Trabajo Fin de Máster

Análisis de datos espaciales de devoluciones

Cristina Martínez Reglero

Máster en Técnicas Estadísticas

Curso 17-18

Propuesta de Trabajo Fin de Máster

Título en galego: Análise de datos espaciais de devolucións.

Título en español: Análisis de datos espaciales de devoluciones.

English title: Analysis of returns spatial data.

Modalidad: B

Autor/a: Cristina Martínez Reglero, Universidad de A Coruña.

Director/a: Rubén Fernández Casal, Universidad de A Coruña.

Tutor/a: Jesús Salceda Sánchez, Responsable de la Oficina de Datos del Departamento de Sistemas de Inditex.

Breve resumen del trabajo:

Inditex es uno de los mayores grupos de distribución de moda del mundo y cuenta con ocho formatos comerciales con más de 6460 tiendas en 88 países. Los artículos comprados en una de las tiendas (físicas o on-line) de un determinado país, pueden ser devueltos por clientes en cualquiera de las tiendas (físicas) de la marca en ese país. La Oficina de Datos del Departamento de Sistemas de Inditex está interesada en disponer de herramientas que permitan analizar este tipo de datos de forma que se pueda utilizar esta información en la distribución de los artículos.

El objetivo principal del TFM es el análisis de devoluciones teniendo en cuenta la localización geográfica de las tiendas (posición espacial y país). Adicionalmente se podrán considerar otras variables explicativas, como por ejemplo el tipo de producto, la temporada, el tipo de empresa del grupo, etc.

Se propone emplear R para desarrollar una aplicación web mediante el paquete shiny. Para el análisis de los datos se empleará el paquete sp y los paquetes de estadística espacial que se consideren oportunos.

Recomendaciones: Es recomendable haber cursado la asignatura del MTE de Estadística Espacial.

Otras observaciones: El alumno dispondrá de ayudas de comedor mientras realiza las prácticas en el centro de Inditex.

Índice general

Capítulo 1

Introducción

La estadística es la rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades. Aunque las primeras manifestaciones estadísticas se remontan al año 3000 A.C., no es hasta los años 30 del pasado siglo cuando se desarrolla vertiginosamente y se define una metodología adecuada para abordar cada situación, teniendo en cuenta las particularidades propias de los datos analizados. Así, la estadística se empieza a dividir en diferentes especialidades, surgiendo, entre otros, el análisis de series temporales, el análisis de supervivencia o, lo que nos interesa en este trabajo, el análisis de datos espaciales.

La estadística espacial es la combinación de técnicas que buscan analizar datos espaciales. Es decir, realizar un análisis estadístico teniendo en cuenta la posición geográfica en la que se recoge cada observación de las variables. Dicha posición será absoluta sobre un sistema de coordenadas y relativa frente a otros elementos del espacio. Al recoger los datos junto con su localización los dotamos de una naturaleza georreferenciada, y esta es la característica principal de los datos espaciales.

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

En esta situación de falta de independencia también nos encontramos cuando queremos analizar series temporales, donde los datos medidos en instantes próximos estarán más relacionados que los observados con una diferencia de tiempo mayor. Sin embargo, los datos espaciales tienen una particularidad que hace imposible analizarlos de manera análoga a los temporales. Y es que, mientras la relación en el tiempo solo puede darse en una dirección, la relación en el espacio puede tomar muchas direcciones, es multidireccional.

En función de la manera en que consideremos el espacio donde se observan los datos, estos pueden dividirse en tres subtipos.

- Si los datos se pueden observar en cualquier posición esteremos trabajando bajo la perspectiva geoestadística. Cualquier localización del espacio en estudio puede seleccionarse para medir en ella la variable. Un ejemplo de datos geoestadísticos es medir los niveles de arsénico de España y relacionarlo con la incidencia del cáncer. Podríamos medir la cantidad de arsénico en cada punto de la superficie de España, pero el proceso no sería viable. Por ello se seleccionan lugares aleatorios para hacer dichas mediciones (Nuñez et al 2016).

- Si el conjunto de posiciones en los que se puede observar la respuesta es discreto, estaremos trabajando bajo el enfoque de datos reticulares. En esta situación cada observación se suele corresponder con agregaciones espaciales. Si nos interesa estudiar el precio de los pisos de alquiler un enfoque podría ser considerar agregaciones espaciales por regiones (provincias, barrios, secciones censales...) y tener una observación por cada región (Gómez et al)

- Si la variable de interés es la localización de un fenómeno concreto y, ocasionalmente, cierta información adicional de dicho fenómeno, nos encontraremos en el contexto de procesos puntuales. Un ejemplo de este tipo de datos será el estudio de incendios en un territorio, en el que las localizaciones que nos interesan son aquellas en las que se produjo el incendio y que se consideran aleatorias (Pompa y Hernández 2012).

Los primeros estudios de datos espaciales se datan en el año 1914 cuando Student estudió la distribución espacial de las partículas en un líquido. Sin embargo, no se reconoce el comienzo de la estadística espacial hasta mediados del siglo XX con los trabajos de Moran y Geary (Moran 1948; Geary 1954). En estos trabajos aparecen los primeros índices para el estudio de la dependencia y se reconoce el error al analizar estos datos sin considerar su naturaleza espacial. En 1969, Cox demostró que existía dependencia en el voto de los estadounidenses; si en un estado ganaba ampliamente el partido demócrata, la probabilidad de que ganara también en los estados cercanos aumentaba (Cox 1969).

El gran desarrollo de la estadística espacial no llega hasta finales del siglo XX. Cliff y Ord y Anselin establecen los fundamentos metodológicos de estos análisis (Cliff y Ord 1981; Anselin 1980). En 1988, Anselin publica el libro "Spatial Econometrics: Methods and Models" que supuso un gran avance en su época (Anselin 1988).

Poco a poco la estadística espacial fue ganando importancia hasta que, en 1993, Cressie publica "Statistics for Spatial Data" (Cressie 1993), libro que fue considerado la guía esencial de la estadística espacial. Con una visión principalmente práctica describe el camino a seguir para usar modelos espaciales con el fin de resolver un gran número de problemas científicos. Este libro está dividido en tres bloques, en cada uno de ellos explica cómo trabajar con cada tipo de dato espacial. Incluso hoy en día, más de 20 años después de su publicación, sigue considerándose un libro de referencia. El manual de Cressie contribuyó al incremento de popularidad de la estadística espacial haciendo que su uso se

generalizara en diversas áreas.

En este trabajo nos centraremos en el análisis de datos reticulares. En los últimos años y gracias al desarrollo de la informática se ha facilitado la realización de este tipo de análisis incrementando su uso. Además, estos análisis pueden aplicarse en áreas muy diversas, ya que todo lo que nos interesa estudiar ocurre en algún lugar. Dentro de la gran variedad de ámbitos en los que puede considerarse la naturaleza espacial de los datos, podemos destacar el área biomédica, dónde puede estudiarse las características de la región con la mayor prevalencia de una enfermedad (Nuñez 2016), el ámbito sociológico donde se estudia la intención de voto antes de unas elecciones (Cottrell et al 2016), el ámbito económico, donde puede estudiarse la distribución del precio de las viviendas (Gómez et al) o el ámbito medioambiental, donde se estudia el comportamiento de los incencios (Pompa y Hernández 2012).

Otra forma de plantear el análisis de este tipo de datos es considerándoles parte de una red (Kolaczyk y Csárdi 2014). Nos referimos a una red cuando queremos representar los elementos de un sistema y sus relaciones entre ellos. En los últimos años es frecuente escuchar que vivimos en un mundo constantemente conectado. Esta frase es la que mejor define el motivo por el que el interés en las redes se ha disparado en los últimos años. Continuamente estamos interactuando con otras personas, formando nosotros mismos parte de una red. En Internet al usar *Facebook* o *Instagram* somos parte de una red, donde nos conectamos unos con otros mediante la relación de amistad. Por algo estas dos plataformas se llaman redes sociales e incluso podemos referirnos a Internet como la red. Pero no solo formamos redes en el mundo digital, también las formamos en el mundo real, estamos conectados con todas las personas con las que nos relacionamos durante el día. La capacidad de estar conectados con otros para formar redes no es exclusiva de las personas; ecosistemas, colecciones de genes que interactúan o conexiones entre neuronas (redes neuronales) son ejemplos de redes.

El inicio del análisis basado en redes se fecha en 1735, cuando Euler dio solución al ahora famoso problema de los puentes de Königsberg (Euler 1735), donde probaba que es imposible pasar por los siete puentes de una ciudad atravesando cada uno solo una vez. A partir de ese momento, y sobre todo a partir de 1850, esta semilla se fue extendiendo en diversas áreas; en matemáticas König (König 1936) estableció sus fundamentos, en química se desarrolló para estudiar la estructura molecular y en electricidad tuvo importancia en el estudio de circuitos eléctricos.

A mediados del siglo XX, gracias al desarrollo de la informática, las redes fueron incorporándose a numerosos problemas de transporte y distribución. Durante este periodo pequeños grupos de sociólogos empezaron a desarrollar el uso de redes para caracterizar las interacciones dentro de los grupos sociales. A partir de 1990 se produjo una explosión del interés en las redes y en técnicas basadas en redes para modelar y analizar sistemas complejos. La física estadística y la ciencia de la computación fueron las dos áreas que más impulsaron este crecimiento.

Hoy en día se utiliza en áreas como la biología computacional (para estudiar la interacción entre genes, proteínas u organismos), en ingeniería (desarrollando

la mejor manera de diseñar y desarrollar redes para detección de dispositivos), en economía (estudiando la interacción entre el banco mundial como parte de la economía del mundo), en publicidad (evaluando la medida en la que la compra de un producto puede ser inducida como un tipo de contagio), en política (estudiando como las preferencias de voto en un grupo evolucionan frente a diversas fuerzas) y la salud pública (estudiando cómo se propaga una enfermedad infecciosa y encontrar como frenarla). Un hito en la biología computacional, que cambió la biología para siempre, fue la obtención del mapa del genoma humano. Para conseguir este importante hallazgo se utilizaron redes que ayudaron a entender en detalle como las componentes del cuerpo humano trabajan juntas.

Cuando decimos que el desarrollo de la informática ha contribuido al aumento de estos análisis (tanto análisis reticulares como en red), nos referimos tanto a la facilidad con la que podemos almacenar los datos como a la disponibilidad de software para realizar dichos análisis. La memoria de los ordenadores actuales nos permite almacenar grandes cantidades de datos y nos resulta más sencillo conseguirlos, por ejemplo, en la página web del Instituto Nacional de Estadística (www.ine.es) podemos acceder y descargar de manera gratuita bases de datos de este tipo. Incluso podemos representar los datos que seleccionemos en el mapa de manera automática. Algunos de los programas desarrollados para analizar estos datos son *SpaceStat, MinitaB* y *ArcView* para los datos reticulares y *Network Workbench* y *SocNetV* para los datos en red. Además, también se han implementado módulos y librerías para estos análisis en el software libre Python (www.python.org) y R (www.R-project.org). Este último programa será el que se utilizará en este trabajo (R v3.3.3). A la hora de presentar los resultados también han disminuido las dificultades para crear aplicaciones web interactivas. En este trabajo se implementará el paquete Shiny para presentar los resultados (Chang 2018).

Ambas metodologías de análisis se proponen para resolver el problema planteado por INDITEX. En cada una de las tiendas pertenecientes a este grupo se vende un gran número de prendas, pero el número de prendas devueltas no puede pasar desapercibido. Las prendas devueltas en una tienda se ponen a la vuelta de nuevo en esa tienda, lo que supone un aumento de stock, pudiendo llegar a reunir en una tienda un stock superior al que podría ser vendido y ocasionando que en otra falte mercancía. Centrándonos en las tiendas de la cadena ZARA en España[1] durante el año 2014, se propone descubrir entre que tiendas se mueven las prendas, para tenerlo en cuenta a la hora de enviar stock y poder reducir costes de transporte y almacenaje.

La extracción de los datos se hizo mediante una consulta en lenguaje SQL a una de las bases de datos de INDITEX. Una vez conseguidos los datos necesarios, estos fueron enmascarados por un trabajador de la compañía, ya que, debido a la política de privacidad de la empresa, no es posible trabajar con los datos reales. Las variables con las que contamos son, por un lado, las características propias de cada tienda; nombre, ciudad, provincia, coordenadas,tipo, unidades vendidas y devueltas, importe vendido y devuelto, y por otro, los datos de devoluciones entre cada para de tiendas; tienda origen (tienda donde se com-

[1] Se excluirán las tiendas de los archipiélagos, de Ceuta y de Melilla

pra), tienda destino (tienda donde se devuelve), importe y unidades compradas en la tienda origen y devueltas en la tienda destino.

Para el análisis desde la perspectiva reticular se decidió, por sugerencia de la empresa, considerar las provincias como unidades territoriales y estudiar las devoluciones entre cada una de ellas. Dentro de cada provincia el comportamiento de devolución es similar entre las tiendas, y nos interesa detectar si existe algún patrón espacial que explique las diferencias entre provincias.

Para el análisis de datos en red se descartó las devoluciones dentro de la misma tienda, ya que en estas devoluciones no tendrían consecuencias sobre el stock. Se estudiará el número de unidades devueltas entre cada par de tiendas, se buscará detectar aquellos establecimientos con un elevado número de devoluciones entre ellos y averiguar si existe alguna pauta que pueda ocasionar dicho comportamiento.

A lo largo de este trabajo se explicarán los pasos para realizar un análisis de datos espaciales. En el capítulo 2 veremos el análisis desde la perspectiva reticular y en el capítulo 3 desde el análisis de redes. El último apartado de cada uno de estos capítulos corresponderá con el análisis de los datos de devoluciones. En el capítulo 4 se explicará la aplicación web desarrollada para la presentación de estos resultados mediante la librería Shiny de R. Para finalizar, en el capítulo 5 se enumerarán las conclusiones obtenidas.

Capítulo 2

Análisis de datos reticulares o *lattice*

Aunque existe tres tipos diferentes de datos espaciales, en este trabajo nos centraremos en los datos reticulares, ya que se corresponden con los datos reales de los que disponemos para analizar. Solo veremos la definición de datos geoestadísticos y de procesos puntuales, si se desea profundizar en estos datos se recomienda consultar la literatura existente (Cressie 1993).

En este primer capítulo del trabajo comenzaremos con definiendo los aspectos más básicos de los datos espaciales e iremos avanzando hasta describir por completo la metodología adecuada para su análisis.

2.1. Conceptos básicos y primeras definiciones

Antes de comenzar a explicar el análisis propiamente dicho, debemos comenzar describiendo el objetivo del problema y alguna característica de los datos que queremos analizar.

2.1.1. Objetivo de la estadística espacial

La estadística espacial es el conjunto de métodos apropiados para el análisis de datos con una componente espacial. El objetivo general de cualquier análisis espacial es explotar al máximo la componente espacial para tratar de explicar el comportamiento de la variable de interés.

2.1.2. Tipos de datos espaciales

Todos los datos tienen una componente espacial y temporal (se observan en un determinado lugar y en un momento concreto), pero si dicha componente puede ser de utilidad debería almacenarse junto con el resto de datos.

La localización de los datos debe hacerse en base a un sistema de referencia que, aunque pueden considerarse muchos y muy diversos, lo usual y lo que uti-

lizaremos en este trabajo será utilizar las coordenadas geográficas (longitud y latitud).

Matemáticamente, todo proceso espacial puede caracterizarse mediante un simple proceso estocástico:

$$\{Y(s); \mathbf{s} \in D\}$$

Siendo s un índice que varía en el conjunto $D \in R^d$ y sea $Y(s)$ un valor aleatorio localizado en s.

Como ya comentamos en la introducción podemos encontrar datos espaciales en multitud de situaciones y cada situación presentará sus particularidades. Por ello, los datos espaciales se dividen en tres grandes grupos en función de la naturaleza del espacio D: datos geoestadísticos, datos reticulares y datos puntuales.

Datos geoestadísticos

Los datos geoestadísticos son aquellos que consideran D como un subconjunto fijo de R^d que contiene a un rectángulo d-dimensional con volumen positivo. Es decir, el índice espacial $\mathbf{s}$ varía dentro del subconjunto de R^d de manera continua, considera el dominio como un conjunto continuo en el que puede observarse la variable en cualquier punto.

La geoestadística, como disciplina que estudia los datos geoestadísticos, surgió en los años 60 gracias a Matheron como una mezcla de ingeniería de minas, matemática, geología y estadística (Matheron 1962). Uno de los problemas geoestadísticos más importantes es el de predecir la cantidad de mineral que hay en un bloque a partir de un número de muestras. Matheron definió ese proceso de predicción "kriging" (Matheron 1963). Los métodos geoestadísticos también son usados por científicos que tratan de conocer las propiedades del suelo a partir de un pequeño número de muestras, datos medioambientales, etc.

Datos reticulares

Llamamos datos reticulares a aquellos datos espaciales que consideran D un subconjunto fijo de R^d, formado por una cantidad contable de elementos. También se les conoce con su denominación inglesa, *lattice*.

En muchas ocasiones los datos *lattice* se corresponden con divisiones del espacio total en regiones y cada una de estas divisiones se comporta como una única observación. Estas regiones son polígonos definidos por vértices y lados. Los lados de las regiones se llamarán fronteras. La definición de las regiones no resulta trivial, ya que el resultado final de nuestro estudio puede variar según como las consideremos.

Normalmente la definición de las regiones es inherente al problema que queremos resolver. Si queremos analizar el paro en las comunidades autónomas o la tasa de alfabetización en los países de la Unión Europea, es lógico considerar como regiones las comunidades autónomas o los países. Según la forma que

presenten estas superficies las llamaremos regulares o irregulares. Las regulares serán aquellas que dividen al espacio total de estudio en subregiones idénticas y, normalmente, rectangulares; como la división de nuestro planeta según los cuadrantes formados por los cruces entre paralelos y meridianos. Las irregulares son aquellas que presentan distintas formas y tamaños, suelen corresponder con divisiones de terreno creadas anteriormente para otros fines. Las provincias o las áreas con un mismo código postal son ejemplos de regiones irregulares.

En la Figura 2.1 aparece representado el mapa de Galicia, a la izquierda aparece dividido según los concejos, que serán regiones irregulares y a la derecha se divide según líneas paralelas verticales y horizontales con una separación de 5º.

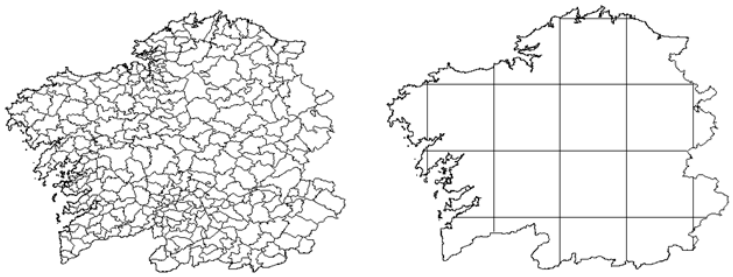

Figura 2.1: Mapa de Galicia dividido por regiones regulares (derecha) y por regiones irregulares (izquierda).

La principal diferencia con los datos geoestadísticos es que estos se pueden encontrar en cualquier punto del espacio de estudio. Para estudiar la concentración de un mineral en la ribera de un río podemos medir la concentración del mineral en cualquier punto. En cambio, si queremos estudiar la prevalencia de cierta enfermedad las observaciones son normalmente agregaciones en distintas regiones y no tendría mucho sentido analizar la prevalencia de cada punto concreto del espacio.

En esta situación existe asociado a D una estructura de vecindad. Ya que los modelos para este tipo de datos suelen considerar que la relación entre zonas vecinas deberá ser mayor que entre otras áreas más alejadas. Hablamos de mayor relación, no de mayor similitud. Aunque de primeras podamos pensar que los valores observados en zonas vecinas deberán ser parecidos, existen situaciones en que ocurre todo lo contrario. Un área donde la variable toma valores altos, puede estar rodeada de zonas de valores bajos (o viceversa).

Patrones puntuales

Cuando D es un proceso puntual en un subconjunto de R^d estamos en la situación de patrones de puntos. El índice espacial es aleatorio y analizarlo será el objetivo del estudio. Muy a menudo la primera pregunta que debe responderse es si el patrón representa aleatoriedad espacial completa, *clusters* o regularidad. Además de la propia variable de localización del evento, podemos considerar

alguna variable adicional que nos proporcione información extra de los eventos.

A partir de aquí centraremos todas las definiciones en el caso de datos reticulares, ya que como hemos explicado anteriormente será en lo que se centre este trabajo.

2.1.3. Autocorrelación espacial

En la estadística básica, y en gran parte de la estadística más avanzada, se asume que las observaciones de una variable se toman bajo condiciones idénticas y de manera independiente. Los datos forman una muestra aleatoria simple, es decir, los datos son independientes e idénticamente distribuidos (i.i.d.). Bajo esta suposición se construye la mayoría de la teoría estadística.

Considerar dependencia en los datos es un gran inconveniente a la hora de trabajar con los modelos usuales. Sin embargo, en muchos casos los modelos que incluyen dependencia son más realistas que los que no lo hacen. La idea de que datos cercanos, en el tiempo o en el espacio, están más correlacionados es natural.

En el contexto espacial, esta falta de independencia recibe el nombre de dependencia o autocorrelación espacial y es una característica que debe ser tratada adecuadamente. La autocorrelación espacial se define como una relación funcional entre lo que ocurre en un punto determinado del espacio y lo que ocurre en otro lugar. Es decir, una variable tendrá autocorrelación espacial cuando el valor observado en un lugar determinado dependa de los valores observados en lugares próximos.

La autocorrelación espacial puede tomar valores negativos, positivos o cero y según en cuál de estas tres situaciones nos encontremos tendremos dependencia negativa, positiva o ausencia de dependencia:

- Autocorrelación espacial negativa: la variable tomará valores diferentes en áreas cercanas. Áreas con alto valor de la variable estarán rodeadas de áreas con valores bajos. Un ejemplo de esto se da en situación de competencia entre plantas por la luz, donde zonas de plantas sanas pueden estar rodeadas de otras con plantas menos fuertes.

- Autocorrelación espacial positiva: la variable tomará valores similares en áreas cercanas. Esta situación representa el efecto contagio, lo que ocurre en un área se "contagia" a zonas próximas. Un área con un valor bajo de la variable estará rodeada de regiones donde la variable también tome valores bajos.

- Autocorrelación espacial nula: en esta situación no existe autocorrelación espacial. Es decir, la variable se distribuye de manera aleatoria en el espacio.

2.1.4. Heterogeneidad espacial

Además del fenómeno de la autocorrelación espacial, otra característica propia de los datos espaciales es la heterogeneidad. Este fenómeno es habitual cuando se trabaja con datos reticulares, y surge debido a las diferencias entre las distintas áreas. Dichas unidades no son homogéneas en determinadas características, lo que hace que ciertos fenómenos no afecten de la misma manera a todas las regiones. Según Anselin la heterogeneidad espacial puede ser definida como la inestabilidad estructural en forma de varianza no constante de los residuos de la regresión o en los coeficientes del modelo, que es posible abordar con herramientas propias del análisis espacial (Anselin 2001).

Las principales causas de la heterocedasticidad en un modelo de regresión son la utilización de datos procedentes de unidades espaciales irregulares, tanto en características propias de la región como en la actitud de la población y algún tipo de especificación errónea del modelo, como la omisión de alguna variable importante. No hay que olvidar que algunas causas que provocan heterogeneidad espacial pueden originar también autocorrelación.

2.1.5. Comparación con las series de tiempo

Los modelos con dependencia temporal, conocidos como series de tiempo, se basan en observaciones idénticamente distribuidas que son dependientes y ocurren en puntos de tiempo equidistantes. Al igual que los espaciales estos datos carecen de independencia ya que los datos próximos en el tiempo estarán más relacionados que los recogidos en tiempos alejados.

Los modelos para series de tiempo están muy estudiados y en un primer momento podemos pensar en utilizarlos sobre datos espaciales. Esto no sería correcto, ya que la correlación temporal y la espacial no son equivalentes. Los datos temporales están ordenados en una única dirección:

$$\text{pasado} \longrightarrow \text{presente} \longrightarrow \text{futuro}.$$

Este flujo unidireccional es esencial de la construcción de modelos temporales, pero pasado, presente y futuro no tienen su equivalente en el espacio, estos datos son multidireccionales. Por ello los modelos espaciales necesitan ser más flexibles. Además, las localizaciones de los datos solo se encuentran regularmente situadas cuando trabajamos con regiones regulares.

Aunque no será objeto de estudio en este trabajo no podemos olvidarnos de situaciones que combinan ambos tipos de dependencia, es el caso de los datos espacio-temporales. Un ejemplo claro de esto es en las predicciones meteorológicas: recogemos la cantidad de lluvia mensual en 30 ciudades durante 5 años. Si consideramos una única ciudad, los datos corresponderán a una serie de tiempo, pero, si estudiamos que ocurre en un determinado momento, el problema se volverá espacial. Por ello, para dar una buena predicción habrá que tener en cuenta ambos tipos de dependencia, que incluso podrían interactuar.

2.1.6. Distancia

En lo que llevamos de trabajo ya hemos hablado en varias ocasiones de localizaciones próximas o cercanas. La cercanía es un concepto relativo que depende de la opinión de la persona que lo define. Para huir de consideraciones subjetivas se hace necesario cuantificar esta proximidad mediante la definición de la distancia entre dos puntos. Desde el punto de vista del análisis matemático una distancia es una aplicación $d : DxD \rightarrow R$ que cumple :

- $d(a, b) \geq 0 \forall a, b \in D$
- $d(a, b) = d(b, a) \forall a, b \in D$
- $\forall a \in D d(a, a) = 0$
- Si $\exists a, b \in D tal\ que\ d(a, b) = 0 \Rightarrow a = b$
- $d(a, b) \leq d(a, c) + d(c, b) \forall a, b, c \in D$

Existen muchas distancias diferentes pero, en este trabajo, siempre que hablemos de distancia nos estaremos refiriendo a la distancia euclidea:

$$d(a, b) := \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

donde $n = 2$, $a = (a_1, a_2)$ y $b = (b_1, b_2)$ serán las coordenadas de cada punto.

2.2. Análisis descriptivo

El primer punto en cualquier análisis de datos es el estudio descriptivo y en el análisis de datos espaciales no iba a ser diferente. Para la descripción de este tipo de datos se utilizan los descriptivos usuales (media, mediana, desviación típica, rango intercuartílico, etc) junto a otros métodos que tienen en cuenta la posición espacial en la que se realizó la medida.

Es común utilizar métodos gráficos como el mapa de colores. Este mapa no es más que una representación del espacio total dividido en las regiones de estudio. Cada región se colorea en un tono en función del valor de la variable observado en dicha región. En la Figura 2.2 vemos un ejemplo de mapa de colores. El número y el rango de valores que representa cada tono se escoge de manera objetiva. Normalmente esta selección se suele hacer de dos maneras; dividiendo el rango en intervalos de igual longitud o utilizando percentiles para que en cada división se encuentren el mismo número de áreas. El principal inconveniente es que el patrón espacial que puede mostrar el mapa depende del criterio de agrupación de los datos.

Cuando la variable que queremos describir es cuantitativa, se suelen utilizar los mapas de gráfico de barras o los mapas de diagrama de sectores. En estos mapas se dibuja en cada región el gráfico de barras o el gráfico de sectores propio. Así, además de ver como se comporta la variable en cada área también podemos ver si es similar en todas las zonas o existe alguna zona con un comportamiento

diferente.

Es muy común encontrar estas representaciones en los medios de comunicación, ya que muestra la información de una manera clara y fácil de entender sin necesidad de conocimientos estadísticos.

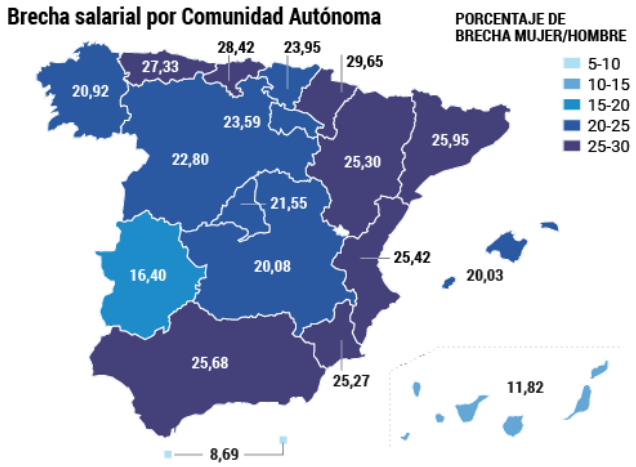

Figura 2.2: Ejemplo de mapa de colores. Fuente: El País .

2.3. Criterio de Vecindad

Como dice la primera ley de la geografía, o principio de autocorrelación espacial, enunciado por el geógrafo Waldo Tobler "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes" (Tobler 1970). Pero ¿qué consideramos cercano? Para dar una respuesta a esta pregunta aparece el concepto de vecindad.

Bajo la perspectiva *lattice* no se considera que todas las regiones influyen sobre el valor que tome la variable en una determinada región, sino, que se trabaja considerando que solo influirán aquellas denominadas vecinas. Definir qué características deben tener dos regiones para que puedan ser consideradas vecinas es una cuestión relevante. Existen varios criterios de vecindad que podemos utilizar y debemos elegir el más apropiado a nuestros datos. Todos los criterios de vecindad deben cumplir que, al seleccionar una región, el resto de ellas queden divididas en dos conjuntos disjuntos, uno compuesto por sus áreas vecinas y otro por las que no lo son.

Una característica importante que tienen los criterios de vecindad es la simetría. Supongamos que A es una región y según el criterio que estamos utilizando la región B es vecina suya. Si el criterio utilizado es simétrico, entonces B también tendrá como vecina a A. Si el criterio no fuera simétrico, B podría no tener a A entre el conjunto de sus regiones vecinas.

En el contexto de datos reticulares en el que estamos trabajando, no podemos olvidar que cada observación representa a una región. Pero para difinir muchos de los criterios de vecindad que veremos a continuación necesitaremos un punto perteneciente a dicha área que la represente. Normalmente dicho punto viene dado por la naturaleza del problema, lo habitual en considerar el centroide cuando estamos trabajando con regiones regulares, y con puntos importantes del área cuando trabajamos con regiones irregulares, como la capital del país, provincia o municipio. Aunque no es lo habitual podemos considerar otros puntos que tengan relevancia en el problema que se estudia, como puede ser el punto con mayor altitud o el más próximo a la costa.

Gráficamente la relación de vecindad se representa mediante una línea que une los dos puntos representativos de cada área vecina.

- **Vecinos por contigüidad.** Este criterio define como áreas vecinas aquellas en las que para ir de una a otra no haya que pasar por una tercera, es decir, que estén contiguas en el mapa. Cuando más de tres áreas coinciden en un único punto se producirán uniones cruzadas. Para controlar si esto puede ocurrir o no se distinguen entre el método Queen y el Rook. Reciben estos nombres porque se asemejan los movimientos de las fichas de ajedrez.

    - La Reina en el ajedrez puede moverse a lo largo de la fila, la columna y las diagonales de la casilla en que se encuentre. Extrapolando esos movimientos a nuestra situación, con este criterio dos áreas serán vecinas si tienen al menos un punto común. En la Figura 2.3 vemos el conjunto total de vecinos del cuadrado y los vecinos de una región determinada, que son aquellos que están en su misma columna, su misma fila o su misma diagonal.

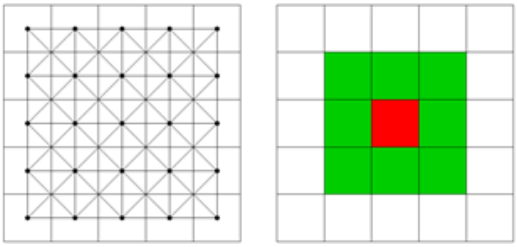

Figura 2.3: Representación del conjunto de vecinos por el criterio de vecindad Queen del cuadrado de lado 5 dividido en cuadrados de lado la unidad (izquierda). Representación de los vecinos (en verde) del cuadrado central (en rojo) según el criterio de contigüidad Queen (derecha) .

    - En el ajedrez, la Torre solo puede moverse a lo largo de la fila y la columna en que se encuentre, no puede moverse en diagonal. Análogamente diremos que dos áreas son vecinas si tienen más de un punto en común. En la Figura 2.4 vemos el conjunto de vecinos total del cuadrado con el criterio Rook. Solo son vecinos de una región aquellos

que se encuentran en su columna o su fila. Con respecto al método Queen se eliminan todos aquellos que están en la diagonal.

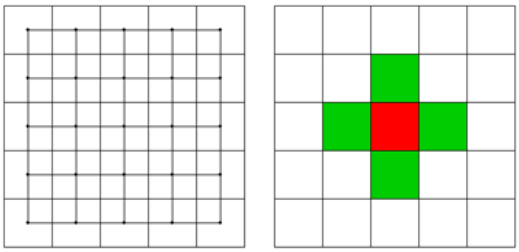

Figura 2.4: Representación del conjunto de vecinos por el criterio de vecindad Rook del cuadrado de lado 5 dividido en cuadrados de lado la unidad (izquierda). Representación de los vecinos en verde del cuadrado central en rojo (derecha) .

Este criterio cumple la condición de simetría ya que si un área tiene un punto o más en común con una segunda, esta segunda también tendrá un punto o más en común con la primera.

- **Vecinos basados en la distancia euclidea**. Este método considera vecinas dos áreas si cumplen cierta propiedad referente a la distancia que las separa. De aquí en adelante nos referiremos a la distancia entre dos regiones como la distancia entre sus puntos representativos. Existen dos variantes:
  - Los k vecinos más próximos. Calcularemos la distancia de una región determinada a todas las demás, y consideraremos vecinos a las k áreas cuya distancia sea menor. Seleccionaremos el número k como sea más apropiado para cada problema. Será una relación asimétrica en la que todas las áreas tendrán el mismo número de vecinos.
  - Considerar como vecinos aquellas áreas que se encuentren a menos de una cierta distancia. Este método funciona bien cuando las áreas tienen una distancia entre ellas similar, ya que si hay una distancia mucho mayor a las otras nos encontraremos con el problema de que dejar esta región sin vecinos, o considerar un número de vecinos demasiado alto en el resto de áreas. Esta relación será simétrica.
- **Vecinos basados en grafos.** El problema de vecindad puede convertirse en un problema de grafos donde los puntos representativos de cada área se denominarán vértices y la unión entre puntos que definirán si son vecinos o no, se llamará arista.

Para los criterios que veremos a continuación es importante tener presente la definición de subgrafo:
Si tenemos un grafo $G = \{V, E\}$ donde V es el conjunto de vértices del grafo y E es el conjunto de aristas del grafo. Diremos que $G' = \{V, E'\}$ es

un subgrafo de G si $V' \subseteq V$ y $E' \subseteq E$.

La triangulación es el uso de la trigonometría para determinar posiciones de puntos, medidas de distancia o áreas de figuras. Consiste en subdividir el área en triángulos, de manera que sus intersecciones sean disjuntas y no contengan a ningún punto del conjunto en su interior. Los métodos de triangulación son muy antiguos, ya se utilizaban en el siglo VI a.C. para medir la altura de las pirámides, y en la actualidad siguen teniendo una importancia notable en los dispositivos GPS. Para nuestro propósito dos puntos conectados por un lado del triángulo serán considerados vecinos.

Veremos a continuación algunos métodos de triangulación que utilizaremos como criterio de vecindad:

- La manera más directa de construir un grafo utilizando métodos de triangulación es mediante la triangulación de Delaunay. Decimos que una triangulación es de Delaunay si y sólo si la circunferencia circunscrita (de centro el punto de corte de las mediatrices) en cada uno de los triángulos no contiene ningún otro vértice en su interior (Delaunay 1934).

Los siguientes criterios se forman a partir de esta triangulación mediante la construción de subgrafos.

- Vecinos en la esfera de influencia: diremos que los puntos x e y son vecinos si las circunferencias C_x y C_y se cortan en dos puntos, siendo C_i la circunferencia de centro i y radio la distancia de i a su vecino más próximo.
- Grafo de Gabriel: dos puntos serán considerados vecinos si en la circunferencia de diámetro la distancia entre ellos no hay ningún otro.
- Grafo de vecindad relativa: dos puntos serán vecinos si la intersección de las circunferencias de centro los puntos y radio la distancia entre ellos no contiene en su interior ningún otro punto. Si dos puntos cumplen esta condición decimos que son vecinos relativos, de ahí el nombre de este grafo.

Los dos primeros métodos garantizaban la simetría, condición que no se cumple en los dos últimos. En la Figura 2.5 vemos los diferentes vecinos del mismo conjunto de puntos según utilicemos los diversos métodos explicados anteriormente. Vemos que utilizando la triangulación de Delaunay como criterio de vecindad obtenemos un número mayor de relaciones de vecindad que si utilizamos los otros tres criterios. Además, el número de vecinos va disminuyendo a medida que pasamos de utilizar como criterio de vecindad la esfera de influencia al grafo de Gabriel y a los vecinos relativos ya que cada uno de estos criterios tiene condiciones más fuertes que el anterior.

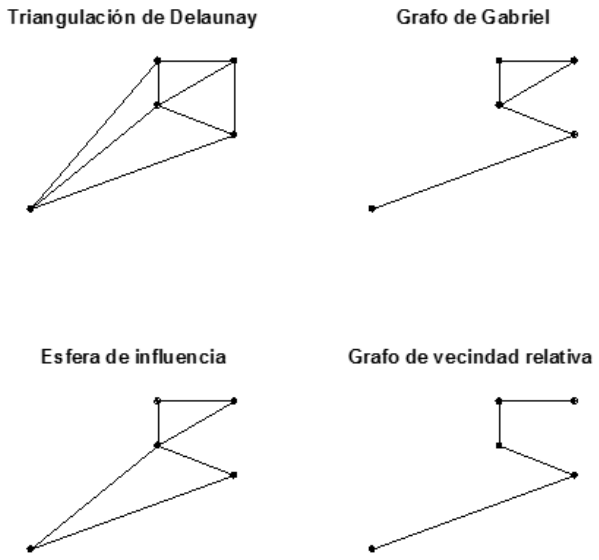

Figura 2.5: Representación del conjunto de vecinos según los criterios de triangulación.

2.4. Pesos espaciales

Una vez que hemos definido el criterio de vecindad que vamos a utilizar, nos puede interesar cuantificar la fuerza de cada unión, esto es lo que conocemos como pesos espaciales. Hasta ahora sabríamos que A tiene dos regiones vecinas B y C, pero lo que desconocemos es si una le influye más que la otra. En esta sección nos encargaremos de explicar los distintos estilos en que pueden definirse los pesos.

Al igual que había que prestar atención a la definición de vecindad, también hay que hacerlo a la hora de definir los pesos, ya que en este caso hacer una consideración u otra puede modificar la solución final.

Los pesos se representan de forma matricial mediante la matriz cuadrada W. Donde cada elemento w_{ij} representa el peso de la relación de vecindad entre las regiones i y j. Cuando $w_{ij} = 0$ las regiones no son vecinas. La diagonal de la matriz será 0 ya que, por convenio, una región no puede ser vecina de ella misma. W es una matriz cuadrada con todos sus elementos mayores o iguales a 0.

Esta matriz será simétrica si el criterio utilizado para definir los vecinos y los pesos lo son. Decimos que los pesos son simétricos cuando una región A ejerce sobre B la misma influencia que B sobre A. Una situación en la que tiene sentido utilizar una relación asimétrica es considerar la influencia de las ciudades grandes sobre los pueblos de alrededor. Las grandes ciudades influyen más en las características de los pueblos que a la inversa.

Dentro de todos los posibles estilos para asignar los pesos, podemos distin-

guir dos grandes grupos; aquellos donde por el mero hecho de ser vecinos cada unión tenga un peso común y aquella en la que la importancia de las uniones variará de unas uniones a otras en base a ciertas características.

- **Estilo binario (B)**. Es la manera más sencilla de trabajar. Asume que $w_{ij} = 1$ cuando i y j son regiones vecinas y $w_{ij} = 0$ cuando no lo sean. Este método es el más utilizado cuando tenemos poca información del proceso espacial. Con este estilo la suma de los pesos de un área es el número de vecinos que tiene.
- **Estandarización por filas (W)**. Este método se basa en que los pesos de cada fila de la matriz sumen 1. Para ello se divide la unidad entre el número de áreas vecinas del área en cuestión. Según este método los pesos de áreas con pocos vecinos serán mayores que los de áreas con un número de vecinos mayor. Es decir, cada vecino de un área con pocos vecinos ejerce gran influencia sobre ella, mientras que los vecino de áreas con muchos vecinos ejercen menor influencia.
- Existen otros dos estilos que consideran el mismo peso para todos los enlaces. El primero de ellos **(C)** define el peso entre dos regiones vecinas como el cociente entre el número total de regiones y el número de enlaces. La suma de todos los pesos será el número de regiones. El segundo **(U)** define el peso como el cociente entre la unidad y el número de enlaces total. Con este estilo la suma de todos los pesos será la unidad.
- **Estabilización de varianza (S).** Este estilo fue propuesto por Tiefelssdorf . Con los estilos vistos anteriormente los pesos de áreas con muchos vecinos varían mucho de utilizar un estilo a otro. Lo que busca este estilo es reducir esta variación. Los pesos variarán menos que con el estilo W. Será siempre asimétrico, pero, al igual que ocurre con el estilo W, si el conjunto de vecinos es simétrico la matriz W estará bastante cerca de ser simétrica (Tiefelsdorf 1999).

Los estilos vistos anteriormente solo tienen en cuenta si las regiones son vecinas o no lo son. En 1981, Cliff y Ord consideraron que la fuerza de la relación de vecindad disminuye con la distancia. Es decir, la distancia entre las regiones vecinas tiene que verse reflejada en los pesos (Cliff y Ord 1981). Entonces se propone el método de los pesos espaciales generales, donde no solo importa si las regiones son vecinas o no, sino que también su distancia. Para ello toma el estilo binario y multiplica los pesos (recordamos que con este estilo todos los pesos son 1) por

$$\frac{1000}{dist(i,j)}$$

Esta condición es bastante lógica, ya que el peso será mayor cuanto menor sea la distancia entre ellas. Este método tiene en cuenta las distancias, pero podrían considerarse otras como el número de habitantes, la superficie o el PIB.

Una vez que hemos seleccionado los pesos que daremos a las uniones de vecindad ya podemos comenzar con el análisis espacial. Para ello comenzaremos con determinar si existe autocorrelación o no.

2.5. Autocorrelación

Ya hemos visto que la autocorrelación espacial se define como la relación entre el valor que toma una variable en una región y la que toma en regiones vecinas. Detectar la presencia de autocorrelación es esencial, ya que utilizar un modelo clásico supondría incurrir en errores importantes como la subestimación de la variabilidad, una estimación de los coeficientes no eficiente y un coeficiente R^2 sobreestimado.

El estudio de la autocorrelación espacial se hace desde dos perspectivas, global y local. La información proporcionada por ambas perspectivas es complementaria y debemos analizar ambas para obtener la mayor información posible de la variable en estudio.

A continuación, se verán los test diseñados para este fin junto con métodos gráficos que pueden ser de utilidad.

2.5.1. Autocorrelación global

El estudio de autocorrelación espacial global busca detectar la presencia de tendencias o estructuras espaciales generales en la distribución de la variable sobre el espacio geográfico completo. Y saber si la variable se distribuye de forma independiente o, si por el contrario, existe algún tipo de asociación entre regiones vecinas.

Para ello de desarrollaron una serie de tests: I de Moran , C de Geary , τ de Mantel y G(d) de Getis y Ord (Moran 1948;Geary 1954; Mantel 1967; Getis 1992). De todos ellos el más usado y el que veremos con detenimiento es el test I de Moran.

El estadístico I de Moran se define como:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde w_{ij} es el elemento (i, j) de la matriz de pesos W, $\bar{y}$ es la media de la variable Y y n el número total de áreas en las que trabajamos.

El valor esperado de este estadístico bajo la hipótesis de ausencia de autocorrelación es $E(I) = -\frac{1}{n-1}$. Este valor es siempre negativo y solo dependerá del número de áreas. I no es un estadístico de interpretación sencilla y para facilitar esta tarea se suele estandarizar,

$$z = \frac{I - E(I)}{\sqrt{var(I)}},$$

que asintóticamente sigue una distribución Normal de media 0 y desviación 1.

Además de conocer valor I, es necesario saber si dicho valor es lo suficientemente grande o pequeño para concluir que existe autocorrelación global. Para ello se lleva a cabo un contraste de hipótesis el cual tiene como hipótesis nula

la igualdad entre el valor I obtenido y el valor esperado bajo la ausencia de correlación $E(I)$ y la hipótesis alternativa será la diferencia entre estos dos valores.

En función del valor estandarizado z y de la significación obtenida, podremos hacer las siguientes interpretaciones:

- Si no se obtuvo significación estadística concluiremos que no podemos rechazar la hipótesis nula de ausencia de autocorrelación.

- Si obtuvimos significación y además $z > 0$, tendríamos autocorrelación positiva. Valores de la variable Y parecidos estarían espacialmente más agrupados que como lo estarían si la distribución tuviera independencia espacial.

- Si obtuvimos significación y, además, $z < 0$, tendríamos autocorrelación espacial negativa. Es decir, una agrupación de valores similares inferior a lo esperado si la distribución fuera independiente.

2.5.2. Autocorrelación local

Con el estudio global habremos detectado o no la presencia de una asociación espacial en todo el territorio, pero no seremos capaces de saber si el esquema detectado se mantiene a nivel local. En el estudio local se calcula un valor para cada unidad espacial, lo que permite analizar el grado de dependencia individual de cada unidad respecto a las demás. Se distinguen dos variantes en función de la existencia o no de autocorrelación global.

La autocorrelación local nos ayuda a detectar "clusters" espaciales, es decir, zonas con una similitud (o diferencia) entre observaciones mayor a la esperada si la distribución fuera completamente independiente. Pueden existir zonas con autocorrelación local aunque no se tenga autocorrelación global, ya que esta tendencia puede no apreciarse en toda la superficie.

Además, cuando tengamos autocorrelación global puede ocurrir que no todas las regiones contribuyan con igual peso al indicador global. Pueden coexisten zonas en las que la variable se distribuye de manera aleatoria junto a otras con importante contribución a la dependencia global existente. Los contrastes globales asumen homogeneidad espacial y bajo el enfoque local podemos ver si es cierto que la tendencia sea homogénea en todo el espacio o si hay puntos atípicos. Estos puntos atípicos serán aquellas regiones con una participación en el estadístico global superior a la media.

Para el estudio de la autocorrelación local se utilizan los indicadores locales de asociación espacial llamados LISA (*Local Indicators of Spatial Association*) (Anselin 1995). Dentro de los contrastes LISA destacan I_i de Moran, C_i de Geary y τ_i de Mantel. Como anteriormente nos centraremos en el test I_i de Moran.

Para cada área i del espacio de estudio definiremos el estadístico:

$$I_i = (y_i - \bar{y}) \sum_{j=1}^{N_i} w_{ij}(y_j - \bar{y})$$

donde w_{ij} es el elemento (i, j) de la matriz de pesos W, $\bar{y}$ es la media de la variable Y y N_i es el conjunto de vecinos de i.

La interpretación del estadístico dependerá de la situación en que nos encontremos.

Si queremos ver la presencia de "clusters" debemos estandarizar la variable I_i que será interpretada de la misma manera que el estadístico global: si no es significativo diremos que en la región i no existe autocorrelación, si es significativo y mayor que 0 habremos encontrado un cluster de valores similares y si es negativo tendremos un cluster de valores diferentes.

Si por el contrario estamos interesados en detectar atípicos, consideraremos la media I_i que será igual a la del estadístico I por un factor de proporcionalidad. Las máximas contribuciones de los valores I_i al estadístico global I pueden identificarse de manera sencilla mediante un diagrama de cajas.

2.5.3. Métodos gráficos de autocorrelación

Normalmente el análisis de autocorrelación se acompaña con Diagrama de Dispersión de Moran (Anselin 1993). Este diagrama es una herramienta visual muy útil que nos proporciona en un único gráfico la información de los dos test de autocorrelación. Nos permite valorar la similitud del valor observado con sus observaciones vecinas.

En el eje X se representan los valores de la variable y en el eje Y se representa el retardo espacial correspondiente. El retardo espacial será el promedio ponderado de los valores que adopta la variable Y en el conjunto de las observaciones vecinas a una dada.

También incluye la recta de regresión lineal del retardo espacial frente a los valores de la variable, cuya pendiente coincide con el índice global de Moran. Cuando mayor sea esta inclinación más fuerte será el grado de autocorrelación espacial global.

Además de la información global, este diagrama también nos proporciona información a nivel local. Cada cuadrante de dicho gráfico tendrá un significado:

- Los puntos que caigan en los cuadrantes I y III tendrán correlación positiva. En el I estarán las regiones con observaciones superiores a la media y en el III regiones con observaciones inferiores.
- Los puntos que caigan en los cuadrantes II y IV representarán autocorrelación negativa. En el II estarán puntos con valores bajos de la variable

rodeados de regiones con valores altos y en el IV se representará la situación contraria.

También nos permite detectar regiones atípicas, que serán aquellas que se encuentren a una distancia del origen superior a dos unidades, a este criterio se le conoce como 2-sigma.

En la Figura 2.6 se muestra un diagrama de Moran con el que podemos obtener cierta información sobre los datos representados. A nivel local predomina la autocorrelación positiva ya que la mayoría de los puntos caen en los cuadrantes I y III. También sabemos que la autocorrelación espacial global no será muy elevada porque la recta de regresión está próxima a la horizontal. El diagrama nos muestra el valor de esta pendiente, que coincide con el valor del índice global de Moran y en este caso es de 0.29.

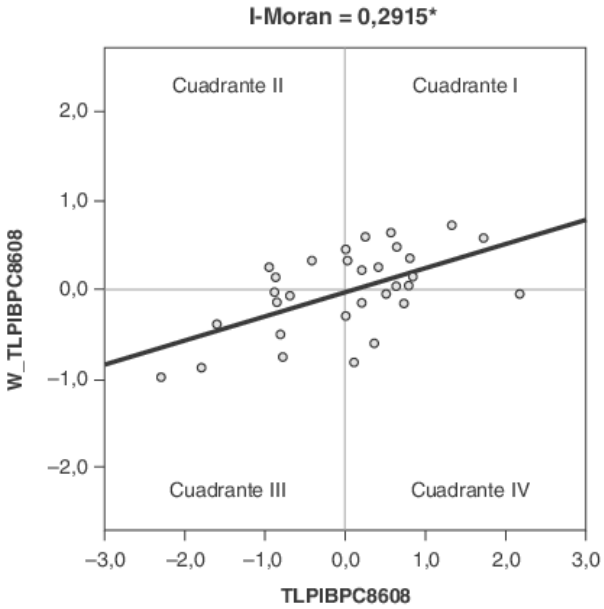

Figura 2.6: Fuente: Crecimiento económico, convergencia y concentración económica espacial en las entidades federativas de México 1970-2008 (Sanen y Quintana 2010).

2.6. Modelización de los datos

Hemos visto que la falta de independencia entre observaciones en datos espaciales es un fenómeno muy frecuente. Si existe independencia el análisis es más sencillo, pero a la hora de realizar predicciones suele ser preferible tenerla en cuenta, ya que obtendremos resultados más fieles a la realidad. En esta sección veremos como la dependencia entre las observaciones puede modelarse para datos reticulares siguiendo la literatura existente (Schabenberger y Gotway 2005).

En este apartado vamos a describir dos enfoques distintos para modelar la autocorrelación en los modelos de regresión espaciales; el modelo autorregresivo simultáneo (SAR) y el modelo autorregresivo condicional (CAR). Estos modelos nos permiten incorporar la estructura de vecindad utilizada cuando trabajamos con este tipo de datos. La principal diferencia que presentan estos dos modelos es que los condicionales se utilizan cuando existe dependencia entre datos cercanos mientras que los modelos SAR tienen cuenta la dependencia de mayor alcance.

2.6.1. Modelo autorregresivo simultanéo (SAR)

Partimos de la idea de un modelo lineal con datos normales, y consideramos que el error tiene autorregresión espacial. Entonces llegamos al modelo

$$Y = X\beta + e, \text{donde } e = Be + u$$

En dicho modelo, la matriz B contiene los parámetros de dependencia espacial. El vector u será el vector de errores de la autorregresión que tendrá de media 0 y de matriz de covarianzas $\Sigma_u = diag[\sigma_1^2, ..., \sigma_n^2]$.

Si $b_{ij} = 0 \; \forall i, j$, no habría autorregresión y el modelo se reduciría a un modelo tradicional de regresión lineal con errores independientes.
Este modelo también se puede escribir como

$$u = (I - B)(Y - X\beta)$$

y de esta expresión se deduce que la varianza de la variable dependiente Y es

$$\Sigma_{SAR} = Var(Y) = (I - B)^{-1}\Sigma_u(I - B^{'})^{-1}$$

siempre que $(I - B)$ sea invertible.

A este modelo le conocemos como modelo autorregresivo simultáneo (SAR) y el término autorregresivo describe las sucesivas autorregresiones que ocurren al mismo tiempo en cada región.

Estimación e inferencia bajo el modelo SAR

Una vez que hemos definido el modelo pasaremos a estimar sus parámetros y a hacer inferencia sobre ellos. Suponiendo que nos encontramos en la situación de normalidad:

$$Y \sim N(X\beta, (I - B)^{-1}\Sigma_u(I - B^{'})^{-1})$$

Podemos reparametrizar la matriz de covarianzas como $\Sigma_u = \sigma^2 V_u$ para considerar una estructura más general. Así podríamos escribir la matriz del modelo SAR como

$$\Sigma_{SAR} = \sigma^2(I - B)^{-1}V_u(I - B^{'})^{-1} = \sigma^2 V_{SAR}(\theta)$$

donde θ contiene los parámetros de dependencia espacial b_{ij} y los parámetros de V_u.

Bajo este modelo los parámetros σ^2 y β pueden estimarse a partir del logaritmo de la función de verosimilitud

$$\varphi(\beta; \theta; Y) = ln(|\Sigma(\theta)|) + nln(2\pi) + (Y - X\beta)^{'}\Sigma(\theta)^{-1}(Y - X\beta).$$

Minimizando esta ecuación se obtienen los estimadores de los parámetros:

$$\hat{\beta}_{ml} = (X^{'}\Sigma_{SAR}(\hat{\theta}_{ml})^{-1}X)^{-1}X^{'}\Sigma_{SAR}(\hat{\theta}_{ml})Y$$

$$\sigma^2_{ml} = \frac{(Y - X\hat{\beta}_{ml})^{'}\Sigma_{SAR}(\hat{\theta}_{ml})^{-1}(Y - X\hat{\beta}_{ml})}{n}$$

donde $\hat{\theta}_{ml}$ es el estimador de máxima verosimilitud de θ.

Estimar $\hat{\theta}$ no es tarea sencilla y en la mayoría de los casos es necesario emplear métodos numéricos. Por ello en la práctica el modelo más utilizado es el modelo SAR de un parámetro. Este modelo supone que $\Sigma_u = \sigma^2 I$. Para facilitar más la estimación también es común reparametrizar $B = \rho W$. Donde W es la matriz de pesos espaciales y ρ es el parámetro que representa la autocorrelación espacial. El modelo SAR se escribirá de la siguiente manera:

$$Y = X\beta + e, \quad e = \rho We + u.$$

Estas dos ecuaciones se pueden combinar dando lugar a las siguientes expresiones:

$$Y = X\beta + (I - \rho W)^{-1}u = X\beta + \rho WY - \rho WX\beta + u$$

Podemos ver como la autoregresión induce a la autocorrelación espacial en la regresión lineal a través del término $(I - \rho W)^{-1}u$. En la última ecuación vemos dos términos, $\rho WX\beta$ y ρWY, que no aparecen en un análisis de regresión con los errores independientes. Estos términos representan los retardos espaciales.

La matriz de covarianzas correspondiente a este modelo sería

$$\Sigma_{SAR} = \sigma^2(I - \rho W)^{-1}(I - \rho W')^{-1}$$

Tanto en la nueva manera de escribir la ecuación del modelo como en la matriz del modelo aparece el término $(I - \rho W)^{-1}$. Para que el modelo este bien definido es necesario que la matriz $(I - \rho W)$ sea invertible.

Si conocemos el valor de ρ y suponemos normalidad, las estimaciones de β y σ^2 se podrían calcular por máxima verosimilitud siendo

$$\hat{\beta} = (X'\Sigma^{-1}_{SAR}X)^{-1}X'\Sigma^{-1}_{SAR}Y$$

$$\hat{\sigma^2} = \frac{(Y - X\hat{\beta})'\Sigma^{-1}_{SAR}(Y - X\hat{\beta})}{n - k}$$

Lógicamente el valor de ρ es desconocido y será necesario estimarlo. Además Y y u no son independientes, por lo que usar la estimación de mínimos cuadrados ordinarios no sería un método consistente para estimar ρ.
Ord planteó el siguiente estimador de ρ (Ord 1975), calculado con el método de mínimos cuadrados generalizado:

$$\hat{\rho} = \frac{Y'W'WY}{Y'W'W^2Y}$$

Así estimador ρ es consistente pero no es eficiente. Una vez que conocemos $\hat{\rho}$ solo tenemos que utilizarlo para estimar β, σ^2 y Σ_{SAR}.

2.6.2. Modelos autorregresivos condicionales (CAR)

El enfoque simultáneo visto anteriormente nos proporciona un modelo multivariante para describir las interacciones espaciales entre los datos. Puede resultar más intuitivo trabajar con un enfoque análogo al análisis de series de tiempo y especificar modelos para las distribuciones de probabilidad de cada observación Y_i condicionada a los valores observados de las demás observaciones. Es decir, modelaremos $f(Y_i|Y_{-i})$, donde Y_{-i} es el vector de todas las observaciones menos la correspondiente a la región i $\forall i$.

En el contexto de las series de tiempo decimos que las variables aleatorias $Y_1, ..., Y_t$ cumplen la propiedad de Markov cuando $f(Y_{t+1}|Y_t, ..., Y_1) = f(Y_{t+1}|Y_t)$, es decir, que el valor en el tiempo $t+1$ solo depende de lo ocurrido en el tiempo inmediatamente anterior t. Una secuencia de variables aleatorias con la propiedad de Markov es un proceso de Markov.

Si trasladamos esta idea a los datos espaciales, diremos que el valor Y_i dependerá solamente de lo que ocurra en sus vecinos. Diremos que Y_i depende de Y_j solamente si la localización j pertenece al conjunto de vecinos de i, N_i. Cuando esto se cumple diremos que el proceso Y es un campo aleatorio de Markov.

Así, con el enfoque autorregresivo condicional construiremos modelos para $f(Y_i|Y_j, j \in N_i)$. Si suponemos que cada una de esas distribuciones condicionadas es normal, podríamos modelarlas usando que :

$$E(Y_i|Y_{-i}) = x_i'\beta + \sum_{j=1}^{n} c_{ij}(Y_j - x_i'\beta)$$

$$Var(Y_i|Y_{-i}) = \sigma_i^2, i = 1, ..., n$$

donde c_{ij} son los parámetros de dependencia espacial que toman un valor distinto a 0 solo si $j \in N_i$.

Estimación e inferencia del modelo CAR

Para estimar los parámetros del modelo y hacer inferencia sobre ellos, necesitamos asegurar que existe la distribución conjunta. El teorema de Hammserley-Clifford (Besag 1974) describe las condiciones necesarias para definir una distribución conjunta $f(Y_1, ..., Y_s)$ a partir de un conjunto de distribuciones condicionadas $f(Y_i|Y_j, j \in N_i)$. Si asumimos que las distribuciones condicionadas son normales, con media y varianza las vistas anteriormente, nos encontramos bajo las condiciones requeridas por el teorema de Hammersley-Clifford y puede demostrarse que dichas distribuciones condicionadas generan una distribución conjunta válida que sigue una normal multivariante con media $X\beta$ y varianza

$$\Sigma_{CAR} = (I - C)^{-1}\Sigma_c$$

donde $\Sigma_c = diag[\sigma_1^2, ..., \sigma_n^2]$. Para asegurarnos de que la matriz de covarianzas sea simétrica se impone la condición $\sigma_j^2 c_{ij} = \sigma_i^2 c_{ji}$.

El modelo condicional es similar al modelo SAR, pero con distinta matriz de varianzas-covarianzas. De hecho, si tenemos $\Sigma_c = \sigma^2 I$ y $\Sigma_u = \sigma^2 I$, cualquier

modelo SAR con matriz de dependencia espacial B podría ser expresado como un modelo CAR con matriz de dependencia espacial $C = B + B' - BB'$.

Al igual que con el modelo SAR, consideraremos la situación con un único parámetro $\Sigma_c = \sigma^2 I$ y los parámetros de dependencia espacial podrán ser escritos en función de un único parámetro de autocorrelación espacial, es decir, $C = \rho W$.

Al contrario de lo que ocurre en el modelo SAR, el estimador de mínimos cuadrados para el parámetro ρ es consistente. Iterando el método de mínimos cuadrados ponderados podemos estimar todos los parámetros del modelo CAR.

$$\hat{\rho} = \frac{\hat{\epsilon}' W \hat{\epsilon}}{\hat{\epsilon}' W^2 \hat{\epsilon}}$$

donde $\hat{\epsilon}$ es el vector de residuos de la regresión por mínimos cuadrados ordinarios.

Para estimar la matriz de covarianzas, podemos reparametrizarla como $\Sigma_c = \sigma^2 V_c$ con V_c conocido. Considerando la reparametrización de C, podemos escribir la matriz del modelo CAR como:

$$\Sigma_{CAR} = \sigma^2 (I - C)^{-1} V_c = \sigma^2 V_{CAR}(\rho).$$

2.6.3. Validación de los modelos

El modelo de SAR puede plantearse como un modelo lineal con errores espaciales autocorrelados. Bajo esta perspectiva todos los test de hipótesis sobre sus parámetros β y θ pueden resolverse con las pruebas utilizados en los modelos lineales, como el test de Wald. Sin embargo, uno de los usos más comunes del modelo SAR de un parámetro es proporcionar una prueba alternativa para comprobar la autocorrelación espacial en los residuos de un modelo de mínimos cuadrados ordinarios.

Para ello supondremos un modelos lineal con errores independientes

$$Y = X\beta + e, \Sigma = \sigma^2 I$$

y lo compararemos con el modelo SAR de un parámetro

$$Y = X\beta + (I - \rho W)^{-1} u, \Sigma = \sigma^2 (I - \rho W)^{-1} (I - \rho W')^{-1}.$$

Si $\rho = 0$ ambos modelos serían equivalentes. Por lo que contrastar la igualdad de estos dos modelos es equivalente a contrastar que $\rho = 0$.
Para ello comparamos los parámetros θ de ambos modelos; el lineal $\theta_1 = [\beta', \sigma^2]'$ y el autorregresivo $\theta_2 = [\beta', \sigma^2, \rho]'$. Las hipótesis a contrastar serán $H_0 : \theta = \theta_1$ frente a $H_1 : \theta = \theta_2$. El estadístico utilizado en este contraste será

$$\varphi(\beta, \theta_1, Y) - \varphi(\beta, \theta_2, Y),$$

donde φ representa menos dos veces el logaritmo de la función de verosimilitud. Cuando el número de datos sea lo suficientemente grande, podremos suponer que este estadístico sigue una distribución χ^2 con 1 grado de libertad

$(dim(\theta_2) - dim(\theta_1)) = 1$. El tamaño necesario para que la aproximación sea adecuada depende de distintos factores, incluida la estructura de la matriz W.

La principal diferencia entre los modelos CAR y SAR (de un parámetro) es la definición de Σ. $\Sigma_{CAR} = \sigma^2(I-\rho W)^{-1}$ y $\Sigma_{SAR} = \sigma^2(I-\rho W)^{-1}(I-\rho W')^{-1}$. Así que el método visto anteriormente también se podrá aplicar al modelo CAR sin más que considerar su matriz de covarianzas correspondiente.

Cuando tengamos más de un modelo autorregresivo nos nteresará saber cuál es el que mejor se ajusta a nuestros datos. Para ello se utilizan criterios de información basados en la verosimilitud. El más utilizado es el Criterio de Información de Akaike (AIC)(Akaike 1974):

$$AIC = \varphi(\beta, \theta, Y) + 2(k+q),$$

donde k es la dimensión de Y y q es el número de parámetros estimados. Generalmente el mejor modelo será aquel que tenga un AIC más pequeño. Alternativamente, si queremos seleccionar modelos más sencillos se puede utilizar el criterio de información Bayesiana (BIC).

2.7. Aplicación a datos reales.

En esta sección vamos a tratar de resolver un problema real aplicando los métodos explicados anteriormente. El objetivo será ver las diferencias entre provincias en el porcentaje de devolución, definido como el cociente entre las unidades devueltas y las unidades compradas en cada provincia.

Con esto pretendemos saber qué provincias son las que reciben mayor número de prendas en función del número de artículos que venden. Además, el estudio por provincias nos interesa para detectar comportamientos diferentes entre zonas de la península. Puede ocurrir que por las características propias de cada región, haya diferencias entre ellos. Mientras que dentro de cada una de ellas el comportamiento de cada tienda sea más homogéneo.

Para conseguir este objetivo podemos abordar el problema desde el punto de vista de datos *lattice*. Consideraremos cada provincia como una región y su punto representativo será su capital.

Comenzamos describiendo la variable porcentaje de devolución, de media se devuelve casi el 14 % de lo comprado y existe una diferencia de 10 puntos porcentuales entre el porcentaje de devolución más alto y el más bajo, lo que nos hace sospechar que habrá diferencias entre provincias. En la Figura 2.7 vemos es histograma de esta variable.

```
library(openxlsx)
Datos_unidades <- readWorkbook("Tasa_provincia.xlsx")
summary(Datos_unidades[,5])
hist(Datos_unidades[,5])
```

```
##  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  6.69   12.57   14.10   13.72   15.43   18.76
```

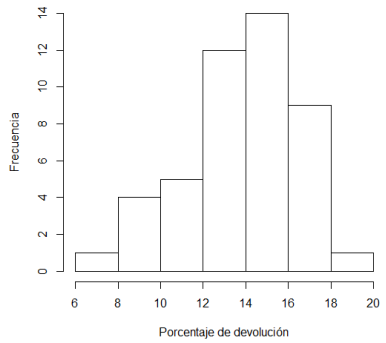

Figura 2.7: Histograma de la variable porcentaje de devolución.

Recurrimos al mapa de colores para añadir la información espacial a los descriptivos y hacernos una idea inicial de la distribución de estos porcentajes en el espacio.

La Figura 2.8 muestra el mapa de colores del porcentaje de devolución, a la izquierda dividiendo el intervalo en 4 trozos iguales y a la derecha dividiendo el rango mediante los cuantiles. En ambos mapas vemos que la mayor parte de las provincias tienen un porcentaje de devolución entre el 13 y el 16 % y que el noreste peninsular parece ser una zona con bajo porcentaje de devolución. Esta situación de baja devolución parece extenderse hacia el centro peninsular si dividimos el rango por cuantiles. De manera opuesta las provincias del sur y noroeste peninsular tienen un alto porcentaje de devolución. Podemos destacar el comportamiento de Valencia que a pesar de tener un porcentaje de devolución bajo se encuentra rodeada de provincias con una tasa más elevada.

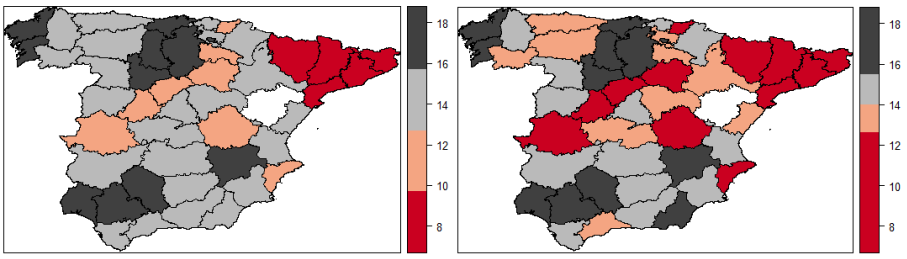

Figura 2.8: Mapa de colores del porcentaje de devolución en las provincias de España. A la izquierda dividiendo el rango en 4 intervalos de igual longitud, y a la derecha dividiéndolo mediante cuantiles.

A la hora de dividir el rango se eligió hacerlo en 4 intervalos ya que parece un número acorde con la cantidad de provincias. Si añadimos más colores el mapa se vuelve menos intuitivo (Figura 2.9).

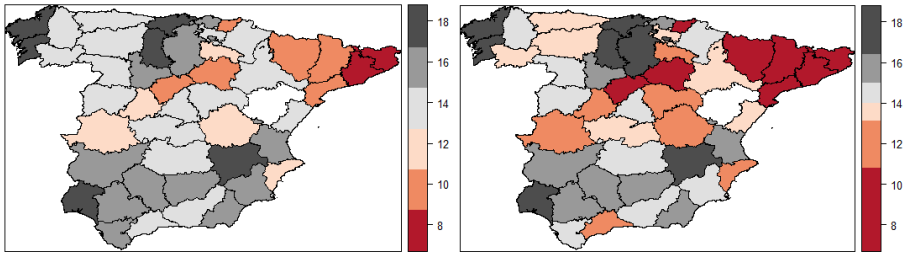

Figura 2.9: Mapa de colores del porcentaje de devolución en las provincias de España. A la izquierda dividiendo el rango en 6 intervalos de igual longitud, y a la derecha dividiéndolo mediante cuantiles.

Las conclusiones obtenidas a partir de estas gráficas son subjetivas, no solo dependen de la manera de representarla, sino, también de la opinión de quién lo interpreta.

Para tener datos objetivos debemos utilizar los test estadísticos explicados teóricamente a lo largo de este capítulo. Pero antes de esto debemos definir los criterios de vecindad y los pesos que utilizaremos. Comenzaremos aplicando los diferentes criterios de vecindad a las provincias y los representaremos gráficamente. Seleccionaremos algunos de estos criterios y les aplicaremos distintos pesos que describiremos para ver sus semejanzas y diferencias. Para ello se utilizará el paquete *spdep* (Bivand 2015).

Los primeros modelos que vimos en teoría son los que llamábamos vecinos por contigüidad y distinguíamos entre el método Rook y el Queen. En nuestra situación ambos métodos son equivalentes, ya que no existe ningún par de provincias cuya frontera sea un único punto. En la Figura 2.10 vemos la representación de este criterio. Lógicamente todas las provincias tienen al menos un vecino.

```
library(spdep)
esp_nb<- poly2nb(esp_pro, row.names=esp_pro@data$NAME_2)
plot(esp_pro, border="grey60", axes=FALSE, main= "Criterio de
 contigüidad ")
```

A continuación, describíamos los métodos de triangulación. La triangulación de Delaunay es la más permisiva de todas, permitiendo considerar vecinos áreas tan alejadas como Pontevedra y Huelva. Los criterios se van volviendo más restrictivos a medida que pasamos a la esfera de influencia, al Grafo de Gabriel y al Grafo de vecindad relativa, que es el más restrictivo. Esta evolución se aprecia en la Figura 2.11.

```
coords <- coordinates(esp_pro)
IDs <- row.names(as(esp_pro, "data.frame"))
esp4_nb <- tri2nb(coords, row.names= IDs))
esp5_nb <- graph2nb(soi.graph(esp4_nb,coords), row.names= IDs)
```

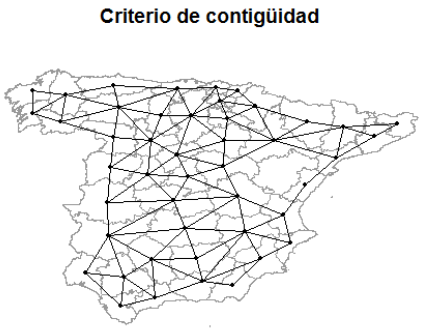

Figura 2.10: Representación de los vecinos de cada provincia de España con el criterio de contigüidad.

```
esp6_nb <- graph2nb(gabrielneigh(esp4_nb,coords), row.names= IDs)
esp7_nb <- graph2nb(relativeneigh(esp4_nb,coords), row.names= IDs)
```

La representación de los k vecinos más próximos aparece en la Figura 2.12, dónde consideramos $k = 1, 2$ y 4. Si $k = 1$ tendremos 13 subgrafos disjuntos, cuando $k = 2$ el número de subgrafos descenderá hasta 2 y cuando $k = 4$ solo habrá 1.

```
esp8_nb <- knn2nb(knearneigh(coords, k=1), row.names=IDs)
esp9_nb <- knn2nb(knearneigh(coords, k=2), row.names=IDs)
esp10_nb <- knn2nb(knearneigh(coords, k=4), row.names=IDs)
```

Por último, definíamos el método basado en las distancias. Calculamos el máximo de la distancia euclidea entre cada par de provincias vecinas según el criterio Queen, que corresponde a la distancia entre Valencia y Albacete (nos referiremos a esta distancia como M), y consideraremos vecinos aquellas provincias cuya distancia entre ellas sea menor que 0.75*M (0.75M), M (1M) y 1.5*M (1.5M). Esta situación está representada en la Figura 2.13.

```
dsts <-  unlist(nbdists(esp_nb, coords))
max_1nn <- max(dsts)
esp11_nb <- dnearneigh(coords,d1=0,d2=0.75*max_1nn,row.names=IDs)
esp12_nb <- dnearneigh(coords,d1=0,d2=1*max_1nn,row.names=IDs)
esp13_nb <- dnearneigh(coords,d1=0,d2=1.5*max_1nn,row.names=IDs)
```

En la Tabla 2.1 aparece un pequeño resumen descriptivo de los diferentes criterios de vecindad aplicados a las provincias españolas. En ella se muestra el número medio de vecinos que tiene cada provincia, el número de provincias que no tienen vecinos, las provincias con mayor y menor número de vecinos y si el criterio es simétrico o no.

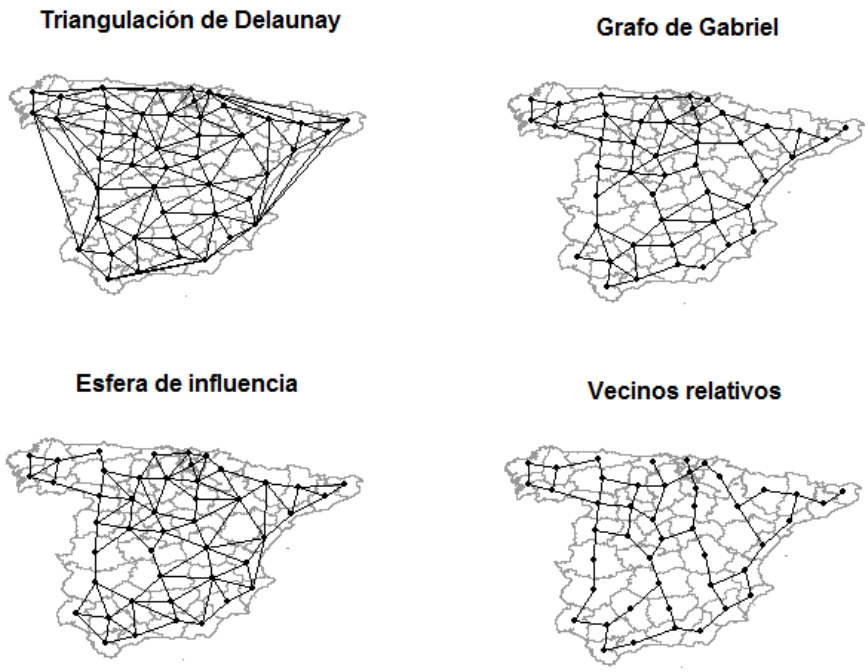

Figura 2.11: Representación de los vecinos de cada provincia de España con los diferentes métodos de triangulación.

Criterio de vecindad	Vecinos medios por provincia	Provincias sin vecinos	Provincias con menos vecinos	Provincias con más vecinos	Criterio Simétrico
Contigüidad	4.56	0	4 Provincias con 2 vecinos	1 Provincia con 8 vecinos	Sí
Triangulación Delanay	5.43	0	2 Provincias con 3 vecinos	2 Provincia con 8 vecinos	Sí
Esfera de influencia	4.35	0	3 Provincias con 2 vecinos	4 Provincia con 7 vecinos	Sí
Gabriel	1.85	8	13 Provincias con 1 vecinos	4 Provincia con 5 vecinos	No
Relativos	1.26	11	15 Provincias con 1 vecinos	3 Provincia con 3 vecinos	No
K=1 vecinos más próximos	1	0	Por definición las 51 provincias tienen 1 único vecino		No
K=2 vecinos más próximos	2	0	Por definición las 51 provincias tienen 2 vecinos		No
K=4 vecinos más próximos	4	0	Por definición las 51 provincias tienen 4 vecinos		No
0.75M	5.13	0	1 Provincias con 2 vecinos	1 Provincia con 10 vecinos	Sí
M	9	0	3 Provincias	1 Provincia	Sí

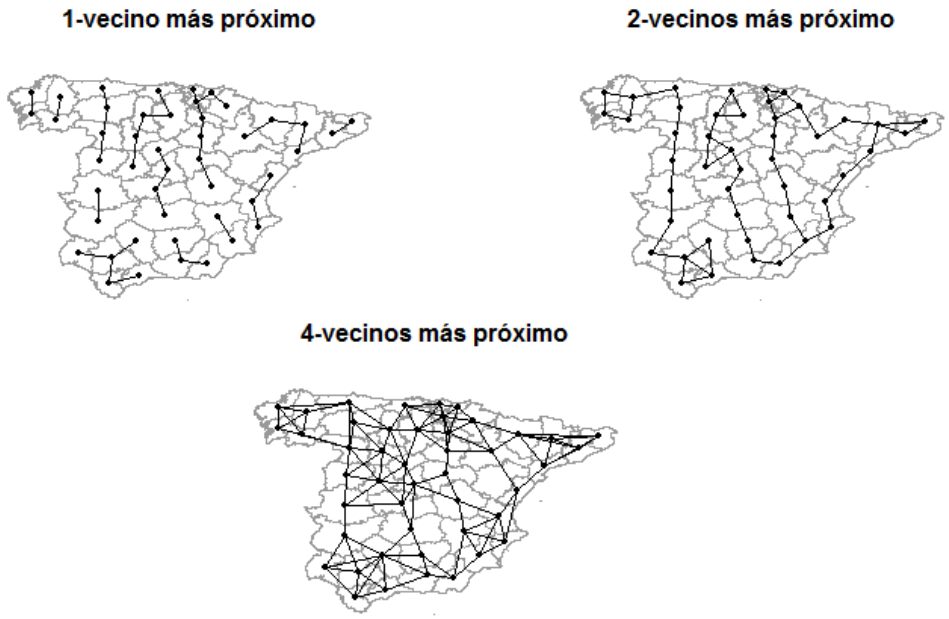

Figura 2.12: Representación de los vecinos de cada provincia de España con los diferentes métodos k vecinos más próximos.

Los criterios que tienen una media de vecinos por provincia más alta son 1M y 1.5M, con 17.35 y 9 vecinos respectivamente. Podemos destacar que con el criterio 1.5M Madrid y Toledo tienen 28 vecinos cada una. Los criterios de la Esfera de Influencia, contigüidad, 0.75M y triangulación Delaunay otorgan una media de vecinos a cada provincia similar, entre 4.3 y 5.4. Estos 6 criterios son simétricos y no deja ninguna provincia aislada.

Los métodos de Gabriel y Vecinos relativos tienen una media por provincia más baja, 1.85 y 1.26 respectivamente, ninguno es simétricos y son los únicos que dejan alguna provincia sin vecinos; 8 con el método de Gabriel y 11 con el de los vecinos relativos.

Mención especial merecen los criterios k-vecinos más próximos. Por definición todas las provincias tendrán k vecinos. Por tanto, no habrá ninguna provincia sin vecinos, la media de vecinos por provincia será el valor k fijado y no serán simétricos.

Observando los criterios de triangulación vemos como se vuelven cada vez más restrictivos. Partimos de los 5.43 vecinos de media según Delaunay y descendemos hasta los 1.26 del grafo de vecindad relativa.

Gracias a este resumen somos conscientes de las grandes diferencias existentes entre los diversos criterios de vecindad. Por ello debemos de prestar atención a cuál elegimos para analizar nuestros datos. Trabajaremos con los siguientes

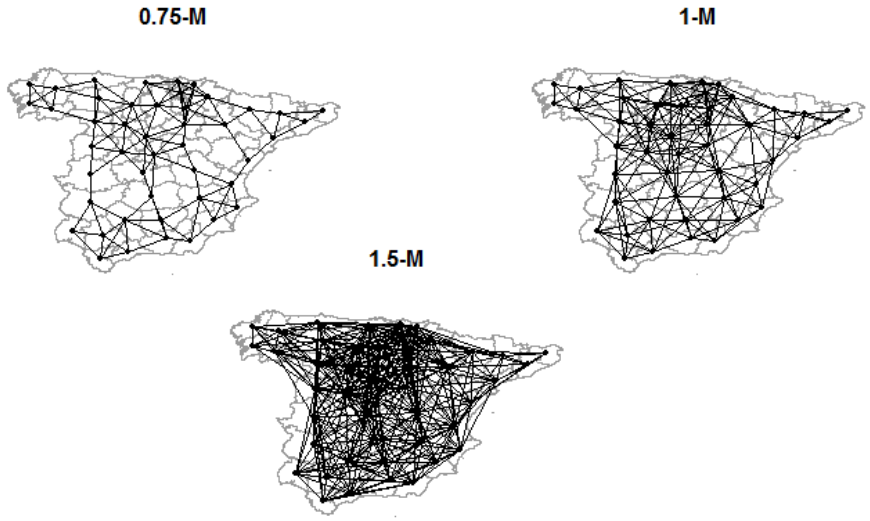

Figura 2.13: Representación de los vecinos de cada provincia de España con los diferentes métodos basados en la distancia.

criterios porque son bastante dispares y cada uno representa a uno de los grandes grupos de criterios de vecindad.

- Criterio Queen/Rook: es un criterio con un número medio de vecinos aceptable; ni demasiado grande ni demasiado pequeño. Además, considerar vecinos a aquellos con una frontera común es bastante apropiado en nuestros datos.
- Trianguación de Delaunay : aunque tienen un número de vecinos por provincia bastante similar a los criterios de contigüidad trabajaremos con este porque une provincias muy alejadas; Pontevedra con Huelva o Girona con Valencia.
- 1 vecino más próximo: es un criterio diferente a los demás y, por definición obliga al número de vecinos que debe tener cada provincia. De los que hemos elegido es el único que no es simétrico.
- 1M: Consideraremos este método para trabajar con un gran número de vecinos por provincia. No elegimos 1.5M porque el número de vecinos resulta excesivo.

Para aplicar los pesos a los conjuntos de vecinos tan solo hay que aplicar la función *nb2listw* y seleccionar el estilo que se desee.

```
esp_pro_W <- nb2listw(esp_nb,zero.policy=TRUE)
esp_pro_B <- nb2listw(esp_nb,zero.policy=TRUE, style="B")
esp_pro_C <- nb2listw(esp_nb,zero.policy=TRUE, style="C")
```

```
esp_pro_U <- nb2listw(esp_nb,zero.policy=TRUE, style="U")
esp_pro_S <- nb2listw(esp_nb,zero.policy=TRUE, style="S")
dsts <-  nbdists(esp_nb, coordinates(esp_pro))
idw  <-  lapply (dsts, function(x) 1/(x/1000))
esp_pro_dist <- nb2listw(esp_nb, glist=idw, style="B",
zero.policy=TRUE)
```

Ahora asignamos a cada unión un peso y en la Tabla 2.2 vemos un resumen descriptivo de los pesos obtenidos al aplicar los distintos métodos sobre los criterios de vecindad seleccionados.

Pesos \ **Vecindad**	Contigüidad	K=1	Delaunay	1M
W				
min-max	0.125-0.500	1-1	0.125-0.333	0.062-0.333
media	0.219	1	0.184	0.111
sd	0.081	0	0.044	0.048
B				
min-max	1-1	1-1	1-1	1-1
media	1	1	1	1
sd	0	0	0	0
C				
min-max	0.219-0.219	1-1	0.184-0.184	0.111-0.111
media	0.219	1	0.184	0.111
sd	0	0	0	0
U				
min-max $*10^{-2}$	0.476-0.476	2.17-2.17	4-4	0.241-0.241
mediana $*10^{-2}$	0.476	2.17	4	0.241
sd	0	0	0	0
S				
min-max	0.168-0.336	1-1	0.153-0.249	0.085-0.196
mediana	0.219	1	0.184	0.111
sd	0.037	0	0.021	0.022
Dist				
min-max	428.2-2288.0	865.9-2288.0	195.2-2288.0	428.0-2288.0
mediana	889.7	1224.0	814.4	717.6
sd	308.15	306.39	340.93	291.78

Tabla 2.2: Resumen de los pesos en cada conjunto de vecinos.

Utilizando el criterio binario todos los pesos son iguales a 1. El criterio W considera que pesan más los vecinos de áreas con menor número de vecinos. Por ello con el criterio de vecindad basado en la distancia aparece un peso muy

pequeño, 0.062, peso asignado a cada unión del área con 16 vecinos.

Los métodos C y U consideran todos los pesos iguales y los criterios con mayor número de uniones tendrán un peso menor. Vemos la diferencia de pesos entre el método con más vecinos 1M y el método con menos $k = 1$; 0,111 y 1 con el criterio C y $0{,}241 \cdot 10^{-2}$ y 0,0217 con el U. La desviación de los pesos con el estilo S es la más pequeña de todas (salvo en los casos en que los pesos son constantes).

El último criterio que consideramos es el basado en la distancia entre los vecinos. Que asignará mayor peso a las uniones de vecinos más próximos. En todos los casos el peso máximo es de 2288.0, este peso corresponde a la unión entre Vizcaya y Álava, que son las capitales de provincias más próximas.

Para los siguientes análisis no utilizaremos el criterio U, ya que trabaja con pesos demasiado bajos.

Ahora que hemos definido los pesos podemos empezar con el estudio de la autocorrelación, para ello utilizaremos las funciones *moran.test* y *localmoran* ambas del paquete *spdep*. Empezaremos con la autocorrelación global para saber si el porcentaje de prendas devueltas en España esta aleatoriamente distribuido por el espacio o, por el contrario, sigue un patrón.

```
moran.test(Datos_unidades$Tasa_unidades_devuelto, listw=esp_pro_W,
           zero.policy=TRUE, na.action=na.pass)
```

Aplicando este test a los pesos en todos los casos obtenemos una autocorrelación espacial significativa y positiva. Lo que significa que desde el punto de vista global el porcentaje de unidades devueltas no es independiente de la posición espacial y que el comportamiento se contagia a áreas próximas. Las zonas que tienen un porcentaje de devolución alto suelen estar rodeadas de zonas con alto porcentaje de devolución y viceversa. Este resultado es compatible con el comportamiento que ya intuimos con la Figura 2.8.

Con el mapa de colores habíamos notado la presencia de un porcentaje de devolución bajo en la zona noreste de la península, veamos si podemos confirmar esta suposición mediante el test local de Moran. Aplicamos este test a todos los pares de pesos y vecinos vistos y en la Tabla 2.3 aparecen representados los valores de la I estandarizada y el p-valor del test solamente de las provincias en las que se obtuvo significación para alguna combinación de vecinos y pesos.

```
localmoran(Datos_unidades$Tasa_unidades_devuelto, listw=esp_pro_W,
           zero.policy=TRUE, na.action=na.pass)
```

Provincias	Contigüidad	K=1	Delaunay		1M
	B	B	B	Dist	B
Huelva	1.87 ($p = 0{,}031$)		2.16 ($p = 0{,}015$)	1.85 ($p = 0{,}032$)	2.94 ($p < 0{,}01$)
Cantabria	2.30 ($p = 0{,}011$)		2.60 ($p < 0{,}01$)	2.81 ($p < 0{,}01$)	
Palencia	3.33 ($p < 0{,}01$)	1.75 ($p = 0{,}040$)	1.84 ($p = 0{,}032$)	2.34 ($p < 0{,}01$)	1.75 ($p = 0{,}040$)
Girona	7.14 ($p < 0{,}01$)	5.63 ($p < 0{,}01$)	8.53 ($p < 0{,}01$)	7.72 ($p < 0{,}01$)	5.63 ($p < 0{,}01$)
Lleida	6.33 ($p < 0{,}01$)	2.87 ($p < 0{,}01$)	7.64 ($p < 0{,}01$)	6.89 ($p < 0{,}01$)	2.87 ($p < 0{,}01$)
Tarragona	3.49 ($p < 0{,}01$)	2.87 ($p < 0{,}01$)	5.28 ($p < 0{,}01$)	4.97 ($p < 0{,}01$)	2.87 ($p < 0{,}01$)
Barcelona	7.34 ($p < 0{,}01$)	5.63 ($p < 0{,}01$)	7.70 ($p < 0{,}01$)	7.67 ($p < 0{,}01$)	5.63 ($p < 0{,}01$)
Huesca		2.93 ($p < 0{,}01$)	3.65 ($p < 0{,}01$)	3.43 ($p < 0{,}01$)	
Burgos		2.07 ($p = 0{,}019$)			2.07 ($p = 0{,}019$)
Gipuzcua			3.35 ($p < 0{,}01$)		
Valladolid		1.75 ($p < 0{,}01$)			1.75 ($p = 0{,}040$

Tabla 2.3: Resumen de la autocorrelación local. Solo se muestran las provincias donde se obtuvo significación.

Con los criterios de vecindad por contigüidad, 1-vecino más próximo y 1M se obtiene significación en las mismas provincias con independencia del peso utilizado. En la tabla se muestra el resumen del peso binario. Con el criterio de vecindad Delaunay, se obtienen las mismas provincias con todos los pesos menos con el basado en la distancia. Por ello en la tabla aparecen los valores obtenidos por los métodos binario y de la distancia.

Según los datos obtenidos aplicando el test local de Moran, podemos concluir que las cuatro provincias de Cataluña tienen autocorrelación local significativa y positiva, formando un cluster de provincias con bajo porcentaje de devolución. Este comportamiento también se extiende a Huesca. Distinguimos otro cluster, esta vez de provincias con alto porcentaje de devolución, en el centro-norte peninsular, con las provincias de Burgos, Valladolid, Palencia y Cantabria.

A partir de aquí solo representaremos los resultados obtenidos con el criterio de contigüidad y pesos binarios, aunque se harán los análisis para todos los criterios de la Tabla 2.2. Se comentaran los resultados en el caso de encontrarse diferencias.

Representamos el diagrama de caja de los valores I_i del test de Moran local (Figura 2.14). Las provincias que más aportan al estadístico global son las cuatro provincias catalanas y Palencia, estas 5 provincias perteneces a un cluster. Principalmente Barcelona, Gerona y LLeida son las que más influyen en la autocorrelación global.

Para completar los resultados analíticos de la autocorrelación, se añade el

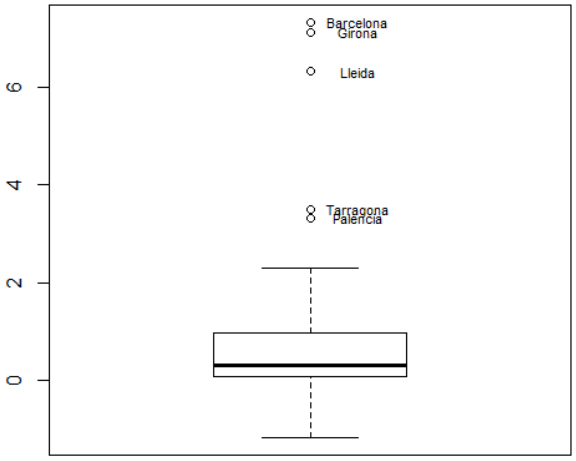

Figura 2.14: Diagrama de cajas de la autocorrelación local.

Diagrama de Dispersión de Moran, Figura 2.15. Este diagrama se realiza mediante la orden *moran.plot* de R. Sin más que proporcionarle como argumentos la variable y la matriz de pesos.

Los puntos se encuentran distribuidos en los cuatro cuadrantes, esto nos dice que existen puntos de autocorrelación local positiva y negativa. Además, la pendiente de la recta de regresión no está demasiado alejada de la horizontal por lo que vemos visualmente que la autocorrelacióno global no será muy alta, en concreto, para este modelo es de 0.367. Existen tres punto atípicos que corresponden a Lleida, Tarragona y Girona, que por encontrarse en el tercer cuadrante sabemos que estas tres provincias tienen un porcentaje de devolución bajo y están rodeadas de provincias con el mismo comportamiento que ellas.

```
moran.plot(Datos_unidades$Tasa_unidades_devuelto, esp_pro_B,
          pch=19, zero.policy=TRUE,xlim=c(5.5,20), ylab= " ",
          xlab= "Porcentaje de unidades devueltas")
```

Por último, vamos a intentar ajustar distintos modelos autorregresivos a nuestros datos. La variable dependiente será el porcentaje de devolución con el que estamos trabajando y las variables independientes son el número de tiendas, el importe vendido, la edad media y la tasa de paro en cada provincia durante el 2014. Estas dos últimas variables se consiguieron en la página web del Instituto Nacional de Estadística.

Para la estimación de los modelos autorregresivos se utilizará la función *spautolm* del paquete *spdep*. Esta función permite ajustar un modelo SAR o CAR, tomando como argumentos las variables incluidas en el modelo y la matriz de pesos espaciales.

La variable *importe vendido* tiene una distribución muy asimétrica con dos valores mucho mayores que el resto (las ventas de Madrid y Barcelona). Esto

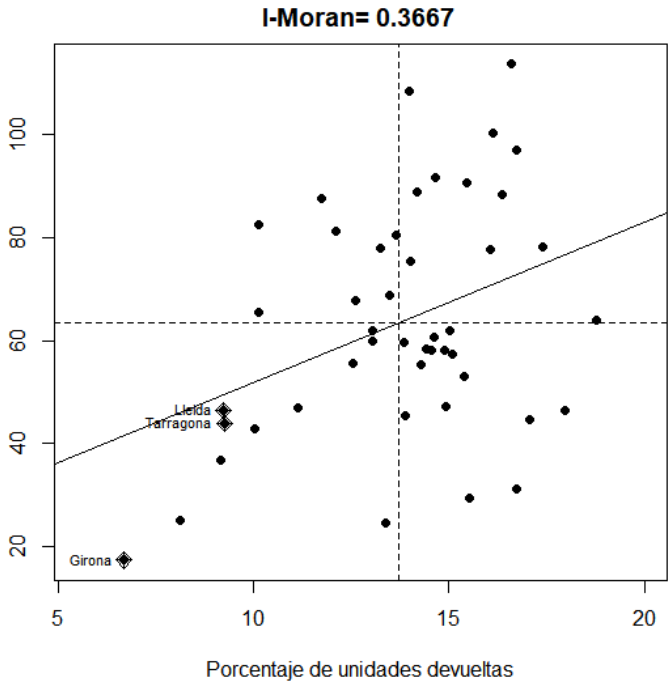

Figura 2.15: Diagrama de dispersión de Moran.

hace que al trabajar con ella se llegue a una matriz singular y, por tanto, no se puedan estimar los coeficientes del modelo. Para solucionar este problema se trabaja con el logaritmo de dicha variable.

```
spautolm(Tasa_unidades~Numero_tiendas+log(Importe_vendido)+Edad+
 +Paro, listw=esp_pro_B , family="SAR", data=Datos_unidades)
```

Empezaremos el análisis ajustando un modelo SAR. El modelo completo aparece en la tabla 2.4.

En este modelo ninguna variable es significativa. Siguiendo un modelo de pasos hacía atrás llegamos al modelo final, tabla 2.5. Este modelo cuenta con el logaritmo del importe vendido como única variable explicativa.

A la vista de los resultados obtenido podemos afirmar que existe correlación espacial significativa en los residuos, su estimación es 0.146. Además, al hacer al comparar este modelo con otro sin autocorrelación espacial ($\rho = 0$) obtenemos significación estadística ($p < 0{,}001$). El logaritmo del importe vendido es la única variable significativa en el modelo final. Al ser su estimación positiva podemos concluir que el porcentaje de devolución aumenta con el importe vendido en la provincia.

Ahora vamos a ajustar un modelo CAR. Los resultados del modelo completo se muestran en la tabla 2.6. Ninguna variable llega a ser significativa, pero el logaritmo del importe vendido y la tasa de paro están cerca de serlo.

	Estimación	Error estándar	p-valor
Constante	0.077	14.14	0.996
Numero_tiendas	-0.03	0.06	0.582
log(Importe_vendido)	0.68	0.44	0.118
Edad_media	-0.02	0.19	0.898
Paro	0.105	0.08	0.118

ρ =0.141, p-valor = $4{,}91 * 10^{-3}$

σ^2=4.13

AIC=215.71

Tabla 2.4: Resumen del modelo SAR ajustado por dos variables.

```
spautolm(Tasa_unidades~Numero_tiendas+log(Importe_vendido)+Edad+
Paro, listw=esp_pro_B , family="CAR", data=Datos_unidades)
```

Utilizando el criterio de selección de variables por pasos hacia atrás llegamos al modelo final, en el que aparecen la variable logaritmo del importe vendido y el paro. El resumen de este modelo se muestra en la tabla 2.7. Al ser los dos coeficientes estimados positivos, podemos decir que el porcentaje de devolución aumenta con la tasa de paro y con el importe vendido.

Los dos modelos estiman el coeficiente del logaritmo del importe vendido cercano a 0.5, ambos presentan un coeficiente de autocorrelación espacial significativo y de valor similar, además en ambos casos el valor σ^2 sobrepasa 4. La principal diferencia entre estos dos modelos es que cuando planteamos un modelo CAR permanece en el modelo la tasa de paro, que aunque no llega a ser significativa el modelo sin esta variable tendría un AIC más alto. El coeficiente asociado a esta variable también se estima con un valor positivo por lo que el porcentaje devuelto aumentará con el paro. Para comparar ambos modelos utilizamos el criterio AIC que, aunque es muy parecido, es menor en el modelo SAR.

En este punto podemos decir que en todas las combinaciones de criterios de vecindad y pesos obtenemos resultados parecidos a los dos modelos propuestos anteriormente. La principal diferencia aparece cuando se considera el peso de estandarización por filas, W. Cuando utilizamos este peso se aparece como única variable significativa la edad. Además, la estimación de su coeficiente es

	Estimación	Error estándar	p-valor
Constante	4.403	3.928	0.262
log(Importe_vendido)	0.499	0.226	0.027

ρ =0.146, p-valor = $1{,}01 * 10^{-4}$

σ^2=4.33

AIC=212.42

Tabla 2.5: Resumen del modelo SAR final.

negativa, la tasa de devolución disminuirá según aumente la edad media de la provincia en cuestión.

En la figura 2.16 vemos la representación gráfica de los residuos del modelo SAR visto en la tabla 2.5, tanto en un boxplot como representados en el mapa. En el diagrama de cajas se distinguen dos puntos atípicos, que corresponden a Palencia y a Girona. En Palencia se devuelve más de lo esperado mientras que en Girona pasa lo contrario, el porcentaje de devolución real es más baja que lo se esperaría según el modelo. En el mapa vemos que aunque la mayoría de residuos están entre -2 y 2, existen algunos que llegan hasta 4. Los residuos positivos corresponden a las provincias en las que la devolución fue más alta a la esperada, mientras que los negativos ocurre lo contrario. Las zonas con devolución mayor a lo esperado están dispersas por toda la península, sin encontrar un patrón claro. Mientras que las zonas con porcentaje de devolución menor corresponden a la zona noreste.

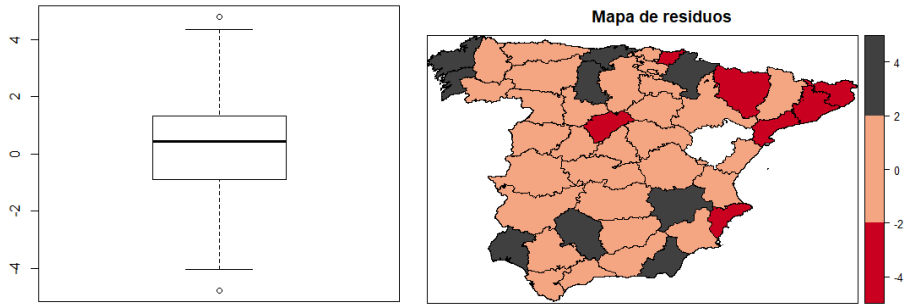

Figura 2.16: Representación de los residuos del modelo SAR mediante un diagrama de cajas (izquierda) y un mapa (derecha) .

En la Figura 2.17 comparamos el mapa de valores ajustados por el modelo con el mapa de valores observados. En un primer vistazo vemos que el modelo tiene a subestimar la devolución, solo León se mantiene con un porcentaje de

	Estimación	Error estándar	p-valor
Constante	-6.12	13.60	0.653
Numero_tiendas	-0.04	0.06	0.433
log(Importe_vendido)	0.82	0.44	0.060
Edad	0.04	0.19	0.820
Paro	0.14	0.08	0.082

ρ =0.183, p-valor = $6{,}81 * 10^{-3}$

σ^2=4.078

AIC=216.3

Tabla 2.6: Resumen del modelo CAR.

devolución mayor al 15.6 %. La autocorrelación espacial que estima el modelo hace que el comportamiento de las provincias catalanas se contagie a sus vecinas y considere un cluster de provincias con bajo porcentaje de devolución mucho mayor que el observado.

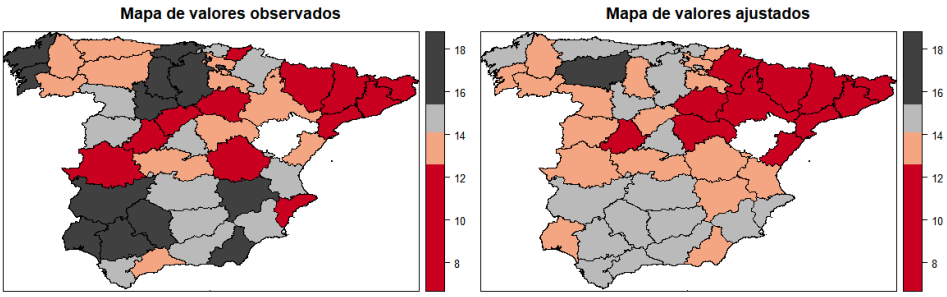

Figura 2.17: Representación en el mapa de los valores observados (izquierda) y de los valores ajustados por el modelo SAR (derecha).

Con todo lo visto podemos concluir que el porcentaje de devolución tiene una clara dependencia espacial. Debido principalmente al comportamiento del noreste peninsular, sobretodo Cataluña, donde las unidades devueltas en función de las compradas es mucho menor que en el resto de la España. El porcentaje de devolución de esta comunidad autónoma es de 8.33 mientras que el del país sin tener en cuenta a Cataluña asciende a 14.22. Existe otra zona peculiar, en este caso con un porcentaje de devolución más alto en el centro norte peninsular,

	Estimación	Error estándar	p-valor
Constante	0.003	4.70	0.999
log(Importe_vendido)	0.539	0.237	0.023
Paro	0.142	0.07	0.054

ρ =0.183, p-valor = $6{,}89 * 10^{-3}$

σ^2=4.13

AIC=212.92

Tabla 2.7: Resumen del modelo CAR.

formada por Burgos, Valladolid, Palencia y Cantabria, en esta zona el porcentaje devuelto es de 17.22 mientras que en el resto de España sin tener esta zona en cuenta disminuye a 13.39. En el cómputo global la comunidad catalana tiene más peso en la autocorrelación global.

Además, planteando modelos autoregresivos vemos que tanto el logaritmo del importe vendido como la tasa de paro tienen relación positiva con el porcentaje de devolución. Aquellas provincias con una tasa de paro alta y en las que se vendió un alto importe, tienen un porcentaje de devolución alto. En la situación en la que obtenemos significación en la edad, esta variable presenta una relación negativa, las provincias más envejecidas tienen un porcentaje de devolución más bajo.

Por otro lado, con el modelo SAR planteado en 2.5 distinguimos dos provincias con un comportamiento atípico. Palencia y Girona tienen un porcentaje de devolución que no les correspondería en según el modelo ajustado. Palencia debería recibir menos devoluciones y Girona más. Estas dos provincias pertenecían a los clusters detectados, por lo que este comportamiento atípico podría deberse a un contagio del comportamiento de sus zonas vecinas.

Capítulo 3

Datos en red

3.1. Conceptos básicos

Definimos una red como el conjunto de elementos y las interacciones entre ellos. Podemos conseguir datos de este tipo en multitud de campos, cada uno tendrá sus propias particularidades. Sin embargo, desde el punto de vista estadístico se ha desarrollado una base metodológica común para las diversas áreas que precisan del estudio de datos en red.

Las redes se suelen representar de manera formal mediante un grafo, debido a esto es frecuente que ambas palabras (grafo y red) se intercambien. A lo largo de este trabajo se utilizará la palabra red para referirnos a su sentido más general, y se utilizará grafo (grafo de la red) para hablar de su representación.

Al utilizar los grafos como la herramienta de representación de redes estamos heredando vocabulario específico de la teoría de grafos. Por ello, antes de comenzar el análisis de la red haremos una breve revisión de algunos conceptos básicos.

- Un grafo es una colección de vértices y de aristas que unen dichos vértices. Desde un punto de vista formal diremos que un grafo $G = \{V, E\}$ es una estructura matemática formada por un conjunto V de vértices (también llamados nodos) y un conjunto E de aristas, donde los elementos de E son pares (u, v) de distintos vértices $u, v \in V$. Cuando el par pares (u, v) está ordenado diremos que el grafo es dirigido, en caso contrario diremos que el grafo es no dirigido.
- El número de vértices $N_v = |V|$ será el orden del grafo y el número de aristas $N_e = |E|$ será el tamaño de G.
- Decimos que $H = \{V', E'\}$ es un subgrafo del grafo $G = \{V, E\}$ si $V' \subseteq V$ y $E' \subseteq E$.
- Un lazo se formará cuando exista una arista que conecte un vértice consigo mismo. Cuando exista un par de vértices unidos por más de una arista diremos que esta arista es múltiple. En caso de que existan alguno de estos dos conceptos estaremos trabajando con un grafo múltiple. Como norma

general trabajaremos con grafos simples, que son aquellos que no tienen ni lazos ni aristas múltiples.

- Dos vértices son adyacentes o vecinos si están unidos por una arista. Equivalentemente, dos aristas son adyacentes si tienen un vértice común.

- Decimos que un grafo es completo si todo par de vértices es adyacente, es decir, si tiene todas las aristas posibles. Este concepto se utiliza sobre todo para buscar subgrafos completos.

- La matriz de adyacencia, A, es una forma muy útil de representar un grafo G. Esta matriz es de tamaño N_v x N_v y se define como

$$A = \begin{cases} 1 \text{ si } (i,j) \in E \\ \\ 0 \text{ en otro caso} \end{cases}$$

  Es decir, tomará el valor 1 en la posición (i,j) cuando los vértices correspondientes sean adyacentes en G, y 0 si no lo son.

A la hora de hablar de movimiento dentro de un grafo aparecen diversas definiciones.

- Un paseo en G de v_0 a v_s es una secuencia que alterna vértices y aristas $v_0, a_1, v_1, a_2, \ldots, v_{s-1}, a_s, v_s$, en la que la arista a_i une los vértices v_{i-1}, v_i. La longitud de dicho paseo, s, es el número de aristas que hay que atravesar para ir desde v_0 a v_s.

- Un paseo en el que no se repite ninguna arista es un camino. Si un camino empieza y acaba en la misma arista será un circuito. Un paseo de longitud ≥ 3 que empiece y acabe en el mismo vértice y no repita ningún vértice (salvo en inicial y el final) será un ciclo.

- Decimos que un vértice $v \in V$ es accesible desde otro $u \in V$ si existe un paseo desde u a v. Un grafo G será conexo si todo par de vértices es accesible.

- Cada uno de los subgrafos conexos reciben el nombre de componente conexa. A veces, una de estas componentes contiene a la mayoría de los ejes del grafo, la llamaremos componente gigante. Es común omitir del análisis los vértices que no se encuentren en la componente gigante, ya que trabajar con un subgrafo conexo disminuye la complejidad de los análisis.

- Para hablar de distancia entre pares de vértices utilizamos la distancia geodésica. Esta distancia se define como la longitud del camino más corto entre los vértices.

- La mayor distancia geodésica del grafo se considera el diámetro de la red. Si no es posible llegar a un vértice desde otro la distancia geodésica será ∞ y este valor no se considerará a la hora de calcular el diámetro.

Todas estas definiciones se han realizado para grafos no dirigidos, para su extensión al caso dirigido solo habrá que tener en cuenta la dirección de las aristas. En el estudio de la conexión de un grafo dirigido se diferenciará entre conexo fuerte si tiene en cuenta la dirección de las aristas y conexo débil en caso contrario.

Estas son las definiciones básicas de la teoría de grafos. Conocer estos conceptos es obligatorio para entender los nuevos conceptos que irán surgiendo a medida que profundicemos en el análisis.

3.2. Visualización y análisis descriptivo

Una vez que tenemos definida la red con la que vamos a trabajar suele ser útil representarla gráficamente. Por lo general los vértices se representan mediante figuras geométricas (puntos, círculos, cuadrados...) y las aristas mediante curvas. En ocasiones se añade a la representación información adicional dada por los atributos.

Si la red es de tamaño reducido su representación es simple, tan solo son figuras geométricas unidas por curvas, incluso puede hacerse de forma manual. Sin embargo, el problema se complica a medida que el tamaño de la red aumenta. Cuando trabajamos con redes de tamaño alto una correcta representación debe considerar una combinación de criterios de optimalidad, algoritmos y aspectos estéticos. Existen infinitas maneras de representar una misma red, pero lo que se intenta es dibujar la red de manera que muestre la información más relevante para nuestro problema.

En la práctica las convenciones del dibujo, la estética y las restricciones se utilizan para plantear una serie de parámetros cuya determinación sirve para definir el método de visualización. El cálculo de dichos parámetros se realiza mediante la resolución de problemas de optimización. Cuando la red es de gran tamaño, no es posible calcular el valor exacto de los parámetros, por lo que se utilizan aproximaciones. Algunos métodos para la representación automática de redes son los siguientes:

- El método del círculo. Este método consiste en dibujar los vértices en una circunferencia.
- El método de Kamada y Kawai. Consiste en dibujar los vértices aproximadamente con la misma distancia entre ellos y con pocos cruces.
- El método Fruchterman y Reingold. Similar al método anterior, pero con otra función a optimizar.
- El método *Distributed Recursive Layout (DrL)* (MArtin et al 2008) Este método consiste en agrupar vértices y está especialmente diseñado para visualizar redes grandes. Los métodos anteriores, en situaciones de 100 o más vértices suelen mostrar una representación desordenada.

Dibujar el grafo de la red, además de ser útil para visualizarla y hacernos una idea general de la red, es imprescindible para llevar a cabo análisis más

complejos. Muchas de las cuestiones de interés que se nos plantean se pueden responder reformulándolas en términos del grafo. Preguntas sobre movimientos de información o mercancías puede plantearse en términos de caminos en la red y flujos en estos caminos; ciertas nociones sobre la "importancia" de un elemento individual del sistema pueden capturarse mediante medidas de como es de central el correspondiente vértice en la red; y la búsqueda de comunidades y grupos dentro de un sistema puede verse como un problema de partición del grafo.

El análisis descriptivo de la red se trata como un análisis estructural del grafo. Las herramientas que se utilizan para este propósito suelen proceder de áreas alejadas de la estadística; algunas se encuentran en la naturaleza de la teoría de grafos y otras tienen sus orígenes en matemáticas o las ciencias de la computación.

En esta sección veremos las características y propiedades de la estructura de las redes, centrándonos en tres partes: la caracterización de vértices y aristas, la caracterización de la cohesión de la red y la conectividad de la red.

3.2.1. Características de los vértices y aristas

Los vértices y las aristas son los elementos fundamentales del grafo, debido a esto hay un gran número de caracterizaciones de la red que se basan en ellos. Comenzaremos viendo las características de los vértices y, posteriormente, extenderemos dichas definiciones a las aristas.

Grado del vértice

Dado un vértice v perteneciente al grafo de la red $G = \{V, E\}$, su grado d_v se define como el número de aristas que inciden sobre él y definimos como f_d como la frecuencia relativa de los vértices $v \in V$ tales que $d_v = d$. El conjunto $\{f_d\}_{d \leq 0}$ será la distribución del grado de la red.

Esta distribución presenta diversas formas, pero, normalmente, suele presentar una alta asimetría positiva; la mayoría de los vértices tienen un grado bajo, mientras que solo unos pocos presentan grados mucho más altos.

Más allá de la propia distribución del grado, es interesante conocer la manera en que los vértices se conectan con otros de diferente grado. Para ello se calcula la media de los grados de sus vértices vecinos. Así podemos saber si un vértice tiende a conectarse con aquellos que tengan un grado alto o a los que tengan grados bajos.

Centralidad del vértice

Muchas preguntas que se nos plantean al trabajar con un grafo están relacionadas con conocer cuál es la "importancia" de un vértice. Para cuantificar la noción de importancia se proponen las medidas de centralidad. A pesar de ser una noción esencial en el análisis de redes, no hay consenso sobre qué significa

que un vértice sea importante ni sobre que medida debe utilizarse para medir dicha centralidad.

Una de las medidas más utilizadas es considerar el grado del vértice que hemos visto anteriormente. Un vértice será más importante cuanto mayor sea su grado, es decir, la importancia del vértice dependerá del número de vértices adyacentes que tenga.

Además de esta medida, hablaremos de otros tipos de medidas de centralidad: centralidad por cercanía, centralidad por intermediación y centralidad por estatus.

- La medida de centralidad por cercanía (Sabidussi 1966) intenta capturar la noción de que un vértice es central si está "cerca" de otros vértices. Esta medida se define como el inverso de la suma de la distancia geodésica del vértice a todos los demás.

$$c_{cer}(u) = \frac{1}{\sum_v dist(u,v)}$$

  Donde $dist(u,v)$ es la distancia geodésica entre los vértices $u, v \in V$. A veces tiene interés comparar esta medida en diferentes grafos o compararla con otras medidas de centralidad, para ello se normaliza multiplicando por $N_v - 1$, así siempre tomará valores en el intervalo $[0,1]$.

- La centralidad por intermediación (Freeman 1966) relaciona la importancia con la posición que ocupa un vértice respecto a los caminos del grafo. Considera que los vértices que se encuentran en más caminos son más importantes. Decimos que un vértice es central por intermediación si sus aristas adyacentes son las que aparecen con mayor frecuencia como parte de caminos más corto entre el resto de vértices del grafo.

$$c_{int}(u) = \sum_{s \neq t \neq u s,t \in E} \frac{\sigma(s,t|u)}{\sigma(s,t)}$$

  Donde $\sigma(s,t|u)$ es el número total de caminos más cortos entre los vértices s y t que pasan por u, y $\sigma(s,t)$ es el número total de caminos más cortos entre los vértices s y t (puede pasar por v o no).

  Como vimos anteriormente, si nos interesa comparar esta medida hay que normalizarla, pero, en este caso, para conseguirlo habrá que dividir por $(N_v - 1)(N_v - 2)/2$ si la red es no dirigida y por $(N_v - 1)(N_v - 2)$ si sí lo es.

- La última medida de centralidad es la centralidad por estatus (Freeman 1977). Un vértice es el más central por estatus si es el vértice con vecinos más centrales. La definición anterior es inherentemente implícita y se suele expresar en términos de autovectores de ciertos sistemas lineales de ecuaciones.

$$c_{est}(v) = \alpha \sum_{u,v \in V} c_{est}(u).$$

El vector $c_{est} = (c_{est}(1), \ldots, c_{est}(N_v))^T$ es la solución al problema $Ac_{est} = \alpha^{-1}c_{est}$, donde A es la matiz de adyacencia del grafo de la red G. La elección óptima de α^{-1} es el mayor autovalor de A y c_{est} es su correspondiente autovector.

Por convenio se toma el valor absoluto de cada valor propio entonces por la propiedad de ortonormalidad de los vectores propios, esta medida de centralidad siempre tomará valores entre 0 y 1.

Caracterización de las aristas.

Todas las medidas vistas anteriormente están definidas para vértices ya que, en la práctica, las cuestiones de centralidad suelen asociarse a los vértices. Sin embargo, hay algunas preguntas más relacionadas con las aristas, por ejemplo conocer que conexiones en las redes sociales son más importantes para la propagación de ciertos rumores.

Una manera natural de medir esto es con la centralidad por intermediación de aristas, que no es más que una extensión de esta centralidad en el caso de vértices. Una arista es el más central por intermediación si dicha arista es la que más veces aparece en los caminos más cortos entre los vértices.

$$c_{int,e}(e) = \sum_{s \neq t s,t \in V e \in E} \frac{\sigma(s,t|e)}{\sigma(s,t)}$$

Donde $\sigma(s,t|e)$ es el número total de caminos más cortos entre los vértices s y t que pasan por e, y $\sigma(s,t)$ es el número total de caminos más cortos entre los vértices s y t.

Hay otras medidas de centralidad de los vértices que no pueden extenderse a los ejes de manera tan sencilla. Una forma de solucionar el problema es aplicar las medidas de centralidad de los vértices al grafo línea de G. El grafo línea de G es un grafo $G' = \{V', E'\}'$ que se obtienen transformando los vértices de G en ejes de G' y los ejes en vértices. Es decir, los vértices de $v' \in V'$ representan los ejes originales $e \in E$, y los ejes $e' \in E'$ indican que los dos ejes correspondientes en E incidían sobre un vértice común en G.

3.2.2. Cohesión de la red

Un alto porcentaje de las preguntas relacionadas con el análisis de redes se reducen a preguntas sobre la cohesión de la red. Preguntas como; ¿los amigos de cierto actor en una red social tienden a ser amigos unos de los otros? o ¿qué colección de proteínas en una célula parecen trabajar juntas? son ejemplos de situaciones en las que necesitamos estudiar la cohesión de la red.

Existen muchas maneras en las que podemos definir la cohesión de la red, dependiendo del contexto de la pregunta. En esta sección veremos la densidad y la detección de comunidades.

La densidad

La densidad de una red se define como el cociente entre el número de aristas que existen y las aristas posibles. En una red no dirigida sin lazos y sin ejes múltiples la densidad del subgrafo H se define como

$$den(H) = \frac{|E_H|}{|V_H|(|V_H|-1)/2}.$$

El valor de la densidad de H toma valores entre 0 y 1. Cuanto más próximo sea a 1 más cerca está el grafo de ser completo. En caso de que H sea dirigido, el denominador de la expresión anterior se sustituirá por

$$|V_H|(|V_H|-1).$$

La sencillez del concepto de densidad se vuelve más interesante a través de la libertad que tenemos para elegir al subgrafo H. Si tomamos $H = G$ tendremos la densidad de la red completa. En cambio, si tomamos $H = H_v$ como el conjunto de vecinos del vértice $v \in V$ y las aristas entre ellos, obtendremos la densidad en los vecinos de v.

Detección de comunidades o partición del grafo

Una partición es la segmentación de conjuntos de elementos en subconjuntos. Formalmente la partición $C = C_1, \ldots, C_k$ de un conjunto finito S es la descomposición de S en K subconjuntos disjuntos no vacíos C_k tales que $U_{k=1}^{K} C_k = S$.

En el análisis de redes, la partición de un grafo, también llamada detección de comunidades, es una herramienta muy útil para encontrar subconjuntos disjuntos cuyos vértices tengan cohesión. Decimos que un subconjunto cohesivo de vértices es un subconjunto de vértices que están bien conectados entre ellos y están suficientemente separados del resto de vértices del grafo.

Existen numerosos algoritmos para la detección de comunidades en un grafo, resumiremos dos de ellos:

1. El algoritmo de Newman-Girvan parte de la idea de que las aristas que conectan diferentes comunidades tendrán un alto valor de cercanía por intermediación, ya que los caminos más cortos deben pasar por dicha arista. Si eliminamos la arista con más cercanía por intermediación dividiremos a la red en dos sub-redes, que a su vez podrán volver a dividirse. El problema fundamental de este algoritmo es su elevado coste computacional.

2. El algoritmo de Reichardt-Bornholdt es un algoritmo bastante complejo basado en la estadística física. Su idea es encontrar una partición de los vértices que optimizan una función llamada energía. Este algoritmo es similar a los métodos de partición de datos multivariantes como el de las K-medias.

3.2.3. Conectividad de la red

Otra pregunta de interés es conocer si un grafo puede separarse en distintos subgrafos, y si no es posible, tratar de cuantificar cómo de cerca está de poder

hacerlo.

Además de la definición de conectividad usual, en la que un grafo es conexo si para cada par de vértices existe un camino entre ellos, existen otras definiciones de conectividad que surgen de preguntarse si al eliminar un subconjunto cualquiera de k vértices o aristas del grafo, el subgrafo continúa siento conexo. Para dar una definición precisa de este concepto necesitamos conocer las definiciones de conectividad por vértices (aristas) y vértice (arista) de corte.

Dado un grafo G, decimos que está k-conectado por vértices si el número de vértices $N_v > k$ y si al eliminar algún subconjunto de vértices $X \subseteq V$ con $|X| < k$ el subgrafo sigue siendo conexo. De manera análoga, dado un grafo G, decimos que estará k-conectado por aristas si $N_v \geq 2$, y si al eliminar algún subconjunto de aristas $Y \subseteq E$ con $|Y| < k$ el subgrafo sigue siendo conexo. La conectividad de vértices (aristas) es el mayor entero k tal que G está k-conectado por vértices (aristas).

Si al eliminar cierto conjunto de vértices (aristas) de un grafo lo desconectamos, decimos que es un conjunto de vértice (aristas) de corte. Si existe un solo vértice que al eliminarlo desconectamos el grafo, lo llamamos punto de articulación. Identificar estos puntos puede ayudarnos a detectar los puntos en los que la red es más vulnerable.

Si el grafo con el que trabajamos es dirigido habrá que sustituir el concepto de conectividad por el de conectividad fuerte en las definiciones anteriores.

3.3. Aplicación a datos reales

En esta sección veremos de manera práctica un análisis de datos en red, para ello utilizaremos los datos de devoluciones de las prendas de ropa durante el año 2014.

Trabajaremos con las unidades compradas en la tienda i y devueltas en la tienda j $\forall i, j \in T$ donde T representa el conjunto de las tiendas de ZARA en la península ibérica durante el 2014. En la red consideraremos como vértices cada una de las tiendas y la existencia de una arista uniendo dos vértices representa que hubo al menos una prenda de ropa comprada en el vértice origen y devuelta en el vértice final. Se podría haber tomado otro criterio y trabajar únicamente con las tiendas entre las que hubiera un número mínimo de devoluciones, lo que simplificaría el problema considerablemente. Para el análisis se utilizó la librería *igraph* de R.

En un primer análisis vamos a ver la distribución de la variable número de prendas devueltas entre cada par de tiendas. De media entre cada par de tiendas se devuelven 82.8 unidades, la mediana es 4 y la desviación estándar es de 603.16 y su rango de valores va desde 1 hasta 18750. Con este rango de valores tan amplio es difícil obtener información de su representación gráfica, por ello representamos el logaritmo de esta variable.

En la Figura 3.1 vemos su boxplot y su función de densidad. En estas gráficas podemos ver que aunque esta variable es muy dispersa la mayoría de los valores es menor que 3 (en escala logarítmica) y que por encima de 5 solo aparecen puntos atípicos. Los puntos atípicos detectados llegan casi a los 3000 y el mayor de los puntos atípicos corresponde a las 18750 devoluciones en el Centro Comercial Vallereal de Camargo (Cantabria) de prendas compradas en la calle Lealtad de Santander.

```
e.tienda<- readWorkbook("DatosRed.xlsx", sheet=2)
boxplot(log(e.tienda$UNIDADES))
plot(density(log(e.tienda$UNIDADES)), main= " ", ylab="Densidad")
```

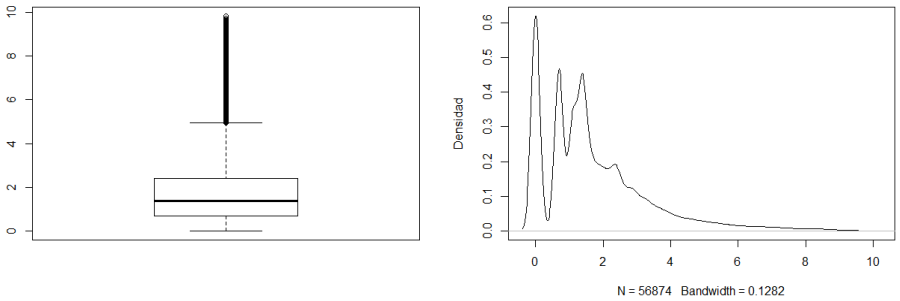

Figura 3.1: Boxplot y función de densidad del logaritmo de la variable número de devoluciones entre cada par de tiendas.

Una vez que hemos descrito esta variable, vamos a definir la red con la que vamos a trabajar. Para ello necesitamos dos archivos de datos, uno con la información de las devoluciones entre las ropas (e.tienda) y otro únicamente con las características de cada tienda (v.tienda).

```
v.tienda<- readWorkbook("DatosRed.xlsx", sheet=1)
library(igraph)
red <- graph.data.frame(e.tienda, directed=TRUE, vertices=v.tienda)
```

En un análisis descriptivo de la red vemos que es una red dirigida de orden 342 y de tamaño 56874. El diámetro es 2 y, además, es fuertemente conexa, por tanto, sabemos que para llegar de una tienda a otra cualquiera habrá una arista que las una, o bien, habrá que pasar por una tienda intermedia. La densidad de nuestra red es de 0.48; este valor nos dice que nuestra red tiene casi el 50 % de todas las aristas posibles.

```
orden <- vcount(red)
tamaño <- ecount(red)
diametro <- diameter(red)
densidad <- edge_density(red)
is.connected(red, mode="strong")
```

```
## [1] TRUE
```

Pasamos ahora al estudio de los grados del vértice, en el que debemos distinguir el grado de entrada del de salida. En la gráfica 3.2 vemos la distribución de los grados. Ambos grados siguen un comportamiento similar, la media de los grados de entrada es 166.3 y su desviación 57.40, mientras que en el grado de salida estos valores son 176 y 69.91 respectivamente.

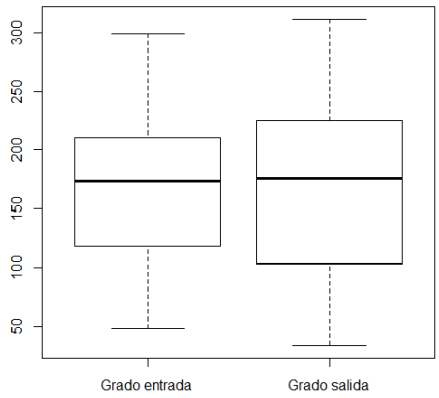

Figura 3.2: Representación de los grados de entrada y salida de los vértices de la red.

```
grado.in<- degree(red, mode="in")
V(red) $name[grado.in==max(grado.in)]
grado.out<- degree(red, mode="out")
V(red) $name[grado.out==max(grado.out)]
```

```
## [1] "93 - BEN-GAMBO"
## [1] "257 - MAD-PARQUE SUR"
```

La tienda con mayor grado de entrada es la tienda situada en la calle Gambó de Benidorm, a ella llegan devoluciones de prendas compradas en 299 tiendas. Por otro lado, la tienda con mayor grado de salida es la tienda situada en el Centro Comercial Parque Sur de Leganés. En esta tienda se compran prendas que posteriormente son devueltas en 311 tiendas distintas. Recordemos que el orden de este grafo es 343, entonces el 90 % de las tiendas reciben devoluciones de prendas de ropa compradas en esta tienda de ZARA.

Pasamos ahora a realizar el estudio de la centralidad de la red, para conocer las tiendas con más importancia en la distribución de las devoluciones.

```
cen.cer.red.in<- closeness(red, mode="in")
v.tienda [ which(cen.cer.red.in==max (cen.cer.red.in)),2]
cen.cer.red.out<- closeness(red, mode="out")
v.tienda [ which(cen.cer.red.out==max (cen.cer.red.out)),2]
```

```
## [1] "93 - BEN-GAMBO"
## [1] "257 - MAD-PARQUE SUR"
```

En la cercanía por centralidad es la única medida de centralidad en la que hay que dividir en cercanía de entrada y de salida. Las tiendas más centrales son las tiendas de la calle Gambó de Benidorm y la tienda del Centro Comercial Parque Sur de Madrid según consideremos entrada o salida. Ambas tiendas coinciden con las tiendas con mayor grado lo que es lógico, ya que son las que están conectadas a más tiendas de manera directa y, por tanto, la distancia geodésica de estas tiendas al resto será menor. El valor de centralidad estandarizado de la tienda de Benidorm es 0,89 y la de Leganés 0,92.

La tienda más central por intermediación es de nuevo la tienda del Centro Comercial Parque Sur. Recordemos que esta medida considera más central aquella que se encuentra en mayor número de caminos más cortos.

```
cen.int.red <-  betweenness(red)
v.tienda [ which(cen.int.red==max (cen.int.red)),2]
```

```
## [1] "257 - MAD-PARQUE SUR"
```

La última medida de centralidad es la centralidad por estatus. La tienda más central según esta medida es la tienda de la Calle Preciados de Madrid.

```
cen.sta.red <-  evcent(directed=TRUE)$vector
 v.tienda [ which(cen.sta.red==max (cen.sta.red)),2]
```

```
## [1] "62 - MAD-PRECIADOS"
```

La visualización de esta red no es sencilla debido a que estamos trabajando con un grafo de gran tamaño. Incluso aplicando las técnicas especiales para grafos de gran tamaño tendremos dificultades. En la Figura 3.3 se representa el grafo de la red mediante los métodos del círculo, Kamada-Kawai, Fruchterman-Reingold y DrL. También aparecen coloreadas de otro color las tiendas más centrales vistas anteriormente. Ninguna de estas tres representaciones nos proporcionan información relevante principalmente porque nuestro grafo, además de un número importante de vértices, tiene un número de aristas muy elevado. Esto hace que se crucen una con otras y nos sea imposible distinguirlas individualmente.

```
plot(red, layout=layout.circle)
plot(red, layout=layout.kamada.kawai)
plot(red, layout=layout.fruchterman.reingold)
plot(red, layout=layout.drl)
```

En cuanto a las aristas, también nos interesa conocer la más central, que será aquella que esté en más caminos más cortos. En la Figura 3.4 vemos los distintos valores de centralidad, destacan 4 aristas con un valor algo más alto que el resto. Tres de ellas partes de la tienda de Zara Kids del Centro Comercial de Cee en A Coruña, y finalizan en tiendas de Centros Comerciales de Madrid; La Vaguada, Isla Azul y Plenilunio. La última parte del Zara Kids de Totana (Murcia) y acaba en la tienda del Centro Comercial Parque Sur.

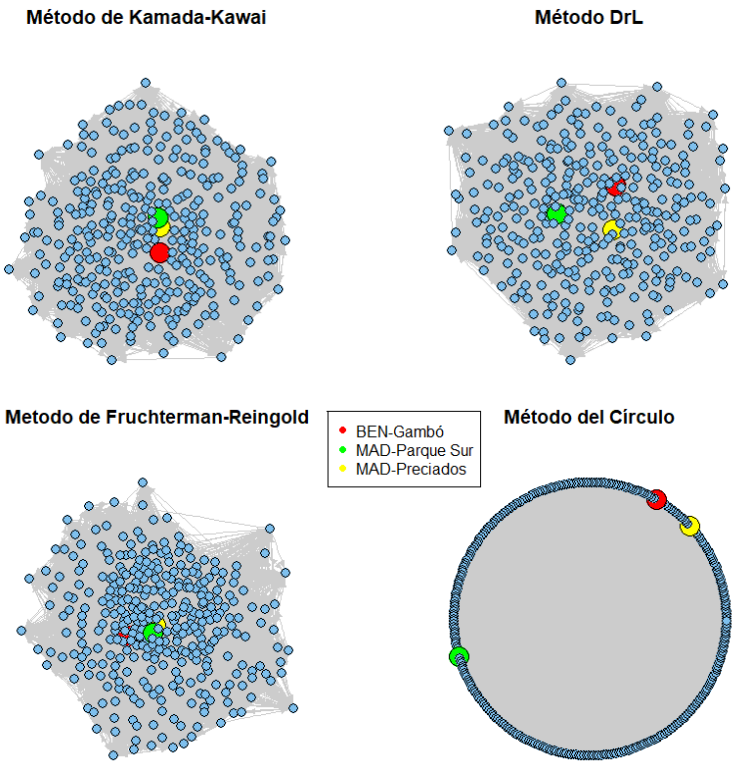

Figura 3.3: Visualización del grafo de la red según el método de Kamada-Kawai (arriba a la izquierda),el método de DrL (arriba a la derecha), el método Fruchterman-Reingold (abajo a la izquierda) y el método Círculo (abajo a la derecha).

Ya hemos visto que esta red es fuertemente conexa, pero profundizando en el estudio de la conectividad del grafo podemos ver que no tiene ningún punto de articulación. Además, será necesario eliminar 34 vértices para que el grafo deje de ser conexo y se descomponga en dos subgrafos.

```
articulation.points(red)
vertex.connectivity(red)
```

```
## [1] 0/342 vertices
## [1] 34
```

Por último, vamos a intentar ajustar un modelo de regresión a nuestros datos. Para ello también calculamos la distancia entre cada par de tiendas. Este cálculo lo hacemos con la orden *distHarvensine* del paquete *geosphere*, que permite calcular la distancia en línea recta entre dos puntos, teniendo en cuenta la esfericidad de la superficie terrestre, a partir de la longitud y la latitud de cada uno de ellos (Hijmans 2017) . También se utilizó la distancia geodésica de la red obteniendo resultados muy parecidos aunque algo peores, lo que hizo que nos

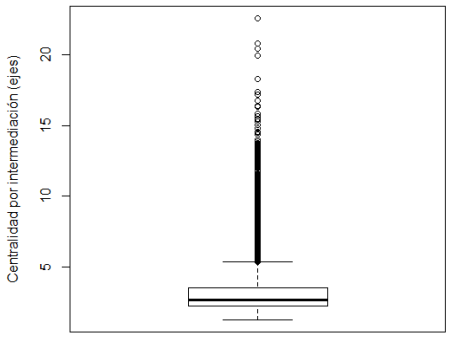

Figura 3.4: Distribución de los valores de centralidad de las aristas.

decantásemos por calcular la distancia mediante la fórmula de *Haversine*.

En la Figrua 3.5 vemos el histograma de la distancia. La distancia más pequeña entre tiendas corresponde a tiendas del mismo centro comercial, y la máxima (1049 km) corresponde a la distancia entre tiendas de Huelva y Gerona.

```
library(geosphere)
for(i in 1: nrow(e.tienda)){
e.tienda$distanciakm[i]<-distHaversine(
      c(e.tienda$lon.fin[i],e.tienda$lat.fin[i]),
      c(e.tienda$lon.origen[i],e.tienda$lat.origen[i]))/1000}
hist(e.tienda$distanciakm)
```

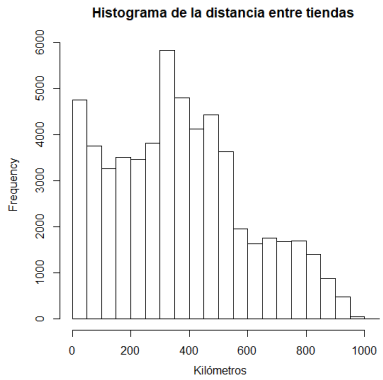

Figura 3.5: Histograma de la distancia entre cada par de tiendas.

El primer modelo que se plantea siguiendo es un modelo gravitacional (Kolaczyk y Csardi 2014). Estos modelos se proponen para modelar redes cuyas aristas sirven como conductores de flujo (en nuestro caso el flujo serán las prendas devueltas). Los modelos gravitacionales parten de la Ley de Gravitación Universal de Newton (de ahí su nombre), ley que asume que la iteración entre dos poblaciones varía proporcionalmente a su tamaño e inversamente propor-

cional a su separación.

El modelo gravitacional general considera el flujo Z una variable de conteo con distribución Poisson,

$$E(Z_{ij}) = h_0(i)h_d(j)h_s(c_{ij})$$

donde h_s, h_d y h_s son funciones positivas del origen, destino y del vector c_{ij} que representa las características de separación entre los nodos origen y destino, como pueden ser el coste o la distancia.

Si $E(Z_{ij}) = \mu_{ij}$ podemos plantear el modelo en escala logarítmica, suponiendo un efecto lineal de las variables en esta escala. El modelo sería:

$$log(\mu_{ij}) = \alpha_i + \beta_j + \theta' c_{ij}$$

donde $\alpha_i = log(h_0(i))$, $\beta_j = log(h_d(j))$ y $\theta' c_{ij} = log(h_s(c_{ij}))$, por lo que se trataría de un modelo lineal generalizado.

Planteamos el modelo considerando como variables independientes el número de unidades compradas en los nodos origen y destino y la distancia entre ellas. Para ello utilizaremos la función *glm*. El resumen de este ajuste se muestra en la Tabla 3.1.

```
formula <-  UNIDADES ~ u.vendidas.origen + u.vendidas.fin + distancia
modelo<- glm(formula, family="poisson", data=e.tienda)
```

	Estimación	Error estándar	p-valor
Constante	5.90	1.07e-3	<2e-16
Unidades origen	1.48e-6	1.43e-9	<2e-16
Unidades destino	1.37e-6	1.47e-9	<2e-16
Distancia	-1.97e-2	1.02e-5	<2e-16
AIC=10271311			

Tabla 3.1: Resumen del modelo gravitacional.

En la Figura 3.6 se muestran las gráficas de diagnóstico del modelo. En ellas se observa un mal ajuste, especialmente cuando el número de devoluciones es grande. También se puede ver que hay mucha heterocedasticidad.

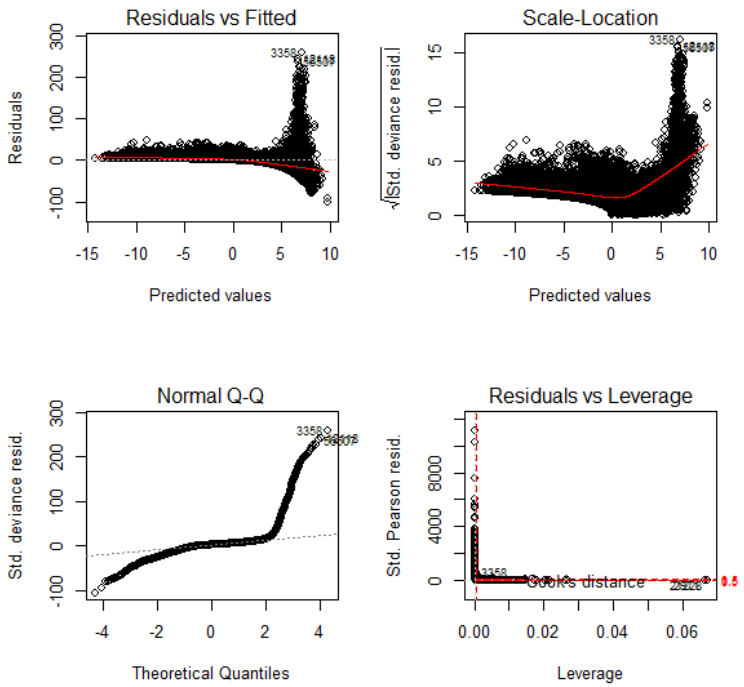

Figura 3.6: Validación del modelo gravitacional.

A la vista de estos resultados es necesario buscar otro modelo para ajustar nuestros datos. El problema de este modelo podría ser la sobredispersión, este fenómeno aparece cuando la varianza de la variable dependiente es mucho mayor a la esperada. La función *dispersiontest* del paquete *AER* nos permite comprobar la sobredispersión de un modelo (Kleiber 2008). Aplicando esta función a nuestros datos aceptaríamos que hay sobredispersión ($p < 0{,}001$).

```
library(AER)
dispersiontest(modelo)
```

```
   Overdispersion test

data:  modelo
z = 8.3221, p-value < 2.2e-16
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
   28086.8
```

Planteamos un modelo que tenga en cuenta esta característica. Sin embargo, como vemos en la Figura 3.7 sigue existiendo problemas de ajuste, con unos residuos demasiado grandes. Además, sigue sin solucionarse el problema de la variabilidad.

```
modelo.sobredisp<-glm(formula.s, family="quasipoisson", data=e.tienda)
par(mfrow=c(2,2))
plot(modelo.sobredisp)
summary(modelo.sobredisp)$dispersion
```

```
[1] 33617.43
```

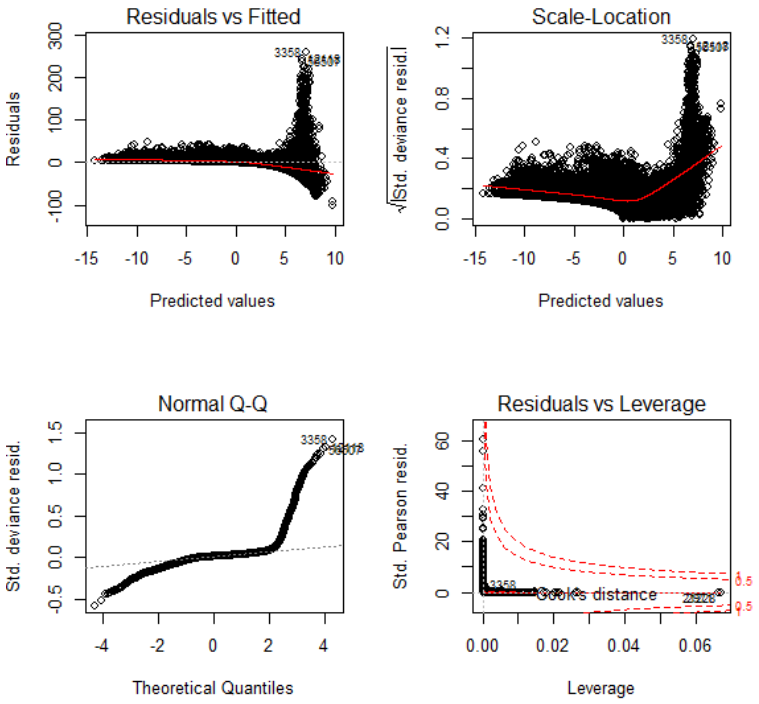

Figura 3.7: Validación del modelo con sobredispersión.

Al analizar el efecto de las variables del modelo se observa (Figura 3.8) que no es adecuado suponer que el efecto de la distancia sea lineal. Además, incluso considerando un efecto no paramétrico de la distancia, sigue apareciendo heterocedasticidad. Después de realizar un análisis exploratorio se decidió considerar el logaritmo del cociente entre las unidades devueltas y las unidades vendidas en el origen. Si representamos esta variable frente a la distancia, Figura 3.9 vemos que se soluciona el problema de la heterocedasticidad.

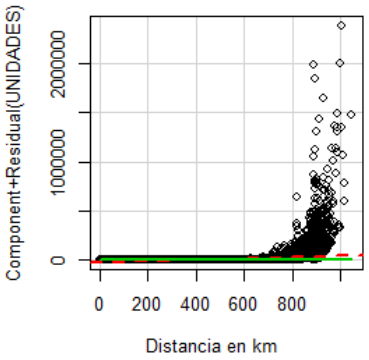

Figura 3.8: Gráfico parcial de residuos de la distancia en el modelo con sobredispersión.

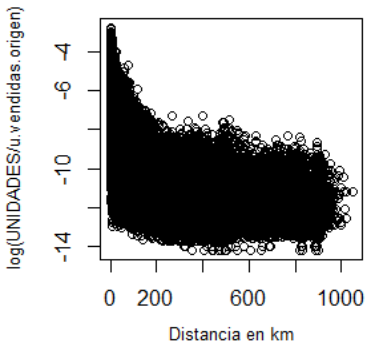

Figura 3.9: Diagrama del logaritmo de las unidades devueltas entre las vendidas, frente a la distancia.

Decidimos plantear un modelo aditivo (Faraway 2006), para ello utilizamos la función *gam* del paquete *mgcv* (Wood 2017). Los modelos aditivos generalizados son un tipo de modelos lineales generalizados en los que el efecto de las variables explicativas puede no ser lineal. El modelo que se plantea es

$$Y = \beta_0 + \sum_{j=1}^{P} f_j(x_j) + \epsilon,$$

donde $f_j(\cdot)$ es el efecto de la variable x_j, que se supone una variable suave cualquiera. Cuando $f_j(x) = \beta_j X$ tendríamos un efecto lineal. El paquete *mgcv* emplea *splines* para estimar las funciones.

Los gráficos de validación del modelo se muestran en la Figura 3.10.

```
library(mgcv)
modelo.gam<-gam(log(UNIDADES/u.vendidas.origen) ~
log(u.vendidas.fin) + s(distancia), data=e.tienda)
gam.check( modelo.gam)
```

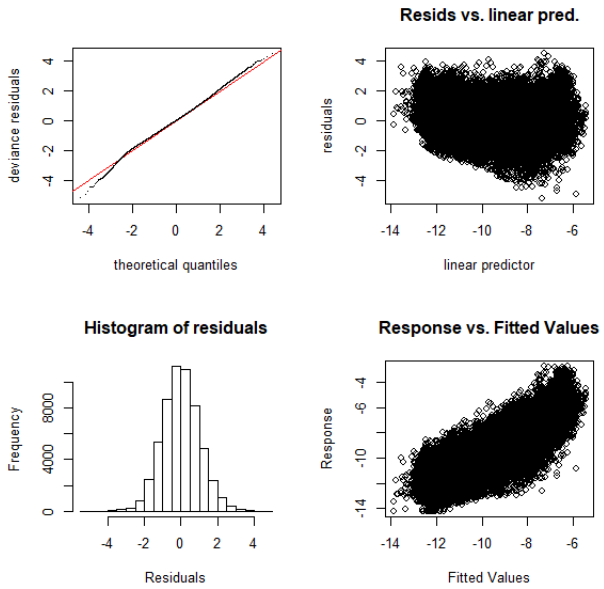

Figura 3.10: Validación del modelo aditivo.

Con este modelo vemos un buen ajuste de los residuos. Se distribuyen sin ningún patrón claro por los distintos valores del predictor lineal y los valores ajustados por el modelo son similares a los observados. Según estas gráficas este modelo podría ser adecuado.

En la Tabla 3.2 aparece el resumen de dicho modelo. Vemos que la distancia debe modelarse con una función de suavizado y tendrá casi 9 grados de libertad efectivos. El modelo incluirá una constante y el logaritmo de las unidades vendidas en la tienda final. Este modelo explica el 62.4 % de la variación y en la Figura 3.11 se muestra el efecto no lineal de la distancia.

	Estimación	Error estándar	p-valor
Constante	-18.49	0.074	<2e-16
log (Unidades destino)	0.63	0.006	<2e-16
Approximate significance of smooth terms:	edf	Red.df	p-valor
s(distancia)	8.97	9	<2e-16

Deviance explained=62.4 %

Tabla 3.2: Resumen del modelo aditivo.

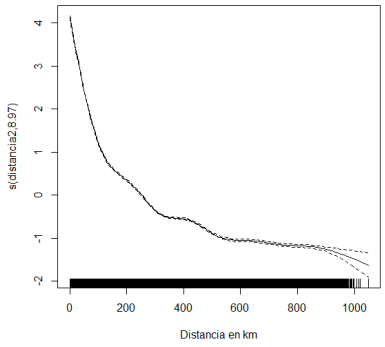

Figura 3.11: Función de suavizado.

Una vez que hemos llegado al modelo final resulta interesante estudiar los residuos para conocer las tiendas con un comportamiento más atípico. Para ello sumamos, en valor absoluto, los residuos de cada tienda de origen. La tienda con una suma de residuos más alta es la que se encuentra en la Calle Gambó de Benidorm. Haciendo lo mismo pero considerando las tienda de destino obtenemos que la tienda de Madrid situada en la calle Ourense tendría un comportamiento más atípico. En la Figura 3.12 se muestra el histograma de los residuos de estas dos tiendas. En ambas se ve un predominio de los residuos positivos.

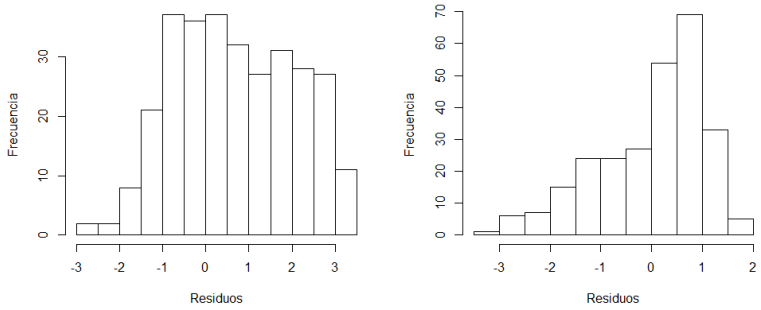

Figura 3.12: Histograma de los residuos de las tiendas con un comportamiento más atípico. A la izquierda con origen en Benidorm y a la derecha con destino de Madrid.

Capítulo 4

Aplicación Shiny

Las aplicaciones web permiten presentar los resultados de manera clara y atractiva. El uso de aplicaciones web ha aumentado en los últimos años debido, en parte, a que al almacenar los datos y archivos en la nube no necesitan ser instaladas en el ordenador.

Shiny es una herramienta con la que podemos crear fácilmente aplicaciones web que permiten a los usuarios interactuar con los datos sin tener que manipular el código. La ventaja de shiny es que no es necesario tener conocimiento de HTML o Jacascrip, solo se necesita conocer R.

Las aplicaciones Shiny constan de dos partes:

- La interfaz de usuario (ui) que controla el diseño y aspecto de la aplicación.
- El servidor (server) que contiene las instrucciones que el equipo necesita para construir la aplicación.

Ambas componentes pueden correrse de manera simultánea gracias a la función **shinyApp(ui, server)**.

Por petición de la empresa, para presentar los resultados de este análisis se desarrollaron dos aplicaciones Shiny. Cada una de ellas consta de una zona de control, desde donde se pueden modificar algunos parámetros de los datos, y otra en la que se muestran los resultados.

En la aplicación para datos lattice (Figura 4.1) se representa el mapa de España con el porcentaje de devolución por colores, el número de clases en que dividimos el intervalo podemos elegirlo en la zona de control. También presenta el resultado del análisis de autocorrelación, tanto global como local, y el modelo final. Esta aplicación permite elegir al usuario el criterio de vecindad, de pesos y si queremos ajustar a un modelo SAR o CAR. También permite seleccionar las variables que introduciremos en el modelo entre la edad, el número de tiendas, el importe vendido y la tasa de paro de cada provincia.

En la aplicación para datos en red (Figura 4.2) se representa la red por el método de Fruchterman-Reingold, el histograma de las unidades devueltas y un

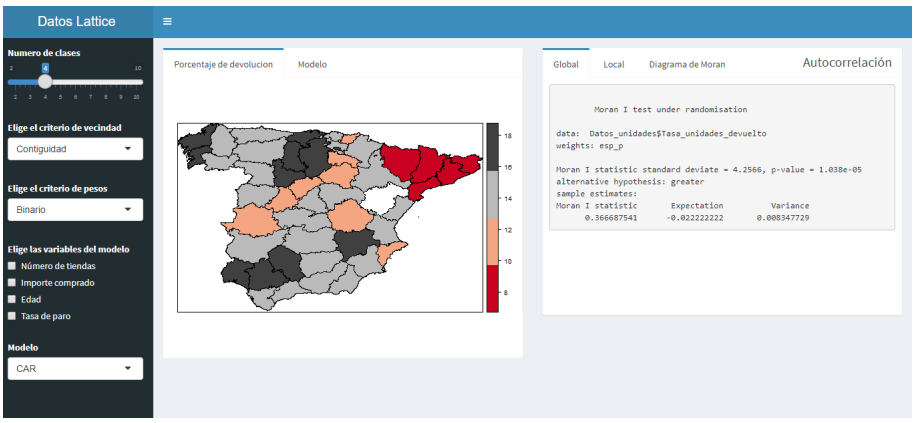

Figura 4.1: Aplicación Shiny para datos latice.

breve resumen de las características más importantes de la red, pudiendo elegir el número mínimo de prendas entre dos tiendas para que sean consideradas y el criterio para elegir a la tienda más central. Además en una caja con tres pestañas se ve el nombre de las tiendas más centrales donde cada pestaña es un tipo de centralidad. También se puede seleccionar si se quieren los análisis para toda la red o para la componente conexa. Por último se muestra el ajuste del modelo y sus gráficos de validación.

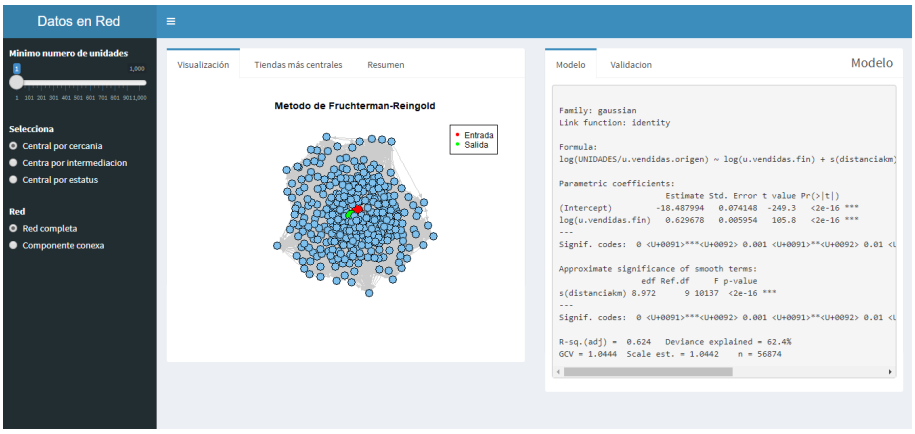

Figura 4.2: Aplicación Shiny para datos en red.

Capítulo 5

Conclusiones

A través de este trabajo se ha planteado la metodología adecuada para llevar a cabo un análisis de datos espaciales. Se han desarrollado las herramientas adecuadas para ello y se han aplicado a los datos de devoluciones facilitados por Inditex.

En la primera parte de este trabajo se ha planteado el problema bajo la perspectiva *latice*. Se ha encontrado autocorrelación global significativa, lo que quiere decir que la localización sí influye en el porcentaje de prensas devueltas.

Con el análisis local hemos detectado en el noreste peninsular un cluster de provincias con autocorrelación espacial positiva. Este cluster lo forman Huesca y las cuatro provincias catalanas, en ellas el porcentaje de devolución es menor que en el resto de España. Además en el centro norte peninsular detectamos otro cluster, pero, en este caso, de provincias con un alto porcentaje de devolución. Detectar este cluster dependerá en mayor medida de los criterios de vecindad y pesos considerados. Las provincias que lo forman serán Cantabria, Palencia, Burgos y Valladolid. En el resto de zonas de la península no se observa ningún otro patrón destacable.

A la hora de aplicar un modelo que explique este comportamiento no llegamos a ninguno que elimine la autocorrelación espacial de los residuos. Esto puede deberse a una de las limitaciones del estudio planteado, las covariables. Estamos intentando explicar un comportamiento complejo con variables relativamente sencillas, como lo son la edad media, el número de tiendas y la tasa de paro. Puede ser que no se estén considerando variables explicativas relevantes. Para conocer estas variables habría que realizar un estudio en profundidad de los aspectos económicos y demográficos de cada provincia, estudio que queda fuera de los objetivos de este trabajo.

La heterogeneidad de las regiones puede ser otro motivo del mal ajuste de dichos modelos. Habría que tener en cuenta características de cada provincia que pudieran hacer que la misma variable no tuviera el mismo efecto en todas ellas.

La conclusión a la que podemos llegar es que el comportamiento en la de-

volución de prendas sí depende de la posición espacial. Así, podría estudiarse aumentar el número de prendas enviadas a la zona noreste peninsular, ya que lo que se recibe de devoluciones es una cantidad pequeña en comparación con las ventas. En cambio, en la zona centro-norte podría plantearse una disminución del envío de mercancías, ya que una parte de las ventas podría cubrirse con las prendas devueltas.

A continuación estudiamos cada tienda por separado, considerándo a cada una de ellas como un nodo de una red. No se tendrán en cuenta las devoluciones realizadas dentro de la misma tienda ya que estas devoluciones no afectan al stock.

Con el estudio descriptivo detectamos que entre dos tiendas de Zara de Santander se produce el mayor flujo de devoluciones, 18750. Además, aunque el número de pares de tiendas entre las que existe alguna prenda devuelta es de casi 60000, no podemos olvidarnos que en la mayor parte el número de prendas devuelto es pequeño, tan solo el 30 % superan las ocho prendas. La red completa es conexa, pero deja de serlo si solo consideramos las devoluciones de más de 61 unidades.

Con este análisis detectamos que la tienda del Centro Comercial Parque Sur de Leganés es la tienda que más distribuye sus prendas de ropa, el 90 % de las prendas de España reciben devoluciones de prendas compradas en este centro comercial. Esta tienda será la más central por cercanía y por intermediación y, además, es el nodo de llegada de una de las aristas más centrales.

Otra tienda que destaca en este análisis es la tienda situada en la Calle Gambó de Benidorm. A esta tienda llegan devoluciones del 87 % de las tiendas de la península. También es una de las tiendas más centrales por cercanía. También hemos visto que no tienen ningún punto de articulación y que para desconectar el gráfico sería necesario eliminar 34 nodos.

En cuanto al ajuste de un modelo gravitacional vemos que hay variables que presentan un efecto significativo. El modelo tiende a subestimar las unidades devueltas y como en el caso del análisis *lattice* habría que buscar variables explicativas que nos proporcionaran información relevante para poder obtener mejores estimaciones.

Bibliografía

[1] Akaike H (1974) A new look at the statistical model identification. IEEE Transaction on Automatic Control 19(6): 716–723.

[2] Anselin L (1980) Estimation methods for spatial autoregressive structures. Tesis, Cornell University.

[3] Anselin L(1988) Spatial econometrics: methods and models. Springer.

[4] Anselin L (1993) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. Regional Research Institute, West Virginia University.

[5] Anselin L (1995) Local Indicators of Spatial Association-LISA. Geographical Analysis 27(2):93-115

[6] Anselin L (2001) Spatial econometrics. En Badi H. Baltagi (ed) A companion to theoretical econometrics. Blackwell Publishing, Oxford, pp 310-330.

[7] Besag J (1974) Spatial and the statistical analysis of lattice systems. Journal of the Royal Statistical Society Series B 36:192–225.

[8] Bivand R, Piras G (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software 63(18):1-36.

[9] Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2018). shiny: WebApplication Framework for R. R package version 1.1.0. https://CRAN.R-project.org/package=shiny. Accedido 19 de Abril de 2018.

[10] Cliff AD, Ord JK (1981) Spatial Processes.Pion, London.

[11] Cottrell David, Herron Michael C, Westwook Sean J (2016) Evaluating Donald Trump's allegations of voter fraud in the 2016 presidential election. Dartmouth Collegue. (INTERNET)

[12] Cox KR (1969) The voting decision in a spatial context. Progress in Geography 1:81-117.

[13] Cressie, N. A. C. (1993) Statistics for Spatial Data. J. Wiley.

[14] Delaunay B (1934) Sur la sphere vide. A la mémoire de Georges Voronoi. Bulletin de l' Académie des Sciences de l ' URSS. Classe des sciences mathématiques et na, no. 6:793–800.

[15] Euler L (1735) Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientarum Petropolitanae 8:128-140.

[16] Faraway JJ (2006) Extending The Linear Model With R. Generalized Linear, Mixed Effects And Nonparametric Regression Models. Chapman and Hall/CRC.

[17] Freeman L (1977) A set of measures of centrality based upon betweenness. Sociometry 40(1): 35-41.

[18] Geary R (1954) The contiguity ratio and statistical mapping. The Incorporated Statistician 5:115-45.

[19] Getis A, Ord J (1992) The analysis of spatial association by use of distance statistics. Geographical Analysis 24:189-206.

[20] Gomez Martínez M, García Rubio N y Alfaro Corté E. Una aplicación de la estadística espacial al comportamiento de la vivienda de alquiler en España. Departamento de Economía y Empresa, Universidad de Castilla-La Mancha.

[21] Hijmans RJ (2017). geosphere: Spherical Trigonometry. R package version 1.5-7. https://CRAN.R-project.org/package=geosphere. Accedido 12 de Junio de 2018.

[22] Kleiber C, Zeileis A (2008) Applied Econometrics with R. Springer-Verlag, New York. https://CRAN.R-project.org/package=AER. Accedido 19 de Febreso de 2018.

[23] Kolaczyk Eric D, Csárdi G (2014), Statistical analysis of network data with R. Springer, Londres.

[24] König D (1936) Theorie der endlichen und unendlichen Graphen. Akademische Verlagsgesellschaft, Leipzig.

[25] Matheron G (1962) Traité de géostatistique appliquée. Mémoires du Bureau de Recherches Géologiques et Minières. Editions Technip, Paris.

[26] Mantel N (1967) The Detection of Disease Clustering and a Generalized Regression Approach, American Association for Cancer Research 27:209-220.

[27] Martin S, Brown WM, Klavans R, Boyack KW (2008) DrL: Distributed Recursive (Graph) Layout. SAND Reports 2936:2-10

[28] Matheron G (1963) Principles of Geostatistics. Economic Geology 58:1246-1266

[29] Moran P (1948) The interpretation of statistical maps. Journal of the Royal Statistical Society Series B 10:243-251.

[30] Nuñez O, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Locutura JF, López-Abente G (2016) Arsenic and chromium topsoil levels and cancer mortality in Spain. Environmental Science and Pollution Research 23(17):17644-75.

[31] Ord K (1975) Estimation methods for models of spatial interaction. Journal of the American Statistical Association 70:120–126.

[32] Pompa García M, Hernández González P (2012) Determinación de la tendencia espacial de los puntos de calor como estrategia para monitorear los incendios forestales en Durango, México. Bosque (Valdivia) 33(1):63-68

[33] Sabidussi G (1966) The centrality index of a graph. Psychometrika 31(4): 581-603.

[34] Sanén N, Quintana L (2010), Crecimiento económico, convergencia y concentración económica espacial en las entidades federativas de México 1970-2008. Investigaciones Regionales 18:83-106.

[35] Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman & Hall, London. CAPITULO!!

[36] Tiefelsdorf M, Griffith DA, Boots B (1999) A variance-stabilizing coding scheme for spatial link matrices. Environment and Planning A 31:165–180.

[37] Tobler W (1970) A Computer Movie Simulation Urban Growth in the Detroit Region. Economic Geography 46(2):234-240.

[38] Wood SN (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC, New York.