**Tesina**

**Indicadores Globales de Autocorrelación Espacial para Unidades de Diferente Tamaño**

Alumno: Ferraro, Sebastián M.

Director: Dr. Pagura, José A.

Codirector: Mignoni César

Carrera: Licenciatura en Estadística

Universidad Nacional de Rosario

Facultad de Ciencias Económicas y Estadística

MES de 2020



**Agradecimientos**

**Resumen**

**Índice**

Agradecimientos…………………………………………………………….……………………………………………………1

Resumen ……………………………………………………………………………………………….……………………………2

Índice ……………………………………………………………………………………………………………………………….…3

1. Introducción…………………………………………………………………………………….………………………………4

2. Objetivos…………………………………………………………………………………………………………………………

3. Materiales ……………………………………………………………………………………………………………………..

3.1 NBI…………………………………………………………………………………………………………………..

3.2 Heridos…………………………………………………………………………………………………………..

4. Metodología…………………………………………………………………………………………………………………..

4.1 Estadística Espacial ………………………………………………………………….

4.2 Criterio de Vecindad…………………………………………………………………………

4.3 Pesos Espaciales……………………………………………………………………………………..

4.4 Autocorrelación Espacial………………………………………………………….

4.5 índice de Moran…………………………………………………………………………………………………

4.6 Efectos de tamaños poblacionales variables………………………………………………………..

4.7 índice de Oden………………………………………………………………………………………………

4.8 Comparación de las pruebas de hipótesis……………………………………………….

4.9 Breve introducción a la Estadística Bayesiana………………………………………………

4.10 Empirical Bayes Index (EBI)……………………………………………………………………………..

5. Aplicación…………………………………………………………………………………………………………………….

5.1 Resultados NBI……………………………………………………………………………………………

5.2 Resultados Heridos………………………………………………………………………………………

6. Conclusiones y Comentarios Finales…………………………………………………………………………..

7. Bibliografía………………………………………………………………………………………………………………..

8. Anexo……………………………………………………………………………………………………………………….

1. **Introducción**

Los índices globales de autocorrelación espacial permiten comprobar 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.

El índice más utilizado es el de Moran, el cualpermite verificar si las unidades se distribuyen aleatoriamente en el espacio teniendo en cuenta la variable bajo estudio y proporciona una medida resumen de la intensidad de la autocorrelación de las unidades.

Podría ocurrir que las unidades espaciales no posean el mismo tamaño, en cuanto a la población en riesgo, dejando en evidencia la necesidad de ponderar por el tamaño poblacional de las áreas consideradas.

Consideremos, por ejemplo, las siguientes dos regiones: p1= = = 0.1 y p2= = = 0.1

Ambas nos arrojan proporciones iguales a 0.1, pero los casos son completamente distintos y el índice de Moran no hace ninguna distinción entre estas dos situaciones, es decir no pondera por el tamaño poblacional de las correspondientes áreas.

De esta manera, surge la necesidad de proponer índices alternativos para las situaciones en las que los tamaños poblacionales de las áreas consideradas sean de diferentes tamaños.

En el presente trabajo de tesina se proponen como alternativa, el índice de Oden y el Empirical Index Bayes (EBI), el primero de ellos parecería funcionar bien en un principio, pero posee otro tipo de desventajas, y se mostrará que el EBI es el que posee las mejores propiedades en general y, principalmente, para este tipo de situaciones.

Estos tres índices serán comparados aplicándolos a los siguientes conjuntos de datos: 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 en el año 2012.

También, se aplicarán sobre un escenario simulado para hogares con NBi, con el fin de poder apreciar situaciones en donde escoger un índice u otro conduce a conclusiones distintas.

1. **Objetivos**
2. Reseña histórica y estudio teórico de los tres índices considerados.
3. Aplicación en ambos conjuntos de datos.
4. Simulación de escenarios que dejen en evidencia cual es el mejor de ellos.
5. Conclusiones de los resultados hallados.
6. **Materiales**

En esta sección se detallarán las características de los conjuntos de datos considerados, junto con el tratamiento que se les realizó previo a su utilización.

También se destaca que, el software que se utiliza para realizar la presente tesina es R Studio. Se crea un programa, específicamente, para calcular el índice de Oden, ya que no se encontró en R ni en ningún otro software. Los paquetes de R que más utilizan son sp y spdep.

* 1. **Hogares con Necesidades Básicas Insatisfechas (NBI)**

Los Hogares con Necesidades Básicas Insatisfechas son aquellos que presentan al menos una de las siguientes condiciones de privación (Indec, 2010):

• Vivienda: Son los hogares que viven en habitaciones de inquilinato, hotel o pensión, viviendas no destinadas a fines habitacionales, viviendas precarias y otro tipo de vivienda. Se excluye a las viviendas tipo casa, departamento y rancho.

• Condiciones sanitarias: Incluye a los hogares que no poseen retrete.

• Hacinamiento: Es la relación entre la cantidad total de miembros del hogar y la cantidad de habitaciones de uso exclusivo del hogar. Operacionalmente se considera que existe hacinamiento crítico cuando en el hogar hay más de tres personas por cuarto.

• Asistencia escolar: Hogares que tienen al menos un niño en edad escolar (6 a 12 años) que no asiste a la escuela.

• Capacidad de subsistencia: Incluye a los hogares que tienen cuatro o más personas por miembro ocupado y tienen un jefe que no ha completado el tercer grado de escolaridad primaria.

El concepto de Necesidades Básicas Insatisfechas (NBI) permite la delimitación de grupos de pobreza estructural y representa una alternativa a la identificación de la pobreza considerada únicamente como insuficiencia de ingresos. Por medio de este abordaje se identifican dimensiones de privación absoluta y se enfoca la pobreza como el resultado de un cúmulo de privaciones materiales esenciales.

Se calcula como el cociente entre los hogares con NBI y sin NBI por radio censal.

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

Para el censo 2010 la ciudad de Rosario se dividió en 87 fracciones y en 1073 radios  
censales y el número de hogares con Necesidades Básicas Insatisfechas se dispone totalizado por radio censal. En esa ocasión el total de hogares con NBI en la ciudad de Rosario resultó igual a 19296.

Es necesario mencionar que se hallaron 3 registros duplicados y que no se dispone de información censal para 1 radio de la ciudad, en consecuencia, los estudios se realizan con los 1069 radios con información.

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

* 1. **Heridos de Arma de Fuego.**

(EN ESTA SECCIÓN, GUIAME BETO PARA “VER QUE HACEMOS” CON LA FUENTE DE LOS DATOS)

**4.1 Estadística Espacial**

Este término hace referencia a un conjunto de técnicas apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios de una región, siendo la ubicación espacial relevante para el problema.

Tiene por objeto la exploración, descripción, visualización y análisis de datos considerando sus características de distribución en el espacio.

Al considerar la localización de los datos, estos últimos adquieren una naturaleza georreferenciada, y esta es la característica principal de los datos espaciales.

En contraposición a la mayoría de los datos estadísticos, los espaciales no suelen cumplir el principal supuesto de la teoría estadística, la independencia. Dos observaciones de una variable son independientes cuando el valor que toma una de ellas no influye en la otra. Este principio no siempre se cumple en la situación espacial, ya que cuanto más próximas estén dos observaciones parece lógico pensar que van a estar más relacionadas. A esta falta de independencia se la denomina autocorrelación espacial, y conocer si aparece o no en el conjunto de datos será parte fundamental del análisis espacial.

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

* Geoestadísticos (o espacialmente continuos): Si los datos se pueden observar en cualquier posición. Cualquiera sea la localización del espacio en estudio puede seleccionarse para medir en ella la variable de interés.
* Reticulares o Lattice (látices): En esta situación cada observación se suele corresponder con agregaciones espaciales, es decir se observa una variable aleatoria sobre cada una de diferentes regiones en las que se divide el territorio que se estudia. Estas regiones 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 subregiones idénticas, las segundas presentan distintas formas y tamaños. La definición de las regiones 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 regiones.
* Puntuales: Un evento ocurre 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.

**4.2 Criterio 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 lattice se considera que no todas las regiones influyen sobre el valor que asume la variable en una determinada región, sino, que solo influirán aquellas que sean vecinas.

Definir qué características deben poseer dos regiones 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 al conjunto de datos y a la naturaleza del problema.

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

Una característica importante que tienen los criterios de vecindad es la simetría. Sea A una determinada región y según el criterio que se está utilizando la región 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 regiones vecinas.

La naturaleza de los conjuntos de datos que se utilizarán (radios censales en la ciudad de Rosario, ya sea para NBI o Heridos de Arma de Fuego) es de la clase “Spatyal Polygons Dataframe”, es decir cada unidad posee la suficiente información espacial, vector de latitud y longitud, como para poder ser representada por un polígono irregular en el espacio. Muchos criterios de vecindad tan sólo requieren de un punto en el espacio que represente a cada una de las regiones, generalmente se utiliza el centroide de las mismas. De esta manera, si se utilizasen criterios de vecindad que requieran tan solo como input el centroide de cada observación, se estaría adaptando el conjunto de datos a la clase “Spatyal Points Dataframe”.

A continuación, se presentan 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, Queen (Reina), Rook (Torre) y Bishop (Alfil), reciben estos nombres porque se asemejan a los movimientos de las fichas en el tablero de ajedrez.
  + Queen: la 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 regiones serán vecinas si tienen al menos un punto común.
  + Rook: la 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.
  + Bishop: el Alfil solo puede moverse a lo largo de la diagonal de la casilla en la que se encuentre. De esta manera, dos regiones 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 euclidea. 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 regiones como la distancia entre sus centroides. Existen dos variantes:
  + Los k vecinos más cercanos. Se calcula la distancia de una región determinada a todas las demás, y serán vecinas las k áreas 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 sea menor a una magnitud fijada a priori. Este método funciona bien 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 región sin vecinos, o considerar un número de vecinos demasiado alto en el resto de áreas. Esta relación será simétrica.

En la presente Tesina, siempre se adoptará el criterio de vecindad por contigüidad, y más específicamente el de tipo Reina (Queen), ya que parecería ajustarse de una manera más adecuada a la naturaleza de nuestro problema y para aprovechar al máximo la información contenida en nuestro conjunto de datos.

**4.3 Pesos Espaciales**

Una vez definido el criterio de vecindad a utilizar, resulta de interés cuantificar la fuerza de cada unión, esto es lo que se conoce como pesos espaciales. Hasta ahora se sabe que A tiene dos regiones 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.

Al igual que resulta indispensable la decisión del criterio de vecindad a utilizar, también lo es definir los pesos espaciales, ya que adoptar un criterio u otro puede modificar los resultados y conclusiones finales.

Los pesos se representan de forma matricial mediante la matriz cuadrada W. Donde cada elemento wij representa el peso de la relación de vecindad entre las regiones i y j. Cuando wij = 0 las regiones no son vecinas. La diagonal de la matriz será 0 ya que, por convenio, una región no puede ser vecina de ella misma (aunque cuando se presente el índice propuesto por Oden se verá una alternativa distinta a esta). 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 una región 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 estilos 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.

* Estilo binario (B). Es el criterio más sencillo, asume que wij = 1 cuando i y j son regiones vecinas y wij = 0 cuando no lo son. Este método es el más utilizado cuando existe poca información del proceso espacial. Bajo este estilo la suma de los pesos de un área es el número de vecinos que tiene.
* Estandarización por filas (W). 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 la región. 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 menor influencia.
* Estandarización Global (C). Considera el mismo peso para todos los enlaces, definiendo el peso entre dos regiones vecinas como el cociente entre el número total de regiones y el número de enlaces. La suma de todos los pesos será el número de regiones.
* Otra forma de estandarización global es el estilo U, donde se define el peso como el cociente entre la unidad y el número de enlaces total. Con este estilo la suma de todos los pesos será igual a uno.
* Estabilización de varianza (S). Los pesos de áreas con muchos vecinos varían mucho de utilizar un estilo a otro. Lo que busca este método es reducir esta variación. Los pesos variarán menos que con el estilo W. Será siempre asimétrico, pero, al igual que ocurre con el estilo W, si el conjunto de vecinos es simétrico la matriz W estará bastante cerca de ser simétrica (Tiefelsdorf, 1999).

En la presente Tesina, se utilizará el estilo Estabilización de variancias (S) para determinar el peso de cada una de las vecindades, ya que se considera el más apropiado para nuestro conjunto de datos.

* 1. **Autocorrelación Espacial**

En general, en la estadística, se asume que las observaciones de una variable se toman bajo condiciones idénticas y de manera independiente. Los datos forman una muestra aleatoria simple, es decir, los datos 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 tiempo o en el espacio, están más correlacionados es natural.

En el contexto espacial, esta falta de independencia recibe el nombre de dependencia o autocorrelación espacial, la cual se define como 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 un punto o región determinada dependa de los valores observados en puntos o regiones vecinas.

La autocorrelación espacial puede ser:

* Negativa: la variable tomará valores diferentes en regiones vecinas. Áreas con valores altos de la variable estarán rodeadas de regiones 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 región se “contagia” a áreas vecinas. Una región con un valor alto de la variable estará rodeada de regiones 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. Es decir, permiten probar la hipótesis de aleatoriedad espacial.

En las siguientes secciones se estudiarán los índices de Moran, Oden y el Empirical Bayes Index, con el fin compararlos posteriormente y determinar cuál de ellos es conveniente utilizar en distintas situaciones.

* 1. **Índice de Moran (I)**

Considere una región S dividida en m áreas o regiones y sean ni y xi el número de casos y la población en riesgo en el área i respectivamente. La proporción observada en la región i se define como pi=

El índice de Moran se define como (Moran, 1950):

I = ∀ *i≠j*, donde, *pi* es el valor de la variable “*P* ” en la unidad *i* ubicada en el punto de coordenadas **s***i*, = es la media de la variable, *wji* es el elemento de la matriz de conectividad 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 regiones del espacio, calculando los productos cruzados entre los valores de pares de unidades y ponderando por una medida de la relación de vecindad entre las unidades de cada par. De esta manera, *I* puede considerarse como una medida de correlación de cada *pi* con el resto de las regiones con las que se encuentra vinculada. Al igual que el índice de correlación de Pearson varía entre -1 y 1, y E[I]= bajo la hipótesis de aleatoriedad espacial.

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

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

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 la variable de interés principal se recurre a los tests permutacionales, en los que se encuentran las m! posibles configuraciones de las unidades espaciales asumiendo que sus valores son aleatorios y sobre cada una de ellas se calcula el valor de *I*, para luego calcular la probabilidad asociada a la hipótesis de aleatoriedad.

Comúnmente se puede utilizar un test basado en el Método de Montecarlo, que consiste en la realización de un test permutacional, pero sólo considerando un subconjunto de configuraciones, y por lo tanto es útil cuando mes muy grande. Este tipo de técnicas obliga la utilización de métodos computacionales ya que para su aplicación intervienen gran cantidad de cálculos. Los softwares suelen utilizar 999 permutaciones, y considerando la muestra observada, resultan 1000 ensayos.

* 1. **Efectos de tamaños poblacionales variables**

Si se considera una región dividida en áreas con poblaciones de diferentes tamaños, al estimar la proporción observada en cada área se estarían utilizando medidas con diferentes variaciones en el cálculo de los índices.

Sea el siguiente caso a modo de ejemplo: p1= = = 0.1 y p2= = = 0.1

En ambos casos se obtiene una razón igual a 0.1 pero las situaciones son completamente diferentes, y el índice de Moran no hace ninguna distinción entre ambos casos, por lo cual se aprecia una potencial falencia del mismo, ya que no se pondera por los tamaños poblacionales de las distintas áreas consideradas.

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 de ambos supuestos 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).

Renato Asuunção 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.

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

* 1. **Índice ajustado por Oden () PODRÍA PROFUNDIZARSE (CÁLCULO)**

Se propone un ajuste al índice de Moran cuando existen tamaños poblacionales distintos (Oden, 1995):

=

Donde n = , x = , , , y

es el peso espacial para el par de observaciones ubicadas en las áreas i y j respectivamente definido por Oden de la siguiente manera:

Mij =

Nota: 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 unidad consigo misma”.

A través de simulaciones, Oden muestra que es mucho más potente que I cuando la variación está presente en los tamaños de las poblaciones consideradas. Él atribuye la superioridad de su índice a su interpretación de la variabilidad de las proporciones entre regiones como evidencia de agregación, incluso en ausencia de correlación espacial. 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.

Se podría agregar la E(I\*pop) y V(I\*pop) conjuntamente con los términos definidos para facilitar su cálculo a la hora de programarlo (paper)

* 1. **Comparación de las pruebas de hipótesis**

A pesar de la gran potencia obtenida por el índice de Oden, se verá con detalle que es lo que realmente se está probando con esta estadística y, por lo tanto, si tiene sentido la comparación con el índice de Moran.

Se considera inapropiada la comparación entre I y presentada por Oden porque las estadísticas no están probando el mismo par de hipótesis nula y alternativa. De esta manera, 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

I: H0) A U B H1) C

: H0) A H1) B U C

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

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

Por lo tanto, queda en evidencia que las pruebas no concuerdan con respecto al estado de B.

Entonces no es sorprendente que tenga mayor potencia, especialmente en estados como B, frente a los cuales Moran debería tener como máxima potencia α, la probabilidad de error de tipo I.

**4.9 Breve introducción a la Estadística Bayesiana.**

El teorema de Bayes fue desarrollado por Thomas Bayes en 1763 y con él se expresa la probabilidad condicional de un evento aleatorio A dado otro evento B, mediante la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal de sólo A.

En otras palabras, sea {𝐴1, 𝐴2, …, 𝐴𝑖, …, 𝐴𝑛} un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales 𝑃(𝐵|𝐴𝑖). Entonces, la probabilidad 𝑃(𝐴𝑖 |𝐵) viene dada por la expresión:

𝑃(𝐴𝑖 |𝐵) =

donde:

* 𝑃(𝐴𝑖 ) son las probabilidades a priori.
* 𝑃(𝐵|𝐴𝑖 ) es la probabilidad de 𝐵 en la hipótesis 𝐴𝑖 .
* 𝑃(𝐴𝑖 |𝐵) son las probabilidades a posteriori.

Por tanto, el teorema de Bayes hace uso de probabilidades a priori, que son probabilidades subjetivas, que se desarrollan a continuación, probabilidad de B en la hipótesis 𝐴𝑖, verosimilitud propia de la muestra, y una distribución a posteriori, que se alcanza mediante el producto de las dos anteriores ponderadas según la verosimilitud propia de la muestra.

Además, cabe destacar que, cuando 𝐴1, 𝐴2, …, 𝐴𝑘 son k sucesos mutuamente excluyentes, uno de los cuales ha de ocurrir necesariamente; entonces, la ley de la probabilidad total establece que:

𝑃(𝐵) =

En el caso continuo, sería:

𝑃(𝐵) =

Donde 𝑓(𝑥) es la función de densidad de una variable aleatoria X evaluada en x, 𝑃(𝐵|𝑥) es la probabilidad de B suponiendo que X=x y Ω es el posible espectro de valores continuos que puede tomar X. Dando lugar a la siguiente modificación de la regla o formula de Bayes:

𝑃(𝐴𝑖 |𝐵) = ó 𝑃(𝐴𝑖 |𝐵) =

**Probabilidad subjetiva**

La estadística Bayesiana se basa en la interpretación subjetiva de la probabilidad. Para ello utiliza la percepción existente, por parte del investigador, como una variable modificadora (distribución a priori) de los datos muestrales, que dan lugar a una distribución (distribución a posteriori) con la que formular inferencias con respecto al parámetro de interés.

El hecho de amoldar los datos muestrales obtenidos en función del criterio del investigador convierte a la estadística Bayesiana en un instrumento altamente controvertido, dado que esto puede interpretarse, como que la estadística bayesiana manipula los datos muestrales con el fin de demostrar lo que uno quiere en lugar de dejar que los datos, por sí solos, demuestren o no el objeto de estudio.

Sin embargo, la aportación subjetiva del investigador no tiene que ser de por sí negativa o ser considerada manipuladora (en su sentido más peyorativo), ya que esta aportación subjetiva que realiza el investigador puede darse a causa de conocimientos previos adquiridos a través de otros estudios anteriores o por la intuición del profesional, que a diario observa la situación objeto de estudio.

Por ello la probabilidad subjetiva no debe ser interpretada, de por sí, como un instrumento inválido que únicamente pretende manipular el método científico, dado que esta puede aportar beneficios al propio método, además de poder ser contrastado a posteriori, con tal de dar validez a la probabilidad subjetiva utilizada en el proceso.

A su vez, una problemática existente en numerosas ocasiones es que las muestras son muy pequeñas, con lo que no se cumple los requisitos exigibles por el Teorema Central del Límite, que nos indica que si 𝑛 es suficientemente grande, la variable aleatoria = /𝑛 tiene aproximadamente una distribución normal con = 𝜇 y = /𝑛. Al mismo tiempo, no siempre se puede conocer la distribución que sigue la muestra y los experimentos no se pueden repetir. Requisitos que no son exigibles para la estadística Bayesiana, con lo que puede ser una herramienta de gran utilidad, si no única, en ciertas condiciones.

Dicho lo anterior, se puede comenzar a discernir el concepto de distribución a priori, ésta se puede comprender como una distribución que modela los datos muestrales en función de los conocimientos previos existentes, como por ejemplo estudios realizados anteriormente sobre la materia de interés o simplemente, por la intención de aportar cierta información que el investigador considera oportuna.

Por ello, si considera una muestra, que puede ser o no aleatoria 𝑿 = (𝑋1, … , 𝑋𝑛) con densidad discreta o continua en la familia 𝑓(𝑥, 𝜽), con 𝜽 = (𝜃1, … , 𝜃𝑘 )𝜖 Θ ⊂ . Suponiendo que se tiene información previa sobre 𝜽. Esta información está expresada por medio de una distribución sobre 𝜽 y es esta distribución la que denominamos distribución a priori.

En conclusión, la distribución a priori lo que pretende es aportar información adicional a los datos extraídos de la muestra, de tal forma que complementen la información obtenida de ellos.

**Distribución a Posteriori**

Por otro lado, la distribución a posteriori 𝑝(𝜽|𝑥) es, por la ley multiplicativa de la probabilidad, el producto de la función de distribución de probabilidad 𝑝(𝜽) y la función de verosimilitud 𝑝(𝑥|𝜽).

Dicho de otro modo, la probabilidad a posteriori es aquella que resulta de aplicarle conjuntamente la probabilidad a priori (probabilidad subjetiva) y la verosimilitud de los datos (transformación de los datos experimentales en función de la probabilidad subjetiva), entre la probabilidad de los propios datos experimentales.

**Intervalo de Probabilidad**

En la estadística bayesiana, se conoce por intervalo de probabilidad a algo similar a lo que se conocería como intervalo de confianza en la estadística Frecuentista.

Del lado Frecuentista, el intervalo de confianza hace referencia a probabilidad de que el estimador calculado se encuentre dentro de dos niveles considerados de confianza, donde la confianza dada a este intervalo suele ser, por ejemplo, del 95%. Dicho de otro modo, si repitiéramos el experimento en multitud de ocasiones, el estimador calculado se encontraría dentro del intervalo en el 95% de las ocasiones, mientras que el 5% de las ocasiones estaríamos estimando erróneamente.

El enfoque bayesiano por el contrario es algo distinto, ya que, el método utilizado para su cálculo sería mediante la curva de la función de densidad que se obtiene a posteriori, donde el área bajo dicha curva y entre unos ciertos valores X e Y con cierta probabilidad (por ejemplo, del 95%) constituyen el intervalo de probabilidad del 95%, entre los mencionados puntos (X, Y).

Fuente: <https://core.ac.uk/download/pdf/78528221.pdf>

**4.10 Empirical Bayes Index (EBI)**

En esta sección se presentará el sustento metodológico del índice empírico de Bayes.

Sean θ1, θ2, ..., θm las proporciones subyacentes desconocidas y posiblemente diferentes de las regiones. 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. La proporción estimada pi posee una media condicional igual a E(pi|θi) = θi y variancia condicional igual a Var(pi|θi) = , de esta manera las proporciones estimadas poseen distintas medias y variancias condicionales.

Adoptando un enfoque de mezcla, se supone que las proporciones subyacentes θi tienen a priori una esperanza y variancia igual a β y α respectivamente. Entonces, la esperanza marginal de pi es β y la variancia marginal es α + . Sólo las variancias difieren entre las áreas y se incrementan a medida que los tamaños poblacionales disminuyen.

Se proponen estimadores de Momentos para los parámetros α y β, los cuales son (Marshall, 1991):

= a = s2 - = b = ,

dónde s2 =

De esta manera, la esperanza y variancia marginal son estimadas por b y = a + , respectivamente. Por convención, si < 0, se define = .

En lugar de utilizar las proporciones pi (Moran), se propone un nuevo índice que utiliza una desviación de la media marginal estimada estandarizada por una estimación de su desviación estándar.

=

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

EBI =

Al igual que el Índice de Moran, EBI tenderá a ser positivo si las proporciones están correlacionadas espacialmente. La prueba de independencia espacial frente a la hipótesis nula H0: A ∪ B depende de la distribución nula de EBI, que se puede obtener por permutación.

Por lo tanto, se permuta independientemente el vector alrededor de las regiones una determinada cantidad de veces, para cada una de las combinaciones obtenidas se calcula el valor de EBI. El valor de la probabilidad asociada al test de hipótesis está dado por la proporción de veces que el EBI permutado excede el EBI observado.

Este índice posee interesantes propiedades que, luego de aplicarlo sobre los conjuntos de datos correspondientes, se detallarán.

1. **Aplicación**

En esta sección se mostrarán los resultados obtenidos, luego de aplicar los distintos índices presentados sobre los conjuntos de datos considerados.

Vale recordar, que se utilizó un criterio de vecindad de tipo Reina (contigüidad) y un estilo de Estabilización de variancias (S) para la obtención de los pesos espaciales.

* 1. **Resultados en Hogares con NBI**

Se calculan los tres índices considerados en Hogares con NBI, donde la proporción que interviene en los cálculos está compuesta por:

* Numerador: Cantidad de Hogares con NBI en cada radio censal de la ciudad de Rosario
* Denominador: Cantidad de Hogares en cada radio censal de la ciudad de Rosario.

**Tabla n° 2**: Comparación de los resultados obtenidos por los distintos índices considerados en Hogares con NBI.

|  |  |  |
| --- | --- | --- |
| Índice | Estadístico | P-Valor |
| Moran (I) | 0.38527 | 0.001 |
| Oden () | 0.08890 | <0.001 |
| EBI | 0.41943 | 0.001 |

Los valores observados para el Índice de Moran resultan iguales ya sea que se utilice el supuesto de normalidad o un test permutacional, además, vale mencionar, que el estimado se obtiene a través del supuesto de normalidad.

Nota: Se utilizaron 999 simulaciones tanto para la obtención del Índice de Moran cómo para el EBI.

A pesar de incluir el índice de Oden en la tabla y notar que es el que arroja la menor probabilidad asociada, hay que recordar que su comparación con el I y con el EBI no corresponde ya que su par de hipótesis nula y alternativa no concuerdan.

Se aprecian que ambos índices arrojan resultados significativos, aunque el EBI muestra una mayor autocorrelación de los hogares con NBI, la cual es aproximadamente igual a 0.42 mientras que el estadístico de Moran es cercano a 0.39.

La pequeña magnitud de la estadística observada 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 de ausencia de autocorrelación espacial se debe a los diferentes tamaños de las regiones consideradas o si, realmente, existe autocorrelación espacial.

* 1. **Resultados en Heridos de Armas de Fuego**

De manera similar a la sección 5.1, se aplicarán los distintos índices sobre el conjunto de datos compuesto por los Heridos de Arma de Fuego en la ciudad de Rosario.

La tasa de heridos por radio censal se obtiene de la siguiente manera:

* Numerador: Cantidad de Heridos de Arma de Fuego en cada radio censal de la ciudad de Rosario
* Denominador: Cantidad de Hogares en cada radio censal de la ciudad de Rosario. Siendo esta una aproximación de la población en riesgo, resulta totalmente lógico pensar que en aquellos radios de la ciudad donde existen más hogares, existirán más personas en riesgo.

**Tabla n° 3**: Comparación de los resultados obtenidos por los distintos índices considerados en los Heridos de Arma de Fuego en Rosario.

|  |  |  |
| --- | --- | --- |
| Índice | Estadístico | P-Valor |
| Moran (I) | 0.21796 | 0.001 |
| Oden () | 0.00549 | <0.001 |
| EBI | 0.24375 | 0.001 |

La estadística observada correspondiente al índice de Moran es aproximadamente igual a 0.22, poco inferior al EBI, cuya estadística resulta cercana a 0.24, ambos significativos.

Nuevamente, el índice de Oden arroja la menor probabilidad asociada, a pesar de poseer una menor magnitud en cuanto a la estadística observada, el motivo por el cual ocurre esto, es el mismo que se explicó anteriormente.

Los resultados de los tres índices muestran una tendencia general similar a la aplicación en hogares con NBI.

1. **Conclusiones y Comentarios Finales**

El EBI y Moran poseen una potencia similar ante escenarios de tamaños poblacionales parecidos de las distintas áreas consideradas, pero a medida que nos alejamos de esta situación, el EBI incrementa su potencia de manera considerable con respecto al índice de Moran (Assunção, 1999).

También el EBI posee interesantes cualidades de robustez, ya que no se ve afectado por valores atípicos, a diferencia de Moran (Assunção, 1999).

La P(e1) del EBI no se ve alterada cuando las poblaciones en las áreas son heterogéneas (Assunção, 1999).

1. **Bibliografía**

* Assunção, R.M.; Reis, E. A. (1999). *“*A new proposal to adjust Moran´s I for population density”. *Statist. Med. 18, 2147-2162.*
* Oden, N. (1995). “Adjusting Moran's I for population density”. *Statistics in Medicine, 14, 17-26.*
* Moran, P. A. P. ‘Notes on continuous stochastic phenomena’, Biometrika, **37**, 17–23 (1950).
* Walter, S.D. (1992). “The analysis of regional patterns in health data. I. Distributional considerations”. *American Journal of Epidemiology, 136, 730-741.*
* Marshall, R. J. ‘Mapping disease and mortality rates using empirical Bayes estimators’, Applied Statistics, **40**, 283–294 (1991).
* Tobler W (1970). A Computer Movie Simulation Urban Growth in the Detroit Region. Economic Geography 46(2):234-240.
* Tiefelsdorf M, Griffith DA, Boots B (1999). A variance-stabilizing coding scheme for spatial link matrices. Environment and Planning A 31:165–180.
* Borra V. (2015). Estadística Espacial. Muestreo y modelización para la aplicación en estudios socioeconómicos.

1. **Anexo**