FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Trabajo Fin de Grado

Modelos de Regresión para Datos Espaciales

José Ángel Borrego Sánchez

Dirigido por:
Juan Manuel Muñoz Pichardo

2018

Índice general

Resumen

En este trabajo nuestro objeto de estudio serán los datos espaciales, de los cuales nos centraremos con detalle en los datos de tipo área. Nuestro objetivo final será proponer diferentes tipos de modelos para los datos de área y aplicarlos a un conjunto de datos real.

Para ello necesitaremos modelar el espacio donde se ubican los datos y modificarlo para conseguir un modelado discreto en el cual podamos distinguir diferentes áreas. A continuación, procederemos a realizar un análisis exploratorio de los datos que consiste en crear una matriz de datos con las diferentes propiedades de los datos y su ubicación en el área de estudio. A partir de esta matriz, y con las técnicas que se expondrán en el capítulo 2, podremos analizar si existe dependencia espacial en los datos, es decir, si los valores de la variable bajo estudio en un punto muestral están relacionados con los valores en puntos muestrales cercanos. Para ver este análisis mostraremos medidas de dependencia espacial y contraste de hipótesis con algunos estadísticos como el índice I de Moran o el c de Geary (capítulo 3). Para finalizar el análisis teórico de los datos, detallaremos diferentes modelos de regresión espacial y la estimación de sus parámetros (capítulo 4).

Por último, el capítulo final está reservado para analizar los datos de área con el programa estadístico R. En él veremos los diferentes paquetes, librerías y funciones que necesitaremos para realizar este análisis y expondremos un ejemplo práctico con datos reales para ver el uso de algunas funciones definidas en este conjunto de datos.

Abstract

In this work our study objetive will be the spatial datas in which we will be centered with detail in the type area data. Our final aim will be to propose differents types of models for area datas and apply them to a real data set.

For this will be need to model the space where datas are located acd modify it to get a discrete modelling in which we can distinguish differents areas. Bellow we will proceed to make an exploratory analysis about datas that consist on creating a datas matrix with the differents datas properties and their location in the study area. From this matrix on, and with techniques that we will see in chapter 2, we could analyze if there is a spatial dependence in the datas, that is, if the variable values under study in a sample point are related to the values in nearby sampling points. To see this analysis we will show measures of spatial dependence and contrast tests with some statistics like Moran's I or Geary's c (chapter 3). To finish the datas theoric analysis, we will detail differents spatial regression models and the estimation of its parameters (chapter 4).

Finally, the last chapter is reserved to analyze the area datas with the statistic programme R. In this programme we will see the different packages, libraries and functions that we are going to be needed to make this analysis and we will expose a practical example with real datas to see the use of some functions which are defined in this datas set.

Capítulo 1

Introducción

En este capítulo comenzaremos introduciendo una definición de datos espaciales proporcionado por Fischer y Wang [5], en el cual nos apoyaremos bastante para la redacción de este trabajo. Dentro de este capítulo veremos la importancia de la ubicación geográfica por lo que necesitaremos modelar nuestra región de estudio y tratarla como un espacio continuo o discreto. Debido a este modelado del espacio podremos distinguir varios tipos de datos que veremos en la sección 1.2. Para finalizar este capítulo de introducción trataremos de alcanzar una fiabilidad y calidad de los datos y veremos algunos de los problemas más comunes que pueden surgir en el análisis a partir de unos datos defectuosos o de baja calidad.

Según M. Fisher y J. Wang, los datos consisten en números o símbolos que, en cierto sentido, son neutrales y, en contraste con la información, casi libres de contexto. Los datos geográficos brutos, como la temperatura en un momento y lugar específicos, son ejemplos de datos. En Longley et al. (2001, p. 64) podemos ver los datos espaciales como construidos a partir de elementos o hechos sobre el mundo geográfico. Un dato espacial vincula una ubicación geográfica (espacio geográfico), a menudo un atributo de tiempo, y un atributo (o propiedad descriptiva) de la entidad entre sí. Estos atributos pueden ser de naturaleza ambiental (por ejemplo; la temperatura), económicos (capital de una región concreta), otros pueden identificar una ubicación (direcciones postales). En particular, en el análisis de datos espaciales el tiempo es opcional, sin embargo la ubicación geográfica es esencial ya que es la que distingue este análisis de otros con datos no espaciales. Para llevar a cabo el análisis de datos espaciales requerimos, al menos, la información de la ubicación y de los atributos, independientemente de cómo se midan los atributos. Este tipo de análisis requiere un marco espacial sobre el cual ubicaremos los fenómenos espaciales que vamos a estudiar.

1.1. Modelado del Espacio Geográfico

Existen dos maneras de modelar y representar información geográfica: una visión discreta y una continua de los fenómenos espaciales. En otras palabras, se hace una distinción entre una concepción del espacio que estudiamos como algo lleno de "objetos discretos" y una visión del espacio como cubierto esencialmente de "superficies continuas". A continuación se recoge una clasificación y discusión de modelados según Vitturini y otros [14]:

1. **Modelado Basado en Entidades:** También denominado modelos de objetos, concibe objetos geográficos "incrustados" en el espacio. En un objeto geográfico se pueden distinguir dos componentes: (1) una descripción y (2) una componente espacial, que corresponde con la forma y ubicación del objeto en el espacio. Esta vista de la información geográfica reúne dentro de un objeto espacial puntos del espacio que comparten propiedades similares, esto es, tienen la misma descripción. Para poder distinguir a unos objetos de otros, a cada objeto se le asigna una identificación. El conjunto entidad completo (identificación, objeto espacial y descripción común) constituye un objeto geográfico. En este modelo los tipos de fenómenos espaciales que se analizan se identifican por su dimensionalidad:

   a) *Punto:* Son objetos cero-dimensionales que ubica entidades cuya superficie es muy pequeña en relación con la del espacio. Ejemplos de puntos son edificios, personas, epicentros de terremotos...

   b) *Línea:* Son objetos unidimensionales que se usan para representar entidades del espacio con forma de redes. Por ejemplo ríos, caminos...

   c) *Superficie:* Son objetos bidimensionales que representan entidades con área. Los polígonos son el principal tipo geométrico para tales objetos. Ejemplos de ello son secciones o regiones.

   d) *Volumen:* Son objetos que tienen longitud, anchura y profundidad, y por tanto son tridimensionales. Se utilizan para representar objetos naturales como cuencas fluviales o fenómenos artificiales, como el potencial de población de los centros comerciales. Este último tipo de fenómeno espacial no lo estudiaremos en este trabajo.

   Por supuesto, cuán apropiado son estos objetos depende de la escala espacial (nivel de estudio, es decir, el nivel de detalle con el que tratamos de representar la "realidad") de estudio. Si estamos mirando la distribución de los asentamientos urbanos a escala nacional, es razonable tratarlos como una distribución de puntos. Los fenómenos como las carreteras se pueden tratar

como líneas pero en los mapas a gran escala de las zonas urbanas, las carreteras tienen un ancho, y esto puede ser importante cuando se trata de temas de navegación de automóviles, por ejemplo. Las líneas también marcan los límites de las áreas que es una característica muy tenida en cuenta a la hora de estudiar datos de área, pues una mala elección de los límites impide la realización de un buen estudio.

2. **Modelado Basado en Campo:** En la aproximación basada en campo cada punto del espacio está asociada con distintos valores de atributos, definidos como una función continua sobre dos coordenadas x e y. Las mediciones sobre los distintos fenómenos se reúnen como valores de atributos variando con la ubicación en el plano. La vista del espacio como un campo continuo es lo que contrasta con el modelo basado en entidades, que identifica como entidad u objeto a un conjunto de puntos. La temperatura, por ejemplo, se muestrea en un conjunto de sitios y se representa como una colección de líneas (las llamadas isotermas). Las características del suelo también se pueden muestrear en un conjunto de ubicaciones discretas y representarse como un campo que varía continuamente. En todos estos casos, se intenta representar la continuidad del espacio a partir de un muestreo discreto de varios puntos de dicho espacio.

1.2. Tipos de Datos Espaciales

Al describir la naturaleza de los datos espaciales, es importante distinguir entre el carácter discreto o continuo del espacio en el que se miden las variables y el carácter discreto o continuo de las variables o mediciones. Si el espacio es continuo (modelo basado en campo), los valores de las variables se deben interpretar de forma continua ya que la continuidad del campo no se puede preservar en variables de valores discretos. Si el espacio es discreto (modelo de objetos) o si un espacio continuo se ha hecho discreto, los valores de las variables se pueden valorar de forma continua o discreta (valores nominales u ordinales). Podemos distinguir cuatro tipos de datos espaciales:

1. **Datos de Patrones Puntuales**: un conjunto de datos que consiste en una serie de ubicaciones de puntos en alguna región de estudio, en la que se han producido eventos de interés, como casos de una enfermedad o incidencia de un tipo de crimen.

2. **Datos de Campo**: también denominados datos geoestadísticos, se relacionan con variables que son conceptualmente continuas (modelo basado en

campo) y cuyas observaciones se han muestreado en un conjunto predefinido y fijo de ubicaciones de puntos.

3. **Datos de Área**: los valores de datos son observaciones asociadas con un número fijo de unidades de área (objetos de área) que pueden formar una red regular, como con imágenes de detección remota, o un conjunto de áreas o zonas irregulares, como condados, distritos, zonas censales e incluso países.

4. **Datos de Interacción Espacial**: también denominados flujo de origen-destino o datos de enlace, consisten en mediciones tal que cada una de las cuales está asociada con un par de ubicaciones de puntos, o un par de áreas.

En este trabajo nos centraremos en estudiar con detalle los datos de tipo área donde las observaciones se relacionan con unidades de área (modelo basado en entidades).

1.3. Fiabilidad de los Datos Espaciales

Para realizar un buen análisis de datos espaciales necesitamos fundamentalmente una buena calidad de los datos, es decir, datos que contengan pocos o ningún error para que los resultados que se obtengan sean fiables. Sin embargo, casi todos los datos tienen errores que pueden surgir al medir tanto la ubicación (puntos, líneas, áreas) como las propiedades de atributo de los objetos espaciales. La solución al problema de la calidad de los datos es tomar las medidas necesarias para evitar tener datos defectuosos que modifiquen esencialmente los resultados de la investigación.

La forma particular de los conjuntos espaciales puede afectar a los resultados del análisis con un cierto grado variable, normalmente desconocido. Este problema se conoce como problema de unidad de área modificable (MAUP, *modifiable areal unit problem*) que deriva del hecho de que las unidades de área no son constructos "naturales"sino usualmente arbitrarios. Este problema, del que profundizaremos a lo largo del trabajo, podría investigarse a través de la simulación de un gran número de sistemas alternativos de unidades de área.

También existen restricciones de confidencialidad que hacen que no se pueda publicar datos de observaciones primarias sino solo de un conjunto de áreas bastantes arbitrarias. El problema surge cada vez que se quiere analizar o modelar datos de área e implica dos efectos: uno se deriva de seleccionar diferentes límites de área manteniendo constante el tamaño general y el número de unidades de área (el efecto de zonificación) y el otro se deriva de reducir el número pero aumentar el tamaño de las unidades de área (el efecto de escala).

Otro problema que nos encontramos a menudo es la falacia ecológica que consiste en que el investigador utiliza los datos en una escala espacial para llegar a conclusiones sobre una relación a una escala más exacta lo que nos lleva a un falso sentido del poder de nuestras técnicas y la utilidad de ellas.

Capítulo 2

Análisis Exploratorio Datos Espaciales

En este capítulo nos centraremos en el análisis teórico previo al modelado de los datos, prestando especial atención a los datos de tipo área. En la primera sección veremos que la forma más fácil e ilustrativa de mostrar los datos espaciales es el mapa, del que daremos algunas características. Para analizar los datos necesitamos tenerlos agrupados en una matriz llamada matriz de pesos que utilizaremos para ver si nuestros datos exhiben autocorrelación espacial o son independientes de la ubicación en la que se encuentran. Por último veremos un estadístico general que por sí solo no proporciona información pero que será muy útil en las medidas para analizar la autocorrelación espacial.

2.1. Visualización de Datos

En el análisis exploratorio de datos espaciales el mapa es el medio más establecido y convencional para mostrar datos de área. El más utilizado es el mapa de coropletas o mapa coroplético que es un mapa donde cada una de las áreas está coloreada o sombreada de acuerdo con una escala discreta basada en el valor de la variable de interés (atributo) dentro de esa área. El número de clases y los intervalos de clases correspondientes pueden basarse en varios criterios diferentes. El número de clases viene dado en función de cuantas observaciones tenemos y como regla general algunos estadísticos recomiendan la siguiente fórmula para este número: $1+3.3ln(n)$, donde n es el número de áreas. En cuanto a la selección del intervalo de clase tenemos cuatro esquemas (ver figura 2.1):

- **Divisiones naturales.** Las clases se definen de acuerdo con algunas agrupaciones naturales de los valores de los datos. Los diferentes cortes pueden imponerse a partir de puntos de ruptura que se sabe que son relevantes (asignación deductiva) o mediante el uso de herramientas del sistema GIS[1] que buscan saltos importantes en los valores de datos.

- **Divisiones por cuantiles.** Cada clase contiene un número igual de observaciones. Las clasificaciones de cuartiles (cuatro categorías) y quintiles (cinco categorías) son las más comúnmente utilizadas en la práctica.

- **Divisiones de intervalos iguales.** Tienen validez cuando las observaciones se distribuyen de manera razonablemente uniforme en su rango. Pero si los datos son marcadamente sesgados, darán un gran número de observaciones en unas pocas clases.

- **División según desviación estándar.** Se basan en intervalos distribuidos alrededor de la media en unidades de desviación estándar.

El uso de mapa de coropletas tiene algunos problemas que deben ser tenidos en cuenta antes de su elaboración:

1. Los mapas de coropletas traen la implicación visual de la uniformidad dentro del área de los valores variables. Además, el mapa de coropletas convencional permite que cualquier área físicamente grande domine la pantalla, de una manera que puede ser bastante inapropiada para el tipo de datos que se mapean. Por ejemplo, al mapear datos socioeconómicos, las áreas rurales grandes y escasamente pobladas pueden dominar el mapa pero el interés real puede estar en áreas físicamente más pequeñas, como las áreas urbanas más densamente pobladas.

2. La variable de interés ha surgido de la agregación de datos individuales a las áreas de forma arbitraria y cualquier patrón que observemos, tanto en los límites de área elegidos como en la distribución espacial, proviene de dicha arbitrariedad de los valores de las variables, no teniendo ningún rigor ni formalidad a la hora de estudiar esos datos con la hipótesis de seguir dicho patrón encontrado. Este es el problema de unidad de área modificable (MAUP) que surge a menudo en el mapa de coropletas.

[1]Las siglas GIS provienen de sistema de información geográfica que es un conjunto de herramientas que integra y relaciona diversos componentes que permiten la organización, almacenamiento, análisis y modelización de grandes cantidades de datos procedentes del mundo real que están vinculados a una referencia espacial.

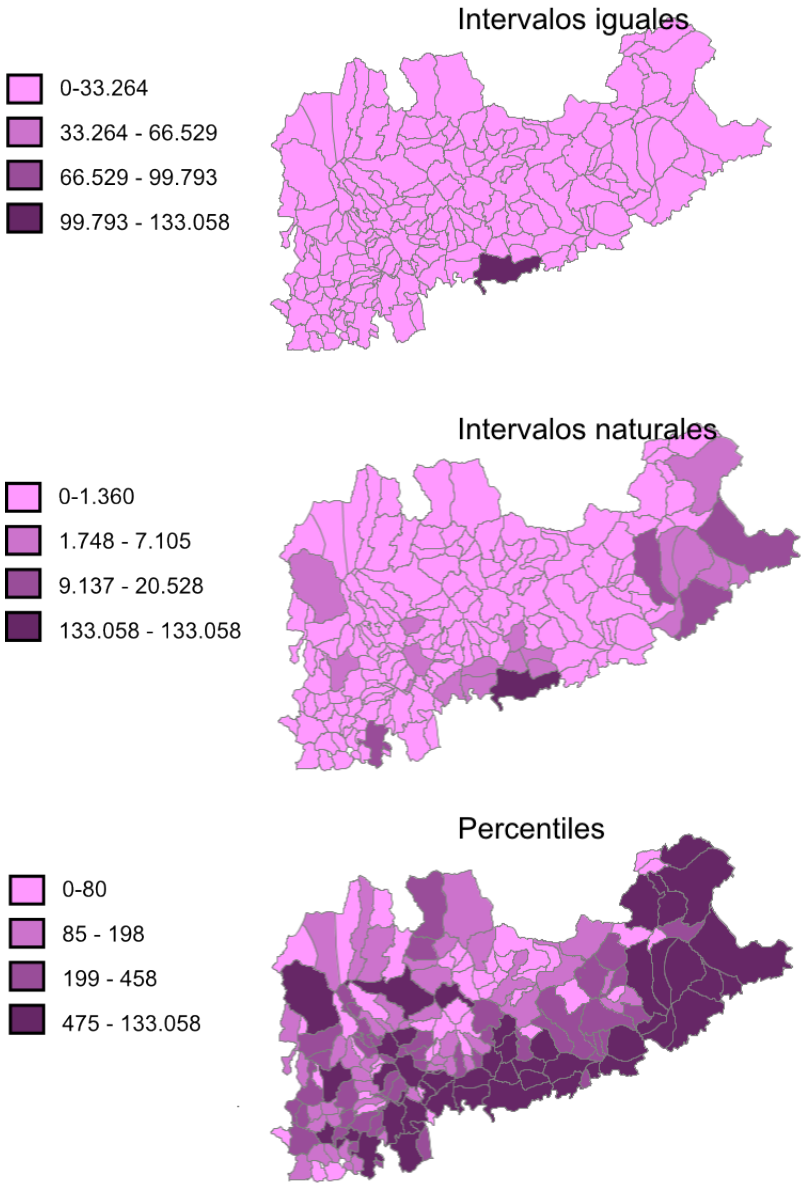


Figura 2.1: Tipos de intervalos de clases para un mapa de coropletas

3. Es importante darse cuenta de que los resultados estadísticos de cualquier análisis de patrones y relaciones dependerán inevitablemente de la configuración de área particular que se utilice, como hemos visto en el párrafo anterior. En general, para conseguir un buen análisis y, por tanto unos resultados fiables de lo que estudiemos, los datos deben analizarse sobre la base de las unidades de área más pequeñas para las cuales están disponibles

y debe evitarse la agregación a áreas más grandes arbitrarias, a menos que haya buenas razones para hacerlo. También es importante verificar cualquier inferencia extraída de los datos mediante el uso de diferentes configuraciones de área de los mismos datos.

2.2. Matriz de Datos Espaciales

Todas las técnicas analíticas que utilizamos en este trabajo usan una matriz de datos que recoge los datos espaciales necesarios para la realización del análisis.

Sean $Z_1, Z_2, \ldots, Z_K$ variables aleatorias y sea S la ubicación del punto o área. La matriz de datos espaciales viene generada por:

$$\left(\begin{array}{cccc|c} z_1(1) & z_2(1) & \cdots & z_K(1) & s(1) \\ z_1(2) & z_2(2) & \cdots & z_K(2) & s(2) \\ \vdots & \vdots & & \vdots & \vdots \\ z_1(n) & z_2(n) & \cdots & z_K(n) & s(n) \end{array}\right)$$

la cual puede ser expresada de la siguente forma

$$\left\{ z_1(i), \quad z_2(i), \quad \cdots \quad z_K(i) \mid s(i) \right\}_{i=1\ldots n} \tag{2.1}$$

donde z_k denota una realización muestral (valor de datos real) de la variable $Z_k (k = 1, \ldots, K)$ mientras que el símbolo i dentro de los paréntesis hace referencia al caso u observación particular (punto, área, polígono o zona espacial). Para cada caso, $i = 1, \ldots, n$, $s(i)$ representa la ubicación del objeto espacial. La referencia implicará dos coordenadas geográficas. Generalmente nos referiremos a un espacio bidimensional, por tanto $s(i) = (s_1(i), s_2(i))^t$.

En el caso de datos referentes a objetos de puntos en un espacio bidimensional, la ubicación del punto i-ésimo puede estar dada por un par de coordenadas cartesianas (ortogonales). Los ejes del sistema de coordenadas generalmente se han construido para el conjunto de datos particular, pero se puede usar cualquier sistema de referencia. En el caso de datos referentes a objetos de área de forma irregular, una opción para representar la ubicación es seleccionar un punto representativo de cada área denominado centroide y luego usar el mismo procedimiento que para un objeto de punto para identificar $s(i)$. Hay situaciones en las que la información de georreferencia proporcionada por $\{s(i)\}$ en la expresión (2.1) tiene que completarse con información de vecindad que define no solo qué pares de áreas son adyacentes entre sí, sino que también puede cuantificar la cercanía de esa adyacencia. Esta información es necesaria para la especificación de muchos modelos

estadísticos espaciales, como los modelos de regresión espacial. Dicha información está recogida en el concepto de autocorrelación espacial que veremos en la siguiente sección.

2.3. Autocorrelación Espacial

Para comenzar con esta sección veamos algunas definiciones que han dado diferentes autores de la autocorrelación espacial que quedan recogidas en el estudio del laboratorio de paleontología y climatología de la universidad de Ottawa [12]. Para Cliff y Ord (1973) es útil ver si la distribución de cierta cualidad o cantidad en los condados o estados de un país hace que su presencia en los condados vecinos sea más o menos probable. En tal caso, exponen que el fenómeno exhibe autocorrelación espacial. Sokal y Oden (1978) argumentaron que el análisis de autocorrelación espacial prueba si el valor observado de una variable nominal, ordinal o de intervalo en una localidad es independiente de los valores de esa misma variable en las localidades vecinas.

Upton y Fingleton (1985) la definen como una propiedad que los datos poseen cuando muestran un patrón de comportamiento. Estos autores exponen que la autocorrelación espacial existe siempre que haya una variación espacial sistemática en los valores a lo largo de un mapa, o patrones en los valores registrados en las localizaciones. Goodchild (1987) y Griffith (1991) destacaron que la autocorrelación espacial trata simultáneamente con la información de ubicación y de atributos. Goodchild (1987) dice que en su sentido más general la autocorrelación espacial se refiere al grado en que los objetos o actividades en algún lugar de la superficie son similares a otros objetos o actividades ubicados cerca y refleja la primera ley de geografía de Tobler [13] *"todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes"*.

Así, según las diferentes aportaciones sobre el concepto de autocorrelación espacial, se puede llegar a la conclusión de que si el valor de una o varias variables en una ubicación son similares a los valores de dichas variables en ubicaciones cercanas, entonces se dice que el patrón en conjunto exhibe una autocorrelación espacial positiva (autocorrelación). Por el contrario, se dice que existe autocorrelación espacial negativa cuando las observaciones que están cerca en el espacio tienden a ser más diferentes en los valores de las variables que las observaciones que están más separadas (en contradicción con la ley de Tobler). La autocorrelación cero se produce cuando los valores variables son independientes de la ubicación (en esta situación los datos no serían espaciales y por tanto no aplicaremos las técnicas que estudiaremos en este trabajo).

En el análisis de autocorrelación espacial se necesita una medida de contigüidad que podemos definirla de manera general como una relación de vecindad. Estas

relaciones pueden ser de tres tipos principalmente, caso de torre (Rook's case), caso de alfil (Bishop's case) y caso de la reina (Queen's case) que reciben este nombre por analogía con los movimientos de las figuras de ajedrez. La contigüidad de tipo torre es por un vecindario de 4 ubicaciones adyacentes a cada celda (posiciones N, S, E, W), la de tipo alfil solo considera las diagonales de cada celda (posiciones NE, SE, NW, SW) mientras que el último tipo considera un vecindario de ocho celdas (posiciones N, S, E, W, NE, NW, SW, SE). En la siguiente figura podemos observar más intuitivamente estos tipos de contigüidad.

Figura 2.2: Tipos de relaciones de vecindad

Un aspecto crucial de la definición de la autocorrelación espacial es la determinación de ubicaciones cercanas, es decir, aquellas ubicaciones que rodean un punto de datos dado que podría considerarse que influyen en la observación en ese punto de datos. Sin embargo, la determinación de este vecindario tiene un cierto grado de arbitrariedad. El número de observaciones en el vecindario establecido para cada ubicación puede expresarse mediante una matriz de ponderaciones **W**:

$$\mathbf{W} = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \vdots & \vdots & & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{pmatrix}$$

donde n representa el número de ubicaciones (observaciones). La entrada en la fila $i(i = 1, \ldots, n)$ y columna $j(j = 1, \ldots, n)$, denotado como W_{ij}, corresponde al par (i, j) de ubicaciones. Los elementos diagonales de la matriz son cero, por convenio, mientras que los elementos no diagonales $W_{ij}(i \neq j)$ toman valores distintos de cero (uno, para una matriz binaria) cuando las ubicaciones i y j se consideran vecinas.

En concreto para los datos de áreas, n representaría el número de áreas donde cada área se identifica con un punto (centroide) del que se conocen sus coordenadas

cartesianas y cada elemento de la matriz **W** corresponde a la relación de dos áreas. Dicha matriz, **W**, a menudo está estandarizada por filas, es decir, cada suma de filas en la matriz se hace igual a uno, así, los valores individuales W_{ij} están proporcionalmente representados. Esto se hace para que cada vecino de un área tenga el mismo peso y la suma de todos los W_{ij} (sobre j) sea igual a uno $\left(\sum_{j=1}^{n} W_{ij} = 1\right)$.

Una forma de representar las relaciones espaciales con datos de área es a través del concepto de contigüidad. Los vecinos contiguos de primer orden se definen como áreas que tienen un límite común. Formalmente:

$$W_{ij} = \begin{cases} 1 & \text{si el área } j \text{ comparte un límite común con el área } i \\ 0 & \text{caso contrario} \end{cases}$$

Alternativamente, dos áreas i y j pueden definirse como vecinas cuando la distancia d_{ij} entre sus centroides es menor que un valor crítico dado, pongamos d, donde las distancias se calculan a partir de la información sobre latitud y longitud, $s(i)$, de las ubicaciones del centroide:

$$W_{ij} = \begin{cases} 1 & \text{si } d_{ij} < d,\ (d < 0) \\ 0 & \text{caso contrario} \end{cases} \tag{2.2}$$

La especificación basada en la distancia (2.2) de la matriz de ponderaciones depende de un valor de distancia crítica dado, d. Sin embargo, cuando hay un alto grado de heterogeneidad en el tamaño de las unidades de área, puede ser difícil encontrar una distancia crítica satisfactoria. En tales circunstancias, una pequeña distancia tenderá a conducir a muchas "islas", mientras que una distancia elegida para garantizar que cada unidad de área tenga al menos un vecino puede producir un tamaño inaceptablemente grande de número de vecinos para las unidades de área más pequeñas. Una solución común a este problema es restringir la estructura contigua a los k-vecinos más cercanos, y por lo tanto excluir las "islas"(áreas que por no estar a una distancia d de otras áreas se podría decir que no tiene vecinos) y forzar a cada unidad de área a tener el mismo número k de vecinos. Formalmente:

$$W_{ij} = \begin{cases} 1 & \text{si el centroide de } j \text{ es uno de los } k \text{ centroides más cercanos al de } i \\ 0 & \text{caso contrario} \end{cases}$$

En este caso el número de vecinos, k, es el parámetro de este esquema de ponderación. También podemos cambiar la ponderación para que los vecinos más distantes obtengan menos peso introduciendo un parámetro θ que permita indicar la tasa de disminución de los pesos.

Un esquema de ponderación continua comúnmente utilizado se basa en la función de distancia inversa, de modo que los pesos están inversamente relacionados con el área de separación de distancia i y el área j donde el parámetro θ se estima o se establece a priori:

$$W_{ij} = \begin{cases} d_{ij}^{-\theta} & \text{si la distancia entre centroides } \ d_{ij} < d \ (d > 0, \theta > 0) \\ 0 & \text{caso contrario} \end{cases}$$

Otro esquema de ponderación continua se deriva de la función exponencial negativa que viene dado por:

$$W_{ij} = \begin{cases} \exp(-\theta d_{ij}) & \text{si la distancia entre centroides } d_{ij} < d \ (d > 0, \theta > 0) \\ 0 & \text{caso contrario} \end{cases}$$

donde θ es un parámetro que puede ser estimado, pero generalmente es elegido a priori por el investigador. Una elección muy común es $\theta = 2$.

Evidentemente, existe una gran cantidad de matrices de ponderaciones espaciales para el mismo modelo espacial. Es importante tener siempre en cuenta que los resultados de cualquier análisis estadístico espacial están condicionados a la matriz de pesos espaciales elegida. Con frecuencia, es una buena práctica verificar la sensibilidad de las conclusiones con la elección de diferentes matrices de ponderaciones espaciales y ver si dichas conclusiones se siguen cumpliendo para la mayoría de las matrices de ponderaciones elegidas, a menos que exista una razón convincente sobre bases teóricas para considerar solo una.

2.4. Estadístico General de Productos Cruzados

La mayoría de medidas de autocorrelación espacial utilizan el producto cruzado de matrices definido de la siguiente forma:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} W_{ij}$$

donde W_{ij} son las entradas de la matriz de pesos espaciales definidas en la sección anterior y M_{ij} es una medida de proximidad entre las áreas i y j en alguna otra dimensión, por ejemplo, la distancia euclídea, distancia esférica o la distancia Manhattan.

Veamos un ejemplo simple y práctico que servirá para visualizar todos estos conceptos más fácilmente.

Ubicación	
A	B
C	D

Valores	
1	4
5	2

En este ejemplo estamos utilizando la contigüidad de tipo torre y utilizaremos la distancia euclídea para el cálculo de la matriz de medida de proximidad entre las áreas. El primer cuadro nos da información de cómo están repartidas cuatro áreas (A,B,C,D), por ejemplo el área A es contigüa a las áreas B y C, mientras que el segundo nos indica los valores de una variable de cada área.

La matriz de pesos **W** la formaríamos creando una matriz binaria 4×4 donde las columnas y las filas están etiquetadas por las ubicaciones de las celdas. Cuando dos celdas son adyacentes en los datos originales, ingresamos un 1 en la entrada (i, j) de la matriz **W**, si dos celdas no son adyacentes ingresamos un 0.

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

El siguiente paso antes de determinar el producto cruzado es crear la matriz **M**, que tiene la misma dimensión que la matriz **W**, donde las entradas se basan en alguna otra medida de distancia entre todas las celdas. Tomemos la distancia euclídea al cuadrado $d(x_i, x_j) = (x_i - x_j)^2$. Para ello escogemos la celda A y restamos su valor al de cada una de las otras celdas y lo elevamos al cuadrado. Por tanto la matriz **M** nos queda:

$$\mathbf{M} = \begin{pmatrix} 0 & 9 & 16 & 1 \\ 9 & 0 & 1 & 4 \\ 16 & 1 & 0 & 9 \\ 1 & 4 & 9 & 0 \end{pmatrix}$$

Veremos ahora la multiplicación de $W_{ij} * M_{ij}$ para después sumar las entradas de esta nueva matriz 4×4 y así obtener el valor del estadístico general de productos cruzados.

$$W_{ij} * M_{ij} = \begin{pmatrix} 0 & 9 & 16 & 0 \\ 9 & 0 & 0 & 4 \\ 16 & 0 & 0 & 9 \\ 0 & 4 & 9 & 0 \end{pmatrix}$$

Tomando la suma de las entradas de la última matriz llega a un producto cruzado de 76 $\left(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}M_{ij} = 76\right)$. Sin embargo, este valor no significa nada porque sería necesario un intervalo de referencia. El estadístico de productos cruzados no proporciona información sobre la estructura del mapa original porque no sabemos en qué contexto poner el valor de 76.

El producto cruzado general es un estadístico en el sentido de que la matriz resulta de una muestra de varias matrices posibles. El valor del producto cruzado se puede comparar con el rango de valores que se pueden producir si se crean varios mapas con el mismo conjunto de valores mediante la asignación aleatoria de esos valores a las áreas que tenemos. En nuestro ejemplo tenemos cuatro valores (1,4,5,2) que asignamos a las áreas (A,B,C,D), respectivamente. Sin embargo este mapa original lo podemos considerar como una posible permutación de los valores en esas áreas. Tenemos 4! mapas distintos que podrían producirse aleatoriamente si cada uno de los cuatro valores se asignan al azar, con reemplazo, a un área individual. Esto nos daría 4! índices de productos cruzados con los que podemos crear una distribución a este estadístico. Obviamente en un mapa con mayor número de áreas no sería computacionalmente posible obtener el valor utilizando todos los índices de productos cruzados que tendríamos con todas las permutaciones posibles de los valores en las áreas del mapa. Para ello se utilizan unas técnicas de muestreo como el enfoque de Monte Carlo donde se toma una muestra aleatoria de todas las permutaciones posibles del mapa. Esta distribución o valores del producto cruzado se utiliza mayoritariamente en las medidas de autocorrelación espacial como por ejemplo el estadístico I de Moran o c de Geary que veremos en profundidad en el siguiente capítulo.

Capítulo 3

Medidas de Autocorrelación Espacial

Las medidas de autocorrelación espacial se distinguen entre globales y locales. Global implica que todos los elementos de la matriz **W** se aplican a una evaluación de la autocorrelación, es decir, todas las asociaciones espaciales de áreas se incluyen en el cálculo de la autocorrelación espacial. En otras palabras, el análisis espacial global es un estadístico para resumir toda la zona de estudio, por tanto asume homogeneidad, y esto permite que se produzca un valor para cualquier matriz de pesos espaciales. Por el contrario, las medidas locales evalúan la autocorrelación espacial asociada con una o algunas unidades de área particulares.

3.1. Medidas globales

Las dos medidas que más se han utilizado para el caso de las unidades de área son los estadísticos I de Moran y c de Geary. Moran [10] introdujo la primera medida de autocorrelación espacial para estudiar fenómenos estocásticos que se distribuyen en el espacio en dos o más dimensiones. La medida I de Moran viene dada por la siguiente expresión:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3.1}$$

donde n es el número de áreas que consideramos y W_{ij} es una matriz de pesos definida anteriormente. En esta medida concreta, la distancia que utiliza para

crear la matriz $\mathbf{M}$ viene dada por $(z_i - \bar{z})(z_j - \bar{z})$ siendo z_i y z_j el valor de la variable Z del área i y j respectivamente, y siendo $\bar{z}$ la media aritmética de todos los valores de las áreas que queremos estudiar.

Aunque es una de las medidas más antiguas, sigue siendo de las medidas más utilizadas para determinar la autocorrelación espacial comparando el valor de Z en el área i con el valor de Z en todas las demás áreas j ($j \neq i$). Observando la ecuación nos fijamos que el sumatorio $\sum_{i=1}^{n}(z_i - \bar{z})^2$ no puede ser 0 pues la ecuación no estaría bien definida. Sin embargo, en el caso de un campo "casi sin varianza", los valores serían similares (iguales), en este caso la medida I de Moran no sería apropiada para ver la autocorrelación espacial. El otro sumatorio que en principio podría dar problema de definición sería $\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$ pero este nos indica la suma de las entradas de la matriz de adyacencia lo cual, para aplicar las técnicas descritas, hemos asumido que existe relación de adyacencia entre las distintas áreas y, por tanto, el término no podrá anularse.

Como consecuencia, la única forma que el estadístico I de Moran sea 0 es que $\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - \bar{z})(z_j - \bar{z}) = 0$. Este estadístico varía entre -1 y 1. Un valor de I cercano a 1 implicará autocorrelación espacial positiva mientras que un valor próximo a -1 nos indica una autocorrelación espacial negativa. En el caso que el valor de I esté cercano a 0, podemos rechazar la hipótesis de que haya autocorrelación espacial y, por tanto, los modelos y técnicas de este trabajo no se podrán utilizar ya que nos proporcionarán información poco fiable de los datos observados.

Sigamos con el ejemplo que vimos en el capítulo anterior para ver el valor del estadístico I de Moran:

Como ya hemos visto, cada entrada de la matriz de valores $\mathbf{M}$ se crea con $(z_i - \bar{z})(z_j - \bar{z})$, esto es, por ejemplo, la entrada $M_{1,1}$ sería $(1-3)(1-3) = 4$ pues $\bar{z} = 3$ o bien la entrada $M_{2,3}$ es $(4-3)(5-3) = 2$. Por tanto la matriz quedaría:

$$\mathbf{M} = \begin{bmatrix} 4 & -2 & -4 & 2 \\ -2 & 1 & 2 & -1 \\ -4 & 2 & 4 & -2 \\ 2 & -1 & -2 & 1 \end{bmatrix}$$

quedando la matriz $W_{ij} * M_{ij}$:

$$W_{ij} * M_{ij} = \begin{bmatrix} 0 & -2 & -4 & 0 \\ -2 & 0 & 0 & -1 \\ -4 & 0 & 0 & -2 \\ 0 & -1 & -2 & 0 \end{bmatrix}$$

Es evidente que para conseguir esta última matriz (de productos cruzados) solo tenemos que tener en cuenta los valores de M_{ij} donde las entradas de la matriz de adyacencia W_{ij} sea distinta de 0 (entradas donde haya un 1 por ser esta una matriz binaria). Podemos ahorrarnos el hacer la matriz de valores y el producto (entrada por entrada) de ella con W_{ij} haciendo el siguiente algoritmo para una matriz de pesos binaria:

```
1. Entrada W_ij
2. Hacer variable W_ijM_ij
3. For i pertenciente a la matriz de datos original
4. For j pertenciente a la matriz de datos original
5. If  W(i,j) = 1 haz:
6.     W_ijM_ij = W_ijM_ij + (z_i - mean(z))(z_j - mean(z))
7. Terminar si
8. Siguiente i
9. Siguiente j
10. Salida W_ijM_ij
```

siendo $mean(z)$ equivalente a $\bar{z}$, es decir, la media aritmética de todos los valores de las distintas áreas del estudio. Si la matriz de pesos no es binaria y tuviésemos una distancia crítica d como ya vimos en los tipos de matrices de pesos solamente habría que cambiar el paso 5. por otro " If " correspondiente.

Para finalizar, nos queda dar el valor del estadístico I:

$$I = \frac{4}{8} \cdot \frac{-18}{10} = \frac{-72}{80} = -0{,}9$$

Al ser un valor cercano a -1 podemos pensar que el fenómeno posee una autocorrelación espacial negativa, es decir, los valores de las áreas cercanas serán distintos en contraposición con la Ley de Tobler.

Otro índice muy comunmente utilizado para la autocorrelación espacial es el índice c de Geary, el cual es un índice de comparaciones por pares de datos entre las diferentes áreas. La expresión del estadístico c viene dada por la siguiente fórmula:

$$c = \frac{(n-1)}{2\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \tag{3.2}$$

donde los símbolos usados en la definición son similares a los del estadístico I de Moran definidos anteriormente. El índice c de Geary toma valores entre 0 y un valor positivo aunque solo en situaciones concretas el valor excede el 2 (véase Gangodagam y otros [6]). Al contrario que el estadístico I de Moran, un valor del estadístico c que tiende a 0 ($c < 1$) nos indica una autocorrelación positiva mientras que un valor que tiende a 2 ($c > 1$) nos estaría indicando una autocorrelación negativa. Un valor cercano a 1 nos estará indicando que hay ausencia de autocorrelación espacial en nuestros datos.

En el ejemplo anterior la matriz de pesos ($\mathbf{W}$) sería la misma que se utilizó para el cálculo de I, mientras que la matriz de valores cambia puesto que utiliza otra medida de distancia. La matriz $W_{ij}M_{ij}$ siendo en este caso $M_{ij} = (z_i - z_j)^2$, es

$$W_{ij} * M_{ij} = \begin{bmatrix} 0 & 9 & 16 & 0 \\ 9 & 0 & 0 & 4 \\ 16 & 0 & 0 & 9 \\ 0 & 4 & 9 & 0 \end{bmatrix}$$

por lo que

$$c = \frac{3}{16} \cdot \frac{76}{10} = \frac{228}{160} = 1{,}4$$

Vemos con este resultado que también interpretamos la presencia de autocorrelación espacial negativa en los datos al igual que el estadístico I de Moran.

Observamos que la interpretación de los valores de c de Geary es opuesta a la de los valores de I de Moran, es decir que altos valores de c equivalen a bajos valores de I y viceversa. Griffith (1987) señala que esta relación inversa entre los índices es básicamente de naturaleza lineal. Las desviaciones de la linealidad se atribuyen a las diferencias en lo que mide cada uno de los dos índices, es decir, la c de Geary se ocupa de las comparaciones pareadas y la I de Moran de las covariaciones. Cliff y Ord (1981) señalan que I parece estar menos afectado por la distribución de los datos que c y adjudican una ligera ventaja estadística a I respecto a c. Sin embargo, el índice I de Moran parece ser más sensible a los valores extremos que el índice c (Legendre y Fortin, 1989).

Estos estadísticos o índices se utilizan en las pruebas de autocorrelación espacial, que son reglas de decisión para evaluar en qué medida la distribución espacial observada de los valores de los datos parte de la hipótesis nula de que el espacio no importa, es decir, las áreas cercanas no se afectan entre sí de forma tal que haya independencia y aleatoriedad espacial mientras que la hipótesis alternativa implica que las áreas se agrupan mediante algún tipo de autocorrelación espacial (positiva o negativa).

Se considera que la autocorrelación espacial está presente cuando el estadístico de autocorrelación espacial calculado para un patrón particular adquiere un valor distinto, en comparación con lo que se esperaría bajo la hipótesis nula de ausencia de asociación espacial. Esta variación depende de la distribución del estadístico de contraste que utilicemos que, generalmente, siguen una aproximación normal o una aproximación por permutación aleatoria. Como vimos en la sección de productos cruzados, si tenemos n observaciones, z_i, para n unidades de áreas dadas, $n!$ permutaciones de esos valores u observaciones son posibles y para cada una de esas permutaciones se puede calcular el valor del estadístico. Sin embargo computacionalmente esto no sería apropiado ya que a medida que aumente las áreas las permutaciones creceran de manera exponencial. Por tanto utilizamos un método de muestreo (Monte Carlo, por ejemplo) para escoger, al azar, un número razonable de permutaciones para calcularles el estadístico. La varianza de los estadísticos I y c bajo permutación aleatoria son respectivamente:

$$Var_P(I) = \frac{(n\lfloor(n^2-3n+3)S_1 - nS_2 + 3S_0^2\rfloor)}{(n-1)(n-2)(n-3)S_0^2} -$$

$$- \frac{k\lfloor(n^2-n)S_1 - 2nS_2 + 6S_0^2\rfloor}{(n-1)(n-2)(n-3)S_0^2} - 2E(I)$$

$$Var_P(c) = \frac{[(n-1)S_1(n^2-3n+3-(n-1)k)] + [S_0^2(n^2-3-(n-1)^2k)]}{(n)(n-2)(n-3)S_0^2} -$$

$$- \frac{[\frac{1}{4}(n-1)S_2(n^2+3n-6-(n^2-n+2)k)]}{(n)(n-2)(n-3)S_0^2}$$

donde los parámetros desconocidos se definen como

n = número de áreas

$E(I) = \dfrac{-1}{(n-1)}$, la esperanza de I.

$S_0 = \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$, la suma de las entradas de la matriz de pesos.

$S_1 = \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} \left(W_{ij} + W_{ji}\right)^2$, si la matriz de pesos es simétrica entonces $S_1 = 2S_0$.

$S_2 = \displaystyle\sum_{i=1}^{n} \left(W_{i.} + W_{.i}\right)^2$, 2 veces la suma de la i-ésima columna y la i-ésima fila.

$$k = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(z_i - \bar{z})^4}{\left[\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(z_i - \bar{z})^2\right]^2}.$$

Bajo ciertas condiciones de normalidad y un número apropiado de áreas, podemos utilizar una aproximación normal de los valores, es decir, cada valor z_i es un valor que se toma de una variable Z_i que sigue una distribución normal. La varianza de los estadísticos I y c bajo aproximación normal son respectivamente:

$$Var_N(I) = \frac{1}{S_0^2(2n-1)}(n^2S_1 - nS_2 + 3S_0^2) - E(I)^2$$

$$Var_N(c) = \frac{1}{2(n+1)S_0^2}((2S_1 + S_2)(n-1) - 4S_0^2)$$

Como conclusión, tanto estos estadísticos I de Moran y c de Geary como los estadísticos locales, que veremos en la siguiente sección, sirven para analizar la autocorrelación espacial de los datos a partir del siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \text{autocorrelación espacial nula} \\ H_1 : \text{autocorrelación espacial no nula} \end{cases} \tag{3.3}$$

Hemos comentado anteriormente que los estadísticos se aproximan, de manera general, a una distribución normal o una aproximación por permutación aleatoria. Esta última, también llamada prueba de asignación al azar, presupone que los valores observados se tratan como una población y por este motivo no tomaremos valores distintos a los observados para analizar la organización de dichos valores en las distintas áreas. Sin embargo, la prueba de distribución de muestreo aproximada hace la suposición de que las observaciones z_i son observaciones sobre variables aleatorias (normales) Z_i, es decir, son una realización de un proceso aleatorio y pueden darse otras realizaciones posibles. La prueba de distribución de muestreo aproximada es, por tanto, una de autocorrelación espacial, que permite suponer que la distribución de las variables aleatorias Z_i puede ser normal.

Para concluir esta sección de estadísticos globales cabe destacar que las fórmulas de la varianza de los estadísticos son más simples en el caso de aproximar los valores a una variable normal por lo tanto este método de aproximación es el que se intenta utilizar siempre que se den las condiciones adecuadas.

3.2. Medidas locales

Con la llegada de grandes conjuntos de datos característicos de sistemas GIS se han desarrollado varios estadísticos, llamados estadísticos locales que son adecuados para identificar la existencia de puntos "calientes" (grupos locales de valores altos) o puntos "fríos" (grupos locales de valores bajos) y para identificar distancias para las cuales habría asociación entre algunas áreas particulares pero que con distancias más grandes la información que tendríamos de asociación de las áreas sería más confusa e irrelevante.

Supongamos que cada área i $(i = 1, \ldots, n)$ se le asocia el valor z_i que representa una observación sobre la variable aleatoria Z_i. Suponemos que los Z_i tienen distribuciones marginales idénticas pero no son independientes ya que, como ya hemos visto, se daría autocorrelación espacial cero y esto implicaría ausencia de autocorrelación espacial. La base para las pruebas locales y las medidas de autocorrelación espacial proviene de una variación del estadístico de productos cruzados:

$$\sum_{j=1}^{n} M_{ij} W_{ij}$$

Describimos brevemente cuatro estadísticos locales: los estadísticos locales de Getis y Ord G_i y G_i^* y las versiones locales de I de Moran y c de Geary. Las estadísticas de Getis y Ord se calculan definiendo un conjunto de vecinos para cada área i como aquellas observaciones que caen dentro de una distancia crítica d de i donde cada $i = 1, \ldots, n$ se identifica con un punto (centroide). Esto se puede

expresar formalmente en una matriz de pesos binarios simétrica $\mathbf{W}(d)$ (véase 2.3), con los elementos $W_{ij}(d)$ indexados por la distancia d.

Los estadísticos G_i y G_i^* miden el grado de asociación local para cada observación i en un conjunto de datos que contiene n observaciones. Formalmente, estas medidas para cada observación (área) i se pueden expresar como:

$$G_i(d) = \frac{\sum_{j \neq i}^{n} W_{ij}(d) z_j}{\sum_{j \neq i}^{n} z_j}$$

$$G_i^*(d) = \frac{\sum_{j=1}^{n} W_{ij}(d) z_j}{\sum_{j=1}^{n} z_j}$$

El estadístico G_i puede interpretarse como una medida de la agrupación de valores similares entorno a una observación particular i, independientemente del valor en esa área, mientras que la estadística G_i^* incluye el valor dentro de la medida de la agrupación. Un valor positivo indica agrupamiento de valores altos y un valor negativo indica un grupo de valores bajos. La distribución de estos estadísticos es normal si la distribución subyacente de las observaciones es normal. Pero si la distribución es asimétrica, la prueba solo se aproxima a la normalidad a medida que aumenta la distancia crítica d. Por tanto, bajo estas circunstancias, la normalidad del estadístico de contraste solo puede garantizarse cuando el número de áreas vecinas es grande.

El estadístico local I_i de Moran para cada área i se define:

$$I_i = (z_i - \bar{z}) \sum_{j \in J_i}^{n} W_{ij}(z_j - \bar{z})^2$$

donde J_i denota el conjunto de vecinos del área i y $\bar{z}$ denota la media de las obseraciones que son vecinas al área i.
Podemos comprobar que la suma de los I_i para todas las observaciones i es proporcional al estadístico I de Moran dado por la ecuación (3.1)

$$\sum_{i=1}^{n} I_i = \sum_{i=1}^{n} (z_i - \bar{z}) \sum_{j \in J_i}^{n} W_{ij}(z_j - \bar{z})$$

El estadístico local c_i de Geary para cada área i se define:

$$c_i = \sum_{j \in J_i}^{n} W_{ij}(z_i - z_j)^2$$

De la misma forma que para el estadístico I_i de Moran es fácil comprobar que la suma de los c_i para todas las observaciones i es proporcional al estadístico c de Geary dado por la ecuación (3.2)

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \sum_{j \in J_i}^{n} W_{ij}(z_i - z_j)^2$$

Estos estadísticos de LISA[1], I_i y c_i, tienen dos propósitos. Por un lado, pueden ser vistos como indicadores de puntos calientes, similares a G_i y G_i^*. Por otro lado, se pueden usar para evaluar la influencia de ubicaciones individuales (observaciones) en la magnitud del estadístico de autocorrelación espacial global correspondiente, I de Moran y c de Geary.

[1]LISA: *(Local Indicator of Spatial Association)*,son estadísticas que descompone el índice global de autocorrelación y verifica en cuánto contribuye cada unidad espacial a la formación del valor general, permitiendo capturar de forma simultánea el grado de asociación espacial y la heterogeneidad resultante del aporte de cada unidad espacial.

Capítulo 4

Modelado de los datos de área

El análisis exploratorio visto en el capítulo 2 es un paso preliminar para abordar el modelo que busca establecer relaciones entre las observaciones de una variable y las observaciones de otras variables, registradas para cada unidad de área. Para estos modelos de regresión espacial se utilizan datos recogidos a partir de un diseño transversal simple, sin tener en cuenta datos de panel[1]. Este diseño toma una muestra de datos de la población objetivo y se obtiene información de esta muestra una única vez. También consideraremos que los datos en cuestión se distribuyen aproximadamente como una distribución normal. Para los datos en los que la variable de interés sea un recuento o una proporción esperamos que los modelos para dichos datos tengan una distribución de Poisson o Binomial ya que a medida que el número de áreas crezca, estas distribuciones tienden asintóticamente a una distribución normal. En este capítulo, a partir del modelo de regresión lineal (múltiple) veremos diferentes modelos de regresión espacial, tales como: modelo de retardo espacial, modelo de error espacial, modelo de orden superior y el modelo espacial de Durbin. En la sección 4.2 podremos ver pruebas para la dependencia espacial que nos ayudará a saber si los datos que estamos analizando poseen autocorrelación espacial. También estudiaremos las estimaciones de los parámetros de algunos de los modelos nombrados.

4.1. Tipos de modelos de regresión espacial

Para comenzar esta sección, recordaremos el modelo de regresión lineal que será nuestro punto de partida para estudiar los principales modelos de regresión espacial.

[1]Un conjunto de datos de panel recoge observaciones sobre múltiples fenómenos a lo largo de determinados períodos.

El modelo de regresión lineal (múltiple) tiene como forma funcional:

$$Y = \sum_{q=1}^{Q} X_q \beta_q + \varepsilon \tag{4.1}$$

y su versión muestral

$$y_i = \sum_{q=1}^{Q} x_{iq} \beta_q + \varepsilon_i \quad i = 1, \cdots n \tag{4.2}$$

donde y_i es una observación de la variable dependiente o de interés, x_{iq} es una observación en una variable explicativa con $q = 1, \ldots, Q$, β_q es el coeficiente de regresión que mide la influencia por sí sola de la q-variable explicativa en la variable dependiente, es decir, mide el cambio en Y por cada cambio unitario en X_q manteniendo las restantes variables explicativas constantes. El término ε_i es el error aleatorio, que puede ser debido a variables no controladas y a la variabilidad muestral. Para estos términos de error asumimos que ε_i son variables independientes e idénticamente distribuidas a una variable normal con media cero y varianza una constante σ^2, esto es, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ donde $E(\varepsilon_i \cdot \varepsilon_j) = E(\varepsilon_i) \cdot E(\varepsilon_j) = 0$.
En notación matricial este modelo de regresión lineal puede ser escrito de la siguiente forma:

$$Y = \mathbf{X}\beta + \varepsilon \tag{4.3}$$

donde Y es un vector $n \times 1$ en el que se encuentran las n observaciones de la variable dependiente, $\mathbf{X}$ es una matriz $n \times Q$ que muestra las observaciones de las variables explicativas, β es el vector $Q \times 1$ de parámetros de regresión asociados a dichas variables explicativas y el vector ε de dimensión $n \times 1$ de términos de error.

Este modelo es válido para entender la relación funcional entre la variable dependiente y las variables explicativas y estudiar cuáles pueden ser las causas de la variación de Y. Para ello, a partir de la información muestral, obtenemos un estimador de β ($\hat{\beta}$) para así obtener una predicción de Y a partir de las variables explicativas $X_1 \cdots, X_Q$.

Podemos ver que bajo la suposición de que las observaciones son independientes, el modelo se simplifica notablemente. Sin embargo a la hora de estudiar los datos espaciales, esta simplificación puede desembocar a unos resultados parciales inconsistentes debido a la dependencia espacial. Esta dependencia espacial puede estar presente en variables explicativas, variable dependiente o en los residuos (términos de error). Cuando la dependecia espacial se encuentra en la varible dependiente los modelos se denominan **modelos de retardo espacial** mientras que si está en los residuos se denominan **modelos de error espacial**. Cuando

está presente en las variables explicativas se llaman **modelos de regresión cruzada** o **modelos X espacialmente retardados** pero, en contraste con los otros dos modelos, no precisan de procedimientos especiales para la estimación.

4.1.1. Modelos de retardo espacial

Estos modelos presentan la correlación espacial (dependencia) en la variable dependiente (Y). Estos modelos son extensiones de modelos de regresión de tipo (4.2). Permiten a las observaciones de la variable dependiente Y en el área i ($i = 1, \ldots, n$) depender de observaciones en áreas vecinas. El modelo de retardo espacial básico, llamado modelo autorregresivo espacial de primer orden (SAR), toma la forma muestral:

$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_j + \sum_{q=1}^{Q} x_{iq} \beta_q + \varepsilon_i \quad i = 1, \cdots n \tag{4.4}$$

donde el término error ε_i es independiente e idénticamente distribuido. W_{ij} es el elemento $(i, j)th$ de la matriz de pesos $n \times n$ $\mathbf{W}$ (véase 2.2). El escalar ρ es un parámetro a estimar que determinará el nivel de relación autorregresiva espacial entre y_i y $\sum_j W_{ij} y_j$ (combinación lineal de observaciones espacialmente relacionadas basadas en elementos distintos de cero en la i-ésima fila de $\mathbf{W}$).
En notación matricial se escribe:

$$Y = \rho \mathbf{W} Y + \mathbf{X}\beta + \varepsilon \tag{4.5}$$

$\mathbf{W}Y$ se conoce como una variable dependiente espacialmente retardada. El dominio de ρ viene definido por $(w_{min}^{-1}, w_{max}^{-1})$ donde w_{min} y w_{max} representan el máximo y mínimo autovalor de la matriz $\mathbf{W}$. Si la matriz de pesos $\mathbf{W}$ es estocástica, se puede probar que los autovalores de dicha matriz están en el intervalo $[-1, 1]$. Podemos ver la prueba de este lema en el capítulo 2 de Baris y Mete [8]. Cuando $\mathbf{W}$ es estocástica por filas, una entrada (i, j) distinta de cero nos indica que la j-ésima observación se usará para ajustar la predicción de la fila i.

Este es un modelo estructural, que puede ser expresado en forma reducida como sigue:

$$Y = (I - \rho \mathbf{W})^{-1} (\mathbf{X}\beta + \varepsilon). \tag{4.6}$$

En consecuencia su valor esperado viene dado por la expresión:

$$E[Y] = (I - \rho \mathbf{W})^{-1} (\mathbf{X}\beta)$$

El término $(I - \rho\mathbf{W})^{-1}$ se denomina multiplicador espacial e indica que el valor esperado de cada observación y_i dependerá de una combinación lineal de valores $\mathbf{X}$ tomados por observaciones vecinas, escalado por el parámetro de dependencia ρ.

4.1.2. Modelos de error espacial

Estos modelos explican la dependencia espacial en el término de error. La dependencia del error espacial puede surgir de variables latentes no observables que están correlacionadas espacialmente. También puede surgir de los límites del área que no reflejan con precisión la vecindad que dan lugar a las variables recopiladas para el análisis. La especificación más común es un proceso autorregresivo espacial de primer orden, dado por:

$$\varepsilon_i = \lambda \sum_{j=1}^{n} W_{ij}\varepsilon_j + u_i \quad i = 1, \cdots n$$

donde λ es el parámetro autorregresivo, y u_i un término aleatorio de error, que asumimos que es independiente e idénticamente distribuido.

Si $|\lambda| < 1$, entonces nos queda:

$$\varepsilon = (I - \lambda\mathbf{W})^{-1}u$$

Insertando esta ecuación en el modelo de regresión estándar obtenemos:

$$Y = \mathbf{X}\beta + (I - \lambda\mathbf{W})^{-1}u \tag{4.7}$$

con $E[uu'] = \sigma^2 I$. En consecuencia la matriz de varianza-covarianza del vector de errores aleatorio es:

$$E[\varepsilon\varepsilon^t] = \sigma^2(I - \lambda\mathbf{W})^{-1}(I - \lambda\mathbf{W^t})^{-1}$$

El modelo (4.7), denominado modelo de error espacial (SEM) puede verse como una combinación de un modelo de regresión estándar con un modelo autorregresivo espacial en el término de error ε.

4.1.3. Modelos de orden superior

Estos modelos usan múltiples matrices de pesos proporcionando una generalización sencilla de los modelos SAR y SEM. Por ejemplo, si usamos dos matrices espaciales $\mathbf{W_1}$ y $\mathbf{W_2}$ (no necesariamente una distinta de la otra) para combinar los modelos de retardo espacial y de error básicos da lugar:

$$Y = \rho \mathbf{W_1} Y + \mathbf{X}\beta + \varepsilon$$

$$\varepsilon = \lambda \mathbf{W_2} \varepsilon + u$$

$$u \sim \mathcal{N}(0, \sigma_u^2 I)$$

En este ejemplo con dos matrices de pesos, si establecemos $\rho = 0$, queda eliminado la influencia de las observaciones de la variable objetivo en los vecinos, resultando el modelo de error espacial dado por (4.7). Por otro lado, si $\lambda = 0$ se elimina el término de perturbación de retardo espacial generando el modelo de retardo espacial dado por (4.4).

4.1.4. Modelo espacial de Durbin

En algunos estudios puede sospecharse que la dependencia influye sobre la variable objetivo o dependiente a través de la propia variable objetivo y de las variables predictoras o explicativas. Se puede construir un modelo tipo SAR, visto antes dado por la ecuación (4.4), aumentado por variables explicativas espacialmente retardadas:

$$Y = \rho \mathbf{W} Y + \mathbf{X}\beta + \mathbf{W}\bar{\mathbf{X}}\gamma + \varepsilon \tag{4.8}$$

donde $\bar{\mathbf{X}}$ es la matriz $n \times (Q-1)$ no constante de variables explicativas. El modelo puede reducirse como:

$$Y = (I - \rho \mathbf{W})^{-1}(\mathbf{X}\beta + \mathbf{W}\bar{\mathbf{X}}\gamma + \varepsilon)$$

con

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

donde γ es un vector $(Q-1)\times 1$ de parámetros que miden el impacto marginal de las variables explicativas de observaciones vecinas (áreas) sobre la variable dependiente Y. $\mathbf{W}\bar{\mathbf{X}}$ produce variables explicativas retardadas espacialmente que reflejan un promedio de observaciones vecinas.

Una motivación para el uso de SDM se basa en dos circunstancias que suelen surgir cuando aplicamos un modelo de regresión espacial a datos de área. Una de esas circunstancia es la dependencia espacial en las perturbaciones de un modelo de regresión MCO (mínimos cuadrados ordinarios). La segunda circunstancia es la existencia de una variable explicativa omitida que muestra una covarianza no nula con una variable incluida en el modelo, y las variables omitidas son probables

cuando se trata de muestras de datos de área. También es importante este modelo de Durbin porque generaliza muchos de los modelos que se utilizan:

- Si imponemos $\gamma = 0$ nos lleva a un modelo tipo SAR (4.5).
- Si imponemos $\gamma = -\rho\bar{\beta}$ nos lleva a un modelo tipo SEM (4.7).
- La restricción $\rho = 0$ da como resultado un modelo de regresión $\mathbf{X}$ espacialmente retardados que asume la independencia entre las observaciones de la variable dependiente, pero incluye características de las áreas vecinas, en forma de variables explicativas espacialmente retardadas.
- Por último, si $\rho = 0$ y $\gamma = 0$ obtenemos el modelo de regresión estándar dado por (4.3).

Si definimos $\mathbf{Z} = [\mathbf{X} \quad \mathbf{W}\bar{\mathbf{X}}]$ y $\delta = \begin{bmatrix}\beta \\ \gamma\end{bmatrix}$ se puede observar de forma más evidente que es un modelo tipo SAR (4.4) quedando el modelo escrito de la siguiente forma:

$$Y = \rho\mathbf{W}Y + \mathbf{Z}\delta + \varepsilon \tag{4.9}$$

que si agrupamos nos queda:

$$Y = (I - \rho\mathbf{W})^{-1}(\mathbf{Z}\delta + \varepsilon)$$

Finalmente, podemos generalizar más el modelo espacial de Durbin de la siguiente manera:

$$Y = \rho\mathbf{W_1}Y + \mathbf{X}\beta + \mathbf{W_1}\bar{\mathbf{X}}\gamma + \varepsilon$$

$$\varepsilon = \lambda\mathbf{W_2}\varepsilon + u$$

$$u \sim \mathcal{N}(0, \sigma_u^2 I)$$

donde $\mathbf{W_1}$ y $\mathbf{W_2}$ son matrices $n \times n$ de pesos espaciales que pueden ser iguales o distintas.

4.2. Pruebas para Dependencia Espacial

La presencia de dependencia espacial en un modelo de regresión se puede detectar mediante pruebas de diagnóstico. A continuación se incluyen tres pruebas para detectar la dependencia espacial en términos de error: el test de Moran y los test de los multiplicadores de Lagrange para los modelos espaciales de retardo (4.4) y de error (4.7) visto en la sección anterior.

- El estadístico más conocido para hacer el contraste de hipótesis de existencia de autocorrelación espacial es el estadístico I de Moran, definido anteriormente por la ecuación (3.1), aplicada a los residuos de regresión. Si denotamos a e como el vector $n \times 1$ de mínimos cuadrados ordinarios (MCO) definido como $e = Y - \mathbf{X}\widehat{\beta}$, el estadístico I de Moran queda redefinido de la siguiente forma:

$$I = \frac{n}{W_0}\frac{e'\mathbf{W}e}{e'e}$$

con W_0 el vector normalizado que aparece en el estadístico I de Moran, $W_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$ y $e'e$ la suma de los cuadrados de los residuos. El estadístico I se interpreta como el coeficiente de una regresión de MCO de $\mathbf{W}e$ sobre e.

En la práctica, la inferencia mediante la prueba I de Moran, se basa en una aproximación normal utilizando un valor estandarizado obtenido restando la media, bajo la hipótesis nula de no dependencia espacial, y dividiendo por la raíz cuadrada de la varianza. El test estadístico queda:

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}} \sim \mathcal{N}(0,1)$$

donde el valor esperado y la varianza del estadístico I de Moran están expresados con más detalle en la sección (3.1)

La causa de la dependencia espacial bajo la hipótesis alternativa no está especificada, por lo tanto, la prueba de Moran es una prueba general para detectar autocorrelación espacial. Al aplicar esta prueba a los residuos, debemos tener cuidado, ya que si estimamos los Q coeficientes de regresión, los residuos observados están sujetos a Q restricciones lineales, lo cual quiere decir que dichos residuos estarán correlacionados y el procedimiento de prueba mediante I de Moran no será válido. Sin embargo, en caso de que el número de coeficientes Q fuese pequeño en comparación con n (número de áreas) podríamos ignorar esta correlación y de este modo validar la prueba.

- Una prueba alternativa se basa en el principio del multiplicador de Lagrange (LM), sugerido por Burridge [4], que se calcula a través de los residuos MCO y que es diferente según qué tipo de dependencia espacial queramos probar. Para una dependencia de tipo error dado por la ecuación (4.7), el estadístico LM(error) viene dado por:

$$LM(error) = \left(\frac{e^t \mathbf{W} e}{\frac{1}{n} e^t e} \right)^2 \frac{1}{tr[\mathbf{W}^t \mathbf{W} + \mathbf{W^2}]}$$

donde tr es el operador traza y $\frac{1}{n} e^t e$ representa la varianza muestral de error.

Con el factor de normalización incluido se logra una distribución asintótica chi-cuadrado con un grado de libertad (χ^2_1) bajo la hipótesis nula de no dependencia espacial. Esto es, a partir del siguiente contraste de hipótesis

$$\begin{cases} H_0 : \lambda = 0 \\ H_1 : \lambda \neq 0 \end{cases}$$

tomamos la decisión de rechazar H_0 si $LM(error) > \chi^2_{1,1-\alpha}$ (α, nivel de signficación).

Bajo la hipótesis alternativa, el logaritmo de la función de verosimilitud de los datos viene dado por:

$$\begin{aligned} L(y|x, \lambda, \beta, \sigma^2) = \; & -\frac{n}{2} ln(2\pi\sigma^2) + ln|I - \lambda \mathbf{W}| \\ & -\frac{1}{2\sigma^2} (y - x\beta)^t (I - \lambda \mathbf{W})^t (I - \lambda \mathbf{W})(y - x\beta) \end{aligned}$$

de modo que la imposición de $\lambda = 0$ produce la función de verosimilitud restringida:

$$L^R() = L() - \alpha\lambda$$

donde α es un parámetro que debe ser estimado.

Las condiciones de primer orden para conseguir el máximo de la función de verosimilitud restringida están dados por:

$$\begin{aligned}
\partial L^R/\partial\beta^t &= \frac{1}{\sigma^2}x^t(I-\lambda\mathbf{W})^t(I-\lambda\mathbf{W})(y-x\beta)=0\\
\partial L^R/\partial\sigma^2 &= -\frac{1}{2\sigma^2}\left[n+\sigma^2(y-x\beta)^t(I-\lambda\mathbf{W})^t(I-\lambda\mathbf{W})(y-x\beta)\right]=0\\
\partial L^R/\partial\lambda &= -\alpha-\sum_{j=1}^{n}\delta_j(1-\lambda\delta_j)^{-1}+\\
&\quad +\frac{1}{2\sigma^2}(y-x\beta)^t\left[(I-\lambda\mathbf{W})^t\mathbf{W}+\mathbf{W}^t(I-\lambda\mathbf{W})(y-x\beta)\right]=0
\end{aligned}$$

donde $\delta_j,\ j=1,\cdots,n$ son los autovalores de la matriz $\mathbf{W}$ y de las ecuaciones anteriores sacamos los estimadores para α y σ^2, inyectando $\hat{\beta}$ el estimador de MCO:

$$\hat{\sigma^2}=\frac{1}{n}(y-x\hat{\beta})^t(y-x\hat{\beta})=\frac{e^te}{n}$$

$$\hat{\alpha}=-tr(\mathbf{W})+\frac{1}{2\sigma^2}(y-x\hat{\beta})^t(\mathbf{W}+\mathbf{W}^t)(y-x\hat{\beta})=\frac{ne^t(\mathbf{W}+\mathbf{W}^t)e}{2e^te}$$

Se puede demostrar fácilmente que bajo la hipótesis H_0, la matriz de información es diagonal y:

$$-E\left(\frac{\partial^2 L}{\partial\lambda\partial\lambda}\right)\Big|_{\lambda=0}=tr(\mathbf{W^2}+\mathbf{W^tW})$$

Por lo tanto, el test puede ser construido tratando a $\dfrac{\hat{\alpha}}{(tr(\mathbf{W^2}+\mathbf{W^tW}))^{-\frac{1}{2}}}$ como una desviación normal estándar.

- Una prueba de dependencia espacial sustantiva (es decir, para un modelo de retardo espacial omitido) la prueba toma la forma:

$$LM(lag)=\left(\frac{e^t\mathbf{W}e}{\frac{1}{n}e^te}\right)^2\frac{1}{H}$$

siendo

$$H = \{(\mathbf{WX}\widehat{\beta})^t[I - \mathbf{X}(\mathbf{X^tX})^{-1}\mathbf{X}]^t(\mathbf{WX}\widehat{\beta})\widehat{\sigma}^{-2}\} + tr(\mathbf{W^tW} + \mathbf{W^2})$$

donde $\widehat{\beta}$ y $\widehat{\sigma}$ denotan los MCO estimados. La prueba $LM(lag)$ sigue también una distribución asintótica chi-cuadrado con un grado de libertad (χ^2_1) bajo la hipótesis nula de no dependencia espacial ($H_0 : \rho = 0$).

Esto es, a partir del siguiente contraste de hipótesis

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases} \tag{4.10}$$

tomamos la decisión de rechazar H_0 si $LM(lag) > \chi^2_{1,1-\alpha}$ (α, nivel de signficación).

4.3. Estimación de modelos espaciales

Para estimar los modelos de regresión espacial se utiliza el método de máxima verosimilitud (MV), el cual se basa en maximizar la probabilidad de distribución conjunta con respecto a unos parámetros relevantes. Este método (MV) tiene propiedades teóricas asintóticas interesantes tales como consistencia, eficiencia o normalidad asintótica, y también se considera robusto para pequeñas variaciones de la suposición de normalidad.

Primero veamos este método en los modelos tipo SAR, que partiremos a partir de la solución de la variable dependiente (Y), que tiene como ecuación (4.6)

Para simplificar las sucesivas expresiones notaremos $\mathbf{A} = I - \rho\mathbf{W}$. Sabemos que ε_i son independientes e idénticamente distribuidos a una distribución normal, $\mathcal{N}(0, \sigma^2)$, por tanto con fdd:

$$\phi(s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}s^2\right)$$

En nuestro caso, el vector ε sigue una distribución normal multivariante y denotando la matriz de covarianza $\Sigma = \sigma^2 I$. Se puede ver que $|\Sigma|^{\frac{-1}{2}} = \sigma^n$, $\Sigma^{-1} = \frac{1}{\sigma^2}I$ y $|\Sigma| = \sigma^{2n}$ y por tanto la distribución de ε queda:

$$\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

Por tanto dado que

$$Y = \mathbf{A^{-1}X}\beta + \mathbf{A^{-1}}\varepsilon$$

entonces

$$Y \sim \mathcal{N}_n(\mathbf{A^{-1}}\mathbf{X}\beta, \sigma^2\mathbf{A^{-1}}(\mathbf{A^{-1}})^t)$$

En consecuencia, la función de verosimilitud queda:

$$\begin{aligned} L(\rho,\beta,\sigma^2) &= (2\pi)^{\frac{-n}{2}}|\sigma^2\mathbf{A^{-1}}(\mathbf{A^{-1}})^t|^{\frac{-1}{2}}\exp\left(-\frac{1}{2\sigma^2}(Y-\mathbf{A^{-1}}\mathbf{X}\beta)^t\mathbf{A}^t\mathbf{A})(Y-\mathbf{A^{-1}}\mathbf{X}\beta)\right) \\ &= (2\pi\sigma^2)^{\frac{-n}{2}}|\mathbf{A}|\exp\left(\frac{1}{2}(\mathbf{A}Y-\mathbf{X}\beta)^t\Sigma^{-1}(\mathbf{A}Y-\mathbf{X}\beta)\right) \end{aligned}$$

Como el modelo espacial de Durbin (ver sección 4.1.4) puede ser escrito como un modelo tipo SAR dado por la ecuación (4.9), sea $\varepsilon \sim \mathcal{N}(0,\sigma^2)$, el logaritmo de la función de probabilidad de la ecuación del modelo SDM queda:

$$\begin{aligned} ln\, L(\rho,\delta,\sigma^2) = \; & -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) + ln|\mathbf{A^*}| \\ & -\frac{1}{\sigma^2}(\mathbf{A^*}Y-\mathbf{Z}\delta)^t(\mathbf{A^*}Y-\mathbf{Z}\delta) \end{aligned} \tag{4.11}$$

donde n es el número de observaciones (áreas) y hemos denominado $\mathbf{A^*} = I - \rho\mathbf{W}$. Por tanto los parámetros a partir de los cuales vamos a maximizar esta probabilidad son ρ, δ y σ^2. Podemos poner los estimadores de los coeficientes regresivos δ y σ^2 en función de ρ de la manera siguiente:

$$\hat{\delta} = \delta_o - \rho\delta_L$$

$$\hat{\sigma^2} = (e_o - \rho e_L)^t(e_o - \rho e_L)\frac{1}{n}$$

donde δ_o y δ_L son los coeficientes de regresión de MCO en una regresión de $\mathbf{Z}$ y $\mathbf{W}Y$ mientras que e_o y e_L son los vectores residuales en la regresión de δ_o y δ_L. En forma de ecuación se espresa como: $\delta_o = (\mathbf{Z^t}\mathbf{Z})^{\mathbf{-1}}\mathbf{Z^t}Y$, $\delta_L = (\mathbf{Z^t}\mathbf{Z})^{\mathbf{-1}}\mathbf{Z^t}\mathbf{W}Y$ $e_o = Y - \mathbf{Z}\delta_o$ y $e_L = \mathbf{W}Y - \mathbf{Z}\delta_L$.

Esto nos permite usar la función de verosimilitud restringida que nos proporcionará el mismo estimador para ρ que maximizando la función de verosimilitud, y por tanto también nos dará el mismo estimador de β y σ^2. La motivación para su uso es simplificar el problema de optimización reduciendo el problema de una optimización multivariable a uno univariable al igual que en el modelo SAR (4.4).

Sustituyendo estas ecuaciones en (4.11) nos da el valor el logaritmo de la función verosimilitud concentrada:

$$ln\, L_{res}(\rho) = k + ln|I - \rho\mathbf{W}| - \frac{n}{2} ln[(e_o - \rho e_L)^t (e_o - \rho e_L)]$$

donde k es una constante que no depende de ρ.

Sea $S(\rho)$ el término de la expresión de la función $ln\, L_{con}(\rho)$, $(e_o - \rho e_L)^t (e_o - \rho e_L) = e_o^t e_o - 2\rho e_o^t e_L + \rho^2 e_L^t e_L$. Para simplificar la optimización de $ln\, L_{con}(\rho)$ con respecto al parámetro ρ, Pace y Barry (1997) proponen evaluar dicha función usando un vector de valores para ρ:

$$\begin{pmatrix} lnL(\rho_1) \\ lnL(\rho_2) \\ \vdots \\ lnL(\rho_q) \end{pmatrix} = k + \begin{pmatrix} ln|I - \rho_1\mathbf{W}| \\ ln|I - \rho_2\mathbf{W}| \\ \vdots \\ ln|I - \rho_q\mathbf{W}| \end{pmatrix} - (n/2) \begin{pmatrix} ln(S(\rho_1)) \\ ln(S(\rho_2)) \\ \vdots \\ ln(S(\rho_q)) \end{pmatrix}$$

Se puede observar que la matriz de pesos $\mathbf{W}$ es uno de los argumentos que, en principio, más pueden dificultar los cálculos de la optimización al ser una matriz $n \times n$. Sin embargo, al ser una matriz de pesos construida a partir de la contigüidad de las áreas, supone que tendrá bastantes entradas nulas. Esto es, realizaremos los cálculos con una matriz hueca o dispersa con lo que su tiempo de operación será lineal ($O(n)$), mientras que para una matriz densa es de ($O(n^2)$).

Por último, veamos la estimación de los parámetros para los modelos SEM cuya ecuación viene dada por (4.7). De la misma forma a partir del método de máxima verosimilitud, asumiendo normalidad en los términos error (ε) y usando el Jacobiano, el logaritmo de la función de probabilidad es:

$$\begin{aligned} ln\, L(\lambda, \beta, \sigma^2) = \; & -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) + ln|I - \lambda\mathbf{W}| \\ & -\frac{1}{2\sigma^2} (Y - \mathbf{X}\beta)^t (I - \lambda\mathbf{W})^t (I - \lambda\mathbf{W})(Y - \mathbf{X}\beta) \end{aligned} \tag{4.12}$$

Si observamos el último término de la ecuación anterior, observamos que el máximo de la función de máxima verosimilitud es equivalente al mínimo de la suma de los residuos al cuadrado de una regresión donde la variable espacialmente dependiente es $Y^* = Y - \lambda\mathbf{W}Y$ en un conjunto de variables explicativas dada por $\mathbf{X}^* = \mathbf{X} - \lambda\mathbf{W}\mathbf{X}$.

Repitiendo el proceso de estimación que utilizamos para los modelos SAR, estimamos β y σ^2 por el método MV:

$$\hat{\beta}_{MV} = [\mathbf{X^t}(I - \lambda\mathbf{W})^t (I - \lambda\mathbf{W})\mathbf{X}]^{-1} \mathbf{X^t} (I - \lambda\mathbf{W})^t (I - \lambda\mathbf{W}) Y$$

$$\hat{\sigma^2}_{MV} = \frac{1}{n} (e - \lambda\mathbf{W}e)^t (e - \lambda\mathbf{W}e)$$

donde $e = Y - \mathbf{X}\hat{\beta}_{MV}$. Para conseguir un estimador consistente de λ, podríamos hacerlo por un método numérico. Para más detalles sobre la estimación de los parámetros para los diferentes modelos de regresión espacial ver *Anselin (1988b)* y el capítulo 4 de LeSage y Pace [9].

Capítulo 5

Implementación de datos espaciales en R

En este último capítulo nos centraremos en ver los modelos de regresión espacial, estudiados en el capítulo anterior, en el programa estadístico R. Analizaremos los paquetes que deben instalarse y las funciones más importantes que utilizaremos, así como algún ejemplo con datos reales con la ayuda de Bivand y otros [1] y la página web The Geospatial and Farming Systems Research Consortium [15]. Los paquetes más esenciales y usados para los datos espaciales de tipo área son *rgdal, raster, sp* y *spdep*. Los tres primeros paquetes son necesarios para utilizar las funciones del paquete *spdep*, el cual, como veremos a continuación, es el que visualmente más se usa para el análisis espacial. Es decir, para utilizar las funciones que analizarán la autocorrelación espacial y los distintos modelos vistos en el capítulo 3 y 4, respectivamente, debemos cargar los demás paquetes y librerias.

5.1. Paquetes previos

- **Paquete *rgdal***

  Este paquete proporciona enlaces a la librería de abstracción de datos geoespaciales ('GDAL') y da acceso a las operaciones de proyección/transformación desde la librería 'PROJ.4'. Su uso más general es leer los datos para devolver un objeto espacial y proporcionar la proyección de coordenadas o matrices de coordenadas. Para más detalles sobre el uso de este paquete en datos espaciales ver Bivand [2].

- **Paquete *ráster***

  El paquete ráster proporciona clases y funciones para manipular datos geográficos (espaciales) en formato 'raster'. Los datos ráster dividen el espacio en

celdas (rectángulos, píxeles) de igual tamaño (en unidades del sistema de referencia de coordenadas). Dichos datos espaciales continuos también se conocen como datos de "cuadrícula" y se contrastan con datos espaciales discretos (basados en objetos como puntos, líneas o polígonos). El paquete es útil cuando se usan conjuntos de datos muy grandes que no se pueden cargar en la memoria de la computadora. Las funciones se ejecutarán correctamente, ya que procesan archivos grandes en trozos, es decir, leen, computan y escriben bloques de datos, sin cargar todos los valores en la memoria a la vez.

El paquete implementa clases para datos ráster y su uso es muy variado:

- Creación de objetos ráster desde cero o desde archivo.
- Manejo de archivos ráster computacionalmente grandes.
- Álgebra ráster y funciones de superposición.
- Funciones de distancia y vecindad.
- Conversión de polígonos, líneas y puntos a datos ráster.
- Modelo de predicciones.
- Resumir valores ráster.
- Fácil acceso a valores de celda rasterizados.
- Dibujar (hacer mapas).
- Cálculo de los números de fila, columna y celda a coordenadas y viceversa.
- Leer y escribir varios tipos de archivos ráster.

Con ayuda de Hijmans [7] vemos las diferentes funciones y sus usos con mayor detalle.

- **Paquete *sp***

  Clases y métodos para datos espaciales. El documento de clases donde reside la información de ubicación espacial, para datos 2D o 3D. Se proporcionan funciones de utilidad, por ejemplo, para trazar datos como mapas, selección espacial, así como métodos para recuperar coordenadas, subconjuntos, impresión, resumen, etc. Este paquete no proporciona muchas funciones para modificar o analizar datos espaciales, pero las clases que se definen son muy utilizadas por otros muchos paquetes de R. El paquete *sp* introduce una serie de clases: Para datos vectoriales, los tipos básicos son `SpatialPoints, SpatialLines y SpatialPolygons`. Para almacenar también los atributos,

las clases están disponibles con estos nombres más DataFrame, por ejemplo, `SpatialPolygonsDataFrame y SpatialPointsDataFrame`. Cuando se hace referencia a cualquier objeto con un nombre que comienza con Spatial, es común escribir Spatial*. Cuando se hace referencia a un objeto `SpatialPolygons o SpatialPolygonsDataFrame`, es común escribir `SpatialPolygons*`.

Para más información sobre este paquete y la utilidad de las funciones mencionadas se recomienda ver Pebesma y otros [11] y Bivand y otros [1].

5.2. Paquete *spdep*

Es uno de los paquetes más importantes a la hora del análisis exploratorio y la inferencia estadística de los datos espaciales. Posee una colección de funciones para crear matrices de pesos espaciales a partir de la contigüidad de los polígonos, para el estadístico general de productos cruzados, para pruebas para el contraste de existencia de autocorrelación espacial como el estadístico I de Moran o c de Geary y sus versiones locales (ver capítulos 2 y 3). También contiene funciones para estimar modelos del tipo SAR o SEM (ver capítulo 4). A continuación veremos las funciones en R más relevantes para llevar a cabo este análisis e inferencia de datos de área. Para mayor detalle de las funciones, a continución descritas y otras funciones que incluye el paquete, se recomienda Bivand y Wong [3].

5.2.1. Vecindad de datos de área

Como hemos visto en teoría, antes de poder hacer un contraste de hipótesis de autocorrelación espacial o modelar los datos, debemos disponer de una organización de dichos datos, es decir, debemos tener en cuenta la contigüidad que se utiliza y a partir de ello, y de los datos, crear una matriz de pesos para su posterior uso en la inferencia estadística de los datos de área. Para conseguir estas matrices de peso, dentro del paquete *spdep* usamos la clase *nb*. Veamos las funciones más frecuentemente usadas en R de dicha clase para crear matrices de pesos a partir de unos datos:

- **poly2nb**: La función crea una lista de vecinos basada en regiones con límites contiguos, que comparte uno o más puntos de límite.

  Uso

  ```
  poly2nb(pl, row.names = NULL, snap=sqrt(.Machine$double.eps),
  queen=TRUE,...)
  ```

Argumentos

`pl`: Lista de polígonos de la clase *SpatialPolygons*.

`row.names`: Vector de identidad de las regiones para ser añadidas a la lista de los vecinos.

`snap`: Los puntos de límites inferiores a la distancia de separación para indicar la contigüidad.

`queen`: Si TRUE el tipo de contigüidad que se utiliza es la de tipo reina. Si FALSE, la contigüidad utlizada es la de tipo torre.

Valor

Nos devuelve una lista de vecinos de la clase *nb*.

- **knearneigh**: La función devuelve una matriz con los índices de puntos que pertenecen al conjunto de los k vecinos más cercanos entre sí.

  Uso

  ```
  knearneigh(x, k=1, longlat = NULL,...)
  ```

  Argumentos

  `x`: Matriz de puntos coordenados o un objeto de *SpatialPoints*.

  `k`: Número de vecinos más cercanos para cada área.

  `longlat`: TRUE si las coordenadas del punto son grados decimales de longitud-latitud, en cuyo caso las distancias se miden en kilómetros; si x es un objeto *SpatialPoints*, el valor se toma del propio objeto.

  Valor

  Una lista con los siguientes elementos:

  `nn`: Matriz entera de identificadores de número de región.

  `np`: Número de puntos de entrada.

  `k`: Valor de entrada requerido.

  `dimension`: Número de columnas de x

  `x`: Coordenadas de entrada

- **knn2nb**: La función convierte un objeto *knn* devuelto por *knearneigh* en una lista de vecinos de clase *nb* con una lista de vectores de números enteros que contienen identificadores de número de región vecina.

  Uso

  ```
  knn2nb(knn, row.names = NULL, sym = FALSE)
  ```

Argumentos

`knn`: Un objeto *knn* devuelto po la función *knearneigh*.

`sym`: Fuerza a la lista de regiones vecinas de salida a ser simétrica.

- **nbdists**: Dada una lista de enlaces vecinos adyacentes (una lista de vecinos de tipo de objeto *nb*), la función devuelve las distancias euclideas a lo largo de los enlaces en una lista de la misma forma que la lista de vecinos. Si longlat = TRUE, se usan las distancias geodésicas (great-circle distance).

  Uso

  ```
  nbdists(nb, coords, longlat = NULL)
  ```

- **nb2listw**: Esta función toma un objeto de lista de vecinos y lo convierte en un objeto o matriz de pesos. El estilo de conversión predeterminado es **W**, donde los pesos para cada entidad de área se estandarizan para sumar por filas igual a la unidad, es decir, está estandarizado por filas.

  Uso

  ```
  nb2listw(neighbours, glist=NULL, style="W",...)
  ```

  Argumentos

  `neighbours`: Un objeto de clase *nb*.

  `glist`: Lista de pesos generales correspondientes a vecinos.

  `style`: A partir de una lista binaria de vecinos, en la que las regiones están listadas como vecinas o están ausentes (por lo tanto, no están en el conjunto de vecinos para alguna definición), la función agrega una lista de ponderaciones con valores dados por el estilo de esquema de codificación elegido. **B** es la codificación binaria básica mientras que **W** está estandarizada por filas. Hay más opciones para el esquema de codificación que no utilizaremos puesto que no lo hemos visto en teoría.

5.2.2. Pruebas para Dependencia Espacial

En la parte de teoría del trabajo hemos detallado las pruebas mediante los estadísticos globales I de Moran y c de Geary, así como sus formas locales y los estadísticos locales de Getis y Ord. En la práctica y, en concreto, en el paquete de R que estamos utilizando, el más usado es el estadístico global I de Moran ya que es efectivo en la mayoría de las situaciones que se pueden plantear. Por este motivo, analizaremos detalladamente las funciones del paquete *spdep* que utilizan este estadístico para la prueba de dependencia espacial. También indicaremos algunas funciones de otros estadísticos pero sin entrar en mucho detalle ya que su uso será similar al del estadístico I.

- **moran.test**: La prueba de Moran para la autocorrelación espacial utilizando una matriz de ponderaciones espaciales en forma de lista de ponderaciones. Las hipótesis que subyacen a la prueba son sensibles a la forma de la gráfica de relaciones entre vecinos y otros factores, y los resultados pueden compararse con la función de "moran.mc"basada en permutaciones.

  Uso

  ```
  moran.test(x, listw, randomisation=TRUE,
  alternative="greater", rank = FALSE,...)
  ```

  Argumentos

  `x`: Vector numérico de la misma longitud que la lista de vecinas de *listw*.

  `litw`: Un objeto tipo *listw* creado por ejemplo por *nb2listw*.

  `randomisation`: Varianza de I calculada bajo la hipótesis de aleatoriedad. Si FALSE, la hipótesis utilizada será normalidad.

  `alternative`: Una cadena de caracteres que especifique la hipótesis alternativa, debe ser una de mayor (por defecto), menor o doble.

  `rank`: Valor lógico. Si FALSE utiliza variables continuas. Si TRUE utiliza la adaptación de I de Moran para los rangos sugeridos por Cliff y Ord.

  Valor

  La función nos devuelve el nombre del método y el de los datos usados, asi como una descripción de la hipótesis alternativa. Nos devuelve el valor del estadístico I y el p-valor del test para saber cuál es el resultado del contraste.

- **lm.morantest**: La prueba I de Moran para autocorrelación espacial en residuos de un modelo lineal estimado.

  Uso

  ```
  lm.morantest(model, listw, alternative = "greater",
  resfun=weighted.residuals,....
  ```

  Argumentos

  `model`: Un objeto de la clase *lm* devuelto por *lm*. Los pesos pueden especificarse en el ajuste de *lm*, pero no se deben usar las compensaciones.

  `resfun`: Por defecto: *weighted.residuals*. La función se utiliza para extraer los residuos del objeto *lm*.

  Valor

  Al igual que la función anterior, esta función nos devolverá los nombres de los métodos y datos utilizados, el p-valor, y el estadístico I, su esperanza y varianza.

- **moran.mc**: Una prueba de permutación para el estadístico I de Moran calculado usando `nsim` permutaciones aleatorias de x para el esquema de ponderación espacial dado, para establecer el rango de la estadística observada en relación con los valores simulados `nsim`.

  Uso

  ```
  moran.mc(x, listw, nsim, alternative="greater",...)
  ```

  Argumentos

  Son los mismos argumentos de la función *moran.test*.

- **localmoran**: El estadístico espacial local I de Moran se calcula para cada zona según el objeto de ponderaciones espaciales utilizado.

  Uso

  ```
  localmoran(x, listw, alternative = "greater",
  p.adjust.method="none", mlvar=TRUE)
  ```

  Argumentos

  `mlvar`: Predeterminado TRUE: los valores de la I de Moran local se informan utilizando la varianza de la variable de interés (suma de desviaciones al cuadrado sobre n), pero se pueden informar como la varianza de la muestra, dividiendo por $(n-1)$ en su lugar.

  `adjust.x`: Si TRUE, los valores de las observaciones de x que no tengan vecinos se omiten para calcular la media de x.

A la hora de utilizar cualquiera de estas funciones, no es necesario usar todos los argumentos, es decir, en cualquiera de ellos los fundamentales son los datos que vamos a utilizar para hacer la inferencia y el tipo de matriz de pesos que vamos a usar, x y *listw* respectivamente.

Para el estadístico c de Geary, el paquete *spdep* también proporciona unas funciones para su utilización. Estas son *geary.test* y *geary.mc* que, como en el caso de I, la primera función es para realizar un test para la autocorrelación espacial y la segunda a partir de un muestreo aleatorio mediantes el enfoque de Monte Carlo para tomar varias matrices de pesos y realizar su posterior análisis.

5.2.3. Modelos de Regresión Espacial en R

En este último apartado del capítulo vamos a aplicar los modelos de regresión, que hemos visto en la parte de teoría, a los datos en el programa R. Dentro del paquete *spdep* podemos encontrar la función *lagsarlm* que se utiliza tanto para el modelo tipo SAR dado por la ecuación (4.4) como para el modelo espacial de

Durbin, ecuación (4.8), ya que el modelo SDM se puede expresar como un modelo tipo SAR. También estudiaremos con detalle la función *errorsarlm* que la usaremos para el modelo tipo SEM dado por la ecuación (4.7).

- **lagsarlm**: Estimación del modelo de regresión espacial de retardo.

  Uso

  ```
  lagsarlm(formula, data = list(), listw, Durbin, method= "___",
  interval=NULL, tol.solve=1.0e-10,...)
  ```

  Argumentos

  `formula`: Una descripción simbólica del modelo a ser ajustado. Los detalles de la especificación del modelo se dan para *lm()*.

  `data`: Un marco de datos opcional que contiene las variables en el modelo. Por defecto, las variables se toman del entorno al que se llama la función.

  `listw`: Una lista de datos creados por ejemplo por *nb2listw*.

  `Durbin`: Si FALSE, el modelo utilizado será el de retraso. Si TRUE, el modelo que utiliza es el de Durbin.

  `method`: El método predeterminado de *method = "eigen"* usa valores propios y, por lo tanto, también puede establecer los límites inferior y superior para la búsqueda de líneas para ρ pero es no es factible para grandes n. El otro método utilizado es *method = "Matrix"* para calcular directamente el determinante de la matriz $I - \rho\mathbf{W}$.

  `interval`: Intervalo de búsqueda para el parámetro autorregresivo ρ.

  `tol.solve`: La tolerancia para detectar dependencias lineales en las columnas de matrices que se van a invertir (por defecto = 1.0e-10).

- **errorsarlm**: Estimación del modelo de regresión espacial de error.

  Uso

  ```
  errorsarlm(formula, data=list(), listw, method="___",
  interval = NULL, tol.solve=1.0e-10,...)
  ```

  Argumentos

  Al ser los mismos argumentos que para el caso de retardo espacial, omitimos su detalle.

Como hemos especificado anteriormente, no es necesario utilizar todos los argumentos descritos en ambos modelos. En el ejemplo que daremos a continuación veremos que a la hora de estudiar los modelos y estimar sus coeficientes utilizaremos los siguientes argumentos: *formula, data, listw* y *tol.solve*.

5.3. Ilustración de los Modelos en R

El conjunto de datos que vamos a utilizar para nuestro ejemplo es `houses2000`[1] que son datos de las viviendas de California del Censo 2000. En este conjunto de datos hay 7049 casos con 29 variables:

- **TRACT**: Identificador del distrito censal
- **GEOID**: Identificador geográfico
- **label**: Etiqueta de identificación
- **HouseValue**: Valor de la vivienda
- **NhousingUn**: Número de viviendas
- **RecHouses**: Número de casas para uso recreativo
- **NMobileHom**: Número de casas móviles
- **NBadPlumbi**: Número de casas con plomería incompleta
- **NBadKitche**: Número de viviendas con cocinas incompletas
- **Population**: Población total
- **Males**: Número de hombres
- **Females**: Número de mujeres
- **Under5**: Número de personas menores de cinco años
- **White**: Número de personas que se identifican como blancas (sólo)
- **Black**: Número de personas que se identifican como afro-ameriano (sólo)
- **AmericanIn**: Número de personas que se identifican como indio americano (sólo)
- **Asian**: Número de personas que se identifican como asiático americano (sólo)
- **Hispan**: Número de personas que se identifican como hispanas (solo)
- **PopInHouse**: Número de personas que viven en hogares
- **nHousehold**: Número de hogares

[1] Datos *house2000*. Fuente: *package spdep*

- **Families**: Número de familias
- **yearBuilt**: Año de construcción de la vivienda
- **nRooms**: Número medio de habitaciones/estancias por casa
- **nBedrooms**: Número medio de habitaciones por casa
- **MedHHinc**: Ingreso medio del hogar
- **MedianAge**: Edad media de la población
- **householdS**: Tamaño medio del hogar
- **familySize**: Tamaño medio de la familia
- **County**: Condado

Comenzaremos viendo los paquetes y librerías que debemos de cargar para poder utilizar este conjunto de datos.

```
>install.packages('devtools')
>install.packages('spdep','rgdal','raster')
>if(!require("rspatial"))devtools::install_github('rspatial/rspatial')

>library(spdep)
>library(rgdal)
>library(rspatial)
>library(rgeos)
```

Cargamos los datos y visualizamos algunos datos como la dimensión y un resumen de las variables de los datos:

```
>h <- sp_data('houses2000')
> dim(h)


[1] 7049   29


> summary(h)
```

```
Object of class SpatialPolygonsDataFrame
Coordinates:
         min        max
x -124.40959 -114.13443
y   32.53416   42.00952
Is projected: FALSE
proj4string :
[+proj=longlat +datum=WGS84 +ellps=WGS84 +towgs84=0,0,0]
Data attributes:
    TRACT               GEOID              label        houseValue
 Length:7049        Length:7049    Length:7049  Min.   :      0
 Class :char        Class :char    Class :char  1st Qu.: 139700
 Mode  :char        Mode  :char    Mode  :char  Median : 188900
                                                Mean   : 243350
                                                3rd Qu.: 296500
                                                Max.   :1000001

 nhousingUn     recHouses        nMobileHom        yearBuilt
 Min.   :   0   Min.   :   0.0   Min.   :   0.00   Min.   :   0
 1st Qu.:1176   1st Qu.:   1.0   1st Qu.:   0.00   1st Qu.:1958
 Median :1602   Median :   4.0   Median :   3.00   Median :1968
 Mean   :1733   Mean   :  33.6   Mean   :  76.38   Mean   :1959
 3rd Qu.:2153   3rd Qu.:  10.0   3rd Qu.:  55.00   3rd Qu.:1976
 Max.   :9905   Max.   :5311.0   Max.   :3606.00   Max.   :1999

 nBadPlumbi        nBadKitche          nRooms       nBedrooms
 Min.   :   0.00   Min.   :   0.00   Min.   :0.00   Min.   :0.00
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:4.00   1st Qu.:1.89
 Median :   8.00   Median :  10.00   Median :4.90   Median :2.36
 Mean   :  15.67   Mean   :  22.62   Mean   :4.80   Mean   :2.36
 3rd Qu.:  20.00   3rd Qu.:  26.00   3rd Qu.:5.60   3rd Qu.:2.83
 Max.   :1109.00   Max.   :2257.00   Max.   :9.10   Max.   :4.78

 medHHinc        Population        Males          Females
 Min.   :     0   Min.   :    0   Min.   :    0   Min.   :    0
 1st Qu.: 33856   1st Qu.: 3418   1st Qu.: 1681   1st Qu.: 1721
 Median : 46597   Median : 4563   Median : 2254   Median : 2304
 Mean   : 51249   Mean   : 4805   Mean   : 2394   Mean   : 2411
 3rd Qu.: 63310   3rd Qu.: 5948   3rd Qu.: 2950   3rd Qu.: 2994
 Max.   :197012   Max.   :36146   Max.   :25573   Max.   :12082
```

```
     Under5         MedianAge          White           Black
 Min.   :   0.0   Min.   : 0.00   Min.   :    0   Min.   :   0.0
 1st Qu.: 202.0   1st Qu.:29.30   1st Qu.: 1712   1st Qu.:  45.0
 Median : 312.0   Median :34.00   Median : 2645   Median : 112.0
 Mean   : 352.8   Mean   :34.31   Mean   : 2861   Mean   : 321.2
 3rd Qu.: 455.0   3rd Qu.:38.70   3rd Qu.: 3721   3rd Qu.: 307.0
 Max.   :4230.0   Max.   :84.80   Max.   :23020   Max.   :6525.0

   AmericanIn          Asian           Hispanic       PopInHouse
 Min.   :   0.00   Min.   :   0.0   Min.   :    0   Min.   :    0
 1st Qu.:  18.00   1st Qu.:  98.0   1st Qu.:  414   1st Qu.: 3343
 Median :  35.00   Median : 265.0   Median :  974   Median : 4466
 Mean   :  47.29   Mean   : 524.5   Mean   : 1556   Mean   : 4689
 3rd Qu.:  60.00   3rd Qu.: 633.0   3rd Qu.: 2227   3rd Qu.: 5828
 Max.   :2230.00   Max.   :7420.0   Max.   :13466   Max.   :23473

 nHousehold      Families      householdS      familySize
 Min.   :   0   Min.   :   0   Min.   :0.000   Min.   :0.000
 1st Qu.:1118   1st Qu.: 788   1st Qu.:2.480   1st Qu.:3.010
 Median :1524   Median :1059   Median :2.880   Median :3.310
 Mean   :1632   Mean   :1124   Mean   :2.966   Mean   :3.411
 3rd Qu.:2044   3rd Qu.:1409   3rd Qu.:3.380   3rd Qu.:3.750
 Max.   :8528   Max.   :5868   Max.   :6.770   Max.   :6.100

  County
 Length:7049
 Class :character
 Mode  :character
```

Ahora queremos organizar los datos a nivel de condados y obtener los valores de interés de las variables a nivel de condado, es decir, en algunos casos, los valores de las variables de los datos de un mismo condado se suman, por ejemplo la variable NhousingUn. Otras variables, sin embargo, usamos la media para cada condado, por ejemplo la variable houseValue.

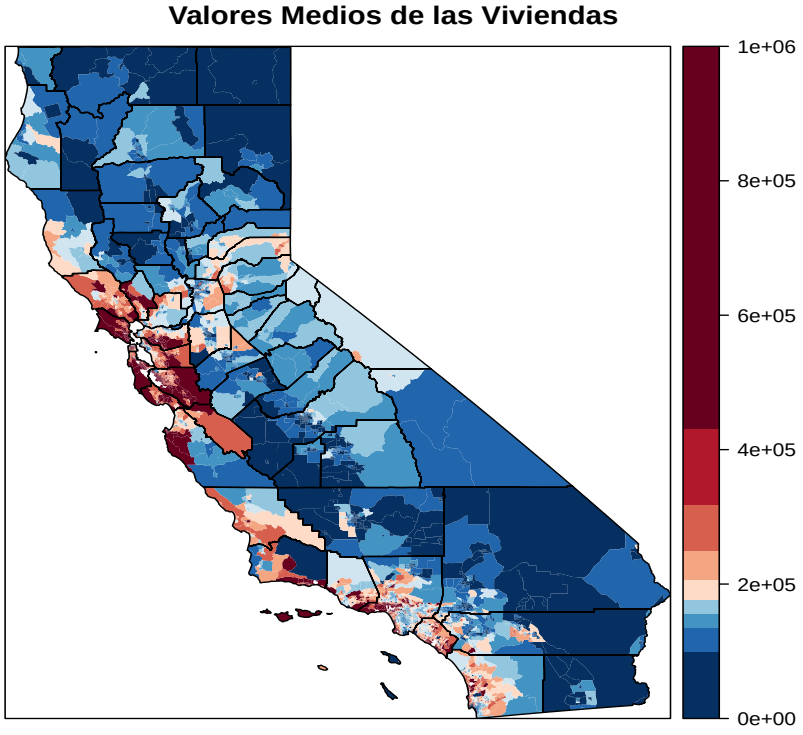
```
> hh <- aggregate(h, "County")

> d1 <- data.frame(h)[, c("nhousingUn", "recHouses", "nMobileHom",
        "nBadPlumbi", "nBadKitche", "Population", "Males", "Females",
        "Under5", "White","Black", "AmericanIn","Asian",
        "Hispanic", "PopInHouse", "nHousehold", "Families")]

>  d1a <- aggregate(d1, list(County=h$County), sum, na.rm=TRUE)

> d2 <- data.frame(h)[, c("houseValue", "yearBuilt", "nRooms",+
+   "nBedrooms","medHHinc", "MedianAge", "householdS",  "familySize")]

> d2 <- cbind(d2 * h$nHousehold, hh=h$nHousehold)

> d2a <- aggregate(d2, list(County=h$County), sum, na.rm=TRUE)

> d2a[, 2:ncol(d2a)] <- d2a[, 2:ncol(d2a)] / d2a$hh

> d12 <- merge(d1a, d2a, by='County')

> hh <- merge(hh, d12, by='County')

> dim(hh)
```

```
[1] 58 27
```

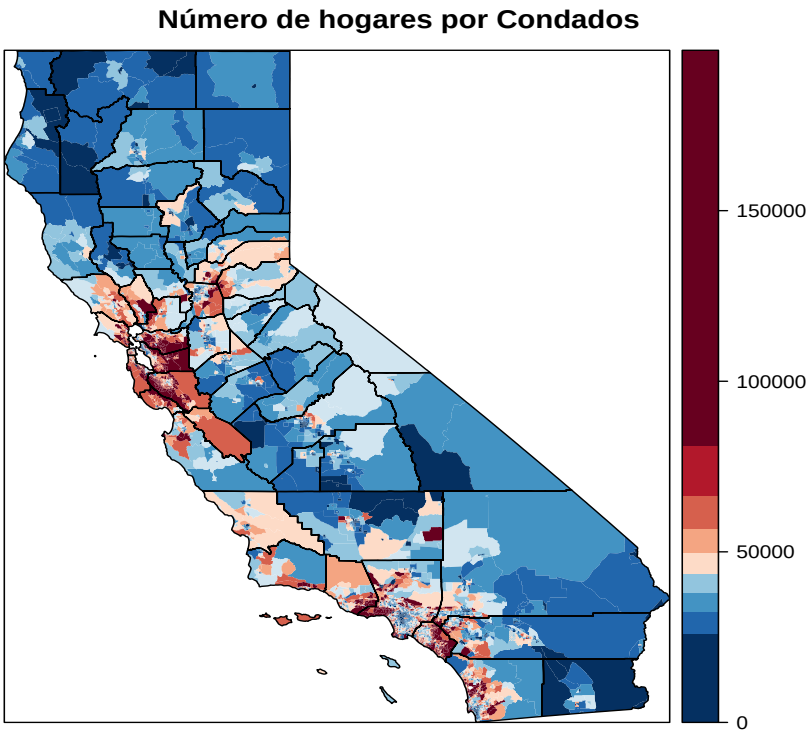
Como podemos comprobar, los datos están ahora organizados por condados (58) y un total de 27 variables que nos dan la información de los condados. Ahora veamos dos gráficas para ver como se distribuyen las variables houseValue y householdS:

```
<install.packages('latticeExtra','RColorBrewer')
<library(latticeExtra)
```

```
-----Plot de valores de las viviendas-----
<grps <- 10

<brksc <- quantile(h$houseValue, 0:(grps-1)/(grps-1), na.rm=TRUE)

<p <- spplot(h,"houseValue", at=brksc, +
+    main="Valores Medios de las Casas",+
+   col.regions=rev(brewer.pal(grps, "RdBu")), col="transparent" )

<p + layer(sp.polygons(hh))
```



```
-----Plot de hogares------
brksh <- quantile(h$medHHinc, 0:(grps-1)/(grps-1), na.rm=TRUE)

<p <- spplot(h, "medHHinc", at=brksh,+
+    main="Número de hogares por Condados",+
+    col.regions=rev(brewer.pal(grps, "RdBu")), col="transparent")

<p + layer(sp.polygons(hh))
```



Para ilustrar los modelos veamos un caso sencillo para este conjunto de datos, veremos el valor de las viviendas a partir de la antigüedad de construcción de la vivienda y del número de habitaciones de ella. Para ello crearemos una nueva variable llamada *age* que serán los años en media que tienen las viviendas en los distintos condados ($houseValue \sim age + nBedrooms$). Realizaremos este simple modelo lineal (múltiple) y así obtendremos los residuos que analizaremos posteriormente:

```
> hh$age <- 2000 - hh$
> f1 <- houseValue ~ age +  nBedrooms
> m1 <- lm(f1, data=hh)
> summary(m1)


Call:
lm(formula = f1, data = hh)

Residuals:
    Min      1Q  Median      3Q     Max
-222541  -67489   -6128   60509  217655

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -628578     233217  -2.695  0.00931 **
age            12695       2480   5.119 4.05e-06 ***
nBedrooms     191889      76756   2.500  0.01543 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94740 on 55 degrees of freedom
Multiple R-squared:  0.3235,    Adjusted R-squared:  0.2989
F-statistic: 13.15 on 2 and 55 DF,  p-value: 2.147e-05
```

El modelo lineal (múltiple) queda de la siguiente forma:

$$houseValue_i = \beta_0 + \beta_1 * age_i + \beta_2 * nBedrooms_i + \varepsilon_i$$

donde los parámetros β_q estimados son $\beta_0 = -628578$, $\beta_1 = 12695$ y $\beta_2 = 191889$.

Por tanto, como habíamos mencionado los años que lleve construida la casa y el número de habitaciones influyen en el valor de ella. Podemos ver incluso que el número de habitaciones hace que aumente cerca de 200.000$ el precio de la vivienda. Veamos si los residuos poseen autocorrelación espacial a partir del índice

I de Moran. Para ello, primero tendremos que crear la matriz de ponderaciones espaciales, $\mathbf{W}$:

```
> nb <- poly2nb(hh)
> nb
```

```
Neighbour list object:
Number of regions: 58
Number of nonzero links: 276
Percentage nonzero weights: 8.204518
Average number of links: 4.758621
```

```
> par(mai=c(0,0,0,0))
> plot(hh)
> plot(nb, coordinates(hh), col='red', lwd=2, add=TRUE)
```

Con la orden anterior se obtiene la figura 5.1 que muestra los enlaces de los condados vecinos. A continuación obtenemos la inferencia sobre la dependencia espacial a través del estadístico I de Moran:

```
> lw<- nb2listw(nb)
> moran.mc(hh$residuals, lw, 999)
```

```
        Monte-Carlo simulation of Moran I

data:  hh$residuals
weights: lw
number of simulations + 1: 1000

statistic = 0.40431, observed rank = 1000, p-value = 0.001
alternative hypothesis: greater
```

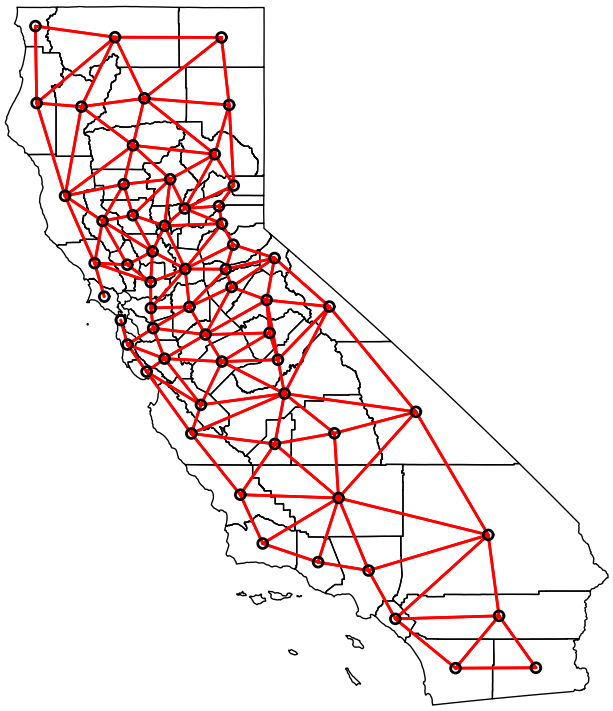


Figura 5.1: Mapa de los enlaces de los condados vecinos

Como tenemos un p-valor de 0,001, se verifica la hipótesis alternativa y, por tanto, los residuos poseen autocorrelación espacial. Lo último que veremos son los tipos de modelos que hemos visto aplicados a estos datos, el modelo tipo SAR (4.4), el modelo de error (4.7) o el modelo de Durbin (4.8):

——Modelo de Retardo Espacial——

$$houseValue_i = \rho * lag.housesValue_i + \beta_0 + \beta_1 * age_i + \beta_2 * nBedrooms_i + \varepsilon_i$$

donde

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$lag.housesValue_i = \sum_{j=1}^{n} W_{ij} * housesValue_j$$

```
> m1s = lagsarlm(f1, data=hh, lw, tol.solve=1.0e-30)
> summary(m1s)
```

```
Call:lagsarlm(formula = f1, data = hh, listw = lw, tol.solve = 1e-30)

Residuals:
      Min        1Q    Median        3Q       Max
-122914.4  -57887.7   -3330.4   41620.1  210803.7

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -488761.6   165471.9 -2.9537 0.0031394
age            6796.5     1831.9  3.7101 0.0002072
nBedrooms    146301.3    54732.7  2.6730 0.0075174

Rho: 0.7291, LR test value: 27.797, p-value: 1.3471e-07
Asymptotic standard error: 0.090506
    z-value: 8.0559, p-value: 8.8818e-16
Wald statistic: 64.897, p-value: 7.7716e-16

Log likelihood: -731.4781 for lag model
ML residual variance (sigma squared): 4494700000, (sigma: 67042)
Number of observations: 58
Number of parameters estimated: 5
AIC: 1473, (AIC for lm: 1498.8)
LM test for residual autocorrelation
test value: 0.46008, p-value: 0.49759
```

Observamos que nos devuelve 5 parámetros estimados que son los que en la parte de teoría vimos que necesitábamos para conseguir el modelo. Estos son: el parámetro ρ que determina el nivel de relación autorregresiva espacial es estimado por 0.7291, los coeficientes β_q dados por $\beta_0 = -488761{,}6$, $\beta_1 = 6796{,}5$ y $\beta_2 = 146301{,}3$ y por último la desviación típica de los residuos generados por este método, es decir, $\sigma = 67042$. También nos devuelve el p-valor de un test realizado a los residuos que generan y como dicho p-valor es 0.497 no podemos rechazar la hipótesis nula de no autocorrelación espacial, es decir, los residuos se pueden considerar independientes y por tanto, tras la aplicación del modelo SAR, la dependencia de los residuos desaparece.

—Modelo de Durbin —

$$houseValue_i = \rho * lag.housesValue_i + \beta_0 + \beta_1 * age_i + \beta_2 * nBedrooms_i +$$

$$+\gamma_1 * lag.age_i + \gamma_2 * lag.nBedrooms_i + \varepsilon_i$$

donde

$$lag.age_i = \sum_{j=1}^{n} W_{ij} * age_j$$

$$lag.nBedrooms_i = \sum_{j=1}^{n} W_{ij} * nBedrooms_j$$

y las variables $lag.housesValue_i$ y ε_i son las mismas que para el modelo SAR.

```
> m1d = lagsarlm(f1, data=hh, lw, Durbin=TRUE,tol.solve=1.0e-30)
>summary(m1d)

Call:lagsarlm(formula = f1, data = hh, listw = lw, Durbin = TRUE,
    tol.solve = 1e-30)

Residuals:
       Min         1Q     Median         3Q        Max
-119926.32  -54437.79    -900.84   38916.33  214614.24

Type: mixed
Coefficients: (asymptotic standard errors)
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -770005.2   389556.9 -1.9766 0.048085
age              5929.2     2263.2  2.6198 0.008799
nBedrooms      135170.0    56760.2  2.3814 0.017246
lag.age          3096.9     4221.2  0.7337 0.463151
lag.nBedrooms  102897.1   135275.1  0.7607 0.446866

Rho: 0.70085, LR test value: 22.59, p-value: 2.005e-06
Asymptotic standard error: 0.099823
    z-value: 7.021, p-value: 2.2036e-12
Wald statistic: 49.294, p-value: 2.2036e-12

Log likelihood: -731.1205 for mixed model
ML residual variance (sigma squared): 4510300000, (sigma: 67159)
Number of observations: 58
Number of parameters estimated: 7
AIC: 1476.2, (AIC for lm: 1496.8)
LM test for residual autocorrelation
test value: 1.2435, p-value: 0.26479
```

En este modelo se estiman 7 parámetros que son, $\rho = 0{,}7$, los coeficientes β_q dados por $\beta_0 = -770005{,}2$, $\beta_1 = 5929{,}2$ y $\beta_2 = 135170$, los coeficientes $\gamma_1 = 3096{,}9$ y $\gamma_2 = 102897{,}1$ y por último la desviación típica de los residuos, $\sigma = 67159$. Al igual que el modelo SAR nos devuelve un p-valor para un test de dependencia espacial para los residuos generados, este es, 0,264, el cual nos indica que los nuevos residuos son independientes desapareciendo así la dependencia espacial de ellos.

——Modelo de Error Espacial——

$$houseValue_i = \beta_0 + \beta_1 * age_i + \beta_2 * nBedrooms_i + \lambda * lag.resid_i + u_i$$

donde

$$u_i \sim \mathcal{N}(0, \sigma^2)$$

$$lag.resid_i = \sum_{j=1}^{n} W_{ij} * \varepsilon_j$$

```
> m1e = errorsarlm(f1, data=hh, lw, tol.solve=1.0e-30)
> summary(m1e)


Call:errorsarlm(formula = f1, data = hh, listw = lw, tol.solve = 1e-30)

Residuals:
        Min          1Q      Median          3Q         Max
-121933.170  -58532.837      47.423   40304.197  216867.129

Type: error
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -302291.4   188876.7 -1.6005 0.109494
age            6989.1     2234.6  3.1277 0.001762
nBedrooms    132526.7    57138.4  2.3194 0.020374

Lambda: 0.75099, LR test value: 23.731, p-value: 1.1078e-06
Asymptotic standard error: 0.088919
    z-value: 8.4457, p-value: < 2.22e-16
Wald statistic: 71.331, p-value: < 2.22e-16
```

```
Log likelihood: -733.5113 for error model
ML residual variance (sigma squared): 4757300000, (sigma: 68973)
Number of observations: 58
Number of parameters estimated: 5
AIC: 1477, (AIC for lm: 1498.8)
```

En este modelo, de nuevo, el número de parámetros estimados son 5 donde el parámetro autorregresivo λ es estimado por 0.75. Los coeficientes son $\beta_0 = -302291{,}4$, $\beta_1 = 6989{,}1$ y $\beta_2 = 132526{,}7$ y la desviación típica de los residuos u_i es $\sigma = 68973$. También nos devuelve unos residuos y aunque no comprobamos que sean independientes, por lo estudiado en teoría han de serlo, de lo contrario habría que modelizar de nuevo los datos teniendo en cuenta estos nuevos residuos.

Como conclusión, los modelos ajustados se resumen en la siguiente tabla y se incluye la medida AIC (Criterio de Información de Akaike)[2] que nos puede permitir seleccionar el modelo más adecuado:

Modelo	**Ecuación**	**AIC**
Lineal (múltiple)	$houseValue = -628578 + \\ + 12695 * age + 191889 * nBedrooms$	1498.8
Retardo Espacial	$houseValue = 0{,}729 * lag.houseValue - 488761{,}6 + \\ + 6796{,}5 * age + 146301{,}3 * nBedrooms$	1473
Espacial de Durbin	$houseValue = 0{,}7 * lag.houseValue - 770005{,}1 + \\ + 5929{,}2 * age + 135170 * nBedrooms + \\ + 3096{,}9 * lag.age + 102897{,}1 * lag.nBedrooms$	1476.2
Error Espacial	$houseValue = 0{,}75 * lag.resid - 302291{,}4 + \\ + 6989{,}1 * age + 132526{,}7 * nBedrooms$	1477

A partir de la tabla anterior podemos comprobar que, siguiendo el criterio AIC, ambos modelos de regresión espacial se ajustan mejor a los datos que el modelo de regresión lineal. Para una elección de los modelos espaciales, utilizaríamos el modelo de retardo espacial ya que el índice AIC es menor.

[2]AIC es una medida de la calidad relativa de un modelo estadístico. Su fórmula viene dado por $AIC = 2k - 2ln(L)$ donde k es el número de parámetros del modelo y L es el máximo de la función de verosimilitud.

Bibliografía

[1] Bivand, R.A.; Pebesma, E. y Gómez-Rubio, V. (2013). *Applied Spatial Data. Analysis with R.* Springer.

[2] Bivand, R; Keitt, T. y Rowlingson, B. (2018). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version1.3-6.* Recuperado de https://CRAN.R-project.org/package=rgdal.

[3] Bivand, R.S y Wong, D. (2018) *Comparing implementations of global and local indicators of spatial association TEST, 27(3), 716-748.* Recuperado de https://doi.org/10.1007/s11749-018-0599-x.

[4] Burridge, P. (1980). On the Cliff-Ord Test for Spatial Correlation. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 42, No. 1.*

[5] Fischer, M.M. y Wang, J. (2011). *Spatial Data Analysis. Models, Methods and Techniques.* Springer.

[6] Gangodagam, C., Zhou, X. y Lin, H. (2008). Autocorrelation Spatial. *Encyclopedia of GIS.* Eds. Shekthar, S. y Xiong, H. Springer.

[7] Hijmans, R.J. (2017). *raster: Geographic Data Analysis and Modeling.* R package version 2.6-7. https://CRAN.Rproject.org/package=raster.

[8] Kazar, B.M. y Celik, M. (2012). *Spatial Autoregression (SAR) Model Parameter Estimation Techniques.* Springer.

[9] LeSage, J. y Pace, R.K., (2009). *Introduction to Spatial Econometrics Statistics.* Chapman and Hall/ CRC.

[10] Moran, P.A.P. (1950). Notes on Continuous Stochastic Phenomena.*Biometrika, Vol. 37, No. 1/2, pp. 17-23.*

[11] Pebesma, E.J. y Bivand, R.S. (2005). *Classes and methods for spatial data in R. R News 5 (2).* Recuperado de https://cran.rproject.org/doc/Rnews/.

[12] Sawada M. *Global Spatial Autocorrelation Indices - Moran's I, Geary's C and the General Cross-Product Statistic*. Laboratory for Paleoclimatology and Climatology, Department of Geography University of Ottawa. Recuperado de http://www.lpc.uottawa.ca/publications/moransi/moran.htm.

[13] Tobler, W.R. (1970). A computer model simulation of urban growth in the Detroit region. *Economic Geaography 46 (2), pp. 234-240.*

[14] Vitturini M. ; Fillottrani P. y Castro S. (2003) Modelo de datos para datos espaciales. *V Workshop de Investigadores en Ciencias de la Computación (WICC 2003).*

[15] The Geospatial and Farming Systems Research Consortium. Recuperado de http://rspatial.org/.