Parte VI

Datos en áreas

15. Introducción

En esta sexta parte introduciremos la teoría para procesos en los que el conjunto de índices $D \subset \mathbb{R}^d$ es un conjunto numerable de localizaciones espaciales fijas (no aleatorias). En este contexto resulta totalmente imposible observa la variable de interés entre dos localizaciones del conjunto D, situación que sí contempla el enfoque de la Geoestadística. Referencias teóricas importantes para datos en áreas:

- Cressie, N. (1993). Statistics for spatial data. Wiley, New York.
- Waller, L.A. y Gotway, C.A. (2004). Applied spatial statistics for public health data. Wiley, New Jersey.

Los datos en áreas suelen observarse asociados a polígonos con contornos definidos, lo cual justifica su nombre.
En la figura 12 se representa un ejemplo de este tipo de fenómeno aleatorio. En este caso el conjunto de índices viene dado por 100 localizaciones correspondientes a las capitales de los municipios de Carolina del Norte, Estados Unidos, y la variable de interés $Z(x) = \{Z(x_1), \ldots, Z(x_{100})\}$ recoge el número de casos de muerte súbita infantil registrados durante los años 1974-78 y 1979-84.

$$D = \{x_1, x_2, \ldots, x_{100}\} \subset \mathbb{R}^2$$

En el caso de conjuntos de índices continuos (caso de la Geoestadística) la topología del espacio está completamente especificada mediante el concepto de distancia métrica, como puede ser la distancia euclídea. Sin embargo, en el caso de conjutos de índices discretos (caso de los datos en áreas) la topología debe ser especificada mediante el concepto de "vecindario de un área" y de "distancia entre áreas vecinas".

El estudio de estos datos puede consultarse en Cressie (1993), páginas 385-402.

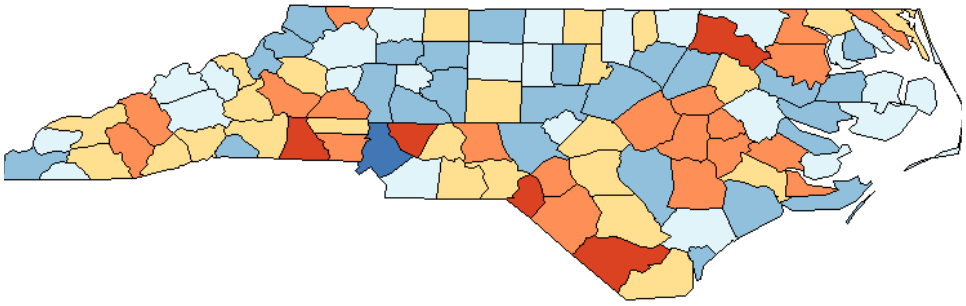

Figura 12: Mapa de los municipios de Carolina del Norte, Estados Unidos, utilizado en el estudio del número de casos de muerte súbita infantil.

Lo primero que hay que entender es que para un área determinada, identificada por una localización x_i, no todas las áreas se consideran vecinas de ella. Dicho de otra forma, cada área establece una partición en el conjunto de índices entre miembros y no miembros de su vecindario.
Por otro lado, el concepto de "distancia entre áreas vecinas" se representa asignando pesos positivos a las áreas pertenecientes a un vecindario.

1. Vecindario del área x_i. Es un conjunto N_i de áreas verificando alguna condición. Por definición $x_i \notin N_i$.

   Ejemplos de vecindarios aplicables al ejemplo ilustrado en la figura anterior :

   - N_i es el conjunto de municipios cuya capital dista de x_i menos de 300 kilómetros.
   - N_i es el conjunto de municipios colindantes al de x_i.
   - N_i es el conjunto de todos los municipios, excepto x_i.

   En la figura 13 se muestra las relaciones entre los municipios de Carolina del Norte si se considera que un vecindario está formado por los municipios colindantes.

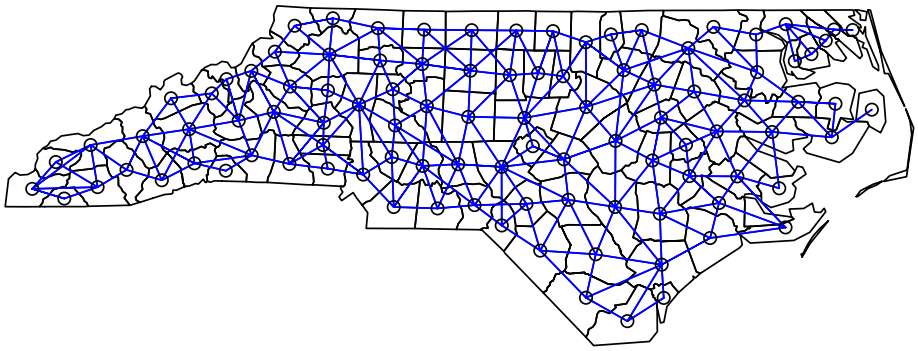

Figura 13: Vecindarios de áreas, considerando la relación de contigüidad

2. Matrices de distacias, o matrices de pesos. Habitualmente son simétricas (pueden no serlo) y representan la conexión entre áreas. $W = (w_{i,j})$, donde $w_{i,j}$ indica la relación entre x_i y x_j. Si $w_{i,j} = 0$, entonces $x_j \notin N_i$, esto es: el área representada por x_j no se considera vecina del área representada por x_i. Si $w_{i,j} > 0$, entonces $x_j \in N_i$, esto es el área representada por x_j está en el vecindario del área representada por x_i.

   La diagonal de la matriz W está formada por ceros: $w_{i,i} = 0$.

   Ejemplos de matrices de distancias pueden ser:

   - W matriz binaria (ceros y unos).
   - W depende de la longitud de la frontera común entre municipios.
   - W recoge distancias euclídeas entre capitales de municipios: los pesos son inversamente proporcionales a la distancia.
   - W recoge un coste de transporte entre municipios: Los pesos son inversamente proporcionales al coste. En este caso la matriz W podría ser no simétrica.

   En algunas ocasiones las matrices de distancias deben ser determinadas por el ususario (posible falta de objetividad), lo que condiciona tanto la dependencia espacial de las áreas como el análisis posterior de los datos.

16. Autocorrelación espacial

Dada una variable espacial $Z(x) = \{Z(x_1), \ldots, Z(x_n)\}$ y una matiz de pesos W asociada interesa conocer el grado de correlación entre las áreas. Esto suele cuantificarse mediante el cálculo de un índice.

- Si la variable $Z(x)$ no tiene tendencia o su tendencia es constante el índice se calcula con las observaciones $\{z(x_1), \ldots, z(x_n)\}$.
- Si $Z(x) = m(x) + Y(x)$, con $m(x) \neq cte$ entonces la variable espacial que puede presentar correlación es $Y(x)$. En este caso hay que estimar $m(x)$ para obtener una aproximación de la variable espacial $Y(x)$ y calcular el índice con esta última.

En esta sección supondremos, sin pérdida de generalidad, que la tendencia es nula o constante.

El índice más utilizado para el cálculo de la autocorrelación espacial con datos en áreas es el denominado índice I de Moran:

$$I = \frac{1}{s^2} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \left(Z(x_i) - \overline{Z}\right) \left(Z(x_j) - \overline{Z}\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}},$$

donde

$$\overline{Z} = \frac{1}{N} \sum_{i=1}^{n} Z(x_i), \quad s^2 = \frac{1}{N} \sum_{i=1}^{n} \left(Z(x_i) - \overline{Z}\right)^2$$

I es una variable aleatoria que, por construcción, depende tanto de la variable de interés $Z(x) = \{Z(x_1), \ldots, Z(x_n)\}$ como de la matriz de pesos W. Una realización de la variable I se obtiene introduciendo las observaciones $\{z(x_1), \ldots, z(x_n)\}$ en la expresión anterior.

Habitualmente $|I| \leq 1$, pero pueden observarse valores $|I| > 1$ si los valores extremos de $z(x_i) - \overline{z}$ vienen acompañados de pesos $\sum_{j=1}^{n} w_{i,j}$ muy elevados.

Este índice recuerda en su construcción al coeficiente de correlación de Pearson y, al igual que éste último, la interpretación suele llevarse a cabo en dos etapas:

1. Interpretación del valor obtenido por el índice I de Moran:

   Si los valores de áreas vecinas son similares obtendremos un valor alto (positivo) del índice mientras que si los valores de áreas vecinas son discordantes, por ejemplo

valores de $Z(x_i)$ altos vienen acompañados de valores $Z(x_j)$ bajos, con $x_j \in N_i$, I tenderá a ser negativo.

Cuando las áreas no presentan correlación I tomará un valor próximo a $-1/(n-1)$.

2. Significación del valor obtenido por el índice I de Moran:

   Cuando no hay correlación, el índice I de Moran sigue una distribución normal $N(\mu_I, \sigma_I)$ con $\mu_I = -1/(n-1)$ y

$$\sigma_I^2 = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{(n-1)(n+1) S_0^2} - \frac{1}{(n-1)^2},$$

   donde

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}, \quad S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{i,j} + w_{j,i})^2, \quad S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{i,j} + \sum_{j=1}^{n} w_{j,i}\right)^2.$$

   Entonces podemos contrastar la hipótesis nula H_0 : Las áreas no presentan correlación espacial, frente a su alternativa comprobando si el valor observado de I está en las colas de la distribución normal. Dicho de otra forma, rechazamos H_0 si $|I - \mu_I| / \sigma_I \geq z_{\alpha/2}$

Ejemplo: La función *moran.test* de la librería *spdep* de *R* calcula el índice de Moran para datos de áreas. Aplicando esta función al caso, mencionado anteriormente, del número de casos de muerte súbita infantil registrados en n =100 municipios de Carolina del Norte se obtienen los siguientes valores:

Indice I de Moran = 0.163740904, μ_I=-0.010101010, σ_I^2= 0.004015196

La matriz W utilizada para estos cálculos es la representada en la figura 13.
La interpretación del valor de I indica cierta concordancia, aunque baja, entre las observaciones. Para ver si este valor es significativo calculamos $|I - \mu_I| / \sigma_I = 2.7435$, cuyo valor p en un contraste bilateral es $2 \times P(Z \geq 2.7435) = 0.00608$

Si tomamos $\alpha = 0.05$ tendremos que concluir que las áreas presentan correlación significativa, es decir rechazamos la hipótesis nula de que las áreas no presentan correlación espacial.