17. Modelos espaciales

En esta sección se presentan los modelos teóricos espaciales más ampliamente utilizados. Estos modelos están basados en la teoría de series de tiempo en donde la observación de un instante dado está correlacionada con las de los instantes pasados. En los datos en áreas la observación de un área dada está correlacionada con las de las áreas de su vecindario.

17.1. Modelos espaciales gausianos

Dada la variable $Z(x) = m(x) + Y(x)$, supongamos que:

- $E[Z(x)] = m(x) = \sum_{l=0}^{L} a_l f^l(x)$, donde f^l son funciones conocidas y a_l son parámetros reales desconocidos.

  Suele suponerse $f^0 = 1$, con lo que se asegura que el caso de media constante está incluido en el modelo.

- $Y(x)$ sigue una distribución gausiana multidimensional $N(0, \Sigma)$, donde los elementos de la matriz Σ son

$$\Sigma_{i,j} = Cov(Z(x_i), Z(x_j)) = Cov(Y(x_i), Y(x_j)).$$

17.1.1. Modelos espaciales gausianos simultáneamente autorregresivos (SAR)

Es similar al modelo de autorregresión de las series de tiempo, pero en el contexto de datos espaciales. Se trata entonces de modelos autorregresivos espaciales.

$$Z(x_i) = m(x_i) + \sum_{j=1}^{n} b_{i,j}[Z(x_j) - m(x_j)] + \epsilon_i;\ i = 1, 2, \ldots, n.$$

Las variables $\epsilon_i;\ i = 1, 2, \ldots, n$ son independientes y distribuidas según una $N(0, \sigma_i)$.
Los coeficientes $b_{i,j}$ están relacionados con la matriz W que define las distancias y el vencindario. Se verifica entonces que $b_{i,i} = 0$ y $b_{i,j} \neq 0$ si $x_j \in N_i$.
La expresión anterior puede ponerse en forma matricial:

$$(I - B)(Z - m) = \epsilon,$$

donde

$$\begin{aligned} B &= (b_{i,j})\,; i,j = 1,2,\ldots,n, \\ Z &= (Z(x_1), Z(x_2), \ldots, Z(x_n))', \\ m &= (m(x_1), m(x_2), \ldots, m(x_n))', \\ \epsilon &= (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)' \sim N(0, \Lambda), \\ \Lambda &= \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}, \end{aligned}$$

I denota la matriz identidad de dimensión n y $'$ representa traspuesto.
Unas sencillas operaciones matriciales nos llevan a deducir que

$$Var(Z) = (I - B)^{-1} \Lambda I (I - B')^{-1}$$

Puesto que estamos bajo hipótesis de normalidad, los parámetros $B\ a_l; l = 0, 1, \ldots, L$ y $\sigma_i; i = 1, 2, \ldots, n$ del modelo se obtienen maximizando una función de verosimilitud:

$$(2\pi)^{-n/2} |\Lambda|^{-1/2} |I - B| \exp\left\{ -\frac{1}{2} (Z - m)' (I - B') \Lambda^{-1} (I - B) (Z - m) \right\}$$

Una observación importante es que $Cov(\epsilon, Z) = \Lambda (I - B')^{-1}$ no es una matriz diagonal, lo que quiere decir que los errores, ϵ, no son independientes de las variables, Z, al contrario de lo que ocurre en los modelos autorregresivos de las series de tiempo. Como consecuencia de esto, los estimadores por mínimos cuadrados de los parámetros pueden ser no consistentes.

Ejemplo: La función *spautolm* de la librería *spdep* de R estima mediante el método de máxima verosimilitud los parámetros de un modelo SAR. Para asegurar la existencia de solución, esta función asume la siguiente relación entre matrices: $B = \lambda W$, con W la matriz de pesos. También supone homocedasticidad en los errores: $\Lambda = \sigma I$.
Utilizaremos esta función a los datos de muerte súbita infantil. Puesto que los modelos SAR se formulan bajo hipótesis de normalidad debemos comenzar comprobando si podemos asumir esta hipótesis. El test no paramétrico de Shapiro-Wilk calcula un valor

$p = 2.080 \times 10^{-11}$, con lo cual debemos rechazar la hipótesis de que los datos siguen una distribución normal.

Con la función *boxcox.fit* podemos estimar los parámetros necesarios para llevar a cabo una transformación de Box-Cox que, para el caso de los datos de muerte súbita es la siguiente:

$$\tilde{Z}(x) = \frac{[Z(x) + 0.0009500]^{0.3195008} - 1}{0.3195008}$$

Con el test de Shapiro-Wilk obtenemos ahora un valor $p = 0.2161$ y aceptamos la normalidad de los datos de la variable transformada. Como habíamos asumido tendencia constante, el modelo a estimar es ahora:

$$\tilde{Z}(x_i) = a_0 + \lambda \sum_{j=1}^{n} w_{i,j} \left[\tilde{Z}(x_j) - a_0\right] + \epsilon_i;\ i = 1, 2, \ldots, n,$$

con tres parámetros desconocidos. La función *spautolm* calcula las siguientes estimaciones:

- $\hat{a}_0 = 3.24960$, con valor $p = 7.805 \times 10^{-13}$ (Rechazamos $H_0 : a_0 = 0$)
- $\hat{\lambda} = 0.50775$, con valor $p = 1.0857 \times 10^{-5}$ (Rechazamos $H_0 : \lambda = 0$). Que este parámetro no sea nulo confirma que, efectivamente, existe correlación entre las áreas.
- $\hat{\sigma} = 2.2327$.

La validez del modelo estimado se puede chequear mediante el cálculo del índice I de Moran sobre los residuos $\left\{\tilde{z}(x_i) - \hat{\tilde{z}}(x_i);\ i = 1, 2, \ldots, n\right\}$, donde

$$\hat{\tilde{z}}(x_i) = \hat{a}_0 + \hat{\lambda} \sum_{j=1}^{n} w_{i,j} \left[\tilde{z}(x_j) - \hat{a}_0\right];\ i = 1, 2, \ldots, n,$$

Indice I de Moran = -0.030435645, μ_I=-0.010101010, σ_I^2= 0.004301033

Para ver si este valor es significativo calculamos $(I - \mu_I)/\sigma_I$ = -0.3101, con valor p = 0.7565 por lo que aceptamos que los residuos no presentan correlación espacial y se asume que esta correlación ha sido recogida por el modelo SAR estimado.

Continuando con el estudio de los residuos realizamos un test de normalidad, obteniendo ahora un valor $p = 0.7588$.

17.1.2. Modelos espaciales gausianos condicionalmente autorregresivos (CAR)

En este caso el modelo autorregresivo se expresa mediante esperanzas condicionadas:

$$E\left[Z\left(x_i\right)\left|_{Z(x_j),j\neq i}\right.\right]=m\left(x_i\right)+\sum_{j=1}^{n}c_{i,j}\left[Z\left(x_j\right)-m\left(x_j\right)\right];\ i=1,2,\ldots,n.$$

Bajo normalidad éste es el mejor predictor, desde el punto de vista del error cuadrático medio de predicción, de $Z\left(x_i\right)$ basado en $\left\{Z\left(x_j\right),j\neq i\right\}$.

Nuevamente la matriz de coeficientes $C=\left(c_{i,j}\right);i,j=1,2,\ldots,n$ está relacionada con la matriz W que define las distancias y el vencindario por lo que que $c_{i,i}=0$ y $c_{i,j}\neq 0$ si $x_j\in N_i$.

Con este modelo es necesario asumir ciertas hipótesis para garantizar que exista la distribución normal multidimensional. Son las siguientes:

1. $(I-C)$ es invertible, con I la matriz identidad de dimensión n.
2. $(I-C)^{-1}M$ es simétrica y definida positiva, con $M=dig\left(Var\left(Z\left(x_i\right)\left|_{Z(x_j)j\neq i}\right.\right)\right)$.

Entonces

$$Z\sim N\left(m,\left(I-C\right)^{-1}M\right).$$

Puesto que estamos bajo hipótesis de normalidad, los parámetros $a_l,l=0,1,\ldots,L$ y las matrices C y M del modelo se estiman maximizando la correspondiente función de verosimilitud:

$$\left(2\pi\right)^{-n/2}\left|M\right|^{-1/2}\left|I-C\right|^{1/2}\exp\left\{-\frac{1}{2}\left(Z-m\right)'M^{-1}\left(I-C\right)\left(Z-m\right)\right\}.$$

Ejemplo: Ajustaremos ahora un modelo CAR a los datos transformados $\tilde{Z}\left(x\right)$:

$$E\left[\tilde{Z}\left(x_i\right)\left|_{\tilde{Z}(x_j),j\neq i}\right.\right]=a_0+\sum_{j=1}^{n}c_{i,j}\left[\tilde{Z}\left(x_j\right)-a_0\right];\ i=1,2,\ldots,n.$$

La función *spautolm* de la librería *spdep* de R estima mediante el método de máxima verosimilitud los parámetros de un modelo CAR asumiendo que $C=\lambda W$ y $M=\sigma I$ obteniendo las siguientes estimaciones:

- $\hat{a}_0 = 3.24960$, con valor $p = 3.745 \times 10^{-6}$ (Rechazamos $H_0 : a_0 = 0$)
- $\hat{\lambda} = 0.85051$, con valor $p = 3.8527 \times 10^{-6}$ (Rechazamos $H_0 : \lambda = 0$). Con lo que volvemos a confirmar la existencia de correlación.
- $\hat{\sigma} = 2.1473$

Es importante señalar que, aunque estamos obteniendo valores similares a los calculados para el modelo SAR esto no tendría por qué ser así ya que estamos ajustando modelos diferentes.

Al igual que en el ejemplo anterior, el estudio de los residuos nos indica que siguen una distribución normal (valor $p = 0.8222$) y no presenentan correlación espacial (valor $p = 0.9998$).

Para responder a la pregunta de cuál de los dos modelos, el SAR o el CAR, es el que mejor ajusta los datos utilizados en estos ejemplos podemos utilizar el criterio de información de Akraike (AIC) calculado también por la función *spautolm*.

- Modelo SAR: AIC = 456.92
- Modelo CAR: AIC = 454.94

El AIC se calcula como una suma ponderada de la función de verosimilitud más el número de parámetros que se han estimado (para más detalles sobre este criterio es preferible consultar Waller y Gotway (2004)). Los mejores modelos son aquellos que alcanzan un menor valor de AIC por lo que, aunque por muy poco margen, podríamos elegir el modelo CAR como el que proporciona el mejor ajuste.

17.1.3. Equivalencia entre ambos modelos

Asumiendo que la media m ha sido modelizada de forma correcta, entonces los dos modelos son equivalentes si y sólo si

$$(I - C)^{-1} M = (I - B)^{-1} \Lambda (I - B')^{-1}$$

En cualquier caso puede demostrarse que siempre un modelos espacial gausiano SAR puede representarse como un modelo espacial gausiano CAR. El recíproco no tiene por qué verificarse.