UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS
DE ELCHE

ANÁLISIS ESTADÍSTICO DE DATOS CON CORRELACIÓN ESPACIAL MEDIANTE EL USO DE MODELOS ECONOMÉTRICOS ESTÁTICOS

Aplicación al estudio de la tasa de paro en las comarcas de la Comunidad Valenciana

Autor: **Tomás Granados Navarro**

Tutor: **Domingo Morales González**

Grado en Estadística Empresarial

Trabajo Fin de Grado

Curso 2016-2017

Índice general

Preámbulo

El presente Trabajo de Fin de Grado pretende tener un carácter didáctico, buscando ilustrarla realización de un análisis espacial desde sus primeras fases hasta el ajuste del modelo apropiado. El trabajo consta de 7 capítulos, los 6 primeros tienen un contenido puramente teórico, explicando conceptos fundamentales como la matriz de pesos espaciales, los efectos espaciales, el test de Moran, etcétera. Además, se explican los modelos más conocidos y utilizados en la econometría espacial. En el último capítulo se realiza un estudio de la tasa de paro en las comarcas de la Comunidad Valenciana con datos reales, adjuntando en uno de los anexos el código necesario para reproducirlo.

Me gustaría acabar estas lineas agradeciendo a mi tutor, el profesor Domingo Morales González su paciencia conmigo y, sobretodo, por todo lo que he aprendido durante estos meses. Además, también quiero agradecer a la Vicedecana del Grado en Estadística Empresarial María Victoria Herranz por su trato hacia mí durante todos estos años.

Capítulo 1

Introducción

Es interesante realizar una introducción al análisis espacial haciendo referencia a los datos espaciales, cuya importancia reside en la ubicación o posición de los mismos, por lo que cuando encontremos en nuestra base de datos una determinada posición en un plano (coordenadas cartesianas) o una georreferenciación con unas coordenadas que nos permiten localizar puntos, líneas y polígonos, podremos realizar este tipo de análisis. No es descabellado pensar que esta localización produce una mayor fuente de variabilidad en nuestros datos[1], pero también permite conocer si localizaciones cercanas entre sí tienen características similares o existe algún tipo de relación entre ellas, en contraposición con otras localizaciones más alejadas.

El análisis de datos espaciales, tal y como se conoce en la actualidad, es un campo "relativamente joven" dentro de la estadística y su origen se remonta a principios de los años 50 con el desarrollo de la geografía cuantitativa y de la ciencia regional. Además, muchas de las técnicas de análisis espacial se desarrollaron durante los años 60 y 70, siendo un período en el que el desarrollo informático era aún muy pobre y las bases de datos ínfimas con respecto a las actuales.

El concepto de autocorrelación espacial se estableció en la Universidad de Washington a finales de los años 50[2], por los geógrafos Michael F. Dacey, William L. Garrison y Edward Ullman. Uno de los estudiantes de los dos últimos fue Waldo Tobler, quien definió en 1970 el enlace entre variables con carácter espacial mediante la

[1] Un ejemplo lo encontramos en la naturaleza, ya que raramente los objetos en ella se encuentran distribuidos con aleatoriedad uniforme.

[2] Hasta 1968 se denominó de varias formas, algunos ejemplos son: dependencia espacial, asociación espacial, interacción espacial. Ese año, gracias al artículo de A. D. Cliff y J.K. Ord "El problema de la autocorrelación espacial", este término entró en el léxico de los analistas espaciales.

primera ley de la geografía: "Todo está relacionado entre sí, pero aquello más cercano en el espacio, tiene una relación mayor que aquello que se encuentre distante".

Los años 70 se pueden considerar una época prolífera, debido a la importancia de los hechos ocurridos en cuanto a la estadística espacial se refiere. Además de la aparición de la primera ley de la geografía en 1973, los investigadores A. D. Cliff y J.K. Ord presentaron su monografía "Spatial autocorrelation", siendo ampliada por los mismos autores en su libro "Spatial Processes: Models and Applications" de 1981, donde explicaban la naturaleza de la matriz de pesos espaciales, de gran importancia en el análisis espacial, y, sobretodo, guiaban paso a paso en la aplicación de los test de la I de Moran[3] (Patrick A. P. Moran, 1950) y la C de Geary (Roy C. Geary, 1954), los dos estadísticos de autocorrelación más importantes, además de realizar las formulaciones para el cálculo de la esperanza y la varianza de los mismos. Es importante señalar que el interés que existe en la actualidad por la estadística espacial viene dado en gran medida a este texto. Además, a finales de los años 70 J. H. P. Paelinck y L. H. Klaassen acuñaron el término "econometría espacial" en su escrito del mismo nombre, no siendo hasta 1988 cuando Luc Anselin en su trabajo "Spatial Econometrics: Methods and Models" lo popularizó al hacer accesible, para todo el mundo, esta disciplina de la econometría[4].

Durante los años 90, la autocorrelación espacial se consideró ya un tema de suma importancia en la literatura relacionada con el análisis espacial, siendo éste un concepto que no se podía obviar al trabajar con datos geolocalizados. A su vez, a finales de esta época el análisis espacial gozó de una popularidad enorme, debido en gran medida al desarrollo de rutinas de estimación y, sobretodo, a las mejoras en la tecnología. Esta popularidad no ha hecho más que crecer en el tiempo hasta nuestros días y es notorio en la cantidad de artículos, investigaciones, literatura y trabajos de doctorado que se presentan a día de hoy, independientemente de la disciplina (epidemiología, geografía, econometría espacial, estadística espacial, etcétera).

En la actualidad, y según establecen Bivand, Pebesma y Gómez-Rubio en su libro

[3]El test de Moran fue el primer test estadístico construido para detectar la posible presencia de relación espacial respecto a una variable dada.

[4]Aunque se popularizó en esta época, puede ser curioso hacer referencia a que ya en 1854 John Snow hizo uso de datos espaciales para acabar con el brote de cólera que afectaba a Londres durante aquel periodo. Lo consigió aislando las zonas donde aparecían los casos, llegando a la conclusión de que era producido por aguas contaminadas que acababan en las bombas de agua donde se recogía para su consumo.

"Applied Spatial Data Analysis with R", podemos distinguir tres tipos de análisis espacial:

1. Procesos puntuales espaciales: Se estudian un conjunto de observaciones en el espacio. Nos preguntamos si los datos muestran patrones de agrupamiento espacial, o si muestran lo que se conoce como aleatoriedad espacial completa (CSR)[5]. Por ejemplo, el análisis en la aleatoriedad o el agrupamiento de los casos y, si es así, si existe algún factor o factores que propicien un cierto patrón, en la ubicación de cánceres en un área urbana.

2. Datos geoestadísticos: Se estudian datos en un conjunto limitado de puntos espaciales. Basados en estas observaciones, estamos interesados en interpolar los datos a los puntos no observados. Por ejemplo, si observamos la calidad del aire en un conjunto de estaciones de monitorización, qué podemos decir sobre la calidad del aire en el conjunto de la región.

3. Datos de área: Se observan datos distribuidos en una región espacial predefinida, pudiendo plantearnos si lo que ocurre en una región se debe a la influencia de lo que ocurre en otra. Pudiendo establecer como ejemplos: Dónde suelen aparecer los tornados con mayor frecuencia y cómo influye su aparición en las zonas cercanas o el poder identificar puntos de delincuencia en un barrio, ciudad o región y comprobar si influyen en zonas colindantes.

De entre todos ellos, el tipo más utilizado en econometría es el referido a datos de área, por lo que el enfoque del presente trabajo está dirigido a ese tipo de análisis. Por lo tanto, los posibles ejemplos y variables vendrán determinadas por conjuntos de datos de regiones espaciales (países, estados, secciones censales, códigos postales, etcétera). Asumiendo, en la mayoría de los casos, que los datos contienen una única observación en cada región (una observación en un momento dado), comprendiendo una sección espacial transversal o de panel[6]y teniendo en cuenta que existen N regiones en el conjunto de datos.

[5]La aleatoriedad espacial completa viene dada cuando dentro de nuestra región de estudio, diversos eventos ocurren de forma totalmente aleatoria e independiente.

[6]Se habla de datos transversales o de panel, cuando existen observaciones repetidas a lo largo del tiempo para una muestra de unidades individuales, es decir, para una variable y_{it} con $i = 1, 2, ..., N$(pudiendo ser i países, regiones, etc) observados a lo largo de $t = 1, 2, ..., T$ períodos de tiempo.

Capítulo 2

Estructura de la adyacencia espacial

Al realizar un análisis espacial, la estructura en los vínculos espaciales se describe mediante la matriz de pesos espaciales. Esta matriz formaliza la proximidad relativa entre las observaciones, basándose en la primera ley de la geografía definida por Tobler.

Una buena construcción de esta matriz es clave a la hora de realizar un estudio estadístico de datos espaciales, puesto que si se construye de forma errónea no servirá de nada si el análisis indica que existe autocorrelación espacial en los datos a estudio o no, ocurriendo lo mismo con el modelo formulado y las conclusiones que alcancemos con el mismo.

En el presente trabajo, se presentan dos enfoques totalmente distintos para la construcción de la matriz de pesos, siendo éstos los más utilizados en la literatura relativa al análisis espacial. El primer enfoque está basado en la distancia y considera que dos o más regiones son adyacentes si, al disponer de datos agregados de un área, establecemos centroides[1] en dichas superficies y tales observaciones se encuentran a una distancia determinada las unas de las otras. También puede darse la situación en la que los datos presenten un determinado número de observaciones en un área determinada, estableciendo que son adyacentes si se encuentran entre sí a una distancia predefinida. El segundo enfoque a estudio es más simple y conlleva menos cálculos por parte del investigador, ya que las observaciones se considerarán adyacentes cuando entre ellas compartan cualquier zona que las delimite.

Para finalizar, y aunque pueda parecer obvio, es interesante hacer hincapié en

[1] Un claro ejemplo de centroide podría ser una capital de provincia si los datos del estudio son provinciales, o la capital de la comarca si son comarcales.

algo. La construcción de la matriz de pesos espaciales puede parecer sencilla, pero debido a su inmensa importancia dentro de un análisis espacial se aconseja su construcción mediante el software estadístico R, al que se hace referencia en el capítulo 6, ya que es enormemente fácil el poder equivocarse debido a lo tedioso que puede resultar ese trabajo.

2.1. Matriz de pesos espaciales basada en distancias

De nuevo, y como en todo el presente trabajo, se debe tener en mente la primera ley de la geografía descrita por Tobler. En nuestro caso, la noción de proximidad se debe basar en la capacidad de poder medir la distancia de separación entre las distintas observaciones en una localización determinada.

El cálculo de la distancia se puede basar en muchas formas de medida, siendo dos de ellas las más utilizadas de cara a un análisis espacial: Distancia Euclídea y distancia Manhattan.
La primera se corresponde con la línea recta que separa los dos observaciones, siendo ésta la longitud de la hipotenusa en el teorema de Pitágoras. Esta distancia se corresponde con la distancia más corta entre dos puntos situados en un plano, pero no tiene en cuenta las posibles restricciones urbanísticas y geográficas que pueden aparecer en el área a estudio, tales como edificios en zonas urbanas o montañas y lagos si el análisis tiene un trasfondo geográfico.
La distancia Manhattan se corresponde con la suma de las longitudes de lo que sería el análogo a los catetos en el teorema de Pitágoras. El principio de esta distancia considera la posible aparición de las restricciones anteriormente expuestas y es tan simple como pensar cómo llegamos de un destino A (como puede ser nuestra casa) a un destino B (universidad, tienda, trabajo, etc), no vamos atravesando objetos por el camino, los bordeamos.

2.1.1. Matriz de distancias

Generalizando el cálculo de la distancia para N observaciones, la medición de las distancias Euclídea (d_{ij}) y Manhattan ($\bar{d_{ij}}$) respecto a dos puntos p_i y p_j con

coordenadas (X_i, Y_i) y (X_j, Y_j), se obtienen las siguientes ecuaciones:

$$d_{ij} = \sqrt{(Y_i - Y_j)^2 + (X_i - X_j)^2}, \quad \forall\, i, j \in \{1, ..., N\},$$

$$\bar{d_{ij}} = |Y_i - Y_j| + |X_i - X_j|, \quad \forall\, i, j \in \{1, ..., N\}.$$

Los resultados obtenidos mediante el cálculo de la distancia deseada se dispondrán en una matriz cuadrada $(N \times N)$ denominada matriz de distancias, en la que cada dato alojado en una determinada celda (i, j) se corresponderá con la distancia entre esa determinada posición i, respecto a una posición j. Para tal propósito, las matrices de distancias obtenidas mediante la distancia Euclídea, denotada en el presente trabajo como D, o mediante la distancia Manhattan, denominada como $\bar{D}$, presentan la siguiente estructura:

$$D = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1j} & \cdots & d_{1N} \\ d_{21} & 0 & \cdots & d_{2j} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \cdots & d_{ij} & \cdots & d_{iN} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nj} & \cdots & 0 \end{pmatrix} \quad \bar{D} = \begin{pmatrix} 0 & \bar{d_{12}} & \cdots & \bar{d_{1j}} & \cdots & \bar{d_{1N}} \\ \bar{d_{21}} & 0 & \cdots & \bar{d_{2j}} & \cdots & \bar{d_{2N}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{d_{i1}} & \bar{d_{i2}} & \cdots & \bar{d_{ij}} & \cdots & \bar{d_{iN}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{d_{N1}} & \bar{d_{N2}} & \cdots & \bar{d_{Nj}} & \cdots & 0 \end{pmatrix}.$$

Como podemos comprobar, la diagonal principal de la matriz $(i = j)$ contiene únicamente ceros y será siempre así, puesto que la distancia entre un punto y sí mismo tendrá siempre este valor. Además, como la distancia entre una determinada observación i y otra j será siempre igual a la distancia entre tal observación j y dicha observación i, la matriz de distancias presentará siempre una estructura simétrica. Dicha simetría y que la matriz esté compuesta por ceros en su diagonal principal, ahorrará mucho tiempo computacional. El cálculo se realizará mediante $(N \times N)/2$ elementos por tratarse de una matriz simétrica y al contener siempre un elemento con valor 0 por fila, solo será necesario calcular $(N - 1)$ distancias para cada una de las observaciones por columna. Por lo que finalmente, el cálculo será de $N \times (N-1)/2$ distancias.

2.1.2. Matriz de pesos espaciales

El cálculo de la distancia entre observaciones, es únicamente un paso en la construcción de una matriz capaz de contener la intensidad de los distintos vínculos espaciales, entre dichas observaciones o puntos. De hecho, el uso de la distancia puede transmitir un mensaje opuesto a las relaciones espaciales descritas en la primera ley de la geografía, ya que cuanto mayor sea la distancia entre observaciones, mayor será el valor en la matriz de distancias. Por lo tanto, es necesario realizar una transformación para obtener una matriz de pesos que tenga en cuenta una determinada estructura de proximidad, en la que cada celda de la matriz contenga un valor que disminuya cuanto más se incremente la distancia entre observaciones. Dicha matriz, es la denominada matriz de pesos espaciales, cuya notación es W y, debido a que se construye a partir de la matriz de distancias ya sea D o $\bar{D}$, seguirá teniendo una dimensión $(N \times N)$. Cada uno de los elementos de la matriz se denota como w_{ij}, $\forall\ i, j \in \{1, ..., N\}$ y cada elemento representa la intensidad en la proximidad espacial entre las observaciones i y j. Además, los elementos serán siempre no estocásticos (no aleatorios), no negativos y finitos, debido a que equivalen a los elementos d_{ij} y $\bar{d_{ij}}$ de la matriz de distancias.
Después de lo anteriormente expuesto, la estructura de la matriz de pesos espaciales queda representada como

$$W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1j} & \cdots & w_{1N} \\ w_{21} & 0 & \cdots & w_{2j} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} & \cdots & w_{iN} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{Nj} & \cdots & 0 \end{pmatrix}.$$

En la literatura se pueden encontrar múltiples formas de transformar una matriz de distancias en una matriz de pesos espaciales y, sin querer profundizar demasiado en muchas de ellas, se presentan las siguientes transformaciones:

- Relaciones basadas en conectividad.
- Relaciones basadas en la inversa de las distancias.

- Relaciones basadas en la inversa de la exponencial o negativa exponencial.
- Relaciones Gaussianas.

2.1.2.1. Relaciones de conectividad

Las relaciones de conectividad se establecen mediante una dicotomización de los elementos de distancia[2], es decir, una determinada observación i adquirirá en su posición de la matriz de pesos espaciales, w_{ij}, el valor 1 cuando una determinada observación j esté localizada a una distancia igual o menor a $dist$ de i. Por supuesto, en caso contrario adquirirá en tal posición un valor 0, siendo

$$w_{ij} = \begin{cases} 1 \text{ si } d_{ij} \leq dist, \quad \forall\ i, j \in \{1, ..., N\},\ i \neq j, \\ 0 \text{ en caso contrario.} \end{cases}$$

De esta forma, las relaciones de conectividad identifican aquellas observaciones localizadas en un determinado radio. En este sentido, puede aparecer la duda de qué longitud de radio, $dist$, utilizar para el análisis y así determinar aquellas relaciones espaciales capaces de capturar o expresar el posible patrón espacial existente en los datos de la forma más correcta posible[3].

Una variante en la relación de conectividad expuesta, puede ser el considerar que el radio que establezcamos varíe para cada una de las observaciones[4], siendo esta distancia $dist_{i(k)}$. De esta forma, la distancia para cada una de las observaciones se ajustará para que cada una de las observaciones esté conectada con un número determinado de k adyacentes, obteniendo

$$w_{ij} = \begin{cases} 1 \text{ si } d_{ij} \leq dist_{i(k)}, \quad \forall\ i, j \in \{1, ..., N\};\ i \neq j, \\ 0 \text{ en caso contrario.} \end{cases}$$

La principal ventaja en esta variante reside en la simplicidad de la interpretación de las relaciones espaciales y, debido a esto, la relación de conectividad entre k adyacentes cercanos está muy extendida en la práctica. En cambio, su principal desventaja

[2]De aquí en adelante ya no se distinguirá en cuanto a la obtención de la matriz de distancias se refiere. Por lo tanto, la distancia entre las observaciones i y j quedará expresada como d_{ij}.

[3]Un punto de partida es tener presente que las observaciones se encuentren conectadas, al menos, con otra observación.

[4]Esta variante se corresponde con lo que en la literatura se denota como el criterio de las adyacencias más cercanas (nearest neighbors criterion).

reside en que no respeta la intensidad en la relación entre distintas observaciones en función de la distancia, es decir, dentro de un radio de acción de 20 kilómetros para una observación i, no va a importar si una observación se encuentra a 5 metros y otra a 19 kilómetros, dado que ambas distancias obtendrán el mismo peso.

2.1.2.2. Relaciones de distancia inversa

Otra posibilidad es tomar la inversa, o incluso el cuadrado inverso, de la distancia que separa las distintas observaciones. Esta transformación permite la construcción una la matriz de pesos espaciales capaz de respetar la ya expuesta ley de Tobler, ya que los pesos son mayores, o menores, cuando las observaciones se encuentran espacialmente cercanas o alejadas, siendo

$$w_{ij} = \begin{cases} 1/d_{ij}^{\gamma} = d_{ij}^{-\gamma}, & \forall\ i, j \in \{1, ..., N\}, \\ 0 & \forall\ i = j. \end{cases}$$

El parámetro γ, con $\gamma > 0$, permitirá penalizar en mayor o menor medida la proximidad espacial, en función de la distancia. Es evidente señalar que cuando se establezca $\gamma = 1$, nos encontraremos en el supuesto relativo a la inversa de la distancia, y que cuando se establezca $\gamma = 2$ estaremos realizando el cuadrado inverso de la distancia. Obviamente, la elección del valor para el parámetro influirá en el valor del peso, w_{ij}, en función de la distancia. La inversa de la distancia, $\gamma = 1$, va a conceder un peso mayor a aquellas variables espacialmente más cercanas, distancia menor, y un peso menor a aquellas más alejadas, distancia mayor. En cambio, el cuadrado inverso de la distancia ($\gamma = 2$) va a otorgar un peso aún mayor en aquellas observaciones más cercanas, facilitando un valor marginal a aquellas más alejadas.

2.1.2.3. Relaciones basadas en la inversa exponencial o negativa exponencial

La transformación de la distancia basada en la inversa de la exponencial, o la negativa exponencial de la distancia puede expresarse como

$$w_{ij} = \frac{1}{e^{d_{ij}}} = e^{-d_{ij}}, \quad \forall\, i \neq j;\ i, j \in \{1, ..., N\}.$$

Como ocurría con la transformación de distancia inversa, esta transformación establece una importancia mayor en observaciones espacialmente más cercanas y, por lo tanto, dando menos peso a aquellas más alejadas. En cambio, la ventaja de este enfoque reside en que el peso converge rápidamente hacia cero cuando la distancia entre observaciones se incrementa.
Cuando comparamos este tipo de relaciones con aquellas basadas en la inversa de la distancia, así como la del cuadrado inverso de la distancia, la transformación de la inversa de la exponencial de la distancia da menos importancia a las observaciones muy cercanas entre sí. Además, esta transformación tiene la ventaja de limitar el efecto a aquellas observaciones que se encuentran cercanas espacialmente, sin introducir una distancia límite de corte para acotar la relación a un pequeño radio de influencia. Otra ventaja que presenta, se basa en el hecho de que las observaciones que comparten la misma localización, las mismas coordenadas geográficas, se consideran como parte de las relaciones espaciales. Una distancia con valor cero conduce a un peso igual a uno, puesto que el exponencial de cero equivale a uno.

Tras lo anteriormente expuesto, se puede añadir que es posible introducir un cierto límite en la distancia, *dist*, para relaciones espaciales basadas en la inversa exponencial de la distancia. Por lo que aquellas relaciones espaciales que se encuentren a una distancia superior a *dist* tendrán un valor equivalente a cero, siendo

$$w_{ij} = \begin{cases} e^{-d_{ij}}, & \forall\ d_{ij} \leq dist;\ i \neq j;\ i,j \in \{1,...,N\}, \\ 0 & \forall\ d_{ij} > dist, \\ 0 & \forall\ i = j. \end{cases}$$

2.1.2.4. Relaciones basadas en la transformación Gaussiana

Es posible tener en cuenta una alternativa a las transformaciones basadas únicamente en la distancia entre observaciones. Para la transformación Gaussiana, la relación entre observaciones se basa en una determinada distancia límite denotada, de nuevo, como *dist* y expresada como

$$w_{ij} = \begin{cases} [1-(d_{ij}/dist)^2]^2, & \forall\ d_{ij} \leq dist;\ i \neq j;\ i,j \in \{1,...,N\}, \\ 0 & \forall\ d_{ij} > dist, \\ 0 & \forall\ i = j. \end{cases}$$

La ventaja de esta trasformación reside en tener en cuenta, explícitamente, una distancia límite por encima de la cual la relación espacial se entenderá con valor cero. Esta distancia límite se puede establecer de acuerdo a la distancia media o, incluso, a la distancia máxima. La selección de una distancia con estas características conlleva que todas las observaciones de la muestra se tienen en cuenta para establecer las posibles relaciones espaciales.
En cambio, una de las desventajas que plantea es que la relación de proximidad no es muy sensible a la distancia inicial, estableciendo así un peso casi idéntico en aquellas observaciones localizadas en un radio más amplio. Sin embargo, la idea de la primera ley de la geografía todavía se tiene en cuenta, puesto que las observaciones más cercanas entre sí reciben un peso mayor que aquellas que se encuentran más alejadas. Otra de las desventajas, recae en la necesidad de encontrar el valor óptimo que establecer como distancia límite.

2.1.2.5. Selección de la transformación para el análisis de nuestros datos

Como ocurre en muchas ocasiones, la selección del método para realizar la transformación de la matriz de distancias en la matriz de pesos espaciales, recaerá en el criterio del investigador o del analista que vaya a realizar el estudio. No obstante, en la actualidad podemos encontrar en la literatura algunas reglas para decantarnos por una transformación en particular:

- El uso de la relación basada en la inversa de la distancia será preferible cuando la escala del sistema geográfico a estudio sea amplia (complejo a gran escala).

- La relación basada en la inversa de la exponencial de la distancia será preferible si la escala del sistema geográfico a análisis es pequeña (escala pequeña o correlación casi local).

- La relación basada en la conectividad deberá tenerse en cuenta cuando se sospeche que las relaciones espaciales son locales (acción local).

2.1.3. Estandarización de la matriz de pesos espaciales

Llegados a este punto, tenemos construida una matriz de pesos espaciales basada en la distancia entre observaciones, pero se debe realizar un último paso, la estandarización de la matriz obtenida.
Este paso se debe realizar antes de llevar a cabo las pruebas de detección de la autocorrelación espacial, así como la estimación de los efectos espaciales autorregresivos. De hecho, la normalización la matriz de pesos espaciales, tiene ciertos efectos positivos a pesar de que su construcción provoca una modificación de la forma en las relaciones espaciales, ya que la matriz ya no tiene estructura simétrica.

La estandarización de las filas de la matriz de pesos, consiste en establecer que la suma de los elementos en la fila i-ésima pueda expresar que la proximidad espacial entre la observación i y las demás observaciones de la muestra sea igual a uno o al 100 % si lo establecemos en términos de porcentaje. El cálculo a realizar no es más que una ponderación, expresada como

$$\tilde{w}_{ij} = \frac{w_{ij}}{\sum_{j=1}^{N} w_{ij}},$$

por lo que

$$\sum_{j=1}^{N} \tilde{w}_{ij} = 1.$$

La matriz que resulta se denota como $\tilde{W}$ y presenta la estructura[5]

$$\tilde{W} = \begin{pmatrix} 0 & \tilde{w}_{12} & \tilde{w}_{13} & \cdots & \tilde{w}_{1N} \\ \tilde{w}_{21} & 0 & \tilde{w}_{23} & \cdots & \tilde{w}_{2N} \\ \tilde{w}_{31} & \tilde{w}_{32} & 0 & \cdots & \tilde{w}_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{w}_{N1} & \tilde{w}_{N2} & \tilde{w}_{N3} & \cdots & 0 \end{pmatrix}.$$

En realidad, estandarizar la matriz de pesos espaciales no es un paso estrictamente necesario, pero sí muy recomendable por las muchas ventajas que presenta. Una de estas ventajas reside en que ayuda a la interpretación de los cálculos en pruebas

[5]Como se apuntó con anterioridad, la matriz de distancias es simétrica y, por ende, la matriz de pesos espaciales. En cambio, la matriz de pesos espaciales estandarizada $\tilde{W}$ no es simétrica

estadísticas, así como la comparación de los resultados obtenidos con otras matrices de pesos espaciales que se hayan obtenido mediante métodos distintos al nuestro. A su vez, podemos interpretar mejor el valor del parámetro ρ en los modelos econométricos; siendo el parámetro que muestra la fuerza de autocorrelación espacial entre observaciones, ya que estará establecido dentro del intervalo $[-1, 1]$, siendo de signo negativo y positivo respectivamente[6]. Para finalizar, el uso de una matriz de pesos espaciales estandarizada nos permitirá calcular los estadísticos espaciales de autocorrelación.

Tras lo expuesto, es fácil llegar a la conclusión de que estandarizar nuestra matriz de pesos es una gran idea pese a que en realidad no sea necesario.

2.2. Matriz de pesos espaciales basada en contigüidades

Como se ha planteado al comienzo del capítulo, el enfoque basado en la construcción de la matriz de pesos espaciales fundamentado en contigüidades es mucho más simple que el expuesto en el apartado anterior.

En este segundo enfoque, construiremos una matriz cuadrada $(N \times N)$, además de simétrica, con $(i,\ j)$ elementos iguales a 1 cuando el área o región i y el área o región j sean adyacentes la una con la otra, se encuentran relacionadas espacialmente, y 0 en caso contrario,[7]es decir

$$w_{ij} = \begin{cases} 1 & \text{si } i \text{ y } j \text{ son regiones adyacentes,} \quad \forall\ i, j = 1, ..., N;\ i \neq j, \\ 0 & \text{en caso contrario.} \end{cases}$$

Por convención, los elementos diagonales de esta matriz binaria de pesos espaciales, o de adyacencia espacial (spatial neighbors matrix), se deben establecer con valor 0.

A la hora de construir este tipo de matrices existe un número increíblemente

[6]Si la matriz de pesos no estuviese estandarizada, el valor de ρ estaría comprendido entre los valores $(1/\omega_{min}) < \rho < (1/\omega_{max})$ para el modelo SAR, siendo ω_{min} y ω_{max} el menor y el mayor autovalor respectivamente.

[7]En una aplicación no regional, de un problema estudiado en sociología o ciencias políticas, podríamos establecer $w_{ij} = 1$ si las regiones i y j pertenecen al mismo entorno social y, así, estudiar su comportamiento bajo tal suposición. En la práctica, la mayoría de las investigaciones de tipo regional comienzan simplemente con la matriz de contigüidad 0/1 de filas estandarizadas.

alto de métodos a utilizar. A continuación se presenta una selección de tales procedimientos:

- Contigüidad lineal este-oeste: i y j son adyacentes si comparten parte de su borde oriental u occidental.[8]

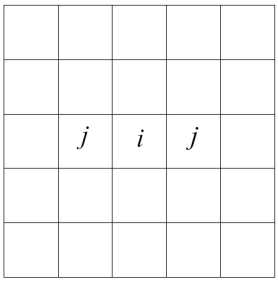


Figura 2.1: Contigüidad lineal este-oeste.
Fuente: Elaboración propia.

- Contigüidad de la torre[9]: i y j son adyacentes si comparten parte de una frontera común, sin importar la orientación.

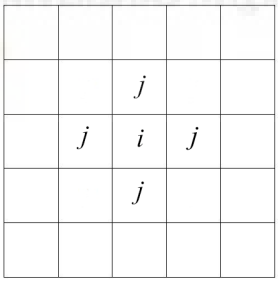


Figura 2.2: Contigüidad de la torre.
Fuente: Elaboración propia.

- Contigüidad del alfil: i y j son adyacentes espacialmente si se encuentran en un "punto". Se puede considerar el análogo espacial de dos elementos de un grafo, al encontrarse en un vértice. En la práctica, se dice que dos regiones son adyacentes mediante la presente método si su frontera común es más corta que la superficie de adyacencia de las contigüidades anteriormente expuestas.

[8]No se encuentra explicación alguna al por qué de establecer este privilegio este-oeste sobre una relación norte-sur en un contexto espacial.

[9]Tanto la contigüidad de la torre, como la del alfil y la reina deben su nombre a las piezas de ajedrez, debido al movimiento en el tablero de éstas.

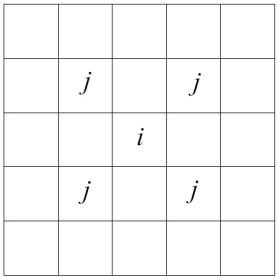


Figura 2.3: Contigüidad del alfil.
Fuente: Elaboración propia.

- Contigüidad de la reina: Es la unión entre la contigüidad de la Torre y la del Alfil. Por lo cual, i y j son adyacentes si comparten alguna parte de la frontera común, sin importar el tamaño de la superficie de contacto entre regiones.

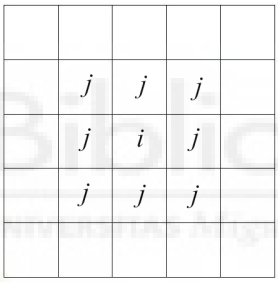


Figura 2.4: Contigüidad de la reina.
Fuente: Elaboración propia.

Además, se puede ir más lejos y definir medidas de contigüidad de "segundo orden". Estas medidas contarían como regiones adyacentes a aquellas que comparten una frontera común con las regiones adyacentes de primer orden, de acuerdo a los criterios enumerados:

- Doble contigüidad lineal este-oeste: i y j son adyacentes si comparten parte de su borde oriental u occidental, generando una contigüidad lineal de segundo orden entre i y k.

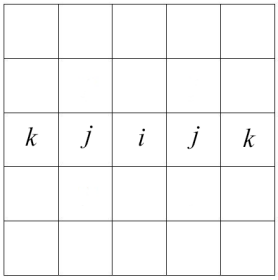


Figura 2.5: Contigüidad lineal este-oeste de segundo orden.
Fuente: Elaboración propia.

- Doble contigüidad de la torre: i y j son adyacentes si comparten parte de una frontera común sin importar la orientación, obteniendo una contigüidad de la torre de segundo orden entre i y k.

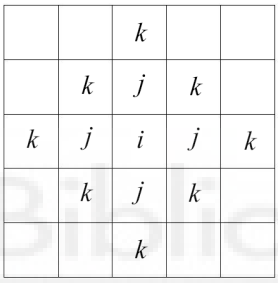


Figura 2.6: Contigüidad de la torre de segundo orden.
Fuente: Elaboración propia.

Al tener tantas posibilidades para construir la matriz de pesos espaciales, podemos dudar sobre cuál de ellas utilizar. Para poder decantarnos por una construcción en particular, deberemos analizar la situación planteada en nuestros datos y elegir la que más se amolde a las necesidades del estudio.
Utilizando un supuesto mencionado por LaSage, supongamos que nuestros datos contienen zonas con distinto código postal, siendo colindantes únicamente en una pequeña parte de sus bordes, es decir, un punto. Además, ese borde de adyacencia se considera como una ciudad dormitorio. En la mayoría de planteamientos se podría no querer usar la contigüidad de la torre, puesto que tales regiones podrían no resultar adyacentes. Una excepción a este problema podría ser la que se plantea cuando no existe una ruta de transporte directa entre las regiones, por ejemplo, si están separadas por un río y el puente más cercano implica tener que pasar por una nueva región. En este caso, la contigüidad de la torre podría ser el método que necesitamos

en nuestro análisis. Es decir, todo depende del contexto que nos indiquen nuestros datos.

La matriz de pesos espaciales construida, W, al igual que la elaborada mediante distancias, queda estructurada como

$$W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1j} & \cdots & w_{1N} \\ w_{21} & 0 & \cdots & w_{2j} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{i1} & w_{i2} & \cdots & w_{ij} & \cdots & w_{iN} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{Nj} & \cdots & 0 \end{pmatrix}.$$

Además, la matriz W se estandarizaría de la misma forma, obteniendo la matriz $\tilde{W}$,

$$\tilde{W} = \begin{pmatrix} 0 & \tilde{w}_{12} & \cdots & \tilde{w}_{1j} & \cdots & \tilde{w}_{1N} \\ \tilde{w}_{21} & 0 & \cdots & \tilde{w}_{2j} & \cdots & \tilde{w}_{2N} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{w}_{i1} & \tilde{w}_{i2} & \cdots & \tilde{w}_{ij} & \cdots & \tilde{w}_{iN} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \tilde{w}_{N1} & \tilde{w}_{N2} & \cdots & \tilde{w}_{Nj} & \cdots & 0 \end{pmatrix}.$$

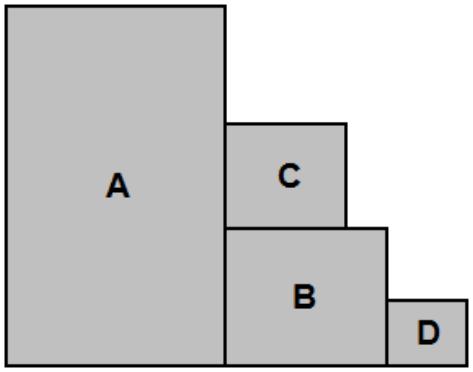
Algo muy importante cuando realicemos una transformación basada en contigüidades es tener muy en cuenta el concepto de isla, entendiendo como isla a aquella conectividad externa limitada o inexistente. Una isla es fácilmente identificable en una matriz de adyacencias o pesos espaciales, puesto que se corresponderá con una fila compuesta únicamente por ceros.[10]

Podemos también señalar que es importante tener las islas en cuenta, debido a que algunas rutinas computacionales pueden fallar si los datos a analizar las contienen, por lo que es mejor, en la medida de lo posible, usar un conjunto de datos que no las contengan.

[10]Si queremos hilar más fino en cuanto a las islas se refiere, podemos señalar que podríamos transformar la fórmula para la estandarización de la matriz de pesos para que quede expresada de la siguiente forma: $\tilde{w}_{ij} = w_{ij}/max(1, \sum_{j=1}^{N} w_{ij})$

2.2.1. Ejemplo práctico

Consideremos el sistema con regiones $N = 4$:



Región	Adyacencia
A	B,C
B	A,C,D
C	A,B
D	B

Tabla 2.1: Sistema de regiones y adyacencias.
Fuente: Elaboración propia.

Podemos obtener la matriz de adyacencias o de pesos espaciales

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

Es notorio como las regiones A y C tienen dos adyacentes, mientras que la región B tiene tres adyacentes y D tiene únicamente uno.

Estandarizando la matriz de pesos espaciales se obtiene la estructura

$$\tilde{W} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Si nos fijamos, cada fila representa el valor de la unidad como un promedio ponderado

de los valores de sus adyacentes

$$\begin{aligned}
[y_1] &= \frac{1}{2}y_2 + \frac{1}{2}y_3, \\
[y_2] &= \frac{1}{3}y_1 + \frac{1}{3}y_3 + \frac{1}{3}y_4, \\
[y_3] &= \frac{1}{2}y_1 + \frac{1}{2}y_2, \\
[y_4] &= y_2.
\end{aligned}$$

Es decir, de forma general, tenemos la ecuación

$$[y_i] = \sum_{j=1}^{N} w_{ij}y_j,$$

mediante la cual podemos obtener el valor de la observación y_i si no la conocemos. Siendo $[y] = ([y_1], [y_2], [y_3], [y_4])'$ el vector del promedio sobre los adyacentes de cada región, podemos estructurar la matriz Wy[11]

$$Wy = \begin{pmatrix} \frac{1}{2}(y_2 + y_3) \\ \frac{1}{3}(y_1 + y_3 + y_4) \\ \frac{1}{2}(y_1 + y_2) \\ y_2 \end{pmatrix}.$$

El vector Wy representa el retardo espacial de la variable dependiente en las observaciones adyacentes a la muestra i. Si la matriz de pesos está estandarizada, las variables de retardo espacial expresan el valor medio de una variable dada en el espacio de adyacencias.

[11]De aquí en adelante, cuando aparezca la notación Wy, W podrá hacer referencia tanto a la matriz de pesos espaciales como a la matriz de pesos espaciales ponderada. Es una cuestión de notación, ya que en todas las fórmulas que veremos más adelante no se distingue entre W y $\tilde{W}$.

Capítulo 3

Efectos espaciales

La localización de unas determinadas observaciones muestrales a lo largo de una superficie definida presenta dos tipos de efectos espaciales, la heterogeneidad espacial y la autocorrelación espacial.

La *heterogeneidad espacial* aparece cuando una determinada variable se distribuye de manera distinta sobre el espacio, siendo palpable en situaciones del tipo centro-periferia o norte-sur, por poner algunos ejemplos. En un contexto formal, podemos definir la heterogeneidad espacial como la inestabilidad estructural en forma de varianza no constante de los residuos de una regresión, heteroscedasticidad, o en los coeficientes del modelo. En gran parte de la literatura se suelen encontrar referencias a la autocorrelación espacial y no a la posible heterogeneidad de los datos espaciales, este hecho se debe a que la heterogeneidad espacial no necesita de unos métodos específicos para su detección y tratamiento, ya que se puede realizar mediante técnicas econométricas tradicionales. En cambio, Luc Anselin plantea tres razones por la que afrontar el efecto de heterogeneidad mediante técnicas específicas de la econometría espacial:

1. La estructura que subyace es de carácter geográfico, ya que la localización de las observaciones es fundamental para determinar la forma o especificación de dicha variabilidad.

2. Dado que la estructura es espacial, la heterogeneidad suele producirse junto con la autocorrelación espacial, con lo que los contrastes típicos de heterocedasticidad pueden contener sesgo en un contexto espacial.

3. La autocorrelación y la heterogeneidad espacial deben estructurarse perfectamente para poder identificar correctamente los parámetros de un modelo con estos problemas.

Las causas de aparición de heterogeneidad espacial en un modelo de regresión espacial serían las siguientes:

- Utilizar datos procedentes de diferentes áreas o extensión territorial, como comunidades autónomas, provincias, comarcas, etcétera.

- Cuando existe un fenómeno en las observaciones muestrales y se distribuye de manera desigual en el espacio, norte-sur o centro-periferia. También puede darse cuando se trabaje con datos relativos a antiguas áreas metropolitanas, áreas censales, etcétera.

- Causas de tipo sociológico, como por ejemplo la existencia de unos determinados gustos políticos.

- Las causas habituales de heterocedasticidad en los modelos de regresión lineal, tales como omisión de variables relevantes u otro tipo de especificación errónea del modelo que produzca en el término de error una varianza no constante.

En cambio, el concepto de *autocorrelación espacial* tiene su origen en otro concepto, el de homogeneidad. La homogeneidad hace referencia a una determinada distribución geográfica en la que los valores de las variables se asemejan, presentan características comunes, estructuras similares, etcétera. Esta similitud suele estar dada por el proceso de generación o construcción de los datos como, por ejemplo, un determinado evento ocurrido en la historia. La medición de la autocorrelación pretende determinar si existe, o no, algún tipo de dependencia espacial para poder comprender el comportamiento de una determinada variable en el espacio. Según establece el geógrafo Michael F. Goodchild en su artículo "A spatial analytical perspective on geographical information systems (1987)", la autocorrelación espacial no es más que la concentración o dispersión de los valores de una variable en un mapa, reflejando el grado en que los objetos o actividades en una unidad geográfica son similares a otros objetos o actividades en unidades geográficas próximas.

Podemos identificar autocorrelación espacial en nuestros datos cuando los valores de una variable están relacionados linealmente, parcialmente o totalmente al valor de esa variable en una localización cercana. Estableciendo un contraste de hipótesis, la hipótesis nula H_0 correspondería a la ausencia de autocorrelación espacial, proponiendo una independencia entre los valores de la variable de estudio en una localización determinada, y_i, y el resto de valores de la misma variable en zonas adyacentes, y_j. A continuación, se exponen los tres casos que pueden aparecer al estudiar la autocorrelación espacial:

- Autocorrelación espacial positiva: Existe una autocorrelación espacial positiva cuando una variable determinada, y_i, que presenta un valor por debajo de la media, está rodeada por otras variables, y_j, que también contienen un valor por debajo de la media. Dentro de este supuesto entra el conocido como efecto desbordamiento (Spillover effect), producido en muchos fenómenos socioeconómicos en los que su presencia en una región, es motivo de propagación a regiones adyacentes beneficiando así la concentración de dicho fenómeno en la zona. En este caso, el diagrama de Moran, del que se hablará posteriormente en la sección 3.1.2, es equivalente a un diagrama de dispersión entre dos variables, x e y, para las que la línea de tendencia es positiva.

- Autocorrelación espacial negativa: Este caso es el antagónico al expuesto anteriormente, es decir, cuando una variable determinada, y_i, presenta valores por encima del valor medio y está rodeada por otras variables, y_j, que presentan valores por debajo de la media, y viceversa. Una analogía la podemos encontrar en un tablero de ajedrez, donde los lados de determinada casilla, blanca o negra, son colindantes a casillas del color opuesto. Gráficamente esta situación se puede comparar a cuando en un diagrama de dispersión entre dos variables x e y, se muestra una tendencia lineal negativa.

- Ausencia de correlación espacial: La presente situación se contempla cuando los valores de las variables a estudio son independientes en el espacio.

La autocovarianza espacial se puede expresar formalmente por la condición

$$Cov[y_i, y_j] = E[y_i y_j] - E[y_i]E[y_j] \neq 0, \quad \forall\, i \neq j,$$

donde y_i representa el valor de la variable dependiente en una determinada localización e y_j representa el valor de la misma variable en los adyacentes a la localización i.

El problema que se plantea puede observarse desde dos perspectivas. En primer lugar, si extraemos una muestra de N localizaciones de un proceso aleatorio autocorrelado espacialmente, obtendremos una muestra de tamaño uno para cada localización[1]. A menos que tengamos datos de panel, no hay forma posible de aumentar el tamaño de la muestra. En segundo lugar, la matriz de covarianzas para nuestra muestra es de tamaño $(N \times N)$ y, claramente, será imposible poder estimar esa cantidad de términos, siendo la única posibilidad el imponer una estructura *a priori* en el problema para hacerlo manejable. Una forma lógica de hacerlo es asumir algún patrón sistemático de covarianzas espaciales (autocorrelaciones), cuidadosamente parametrizadas.

Para finalizar, es importante realizar unas precisiones respecto a la autocorrelación espacial:

- En primer lugar, la autocorrelación espacial es similar a la autocorrelación temporal pero más complicada. La razón, es que la autocorrelación temporal tiene la peculiaridad de que es unidireccionalidad, es decir, el pasado explica el presente. Pero la autocorrelación espacial no tiene esta limitación ya que es multidireccional y, siguiendo el ejemplo, podemos decir que lo ocurrido en ese espacio temporal puede estar influenciado por lo ocurrido en el pasado y lo que sucederá en un futuro. Debido a esta peculiaridad, no podemos trasladar sin más modelos de autocorrelación temporal al contexto espacial.

- La detección de autocorrelación espacial en una variable no tiene por qué suponer un problema, puesto que la única información que obtenemos de esta situación, por el momento, es que esa variable está estructurada espacialmente. La autocorrelación espacial será un problema cuando la detectemos en nuestros residuos, puesto que se incumple el supuesto de independencia entre las observaciones muestrales, tal y como ocurre en la estadística tradicional. También se debe señalar que J. LeGallo establece en su libro "Econométrie spatiale:

[1] Comparando una única extracción de una distribución normal bivariada obtenemos dos valores, pero de una única extracción.

L'autocorrélation spaitale dans les modèles de régression linéaire (2002)", que la presencia de autocorrelación espacial puede provenir de dos fuentes. La primera está relacionada con el hecho de que los datos espaciales estén generados por procesos que impliquen una relación entre las diversas localizaciones, por ejemplo, la construcción de un vecindario con casas adosadas (misma estructura e infraestructura, precio muy similar, etcétera). La segunda fuente está relacionada con una mala especificación en la forma o en la transformación de la variable a estudio. Esta mala especificación debería acarrear la generación de una autocorrelación espacial. Una opción es tratar de identificar variables que se hayan podido omitir e incluirlas en una regresión múltiple, ya que variables perdidas u omitidas pueden ser causa de autocorrelación espacial.

- El concepto de autocorrelación espacial tiene una importancia altísima en cuanto a la construcción de un modelo espacial se refiere, puesto que determina la fuerza del efecto espacial en cualquier variable utilizada en el modelo, identifica una posible interacción espacial en un modelo autorregresivo, permite identificar la fuerza de asociación de una variable entre unidades espaciales y ayuda en el estudio de valores perdidos, entre otras muchas cuestiones.

3.1. Detección de autocorrelación espacial. Test de Moran

Existen dos tipos de perspectivas para la correcta detección de autocorrelación espacial en nuestros datos, una perspectiva global y otra local. Éstas, se diferencian por el alcance o escala del análisis que se necesite realizar y ambas contienen los estadísticos necesarios para evaluar la existencia de grupos, en la distribución espacial para una determinada variable. Los estadísticos más utilizados son el estadístico I[2] de Moran y el C de Geary. Debido a que el estadístico de Moran es la técnica más antigua y típica para la detección y medición de la autocorrelación espacial, será el que se exponga en profundidad en el presente trabajo.

[2]Para poder diferenciar la notación de la matriz identidad, I, que aparecerá en la formulación de diversos modelos estadísticos de la notación del estadístico o índice de Moran, I, denotaremos este último en cursiva.

Es interesante señalar que los estadísticos utilizados en la localización de autocorrelación espacial poseen un origen común en su construcción, basado en un término de producto cruzado, el estadístico de Mantel y cuya formulación queda expresada como

$$\Gamma = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} c_{ij}.$$

El test de Mantel es una regresión en la que las variables son, a su vez, matrices de distancia o disimilitud que resumen las similitudes de parejas entre las ubicaciones de la muestra. Donde w_{ij} es, por supuesto, el elemento i-ésimo y j-ésimo ubicado en la matriz de pesos espaciales y c_{ij} es un índice de semejanza, similitud o disimilitud, que nos permitirá estimar la diferencia entre los valores que contiene la variable explicada y, como puede ser la tasa de paro, sueldo medio, etcétera, en la interacción entre el entorno i, y_i, y su entorno adyacente j, y_j, siendo por supuesto $i \neq j$.
La diferencia entre los distintos estadísticos viene marcada en la forma en la que calculan el índice de semejanza c_{ij}. Por ejemplo, el estadístico c de Geary se construye utilizando una medida de asociación, en la que la similitud de los valores de la variable y se mide mediante la diferencia entre los valores de dicha variable en las diferentes localizaciones y_i e y_j, siendo la diferencia al cuadrado para evitar un posible valor negativo, siendo

$$c_{ij} = (y_i - y_j)^2.$$

En cambio, el estadístico de Moran está basado en una medida de tipo covarianza, donde la similitud entre los distintos valores de la variable y queda definido como el producto de las diferencias entre los valores de la variable dependiente, en los distintos entornos i, y_i, y j, y_j, y el valor medio de la variable objetivo, siendo

$$c_{ij} = (y_i - \bar{y})(y_j - \bar{y}). \tag{3.1}$$

3.1.1. Autocorrelación espacial global

El análisis de autocorrelación espacial global, implica la realización de un estudio centrado en la asociación de todas las observaciones existentes en los datos, para poder determinar si la composición de las unidades espaciales es aleatoria o, en

cambio, están localizadas de acuerdo a un tipo de patrón o clúster.

Rosina Moreno y Esther Vayá establecen en su artículo "La utilidad de la econometría espacial en el ámbito de la ciencia regional (2000)" que los estadísticos de autocorrelación espacial global, son siempre las primeras formulaciones que aparecen en la literatura para la medición del efecto de autocorrelación espacial y de entre todas ellas, destaca el estadístico I de Moran.

El estadístico I de Moran se puede aplicar a todas las variables que miden fenómenos continuos. Además, al tratarse de un estadístico deductivo, interpretaremos los resultados obtenidos dentro del contexto de hipótesis nula, H_0, estableciendo que la variable a estudio está distribuida de forma aleatoria, es decir, no existe autocorrelación espacial. Dicho estadístico presenta la siguiente formulación

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \quad \forall\, i \neq j.$$

Si se observa la fórmula podemos comprobar que no es más que un índice de correlación, pudiendo apreciarlo de forma mucho más clara si estructuramos la ecuación de la siguiente forma

$$I = \frac{\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}, \quad \forall\, i \neq j,$$

donde, en efecto, el numerador de la ecuación se corresponde a una covarianza que tiene en cuenta el peso existente entre las relaciones de adyacencia de la variable objetivo en nuestros datos y en el denominador se realiza una división por la varianza de dicha variable.

Además, si trabajamos con una matriz de pesos espaciales estandarizada, esta es otra de sus grandes bondades, el término $\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$ será equivalente al tamaño de las observaciones en nuestros datos, es decir, N. Por lo que el tiempo de computación quedaría reducido al quedar la fórmula establecida[3]

$$I = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \quad \forall\, i \neq j. \tag{3.2}$$

[3]Si nos fijamos en el numerador de la ecuación (3.2) podemos comprobar como no son más que las similitudes de parejas entre las ubicaciones de la muestra, es decir, el estadístico de Mantel donde se ha sustituido el término c_{ij} por su valor en la ecuación (3.1).

Cabe destacar que la media teórica del índice de Moran no se encuentra centrada en cero, como ocurre con el coeficiente de correlación de Pearson, dado que su valor esperado se obtiene,bajo la hipótesis nula de no autocorrelación espacial, mediante la expresión

$$E(I) = \frac{-1}{N-1}.$$

Es decir, el valor medio del estadístico de Moran tendrá un valor negativo que tenderá a cero cuanto más grande sea el tamaño de la muestra. Este valor esperado puede ser un indicador inicial del tipo de autocorrelación existente en nuestros datos, puesto que un coeficiente I de Moran mayor que su valor esperado indicaría una autocorrelación espacial positiva, en cambio, un valor inferior al de la media indicaría la presencia de autocorrelación espacial negativa.

La varianza del estadístico de Moran tiene una formulación mucho más compleja que la que observada en su media, siendo[4]

$$Var(I) = \frac{N^2 S_1 - N S_2 + 3S_0^2}{S_0^2(N^2-1)} - [E(I)]^2, \quad \forall\, i \neq j.$$

Siendo

$S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$,

$S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_{ij} + w_{ji})^2$,

$S_2 = \sum_{i=1}^{N} (w_{i.} + w_{.i})^2$,

$w_{i.} = \sum_{j=1}^{N} w_{ij}$,

$w_{.i} = \sum_{j=1}^{N} w_{ji}$.

Debemos tener en cuenta que este test tiene únicamente un valor asintótico, por lo que su aplicación requiere de un considerable número de observaciones. De acuerdo con A. Cliff y J. Ord en su libro de 1981 "Spatial processes, models and applications", cuando el tamaño muestral es suficientemente grande, la expresión estandarizada del estadístico de Moran se distribuye como una normal estándar,

[4]En la literatura se pueden observar infinidad de variantes para la ecuación de la varianza, siendo ésta la de más fácil interpretación a criterio del autor del presente trabajo.

$N(0, 1)$, tomando la forma

$$z = \frac{I - E(I)}{\sqrt{Var(I)}}.$$

Tras la obtención del estadístico z, rechazaremos la hipótesis nula, H_0, cuando su valor absoluto sea mayor que el cuantil de la distribución normal estándar que proporcione el error de tipo I, α, que se haya establecido previamente. El valor de α que se suele utilizar con mayor frecuencia, es aquel con un valor de significación del 95 %, es decir, $\alpha = 0{,}05$ con $z_\alpha = 1{,}96$. Una vez conocida la existencia de autocorrelación espacial, ésta será positiva cuando el valor z obtenido sea positivo y, por lo tanto, negativa cuando su valor sera negativo. En la tabla 3.1 se puede encontrar un breve resumen de lo anteriormente expuesto:

No se rechaza H_0 ($\lvert z\rvert < z_\alpha$)	Se rechaza H_0 $z > z_\alpha$	Se rechaza H_0 $z < -z_\alpha$
No existe autocorrelación espacial	Autocorrelación espacial positiva	Autocorrelación espacial negativa

Tabla 3.1: Interpretación de los valores estandarizados del estadístico de Moran. Fuente: Elaboración propia.

Para concluir el presente apartado, podemos señalar las siguientes cuestiones:

- Debemos tener cuidado cuanto utilicemos el estadístico I de Moran en un entorno muestral pequeño, ya que como se ha comentado, éste tiene un valor asintótico. Aunque L. Anselin y R. Florax en su libro "New Directions in Spatial Econometrics (1995)" presentan un análisis basado en la simulación de las propiedades del estadístico de Moran en muestras pequeñas, concluyendo que funciona bastante bien en dicho supuesto.

- El estadístico de Moran asume que cualquier otro tipo de tendencia existente en nuestros datos ha sido eliminada, ya que en caso contrario el test podría arrojar resultados falsos. Los autores R. Bivand, E. Pebesma y V. Gómez-Rubio muestran en su texto "Applied Spatial Data Analysis with R (2008)" un ejemplo donde no existe autocorrelación espacial en los datos. En cambio, existe una tendencia lineal simple en la que el test de Moran indica, incorrectamente, la presencia de autocorrelación espacial[5].

[5]En tal supuesto el problema queda resuelto al eliminar tal tendencia en los datos.

3.1.2. Autocorrelación espacial local. Indicadores locales de asociación espacial

La dependencia o autocorrelación espacial local puede definirse como la concentración de valores, ya sean altos o bajos, de la variable dependiente en el análisis respecto de su valor medio. Tanto los análisis locales como los globales, no tienen por qué ser excluyentes entre sí, puesto que cabe la posibilidad de no detectar una dependencia espacial global en la variable objetivo, pero sí existan pequeños clústers espaciales en los que la variable en cuestión presente una concentración o escasez importante, los conocidos como puntos calientes y fríos. Además, si comprobamos en el estudio la existencia de autocorrelación espacial global, podemos realizar un análisis local para comprobar qué observaciones en ese clúster son las que aportan valores más altos o más bajos a la autocorrelación espacial observada.

L. Anselin definió en su artículo "Local Indicators of Spatial Association-LISA (1995)" un conjunto de indicadores locales de asociación espacial, conocidos comúnmente como LISA por su acrónimo en inglés, capaces de descomponer los estadísticos de autocorrelación espacial globales y así realizar un análisis para cada una de las observaciones i. Para el autor, un LISA es aquel estadístico que pueda reunir los siguientes requisitos:

1. Para una determinada observación, LISA debe ser capaz de determinar la extensión de la agrupación espacial significativa, clúster, alrededor de la misma.

2. La suma de los indicadores locales de asociación espacial, debe proporcionar un índice proporcional al indicador global de asociación espacial.

Uno de estos estadísticos LISA, es el I de Moran local denotado como I_i y definido mediante la ecuación,

$$I_i = (y_i - \bar{y}) \sum_{j=1}^{N} w_{ij}(y_j - \bar{y}), \quad \forall\, i \neq j.$$

Además, bajo la hipótesis nula, H_0, la esperanza y la varianza del I_i de Moran quedan definidas, respectivamente como

$$E(I_i) = \frac{-\sum_{j=1}^{N} w_{ij}}{(N-1)}, \quad \forall\, i \neq j,$$

$$Var(I_i) = \frac{\sum_{j=1,j\neq i}^{N} w_{ij}^2(N-b_2)}{N-1} + \frac{2w_{i(kh)}(2b_2-N)}{(N-1)(N-2)} - \frac{(\sum_{j=1}^{N} w_{ij})^2}{(N-1)^2}, \quad \forall\, i \neq j,$$

donde

$$b_2 = m_4/m_2^2,$$

$$m_4 = \sum_{i=1}^{N}(y_i - \bar{y})^4/N,$$

$$m_2 = \sum_{i=1}^{N}(y_i - \bar{y})^2/N,$$

$$2w_{i(kh)} = \sum_{k=1}^{N}\sum_{h=1}^{N} w_{ik}w_{ih}, \quad \forall\, k, h \neq i.$$

Para finalizar, y al igual que en el estadístico a nivel global, el valor z local, z_i, es el que viene determinado por la ecuación

$$z_i = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}}.$$

La dependencia local puede ser graficada mediante el diagrama de dispersión de Moran. Este diagrama permite graficar en un plano cartesiano la variable dependiente normalizada en el eje de abscisas y su retardo espacial en el eje de ordenadas, Wy. Además, la pendiente de la recta de regresión es el valor del estadístico I de Moran de autocorrelación espacial global, de forma que cuanto mayor sea el ángulo que forme con el eje de abscisas, más fuerte será el grado de autocorrelación espacial, y viceversa. El diagrama queda dividido en cuatro cuadrantes que reproducen los diferentes tipos de asociación espacial. A continuación se explican los distintos cuadrantes y se muestra una simulación realizada en R, mediante el paquete ggplot2, del diagrama de dispersión de Moran. En el diagrama se represente la variable objetivo, precio de la vivienda, respecto al valor de sus adyacentes:

- El cuadrante Alto-Alto[6] (AA) se corresponde al caso en el valor de la variable objetivo en la observación i es alto y está rodeado de observaciones con un valor alto.

- El cuadrante Bajo-Bajo (BB) se corresponde al mismo caso que el anterior pero en un contexto de valores negativos.

[6]Las referencias "Alto" y "Bajo" hace referencia al valor que toman las distintas variables que se grafican.

- El cuadrante Alto-Bajo (AB) se corresponde a la situación donde un valor alto en el valor de la variable objetivo en la observación i es alto y el valor medio de sus adyacentes es bajo.

- El cuadrante Bajo-Alto (BA) se corresponde a la situación inversa descrita para el cuadrante (AB), donde un valor bajo de la variable explicada está asociado a valores altos en sus adyacentes.

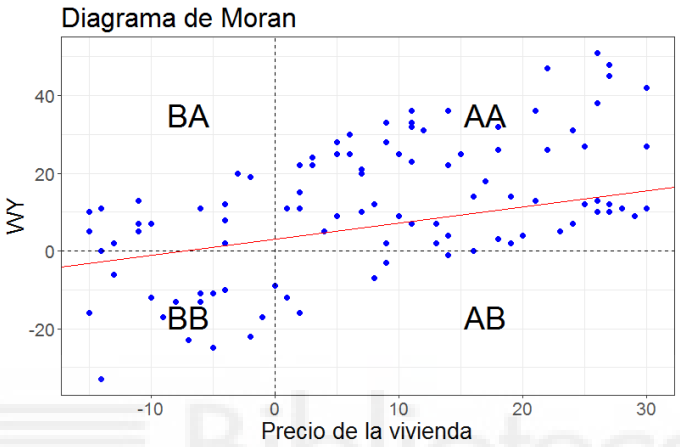


Figura 3.1: Diagrama de dispersión Moran (simulado).
Fuente: Elaboración propia.

Es fácil observar que cuando la mayoría de las observaciones se encuentran alojadas en los cuadrantes AA y BB existe una autocorrelación espacial global positiva. Mientras que cuando la mayoría de las observaciones estén ubicadas en los cuadrantes BA y AB existe una autocorrelación espacial global negativa. Además, en caso de tener una correlación espacial positiva, las observaciones que se encuentren presentes en los cuadrantes AB y BA se considerarán como atípicas. Aquellas localizadas en el cuadrante AB se suelen denotar como diamantes en bruto (diamonds in the rough) y a las ubicadas en el cuadrante BA como ovejas negras (black sheeps)[7].

[7]Diamantes en bruto son aquellas observaciones que tienen un valor alto para la variable dependiente y sus adyacentes un valor bajo. En cambio las ovejas negras es el caso invereso

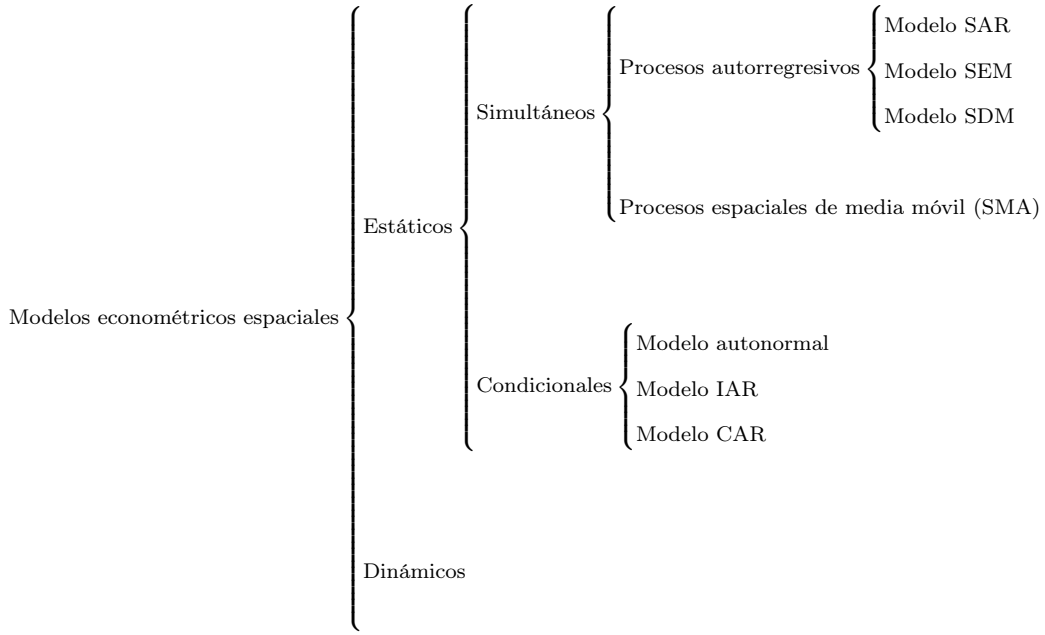
Capítulo 4

Modelos de procesos espaciales

Como se expuso en la introducción del trabajo, multitud de disciplinas utilizan la estadística espacial para sus análisis, como la econometría, ecología, geografía, etcétera. Por una parte, esta cuestión presenta grandes ventajas debido a la sinergia que se puede establecer entre las distintas disciplinas, por ejemplo, en cuanto a la investigación se refiere. En cambio, también presenta una gran desventaja, la exposición teórica que podemos encontrar de este tipo de análisis en la literatura. Esta confusión tiene un auge importante en la modelización, llegando al punto en el que un número importante de libros y artículos revisados para la elaboración del presente trabajo, algunos incluidos en la bibliografía, establecen como semejantes a modelos tan diferentes en su definición teórica como son el modelo de autorregresión espacial (SAR) y el modelo condicional autorregresivo (CAR).

Para la elaboración del presente trabajo, se ha pretendido establecer una estructuración lo más clara posible para la exposición de los distintos modelos econométricos espaciales. Tal estructuración se ha basado, principalmente, en el capítulo 15 del texto "Handbook of Spatial Statistics (2010)" elaborado por R.K. Pace y J. LeSage, en el que se propone una primera división entre modelos estáticos y modelos dinámicos y, a su vez, las clases impartidas para la asignatura "Spatial Regression Analysis" por el profesor L. Anselin en la Universidad de Chicago durante la primavera de 2017. En dicha asignatura, se establecía una clara división en los modelos estáticos, siendo clasificados entre modelos de procesos espaciales simultáneos y modelos de procesos espaciales condicionales.

Mediante el esquema que se muestra a continuación, el lector podrá tener una visión clara del contenido del presente capítulo[1]:



Fuente: Elaboración propia.

La división entre modelos estáticos y dinámicos reside en la cantidad de dimensiones que componen la muestra a estudio, ya que consideraremos *modelos estáticos* a aquellos que trabajan bajo una dimensión espacial de los datos, es decir, están establecidos en un espacio previamente delimitado. En cambio, los *modelos dinámicos* se utilizan cuando los datos a estudio presentan, además de la dimensión espacial, una dimensión temporal. El añadir un ámbito temporal a los datos, aporta una complejidad considerable dentro del análisis espacial ya que se requiere de la utilización de modelos espacio-temporales, como el modelo autorregresivo espacio-temporal (STAR), por lo que este tipo de modelos no se estudiarán en el presente trabajo.

Antes de comenzar los apartados referidos tanto a los modelos de procesos espaciales simultáneos como a los modelos de procesos espaciales condicionales, puede resultar interesante señalar qué es un modelo de proceso espacial. Un *modelo de proceso espacial* o, formalmente denominado, modelo de proceso estocástico espacial es

[1]En la literatura se puede encontrar una gran variedad de modelos econométricos espaciales pero, la mayoría de los mismos, tienen sus raíces en un determinado número de modelos que no son otros que los que se presentan en este capítulo.

aquel capaz de cuantificar la variabilidad aleatoria de una variable determinada en el espacio. También, en un contexto más formal, podemos señalar que la estructura de la variable de interés debe seguir lo que especifica el proceso estocástico, ya que existe una diferencia entre la especificación del proceso, el modelo y la estructura de covarianzas que resulta del mismo. Dicha estructura de covarianzas es más densa debido a que relaciona todas las observaciones entre sí, no como ocurre con la matriz de pesos en la que podemos comprobar la estructura relacional existente entre las distintas observaciones a estudio, es decir, la dispersión existente entre observaciones. Por lo tanto, la estructura de covarianzas será más compleja que la presentada por los pesos especificados en el modelo.

4.1. Modelos de procesos espaciales simultáneos

Los procesos simultáneos son aquellos que explican de forma paralela una misma variable aleatoria, ya sea la variable dependiente, el término del error, etcétera. Es decir, en estos modelos podemos encontrar a la variable objetivo en ambos lados de la ecuación, tanto en el izquierdo reservado típicamente para esta variable como en el derecho donde tradicionalmente se encuentran las variables explicativas o independientes. Esta peculiaridad es la que lleva a muchos autores a confundir los procesos simultáneos con los condicionales, puesto que al tener ubicada la variable dependiente en ambos lados de la ecuación se puede pensar que el valor de la variable dependiente en la parte derecha está explicando a su homóloga en la parte izquierda de la ecuación, es decir, que con su valor está condicionando el valor de la variable dependiente izquierda. En cambio, la explicación es mucho más sencilla y reside en que para poder explicar el valor de la variable dependiente, de la parte izquierda, necesitamos conocer el valor de dicha variable dependiente en localizaciones adyacentes a la misma y, a su vez, estas variables en localizaciones adyacentes se encontrarán en la parte izquierda de la ecuación en otras observaciones, es decir, soy el vecino de mi vecino. Nunca vamos a poder explicar el comportamiento de una variable dependiente sin conocer el comportamiento de dicha variable en sus adyacentes, ahí recae el concepto de simultaneidad. También debemos realizar una apreciación muy importante en cuanto a la ubicación de las variables dependientes,

puesto que lo comentado hasta ahora hace referencia a la parte inicial del modelo, ya que siempre se establecerá una forma reducida del mismo para, de esta manera, no tener ubicada la variable dependiente en ambos lados de la ecuación y así poder explicar únicamente en función de las variables independientes.

Podemos distinguir entre dos tipos de modelos de procesos espaciales simultáneos, los *procesos autorregresivos espaciales*[2] y los *procesos espaciales de media móvil*. La diferencia entre ambos recae en que los procesos autorregresivos espaciales contienen la misma variable aleatoria dependiente en el lado izquierdo de la ecuación y en el derecho de la misma. En cambio, en los procesos espaciales de media móvil lo que se busca es suavizar el efecto del término de error en la observación, por lo que se modeliza junto a un error medio en observaciones adyacentes, tomando la formulación

$$y_i = \gamma \sum_{j=1}^{N} w_{ij}\epsilon_j + \epsilon_i, \quad \forall\, i \neq j. \tag{4.1}$$

Además, puede ser expresada matricialmente mediante la ecuación

$$y = \gamma W \epsilon + \epsilon,$$

teniendo su forma reducida en la expresión

$$y = (\mathrm{I} + \gamma W)\epsilon.$$

Como se puede observar en la ecuación (4.1), tenemos una variable aleatoria, y, en una determinada observación, i, en función de dos tipos de errores, uno en dicha observación, ϵ_i, y también una media suavizada de los errores en las observaciones adyacentes, $\sum_{j=1}^{N} w_{ij}\epsilon_j$. Además, incluye el parámetro γ como efecto de dependencia espacial entre las distintas observaciones i y j. Expuesta la ecuación, podemos entender que tal especificación puede resultar inútil debido a que únicamente contamos con errores en la parte explicativa, no tenemos datos, por lo que se le considera un modelo de influencia espacial. Ésto es lo que lo hace tan diferente de un modelo

[2]Debido al uso generalizado de los procesos autorregresivos espaciales, el presente trabajo contempla una sección dedicada a este tipo de modelos. Además, se debe puntualizar que por tener un uso más generalizado no son mejores que los procesos espaciales de media móvil, ni tampoco peores. Todo recae en el criterio del investigador, siendo recomendable el plantear una modelización de los datos respecto a ambos procesos espaciales y decidir cuál utilizar en función de los resultados obtenidos y unos criterios previamente establecidos.

autorregresivo.

4.1.1. Modelos autorregresivos espaciales

Como se ha explicado, este tipo de modelo tiene la peculiaridad de contar con la variable aleatoria, dependiente, en ambos lados de la ecuación y la mejor forma de exponer este tipo de modelos es desde su origen, que no es otro que el modelo normal de regresión múltiple, cuya expresión viene dada por

$$\begin{aligned} &y_i = x_i\beta + \epsilon_i, \\ &\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall\, i \in \{1, ..., N\}. \end{aligned} \tag{4.2}$$

En la presente ecuación, y_i representa el valor de la variable dependiente en una determinada localización, β es el parámetro a estimar que representa el coeficiente de regresión y, por último, ϵ_i es la variable aleatoria que encarna el término de error o perturbación en el modelo para dicha localización, con media 0 y varianza constante σ^2.

En esta situación, los valores en la observación i serán siempre independientes de aquellos en otra observación, j, por lo que las perturbaciones serán independientes entre ellas. Por lo que en un contexto más formal, podemos señalar que observaciones independientes o estadísticamente independientes implican que

$$E[\epsilon_i\epsilon_j] = E[\epsilon_i]E[\epsilon_j] = 0.$$

La suposición de independencia simplifica mucho los modelos, pero en un contexto espacial representa un sinsentido, únicamente tenemos que pensar en la primera ley de la geografía de Tobler que explica que todo está relacionado entre sí, pero lo está más aún con aquello más cercano.

La dependencia espacial refleja una situación donde valores observados en una localización i depende de los valores que contienen localizaciones adyacentes. Suponiendo que las observaciones $i = 1$ y $j = 2$ son adyacentes, en tal caso, podemos

ampliar la ecuación (4.2) de la siguiente manera:

$$\begin{aligned} y_i &= \alpha_i y_j + x_i\beta + \epsilon_i, \\ y_j &= \alpha_j y_i + x_j\beta + \epsilon_j, \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2), \quad i = 1, \\ \epsilon_j &\sim \mathcal{N}(0, \sigma^2), \quad j = 2. \end{aligned}$$

En esta situación, el valor que toma la variable y_i depende de la variable y_j y viceversa, por lo que el proceso generador de datos es simultáneo. Añadiendo una nueva observación, k, adyacente a las observaciones i y j seremos conscientes del problema que se plantea en estos casos:

$$\begin{aligned} y_i &= \alpha_{i,j} y_j + \alpha_{i,k} y_k + x_i\beta + \epsilon_i, \\ y_j &= \alpha_{j,i} y_i + \alpha_{j,k} y_k + x_j\beta + \epsilon_j, \\ y_k &= \alpha_{k,i} y_i + \alpha_{k,j} y_j + x_k\beta + \epsilon_k, \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2), \quad i = 1, \\ \epsilon_j &\sim \mathcal{N}(0, \sigma^2), \quad j = 2, \\ \epsilon_k &\sim \mathcal{N}(0, \sigma^2), \quad k = 3. \end{aligned}$$

Como podemos comprobar, deriva en un sistema con muchos más parámetros que observaciones. Esto se debe a que una vez habiendo permitido una relación de dependencia entre un conjunto de observaciones N, pueden aparecer una cantidad de $N^2 - N$ relaciones.[3] La solución al problema de sobreparametrización surgido al permitir que cada observación posea una relación de dependencia con observaciones adyacentes, es imponer una estructura de dependencia espacial en tales relaciones. En 1975, J.K. Ord propuso una parametrización parsimoniosa para las relaciones de dependencia basada en los primeros trabajos de Whittle en 1954 al respecto. Esta estructuración da lugar a un proceso generador de datos conocido como *proceso autorregresivo espacial*. Este proceso aplica las relaciones de dependencia entre las distintas observaciones respecto a una determinada variable aleatoria, obteniendo la

[3]La diferencia entre N^2 y N se realiza para eliminar la dependencia entre una relación y sí misma.

expresión

$$y_i = \rho \sum_{j=1}^{N} w_{ij} y_j + \epsilon_i,$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall\, i \neq j;\ i, j \in \{1, ..., N\},$$

donde se ha eliminado el término intercepto al suponer que el vector de observaciones sobre la variable y se encuentra desviado de la media. El término $\sum_{j=1}^{N} w_{ij} y_j$ es el denominado *retardo espacial* y no es más que una combinación lineal que representa la media de los valores de la variable y en las adyacencias de la observación i. Además, el parámetro escalar ρ describe la fuerza de dependencia espacial en la muestra de observaciones.

Podemos exponer una versión matricial de la ecuación del proceso autorregresivo espacial de la siguiente forma:

$$\begin{aligned} y &= \rho W y + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N). \end{aligned} \tag{4.3}$$

Como se puede comprobar, los términos de error siguen una distribución normal de media 0 y matriz de varianzas covarianzas, con covarianzas 0 entre las distintas observaciones. Además, podemos especificar la ecuación en su forma reducida mediante la expresión

$$y = (\mathrm{I}_N - \rho W)^{-1} \varepsilon.$$

4.1.1.1. Modelo de autorregresión espacial (SAR)

La ecuación referente a un proceso autorregresivo espacial (4.3) puede combinarse con la ecuación matricial de un modelo de regresión estándar para, de esta manera, conseguir que el valor de la variable objetivo en la observación i dependa del valor de la variable independiente, x, en dicha observación y del valor de la variable objetivo en el resto de localizaciones, j, de la muestra. Mediante esta combinación obtenemos el modelo de autorregresión espacial (SAR), también conocido como modelo de retardo espacial (SLM) o, incluso, modelo mixto autorregresivo de regresión espacial,

tomando la forma matricial

$$\begin{aligned} y &= \rho W y + X\beta + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N), \end{aligned} \tag{4.4}$$

cuya notación viene dada por:

y: Vector de tamaño $(N \times 1)$ que contiene la variable dependiente.

X: Matriz de observaciones de tamaño $(N \times K)$ de la variable independiente.

β: Vector de tamaño $(K \times 1)$ de coeficientes de regresión.

ε: Vector de términos del error aleatorios, independientes e identicamente distribuidos de tamaño $(N \times 1)$.

W: Matriz de pesos espaciales de tamaño $(N \times N)$ que especifica la estructura espacial existente en los datos.

Wy: Vector de tamaño $(N \times 1)$ que contiene las variables dependientes espacialmente retardadas. Es decir, representa el retardo espacial de la variable dependiente en las localizaciones adyacentes a cada observación.

ρ: Parámetro o efecto de dependencia espacial[4] que aporta la intensidad de la dependencia entre adyacentes a cada observación respecto a la variable objetivo. Algo a señalar en cuanto a este parámetro, es que cuando posee un valor equivalente a cero, el modelo SAR se convierte en un modelo de regresión estándar.

Quedando expresada en su forma reducida mediante la igualdad

$$y = (\mathrm{I}_N - \rho W)^{-1} X\beta + (\mathrm{I}_N - \rho W)^{-1}\varepsilon.$$

Al término $(\mathrm{I}_N - \rho W)^{-1}$ se le conoce como *multiplicador espacial* o *filtro espacial*, ya que puede ser ampliado en una serie de potencias, tal que

$$(\mathrm{I}_N - \rho W)^{-1} = \mathrm{I}_N + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots ,$$

[4]Por tradición se utiliza el parámetro ρ en el modelo SAR para denotar al parámetro de dependencia espacial, así como en el modelo SEM es el parámetro λ y en el modelo de media móvil γ.

siendo W los pesos espaciales de los adyacentes de una región dada, W^2 son los pesos de los adyacentes de los adyacentes, es decir, los pesos de los adyacentes de segundo orden y W^3 son los pesos de los adyacentes de tercer orden. Por lo tanto, cuando se observa la expresión

$$(\mathrm{I}_N - \rho W)^{-1} X\beta,$$

en realidad se examina la suma de una serie de influencias decrecientes de todo el sistema espacial

$$X\beta + \rho W X\beta + \rho^2 W^2 X\beta + \rho^3 W^3 X\beta + \dots .$$

Para finalizar, incidiremos en que el modelo SAR controla principalmente la autocorrelación espacial en la variable dependiente, por lo que se recomienda su uso para la interpretación y medida del denominado efecto desbordamiento (spillover effect).

4.1.1.2. Modelo de error autorregresivo espacial (SEM)

El modelo de error autorregresivo espacial considera que existen ciertos factores o variables no consideradas en el modelo que trasladan hacia los términos del error la configuración de autocorrelación presente en la variable dependiente, es decir, plantea que quien lleva el peso del retardo en el modelo son los términos del error en una determinada observación i y no la variable dependiente en dicha observación, quedando expresado mediante la formulación

$$\begin{aligned} y &= X\beta + u, \\ u &= \lambda W u + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N). \end{aligned}$$

El término u representa el vector de tamaño $(N \times 1)$ que contiene los términos del error en las observaciones i y, evidentemente, los dos vectores de error que aparecen en el modelo, u y ε, se asumirán siempre incorrelados. Además, $\lambda W u$ representa la dependencia espacial ponderada de los términos de error de las observaciones

adyacentes.
Si nos fijamos la formulación del modelo, la primera ecuación no es más que un modelo de regresión lineal en la que hemos proporcionado una segunda ecuación con la estructura correlada en el error. Podemos expresar su forma reducida,

$$y = X\beta + (\mathrm{I}_N - \lambda W)^{-1}\varepsilon.$$

Como podemos comprobar en esta última expresión, el valor de la variable dependiente para cada localización i está afectada por los errores estocásticos en todas las localizaciones a través del multiplicador espacial.

Para concluir, el modelo de error autorregresivo espacial es indicado para aquellas situaciones en la que contemos con el efecto de heterogeneidad espacial o cuando puedan existir variables omitidas en nuestro modelo.

4.1.1.3. Modelo espacial de Durbin (SDM)

El modelo espacial de Durbin ocupa un lugar importante en el campo del análisis espacial debido a que aglutina a varios de los modelos más utilizados en la literatura. De hecho, podemos obtenerlo tanto del modelo SAR como del modelo SEM. Luc Anselin lo desarrolló a partir de la forma reducida del modelo SEM, conociéndose como el modelo espacial de Durbin clásico y cuya formulación viene dada por la expresión

$$\begin{aligned} y &= \rho W y + X\beta - \rho W X\beta + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N). \end{aligned}$$

Como se puede comprobar, en este modelo la variable dependiente en cada observación, queda explicada por el retardo espacial de la variable dependiente en sus adyacentes Wy, la variable exógena X y un nuevo término compuesto por el retardo espacial de las variables independientes WX, además del omnipresente término del error ε. También contiene como parámetros a estimar a ρ como coeficiente del retardo espacial Wy, a β como coeficiente de las variables independientes X y a ρ junto a β como coeficientes del retardo espacial WX, lo que conlleva que este modelo no sea lineal por contener el producto de dos parámetros. Aplicando la restricción de factor

común $\theta = -\rho\beta$ el modelo espacial de Durbin queda representado como[5]:

$$y = \rho W y + X\beta + WX\theta + \varepsilon,$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N).$$

La versatilidad que plantea este modelo y que lleva a diversos a autores a apostar por él, son los modelos que se pueden obtener si aplicamos una serie de restricciones:

- Imponiendo la restricción $\theta = 0$ se obtiene el modelo de autorregresión espacial (SAR) en su forma no reducida (4.4).

- La restricción $\rho = 0$ produce un modelo no expuesto en el presente trabajo, el modelo SLX (modelo de retardo espacial en la variable independiente), idóneo cuando los datos presentan externalidades espaciales.

- Además, impuesta la restricción $\rho = 0$ y $\gamma = 0$ se obtiene un modelo de regresión lineal.

Estas restricciones son también el motivo de poder utilizar este modelo tanto en aquellos casos en los que podamos observar externalidades espaciales, omisión de variables, efecto de desbordamiento o heterogeneidad espacial.

Por supuesto, podemos expresar el modelo espacial de Durbin en su forma reducida,

$$y = (\mathrm{I}_N - \rho W)^{-1}(X\beta + WX\theta) + (\mathrm{I}_N - \rho W)^{-1}\varepsilon.$$

Es notorio como el modelo permite determinar el valor de la variable dependiente en observaciones i, de acuerdo a los valores del resto de variables en observaciones adyacentes j y de los términos de error en todas las localizaciones.

[5]En la literatura podemos encontrar esta ecuación formulada de diversas formas, en cuanto a su estructura y su nomenclatura En cuanto a su estructura, porque muchos autores apuestan por extraer el término del intercepto de la matriz de variables independientes X apareciendo multiplicado por un vector de unos, $\alpha\iota_N$, quedando finalmente sumado al resto de términos de la ecuación y en cuanto a su nomenclatura, el autor de este trabajo ha apostado por utilizar el parámetro θ para la restricción de factor común, pero podemos encontrar refiriéndose a ese mismo parámetro a γ e incluso a β_2 (siendo el parámetro β "clásico", β_1), entre otros.

4.2. Modelos de procesos espaciales condicionales

Una gran diferencia entre los distintos modelos simultáneos y los condicionales recae en que los primeros tienen un carácter más inferencial al estar más centrados en la estructura espacial existente en los datos y los segundos tienen un trasfondo más predictivo, ya que realizan un análisis mucho más global de las observaciones. Visualmente comprobaremos el por qué mediante el siguiente esquema:

$$\text{Modelos simultáneos} \longrightarrow y = f(x)$$
$$\text{Modelos condicionales} \longrightarrow y_i = f(y_{i-}).$$

En los modelos condicionales la variable dependiente para una determinada observación i queda explicada en función de los valores que posee la variable dependiente en el resto de observaciones, es decir, sus adyacentes y en los modelos simultáneos, los valores en todas las localizaciones están explicados en función de los valores de las variables exógenas o independientes[6] en sus adyacentes. Los modelos simultáneos se pueden utilizar para realizar predicciones, pero se obtendrán resultados sesgados respecto a los que se pueden obtener con los modelos condicionales.
El carácter predictivo de los modelos de procesos espaciales condicionales les lleva a ser muy utilizados cuando los datos están formados por un conjunto de puntos en un espacio determinado y mediante el valor de dichos puntos se pretende predecir una superficie, es decir, tratas de modelizar una superficie continua en base a unas determinadas observaciones, es el denominado *kriging*. También son ampliamente utilizados con la modelización jerárquica bayesiana, donde existe una estructura jerárquica de parámetros en la que todo se considera aleatorio, los datos de las observaciones (puntos), los parámetros, etcétera. Esta distribución jerarquizada de parámetros dependientes entre sí es la que propicia que los modelos condicionales sean los idóneos para ser utilizados.

Los modelos condicionales son mucho más complejos y conceptuales que los modelos simultáneos como comprobaremos, debido a que la probabilidad estimada de los valores para cualquier observación, están condicionados a los valores en sus adyacentes.

[6] Hablamos en todo momento de la forma reducida del modelo.

Mediante el método de cadenas de Markov Monte Carlo (MCMC), a partir de una determinada muestra de distribuciones condicionales podemos obtener una muestra de distribuciones conjuntas, podemos plasmarlo en el contexto espacial que nos ocupa mediante el lema de Brook (Brook's Lemma). Este lema establece un enlace entre las distintas distribuciones condicionales y una distribución conjunta de las mismas a partir de una observación determinada. El lema de Brook señala que

$$\begin{aligned} p(y_1, ..., y_N) = & \frac{p(y_1|y_2, ..., y_N)}{p(y_{1,0}|y_2, ..., y_N)} \cdot \frac{p(y_2|y_{1,0}, y_3, ..., y_N)}{p(y_{2,0}|y_{1,0}, y_3, ..., y_N)} \cdot \\ & ... \frac{p(y_N|y_{1,0}, ..., y_{N-1,0})}{p(y_{N,0}|y_{1,0}, ..., y_{N-1,0})} \cdot p(y_{1,0}, ..., y_{N,0}), \end{aligned} \tag{4.5}$$

donde se fijan todos los condicionales a una distribución conjunta de una observación determinada, $y_0 = (y_{1,0}, ..., y_{N,0})'$, que denominaremos punto de fijación. Gracias a este punto de fijación podremos tener proporciones de distribuciones condicionales que nos ayudarán a obtener la distribución conjunta deseada para esa observación cero.

Si se observa la ecuación (4.5), al final de la misma se encuentra el punto de fijación, comprobando como los condicionales se van construyendo a partir de un elemento del punto de fijación y se van añadiendo gradualmente nuevos elementos del punto de fijación hasta conseguir que todos los elementos de dicho punto se encuentren en la condición. Este es un concepto bastante complejo, por lo que con un ejemplo se podrá visualizar mejor. Pensemos en las 34 comarcas de la Comunidad Valenciana, cada una de ellas estará condicionada a las 33 restantes y tomaremos un punto de fijación, por ejemplo, la comarca del Bajo Vinalopó; la idea que debemos tener en mente ahora es que ahora las demás comarcas estarán más y más condicionadas a lo que podemos observar respecto a la comarca del Bajo Vinalopó, el punto de fijación.

Existe un concepto mucho más sencillo que se puede utilizar a tal efecto, el *campo aleatorio de Markov*(MRF), mediante el cual se restringen las observaciones que se van a condicionar, es decir, se especifica la distribución condicional en las observaciones adyacentes. La idea principal de este concepto es que mediante la utilización de una especificación local, una observación condicionada sobre sus adyacentes, se

va a ser capaz de determinar una distribución conjunta o global:

$$p(y_i|y_{-i}) = p(y_i|y_j), \quad j \in \partial_i.$$

Siendo $-i$ todas las observaciones menos la observación i y ∂_i el conjunto de observaciones adyacentes a i.

4.2.1. Modelo autonormal

El modelo autonormal es un modelo que se puede catalogar como teórico o conceptual, ya que no se presenta como un modelo muy práctico. Especifica, directamente, la distribución condicional de y_i condicionada por el resto de observaciones como promedio ponderado de los valores en tales observaciones y heterocedasticidad, puesto que para este modelo tenemos un término de varianza distinto para cada observación, en esta peculiaridad reside el tenerlo presente como un modelo más conceptual que práctico. Teniendo en cuenta que la notación b_{ij} es equivalente a w_{ij} y que τ_i^2 es la varianza condicionada de y_i respecto de y_j, $\forall j \neq i$, podemos presentar el modelo como

$$y_i|y_j \sim \mathcal{N}\left(\sum_{j=1}^{N} b_{ij}y_j,\, \tau_i^2\right), \quad \forall\, i \neq j,\, i \in \{1, ..., N\}. \tag{4.6}$$

Mediante el lema de Brook obtenemos una densidad conjunta que toma la forma

$$p(y_1, ..., y_N) \propto \exp\left\{-\frac{1}{2}\mathrm{y}'D^{-1}(\mathrm{I} - B)\mathrm{y}\right\}, \tag{4.7}$$

donde B es la matriz que contiene los términos b_{ij} y D es una matriz diagonal con los términos de la varianza, es decir, $D_{ii} = \tau_i^2$. La expresión (4.7) sugiere una distribución normal multivariante para y, con media cero y matriz de varianzas covarianzas $D^{-1}(\mathrm{I} - B)$. En la matriz de varianzas y covarianzas se nos plantea un problema, ya que en una distribución normal multivariante debe ser siempre simétrica, por lo que para conseguirlo debemos establecer la restricción

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}, \quad \forall\, i, j. \tag{4.8}$$

4.2.2. Modelo intrínsecamente autorregresivo

Utilizando la matriz de pesos espaciales en su estructura binaria, sin estandarizar, supongamos que establecemos $b_{ij} = w_{ij}/w_{i+}$ y $\tau_i^2 = \tau^2/w_{i+}$, siendo $w_{i+} = \sum_{j=1}^{N} w_{ij}$. Entonces, la restricción (4.8) se cumple y podemos expresar la ecuación (4.6) como

$$y_i|y_j \sim \mathcal{N}\left(\sum_{j=1}^{N} w_{ij}y_j/w_{i+},\, \tau^2/w_{i+}\right), \quad \forall\, i \neq j, \tag{4.9}$$

expresando la ecuación (4.7) mediante la formulación

$$p(y_1, ..., y_N) \propto \exp\left\{-\frac{1}{2\tau^2}\mathrm{y}'(D_w - W)\mathrm{y}\right\},$$

donde D_w es una matriz diagonal con $(D_w)_{ii} = w_{i+}$.

El modelo intrínsecamente autorregresivo, no es un modelo para estudiar una distribución puesto que su distribución no tiene varianza debido a que la matriz de pesos espaciales es singular y no puede ser invertida, es un modelo apropiado para una distribución a priori bayesiana en aquellos parámetros que muestren un comportamiento de autorregresión espacial, agrupamiento espacial.

4.2.3. Modelo condicional autorregresivo (CAR)

El modelo condicional autorregresivo introduce en la formulación el parámetro espacial autorregresivo ρ, es decir, aquel que describe la fuerza de dependencia espacial entre las observaciones. De hecho, la formulación del modelo es exactamente la misma que la del modelo intrínsecamente autorregresivo (4.9), salvo la inclusión del parámetro autorregresivo

$$y_i|y_j \sim \mathcal{N}\left(\rho\sum_{j=1}^{N} w_{ij}y_j/w_{i+},\, \tau^2/w_{i+}\right), \quad \forall\, i \neq j.$$

Lo que se consigue al introducir el parámetro en la ecuación es que la matriz de pesos ya no sea singular al estar escalada por dicho parámetro.

Podemos plantear una especificación lineal del modelo, para observarlo como un

modelo de regresión, de la siguiente forma

$$E(y_i|y_j) = \mu_i + \rho \sum_{j=1}^{N} w_{ij}(y_j - \mu_j), \quad \forall\, i \neq j,$$

donde μ_i y μ_j son los valores esperados en la observación i y en sus adyacentes, respectivamente. Además, en el modelo CAR, la matriz de varianzas y covarianzas toma la forma

$$Var(y) = \sigma^2(\mathrm{I}_N - \rho W)^{-1}.$$

Capítulo 5

Estimación

Una vez seleccionado el modelo, se deben estimar aquellos parámetros que contenga, así como evaluar la precisión de los estimadores. Es importante señalar que la estimación de los parámetros por mínimos cuadrados ordinarios no es apropiada en un contexto espacial, debido a que los estimadores son sesgados e inconsistentes, independientemente del comportamiento de los errores en el modelo. Este problema viene dado fundamentalmente por el carácter multidireccional de la dependencia en el espacio muestral. Dado que el parámetro autorregresivo debe estimarse simultáneamente con otros parámetros contenidos en la ecuación, en su forma reducida, la estimación mediante máxima verosimilitud es la utilizada con mayor regularidad bajo ciertas condiciones en los estimadores, como son la convergencia, eficiencia y normalidad asintótica de los mismos.

La estimación por máxima verosimilitud busca maximizar la verosimilitud, siendo ésta la densidad conjunta de la muestra. Para observaciones independientes, la probabilidad conjunta viene dada por el producto de las probabilidades individuales o, dicho de otra forma, el logaritmo de la probabilidad conjunta vendrá dado por la suma de los logaritmos de las probabilidades individuales. En cambio, en un contexto espacial no es así debido a que las observaciones no son independientes, por lo que la probabilidad conjunta no será el producto de las probabilidades marginales, es decir, trabajaremos con un todo y no con observaciones individuales.
Algo también muy importante es que para calcular la densidad conjunta de la muestra se requiere asumir una función de densidad de probabilidad, siendo la utilizada para tal efecto por los modelos autorregresivos espaciales la presentada en la distri-

bución normal multivariante y expresada mediante

$$f(y) = \frac{1}{(2\pi)^{N/2}|\sum|^{1/2}} \exp\left\{-\frac{1}{2}(y-\mu)^t\Sigma^{-1}(y-\mu)\right\}, \; y \sim \mathcal{N}(\mu, \Sigma).$$

Donde Σ es una matriz de varianzas y covarianzas simétrica y definida positiva e y es un vector aleatorio N-dimensional, $y = [y_1, ..., y_N]$. Además, podemos expresar su función de log-verosimilitud como

$$l(\mu, \Sigma; y) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(y-\mu)^t\Sigma^{-1}(y-\mu),$$

siendo $|\Sigma|$ es el determinante de la matriz de varianzas covarianzas.

5.1. Estimación por máxima verosimilitud del modelo SAR

Recordemos que la forma reducida del modelo de autorregresión espacial es

$$y = (\mathrm{I}_N - \rho W)^{-1}X\beta + (\mathrm{I}_N - \rho W)^{-1}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{I}_N). \tag{5.1}$$

Si prestamos atención a (5.1), es fácil darse cuenta que la variable dependiente de retardo espacial está correlada con el término de error, violando así una de las suposiciones básicas en la regresión, ya que los regresores necesitan estar incorrelados con el término de error. Este es el denominado sesgo de ecuaciones simultáneas o sesgo de simultaneidad. Para poder trabajar con la endogeneidad presentada en el modelo, la estimación mediante máxima verosimilitud lidia con este problema especificando una distribución completa para las perturbaciones en el modelo. Para tal efecto, se asumirá una normalidad multivariante conjunta de los términos de error y, por simplicidad, no permitiremos heterocedasticidad en los mismos, por lo que seguirán una distribución normal multivariante con vector de medias cero y matriz de varianzas covarianzas $\sigma^2\mathrm{I}_N$, es decir, $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathrm{I}_N)$.

En general, la log-verosimilitud para una distribución normal multivariante se ex-

presa para la densidad conjunta, por ejemplo, $\varepsilon \sim \mathcal{N}(0, \Sigma)$ y tomará la expresión

$$l\left(\mu, \Sigma; \varepsilon\right) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\varepsilon'\Sigma^{-1}\varepsilon.$$

En el modelo de autorregresión espacial necesitamos la verosimilitud, la densidad conjunta, para la variable dependiente y. En cambio, la suposición de normalidad multivariante es para los términos del error que no están observados, ya que es todo aquello en el modelo de regresión que no va a poder explicar la variable independiente X. Para solventar este problema será necesario realizar una transformación de variables, por lo que en primer lugar expresamos el modelo SAR respecto al vector de términos de error

$$\varepsilon = (\mathrm{I}_N - \rho W)y - X\beta. \tag{5.2}$$

Esto equivale a escribir el término de error en función de la variable dependiente, $\varepsilon = f(y)$. Para conseguir la distribución conjunta de la variable dependiente teniendo la distribución conjunta de los términos de error, calculamos el jacobiano de la transformación, es decir, realizamos la derivada parcial entre los términos de error y la variable dependiente. El jacobiano forma parte de la función de log-verosimilitud, quedando construida con la expresión

$$\begin{aligned} l(\sigma^2, \rho, \beta; y, X) = &-\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) + \log|\mathrm{I}_N - \rho W| \\ &- \frac{(y - \rho W y - X\beta)'(y - \rho W y - X\beta)}{2\sigma^2}, \end{aligned} \tag{5.3}$$

donde $(-N/2)\log(2\pi)$ es una constante que no debemos ignorar en la práctica, puesto que algunos programas la tienen en cuenta y otros no, por lo que la log-verosimilitud maximizada dependerá de esta cuestión. El segundo término de la ecuación viene dado por la sustitución de la matriz de varianzas covarianzas por $\sigma^2\mathrm{I}_N$, ubicándose el tamaño de la muestra en el numerador de la fracción que la antecede. Además, aparece $\log|I_N - \rho W|$ que es el determinante del jacobiano entre el término de error y la variable dependiente. Para finalizar, la última expresión en la log-verosimilitud no es más que la sustitución de ε por su expresión en la ecuación (5.2) y la sustitución de la inversa de la matriz de varianzas covarianzas por su valor, es decir, $\Sigma^{-1} = \sigma^{-2}\mathrm{I}_N$.

Una vez calculada la función de log-verosimilitud para la variable dependiente del modelo, necesitamos realizar la estimación de los distintos parámetros que aparecen en la misma, como son β, ρ y σ^2. Utilizaremos para tal efecto lo expresado por Luc Anselin en su libro "Spatial Econometrics: Methods and Models (1988)", donde sugiere la realización de la estimación de los distintos parámetros centrándonos primero en β, ya que su estimador mediante máxima verosimilitud viene dado por

$$\hat{\beta} = (X'X)^{-1}X'Ay = (X'X)^{-1}X'y - \rho(X'X)^{-1}X'Wy = \hat{\beta}_0 - \rho\hat{\beta}_L,$$

donde Ay viene dada por la igualdad

$$\begin{aligned}(\mathrm{I}_N - \rho W)y &= X\beta + \varepsilon,\\ Ay &= X\beta + \varepsilon.\end{aligned}$$

Si nos fijamos en $\hat{\beta}_0$, $\hat{\beta}_0 = (X'X)^{-1}X'y$, y $\hat{\beta}_L$, $\hat{\beta}_L = (X'X)^{-1}X'Wy$, podemos vislumbrar la estructura del vector de coeficientes en una regresión por mínimos cuadrados ordinarios de X en y y de X en Wy respectivamente. Por lo que podremos calcular el estimador de máxima verosimilitud de β si el valor de ρ es conocido. Además, mediante los vectores de coeficientes $\hat{\beta}_0$ y $\hat{\beta}_L$ podemos establecer los conjuntos de residuos $\hat{\varepsilon}_0$ y $\hat{\varepsilon}_L$ que dependerán de X, además de Wy,

$$\begin{aligned}\hat{\varepsilon}_0 &= y - X\hat{\beta}_0,\\ \hat{\varepsilon}_L &= Wy - X\hat{\beta}_L.\end{aligned}$$

Teniendo en cuenta los residuos obtenidos, se puede obtener el estimador de la varianza de los términos de error,

$$S^2 = \frac{1}{N}(\hat{\varepsilon}_0 - \rho\hat{\varepsilon}_L)'(\hat{\varepsilon}_0 - \rho\hat{\varepsilon}_L).$$

De nuevo podremos estimar σ^2 si ρ es conocida. Además, con todos los cálculos realizados hasta el momento, podemos sustituir los estimadores tanto de β como de σ^2 en la función de log-verosimilitud (5.3), obteniendo la función de log-verosimilitud

concentrada

$$l^*(\sigma^2, \rho, \beta; y, X) = C - \frac{N}{2} \log \left[\frac{1}{N} (\hat{\varepsilon}_0 - \rho \hat{\varepsilon}_L)'(\hat{\varepsilon}_0 - \rho \hat{\varepsilon}_L) \right] + \log |A|,$$

donde C es el término constante de la función de log-verosimilitud (5.3). Con esta función de log-verosimilitud concentrada, y con un único parámetro, podemos maximizar respecto de ρ y así obtener el estimador de máxima verosimilitud para este parámetro y trabajar de forma inversa a la expuesta en estas líneas para conseguir el resto.

Para finalizar, se expone el último elemento necesario para implementar el algoritmo de computación, es decir, la matriz de información de Fisher

$$I(\theta) = -E \left[\partial^2 l(\sigma^2, \rho, \beta; y, X) / \partial \theta \partial \theta' \right].$$

5.2. Estimación por máxima verosimilitud del modelo SEM

En primer lugar, recordemos cómo se estructura la ecuación reducida del modelo SEM, siendo ésta

$$y = X\beta + (\mathrm{I} - \lambda W)^{-1} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N),$$

y dejando la expresión respecto al vector de términos de error se obtiene

$$\varepsilon = (\mathrm{I}_N - \lambda W)(y - X\beta). \tag{5.4}$$

La estimación por máxima verosimilitud del modelo de error autorregresivo espacial utiliza el mismo principio que la estimación por máxima verosimilitud del modelo SAR, debido a que la variable dependiente de retardo espacial está correlada con el término de error en ambos modelos. Por lo que tras realizar los cálculos detallados en el apartado anterior, obtenemos la siguiente función de log-verosimilitud conjunta

para el modelo SEM

$$l(\sigma^2, \lambda, \beta; y, X) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) + \log|\mathrm{I}_N - \lambda W| - \frac{((\mathrm{I}_N - \lambda W)(y - X\beta))'(\mathrm{I}_N - \lambda W)(y - X\beta)}{2\sigma^2}. \tag{5.5}$$

Se debe estimar el parámetro β utilizando mínimos cuadrados generalizados (MCG) debido a que, si calculamos la matriz de varianzas covarianzas a partir de la ecuación (5.4), se obtiene una matriz de varianzas covarianzas no esférica

$$\Sigma = \sigma^2[(\mathrm{I}_N - \lambda W)'(\mathrm{I}_N - \lambda W)]^{-1}, \tag{5.6}$$

ya que los términos no diagonales de la misma tendrán un valor que diferirá de cero. Por lo que, utilizando la estimación por mínimos cuadrados generalizados podemos estimar β mediante la ecuación

$$\hat{\beta}_{MCG} = [X'\Sigma^{-1}X]^{-1}X'\Sigma^{-1}y. \tag{5.7}$$

Si se auna la expresión (5.7) con el valor de la matriz de varianzas covarianzas (5.6) se obtiene la expresión conocida como estimación por mínimos cuadrados ponderados espacialmente (SWLS), aunque también se puede encontrar esta expresión denotada como Cochrane-Orcutt:

$$\hat{\beta}_{SWLS} = [X_L'X_L]^{-1}X_L'y_L,$$

siendo $X_L = (\mathrm{I}_N - \lambda W)X$ e $y_L = (\mathrm{I}_N - \lambda W)y$.
La estimación de σ^2 sigue un procedimiento similar al utilizado para el modelos SAR, siendo formulada como

$$S^2 = \frac{1}{N}\hat{\varepsilon}'(\mathrm{I}_N - \lambda W)'(\mathrm{I}_N - \lambda W)\hat{\varepsilon},$$

con $\hat{\varepsilon} = y - X\hat{\beta}_{MCG}$.

Se puede apreciar como los estimadores $\hat{\beta}_{MCG}$ y S^2 se encuentran, ambos, en función del valor de λ. Utilizando el enfoque del modelo SAR, podemos obtener la

expresión

$$\begin{aligned} l^*(\sigma^2, \lambda, \beta; y, X) =& C - \frac{N}{2} \log \left[\frac{1}{N} \hat{\varepsilon}'(\mathrm{I}_N - \lambda W)'(\mathrm{I}_N - \lambda W)\hat{\varepsilon} \right] \\ & + \log |\mathrm{I}_N - \lambda W|. \end{aligned} \tag{5.8}$$

para la función de log-verosimilitud concentrada. Además, es importante señalar que la función (5.8) plantea un problema, ya que el término de residuos $\hat{\varepsilon}$ se encuentra indirectamente en función de λ, al haber sido obtenido el estimador $\hat{\beta}_{MCG}$ mediante el valor de dicho parámetro, algo que no ocurría con el modelo SAR. Para solventar tal situación, Anselin establece el siguiente proceso iterativo:

1. Llevamos a cabo una regresión mediante mínimos cuadrados ordinarios de X en y, obteniendo el estimador $\hat{\beta}_{MCO}$, con el que podemos obtener los residuos mediante la ecuación $\hat{\varepsilon}_{MCO} = y - X\hat{\beta}_{MCO}$.

2. Dado $\hat{\varepsilon}$, lo utilizamos en la función de log-verosimilitud concentrada (5.8) y maximizamos para encontrar λ.

3. Utilizamos λ para calcular el estimador de β mediante mínimos cuadrados generalizados, $\hat{\beta}_{MCG}$, y poder calcular el nuevo vector de residuos ε mediante la igualdad $\hat{\varepsilon} = y - X\hat{\beta}_{MCG}$.

4. Si se cumple el criterio de convergencia seguimos, si no es así, volvemos al paso 2 y reestimamos λ.

5. Dados $\hat{\varepsilon}$ y λ, calculamos el estimador de la varianza $S^2 = \frac{1}{N}\hat{\varepsilon}'(\mathrm{I}_N - \lambda W)'(\mathrm{I}_N - \lambda W)\hat{\varepsilon}$.

5.3. Estimación por máxima verosimilitud del modelo SDM

Si recordamos lo expuesto en el capítulo anterior, el modelo espacial de Durbin se puede construir a partir del modelo SAR como del modelo SEM, siendo esta última opción la utilizada en el presente trabajo. La forma reducida del SDM queda

expresada como sigue:

$$y = (\mathrm{I}_N - \rho W)^{-1}(X\beta + WX\theta) + (\mathrm{I}_N - \rho W)^{-1}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_N).$$

James P. LeSage establece una forma rápida y sencilla para conseguir la estimación por máxima verosimilitud del presente modelo, de acuerdo a sus apuntes para el curso "Course on Spatial Econometrics, University of Toledo (2004)". En dicho curso, LeSage, construye la estimación utilizando exactamente el mismo procedimiento que el expuesto en el presente capítulo para la estimación por máxima verosimilitud del modelo SAR, con la salvedad de la sustitución de la igualdad $X = [X \quad WX]$ en las ecuaciones

$$\begin{aligned}
&\hat{\beta} = \hat{\beta}_0 - \rho\hat{\beta}_L, \quad \text{siendo } \hat{\beta}_0 = (X'X)^{-1}X'y,\ \hat{\beta}_L = (X'X)^{-1}X'Wy,\\
&\hat{\varepsilon}_0 = y - X\hat{\beta}_0,\\
&\hat{\varepsilon}_L = Wy - X\hat{\beta}_L.
\end{aligned}$$

Capítulo 6

Software

En la actualidad se puede encontrar diverso software con el que realizar un análisis espacial, como Matlab, GeoDa (desarrollado por Luc Anselin) y R. La elección del software es una decisión totalmente personal que debe tomar el investigador o analista según sus gustos, preferencias o conocimientos de programación. Aun así, se recomienda el uso del programa estadístico R de entre todos ellos, debido a la gran cantidad de paquetes desarrollados y listos para poder ser utilizados en este tipo de análisis. Otra de sus grandes virtudes es que se trata de un software libre, por lo que es totalmente gratuito y existe una gran comunidad de usuarios que lo mejoran y, que además, corrigen errores que puedan aparecer. Además, crean nuevos paquetes y exponen y solucionan dudas en multitud de foros, blogs y páginas de internet dedicadas exclusivamente a este software. Por último, pero no por ello menos importante, se debe destacar la gran potencia que presenta R y su gran versatilidad a la hora de realizar descriptivos, construir gráficos a gusto del usuario y trabajar con datos en un contexto general.

La localización de una observación determinada, puede expresarse mediante tres tipos de objetos espaciales: puntos, líneas y polígonos. Un punto tiene asociado un par de coordenadas y un ejemplo podría ser la localización de una central eléctrica. Las líneas son objetos, abiertos, que cubren cierta distancia y comunican puntos o nodos y se corresponderían con las líneas de transmisión de energía entre centrales eléctricas. Por último, los polígonos, son aquellos objetos formados por un conjunto cerrado de líneas que cubren un área determinada y siguiendo el ejemplo, serían

aquellas zonas que contienen centrales eléctricas.

R tiene asignado para cada tipo de objeto espacial un tipo de vector de datos distinto, como son: SpatialPoints (puntos espaciales), SpatialLines (líneas espaciales) y SpatialPolygons (polígonos espaciales). Puede resulta interesante conocer que estos vectores pertenecen a una clase distinta a la que se utilizan en la mayoría de análisis estadísticos, la denominada clase S4.

La inmensa mayoría de objetos con los que trabaja R pertenecen a la clase S3, compuesta por vectores atómicos, arrays y listas. Estos objetos, que no están formalmente definidos, siempre se podrán dividir mediante las operaciones normales que contiene R para tal efecto. Además, podemos conocer su contenido mediante la función *str()*. En cambio, los objetos pertenecientes a la clase S4 tienen la ventaja de poseer una definición formal, especificando nombre y tipo de sus componentes, denominados ”slots“. Además, los objetos de la clase S4 incluyen dos operadores de subconjuntos adicionales a los de la clase S3, los típicos que se pueden utilizar en R, como son @ que tiene una cierta equivalencia al operador *$* de la clase S3 y *slot()* equivalente a los dobles corchetes utilizados para trabajar con listas en la clase S3.

Capítulo 7

Estudio de la tasa de paro en las comarcas de la Comunidad Valenciana

7.1. Resumen

En el presente capítulo se pretende realizar un supuesto práctico con datos reales descargados del portal de Información de la Dirección de Análisis y Políticas Públicas de la Presidencia de la Generalitat valenciana (ARGOS). El estudio pretende analizar la posible dependencia espacial de la tasa de paro en las comarcas de la Comunidad Valenciana. El estudio cuenta con 56 variables relativas a las 34 comarcas de las que consta la Comunidad Valenciana. El análisis, que comenzó con el filtrado de datos mediante regresiones lineales múltiples, permitió establecer un modelo cuyas variables explicativas eran el índice de dependencia de la población en la comarca, en tanto por cien, el porcentaje de paro representativo del sector industrial y el porcentaje de empresas de este sector en la comarca. Además se demuestra la existencia de autocorrelación espacial global y local, ubicando aquellas comarcas que poseen una autocorrelación local positiva. Además, tras diversos test, el modelo idóneo para realizar el estudio es el modelo de error espacial autorregresivo, demostrando que es el idóneo debido a la omisión de variables en nuestro modelo.

7.2. Obtención y descripción de los datos

La base de datos con la que se ha contado para el estudio, se obtuvo del portal de Información de la Dirección de Análisis y Políticas Públicas de la Presidencia de la Generalitat valenciana (ARGOS)[1]. Del mismo, se recopiló información de las 34 comarcas de las que consta la Comunidad Valenciana, pudiendo ser dividida en cuatro tipos de datos:

- Datos de carácter general: Incluye la información más básica de la comarca, como su denominación, código comarcal y provincia a la que pertenece, por poner algunos ejemplos. Se recopilaron los datos 6 variables, además, se añadió la capital comarcal por si se necesitaba para la elaboración de gráficos. Por lo tanto, aporta 7 variables al estudio.

- Datos estadísticos de carácter demográfico: Incluye información relativa a la población y a sus características. Aporta 15 variables al estudio y algunos ejemplos de éstas son la variación del padrón en tanto por cien, los españoles residentes en el extranjero o el porcentaje de residentes nacidos en la Comunidad Valenciana.

- Datos estadísticos de carácter laboral: Aporta 24 variables al estudio, siendo los datos más numerosos. Estas variables hacen referencia al paro registrado en la comarca, al paro registrado por sectores de actividad, afiliados a la Seguridad Social y afiliados a la Seguridad Social residentes en la comarca.

- Datos estadísticos de carácter socio-económico: Incluyen 9 variables al estudio, referentes a los indicadores de la actividad económica y al DIRCE, Directorio Central de Empresas. Además, los datos se completaron con una nueva variable para así obtener el 100 % de los sectores empresariales en la comarca, por lo tanto, estos datos cuentan con 10 variables. Algunos ejemplos son el valor catastral medio en euros, la deuda viva del municipio en euros por habitante y el tanto por cien de empresas en el sector industria.

Una vez expuestos los distintos tipos de datos, resulta oportuno realizar algunas puntualizaciones al respecto. En primer lugar, se sustituyó el código comarcal de la

[1] `http://www.argos.gva.es/bdmun/pls/argos_mun/DMEDB_UTIL.INDEXC`

Comunidad Valenciana por el código comarcal estatal, se hizo para que no hubiese ningún tipo de problema a la hora de manipular los datos que contiene el dataframe de polígonos espaciales. Este dataframe es de suma importancia, ya que contiene la información relativa a la localización geográfica de las distintas comarcas y su representación. Además, el portal ARGOS actualiza sus datos de forma mensual, por lo tanto el estudio cuenta con información relativa al mes de febrero para una numerosa cantidad de variables, así como de información relativa al mes de diciembre en otras. Las distintas variables y su descripción, se pueden encontrar en el Anexo I: Bases de datos utilizadas. Además, en dicho anexo también se encuentra un link para descargar un archivo comprimido con todos los datos que se han requerido para realizar el presente estudio. Para finalizar, resulta oportuno también señalar que la comarca de Castellón El Alto Mijares, no contaba con los datos relativos al porcentaje de empresas en el sector industria, porcentaje de empresas en el sector de la construcción, porcentaje de empresas en el sector comercio, transporte y hostelería y el porcentaje de empresas en el sector servicios. Para la obtención de dichas variables se utilizó la base de datos SABI, accesible desde el acceso identificado de la Universidad Miguel Hernández de Elche. SABI es el Sistema de Análisis de Balances Ibéricos que incluye información general y análisis financieros de 2.000.000 empresas españolas; realizando una búsqueda con unos parámetros determinados se pudieron cumplimentar tales variables[2].

Como se puede comprobar, el estudio cuenta con 56 variables relativas a 34 observaciones de un fuerte carácter espacial. Con tales variables se intentará explicar qué tipo de influencia tienen respecto a la tasa de paro, en tanto por cien, en la Comunidad Valenciana. Además, se estudiará la posible existencia de efectos espaciales, es decir, autocorrelación espacial o heterogeneidad espacial. También se intentará comprobar si es apreciable algún tipo de autocorrelación espacial local.

[2]Se puso mucho énfasis en conseguir los datos faltantes para el estudio, ya que al tratarse de un análisis espacial de datos es imprescindible poder contar con todos los datos en todas las localizaciones.

7.3. Selección de variables para el análisis

Antes de una selección "estadística", se estudiaron las variables que componían las distintas categorías. Se observaron algunas variables redundantes en cuanto a la tasa de paro, nuestra variable dependiente, por lo que se eliminaron estas 7 variables del estudio, siendo un claro ejemplo la variable relativa al paro registrado. También se eliminaron aquellas que aportarían una clara multicolinealidad, puesto que representaban la suma de otras e incluso se eliminó una variable relativa al porcentaje de personas mayores de 64 años. Ésta última se eliminó porque la edad de jubilación de los españoles, en la actualidad, es a los 65 años[3]. Por lo tanto, con este filtrado el estudio quedó reducido a 48 variables. Dicho esto, para la selección de las variables no se optó por un estudio conjunto de las 48 variables, se realizaron tres análisis estadísticos, para las variables demográficas, para las variables de carácter laborar y para las socio-económicas respectivamente, de esta forma también se evitaría el realizar un estudio con más variables que observaciones. Utilizando un modelo de regresión lineal múltiple en cada estudio se obtendrían las variables significativas para estos grupos de datos y, entonces sí, se homogeneizarían.
La obtención de los distintos modelos se llevó a cabo con el siguiente proceso iterativo:

1. Estudio de la correlación entre las distintas variables para buscar una posible colinealidad entre las variables independientes.

2. Creación de un modelo cuya variable dependiente es la tasa de paro, en tanto por cien, y realización de una regresión lineal múltiple.

3. Selección del modelo mediante el criterio de información de Akaike (AIC), utilizando una regresión por pasos a ambos lados.

4. Al modelo resultante se le aplica un análisis de la varianza (ANOVA) para comprobar la significatividad de las variables incluidas en el modelo. Si las variables son significativas obtenemos el modelo deseado, si no, realizamos el paso 2 de nuevo.

[3]Todas estas variables aparecerán en el Anexo I: Bases de datos utilizadas y se distinguirán del resto.

Tras los análisis se obtuvieron los modelos

$$tasa\ de\ paro \sim indice\ de\ dependencia,$$

$$tasa\ de\ paro \sim paro\ registrado\ en\ industria + total\ de\ afiliados\ Seg.Social +$$
$$tasa\ de\ afiliacin\ Seg.Social + afiliados\ Seg.Social\ en\ RgimenGeneral +$$
$$tasa\ de\ afiliacin\ Seg.Social\ de\ residentes\ en\ la\ comarca,$$

$$tasa\ de\ paro \sim empresas\ sector\ industrial + empresas\ sector\ servicios.$$

Una vez obtenidas todas las variables significativas para cada uno de los conjuntos de datos, elaboramos el modelo final. En éste no se van a incluir las variables tasa de afiliación a la Seguridad Social y tasa de afiliación de los residentes en la comarca. Su eliminación viene dada porque van a ser variables excesivamente buenas, muy significativas para el modelo, lo que acarreará que quitarán capacidad explicativa del resto de variables. Antes de comenzar con la modelización del mismo, se debe estudiar la correlación existente entre las variables obtenidas. Como se puede

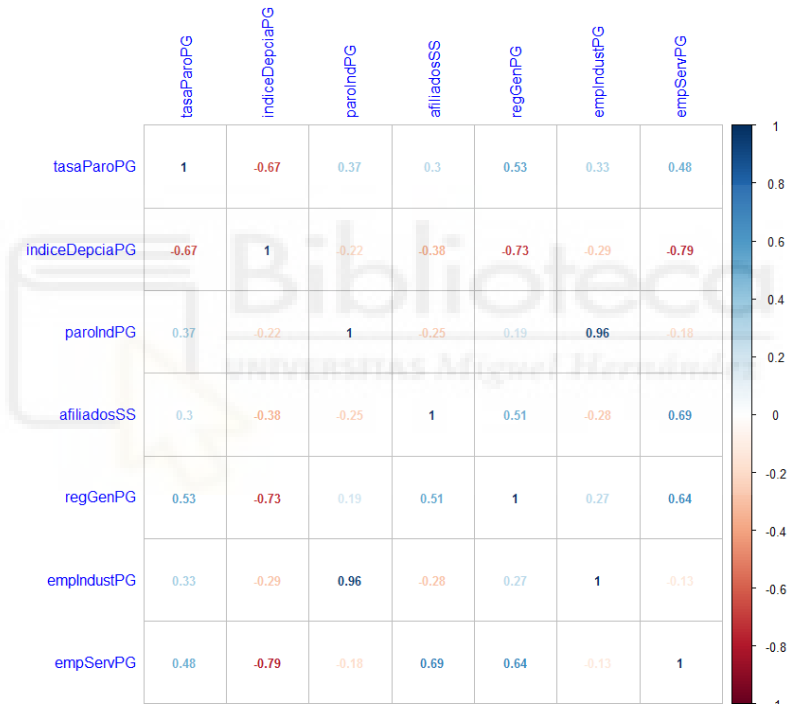


Figura 7.1: Matriz de correlaciones entre variables.
Fuente: Elaboración propia.

comprobar no existe ninguna correlación alarmante, sí es cierto que existen ciertas variables muy correladas, como por ejemplo empIndustPG y paroIndPG. En este caso se debe a que este tipo de empresas tienen los procesos muy mecanizados, por lo que cada vez se necesita de menos personal.

Planteado el modelo y tras utilizar el criterio AIC, con la respectiva comprobación

de significatividad en las variables, se obtuvo el modelo global, siendo

$$\begin{aligned} \textit{tasa de paro} \sim & 25{,}518 - 0{,}2554(\textit{indice de dependencia}) + \\ & 0{,}3552(\textit{paro registrado en industria}) - \\ & 0{,}5669(\textit{empresas sector industrial}). \end{aligned}$$

Además, el modelo presenta un p-valor de $1{,}222e-05$, muy inferior a 0,05 por lo que es significativamente mejor que el modelo planteado inicialmente. También se puede señalar que el valor del R^2 indica que la recta de regresión explica el 52,36 % de la variabilidad del modelo, un valor no excesivamente bueno pero aceptable.

Tras la obtención del modelo se debe realizar una diagnosis del mismo para comprobar cómo se comportan sus residuos. A simple vista se cumple tanto la nor-

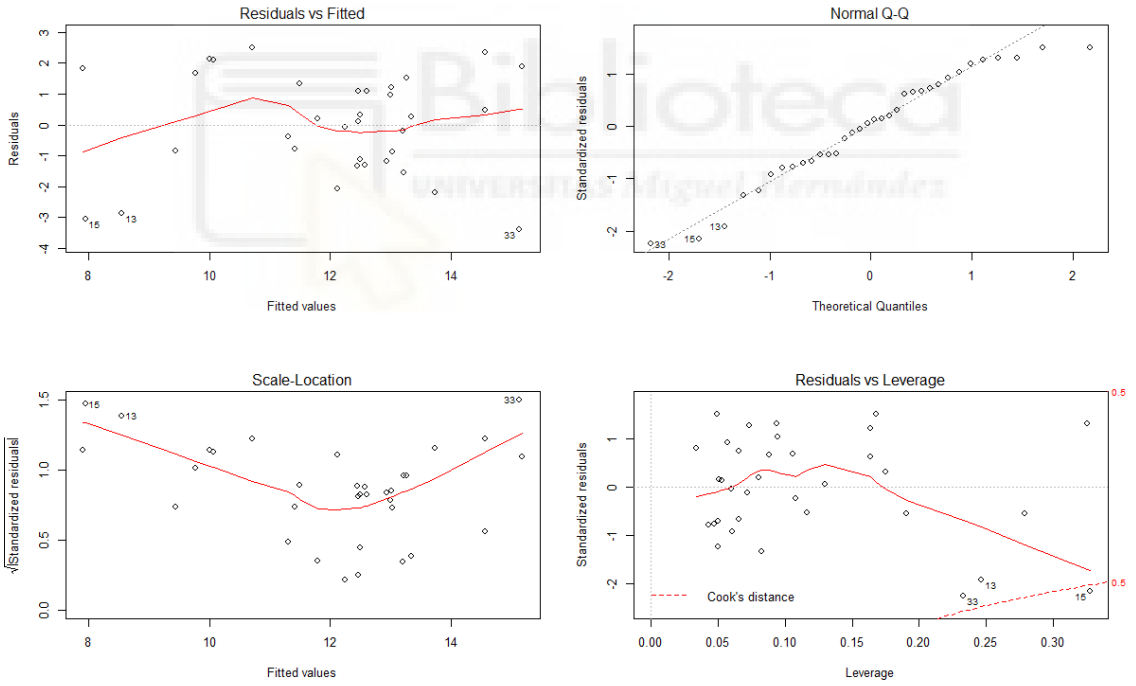


Figura 7.2: Gráficos de diagnóstico de los residuos.
Fuente: Elaboración propia.

malidad como la linealidad en los residuos, en cambio sí se aprecia un ligero patrón de los residuos en el gráfico de valores ajustados y residuos. También se observan algunas observacioness que pueden ser influyentes, sin duda es algo a estudiar.
De forma empírica comprobaremos el resultado del diagnóstico del modelo mediante la siguiente tabla:

Test	P-Valor	Conclusión
Reset de Ramsey	0.2948 >0.05	Existe linealidad
Shapiro	0.271 >0.05	Existe normalidad
Breusch-Pagan	0.1343 >0.05	Varianza constante
Durbin-Watson	0.9739 >0.05	Residuos no correlados

Tabla 7.1: Test para la diagnosis del modelo.
Fuente: Elaboración propia.

En cuanto a las observaciones que parecían ser atípicas, no debemos preocuparnos, ya que como comprobamos en el siguiente gráfico, ninguno de ellas presenta una distancia de Cook superior a 1 y por lo tanto, no hace falta eliminar ninguna de ellas:

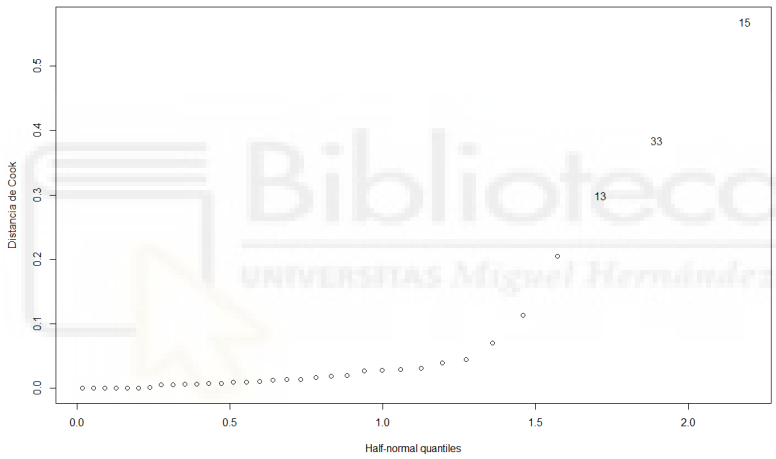


Figura 7.3: Gráfico de distancias de Cook.
Fuente: Elaboración propia.

7.4. Análisis econométrico espacial

7.4.1. Descriptivo de los datos

Siempre es interesante comenzar a describir los datos con una tabla resumen de las variables que se estudian. Como podemos comprobar, la tasa de paro se mueve en las comarcas de la Comunidad Valencianaentre un 4.9 % y un 17.08 %. Además, posee un valor medio de 12.07 % con una varianza de un 2.49 %. De todas las variables, la que posee una varianza mayor es la variable relativa al paro en el sector industrial, donde la comarca con menos paro en este sector tiene un valor de un 4.35 % y el

	Tasa de paro	**Índice de dependencia**	**Paro sector industrial**	**Empresas sector industrial**
Mínimo	4.90 %	48.62 %	4.35 %	2.4 %
Media	12.07 %	56.25 %	15.66 %	8.18 %
Desviación típica	2.49 %	6.8 %	7.64 %	4.02 %
Máximo	17.08 %	75.92 %	32.13 %	16.34 %

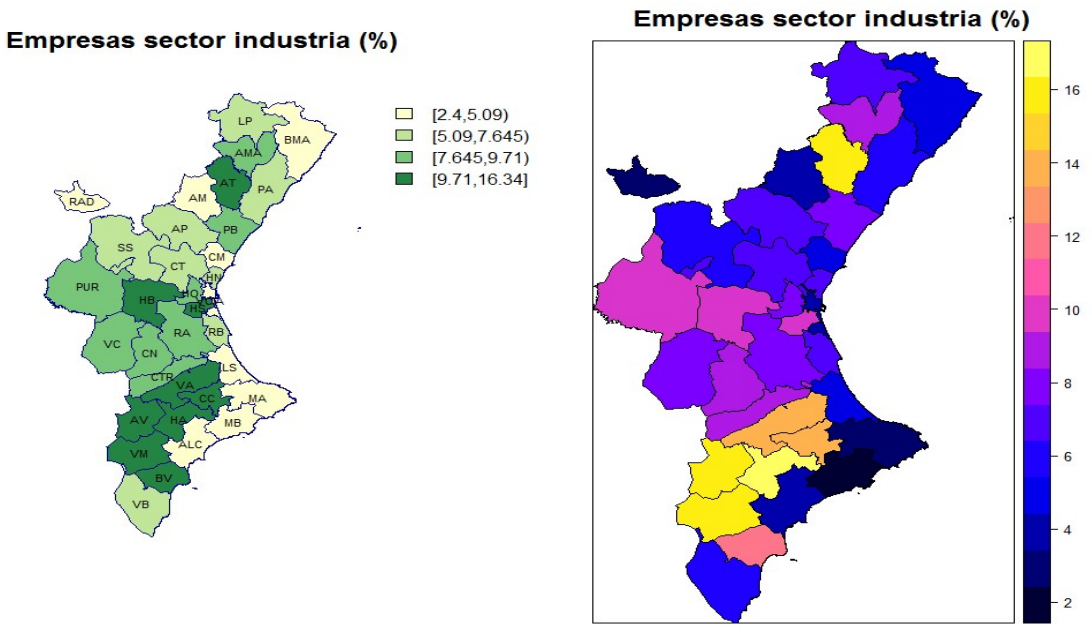
Tabla 7.2: Resumen de las variables pertenecientes al modelo.
Fuente: Elaboración propia.

porcentaje máximo se encuentra en un 32.12 %.

A continuación representaremos las distintas variables en el mapa comarcal de la Comunidad Valenciana. Para tal efecto, se han establecido los siguientes acrónimos para comarcas:

Comarca	**Acrónimo**	**Comarca**	**Acrónimo**	**Comarca**	**Acrónimo**	**Comarca**	**Acrónimo**
CAMPO DE ALICANTE	ALC	ALTO PALANCIA	AP	HOYA DE BUNYOL	HB	VALLE DE COFRENTES-AYORA	VC
HOYA DE ALCOY	HA	ALTO MAESTRAZGO	AMA	HUERTA NORTE	HN		
ALTO VINALOPO	AV	BAJO MAESTRAZGO	BMA	HUERTA OESTE	HO		
VEGA BAJA	VB	LOS PUERTOS	LP	HUERTA SUR	HS		
CONDADO DE COCENTAINA	CC	PLANA ALTA	PA	SAFOR	LS		
MARINA ALTA	MA	PLANA BAJA	PB	PLANA DE UTIEL REQUENA	PUR		
MARINA BAJA	MB	CAMPO DE MURVIEDRO	CM	RIBERA ALTA	RA		
VINALOPO MEDIO	VM	CAMPO DE TURIA	CT	RIBERA BAJA	RB		
BAJO VINALOPO	BV	CANAL DE NAVARRES	CN	RINCON DE ADEMUZ	RAD		
ALCALATEN	AT	VALENCIA	VCIA	SERRANOS	SS		
ALTO MIJARES	AM	COSTERA	CTR	VALLE DE ALBAIDA	VA		

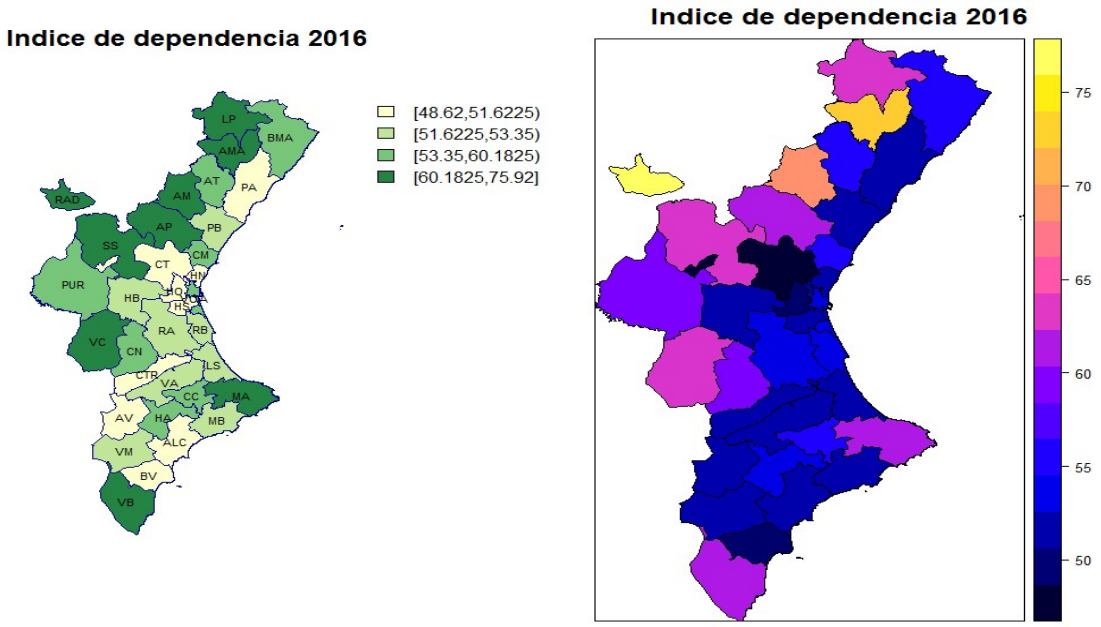
Tabla 7.3: Acrónimos comarcales de la Comunidad Valenciana.
Fuente: Elaboración propia.



(a) Mapa de coropletas (b) Mapa de calor

Figura 7.4: Representación del sector industrial (%).
Fuente: Elaboración propia.

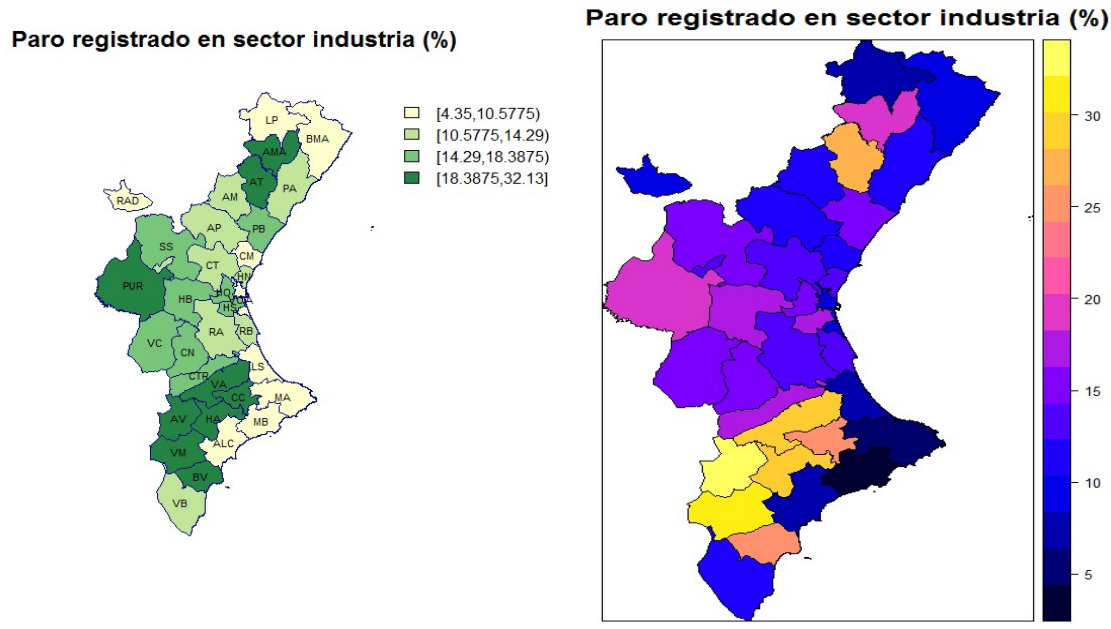
Las comarcas con mayor porcentaje de empresas dedicadas al sector industrial se ubican en la zona suroeste de la Comunidad. Siendo el Alto y Bajo Vinalopó las que destacan entre ellas. Además al norte de la Comunidad también encontramos la comarca de Alcalatén con un buen porcentaje de este tipo de empresas. En cuanto al porcentaje más bajo, se ubica, de forma más global, en el litoral mediterráneo desde el Campo de Alicante hasta La Safor.



(a) Mapa de coropletas (b) Mapa de calor

Figura 7.5: Representación del índice de dependencia comarcal (%).
Fuente: Elaboración propia.

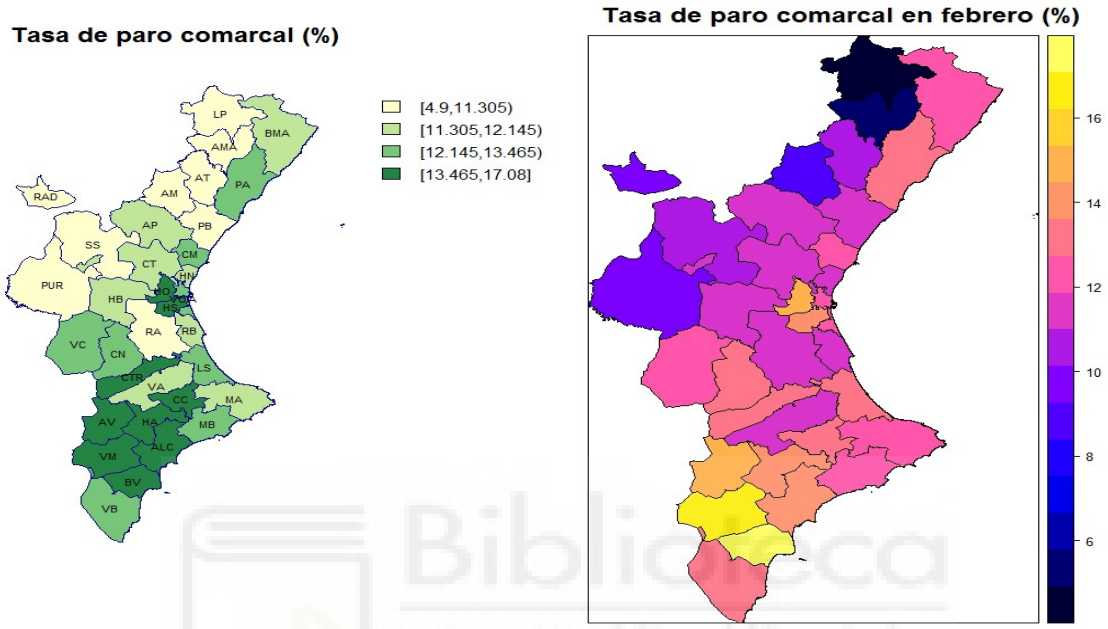
El índice de dependencia se encuentra más disperso, pero los mayores porcentajes se ubican al noroeste.



(a) Mapa de coropletas (b) Mapa de calor

Figura 7.6: Representación del porcentaje de paro en el sector industrial.
Fuente: Elaboración propia.

Respecto al paro en el sector industrial, encontramos prácticamente los mismos mapas que para la variable referente a la cantidad de empresas de este sector, como es totalmente normal.



(a) Mapa de coropletas (b) Mapa de calor

Figura 7.7: Representación de la tasa de paro comarcal.
Fuente: Elaboración propia.

Para finalizar, la tasa de paro se encuentra más concentrada al sur y suroeste de la Comunidad y aquellas con una tasa de paro menor se ubican al norte y al este de la misma.

Para finalizar el análisis descriptivo se realizará el test de Moran para estudiar la posible autocorrelación espacial que presenta la tasa de paro en la Comunidad Valenciana. Pero antes de exponer los resultados se debe hacer hincapié en la estructura de adyacencias espaciales de nuestras observaciones. Para la realización del estudio se necesitaba de una adyacencia total entre las distintas observaciones, por lo que se optó por crear una matriz de pesos espaciales basada en la contigüidad de la reina, pueden observarse las distintas adyacencias entre observaciones en el presente mapa y la isla que contiene:

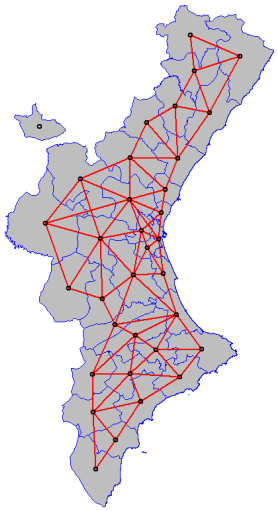


Figura 7.8: Gráficos de adyacencias entre comarcas.
Fuente: Elaboración propia.

También se optó por una matriz de pesos espaciales estandarizada, para poder aprovechar todas las ventajas que representa, entre ellas, poder analizar el valor de la variable de dependencia espacial como un simple índice de correlación. Dicho lo cual, pasamos a analizar los resultados obtenidos mediante el test de Moran: El valor

TEST DE MORAN				
Z	P-VALOR	I de Moran	E[I]	Var[I]
4.2572	1.035e-05	0.4460953	-0.03125	0.0125722

Tabla 7.4: Resultados obtenidos del test de Moran

del estadístico Z es muy superior al 1.96 requerido para asegurar que existe autocorrelación espacial en la variable dependiente, la tasa de paro en las comarcas de la Comunidad Valenciana, justificando la utilización de un modelo espacial. Además, al tener el estadístico Z un valor positivo, se trata de una autocorrelación espacial positiva.

Veamos que resultados obtenemos si aplicamos el test de Moran a nuestro modelo lineal: Como se puede comprobar, todos los valores disminuyen, pero sigue recha-

TEST DE MORAN				
Z	P-VALOR	I de Moran	E[I]	Var[I]
2.2508	0.0122	0.21039459	-0.04007831	0.01238409

Tabla 7.5: Resultados obtenidos del test de Moran

zándose la hipótesis nula de distribución aleatoria de la variable dependiente.

Una vez demostrada la existencia de una autocorrelación espacial positiva, comprobemos si existe una autocorrelación espacial local entre las comarcas de la Comunidad Valenciana. Para tal efecto, graficaremos el diagrama de dispersión de Moran: Es palpable también la existencia de una autocorrelación espacial local, y además

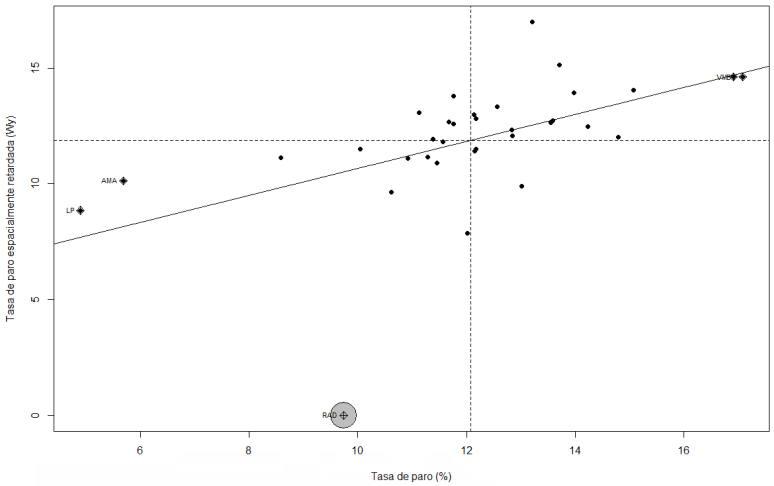


Figura 7.9: Gráfico de dispersión de Moran.
Fuente: Elaboración propia.

positiva, de la tasa de paro en las comarcas del Alto Vinalopó, Vinalopó Medio y Bajo Vinalopó claramente representadas en el cuadrante (AA). También existe este tipo de autocorrelación espacial para las comarcas de Los Puertos y El Alto Maestrazgo. También se puede observar la isla en nuestros datos, es decir, la comarca del Rincón de Ademúz.

7.4.2. Identificación del proceso autorregresivo y modelización

En la práctica, tras el estudio del estadístico de Moran se suele realizar el *test de los multiplicadores de Lagrange* para identificar el tipo de proceso autorregresivo. El test de los multiplicadores de Lagrange no es más que el *estadístico de puntuaciones de Rao.*

Este test se basa en observar la inclinación de la pendiente de la función de verosimilitud, comprobando si el gradiente es significativamente distinto de cero. Recordemos que en el método de máxima verosimilitud, buscamos encontrar un estimador que maximice la probabilidad conjunta de la muestra y encontraremos ese valor cuando

	TEST DE MULTIPLICADORES DE LAGRANGE	
	Estadístico	P-Valor
SEM	3.23279	0.07218
SAR	0.87577	0.34936

Tabla 7.6: Test de los multiplicadores de Lagrange.
Fuente: Elaboración propia.

la pendiente tenga como valor cero. En un contexto espacial, la hipótesis nula que queremos comprobar es si el parámetro autorregresivo espacial es significativo o no, es decir, cuando el valor obtenido sea inferior a 0.05.

El test de los multiplicadores de Lagrange se ha pogramado para contrastar la hipótesis nula con los modelos SAR y SEM, obteniendo: En efecto, existe significatividad respecto al modelo de error autorregresivo espacial y, por lo tanto, existen factores o variables omitidas en el modelo planteado que trasladan hacia los términos de error la configuración de autocorrelación presente en la variable dependiente, la tasa de paro.

A continuación se muestran los resultados obtenidos tanto para el modelo SEM, el ideal para nuestro análisis, el modelo SAR y el modelo lineal[4]: El modelo obtenido

	Regresión Lineal	**Modelo SEM**	**Modelo SAR**
Intercepto	25.51799*	23.904550*	22.998283*
Índice de dependencia	-0.25538*	-0.220974*	-0.231103*
Paro en el sector industrial	0.35520*	0.235912*	0.341822*
Empresas en el sector industrial	-0.56694*	-0.358849*	-0.550563*
rho			0.75692
lambda		0.47015*	

Tabla 7.7: Comparación resultados obtenidos para el modelo lineal, el modelo SAR y el modelo SEM.
Fuente: Elaboración propia.

es

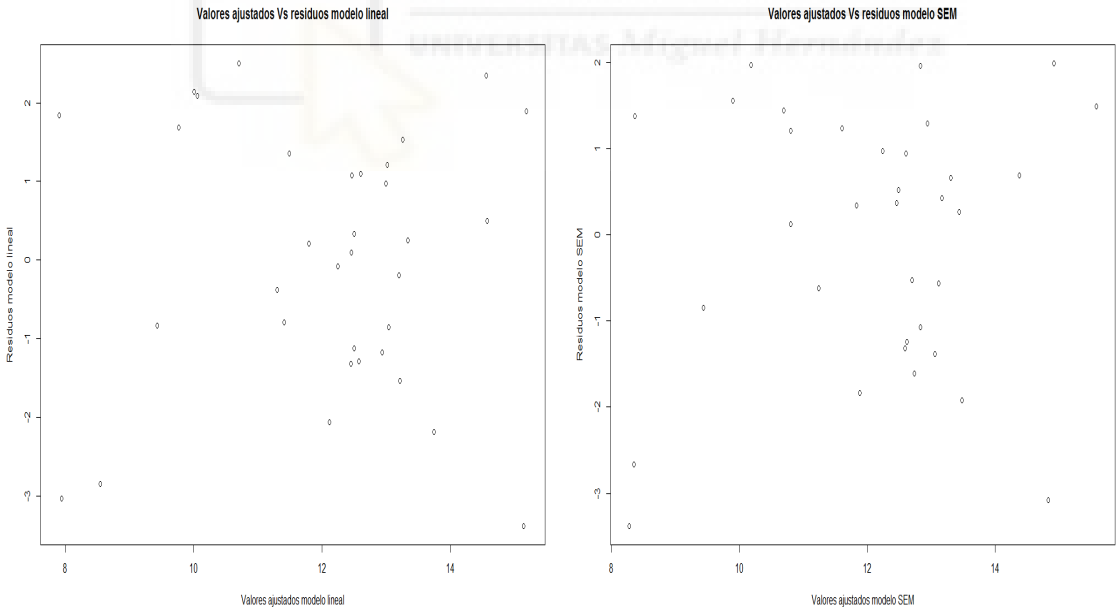
$$\begin{aligned} \textit{tasa de paro} \sim & 23{,}9 - 0{,}221(\textit{indice de dependencia}) + \\ & 0{,}236(\textit{paro registrado en industria}) - \\ & 0{,}359(\textit{empresas sector industrial}). \end{aligned}$$

[4]El asterisco denota significatividad.

Por lo que, si se incrementa en una unidad el índice de dependencia en la población de la comarca y por cada empresa de sector industrial que se constituya, la tasa de paro decrecerá en un 0,22 % y en un 0,36 % respectivamente. El caso opuesto lo encontramos con el paro en el sector industrial, puesto que por cada unidad que crezca éste, la tasa de paro aumentará en un 0,23 %.

Además, constatamos que el p-valor asociado a ρ no es significativo, en cambio el asociado a λ sí lo es. Además, como el modelo de error espacial autorregresivo es el indicado para aquellas situaciones que presentan heterogeneidad espacial o cuando existen variables omitidas, se ha realizado el test de Breusch-Pagan para modelos autorregresivos y hemos obtenido un p-valor de 0,271 que es mayor que 0,05, indicando que la varianza en los residuos es constante. Por lo que nuestros datos presentan un problema de variables omitidas en el estudio.

Para finalizar, comparemos los gráficos de residuos frente valores ajustados del modelo SEM y el modelo lineal: En efecto, el gráfico de los valores ajustados frente a



(a) Mapa de coropletas (b) Mapa de calor

Figura 7.10: Comparación en el comportamiento de los residuos frente a los valores ajustados.
Fuente: Elaboración propia.

los residuos del modelo SEM presenta unos residuos mucho más compactos y no tan dispersos como en el modelo lineal, sin apreciarse ningún tipo de tendencia.

7.4.3. Conclusiones

Tras construir una matriz de pesos espaciales cuyas adyacencias se basan en la contigüidad de la reina, para tener presente cualquier tipo de posible adyacencia entre las comarcas. Gracias al uso de la matriz de pesos, que además se estandarizó, el estudio ha arrojado la existencia de autocorrelación espacial global en nuestra variable dependiente, la tasa de paro, además de una autocorrelación espacial local en diversas comarcas del suroeste de la Comunidad y en el norte de la misma. Por este motivo, y aunque el modelo lineal era relativamente aceptable, debemos utilizar en nuestro estudio un modelo autorregresivo espacial que tenga en cuenta la dependencia espacial existente en nuestros datos.

El modelo idóneo para estudiar la tasa de paro es el modelo de error espacial autorregresivo, en el que variables omitidas en el modelo trasladan hacia los términos del error la autocorrelación existente en la tasa de paro.

Anexo I: Bases de datos utilizadas

Los archivos necesarios para realizar el análisis se pueden descargar del siguiente enlace: `https://drive.google.com/file/d/0BzPIuABKAuzRTGFaUUdNejN1dTQ/view?usp=sharing`. A continuación, se presentan las distintas variables a estudio, siendo aquellas sombreadas en gris las que no entraron en el estudio:

CARACTERÍSTICAS GENERALES			
Variable	*Declaración en R*	*Medida*	*Descripción*
Código comarca	cod_com	-	Código a nivel estatal
Comarca	comarca	-	Nombre de la comarca
Superficie	superficie	Km2	Superficie de la comarca
Población	padron	Habitantes	Número de habitantes (empadronados)
Municipios	cant_mun	Unidades	Cantidad de municipios
Capital	capital	-	Capital comarcal
Provincia	provincia	-	Provincia a la que pertenece

Tabla 8: Datos de carácter general
Fuente: Elaboración propia.

DATOS SOCIO-ECONÓMICOS			
Variable	*Declaración en R*	*Medida*	*Descripción*
Presupuesto medio por habitante (2016)	prepto_medio	euros/hab	-
Gasto medio por habitante (Liquidación presup. 2015)	gasto_medio	euros/hab	-
Deuda viva de la comarca (2015)	deudaviva	euros/hab	Deuda bancaria de la comarca
Valor catastral medio (2016)	valor_cat_medio	euros	Valor medio de bienes inmuebles
Número total de empresas (2016)	-	Unidades	
Empresas en el sector Industria (2016)	empIndustPG	%	
Empresas en el sector Construcción (2016)	empConstrucPG	%	
Empresas en el sector Comercio, transporte y hostelería (2016)	empCTHostPG	%	
Empresas en el sector Servicios (2016)	empServPG	%	Empresas de la información y comunicación, financieras y de seguros, inmobiliarias, relativas a actividades profesionales y técnicas, del sector de la educación, sanidad y servicios, además de aquellas relativas a otros servicios personales
Empresas de otros sectores (2016)	empOtrasPG	%	Porcentaje de empresas no expresado por las anteriores

Tabla 9: Datos de carácter Socio-económico. Fuente: Elaboración propia.

DATOS DEMOGRÁFICOS			
Variable	*Declaración en R*	*Medida*	*Descripción*
Padrón (2016)	-	-	-
Variación padrón (2015-2016)	difPadronPG	%	Diferencia entre los residentes empadronados
Población respecto a la provincia (2016)	poblComarcProvPG	%	-
Población respecto a la Comunidad (2016)	-	-	-
Densidad de población (2016)	densidadPobl	hab/Km2	Cantidad de habitantes de acuerdo a la superficie de la comarca
Españoles residentes en el extranjero (2017)	expatriados	Unidades	-
Menores de 16 años	menores16PG	%	-
De 16 a 29 años	edad16_29PG	%	-
De 30 a 64 años	edad30_64PG	%	-
Mayores de 64 años	-	-	
Índice de dependencia	indiceDepciaPG	%	Proporción entre población dependiente y la activa de la que aquella depende
Nacidos en la C.V.	-	Unidades	-
Extranjeros	-	Unidades	-
Nacidos en la C.V.	nacidosCVPG	%	
Extranjeros	extranjerosPG	%	-

Tabla 10: Datos de carácter demográfico. Fuente: Elaboración propia.

Los datos con carácter laboral tienen fecha de 28/02/2017, ya que los datos se recopilaron durante el mes de marzo, salvo las variables: total de afiliados residentes en la comarca, mujeres afiliadas residentes en la comarca y tasa de afiliación en la comarca. Estas variables tenían fecha de 31/12/2016.

DATOS LABORALES			
Variable	*Declaración en R*	*Medida*	*Descripción*
Paro registrado (27/02/2017)	-	-	-
Paro registrado en menores de 25 años	paroInf25PG	%	-
Paro registrado en mujeres	paroMujeresPG	%	-
Tasa de paro	tasaParoPG	%	-
Contratación registrada	-	-	-
Índice de rotación contractual	indceRotContrac	contratos/personas	Contratos que se realiza a una persona en un periodo determinado
Paro registrado en Agricultura	paroAgricPG	%	-
Paro registrado en Industria	paroIndPG	%	-
Paro registrado en Construcción	paroConstPG	%	-
Paro registrado en Servicios	paroServPG	%	-
Contratación registrada en Agricultura	contratAgriPG	%	-
Contratación registrada en Industria	contratIndPG	%	-
Contratación registrada en Construcción	contratConstPG	%	-
Contratación registrada en Servicios	contratServPG	%	-
Total de afiliados	afiliadosSS	Unidades	-
Tasa de afiliación	tasaAfilPG	%	-
Régimen General	regGenPG	%	Trabajadores que no están en regímenes especiales, autónomos, minería y mar
Régimen General- Sistema especial hogar	regGenHogarPG	%	-
Régimen General- Sistema Especial agrario	regAgrarPG	%	-
Régimen Especial- Trabajadores autónomos	regAutonomPG	%	-
Régimen Especial- Mar	regMarPG	%	-
Total de afiliados residentes en la comarca	afilResidComar	%	-
Mujeres afiliadas residentes en la comarca	mujersAfilResComarPG	%	-
Tasa de afiliación residentes en la comarca	tasaAfilResComarPG	%	-

Tabla 11: Datos de carácter laboral. Fuente: Elaboración propia.

Anexo II: Código R

```
##############################################
### FILTRADO DE VARIABLES Y MODELO LINEAL ###
##############################################

###LIBRERIAS
library(readxl)
library(corrplot)
library(MASS)
library(lmtest)
library(car)
library(faraway)

#LECTURA DE DATOS

#Si la base de datos se encuentra en el directorio de trabajo se puede
#utilizar:

datosGen<-read_excel("Datos_comarcas_CV_R.xlsx",
sheet = "Datos Generales")
datosSocioEc<-read_excel("Datos_comarcas_CV_R.xlsx",
sheet = "Datos Socioeconomicos")
datosDemo<-read_excel("Datos_comarcas_CV_R.xlsx",
sheet = "Datos Demograficos")
datosLaboral<-read_excel("Datos_comarcas_CV_R.xlsx",
sheet = "Datos Laborales")

#Con la funcion file.choose() se abre la ventana de documentos de
```

```
#windows. Seleccionar el archivo de excel "Datos_comarcas_CV_R.xlsx" y
#aceptar, se cargara la base de datos que aparece en el codigo.

#datosGen<-read_excel(file.choose(), sheet = "Datos Generales")
#datosSocioEc<-read_excel(file.choose(), sheet = "Datos Socioeconomicos")
#datosDemo<-read_excel(file.choose(), sheet = "Datos Demograficos")
#datosLaboral<-read_excel(file.choose(), sheet = "Datos Laborales")

#ELIMINAR VARIABLES
names(datosDemo)
datosDemo<-datosDemo[,-1]
names(datosSocioEc)
datosSocioEc<-datosSocioEc[,-1]
names(datosLaboral)
datosLaboral<-datosLaboral[,-1]

#INTRODUCCION VAR. DEPENDIENTE EN DATOS DEMOGRAFICOS Y SOCIOECONOMICOS
datosDemo$tasaParoPG<-datosLaboral$tasaParoPG
datosSocioEc$tasaParoPG<-datosLaboral$tasaParoPG

###ANALISIS DE LAS DISTINTAS BASES DE DATOS CARGADAS
##DATOS DEMOGRAFICOS

#CORRELACIONES
corrDemog<-cor(datosDemo)
corrplot(corrDemog, method = "number",tl.pos = "lt",tl.cex=1,
tl.col="blue",number.cex=0.9)



#MODELO INICIAL y AIC
modDemog1<-lm(tasaParoPG~difPadronPG+poblComarcProvPG+densidadPobl+
expatriados+menores16PG+edad16_29PG+edad30_64PG+indiceDepciaPG+
nacidosCVPG+extranjerosPG,data = datosDemo)
```

```
modAICDemog1<-stepAIC(modDemog1, direction ="both", k=2)

#ANOVA DEL MODELO AIC PLANTEADO (Comprobacion significatividad variables)
anova(modAICDemog1)

#NUEVO MODELO (modelo 2)
modDemog2<-lm(tasaParoPG~edad30_64PG+indiceDepciaPG,
data = datosDemo)
modAICDemog2<-stepAIC(modDemog2,direction = "both",k=2)

#ANOVA NUEVO MODELO (modelo 2)
anova(modAICDemog2)
summary(modAICDemog2)

##DATOS LABORALES

#CORRELACIONES
corrLaboral<-cor(datosLaboral)
corrplot(round(corrLaboral,1), method = "number",tl.pos = "lt",tl.cex=1,
tl.col="blue",number.cex=0.9)

#MODELO INICIAL y AIC
modLabo1<-lm(tasaParoPG~paroInf25PG+paroMujeresPG+indceRotContrac+
paroAgricPG+paroIndPG+paroConstPG+paroServPG+contratAgriPG+contratIndPG+
contratConstPG+contratServPG+afiliadosSS+tasaAfilPG+regGenPG+
regGenHogarPG+regAgrarPG+regAutonomPG+regMarPG+afilResidComar+
mujersAfilResComarPG+tasaAfilResComarPG,data = datosLaboral)

modAICLabo1<-stepAIC(modLabo1, direction ="both", k=2)

#ANOVA DEL MODELO AIC PLANTEADO (Comprobacion significatividad variables)
anova(modAICLabo1)
```

```
#NUEVO MODELO (modelo 2)
modLabo2<-lm(tasaParoPG~paroInf25PG+indceRotContrac+paroIndPG+
paroConstPG+paroServPG+ contratServPG+afiliadosSS+tasaAfilPG+
regGenPG+tasaAfilResComarPG,data = datosLaboral)

modAICLabo2<-stepAIC(modLabo2, direction ="both", k=2)

#ANOVA NUEVO MODELO (modelo 2)
anova(modAICLabo2)

#NUEVO MODELO (modelo 3)
modLabo3<-lm(tasaParoPG~indceRotContrac+paroIndPG+paroServPG+
contratServPG+afiliadosSS+tasaAfilPG+regGenPG+tasaAfilResComarPG,
data = datosLaboral)

modAICLabo3<-stepAIC(modLabo3, direction ="both", k=2)

#ANOVA NUEVO MODELO (modelo 3)
anova(modAICLabo3)

#NUEVO MODELO (modelo 4)
modLabo4<-lm(tasaParoPG~paroIndPG+afiliadosSS+tasaAfilPG+
regGenPG+tasaAfilResComarPG,data = datosLaboral)

modAICLabo4<-stepAIC(modLabo4, direction ="both", k=2)

#ANOVA NUEVO MODELO (modelo 4)
anova(modAICLabo4)
summary(modAICLabo4)

##DATOS SOCIOECONOMICOS

#CORRELACIONES
corrSocioEC<-cor(datosSocioEc)
```

```
corrplot(corrSocioEC, method="number",tl.pos = "lt",tl.cex=1,
tl.col="blue",number.cex=0.9)

#MODELO INICIAL y AIC
modSocioEc1<-lm(tasaParoPG~prepto_medio+gasto_medio+deudaviva+
valor_cat_medio+empIndustPG+empConstrucPG+empCTHostPG+empServPG+
empOtrasPG,data = datosSocioEc)

modAICSocioEc1<-stepAIC(modSocioEc1, direction ="both", k=2)

#ANOVA DEL MODELO AIC PLANTEADO (Comprobacion significatividad variables)
anova(modAICSocioEc1)
summary(modAICSocioEc1)

####################
# MODELO AGRUPADO #
####################

#CREACION DATOS AGRUPADOS POR LAS VARIABLES SIGNIFICATIVAS DE LOS
#DISTINTOS MODELOS LINEALES

#Se quitaron las dos variables dependientes relativas a la tasa de
#afiliacion (tasaAfilPG y tasaAfilResComarPG) ya que explicarian
#demasiado bien a la variable independiente por tener una relacion
#tan directa que restarian significatividad al resto de variables.
datosnuevomodelo<-c(datosLaboral$tasaParoPG,datosDemo$indiceDepciaPG,
datosLaboral$paroIndPG,datosLaboral$afiliadosSS,datosLaboral$regGenPG,
datosSocioEc$empIndustPG,datosSocioEc$empServPG)
nombvar<-c("tasaParoPG","indiceDepciaPG","paroIndPG","afiliadosSS",
"regGenPG","empIndustPG","empServPG")

datosAnalisismatrix<-matrix(c(datosnuevomodelo),nrow = 34,byrow = F)
dimnames(datosAnalisismatrix)<-list(c(),nombvar)
datosanalisis<-data.frame(datosAnalisismatrix)
```

```
rm(datosAnalisismatrix,datosnuevomodelo,nombvar)

#ANALISIS MODELO AGRUPADO

#CORRELACIONES
corrdatanalisis<-cor(datosanalisis)
corrplot(corrdatanalisis, method="number",tl.pos = "lt",
tl.cex=1,tl.col="blue",number.cex=0.8)

#MODELO INICIAL y AIC
modglobal1<-lm(tasaParoPG~indiceDepciaPG+paroIndPG+afiliadosSS+
regGenPG+empIndustPG+empServPG,data = datosanalisis)

modAICglobal1<-stepAIC(modglobal1, direction ="both", k=2)

#ANOVA DEL MODELO AIC PLANTEADO (Comprobacion significatividad variables)
anova(modAICglobal1)
summary(modAICglobal1)

##DIAGNOSTICO MODELO OBTENIDO

#GRAFICOS:
par(mfrow=c(2,2))
plot(modAICglobal1)
par(mfrow=c(1,1))

#CONSTRUCCION RESIDUOS DEL MODELO
resid_modAICglobales<-residuals(modAICglobal1)

#TESTS DEL MODELO

#LINEALIDAD Test Reset de Ramsey
resettest(modAICglobal1)
```

```
#NORMALIDAD Test de Shapiro
shapiro.test(resid_modAICglobales)

#HOMOGENEIDAD (varianzas constantes)
bptest(modAICglobal1)
#(homocedasticidad)

#AUTOCORRELACION Test de Durbin-Watson
dwtest(modAICglobal1, alternative = "two.sided")

#VALORES ATIPICOS
outlierTest(modAICglobal1)
influencePlot(modAICglobal1, id.n = 2, col="blue")
labels<-rownames(datosanalisis)
cook<-cooks.distance(modAICglobal1)
halfnorm(cook, 3, labs = labels, ylab = "Distancia de Cook",col="red")

################################################
### DESCRIPTIVO DE LAS VARIABLES A ESTUDIO ###
################################################

###LIBRERIAS
library(readxl)
library(sp)
library(RColorBrewer)
library(classInt)
library(knitr)


###CARGAR BBDD CON LA QUE REALIZAR EL ANALISIS ESPACIAL
#Si la base de datos esta en el directorio de trabajo se puede utilizar:
spatdata<-read_excel("Variables_para_rds_R.xlsx")
```

```
#Se pasaron todas las variables que salieron significativas a un excel,
#por lo que se debe seleccionar el archivo "Variables_para_rds_R.xlsx"
#spatdata<-read_excel(file.choose())

#CARGAR ARCHIVO RDS CON LAS COMARCAS A NIVEL NACIONAL.
#El archivo tiene separada la comarca de valencia en dos: valencia y
#huerta sur-valencia. Se eliminara esta ultima.
archrds<-readRDS("ESP_adm3.rds")
View(archrds)

#UTILIZAMOS SOLO LAS COMARCAS DE LA COMUNIDAD VALENCIANA
BORRARCOMARCAS<-which(archrds$NAME_1!="Comunidad Valenciana")
RDS_CV<-archrds[-BORRARCOMARCAS,]
View(RDS_CV)
rm(archrds,BORRARCOMARCAS)

#ELIMINAMOS OBSERVACION 26 DEL RDS POR CONTENER DOS COMARCAS PARA
#VALENCIA , VALENCIA Y HUERTA SUR-VALENCIA
RDS_CV<-RDS_CV[-26,]
#CAMBIAMOS EL NOMBRE DE LAS COMARCAS DE VALENCIANO A CASTELLANO
Comarcas<-c("CAMPO DE ALICANTE", "HOYA DE ALCOY","ALTO VINALOPO",
"VEGA BAJA","CONDADO DE COCENTAINA","MARINA ALTA","MARINA BAJA",
"VINALOPO MEDIO","BAJO VINALOPO","ALCALATEN","ALTO MIJARES",
"ALTO PALANCIA","ALTO MAESTRAZGO","BAJO MAESTRAZGO","LOS PUERTOS",
"PLANA ALTA","PLANA BAJA","CAMPO DE MURVIEDRO", "CAMPO DE TURIA",
"CANAL DE NAVARRES","VALENCIA","COSTERA","HOYA DE DE BUNYOL",
"HUERTA NORTE","HUERTA OESTE","HUERTA SUR","SAFOR",
"PLANA DE UTIEL REQUENA","RIBERA ALTA","RIBERA BAJA",
"RINCON DE ADEMUZ","SERRANOS","VALLE DE ALBAIDA",
"VALLE DE COFRENTES-AYORA")

RDS_CV$NAME_3<-Comarcas
rm(Comarcas)
```

```
#INTRODUCCION VARIABLES DE SPATDATA A RDS_CV
#Como tienen el mismo orden podemos hacerlo sin problemas
RDS_CV@data$tasaParoPG<-spatdata$tasaParoPG
RDS_CV@data$indiceDepciaPG<-spatdata$indiceDepciaPG
RDS_CV@data$paroIndPG<-spatdata$paroIndPG
RDS_CV@data$empIndustPG<-spatdata$empIndustPG

#DECLARACION DE VARIABLES POR COMODIDAD
#1. Tasa de paro(%)
tasa_de_paro<-RDS_CV@data$tasaParoPG
tasa_de_paro<-data.frame(tasa_de_paro)

#2. Indice de dependencia (%)
indice_depcia<-RDS_CV@data$indiceDepciaPG
indice_depcia<-data.frame(indice_depcia)

#3. Paro sector industria (%)
paro_indust<-RDS_CV@data$paroIndPG
paro_indust<-data.frame(paro_indust)

#5. Empresas sector industria (%)
empresas_ind<-RDS_CV@data$empIndustPG
empresas_ind<-data.frame(empresas_ind)

#QUITAMOS COLUMNAS INSERVIBLES DEL ARCHIVO RDS
columnasinservibles<-c(11,12,15,16)
RDS_CV@data<-RDS_CV@data[,-columnasinservibles]
View(RDS_CV)
rm(columnasinservibles)

#MAPA CON ACRONIMOS DE LAS COMARCAS
#Acronimos provincias:
#CAMPO DE ALICANTE="ALC",HOYA DE ALCOY="HA",ALTO VINALOPO="AV",
#VEGA BAJA="VB",CONDADO DE COCENTAINA="CC",MARINA ALTA="MA",
```

```
#MARINA BAJA="MB",VINALOPO MEDIO="VM",BAJO VINALOPO="BV",
#ALCALATEN="AT",ALTO MIJARES"AM",ALTO PALANCIA="AP",
#ALTO MAESTRAZGO="AMA",BAJO MAESTRAZGO="BMA",LOS PUERTOS="LP",
#PLANA ALTA="PA",PLANA BAJA="PB",CAMPO DE MURVIEDRO="CM",
#CAMPO DE TURIA="CT",CANAL DE NAVARRES="CN",VALENCIA="VCIA",
#COSTERA="CTR",HOYA DE DE BUNYOL="HB",HUERTA NORTE="HN",
#HUERTA OESTE="HO",HUERTA SUR="HS",SAFOR="LS",
#PLANA DE UTIEL REQUENA="PUR",RIBERA ALTA="RA",RIBERA BAJA="RB",
#RINCON DE ADEMUZ="RAD",SERRANOS="SS",VALLE DE ALBAIDA="VA",
#VALLE DE COFRENTES-AYORA="VC".

#VECTOR CON ACRONIMOS
AcrCom<-c("ALC","HA","AV","VB","CC","MA","MB","VM","BV","AT","AM","AP",
"AMA","BMA","LP","PA","PB","CM","CT","CN","VCIA","CTR","HB","HN","HO",
"HS","LS","PUR","RA","RB","RAD","SS","VA","VC")

#Creamos unos centroides con la funcion coordinates y asi representar
#los acronimos en las comarcas
centroides<-coordinates(RDS_CV)
plot(RDS_CV)#probamos si se corresponde todo
title(main = "Acrónimos de las comarcas",cex=3,cex.main = 1.5,
cex.main = 1.5)
text(centroides,AcrCom,cex=0.7,col = "darkblue")

###REPRESENTACION MAPAS

#MAPA POBLACION, CANTIDAD DE MUNICIPIOS COMARCAS Y DENSIDAD DE POBLACION
poblmun<-datosGen[,c(1,4,5)] #seleccion cod. comarca, padron y
#cantidad municipios
poblmun<-data.frame(poblmun)
densidadPoblacion<-datosDemo[,3] #seleccion densidad de poblacion
densidadPoblacion<-data.frame(densidadPoblacion)

#ASIGNAMOS LOS NOMBRES DE FILAS de RDS_CV A POBLMUN Y DENSIDADPOBLACION
```

```
#(son distintos entre si). ESTABLECEMOS VARIABLES COMO
#SPATIALPOLYGONSDATAFRAMES
row.names(poblmun)<-row.names(RDS_CV)
row.names(densidadPoblacion)<-row.names(RDS_CV)
poligonos.datapobl<-SpatialPolygonsDataFrame(RDS_CV,poblmun)
poligonos.datadens<-SpatialPolygonsDataFrame(RDS_CV,densidadPoblacion)
plotvarpobl<-poligonos.datapobl$padron
plotvardens<-poligonos.datadens$densidadPobl
plotvarmun<-poligonos.datapobl$cant_mun

#Poblacion comarcas
nclr<-5 # Numero de colores
plotclr<-brewer.pal(nclr,"Accent")
class<-classIntervals(plotvarpobl,nclr,n=5,style="quantile") #Quintiles

colcode<-findColours(class,plotclr) # defino paleta de colores

plot (poligonos.datapobl, col=colcode, border="darkblue")
title(main = "Población en las distintas comarcas (habitantes)",cex=3,cex
    .main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.datapobl,"padron",
main=list(label="Población en las distintas comarcas (habitantes)",
cex=1.5))

#Densidad poblacion comarcas
nclr<-5 # Numero de colores
plotclr<-brewer.pal(nclr,"Accent")
class<-classIntervals(plotvardens,nclr,n=5,style="quantile") #Quintiles

colcode<-findColours(class,plotclr) # defino paleta de colores
```

```
plot (poligonos.datadens, col=colcode, border="darkblue")
title(main = "Densidad de población comarcal (hab/km2)",
cex=3,cex.main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.datadens,"densidadPobl",
main=list(label="Densidad de población comarcal(hab/km2)",cex=1.5))

#Municipios comarcas
nclr<-10 # Numero de colores
plotclr<-brewer.pal(nclr,"Spectral")
class<-classIntervals(plotvarmun,nclr,style="equal") #Establezco 10

colcode<-findColours(class,plotclr) # defino paleta de colores

plot (poligonos.datapobl, col=colcode, border="darkblue")
title(main = "Cantidad de municipios por comarca",cex=3,cex.main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.datapobl,"cant_mun",
main=list(label="Cantidad de municipios por comarca",cex=1.5))

#VARIABLES FIJAS EN MAPAS DE VARIABLES MODELO
nclr<-4 #cantidad de colores a graficar
n<-4 #cuartiles
plotclr<-brewer.pal(nclr,"YlGn")

#1.Tasa de paro (%)
row.names(tasa_de_paro)<-row.names(RDS_CV)
```

```
poligonos.data<-SpatialPolygonsDataFrame(RDS_CV,tasa_de_paro)
plotvartasaparo<-poligonos.data$tasa_de_paro#variable a graficar
#intervalos por cuartiles (n=4)
class<-classIntervals(plotvartasaparo,nclr,n, style="quantile")
#defino paleta de colores para los datos
colcode<-findColours(class,plotclr)
plot(poligonos.data, col=colcode, border="darkblue")
title(main = "Tasa de paro comarcal (%)",cex=3,cex.main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.data,"tasa_de_paro",
main=list(label="Tasa de paro comarcal en febrero (%)",cex=1.5))
#2. Indice de dependencia (%)
row.names(indice_depcia)<-row.names(RDS_CV)
poligonos.data<-SpatialPolygonsDataFrame(RDS_CV,indice_depcia)
plotvarindiceDepciaPG<-poligonos.data$indice_depcia
# establezco intervalos por cuartiles (n=4)
class<-classIntervals(plotvarindiceDepciaPG,nclr,n, style="quantile")
#defino paleta de colores para los datos
colcode<-findColours(class,plotclr)
plot(poligonos.data, col=colcode, border="darkblue")
title(main = "Indice de dependencia 2016",cex=3,cex.main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.data,"indice_depcia",
main=list(label="Indice de dependencia 2016",cex=1.5))
#3. Porcentaje paro registrado en industria
row.names(paro_indust)<-row.names(RDS_CV)
poligonos.data<-SpatialPolygonsDataFrame(RDS_CV,paro_indust)
plotvarparoindust<-poligonos.data$paro_indust
```

```
#establezco intervalos por cuartiles (n=4)
class<-classIntervals(plotvarparoindust,nclr,n, style="quantile")
#defino paleta de colores para los datos
colcode<-findColours(class,plotclr)
plot(poligonos.data, col=colcode, border="darkblue")
title(main = "Paro registrado en sector industria (%)",cex=3,
cex.main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.data,"paro_indust",
main=list(label="Paro registrado en sector industria (%)",cex=1.5))
#4. Porcentaje empresas sector industria
row.names(empresas_ind)<-row.names(RDS_CV)
poligonos.data<-SpatialPolygonsDataFrame(RDS_CV,empresas_ind)
plotvarempresasind<-poligonos.data$empresas_ind
#establezco intervalos por cuartiles (n=4)
class<-classIntervals(plotvarempresasind,nclr,n, style="quantile")
#defino paleta de colores para los datos
colcode<-findColours(class,plotclr)
plot(poligonos.data, col=colcode, border="darkblue")
title(main = "Empresas sector industria (%)",cex=3,cex.main = 1.5)
legend(0.8,40.8,legend = names(attr(colcode,"table")),
fill= attr(colcode,"palette"), bty = "n")
text(centroides,AcrCom,cex=0.7)
#Mapa de calor
spplot(poligonos.data,"empresas_ind",
main=list(label="Empresas sector industria (%)",cex=1.5))

# TABLA DESCRIPTIVA DE LOS DATOS MODELIZADOS
kable(summary(spatdata[,-1]))
#Desviaciones estandar de las variables
sd(spatdata$tasaParoPG)
```

```
sd(spatdata$indiceDepciaPG)
sd(spatdata$paroIndPG)
sd(spatdata$empIndustPG)

#########################
### ANALISIS ESPACIAL ###
#########################
#LIBRERIAS
library(spdep)
###ESTRUCTURA ESPACIAL
#Objeto de adyacencias basado en la contiguidad de la reina
lista_adyac<-poly2nb(RDS_CV, row.names=RDS_CV$OBJECTID, queen=TRUE)
lista_adyac

#MAPA CV CON LAS ADYACENCIAS ENTRE COMARCAS MEDIANTE LA CONTIGUIDADDE LA
#REINA
par(mai=c(0,0,0,0))
plot(RDS_CV, col='gray', border='blue')
x <-coordinates(RDS_CV)
plot(lista_adyac, x, col='red', lwd=2, add=TRUE)

#Matriz de pesos espaciales calculando los pesos ponderados (style="W")
lista_adyac #Sabemos que existe una isla, aun asi se comprueba
WW<-nb2listw(lista_adyac, style="W",zero.policy=TRUE)
#Establecemos zero.policy=TRUE debido a la existencia de una isla

#CALCULO DEL I DE MORAN
#Por supuesto, se utiliza la variable dependiente
moran(RDS_CV@data$tasaParoPG,WW,n=length(WW$neighbours),S0=Szero(WW),
zero.policy=TRUE)
#Residuos del modelo lineal
moran(resid_modAICglobales,WW,n=length(WW$neighbours),Szero(WW),zero.
    policy=TRUE)
```

```
#CALCULO DEL ESTADISTICO DE SIGNIFICACION Z
#(Deteccion de autocorrelacion espacial en residuos)
#(randomisation=FALSE para calcular varianza bajo normalidad)
moran.test(RDS_CV@data$tasaParoPG,WW,zero.policy = TRUE,
randomisation = FALSE)

#ESTUDIO RESIDUOS MODELO LINEAL
lm.morantest(modAICglobal1,WW,zero.policy = TRUE)

#DIAGRAMA DE DISPERSION DE MORAN
moran.plot(RDS_CV@data$tasaParoPG,WW,zero.policy = TRUE, labels=AcrCom,
xlab = "Tasa de paro (%)",
ylab = "Tasa de paro espacialmente retardada (Wy)",pch=19)

#TEST LOCAL DE MORAN
localmoran(RDS_CV@data$tasaParoPG,WW,zero.policy = TRUE)

#DETECCION DEL TIPO DE PROCESO AUTORREGRESIVO PRESENTE EN LOS RESIDUOS
#MEDIANTE LOS MULTIPLICADORES DE LAGRANGE
#(Estadistico de puntuaciones de Rao) EN LOS MODELOS SAR Y SEM
lagrange_multiplier_test<-lm.LMtests(modAICglobal1,WW,zero.policy = TRUE,
test = c("LMerr", "LMlag"))
summary(lagrange_multiplier_test) #Debemos utilizar el modelo SEM

#MODELOS ESPACIALES SIMULTANEOS AUTORREGRESIVOS
#Modelo SEM
spatial_error_model<-errorsarlm(tasaParoPG ~ indiceDepciaPG +paroIndPG +
empIndustPG,data = datosanalisis,WW,zero.policy = TRUE)
summary(spatial_error_model)

#Modelo SAR
spatial_lag_model<-lagsarlm(tasaParoPG ~ indiceDepciaPG +paroIndPG +
empIndustPG,data = datosanalisis,WW,zero.policy = TRUE)
summary(spatial_lag_model)
```

```
#TEST DE BREUSCH-PAGAN PARA MODELOS ESPACIALES
bptest.sarlm(spatial_error_model)

#GRAFICOS DE RESIDUOS
#Residuos modelo SEM
plot(spatial_error_model$fitted.values,spatial_error_model$residuals,
main="Valores ajustados Vs residuos modelo SEM",
xlab="Valores ajustados modelo SEM",
ylab="Residuos modelo SEM")

#Residuos modelo lineal
plot(modAICglobal1$fitted.values,modAICglobal1$residuals,
main="Valores ajustados Vs residuos modelo lineal",
xlab="Valores ajustados modelo lineal",
ylab="Residuos modelo lineal")
```

Bibliografía

[1] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers. Capítulo 12 páginas [181-183].

[2] Anselin, L. (2001). *Spatial Econometrics. A Companion to Theoretical Econometrics*. Blackwell Publishing Ltd. Capítulo 14, página 312.

[3] Anselin, L. (2015). Lecture: Maximum Likelihood Estimation - Spatial Error Model. Subject: *Spatial Regression Analysis*. Canal GeoDa Software. URL: `https://www.youtube.com/watch?v=GhPGgPOxcmM`

[4] Anselin, L. (2015). Lecture: Maximum Likelihood Estimation - Spatial Lag Model. Subject: *Spatial Regression Estimation*. Canal GeoDa Software. URL: `https://www.youtube.com/watch?v=GHP-wnzAARk`

[5] Anselin, L. (2017). Class: Spatial Process Models (Part I). Subject: *Spatial Regression Analysis*. University of Chicago. URL: `https://www.youtube.com/watch?v=fdCNctqDyo4&index=4`

[6] Anselin, L. (2017). Class: Spatial Process Models (Part II). Subject: *Spatial Regression Analysis*. University of Chicago. URL: `https://www.youtube.com/watch?v=FLQcE7FZIxU&index=5`

[7] Anselin, L. (2017). Class: Specification of Spatial Dependence. Subject: *Spatial Regression Analysis*. University of Chicago. URL: `https://www.youtube.com/watch?v=b2ceaoe2Wjs&index=6`

[8] Arellano, M. (1991). *Introducción al análisis econométrico con datos de panel*. Servicio de estudios del Banco de España (documento de trabajo nº 9222). Página 3.

[9] Banerjee, S., Carlin, B. P., Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data, second edition*. CRC Press. Capítulo 4, páginas [78-83].

[10] Chasco Yrigoyen, C. (2002). Tesis doctoral: *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Universidad Autónoma de Madrid. Capítulo 2, página 17; capítulo 3, páginas [63, 65, 71], capítulo 4, páginas [104-106, 129].

[11] Dubé, J. & Legros, D. (2014). *Spatial Econometrics Using Microdata*. ISTE & Wiley. Capítulo 1, páginas [13, 18]; capítulo 2, páginas [29, 31-43, 45, 47, 49-51]; capítulo 3, páginas [60-68, 70-72 , 80, 84-85]; Anexo 3, páginas [200-207].

[12] Dubin, R. (2009). Spatial Weights. *The SAGE Handbook of Spatial Analysis*. Capítulo 8, página 125.

[13] Fischer, M. M. (2005). Spatial analysis: retrospect and prospect. *Geographical Information Systems: Principles, Techniques, Management and Applications, second edition*. Abridged. Capítulo 19, página 285.

[14] Fischer, M. M., Bartkowska, M., Riedl, A., Sardadvar, S., kunnert, A. (2009). The Impact of Human Capital on Regional Labor Productivity in Europe. *Handboook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Springer. Part E (E1), página 588.

[15] Getis, A. (2005). Spatial Statistics. *Geographical Information Systems: Principles, Techniques, Management and Applications, second edition*. Abridged. Capítulo 16, página 241.

[16] Getis, A. (2008). A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective. *Geographical Analysis, 40*. Páginas [299, 300, 302].

[17] LeSage, J. P. (1998). *Spatial Econometrics*. Department of Economics University of toledo. Páginas 8-10.

[18] LeSage, J. P. (2004). Lecture 1: Maximum likelihood estimation of spatial regression models. *Course on Spatial Econometrics*. University of Toledo, Department of Economics.Página 7. URL: `http://www4.fe.uc.pt/spatial/doc/lecture1_slide.pdf`

[19] Lloyd, C. D. (2011). *Local Model for Spatial Analysis, second edition*. CRC Press. Capítulo 4, páginas [86, 88].

[20] Miller, H. J. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers, 94 (2)*. Página 284.

[21] Pace, R.K. & LeSage, J. (2009). *Introduction to Spatial Econometrics*. CRC Press. Capítulo 1, páginas [2, 8-9].

[22] Pace, R.K. & LeSage, J. (2010). Spatial Econometrics. *Handbook of Spatial Statistics*. CRC Press. Capítulo 15, páginas [246, 248, 250].

[23] Plant, R. E. (2012). *Spatial Data Analysis in Ecology and Agriculture Using R*. CRC Press. Capítulo 13, página 432.

[24] Urban, D. L. (2003). *Spatial Analysis in Ecology, Mantel's test*. National Center for Ecological Analysis and Synthesis (NCEAS). Página 1. URL: `https://www.nceas.ucsb.edu/files/scicomp/doc/SpatialEcologyMantelTest.pdf`

[25] Vilalta, C. J. (2005). *Cómo enseñar autocorrelación espacial*. Universidad de Monterrey, México. Pag. 325

[26] Viton, P. A. (2010). *Notes on Spatial Econometric Models*. City and regional planning 870.03. Capítulo 1, página 2; capítulo 2, página 3; capítulo 3, páginas 4-6; capítulo 4, página 9; capítulo 5, página 12.

[27] Wickham, H. (2014). *Advanced R*. Chapman and Hall/CRC, The R Series. Capítulo 3, página 39; capítulo 7, página 111.