**4. Metodología**

**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 es 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 de datos reticulares 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.

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 relación, esto es lo que se conoce como pesos espaciales. Por el momento 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 estadística, se asume que las observaciones de una variable se toman bajo condiciones idénticas y de manera independiente. 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, 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, en cierta manera, de los valores observados en puntos o regiones vecinas.

La autocorrelación espacial puede ser:

* Negativa: Se presenta una relación inversa entre las regiones vecinas. Áreas con valores altos de la variable serán vecinas 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)**

Se considera una región R dividida en m áreas ri, i=1…, m.

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

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, por ejemplo, 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 de la matriz conocida como “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 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 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.

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 en estudio se utiliza un test permutacional, donde 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.

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.

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 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 de este supuesto 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 (Asuunção, 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 regiones consideradas.

* 1. **Índice ajustado por Oden ()**

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

=

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 =

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.

En sus simulaciones, 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.

* 1. **Comparación de las pruebas de hipótesis entre Moran y Oden**

A pesar de la gran potencia obtenida por el índice de Oden, años más tarde Assunção (1999) advierte sobre ciertas características que posee este índice y cuestiones a considerar si se decide utilizarlo, casi opacando la mejoría en potencia obtenida por Oden.

Para fundamentar porque se considera inapropiada la comparación entre I y presentada por Oden, 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, la tabla permite apreciar que las pruebas no concuerdan con respecto al estado de B, en otras palabras, las estadísticas no están probando el mismo par de hipótesis nula y alternativa.

Entonces no es sorprendente que 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.

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

En esta sección se mencionarán ciertos conceptos referidos a la estadística bayesiana que resultan importantes para el estudio del último índice considerado, el índice empírico de Bayes.

El teorema de Bayes fue desarrollado por Thomas Bayes en 1763, donde se expresa la probabilidad condicional de un suceso aleatorio A dado un evento B, mediante la distribución de probabilidad condicional del evento B dado A y la distribución de probabilidad marginal sólo de 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 (Bayes, 1763):

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.

Además, cabe destacar que, cuando 𝐴1, 𝐴2, …, 𝐴n son n 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 la variable aleatoria X evaluada en 𝑥, 𝑃(𝐵|𝑥) es la probabilidad de B suponiendo que X=𝑥 y Ω es el espacio paramétrico de X. Dando lugar a la siguiente modificación a la fórmula de Bayes:

𝑃(𝐴𝑖|𝐵) = ó 𝑃(𝐴𝑖|𝐵) =

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

La intervención directa del criterio del investigador en los datos muestrales convierte a la estadística Bayesiana en un instrumento muchas veces considerado controvertido, dado que esto puede interpretarse, como que la 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 considerada negativa, ya que esta aportación puede darse a causa de conocimientos previos adquiridos por medio de otros estudios anteriores o por la intuición experta del profesional, que a diario observa la situación que interesa estudiar.

La distribución de los datos en función de la distribución a priori es conocida como verosimilitud de los datos.

Por otro lado, la probabilidad a posteriori es aquella que resulta de aplicar conjuntamente la probabilidad a priori y la verosimilitud de los datos .

Un problema que se presenta en muchas ocasiones es que las muestras son pequeñas, con lo que no se cumple los requisitos necesarios para utilizar el Teorema Central del Límite, esto no es necesario en la estadística Bayesiana, con lo que puede ser una herramienta de gran utilidad, si no única, en ciertas condiciones.

**4.10 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, ..., θm las proporciones de interés desconocidas de las áreas bajo estudio. Se realiza el supuesto de que el número de eventos observados ni durante un período de referencia tiene un comportamiento Poisson (distribución a priori) 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.

Se supone que las proporciones 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 las poblaciones disminuyen.

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

= 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 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, 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 mediante permutaciones.

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