Índice

Capítulo 6

Análisis de Datos Espaciales agregados o regionales

6.1. Introducción

En ocasiones, los datos espaciales observados son *áreas* o *zonas*, datos que hemos denominados en el título del capítulo como datos agregados o regionales. Estas áreas serán habitualmente las unidades de investigación, es decir, los datos observados, los cuales deben de tener límites bien definidos. No obstante, por su propia definición, varias áreas pueden unirse y formar un nuevo "dato" por lo que es de gran interés analizar si existe *autocorrelación espacial* entre varias unidades ya que, en muchas ocasiones, lo que pase en una zona está relacionado con lo que pase en la zona adyacente. Este problema se conoce como *problema de Galton* que consiste básicamente en establecer cuántas observaciones (zonas) independientes hay en la muestra cuando se han utilizado límites arbitrarios en la definición de las áreas en estudio. Por esta razón, la primera sección de este capítulo se dedica a estudiar cómo definir pesos a las áreas o zonas consideradas.

6.2. Entornos y pesos de Áreas

Asignar pesos a las zonas en estudio es necesario si queremos realizar un Análisis de datos espaciales supuesto que éstos sean Áreas, fundamentalmente con el propósito de conseguir residuos totalmente independientes (i.e., sin autocorrelación espacial) cosa que será necesaria cuando hagamos el ajuste de un modelo a este tipo de datos.

La determinación de las zonas (o clusters) a considerar ya es un problema en sí mismo aunque supondremos que las zonas están ya definidas. Después, a la hora de fijar pesos a esas zonas se recomienda asignar un peso igual a 1 a

las zonas limítrofes y un peso igual a 0 a las zonas que no son limítrofes a una dada. Es decir, si una zona A tiene sólo una zona como vecina, el peso de A será 1; si A tiene dos zonas vecinas, su peso será 2. Esta forma de fijar pesos se denomina *binaria* y, con ella, si una zona tiene a su alrededor 2 zonas, su pero será doble que el de otra zona con sólo una zona limítrofe.

La forma natural de contar cuántos vecinos tiene una zona es unir los centroides de las zonas. El número de conexiones entre los centroides nos dará su peso.

La forma binaria de asignar pesos hará que algunas áreas tengan peso 2 y otras pesos, por ejemplo, 7. Otra forma de asignar pesos es una forma *ponderada*, en donde los pesos de cada área son estandarizados de forma que todos los pesos sumen 1.

La función `nb2listw` de la librería de R, `spdep` es la que calcula los pesos de las zonas de las dos maneras antes mencionadas, las cuales se indican con el argumento `style`, asignando el valor `W` en el caso de la forma ponderada y asignando el valor `B` cuando lo hace de forma binaria.

En todo caso, representaremos por w_{ij} al *peso espacial* de la unión entre el individuo i-ésimo de la matriz de datos (fila i-ésima) y el individuo j-ésimo en donde se habrá medido la variable de interés, obteniendo respectivamente los valores y_i e y_j; es decir, por y_i representaremos el valos de Z en $\mathbf{s}_i$, es decir, $Z(\mathbf{s}_i)$.

6.3. Contraste global de autocorrelación espacial: Estadístico I de Moran

El estadístico (índice) I de Moran se define como el cociente del producto de la variable de interés y su retardo (*lag*) de la forma

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

Este índice toma un valor comprendido entre -1 y 1. Si es $I = 0$ interpretamos que los datos están distribucidos al azar (del estilo del CSR que vimos); si es positivo habrá concentración y, si toma un valor negativo, entendemos que hay una dispersión mayor de la que tendríamos si los datos se distribuyeran al azar.

El valor esperado del índice I, bajo la hipótesis nula de ausencia de autocorrelación espacial es $E[I] = -1/(n-1)$. Habitualmente se calculan los valores de este índice, de su media, de su varianza, de la desviación estándar $(I - E(I))/\sqrt{Var(I)}$, así como del p-valor del test de la hipótesis nula de

ausencia de autocorrelación espacial.

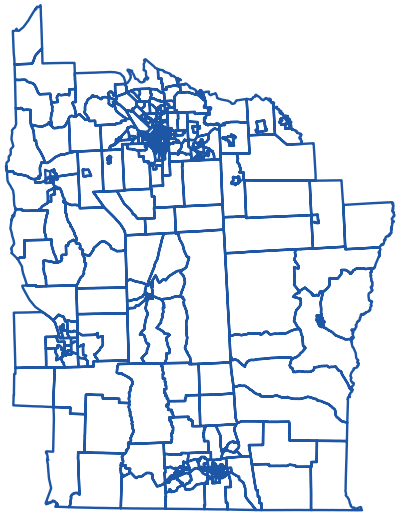

Figura 6.1 : Mapa del estado de Nueva York

Ejemplo 6.1

Supongamos que tenemos interés en estudiar el estado americano de Nueva York estado compuesto por varios condados. En el directorio `d:/datos` tenemos los 3 ficheros asociados a un GIS (el de extensión shp, el de extensión shx y el de extensión dbf), los tres con el nombre `NY` (datos basados en Waller y Gotway, 2004 y Bivand, et al., 2013).
Esta matriz de datos sobre casos de leucemia en este estado está formado por 281 individuos y 12 variables. Es la que aparece en el fichero dbf antes mencionado.
Como hicimos en el Capítulo 4, incorporamos estos datos a R mediante la función `readOGR` ejecutando `(1)`. Si queremos, podemos representar el mapa ejecutando `(2)` en donde hemos elegido un color azul utilizando el argumento `border=4` y un grosor 2 con el argumento `lwd=2` obteniendo así la Figura 6.1.
Si queremos utilizar QGIS podemos importar directamente el fichero `NY.shp` sabiendo que estamos en coordenadas geográficas UTM zona 18N. La representación sería la Figura 6.2.
Con `(3)` incorporamos a R las zonas del estado de Nueva York y con `(4)` las unimos creando la Figura 6.3.
Con `(5)` asignamos los pesos a esas zonas eligiendo aquí la forma binaria.

```
> library(spdep)
> library(rgdal)
> NY<-readOGR("d:/datos","NY")                                    (1)
> plot(NY,border=4,lwd=2)                                         (2)
```

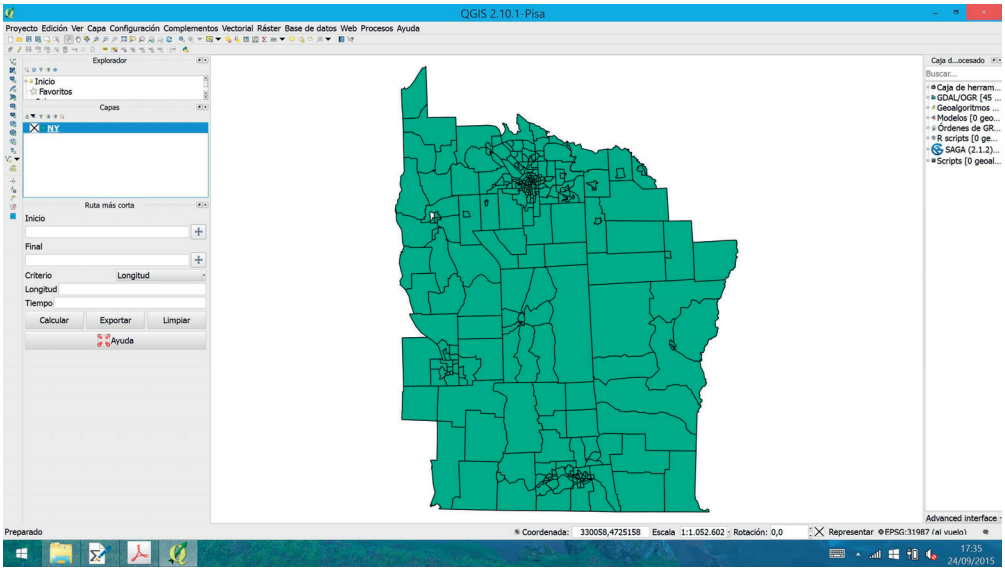

Figura 6.2 : GIS del problema con QGIS

```
> NYzonas<-read.gal("d:/datos/NY.gal",region.id=row.names(NY))                (3)
> plot(NYzonas,coordinates(NY),pch=16,cex=0.5,add=TRUE)                       (4)

> pesoszonas<-nb2listw(NYzonas,style="B")                                     (5)

> moran.test(NY$Casos,listw=pesoszonas)                                       (6)

        Moran's I test under randomisation

data:  NY$Casos
weights: pesoszonas

Moran I statistic standard deviate = 3.1862, p-value = 0.0007207            (7)
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.110387402      -0.003571429       0.001279217
```

Entre los datos del censo está el número de casos de leucemia observados. Si queremos contrastar que estos casos se producen al azar, es decir no dependiendo del lugar (ausencia de correlación espacial) se puede ejecutar el test global de Moran. Este test sobre la hipótesis nula de ausencia de autocorrelación espacial es calculado ejecutando `(6)`, test cuyo p-valor se da en `(7)`, el cual es suficientemente pequeño como para rechazar la ausencia de autocorrelación espacial y concluir con la existencia de dicha relación. Es decir, hay una concentración espacial mayor de la cabría esperar si los casos se repartieran al azar en todo el estado.

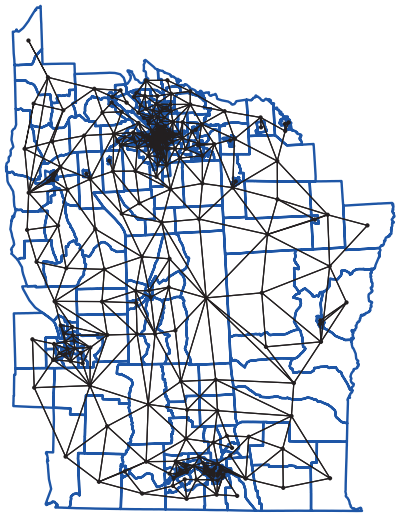

Figura 6.3 : Mapa del estado de Nueva York con zonas unidas

6.4. Contraste local de autocorrelación espacial: Gráfico de dispersión de Moran

El test de Moran estudiado más arriba es un test global de autocorrelación espacial. El valor obtenido con este test global se puede dividir para conseguir tests locales que permitan detectar clusters en donde las observaciones sean similares a las de su entorno así como detectar outliers locales, también denominados *puntos calientes* o *hotspots*.

Comencemos estudiando el *Gráfico de dispersión (scatterplot) de Moran*. Este gráfico es un gráfico de dispersión en donde aparecen en el eje de abscisas los valores de la variable de interés y en el eje de ordenadas esos mismos valores retardados espacialmente, lo que representa a sus entornos.

Por tanto, este gráfico recoge el grado de asociación espacial de cada observación con su entorno y se divide en cuatro cuadrantes que expresan ese grado de asociación. Los cuatro cuadrantes son pares de valores de tipo (bajo,bajo), (alto,alto), (bajo,alto) y (alto,bajo). Los dos cuadrantes que recogen a los dos

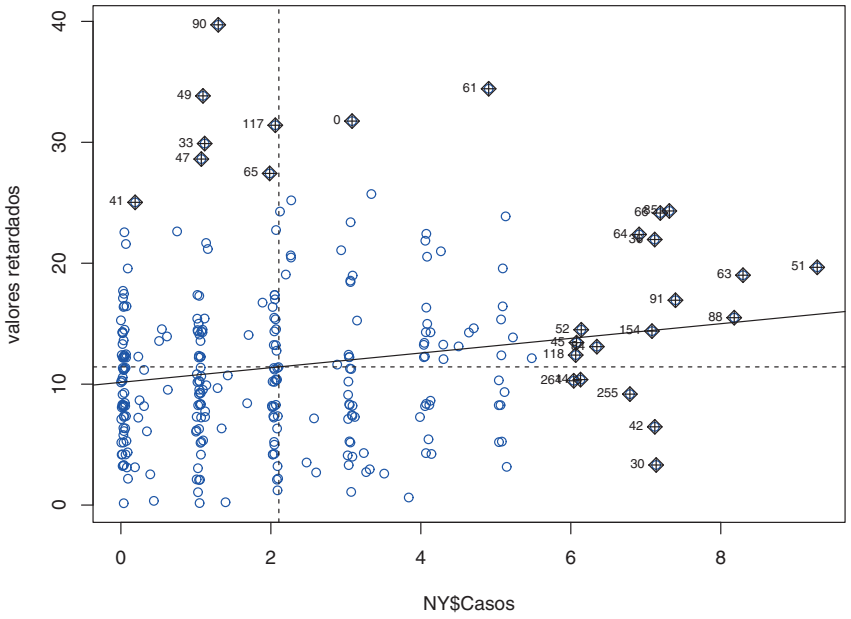

Figura 6.4 : Scatterplot de Moran

últimos tipos de datos, son valores anómalos espacialmente pues presentan poca correlación espacial con las observaciones de su entorno.

En R este gráfico es algo distinto ya que se añade un recta con pendiente igual al índice de Moran I tratando de expresar de esta forma una relación lineal con correlación el índice I de manera que se aprecian los puntos del gráfico que influyen en la *recta de regresión* así construida. Estos son sospechosos de autocorrelación espacial (*puntos calientes*).

Si definimos los *índices locales de Moran* como

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

se pueden hacer tests de zonas locales y contrastar esos puntos sospechosos.

Podemos escribir que es

$$\frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \sum_{i=1}^{n} I_i$$

con lo que, dado que el valor $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ es una constante de normalización, podemos decir que con los índices locales de Moran descomponemos el índice

global.

Ejemplo 6.1 (continuación)

El *scatterplot* de Moran se obtiene ejecutado `(1)`, obteniendo así Figura 6.4 en donde aparecen numerados los *puntos calientes*.
Los tests locales se ejecutan con `(2)`. Los p-valores aparecen en la última columna.

```
> moran.plot(NY$Casos,listw=nb2listw(NYzonas,style="B"),col=4,
+ ylab="valores retardados")                                          (1)

> localmoran(NY$Casos, listw = nb2listw(NYzonas,style="B"))           (2)
               Ii         E.Ii     Var.Ii         Z.Ii    Pr(z > 0)
0    3.6835835873 -0.028571429  7.9485633  1.316684690 9.397217e-02
1    3.9539648449 -0.021428571  5.9606995  1.628289043 5.173181e-02
2   -2.0568022902 -0.010714286  2.9787125 -1.185523086 8.820947e-01
.....................................................................
90  -4.6810881998 -0.028571429  7.9485633 -1.650226776 9.505517e-01
.....................................................................
278 -0.4781876360 -0.014285714  3.9727337 -0.232745582 5.920205e-01
279  0.1852043194 -0.014285714  3.9727337  0.100086725 4.601377e-01
280  0.0512836166 -0.021428571  5.9606995  0.029782325 4.881203e-01
```

6.5. Ajuste de Modelos

En la mayoría de datos espaciales no habrá independencia, es decir, presentarán autocorrelación espacial. Una forma habitual de modelizar este problema es la de suponer que nuestras observaciones multivariantes $\mathbf{Y}$ (o $\mathbf{Z}$ si seguimos la notación anterior de este capítulo) se pueden expresar en función de covariables de la forma

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + e$$

en donde e una variable de error con distribución normal multivariante de vector de medias cero y matriz de varianzas-covarianzas $\mathbf{V}$, aunque esta modelización podrá variar según el modelo considerado como por ejemplo los modelos Autorregresivos, similares a los utilizados en series temporales, capaces de recoger la dependencia espacial de la matriz $\mathbf{V}$

Nosotros, no obstante, nos decantamos por utilizar alguno de los modelos estudiados en los siguientes capítulos. Como muestra, vamos a utilizar una regresión lineal en el ejemplo que hemos tratado en este capítulo.

Ejemplo 6.1 (continuación)

Continuando con el ejemplo del estado americano de Nueva York, vamos a modelizar, en lugar de los valores observados Y_i los valores

$$Z_i = \log \frac{Y_i + 1}{n_i}$$

que ya están en la base de datos bajo el nombre de

```
NY$Z
```

Vamos a considerar como covariables `PEXPOSURE` (*Distancia inversa al Tricloroetileno más cercano*), `PCTAGE65P` (*Proporción de personas mayores de 65 años*) y `PCTOWNHOME`, (*Proporción de personas dueñas de su casa*).
Si ajustamos una regresión lineal múltiple ejecutando `(1)`, vemos con `(2)` y `(3)` que los residuos presentan autocorrelación espacial ya que el p-valor dado en `(4)` rechaza la hipótesis nula de ausencia de autocorrelación espacial.
Suele utilizarse mejor la sentencia `(5)`.

```
> recta <- lm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY)          (1)
> NY$residuos <- residuals(recta)                                         (2)

> library(spdep)
> moran.test(NY$residuos,list=pesoszonas)                                 (3)

        Moran's I test under randomisation

data:  NY$residuos
weights: pesoszonas

Moran I statistic standard deviate = 2.4457, p-value = 0.007229          (4)
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.083090278      -0.003571429       0.001255603

> lm.morantest(recta,list=pesoszonas)                                     (5)

        Global Moran's I for regression residuals

data:
model: lm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY)
weights: pesoszonas

Moran I statistic standard deviate = 2.638, p-value = 0.004169
alternative hypothesis: greater
sample estimates:
Observed Moran's I        Expectation           Variance
       0.083090278       -0.009891282        0.001242320
```

Observamos en `(6)` que la covariable `PEXPOSURE` no es significativa por lo que la quitamos y repetimos el análisis.