

Índice

1. Introducción al QGIS	13
1.1. Introducción	13
1.2. Sistemas de Información Geográfica	13
1.2.1. Utilidad de los Sistemas de Información Geográfica . . .	16
1.2.2. Aplicaciones de los Sistemas de Información Geográfica	16
1.2.3. Sistemas de Información Geográfica más utilizados . . .	17
1.3. Instalación de QGIS	18
1.3.1. Descripción del área de trabajo	18
1.4. Tipos de datos GIS	19
1.4.1. GIS vectorial	20
1.4.2. Ejemplo de QGIS vectorial	23
1.4.3. GIS raster	26
1.4.4. Ejemplo de QGIS raster	31
2. Utilización y Manejo de QGIS	43
2.1. Introducción	43
2.2. Incorporación de Tablas de Datos	43
2.3. Selección Espacial	49
2.4. Análisis Espacial de Proximidad	58
2.5. Presentación e Impresión	64
3. Interacción entre QGIS y R	69
3.1. Introducción	69
3.2. Configuración de QGIS	70
3.3. Ejecución de programas de R a través de QGIS	73
3.4. Ejecución de SAGA a través de QGIS	76
4. Análisis de Datos Espaciales de tipo discreto. Procesos Puntuales	79
4.1. Introducción. Tipos de datos espaciales	79
4.2. Datos espaciales y su representación	81

4.3.	Procesos Puntuales Espaciales	83
4.3.1.	Análisis de la distribución espacial	85
4.3.2.	Aleatoriedad Espacial Completa (<i>CSR</i>)	89
4.3.3.	Ajuste de Modelos Espaciales Puntuales	94
4.3.4.	Análisis de la densidad espacial	105
5.	Análisis de Datos Espaciales de tipo continuo. Geoes-	
tadística		109
5.1.	Introducción	109
5.2.	Variograma	110
5.2.1.	Interpretación del Variograma	111
5.2.2.	Modelos de Variograma	114
5.2.3.	Estimación clásica del Variograma	117
5.2.4.	Utilización de covariables	118
5.2.5.	Estimación clásica del Variograma con R	118
5.2.6.	Nube Variograma	120
5.2.7.	Estimación del Modelo Variograma	123
5.3.	Interpolación espacial	126
5.3.1.	Kriging	128
5.4.	Variograma multivariante	145
6.	Análisis de Datos Espaciales agregados o regionales	153
6.1.	Introducción	153
6.2.	Entornos y pesos de Áreas	153
6.3.	Contraste global de autocorrelación espacial: Estadístico I de Moran	154
6.4.	Contraste local de autocorrelación espacial: Gráfico de dispersión de Moran	157
6.5.	Ajuste de Modelos	159
7.	Modelos Lineales Generalizados GLM	165
7.1.	Introducción	165
7.2.	Definición de Modelo Lineal Generalizado univariante	167
7.3.	Estimación y Contrastes basados en la verosimilitud	172
7.3.1.	Estimador de máxima verosimilitud de los β_i	173
7.3.2.	Estimador del parámetro de escala ξ	175
7.3.3.	Contrastes de hipótesis sobre los parámetros	175
7.3.4.	Contraste de bondad de ajuste del modelo	177
7.3.5.	Diagnóstico del Modelo	177
7.4.	Cálculo con R	178
7.4.1.	Regresión Logística y Regresión Binomial	179

Interpretación de los coeficientes del Modelo de Regre-	
sión Logística ajustado	185
Dispersión excesiva (<i>Overdispersion</i>)	190
7.4.2. Regresión Logística Multinomial	192
7.4.3. Regresión Poisson	193
7.5. Métodos basados en la cuasi-verosimilitud	197
7.6. Métodos Bayesianos	198
7.7. Métodos robustos	199
7.7.1. <i>M</i> -estimadores basados en la cuasi-verosimilitud	199
7.7.2. Contraste robusto de bondad de ajuste del modelo . . .	201
7.7.3. Cálculo con R^{mo}	202
7.8. Ajuste de modelos GLM para datos espaciales	208
8. Modelos Aditivos Generalizados GAM	211
8.1. Introducción	211
8.2. Modelos GAM clásicos	213
8.2.1. Estimación	213
8.2.2. Validación Cruzada (<i>Cross validation</i>)	215
8.2.3. Cálculo con R	217
8.3. Modelos GAM robustos	220
9. Robustez en el Análisis de Datos Espaciales	229
9.1. Introducción	229
9.2. Estimador robusto del Variograma	230
9.3. Detección de outliers locales	236
10. Análisis de Formas	241
10.1. Introducción	241
10.2. Análisis Morfométrico Clásico desde un punto de vista Descriptivo	242
10.2.1. Eliminación del Efecto Tamaño	244
10.2.2. Eliminación de la Localización por Traslación	246
10.2.3. Eliminación de la Orientación por Rotación	246
10.2.4. Más de dos Configuraciones (Análisis Procrustes Gene-	
ralizado)	248
10.2.5. Proyección de la Configuración en el Espacio Tangente .	249
10.3. Análisis Morfométrico Robusto desde un punto de vista Des-	
criptivo	249
10.3.1. Eliminación del Efecto Tamaño de manera robusta . . .	250
10.3.2. Eliminación la Localización de manera robusta	254
10.3.3. Más de dos Configuraciones	255
10.4. Análisis Morfométrico Clásico desde un punto de vista Inferencial	256
10.5. Análisis Morfométrico Robusto desde un punto de vista Inferencial	258

10.5.1. Aproximación von Mises para el p-valor del Estadístico Procrustes	258
10.5.2. Aproximación Saddlepoint para el p-valor del estadístico Procrustes	260
10.6. Aplicaciones	264
11. Análisis Espacio-Temporal de datos	271
11.1. Introducción	271
11.2. Visualización de datos espacio-temporales	272
11.2.1. Representaciones de tipo descriptivo	272
11.2.2. Visualización de datos espacio-temporales mediante ani- mación	274
11.3. Estimadores espacio-temporales marginales	276
11.4. Variograma espacio-temporal	278
11.4.1. Datos en formato STFDF	280
11.4.2. Estimador del variograma espacio-temporal	283
11.4.3. Modelos de Variograma espacio-temporal	286
11.5. Modelos de Regresión Espacio-Temporales	289
12. Problemas Resueltos	291
13. Bibliografía	325

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 denominado 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 peso 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 valor de Z en s_i , es decir, $Z(s_i)$.

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

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

Este índice toma un valor comprendido entre -1 y 1 . Si es $I = 0$ interpretamos que los datos están distribuidos 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.

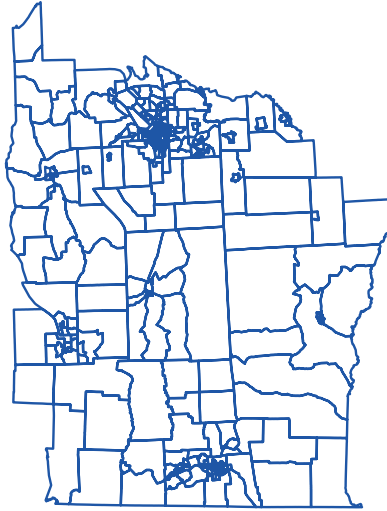


Figura 6.1 : Mapa del estado de Nueva York

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)
```

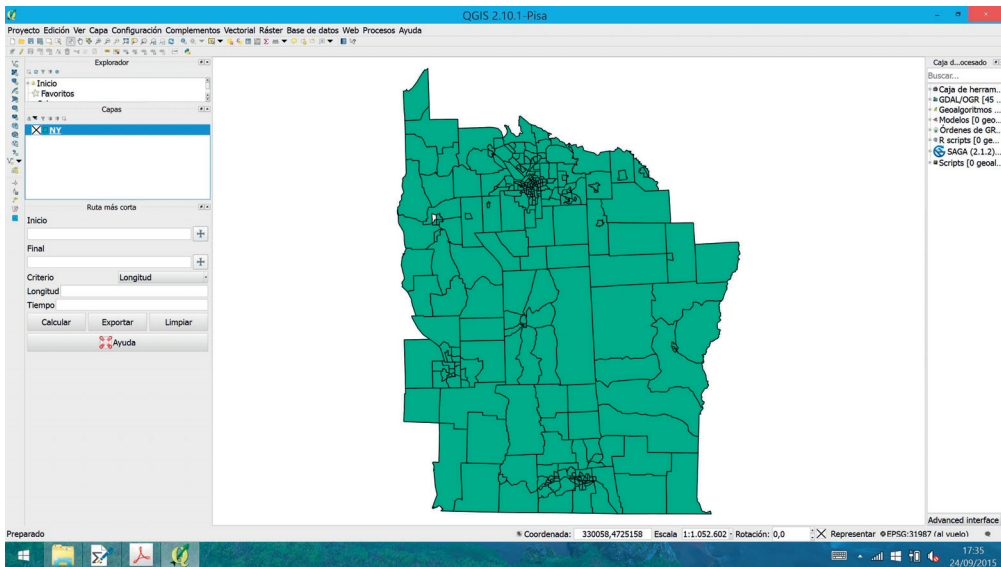


Figura 6.2 : GIS del problema con QGIS

```
> 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.

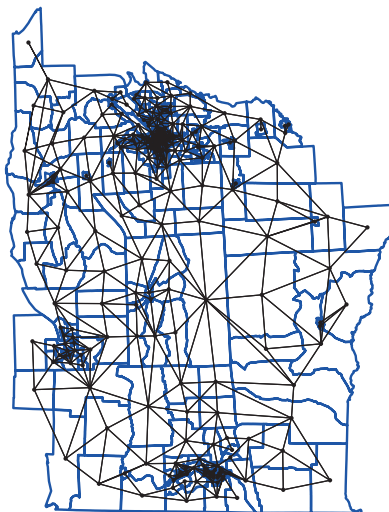


Figura 6.3 : Mapa del estado de Nueva York con zonas unidas

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

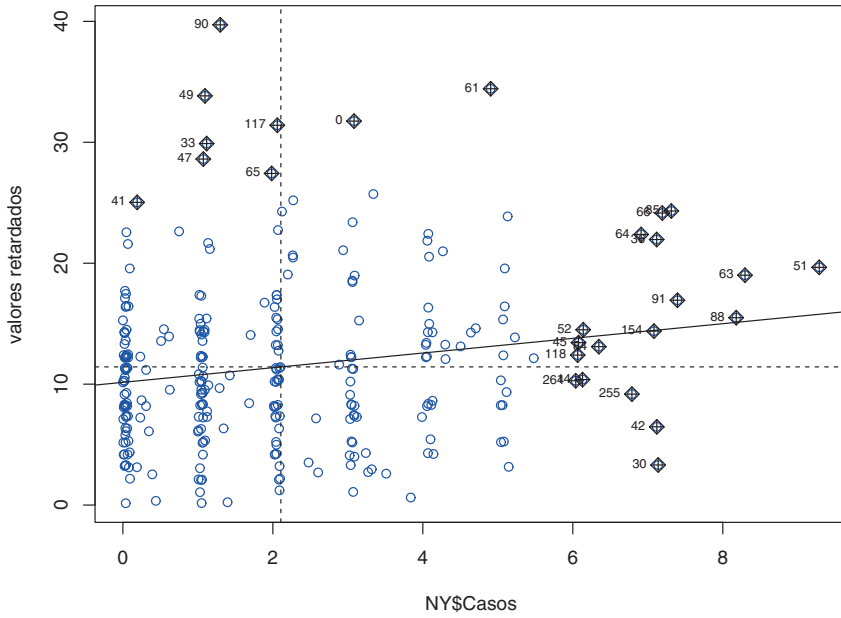


Figura 6.4 : Scatterplot de Moran

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 una 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

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

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

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

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.