre en una unidad determinada del espacio y en sus unidades vecinas. En otras palabras, existirá autocorrelación espacial cuando el valor observado de una variable en una unidad o área determinada dependa, en cierta manera, de los valores observados en unidades o áreas vecinas.

La autocorrelación espacial puede ser:

- Negativa: Se presenta una relación inversa entre los valores de la variable medida en la unidad con respecto a los valores en las unidades vecinas. Áreas con valores altos de la variable serán vecinas de áreas con valores bajos. A modo de ejemplo, se puede considerar la competencia entre plantas por la luz, donde zonas de plantas fuertes pueden estar rodeadas de otras con plantas menos fuertes.
- Positiva: la variable asumirá valores similares en unidades cercanas. Esta situación representa el efecto contagio, lo que ocurre en una unidad se "contagia" a áreas vecinas. Un área con un valor alto de la variable estará rodeada de unidades donde la variable también asuma valores altos.
- Nula: en esta situación no existe autocorrelación espacial, en otras palabras, la variable se distribuye de manera aleatoria en el espacio.

El cálculo de los índices de autocorrelación espacial permite 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. Para ello se necesita proporcionar criterios de cercanía entre unidades y pesos que reflejen la fuerza de la influencia en la relación entre las mismas.

2.2.1 Criterios de Vecindad y Pesos Espaciales

La primera ley de la geografía, o principio de autocorrelación espacial enuncia: "Todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes" (Tobler, 1970). Para poder entender cuando se considera que dos unidades se encuentran "cercanas" en el espacio nace el concepto de vecindad.

Bajo la perspectiva de datos reticulares se considera que no todas las áreas influyen sobre el valor que asume la variable en una determinada área, sino, que solo influirán aquellas que sean vecinas.

Definir qué características deben poseer dos áreas o unidades para que sean consideradas vecinas es una cuestión de suma relevancia. Existen varios criterios de vecindad que pueden utilizarse y se debe escoger el más apropiado para el conjunto de datos y según la naturaleza del problema.

Todos los criterios de vecindad deben cumplir que, al seleccionar una unidad o área, el resto de ellas queden particionadas en dos conjuntos mutuamente excluyentes, uno compuesto por unidades vecinas y otro por las que no lo son.

Una característica importante a considerar en los criterios de vecindad, es la simetría. Sea A un área determinada y, según el criterio que se está utilizando, la unidad B es vecina suya. Si el criterio utilizado es simétrico, entonces B también tendrá como vecina a A. Si el criterio no fuera simétrico, B no necesariamente tendrá a A entre el conjunto de sus unidades vecinas.

Criterios de vecindad

Los criterios de vecindad más utilizados y divulgados en la bibliografía son:

- Vecinos por contigüidad. Se define como áreas vecinas a aquellas en las que para ir de una a otra no haya que pasar por una tercera, es decir, que estén contiguas en el mapa. Existen tres criterios de vecindad basados en contigüidades: Reina, Torre y Alfil. Reciben estos nombres porque las unidades vecinas son aquellas a las que se accede según el movimiento de las piezas en el tablero de ajedrez.
    - Reina: en el ajedrez puede moverse a lo largo de la fila, la columna y las diagonales de la casilla en que se encuentre. Extrapolando esos movimientos a la situación de interés, dos áreas serán vecinas si tienen aristas o vértices en común.
    - Torre: solo puede moverse a lo largo de la fila y la columna en que se encuentre, no puede moverse en diagonal.

Análogamente se considera que dos áreas son vecinas si tienen aristas en común.

- Alfil: solo puede moverse a lo largo de la diagonal de la casilla en la que se encuentre. De esta manera, dos unidades en el espacio serán vecinas si y solo si tienen vértices en común.

Este criterio cumple la condición de simetría ya que, si un área tiene aristas o vértices en común con una segunda, esta segunda tendrá las mismas aristas o vértices en común con la primera.

- Vecinos basados en la distancia euclídea. Este método considera vecinas dos áreas si cumplen cierta condición referente a la distancia que las separa. Una forma usual de referirse a la distancia entre dos áreas es hacer que ella sea la distancia entre sus centroides. Existen dos variantes:
  - Los k vecinos más cercanos. Se calcula la distancia de la unidad considerada a todas las demás: serán vecinas las k unidades cuyas distancias sean las k menores. El número k se determina en base a la naturaleza de cada problema. Será una relación asimétrica en la que todas las áreas tendrán el mismo número de vecinos.
  - Dos áreas serán vecinas si y solo si la distancia entre ellas es menor a una magnitud fijada a priori. Este método es adecuado cuando las áreas tienen una distancia similar entre ellas, ya que si hay una distancia mucho mayor a las otras se presentará el problema de dejar esta unidad sin vecinos, o considerar un número de vecinos demasiado alto en el resto de áreas. Esta relación es simétrica.

Pesos espaciales

Una vez definido el criterio de vecindad a utilizar, resulta de interés cuantificar la fuerza de cada relación, esto es lo que se conoce como pesos espaciales. Por ejemplo, se sabe que A tiene dos áreas vecinas B y C, pero lo que se desconoce es si ambas influyen de la misma manera. En esta sección se explicarán los distintos criterios que pueden adoptarse a la hora de definir los pesos espaciales.

Los pesos se representan de forma matricial mediante la matriz cuadrada W. Cada elemento w_{ij} representa el peso de la relación de vecindad entre las áreas i y j . Cuando las unidades no son vecinas será w_{ij} = 0. La diagonal de la matriz de conectividad W será 0 ya que, por convención, una unidad no se considera vecina de ella misma por lo que W es una matriz cuadrada con todos sus elementos mayores o iguales a 0.

Esta matriz será simétrica si el criterio utilizado para definir los vecinos y los pesos lo son. Los pesos resultan simétricos cuando un área A ejerce sobre B la misma influencia que B sobre A. Un ejemplo de una situación en la que tiene sentido utilizar una relación asimétrica es considerar la influencia de las ciudades grandes sobre los pueblos de alrededor. Las grandes ciudades influyen más en las características de los pueblos que a la inversa.

Dentro de todos los posibles criterios para asignar los pesos, se distinguen dos grandes grupos; aquellos donde por el simple hecho de ser vecinos cada unión tendrá un peso común (simétrico) y aquellos en los que la importancia de las uniones variará en base a ciertas características (no simétrico).

Tres de los criterios más usuales son:

- Binario: es el criterio más sencillo, asume que w_{ij} = 1 cuando i y j son unidades vecinas y w_{ij} = 0 cuando no lo son. Este método simétrico es el más utilizado cuando existe poca información del proceso espacial. Bajo este criterio la suma de los pesos de un área es el número de vecinos que tiene.
- Estandarización por filas: este método no simétrico se basa en que los pesos de cada fila de la matriz sumen 1. Para ello se divide la unidad entre el número de vecinos que posee el área considerada. Según este método los pesos de áreas con pocos vecinos serán mayores que los de áreas con un número de vecinos mayor. Es decir, cada vecino de un área con pocos vecinos ejerce gran influencia sobre ella, mientras que los vecinos de áreas con muchos vecinos ejercen una influencia menor.
- Estandarización Global: considera el mismo peso para todas las relaciones de vecindad, definiendo el peso como el cociente entre la unidad y el

número total de vecinos (método simétrico). De esta manera, la suma de todos los pesos será igual a uno.

Existen otros criterios más específicos para determinar los pesos espaciales de la matriz de vecindad que han sido estudiados, pero en los párrafos anteriores se han mencionado solamente los más utilizados.

2.2.2 El índice de Moran (I)

Es el índice más usual. Fue desarrollado por Moran en 1950 y en la casi totalidad de los programas geoestadísticos se incluye su cálculo. Su interpretación es sencilla ya que es similar a la del coeficiente de correlación de Pearson. Se conocen sus propiedades estadísticas y pueden hacerse pruebas de hipótesis sobre su significación estadística. Sin embargo, cuando los tamaños de las unidades son diferentes y la variable para la cual se quiere evaluar la correlación espacial es una proporción o razón, este índice puede conducir a resultados erróneos como se comenta más adelante.

Se considera una región R dividida en m áreas r_i, i=1…, m. Por ejemplo, si la región es la ciudad de Rosario, las áreas o unidades podrían ser los radios censales.

Sea x_i el valor de una variable que refleja el tamaño del área i, por ejemplo, el total de hogares en el radio censal i.

Sea n_i el valor de la variable de interés en el área i, que en el primer problema a tratar es el número de hogares con NBI en el radio censal i.

La razón (o proporción) observada en el área i se define como $p_i=\frac{n_i}{x_i}$

El índice de Moran para la razón p_i, se define como (Moran, 1950):

$I = \frac{m}{\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij}(p_i-\bar{p})(p_j-\bar{p})}{\sum_i (p_i-\bar{p})^2}$ $\forall\ i \neq j$, donde

p_i es el valor de la razón en la unidad i a la que se le asocia el conjunto de coordenadas s_i, vector cuyas componentes son las coordenadas espaciales,

$\bar{p}=\frac{\sum_{i=1}^{m} p_i}{m}$ es la media de las razones p_i

w_{ij} es el elemento i,j de la matriz de conectividad que corresponde al peso entre las unidades i y j definida de la siguiente manera:

$$W = \begin{cases} w_{ij} & si\ i\ y\ j\ son\ vecinos,\ \mathrm{i} = 1,..,m,\ \mathrm{j} = 1,..,m \\ 0 & en\ otro\ caso \end{cases}$$

El índice de asociación de Moran resume la intensidad y dirección de la dependencia entre los valores de una variable observados en distintas unidades del espacio. Se obtiene calculando los productos cruzados de las diferencias entre las razones y su media para cada par (i,j) de unidades, ponderados por el peso w_{ij} correspondiente. Por lo tanto, el índice de Moran puede considerarse como una medida de correlación de cada p_i con el resto de las áreas con las que se encuentra vinculada. Al igual que el índice de correlación de Pearson, varía entre -1 y 1, y E[I]=$\frac{-1}{m-1}$ bajo la hipótesis de aleatoriedad espacial.

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

Para probar la significación estadística de I y así comprobar la hipótesis de no autocorrelación espacial se puede utilizar un test de hipótesis basado en supuestos de normalidad.

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

- E[I] =$\frac{-1}{m-1}$
- Var[I] =$\frac{m^2 \sum_{ij}^{m}(w_{ij}-w_{ji})^2 - m\sum_{i}^{m}(\sum_{j}^{m} w_{ij}+\sum_{j}^{m} w_{ji})^2 + 3[\sum_{ij}^{m} w_{ij}^2]}{(m^2-1)[\sum_{ij}^{m} w_{ij}]^2} - \frac{1}{(m-1)^2}$

Cuando no se cumple el supuesto de normalidad de p_i se puede utilizar un test permutacional: se encuentran las m factoriales 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.

Si m es grande, el trabajo computacional puede volverse dificultoso. Un recurso utilizado en estos casos es un test basado en el Método de Montecarlo que consiste en la realización de un test permutacional, pero sólo considerando un subconjunto de configuraciones elegidas al azar. Los paquetes computacionales suelen utilizar 999 permutaciones, y considerando la muestra observada, resultan 1000 valores de I para construir la distribución de aleatorización.

Un instrumento gráfico habitual en el análisis de autocorrelación espacial es el diagrama de dispersión de Moran, el cual representa en el eje de abscisas la variable de interés estandarizada y en el de ordenadas los retardos espaciales de dicha variable estandarizada, el cual es igual al promedio ponderado de la variable estandarizada en las áreas vecinas; las ponderaciones son los pesos asignados a cada unidad vecina. Este gráfico 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 se ajusta una recta de regresión lineal sobre los puntos del gráfico, el índice de Moran será igual a su pendiente. Si la nube de puntos está dispersa en los cuatro cuadrantes es indicio de ausencia de autocorrelación espacial.

Efecto de unidades de diferentes tamaños

Si se considera una región dividida en áreas con tamaños diferentes, como ocurre en los problemas que se estudian en esta tesina, al calcular la razón observada en cada unidad se estarían utilizando cantidades con diferentes denominadores en el cálculo de dichas razones.

Por ejemplo, si se desea estudiar la autocorrelación espacial para la variable proporción de hogares NBI observada en los radios censales de la ciudad de Rosario, tal como se mencionó con anterioridad, puede ocurrir que un radio censal con 100 hogares tenga 10 con NBI y otro radio censal con 10000

hogares tenga 1000 con NBI. En ambos casos, la proporción de hogares con NBI es 0,10 pero evidentemente la situación es diferente.

Si se utiliza el índice de Moran con las razones, no hay distinción alguna entre estas dos situaciones al momento de hacer los cálculos, es decir no se tiene en cuenta el "tamaño" de las correspondientes áreas.

Assunção y Reis (1999) estudiaron los efectos de tener diferentes tamaños de unidades sobre la probabilidad de error de tipo I en el test de significación del índice de Moran. Si se aplica la primera de las pruebas de hipótesis mencionadas se debe asumir que las razones están distribuidas normal, idénticamente y son independientes. Si las unidades tienen diferente tamaño y la variable aleatoria que se considera para cada unidad es una razón, la variancia será diferente ya que dependerá del tamaño de la unidad y la distribución posiblemente sea Poisson o Binomial. Por otra parte, las medias podrían ser diferentes ya que la inexistencia de autocorrelación espacial no descarta la heterogeneidad espacial.

Si se aplica un test permutacional y la situación es la descripta en el párrafo anterior, las distribuciones no son "intercambiables", condición requerida por esa clase de pruebas. Para entender un poco más el concepto de "intercambiabilidad" es importante mencionar que el test permutacional, a pesar de ser una prueba a distribución libre, no está excepta de los supuestos estadísticos de independencia e igual distribución de las unidades. Por lo tanto, al permutar los valores sobre las áreas se asume, bajo la hipótesis nula, que cada una de ellas posee la misma probabilidad de asumir uno de los valores observados de las razones. Utilizar distintos denominadores, al calcular las razones, en cada una de las unidades provoca que aquellas que sean de menor tamaño sean más propensas a asumir un valor extremo, esta situación pone de manifiesto la fragilidad del supuesto anteriormente mencionado. El test permutacional es sensible al tamaño de la muestra, variancias distintas y el tipo de distribución (Torres y otros, 2009).

2.2.3 El índice de Oden (I^*_{pop})

Cuando las unidades son de diferente tamaño, Oden (1995) propuso una corrección al índice de Moran. El conocido como Índice de Oden es:

$$I^*_{pop} = \frac{n^2 \sum_{ij}^m M^*_{ij}(e_i - d_i)(e_j - d_j) - n(1-2\bar{b})\sum_i^m M^*_{ii} e_i - n\bar{b}\sum_{ii}^m M^*_{ii} d_i}{\bar{b}(1-\bar{b})(x^2 \sum_{ij}^m d_i d_j M^*_{ij} - x\sum_i^m d_i M^*_{ii})}$$, donde: *(1)*

$n = \sum_i^m n_i$ total de la variable en estudio en la región. Por ejemplo, el total de hogares con NBI en la ciudad de Rosario.

$x = \sum_i^m x_i$ es el total en la región de la variable utilizada como denominador, siguiendo con el ejemplo: total de hogares en la ciudad de Rosario.

$\bar{b} = \frac{n}{x}$, proporción de interés en la región.

$e_i = \frac{n_i}{n}$, proporción de la variable en estudio en la unidad i, con respecto al total en la región de la misma.

$d_i = \frac{x_i}{x}$ proporción con respecto al total en la región de la variable utilizada como denominador en la unidad i.

Se llama M_{ij} al elemento i, j de la matriz M de pesos espaciales definida por Oden de la siguiente manera:

$$M = \begin{cases} w_{ij} & si\ i\ y\ j\ son\ vecinos, \quad \mathrm{i} = 1,..,m,\ \mathrm{j} = 1,..,m \\ 2 & si\ i = j, \quad \mathrm{i} = 1,..,m,\ \mathrm{j} = 1,..,m \\ 0 & en\ otro\ caso \end{cases}$$

Como puede verse, los elementos de la diagonal principal de la matriz de vecindad tendrán valores iguales a 2, a diferencia de las matrices de vecindad ya vistas donde la diagonal está compuesta por elementos iguales a 0. Oden propone esta modificación, con el objeto de diferenciar la situación de dos áreas que no son vecinas a aquella cuando se compara un área consigo misma.

La cantidad M^*_{ij} incluida en el índice de Oden es: $M^*_{ij} = \frac{M_{ij}}{\sqrt{(d_i d_j)}}$

Al igual que el índice de Moran, la prueba de significación estadística de la existencia de autocorrelación espacial con el índice de Oden se realiza bajo el supuesto de normalidad o mediante un test permutacional.

Bajo la hipótesis nula de inexistencia de autocorrelación espacial, E[I^*_{pop}]=$-(1-x)^{-1}$ y la variancia de I^*_{pop} puede obtenerse mediante la igualdad Var[I^*_{pop}] = E[$I^*_{pop}{}^2$] – E^2[I^*_{pop}]

El cálculo de E[$I^*_{pop}{}^2$] es complejo pero puede obtenerse mediante la igualdad E[$I^*_{pop}{}^2$] =$\frac{x[(x^2-3x+3)S_1-xS_2+S_0^2]-b_2[x_{(2)}S_1-2xS_2+6\,S_0^2]}{(x-1)_{(3)}S_0^2}$, donde:

- $S_0 = x^2A - xB$
- $S_1 = x^2C/2 - 2xD$
- $S_2 = x^3E - 4x^2F + 4xd$, donde:
  - $A = \sum_{ij}^m d_i d_j M_{ij}$
  - $B = \sum_i^m d_i M_{ii}$
  - $C = \sum_{ij}^m (M_{ij} + M_{ji})^2$
  - $D = \sum_i^m d_i M_{ii}{}^2$
  - $E = \sum_i^m (d_i[\sum_j^m (M_{ij} + M_{ji})]^2$
  - $F = \sum_i^m d_i M_{ii} \sum_j^m d_j (M_{ij} + M_{ji})$

Los valores que se encuentran indicados de la forma $x_{(k)}$ denotan factoriales descendientes $x_{(k)} = x\ (x - 1)\ (x - 2)\ldots(x - k + 1)$.

Bajo la hipótesis nula: E[I^*_{pop}] = $-(1-x)^{-1}$, y si m es suficientemente grande, la estadística Z = $\frac{I^*_{pop}-E[I^*_{pop}]}{\sqrt{Var[I^*_{pop}]}}$ se distribuye como una variable aleatoria normal estándar.

Oden (1995) muestra, mediante estudios por simulación, que el test que utiliza a I^*_{pop} es más potente que la prueba realizada con la estadística de Moran, cuando los tamaños de las unidades espaciales consideradas son distintos. Esta capacidad de capturar la variabilidad se debe al efecto del primer término en el numerador de *(1)*, que es una versión espacial de la prueba chi-cuadrado convencional para la heterogeneidad de proporciones.

Las hipótesis estadísticas para los índices de Moran y Oden

A pesar de la mayor potencia obtenida por el índice de Oden con respecto al índice de Moran se debe poner atención a la observación realizada por Assunção y Reis (1999), quienes advierten sobre lo inapropiado de la comparación entre las pruebas de hipótesis que se realizan con I e I^*_{pop}.

Para realizar la comparación entre los test se definen tres estados con respecto a la configuración espacial de las razones subyacentes de las áreas:

A: Razones espaciales homogéneas o constantes

B: Razones heterogéneas sin correlación espacial

C: Razones heterogéneas correlacionados espacialmente

La prueba de existencia de correlación espacial que propone Moran considera la hipótesis nula A U B y alternativa C, es decir solo será significativa la prueba cuando las razones sean heterogéneas y correlacionadas espacialmente.

Mientras que la prueba propuesta por Oden plantea la hipótesis nula A y alternativa B U C, es decir la prueba será significativa si las razones son heterogéneas, estén o no correlacionadas espacialmente (tabla 2.1).

Tabla 2.1: Hipótesis probadas por Moran y Oden.

Índices/Hipótesis	H_0	H_1
I	A U B	C
I^*_{pop}	A	B U C

Puede apreciarse que el estado B, razones heterogéneas sin correlación espacial, en el índice de Moran se acepta como parte de la H_0 y en el índice de Oden cuando ella se rechaza. El test basado en el índice de Oden conduce a aceptar la existencia de correlación espacial cuando la situación es B, es decir heterogeneidad espacial pero no autocorrelación.

En consecuencia no sorprende que el test basado en I^*_{pop} tenga mayor potencia, especialmente en estados como B, frente a los cuales el índice de Moran debería tener como máxima potencia la probabilidad de error de tipo I (Assunção y Reis, 1999).

De esta manera, resurgió la necesidad de encontrar un índice que tenga en cuenta el tamaño de las distintas áreas consideradas de una región a la hora de determinar si existe autocorrelación espacial de una variable aleatoria.

2.2.4 Índice Empírico de Bayes (EBI)

Assunção y Reis (1999) proponen un indicador alternativo para probar la existencia de autocorrelación espacial que llaman Índice Empírico de Bayes. En el artículo citado, realizan una reseña metodológica de los índices de Moran y Oden y una comparación de los tests destinados a comprobar la existencia de autocorrelación espacial mediante simulaciones en diferentes escenarios. Consideran la ciudad de Belo Horizonte, Minas Gerais, Brasil y la variable que se estudia es la tasa de homicidios en 1994, calculada en cada uno de los 81 distritos como el cociente entre el número de homicidios ocurridos y la población del distrito. La población de los distritos varía entre 31 y 70870 habitantes. Simulan para cada distrito el número de casos de acuerdo a una distribución de Poisson y se estudian tres escenarios:

- Población de tamaño constante en cada distrito y sin estructura espacial.
- Población de tamaño variable en cada distrito de acuerdo a un patrón no explicitado sin estructura espacial.
- Población de tamaño variable en cada distrito con estructura espacial.

Realizan 1000 simulaciones en cada escenario, llevando a cabo los tests correspondientes con un nivel de significación del 5% y observan la proporción de veces que rechazan la H_0 (no hay autocorrelación espacial), encontrando que el menos afectado por las diferencias en el tamaño de las unidades es el índice de Moran. Luego de proponer el EBI, comparan mediante un estudio por simulación en escenarios similares, su desempeño con respecto al del índice de Moran, evaluando la potencia de los tests y encontrando diferencias a favor del

EBI en los casos de heterogeneidad en el tamaño de las unidades. El artículo finaliza con la aplicación de los índices al problema de la distribución espacial de homicidios en Belo Horizonte mostrando los resultados que se obtienen y destacando la importancia del EBI.

A continuación, se describen brevemente los aspectos teóricos del EBI.

Un fenómeno en el espacio se trata como un proceso estocástico, es decir como una colección de variables aleatorias para las que se indica su ubicación en la región de estudio. En cada una de las áreas que se consideran, se tendrá una variable aleatoria distinta. En los estudios aquí abordados, la variable aleatoria es una razón o una proporción y se consideran los $\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_m$ parámetros - razones o proporciones- en las m áreas en estudio. Se realiza el supuesto de que el número de eventos observados n_i sigue una distribución Poisson[1] con media condicional $E(n_i|\theta_i) = Var(n_i|\theta_i) = x_i\theta_i$ siendo x_i el "tamaño" poblacional del área i. De esta forma, la media condicional de la razón estimada p_i es $E(p_i|\theta_i) = \theta_i$ y su variancia condicional es igual a $Var(p_i|\theta_i) = \frac{\theta_i}{x_i}$, por lo tanto las razones estimadas poseen distintas medias y variancias condicionales.

Assunção y Reis (1999), considerando el enfoque bayesiano, tratando a los parámetros θ_i como variables aleatorias y realizando el supuesto de que las razones θ_i tienen una distribución a priori con esperanza $E(\theta_i) = \beta$ y variancia $Var(\theta_i) = \alpha$, encuentran que la esperanza marginal de p_i es $E(p_i) = \beta$ y su variancia marginal es $Var(p_i) = \alpha + \frac{\beta}{x_i}$. Puede verse que las razones poseen la misma esperanza marginal (β) y las variancias marginales difieren entre ellas dependiendo de los tamaños de las unidades (x_i). Las variancias marginales de las razones p_i se incrementan a medida que los tamaños de las áreas disminuyen.

Marshall (1991), utilizando el método de los Momentos, propone los siguientes resultados para estimar los parámetros α y β:

[1] Por las características de las poblaciones con las que se trabaja en la presente tesina se considera la distribución de Poisson. En otras situaciones puede emplearse un razonamiento similar con la distribución Binomial.

$\widehat{\alpha} = a = s^2 - \frac{b}{(x/m)}$ $\widehat{\beta} = b = \frac{n}{x}$, donde

$s^2 = \sum_i^m \frac{x_i(p_i-b)^2}{x}$, x y n son el "tamaño" de la región y el total de la variable de interés en la región respectivamente.

Por lo tanto, la esperanza y variancia marginales de p_i son estimadas por b y $v_i = a + \frac{b}{x_i}$, respectivamente. Por convención, si $v_i < 0$, se define $v_i = \frac{b}{x_i}$.

En lugar de utilizar las razones estimadas p_i (como se emplean en el índice de Moran), se propone un nuevo índice que toma las razones estandarizadas utilizando las estimaciones presentadas anteriormente: $y_i = \frac{p_i-b}{\sqrt{v_i}}$

El Índice Empírico de Bayes se define de la siguiente manera:

$$EBI = \frac{m}{\sum_{ij}^m w_{ij}} \frac{\sum_{ij}^m w_{ij} y_i y_j}{\sum_i^m (y_i-\bar{y})^2}$$

Al igual que el índice de Moran, el EBI será significativamente distinto de su valor esperado si las razones están correlacionadas espacialmente. La prueba de independencia espacial depende de la distribución bajo la hipótesis nula del EBI, la cual se puede obtener mediante permutaciones.

Por lo tanto, se permuta independientemente el vector $(y_1,\ y_2,.,y_m)$ y se asignan aleatoriamente a las áreas una determinada cantidad de veces (en general se utilizan 999 permutaciones al igual que en el índice de Moran). Para cada uno de los ensayos obtenidos se calcula el EBI. El valor de la probabilidad asociada al test de hipótesis del EBI está dado por el cociente entre la cantidad de veces que el EBI permutado excede el EBI observado (numerador) y la cantidad de permutaciones utilizadas (denominador).

Es posible construir un gráfico de dispersión del EBI, el cual se obtiene de igual manera que para el índice de Moran, pero utilizando los y_i en lugar de la variable de interés estandarizada, es decir en el eje de las abscisas se grafican los y_i y en el eje de las ordenadas los retardos espaciales asociados a los y_i. La forma de interpretar este diagrama es equivalente a la del diagrama de dispersión de Moran.

Assunção y Reis (1999) estudian el efecto de “tamaños” de unidades heterogéneos sobre la potencia del test, es decir evalúan el impacto de la variación de los “tamaños” de las áreas cuando existe una correlación espacial entre las razones.

2.3 Problemas de aplicación

Para poder comparar los valores de los índices estudiados se consideran dos problemas de aplicación: el estudio de la distribución espacial de los hogares con NBI en la ciudad de Rosario en el año 2010 y de los heridos por delitos con armas de fuego en la ciudad de Rosario durante un determinado período.

Es oportuno que las comparaciones entre los índices de Moran, Oden y el EBI se realicen sobre una región formada por unidades de “tamaños” distintos, tal como lo son los radios censales de la ciudad de Rosario, motivo principal por lo que se escogieron dichas aplicaciones.

Para el primer problema, se dispone de un archivo de datos georreferenciado que contiene, entre otras variables, el número de hogares con NBI y el total de hogares para cada radio censal de la ciudad de Rosario en el año 2010, obtenido del sitio web del Instituto Nacional de Estadística y Censos.

En cuanto al segundo problema a considerar, se cuenta con un conjunto de datos que corresponden al número de heridos por armas de fuego en la ciudad de Rosario en un determinado año, problema similar al utilizado en Assunção y Reis (1999).

En base a los distintos tipos de datos espaciales que fueron presentados en la presente tesina es importante destacar tal como se mencionó con anterioridad que las unidades de análisis, en los conjuntos de datos considerados, son reticulares irregulares, donde cada una de las áreas o agregaciones espaciales corresponden a los radios censales de la ciudad de Rosario.

Para determinar la matriz de conectividad, se adopta el criterio de vecindad por contigüidad y más específicamente el de tipo Reina, ya que parecería ajustarse de una manera más adecuada a la naturaleza de los

problemas que se consideran. Por otro lado, se emplea el criterio de "estandarización por filas" para determinar el peso de cada una de las relaciones entre vecinos.

Se empleó el software R para efectuar los procedimientos estadísticos necesarios para realizar la presente tesina y en particular los paquetes *sp* y *spdep*. Se destaca que debió desarrollarse una función en R para el cálculo del índice de Oden, ya que no se encontró en ningún paquete de R ni en otro software. Los programas utilizados se encuentran disponibles para ser consumidos de manera colaborativa en el siguiente repositorio de código en GitHub: *https://github.com/Seferra18/Tesina*

3. APLICACIONES

Este capítulo está dedicado a tratar dos problemas en los que es oportuno el cálculo y comparación de los índices de Moran, Oden y el EBI. Primeramente, se describen las particularidades de los problemas, se presenta un breve análisis descriptivo y luego, se calculan los índices mencionados anteriormente junto con los resultados de las pruebas de hipótesis para evaluar la significación estadística de la existencia de autocorrelación espacial.

El primer problema es el estudio del comportamiento de la variable "número de hogares con necesidades básicas insatisfechas (NBI)" observada en cada radio censal de la ciudad de Rosario en el año 2010. En este caso, corresponde no solo estudiar la distribución de frecuencias y medidas descriptivas de la variable mencionada, sino que se debe tener en cuenta que la variable se encuentra referida a ubicaciones espaciales y por lo tanto se deberá dar el tratamiento adecuado para la descripción de datos espaciales. Un resultado importante en estos estudios, es la determinación de la existencia de autocorrelación espacial; en caso de darse este fenómeno, se pasaría a una segunda fase del estudio que es el modelamiento de dicha estructura espacial. La ciudad de Rosario cuenta en el año 2010, con 1073 radios censales con cantidades de hogares muy diferentes. Por ese motivo será más importante estudiar la existencia de autocorrelación espacial para la variable "proporción de hogares con NBI". Como se ha visto en el capítulo anterior, el desbalanceo en los denominadores de estas proporciones constituye un problema en el caso de utilizar el índice de Moran ya que se observa un aumento en la probabilidad de error de tipo I y una disminución en la potencia del test de significación (Assunção y Reis, 1999).

El segundo caso en consideración, es el estudio de la distribución espacial del número de heridos por delitos con armas de fuego en la ciudad de Rosario en un determinado período. Dadas las diferencias existentes en los valores de la población en cada radio censal es usual estudiar la razón de heridos por armas de fuego definida como el número de eventos ocurridos en un radio censal sobre la población en el mismo. En este caso se presenta una situación similar al problema anterior y se hace notar que se trata de un estudio muy parecido al que se encuentra como ejemplo de aplicación del EBI, en el artículo en el que se

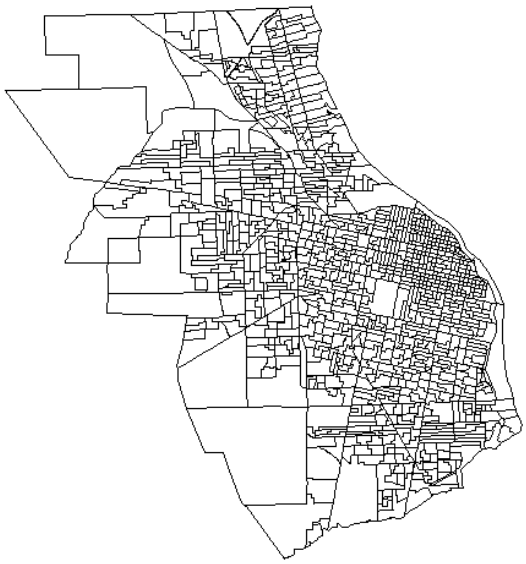
propone este índice (Assunção y Reis, 1999). En dicho trabajo científico se estudia la distribución espacial del número de homicidios en las diferentes áreas administrativas de Belo Horizonte en un determinado período.

En ambos casos se realiza un estudio descriptivo, usual para datos espaciales, y luego se obtienen los tres índices presentados en el capítulo Material y Método.

3.1. Radios censales, número de hogares y población en la ciudad de Rosario

La ciudad de Rosario en el año 2010, tal como se menciona anteriormente, se encontraba dividida en 1073 radios censales cuya distribución geográfica se muestra en la figura 3.1. La población de esta ciudad según el Censo Nacional de Población, Hogares y Viviendas 2010 era de 951856 habitantes y existían 321715 hogares.

Figura 3.1: Radios censales de la ciudad de Rosario.



La distribución de frecuencias del número de hogares por radio censal se presenta en la figura 3.2 donde, en complemento con la tabla 3.1, se puede apreciar que el número medio de hogares por radio censal es 300 ($\overline{x}$) y la desviación estándar 107 (s) . La mediana (Q_2) es un valor parecido a la media (292) y el 25% de los radios censales tiene 234 hogares o menos (Q_1). El 25% de los radios más poblados tienen entre 352 (Q_3) y 1329 hogares. Todo esto

permite apreciar el desequilibrio en el tamaño de los radios censales expresado en término del número de hogares.

Figura 3.2: Cantidad de radios censales según número de hogares, en la ciudad de Rosario en el año 2010.

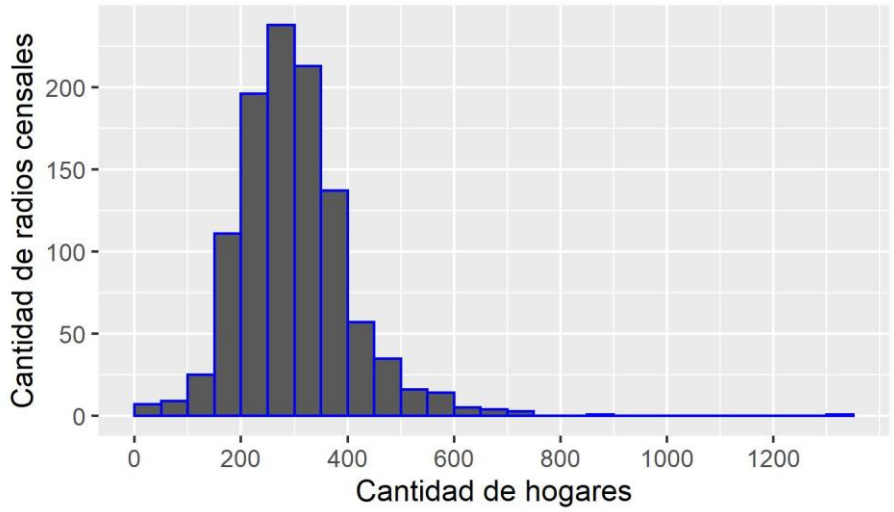


Tabla 3.1: Medidas descriptivas de la cantidad de radios censales según número de hogares, en la ciudad de Rosario en el año 2010.

Mínimo	**Q_1**	**Q_2**	**Q_3**	**Máximo**	$\bar{x}$	s
4	234	292	352	1329	300	107,5

En la figura 3.3. se presenta la distribución de frecuencias del número de habitantes en los radios censales de la ciudad de Rosario donde, en complemento con la tabla 3.2, se observa que el número medio de habitantes por radio censal es 888 ($\bar{x}$), superior al valor mediano 811 (Q_2). La desviación estándar es 427 (s) y el 25% de los radios censales tiene 603 habitantes o menos (Q_1). El 25% de los radios más poblados tienen entre 1070 (Q_3) y 4663 hogares lo que, nuevamente, permite apreciar el desequilibrio en el tamaño de los radios censales, en este caso, expresado en término del número de habitantes.

Figura 3.3: Cantidad de radios censales según número de habitantes, en la ciudad de Rosario en el año 2010.

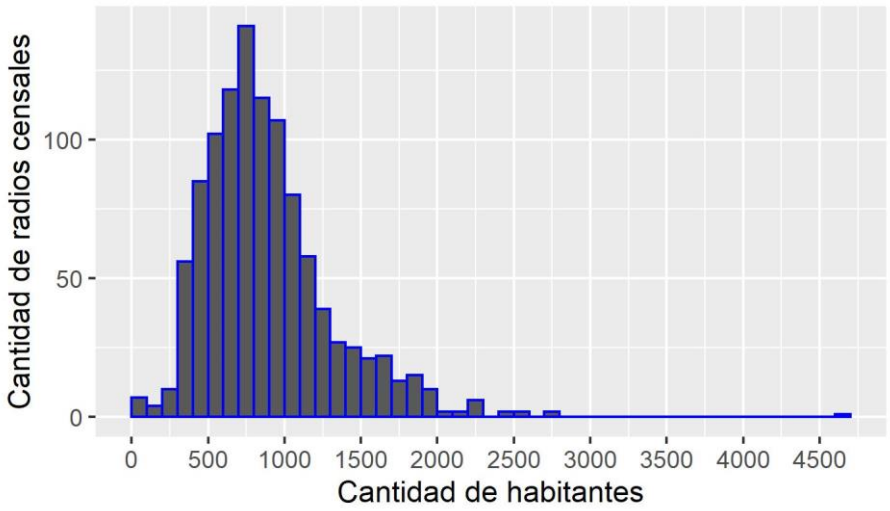


Tabla 3.2: Medidas descriptivas de la cantidad de radios censales según número de habitantes, en la ciudad de Rosario en el año 2010.

Mínimo	**Q_1**	**Q_2**	**Q_3**	**Máximo**	$\bar{x}$	s
12	603	811	1070	4663	888	427

3.2. Criterio de vecindad

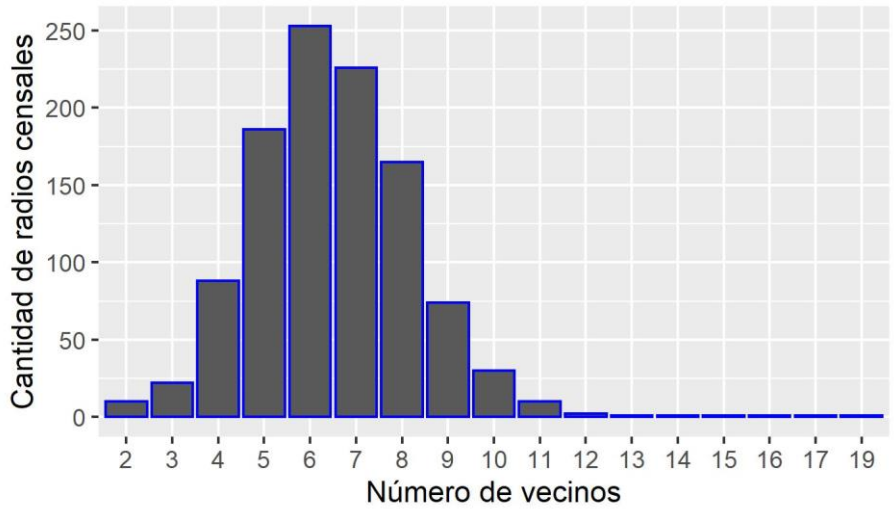
Previo a la realización de los análisis destinados a evaluar la existencia de autocorrelación espacial y teniendo en cuenta que en ambos problemas la unidad observacional es el radio censal, se trata la definición del criterio de vecindad que se aplicará en los dos problemas.

Como se mencionó, se opta por un criterio de contigüidad y se utiliza el tipo Reina. Los pesos de las unidades vecinas se establecen en base al criterio de estandarización por filas.

Cada par de radios censales que compartan al menos una arista o vértice en el espacio serán vecinos, ya que así lo establece el criterio de vecindad elegido.

En la figura 3.4 se presenta la distribución de frecuencias del número de vecinos que tiene cada radio censal. El 50% de los radios censales tienen 6 o menos vecinos. Entre los radios censales con más de 6 vecinos se destacan 7 unidades que poseen entre 12 y 19 vecinos.

Figura 3.4: Distribución del número de vecinos de los radios censales de la ciudad de Rosario.



3.3. Autocorrelación espacial en los hogares con NBI

Teniendo en cuenta las diferencias existentes en la cantidad de hogares en cada radio censal, en lugar de estudiar la autocorrelación espacial del número de hogares con NBI, se estudia el comportamiento espacial de la proporción de hogares con NBI. El primer paso que usualmente se lleva a cabo en el análisis exploratorio de datos espaciales es la construcción de representaciones gráficas que permitan apreciar el comportamiento de la variable en estudio en la región que se considera. La distribución de frecuencias de la proporción de hogares con NBI se presenta en la figura 3.5. En la tabla 3.3 puede apreciarse que la mediana de la proporción de hogares con NBI es 0,025 (Q_2), es decir la mitad de los radios censales tiene un 2,5% o más de hogares con NBI. En el 25% de los radios censales con menores valores de la variable estudiada, existe una proporción de hogares con NBI menor a 0,007 (Q_1), mientras que en el 25% superior se observa una proporción de hogares con NBI mayor a 0,068 (Q_3). Cabe comentar que hay tres radios censales con pocos hogares (23, 8 y 4) cuyas proporciones de hogares

con NBI resultan iguales a 1. Estos se reconocen como outliers superiores, con respecto a la proporción mencionada anteriormente, definidos como los valores que superan la cantidad Q_3+1,5*RI, donde RI=(Q_3-Q_1). Para la proporción de hogares con NBI aquellos radios censales con una menor cantidad de hogares son más propensos a asumir un valor extremo, tal como se mencionó en el capítulo de Material y Método.

Figura 3.5: Distribución de la proporción de hogares con NBI en los radios censales de la ciudad de Rosario.

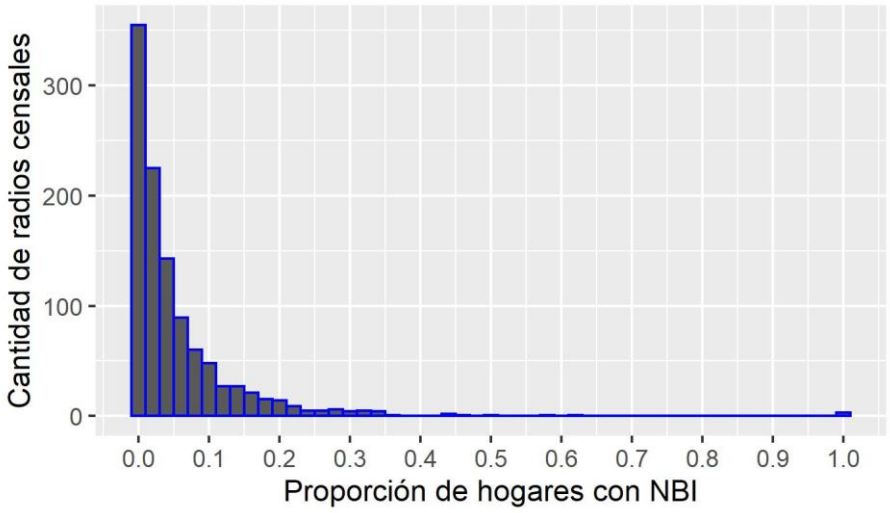


Tabla 3.3: Medidas descriptivas de la proporción de hogares con NBI en los radios censales de la ciudad de Rosario.

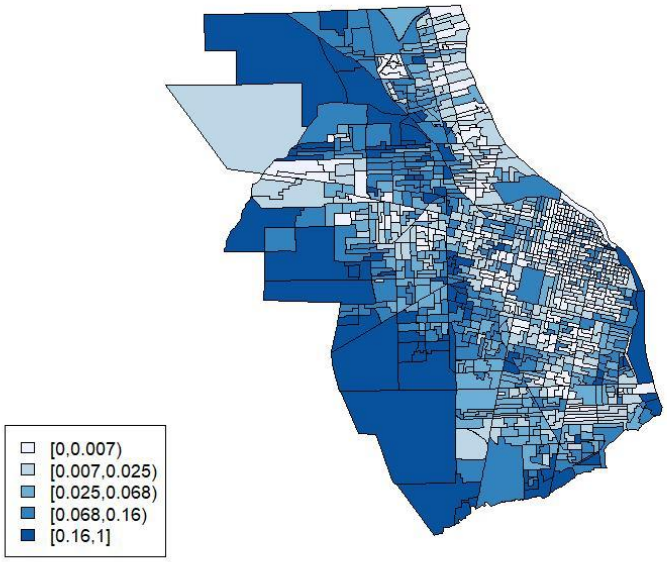
Mínimo	**Q_1**	**Q_2**	**Q_3**	**Q_3 + 1,5*RI**	**Máximo**	$\bar{p}$	s
0,000	0,007	0,025	0,068	0,160	1,000	0,055	0,088

El Box Map, figura 3.6, muestra la agrupación de radios con menor proporción de hogares con NBI, los que se corresponden a áreas con tonalidades más claras, y se encuentran principalmente en la zona céntrica de la ciudad y en un sector de la zona norte. En la zona noroeste se distingue un grupo de radios censales con baja proporción de hogares con NBI, pero rodeados de radios con valores bajos de esta variable. Por otro lado, los radios con mayor proporción de

hogares con NBI se han representado con tonalidades más oscuras y se concentran mayormente en la zona sur y oeste de la ciudad. Las categorías definidas para armar el Box Map son:

- [0; 0,007): corresponde a los radios con proporciones de NBI inferiores al primer cuartil.
- [0,007; 0,025): agrupa los radios con proporciones de hogares con NBI comprendidos entre el primer cuartil y la mediana.
- [0,025; 0,068): corresponde al 25% de radios censales con proporciones entre la mediana y el tercer cuartil.
- [0,068; 0,16): agrupa los radios con proporciones de hogares con NBI iguales o mayores que el tercer cuartil y menores que el valor máximo habiendo excluido los "outliers".
- [0,16; 1): corresponde a "outliers" superiores.

Figura 3.6: Box Map de la proporción de hogares con NBI en los radios censales de la ciudad de Rosario.



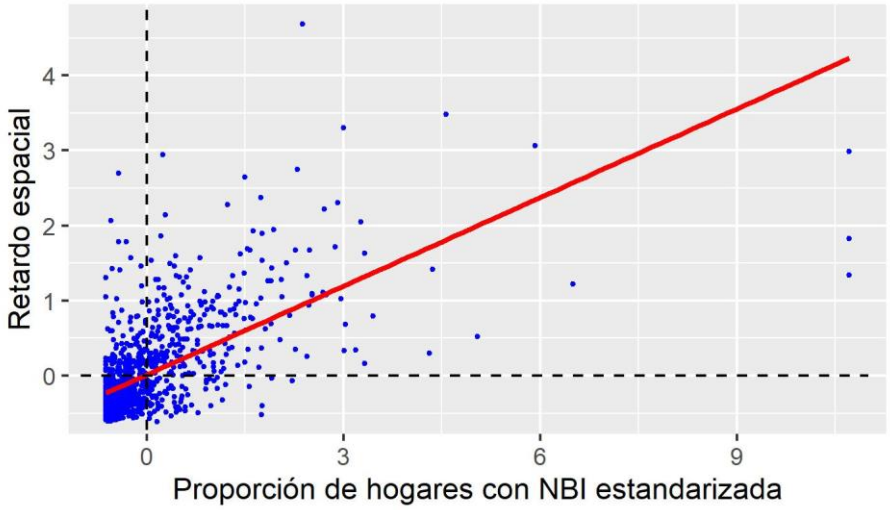
El próximo paso en un análisis exploratorio de datos espaciales es evaluar si la variable analizada tiene una estructura espacial, para esto se utiliza un índice de autocorrelación espacial, se comenzará con el cálculo del índice de Moran.

La disposición de los puntos en la figura 3.7 indica una autocorrelación espacial positiva de la proporción de hogares con NBI, ya que hay un predominio de puntos en los cuadrantes I y III. La pendiente de la recta de mínimos cuadrados ajustada sobre la nube de puntos coincide con el estadístico de Moran, el cual es igual a 0,394.

Los tres puntos que asumen un valor de la proporción igual a 1 correspondientes a radios censales con pocos hogares, mencionados ya en el Box Map, fuerzan la pendiente de la recta de regresión hacia la dirección que resulta en la representación gráfica.

Se realizó un test de hipótesis para evaluar la existencia de autocorrelación espacial, obteniéndose una probabilidad asociada igual a 0,001, por lo que se rechazó que I sea igual a -1/1072 (H_0) siendo 1072 el número de radios censales de la región considerada menos 1. El valor del índice de Moran, igual a 0,394, indica la existencia de autocorrelación espacial positiva, es decir que la proporción de hogares con NBI no se distribuye de manera aleatoria en la ciudad de Rosario.

Figura 3.7: Gráfico de dispersión de Moran para la proporción de hogares con NBI.



Otro índice planteado para la detección de la autocorrelación espacial es el índice propuesto por Oden en 1995. En el apartado 2.2.3 del capítulo Material y Método se realiza una presentación de los fundamentos de dicho índice.

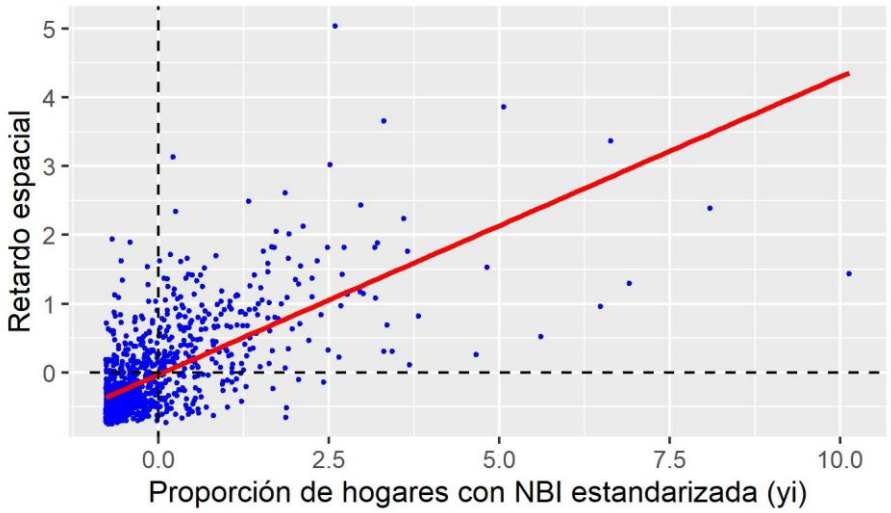
El I^*_{pop} resultó igual a 0,090 con una probabilidad asociada a la prueba de hipótesis menor a 0,001, por lo que existe evidencia estadística para rechazar la hipótesis de que I^*_{pop} = -1/(321715-1), donde 321715 es el número de hogares en la ciudad de Rosario. Esto se traduce, en términos del problema, en que la proporción de hogares con NBI no es homogénea a lo largo de los radios censales, pero no se tiene certeza de que exista una estructura espacial.

Es importante recordar, que la prueba asociada al índice de Oden es más potente que Moran pero que sus hipótesis nula y alternativa no concuerdan.

La pequeña magnitud de la estadística de Oden junto con su probabilidad asociada casi nula puede explicarse por el hecho de que la prueba es muy potente, por lo tanto, pequeños alejamientos del valor esperado serán detectados, aunque sin poder diferenciar si el rechazo de la hipótesis nula se debe a la existencia de autocorrelación espacial o de variabilidad espacial.

Por último, se calculó el EBI, el cual tiene mejores propiedades estadísticas que el índice de Moran, de acuerdo a lo presentado por Assunção y Reis (1999). La magnitud encontrada para este índice fue 0,433 proporcionando valores estadísticamente significativos que permiten concluir que la proporción de hogares con NBI no se distribuye de manera aleatoria en la ciudad de Rosario, sino que existe autocorrelación espacial positiva. Se acompaña el cálculo del EBI con su diagrama de dispersión correspondiente, el cual se presenta en la figura 3.8, donde se visualiza que la mayoría de puntos se encuentran en los cuadrantes I y III, indicando una relación directa entre los radios censales con respecto a la proporción de hogares con NBI.

Figura 3.8: Gráfico de dispersión del EBI para la proporción de hogares con NBI



Puede apreciarse que ambos índices arrojan resultados significativos, siendo el EBI aproximadamente igual a 0,43 mientras que el estadístico de Moran es cercano a 0,39.

Es importante mencionar que, el hecho de considerar el tamaño de los radios censales a la hora de calcular un índice de autocorrelación espacial proporciona una ventaja a la hora de trabajar con radios censales, como los mencionados anteriormente, que poseen pocos hogares y una proporción de hogares con NBI igual a la unidad.

Debe también realizarse un comentario acerca de la situación observada en la figura 3.7 donde se distinguen claramente por lo separado de la nube de puntos, tres radios con muy pocos hogares, disposición que no se observa en la figura 3.8. Esto puede vincularse con una referencia, que hacen los autores del artículo que se viene mencionando, acerca de una "cualidad adicional de robustez" del EBI. Assunção y Reis (1999) visibilizan esta propiedad en el estudio espacial de los homicidios en Belo Horizonte, donde se presenta un dato anómalo. Obtienen los índices de Moran, Oden y EBI con el conjunto de datos que incluye al dato anómalo y luego excluyéndolo, llegando en un caso a aceptar la existencia de autocorrelación espacial con I e I^*_{pop} y en otro a rechazar. Mientras que al emplear EBI la conclusión se mantiene en ambos casos, de esta manera presentan la "cualidad adicional de robustez" de este índice.

En este trabajo se calcularon los índices incluyendo y excluyendo los radios censales extremos en cuanto al tamaño, obteniendo cambios importantes en el índice de Moran y diferencias menores en el EBI.

La tabla 3.4 contiene los valores de los índices, incluyendo y excluyendo los tres radios censales anómalos, junto con la probabilidad asociada a sus correspondientes pruebas de hipótesis.

Tabla 3.4: índices de autocorrelación espacial calculados para la proporción de hogares con NBI.

Índice	Estadístico	P-Valor
Conjunto completo de datos		
Moran (I)	0,39364	0,001
Oden (I^*_{pop})	0,08953	<0,001
EBI	0,43339	0,001
Conjunto excluyendo tres radios con información anómala		
Moran (I)	0,47990	<0,001
Oden (I^*_{pop})	0,08838	<0,001
EBI	0,48152	0,001

3.4 Autocorrelación espacial en los heridos por armas de fuego

De manera similar a la sección 3.3, se aplicarán los distintos índices sobre el conjunto de datos compuesto por la localización de los heridos por arma de fuego en la ciudad de Rosario. Teniendo en cuenta las diferentes cantidades de habitantes en cada radio censal, se estudia la distribución espacial de la razón de heridos por delitos con armas de fuego en la ciudad de Rosario, cantidad resultante del cociente, en cada radio censal, del número de heridos por arma de fuego y el total de habitantes.

De manera similar a la aplicación de hogares con NBI, se construirán representaciones gráficas que ayuden a estudiar el comportamiento espacial de la razón de heridos por arma de fuego en Rosario.

La distribución de frecuencias de la razón de heridos por delitos con armas de fuego se presenta en la figura 3.9; puede verse que la mayoría de los radios censales (834) no poseen heridos, provocando que todos los cuartiles (primero, segundo y tercero) sean iguales a 0.

Las medidas descriptivas de la razón de heridos por delitos con armas de fuego se presentan en la tabla 3.5, donde puede apreciarse que la razón promedio de heridos por arma de fuego es 0,0005 ($\bar{r}$) con una desviación estándar de 0,0013 (s).

Figura 3.9: Distribución de la razón de heridos por arma de fuego en los radios censales de la ciudad de Rosario.

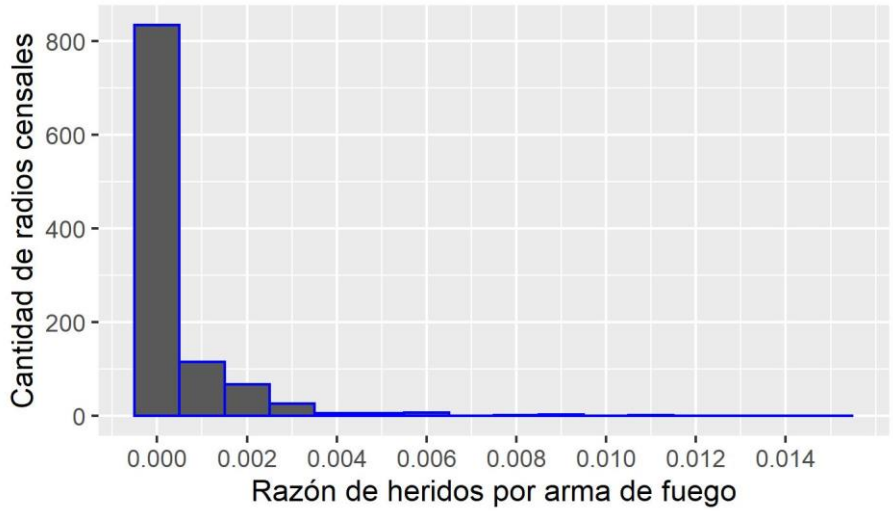


Tabla 3.5: Medidas descriptivas de la razón de heridos por arma de fuego en los radios censales de la ciudad de Rosario.

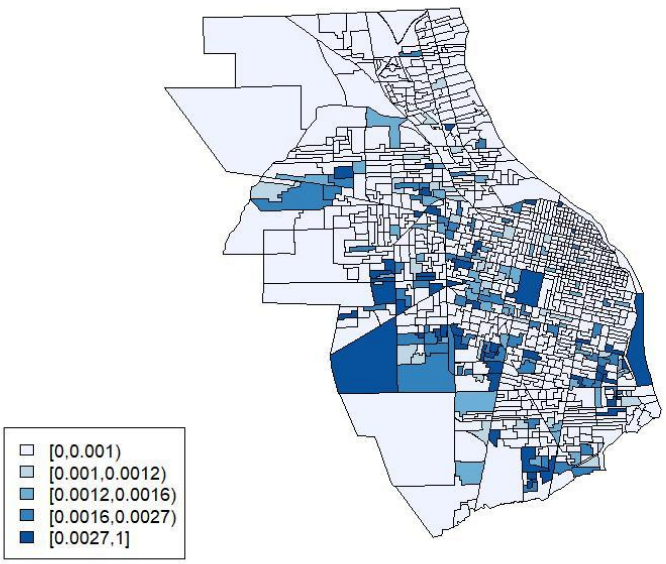
P_{75}	P_{80}	P_{85}	P_{90}	P_{95}	Máximo	$\bar{r}$	s
0,000	0,001	0,0012	0,0016	0,0027	0,015	0,0005	0,0013

En lugar de construir un Box Map, se presenta un mapa de percentiles, determinados de 5 en 5 comenzando por el percentil P_{75}, dando lugar a las siguientes categorías:

- [0; 0,001): valores inferiores al percentil P_{80}.
- [0,001; 0,0012): valores comprendidos entre los percentiles P_{80} y P_{85}.
- [0,0012; 0,0016): valores comprendidos entre los percentiles P_{85} y P_{90}.
- [0,0016; 0,0027): valores comprendidos entre los percentiles P_{90} y P_{95}.
- [0,0027; 1): valores superiores al percentil P_{95}.

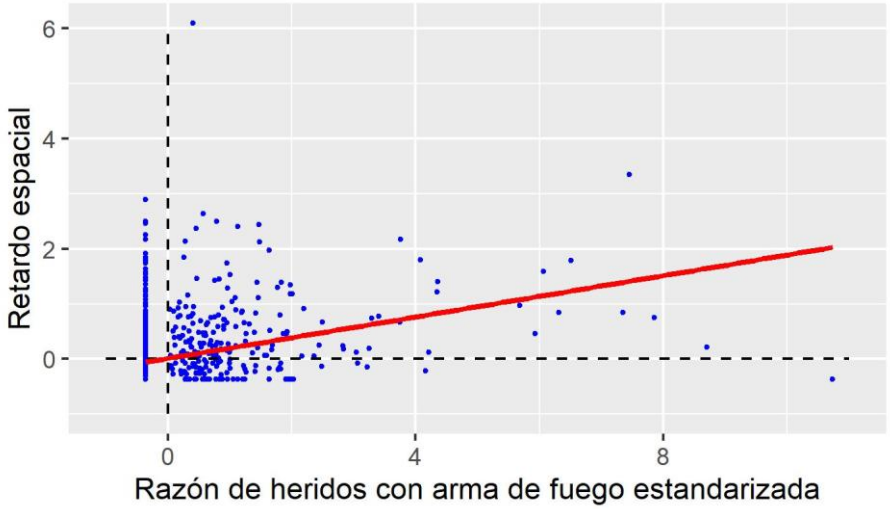
El mapa de la figura 3.10 muestra la agrupación de radios con menor razón de heridos por delitos con armas de fuego en las áreas representadas con una tonalidad más clara, por otro lado, los radios con mayor razón de heridos por delitos con armas de fuego asumen un color más oscuro. Si bien no se observa un patrón claro, puede identificarse mayoritariamente concentraciones de radios censales con razones altas en las zonas oeste y sur de la ciudad de Rosario.

Figura 3.10: Mapa de percentiles de la razón de heridos por arma de fuego en los radios censales de la ciudad de Rosario.



Se calcula el índice de Moran (0,188) y se acompaña con el diagrama de dispersión para el mismo en la figura 3.11. Este índice muestra la existencia de autocorrelación espacial positiva, la que se confirma mediante el correspondiente test de hipótesis, obteniéndose una probabilidad asociada igual a 0,001 y rechazando así la hipótesis de que en la ciudad de Rosario la razón de heridos por arma de fuego se comporta de manera aleatoria a lo largo de toda la región.

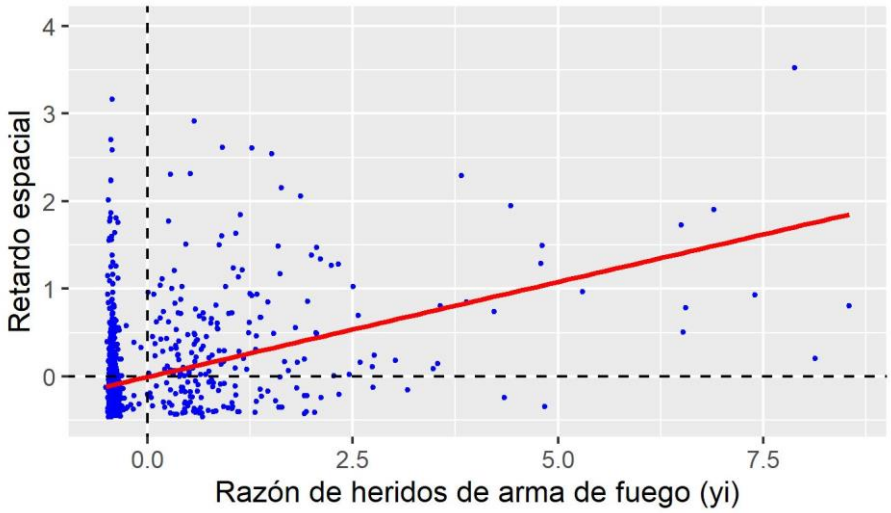
Figura 3.11: Gráfico de dispersión de Moran para la razón de heridos por arma de fuego en los radios censales de la ciudad de Rosario.



El segundo índice que se calcula es el propuesto por Oden, el cual resultó igual a 0,00155 con una probabilidad asociada a la prueba de hipótesis menor a 0,001, por lo que existe evidencia estadística para rechazar la hipótesis de que la razón de heridos por arma de fuego es homogénea a lo largo de los radios censales de la ciudad de Rosario, pero sin saber si existe un patrón espacial en la variable analizada.

Por último, se calculó el EBI cuyo valor fue 0,21715 proporcionando valores estadísticamente significativos que permiten rechazar la inexistencia de autocorrelación espacial. Se acompaña este indicador con su gráfico de dispersión correspondiente (figura 3.12), que pone en evidencia la relación directa que existe entre los radios censales de la ciudad de Rosario para la razón de heridos por armas de fuego.

Figura 3.12: Gráfico de dispersión del EBI para la razón de heridos por arma de fuego en los radios censales de la ciudad de Rosario.



La tabla 3.6 contiene los valores de los índices junto con la probabilidad asociada a sus correspondientes pruebas de hipótesis para los dos conjuntos de datos: completo y excluyendo outliers.

Tabla 3.6: índices de autocorrelación espacial calculados para la razón de heridos por arma de fuego.

Índice	Estadístico	P-Valor
Conjunto completo de datos		
Moran (I)	0,18779	0,001
Oden (I^*_{pop})	0,00155	<0,001
EBI	0,21715	0,001
Conjunto excluyendo tres radios con información anómala		
Moran (I)	0,21887	0,001
Oden (I^*_{pop})	0,00156	<0,001
EBI	0,22653	0,001

En términos generales, los resultados de los tres índices muestran una tendencia general similar a la aplicación en hogares con NBI.

Se ha considerado conveniente incluir al final de este capítulo algunas conclusiones específicas a partir de las aplicaciones realizadas.

En las pruebas de hipótesis para los tres índices en los dos problemas se rechazó la hipótesis nula, aceptando la existencia de autocorrelación espacial; es decir, a pesar de haber utilizado metodologías diferentes arriban a la misma conclusión.

Debe destacarse que, al utilizar el índice de Oden no se sabe si la hipótesis nula se rechaza por la existencia de autocorrelación espacial o porque las razones son diferentes.

Si bien el EBI y el índice de Moran conducen a la misma conclusión en los casos estudiados, el primero de ellos debe considerarse como una alternativa válida frente a las desigualdades en los denominadores de las razones, debido a que el test es más potente. Además, la probabilidad de error de tipo I cuando no hay autocorrelación espacial es mayor que el valor nominal en el caso del índice de Moran y es similar al valor nominal en el caso de EBI, de acuerdo a los resultados presentados por Assunção y Reis, 1999.

4. COMENTARIOS FINALES

Uno de los primeros pasos en el análisis de datos espaciales es el diagnóstico de la existencia de autocorrelación espacial. Cuando las unidades son áreas en las que se ha dividido el territorio de interés, el índice que se aplica en la mayoría de los casos es el de Moran. Si las unidades son de tamaño diferente y dicho tamaño, como sucede frecuentemente, está relacionado con la variable de interés principal se estudia la distribución espacial de razones que tienen como denominador el tamaño del área. En esos casos, las pruebas de hipótesis sobre el índice de Moran pierden potencia con respecto al caso en que las áreas son de igual tamaño.

Como alternativas se han propuesto los índices de Oden y el EBI cuyas características fueron presentadas en el capítulo Material y Método. Entre ellas se pueden mencionar:

- Los test de hipótesis para el EBI y el índice de Moran presentan una potencia similar ante escenarios de tamaños parecidos de las distintas áreas consideradas, pero a medida que se alejan de esta situación el EBI incrementa su potencia de manera considerable con respecto al índice de Moran (Assunção y Reis, 1999). En cuanto a la potencia del test para el índice de Oden no es comparable con los otros dos casos, ya que se ha destacado que prueba hipótesis diferentes.

- La probabilidad de error de tipo I de la prueba asociada al EBI no se ve alterada cuando los tamaños de las áreas son heterogéneos (Assunção y Reis, 1999), mientras que la probabilidad de error de tipo I en la prueba de hipótesis que utiliza el índice de Moran se incrementa cuando los tamaños de las áreas son diferentes (Walter, 1992).

- El EBI posee cualidades de robustez viéndose poco afectado por valores atípicos, según señalan Assunção y Reis, 1999.

Se consideraron dos problemas en los que las unidades son de diferente tamaño:

- El estudio de la distribución espacial de la proporción de hogares con necesidades básicas insatisfechas en la ciudad de Rosario en el año 2010. Esta variable está observada en cada radio censal de la ciudad, observando una importante heterogeneidad en el tamaño de las unidades

- El estudio del comportamiento espacial de la razón de heridos por delitos con armas de fuego en la ciudad de Rosario, también referida a cada radio censal.

Luego de un análisis descriptivo, el cual permite apreciar la heterogeneidad del tamaño de los radios censales y la existencia de correlación espacial, se calcularon los tres índices estudiados.

Para la realización de los cálculos se utilizó el software R y hubo que desarrollar un programa para obtener el índice Oden, el cual se puede encontrar en el repositorio de GitHub situado en la sección 2.3 de la presente tesina.

Las pruebas de hipótesis para los tres índices condujeron al rechazo de la H_0, es decir que no hubo resultados contradictorios, aunque en el caso del índice de Oden no se puede confirmar si lo que se detectó fue autocorrelación o heterogeneidad espacial.

En la etapa descriptiva se aprecia la existencia de tres radios censales con datos anómalos, tratándose de radios de tamaño muy pequeño. Al retirar estos datos, se observa un cambio importante en el valor del índice de Moran, no así en el EBI. Esto también se observa en la recta que ajusta los puntos de ambos diagramas de dispersión, encontrando que la correspondiente al EBI se encuentra menos afectada que la del gráfico de dispersión de Moran.

Por último, se considera de interés la profundización del estudio sistemático de las propiedades del EBI y su comparación con Moran para diferentes anomalías y evaluando otras distribuciones subyacentes, teniendo en cuenta que el trabajo citado de presentación del índice, lo hace utilizando la distribución de Poisson.

REFERENCIAS BIBLIOGRÁFICAS

Anselin, L.; Ibnu, S.; Younggihn, K. (2006). GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis 38 (1), 5-22.*

Assunção, R. M.; Reis, E. A. (1999) *"*A new proposal to adjust Moran´s I for population density". *Statist. Med. 18, 2147-2162.*

Borra, V. (2015) "Estadística Espacial. Muestreo y modelización para la aplicación en estudios socioeconómicos".

Marshall, R. J. (1991) "Mapping disease and mortality rates using empirical Bayes estimators". *Applied Statistics, 40, 283–294.*

Moran, P. A. P. (1950) "Notes on continuous stochastic phenomena". *Biometrika, 37, 17–23.*

Oden, N. (1995) "Adjusting Moran's I for population density". *Statistics in Medicine, 14, 17-26*.

R Core Team (2020) "*R: A language and environment for statistical computing".*

Tiefelsdorf, M.; Griffith, D. A.; Boots B. (1999) "A variance-stabilizing coding scheme for spatial link matrices". *Environment and Planning A 31,165–180.*

Torres, P.S.; Quaglino, M.B.; Pillar, V.D. (2009) "Properties of a randomization test for multifactor comparisons of groups". *Journal of Statistical Computation and Simulation.*

Tobler, W. (1970) "A Computer Movie Simulation Urban Growth in the Detroit Region*". Economic Geography 46(2),234-240.*

Walter, S. D. (1992) "The analysis of regional patterns in health data. I. Distributional considerations". *American Journal of Epidemiology, 136, 730-741.*