**3. Materiales y Métodos**

En este capítulo se introduce al concepto de datos espaciales, se mencionan conceptos importantes a considerar a la hora de trabajar con índices de autocorrelación espacial y se presentan los fundamentos teóricos de los índices de Moran, Oden y el índice Empírico de Bayes, junto con sus aspectos más relevantes. Estos tres índices mencionados serán comparados aplicándolos a los siguientes problemas: el estudio de la distribución espacial de los Hogares con Necesidades Básicas Insatisfechas (NBI) en la ciudad de Rosario en el año 2010 y Heridos de armas de fuego en la ciudad de Rosario durante un determinado año.

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 INDEC. Resulta de interés mencionar, que se hallaron tres registros duplicados y que no se cuenta con información censal para un radio de la ciudad de Rosario, de esta manera los estudios se realizaron utilizando 1069 radios censales de la ciudad de Rosario.

En cuanto al segundo problema a considerar, se cuenta con un conjunto de datos ficticios que corresponden al número de Heridos de armas de fuego en la ciudad de Rosario en un determinado año. Este problema fue tratado debido a que, en el artículo referenciado en la presentación del EBI, se aplica este indicador al mismo problema en la ciudad de Belo Horizonte, Minas Gerais, Brasil.

El software que se utiliza para realizar la presente tesina es R y en particular los paquetes sp y spdep. Se desarrolla un programa, para el cálculo del índice de Oden, ya que no se encontró en ningún paquete de R ni en otro software.

**3.1 Datos Espaciales**

En muchos problemas estadísticos, los datos con los que se debe tratar, se observan unidades que se encuentran situadas en el espacio y su posición 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 de los datos espaciales es la existencia de autocorrelación espacial. Se denomina Estadística Espacial al conjunto de métodos que tienen por objeto la exploración, descripción, visualización, y análisis de datos considerando sus características de distribución en el espacio.

En un análisis de esta clase 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. Se emplean métodos descriptivos convencionales e índices de autocorrelación espacial. El índice 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 predecir 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 índice más usual es el de Moran, qué se disponible en cualquier software de Geoestadística, pero presenta inconvenientes en situaciones particulares y por ese motivo se han propuesto otras alternativas, menos difundidas, que se tratan en la presente tesina.

**3.2 Tipos de datos espaciales**

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

* 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 subáreas 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 variar por este motivo. 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.

Los conjuntos de datos considerados en la presente tesina, se corresponden con datos reticulares irregulares, donde cada una de las áreas o agregaciones espaciales se corresponden con los radios censales de la ciudad de Rosario.

**3.3 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.

Considerar 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, 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 las unidades vecinas. Áreas con valores altos de la variable serán vecinas de áreas con valores bajos. A modo de ejemplo, considerar la competencia entre plantas por la luz, donde zonas de plantas sanas pueden estar rodeadas de otras con plantas menos fuertes.
* Positiva: la variable asumirá valores similares en regiones 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.

Quienes permiten verificar si se cumple la hipótesis de que una variable se encuentra distribuida en forma aleatoria en el espacio o si, por el contrario, existe asociación significativa entre unidades vecinas son los índices de autocorrelación espacial. Para su cálculo se necesita proporcionar criterios de cercanía entre unidades y pesos que reflejen la fuerza de la influencia en la relación entre las mismas.

**3.3.1 Criterios de Vecindad**

Como dice la primera ley de la geografía, o principio de autocorrelación espacial “Todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes” (Tobler, 1970). Pero ¿qué se considera cercano? Para poder responder a esta pregunta 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 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.

Los criterios de vecindad más utilizados y aceptados por la bibliografía:

* 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 alternativas dentro del vecindario determinado por contigüidades, Reina, Torre y Alfil, y 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 al menos un punto 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 más de un punto 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 tan solo un punto en común.

Este criterio cumple la condición de simetría ya que, si un área tiene un punto o más en común con una segunda, esta segunda también tendrá un punto o más 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. De aquí en adelante se refiere a la distancia entre dos áreas como la distancia entre sus centroides. Existen dos variantes:
  + Los k vecinos más cercanos. Se calcula la distancia de una unidad determinada a todas las demás, y serán vecinas las k unidades que posean una distancia menor. 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.

En la presente Tesina, se adoptará 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.

**3.3.2 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 wij representa el peso de la relación de vecindad entre las áreas i y j. Cuando wij = 0 las unidades no son vecinas. La diagonal de la matriz será 0 ya que, por convención, una unidad no puede ser 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 y aquellos en la que la importancia de las uniones variará en base a ciertas características.

* Binario: Es el criterio más sencillo, asume que wij = 1 cuando i y j son unidades vecinas y wij = 0 cuando no lo son. Este método 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 se basa en que los pesos de cada fila de la matriz sumen 1. Para ello se divide la unidad entre el número de áreas vecinas 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, definiendo el peso como el cociente entre la unidad y el número total de vecinos. De esta manera, la suma de todos los pesos será igual a uno.
* Estabilización de variancia: Los pesos de áreas con muchos vecinos varían mucho al utilizar un criterio u otro, de los mencionados anteriormente. Lo que busca este método es reducir esta variación. Puede pensarse como un híbrido del método de estandarización por filas y el de estandarización global. Será siempre asimétrico, pero si el criterio de vecindad es simétrico la matriz W estará bastante cerca de ser simétrica.

En la presente Tesina, se utilizará el criterio de Estabilización de variancia para determinar el peso de cada una de las relaciones.

**3.3.3 El índice de Moran (I)**

Este 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, este índice puede conducir a resultados erróneos.

Se considera una región R dividida en m áreas ri, i=1…, m. En los casos considerados en este estudio, la región es la ciudad de Rosario y las áreas o unidades son los radios censales.

Sea xi el “tamaño” del área i, por ejemplo, el total de hogares en el radio censal i.

Sea ni el valor de la variable de interés en el área i, en el primer problema que se estudia es el número de hogares con NBI en el radio censal i.

La proporción observada en la región i se define como pi=

El índice de Moran para la proporción pi, se define como (Moran, 1950):

I = ∀ *i≠j*, donde

*pi* es el valor de la variable “*P”* en la unidad ria la que se le asocia el conjunto de coordenadas **s***i*,

= es la media de la variable,

*wji* es el elemento i,j de la matriz conocida como “de conectividad” la que recoge la relación de vecindad entre las unidades *i* y *j* definida de la siguiente manera:

=

El índice de asociación global de Moran resume la intensidad y dirección de la dependencia entre los valores de una variable observados en distintas unidades del espacio, calculando los productos cruzados entre los valores de pares de unidades y ponderando por una medida de la relación de vecindad entre las unidades de cada par. Por lo tanto, *I* puede considerarse como una medida de correlación de cada *pi* 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]= 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 *n* crece) indica ausencia de autocorrelación.

Para probar la significación estadística del índice *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* ∼ *N*(*µ, σ*2), o si mes suficientemente grande, la estadística Z = sigue una distribución normal estándar donde:

* E[I] =
* Var[I] = -

Cuando no se cumple el supuesto de normalidad de pi se utiliza un test permutacional: se encuentran las m! 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 se vuelve 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. Los softwares suelen utilizar 999 permutaciones, y considerando la muestra observada, resultan 1000 valores de I para construir la distribución de aleatorización.

**3.3.4 Efectos de unidades de diferentes tamaños**

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

Por ejemplo, si se desea estudiar la autocorrelación espacial para la variable proporción de hogares con necesidades básicas insatisfechas (NBI) observada en los radios censales de la ciudad de Rosario, puede ocurrir que un radio censal con 100 hogares tenga 10 con necesidades básicas insatisfechas, 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 proporciones, 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.

Teniendo en cuenta el supuesto de normalidad cuando las poblaciones son de diferentes tamaños, las áreas con menos población tienen proporciones más variables y, por lo tanto, es más probable que asuman un valor extremo.

Se han estudiado los efectos de las desviaciones del supuesto de normalidad en las pruebas de autocorrelación espacial. La P(e1) en estos casos se incrementa cuando las poblaciones en las áreas son heterogéneas y la proporción es constante (Walter, 1992).

Assunção estudia el efecto de tamaños poblaciones heterogéneos sobre la potencia del test, es decir evalúa el impacto de la variación de los tamaños poblacionales cuando existe una correlación espacial entre las proporciones (Assunção, 1999).

En la presente tesina, se estudiarán y aplicarán distintos índices de autocorrelación espacial, con el fin de dejar en evidencia las características de los mismos y cuando es conveniente utilizar uno u otro, principalmente en un escenario de tamaños poblacionales variables de las distintas áreas consideradas.

**3.3.5 El índice de Oden ()**

Cuando las unidades son de diferente tamaño Oden, (1995) propone un ajuste al índice de Moran. El conocido como Índice de Oden es:

=

Donde:

n = total de la variable en estudio en la región,. En el primer problema que se trata es el total de hogares con NBI en la ciudad de Rosario.

x = es el total en la región de la variable utilizada como denominador, siguiendo con el ejemplo: total de hogares en la ciudad de Rosario.

, proporción de interés, en la región.

, proporción con respecto al total en la región, de la variable en estudio en la unidad i.

proporción con respecto al total en la región de la variable utilizada como denominador en la unidad i.

Llamando al peso espacial definido por Oden de la siguiente manera:

Mij =

Generalmente, por convención, la diagonal principal de la matriz de vecindad es igual a 0. Oden propone esta modificación, de alguna manera intentando diferenciar las situaciones de dos regiones que no son vecinas y cuando se compara a la región consigo misma.

valor incluido en la fórmula del índice de Oden

Oden muestra que el test que utiliza a es más potente que la prueba asociada a I cuando la tamaños de las unidades espaciales consideradas son muy variables. Esta capacidad de capturar la variabilidad se debe al primer término en el numerador, que es una versión espacial de la prueba chi-cuadrado convencional para la heterogeneidad de proporciones.

**3.3.6 Comparación de las pruebas de hipótesis entre Moran y Oden**

A pesar de la mayor potencia obtenida por el índice de Oden con respecto al índice de Moran, Assunção (1999) advierte sobre ciertas particularidades que posee este índice las que se deben considerar en caso de decidir su utilización.

La consideración más importante que realiza Assuncao es lo inapropiado de la comparación entre las pruebas de hipótesis que se realizan con I e presentada por Oden.

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

1. Proporciones espaciales homogéneas o constantes
2. Proporciones heterogéneas sin correlación espacial
3. Proporciones heterogéneas correlacionados espacialmente

La prueba de existencia de correlación espacial que propone Moran considera las siguientes hipótesis nula y alternativa:

H0) A U B H1) C

Mientras que la prueba propuesta por Oden plantea las siguientes hipótesis:

H0) A H1) B U C

Como se muestra en la Figura 1.

**Tabla n°1**: Hipótesis probadas por Moran y Oden.

|  |  |  |
| --- | --- | --- |
| Índices/Hipótesis | H0 | H1 |
| I | A U B | C |
|  | A | B U C |

Puede apreciarse que el estado B, proporciones heterogéneas sin correlación espacial, en el índice de Moran se acepta como parte de la H0 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 a una situación en la que no hay (B).

En consecuencia no sorprende que el test basado en 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.

A medida que se fueron desarrollando las investigaciones correspondientes que desencadenaron en las propiedades mencionadas anteriormente con respecto al índice de Oden, este mismo dejo de utilizarse.

De esta manera, volvió a resurgir 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.

**3.3.7 Empirical Bayes Index (EBI)**

Una nueva alternativa que se considera para probar la autocorrelación espacial, es el Empirical Bayes Index (EBI), en esta sección se presentará el sustento metodológico del mismo.

Sean θ1, θ2, …, θi, ..., θm, donde i=1…, m las proporciones de interés desconocidas de las áreas en estudio. Se realiza el supuesto de que el número de eventos observados ni durante un período de referencia tiene una distribución Poisson con media condicional E(ni|θi) = Var(ni|θi) = xi θi siendo xi el tamaño poblacional del área i. De esta manera, la media condicional de la proporción estimada pi es E(pi|θi) = θi y su variancia condicional es igual a Var(pi|θi) = , por lo tanto las proporciones estimadas poseen distintas medias y variancias condicionales.

Se realiza el supuesto de que las proporciones θi tienen a priori una esperanza y variancia igual a β y α respectivamente. De esta manera, la esperanza marginal de pi es β y su variancia marginal es α + (Assunção, 1999). Dicho de otra manera, las áreas poseen la misma esperanza marginal y sólo las variancias difieren entre las unidades, las cuales se incrementan a medida que lostamaños poblaciones disminuyen.

Para estimar los parámetros α y β desconocidos, Marshall (1991) propone utilizar el método de los Momentos, el cual conduce a los siguientes resultados:

= a = s2 - = b = ,

dónde s2 = , x y n son el tamaño poblacional de la región y el total de la variable de interés en la región respectivamente.

Por lo tanto, la esperanza y variancia marginales de pi son estimadas por b y = a + , respectivamente. Por convención, si < 0, se define = .

En lugar de utilizar las proporciones estimadas pi (como se emplean en el índice de Moran), se propone un nuevo índice que toma las proporciones estandarizadas, utilizando las estimaciones presentadas anteriormente.

=

El índice Empírico de Bayes (EBI) se define de la siguiente manera:

EBI =

Al igual que el índice de Moran, el EBI será significativamente distinto de cero si las proporciones 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 se asignan aleatoriamente a las áreas una determinada cantidad de veces (en general, se utilizan 999 permutaciones al igual que en Moran). Para cada una de las combinaciones obtenidas se calcula el EBI. El valor de la probabilidad asociada al test de hipótesis 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).