Universidad Politécnica de Cartagena
E.T.S. de Ingeniería de Telecomunicación

Espacio-Tele□o n^{0}2 (2011)
Revista de la ETSIT-UPCT

Estadística espacial

Pilar Sanmartín Fita[1]

1. Departamento de Matematica Aplicada y Estadística.
Universidad Politécnica de Cartagena.

pilar.sanmartin@upct.es

Resumen

En este artículo se muestra una breve introducción sobre la estadística espacial y sus aplicaciones en algunos campos como es el caso de la epidemiología y el medio ambiente.

Proyecto/Grupo de investigación: Métodos Bayesianos Objetivos en Salud Pública y Medio Ambiente. Entidad financiadora: Ministerio de Educación y Ciencia. Código: MTM2007-61554.

Líneas de investigación: *Estadística Espacial*; *Procesos Estocásticos* ; *Campos aleatorios Markovianos*; *Inferencia Bayesiana*; *Modelos gráficos*

1. Introducción

Existen numerosas situaciones donde los datos objetivo de análisis presentan una estructura espacial (y/o temporal). La figura 1 ilustra alguna de estas situaciones. Un mapa de la costa NE de USA en los puntos en que se midió el número de capturas pesqueras, un mapa donde se establece mediante diferentes niveles de color el número de casos de muerte súbita infantil (SID) en Estados Unidos de 1974 a 1978 y la distribución de dos especies de árboles en un area de bosque (ver [5] y [19]).

En estudios epidemiológicos también es frecuentemente que los datos objeto de estudio sean observados en distintas localizaciones espaciales a lo largo de un determinado periodo temporal. En este contexto surge la necesidad de un análisis espacial a la vez que temporal para entender la evolución de las

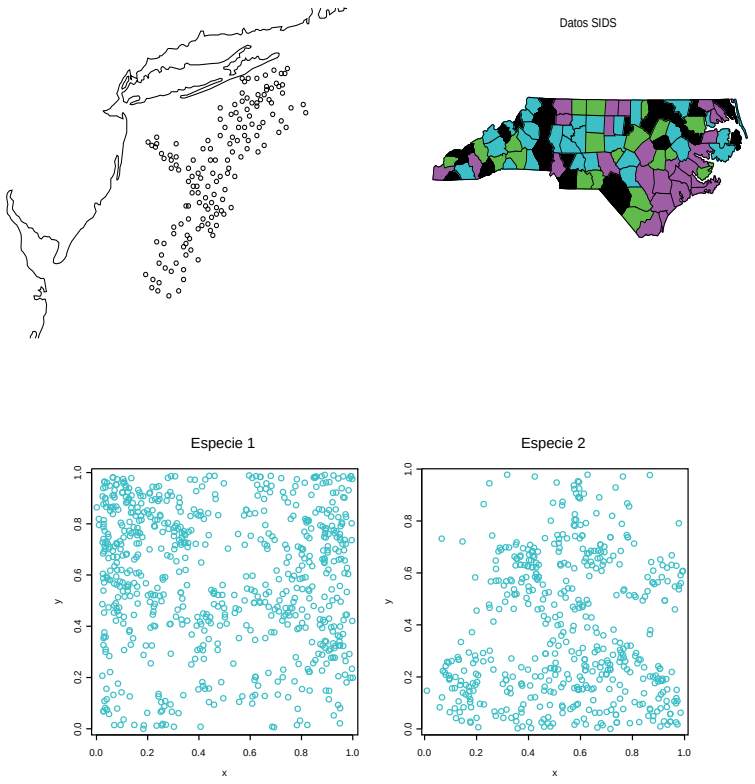

Figura 1: Algunos ejemplos.

interdependencias espaciales a lo largo del tiempo. En la figura 2 se muestra la serie temporal de la mortalidad por meningitis en España de 1950 a 1990 y los valores recogidos en las provincias de Álava y Guipúzcoa respectivamente (ver [6]). Los datos están agregados a nivel provincial, se pretende estudiar la evolución espacio-temporal de la enfermedad y y la influencia de los movimientos migratorios en la aparición de los brotes epidémicos.

En este trabajo introducimos brevemente el modelo espacial general para centrarnos posteriormente en datos observados en retículos, en ese contexto estudiamos los campos aleatorios markovianos y posteriormente analizamos los modelos jerárquicos y su aplicación a la epidemiología y el medio ambiente. Terminamos con algunos ejemplos de extensión al caso espacio-temporal.

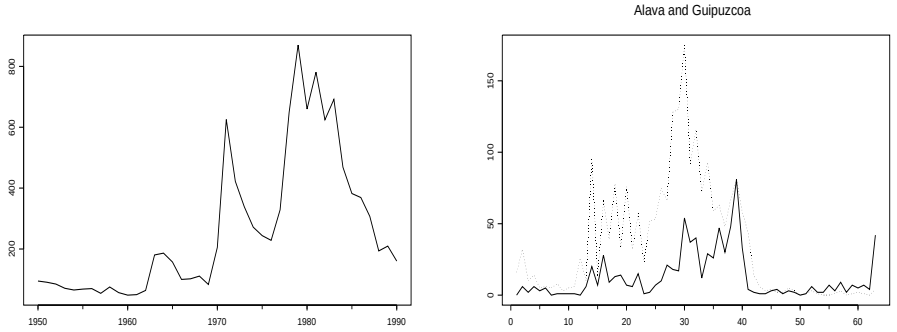

Figura 2: mortalidad por meningitis en España 1950-1991.

1.1. El modelo Espacial General

El modelo espacial general se puede establecer en los siguientes términos, consideremos $\mathbf{s} \in R^d$ una localización en el espacio euclideo d-dimensional y $\mathbf{Z}(\mathbf{s})$ una cantidad aleatoria observada en $\mathbf{s}$. Al variar $\mathbf{s}$ sobre un conjunto de índices $D \subset R^d$, tenemos el proceso aleatorio $\{\mathbf{Z}(\mathbf{s}) : \mathbf{s} \in D \subset R^d\}$ y $\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in D \subset R^d\}$ una realización de dicho proceso. Según sea el conjunto de índices D, los datos objetivo de análisis se pueden considerar de tres tipos, *datos Geoestadísticos*, si D es un "continuo"(fijo), *Datos es Retículo*, si D es un conjunto numerable (fijo) y *procesos puntuales* si D es aleatorio. En este artículo nos centraremos en el caso de análisis de datos en retículo y su aplicación dentro del ámbito de la epidemiología y el medio ambiente (Para un estudio del resto de casos y otras posibles aplicaciones ver ([5])). También cabe comentar que todas estas técnicas de análisis espacial han tenido un gran desarrollo gracias también a la ayuda de programas informáticos tales como el R ([19]).

2. Campos aleatorios Markovianos.

Consideramos una colección numerable de localizaciones espaciales (regular o irregular), de forma que $D \equiv \{(x_i, y_i) \equiv i : i = 1, \ldots, N\} \equiv \mathcal{N}$. Para este retículo se establece un sistema de vecindades δ, $\delta_i = \{j : j \text{ vecino de } i\}$ $i = 1, \ldots, N$ basado en algún criterio tal como en distancias entre localizaciones, contigüidad, etc... de forma que $\{N, \delta\}$ se puede considerar equivalente a un Grafo no dirigido (UDG) (ver [6], [3] y [7]).

Una forma de abordar el estudio de datos espaciales en retículos es a través de distribuciones de Gibbs (ver [5]) y en particular de los llamados automodelos de Besag (1974) (ver [1]). Las distribuciones de Gibbs (equivalentemente campos aleatorios Markovianos) definen la distribución conjunta a partir de

las distribuciones condicionales:

$$p(z_i|\boldsymbol{z}_{-i}) \tag{1}$$

donde $\boldsymbol{z}_{-i}$ es el subvector de observaciones en todas las localizaciones salvo la i-ésima y p es la función de densidad con respecto a la medida considerada (de conteo o de Lebesgue) . A partir de estas distribuciones condicionales se considera la relación de vecindad:

$$i \sim j \Leftrightarrow (1) \text{ involucra a } z_j$$

Si δ_i es el conjunto de localizaciones vecinas de la localización i. Un conglomerado se define como un conjunto de localizaciones formado por una sola localización i o por un conjunto de localizaciones que son todas vecinas entre sí. Denotaremos por $\mathcal{C}$ el conjunto de conglomerados para un conjunto de localizaciones $\mathcal{L}$. Suponiendo que el conjunto soporte de la distribución de $\mathbf{y}$ coincide con el producto de los conjuntos soportes de cada una de las distribuciones de z_i $i \in \mathcal{L}$ (condición de positividad):

$$p(\mathbf{z}) \propto \exp(Q(\mathbf{z})) \tag{2}$$

$Q(.)$ recibe el nombre de función negpotencial y por el teorema de Hammersley-Clifford (ver [8])):

$$\exp(Q(\mathbf{z})) = \sum_{C \in \mathcal{C}} V_C(\mathbf{z}_C) \prod_{i \in C} z_i \tag{3}$$

siendo $V_C(\mathbf{z}_C)$ una función que depende de $\mathbf{z}$ únicamente a través de los valores en el subvector $\mathbf{z}_C$ y que llamaremos función de interacción. Los automodelos de Besag ([1]) son un caso particular de distribuciones de Gibbs caracterizados por el hecho de que las distribuciones condicionales (1) pertenecen a la llamada familia exponencial (que engloban la distribución binomial, Poisson y Normal o Gaussiana como casos particulares) y $V_C(.)$ es nula $\forall$ C con $|C| > 2$. En este caso:

$$p(z_i|\boldsymbol{z}_{-i}) \propto \exp\{(\alpha_i + \sum_{j \in \delta_i} \beta_{ij} z_j) + h_i(z_i)\} \tag{4}$$

en la expresión (4) β_{ij} son parámetros de interacción espacial y α_i es un parámetro específico de cada localización, en el que puede recogerse la posible influencia de un vector de covariables $\boldsymbol{x}_i \in \mathbb{R}$ observadas, mediante:

$$\alpha_i = \boldsymbol{x}_i^t \vec{\alpha}_i \tag{5}$$

con lo cuál tendríamos un modelo lineal generalizado espacial (ver por ejemplo [11], y [12]). Si consideramos dentro de la familia exponencial la distribución de Poisson, el modelo auto-Poisson quedaría:

$$p(z_i|\boldsymbol{z}_{-i}) = exp(-\lambda_i(\mathbf{z}_{-i}))\lambda_i(\mathbf{z}_{-i})^{z_i}/z_i!$$

$$log(\lambda_i(\mathbf{z}_{-i})) = \alpha_i + \sum_{j\in\delta_i} \beta_{ij} z_j$$

$$Q(\mathbf{z}) = \sum_i (\alpha_i + \sum_{j\in\delta_i} \beta_{ij} z_j) z_i - \sum_i \log(z_i!)$$

En este caso, hay que suponer $\beta_{ij} < 0$ (modelos de inhibición) para garantizar sumabilidad o considerar una 'Poisson truncada' para interacciones positivas (ver [11]). En [11] se analiza un modelo de Poisson aplicado al análisis de mortalidad por cáncer en Valencia.

Cuando usamos la distribución Gaussiana para modelizar las distribuciones condicionales tenemos el modelo auto-Gaussiano condicionalmente especificado (CAR):

$$z_i|\boldsymbol{z}_{-i} \sim Gau(\mu_i + \sum_{j\in\delta_i} \beta_{ij}(z_j - \mu_j), \sigma_i^2)$$

con $\beta_{ij}\sigma_j^2 = \beta_{ji}\sigma_i^2$, $\beta_{ii} = 0$ y $\beta_{ij} = 0$ si $j \notin \delta_i$.

$$\mathbf{Z} \sim Gau(\mu, (I - B)^{-1} diag(\sigma_i^2))$$

La matriz B tiene ceros en su diagonal y el resto son los coeficientes β_{ij}. $diag(\sigma_i^2)$ es una matriz diagonal con elementos σ_i^2. Para el caso en que se quieran introducir covariables, hablaríamos de modelos de Regresión Condicionalmente Gaussianos (CARX) (ver [5] y [17]):

$$z_i|\boldsymbol{z}_{-i} \sim N(\boldsymbol{x}_i^t \vec{\alpha}_i + \sum_{j\in\delta_i} \beta_{ij}(z_j - \boldsymbol{x}_j^t \vec{\alpha}_j), \sigma_i^2) \tag{6}$$

de donde se deduce la distribución conjunta:

$$\boldsymbol{z} \sim N(X\boldsymbol{\alpha}, (I - B)^{-1} diag(\sigma_i^2)) \tag{7}$$

Las columnas de la matriz X corresponden a las covariables incluidas en el modelo y $\boldsymbol{\alpha} = (\vec{\alpha}_1^t, \ldots, \vec{\alpha}_N^t)^t$.

Para estimar estos modelos, una estimación basada en la función de verosimilitud

$$p(\mathbf{z}|\theta) = \frac{\exp Q(\mathbf{z}|\theta)}{\int \exp Q(\mathbf{z}|\theta) d\mathbf{z}} = K(\theta) \exp Q(\mathbf{z}|\theta)$$

(donde θ denota los parametros a estimar) plantea el problema de que la constante $K(\theta)$ no tiene expresión cerrada y es complicado en general (ver por ejemplo [9]). Una alternativa es la estimación basada en la pseudo-verosimilitud

$$p(\mathbf{z}|\theta) = \prod_{i=1}^{N} p(z_i|z_{-i}, \theta)$$

.

En el caso de los modelos gaussianos la estimación por el método de máxima verosimilitud se puede realizar de forma iterativa de la siguiente forma, si consideramos el logaritmo de la verosimilitud:

$$L(\alpha, \sigma^2, B) \quad = (N/2)log(2\pi\sigma^2) - (1/2)log(|I - B|)) +$$

$$(1/2)(z - X\beta)'(I - B)(z - X\beta)/\sigma^2$$

- para B fijo, resolviendo el problema de mínimos cuadrados generalizados

$$\hat{\alpha} = (X'(I - B)X)^{-1}X'(I - B)\mathbf{z}$$

- para B fijo, optimizando respecto de σ^2

$$\hat{\sigma}^2 = \frac{1}{N}(\mathbf{z} - X\hat{\alpha})'(I - C)(\mathbf{z} - X\hat{\alpha})$$

- finalmente, maximizando la log-verosimilitud *perfil* de C

$$\hat{B} = \text{argmax}\, L(\hat{\alpha}, B, \hat{\sigma}^2)$$

3. Modelos jerárquicos

Una alternativa a los modelos antes considerados consiste en introducir la estructura espacial en un segundo nivel de jerarquía, es decir,

- $\mathbf{z}_i|\theta$ Independientes
- $\theta \sim$ estructura espacial

Estos modelos han tenido un desarrollo notable, especialmente en la elaboración de mapas de riesgo, gracias a la disponibilidad de métodos de estimación MCMC (Monte Carlo Markov Chain) y programas para su implementación tales como el WinBUGS ([10], [18]). Para el caso en que los datos sigan una distribución de Poisson,

- $\mathbf{z}_i|\theta \sim Poiss(exp(\theta))$ Independientes
- $\theta \sim$ estructura espacial+ covariables.

La estructura espacial de θ se suele modelizar a través de estructuras gaussianas, como el modelo CAR [4] o el conocido como Modelo Auto-regresivo Intrínseco ([2]),

$$\theta_i = u_i + v_i + covariables$$

$$\mathbf{u} \sim Gau(\mathbf{0}, \sigma^2\mathbf{I})$$

$$\mathbf{v} \sim Gau(\mathbf{0}, \tau^2\mathbf{W}^{-1})$$

$$w_{ii} = -\sum_{i \neq j} w_{ij}$$

Como ya avanzabamos este tipo de modelos se usan frecuentemente en la elaboración de mapas de riesgo cuando se representan datos de mortalidad/ morbilidad en epidemiología (ver [14]). Partimos de un retículo irregular en el que cada localización se identifica con una de las areas de estudio (municipios, provincias...), Z_i representa el número de casos (o muertes) en la localización i-esima, E_i número esperado de casos en esa región suponiendo un riesgo común (estandarizadas por grupos de edad), r_i riesgo relativo específico de cada area.

$$Z_i|r_i \sim \text{Pois}(E_i r_i)$$

La estimación Máximo verosímil para la tasa de mortalidad estandarizada (SMR)

$$SMR = \hat{r}_i = z_i/E_i \quad , \quad s_{r_i} = \sqrt{z_i}/E_i$$

suele mostrar resultados muy crudos que no tienen en cuenta el tamaño de la población que varía de un area a otra. Para este caso el uso de modelos jerárquicos permiten una suavización de los resultados que mejora considerablemente las estimaciones, combinando la información de cada area (verosimilitud), $(Z_i|r_i \sim \text{Pois}(E_i r_i))$ con información adicional sobre la variabilidad de r_i $(p(\mathbf{r}|\gamma))$(distribución a priori) lo que coloca estos modelos dentro de la metodología bayesiana, siendo la distribución "final":

$$p(\mathbf{r}|\mathbf{z}, \gamma) \propto p(\mathbf{r}|\gamma) \prod_i \text{Pois}(z_i|E_i r_i)$$

tomando como estimador $\hat{\mathbf{r}}$ la media o moda de esta distribución.

4. Algunos ejemplos de extensión al caso espacio-temporal

Una vez que hemos considerado la dependencia espacial, el siguiente paso sería estudiar la posible dependencia temporal de los datos e incorporarla a nuestro modelo. La estructura de vecindad espacio-temporal podría entenderse como un grafo cadena (Chain Graph) (ver [6]). Para la modelización directa de la dependencia espacio-temporal en [13] se propone la concatenación temporal de modelos auto-regresivos espaciales gaussianos condicionalmente especificados para abordar el análisis de mortalidad por meningitis en España,

en el periodo 1950-1990, con datos agregados a nivel provincial. Se estudia también la influencia de los movimientos migratorios en la aparición de los brotes epidémicos, que alteraría considerablemente la estacionalidad natural de fenómeno. Los datos considerados corresponden a las 50 provincias españolas salvo Ceuta y Melilla que no se incluyeron en el estudio por su heterogéneo tratamiento en las estadísticas oficiales a lo largo del periodo. Junto con las tasas de mortalidad, se recogieron los datos de inmigración anual provincial. El modelo planteado es una extensión de (4 y 6):

$$p(z_i(t)|\boldsymbol{z}(t)_{-i}, \boldsymbol{z}(t)^*, \boldsymbol{z}(t), \boldsymbol{x}(t)^*) \propto \exp\{(\alpha_i(t) + \sum_{j\in\delta_i} \beta_{ij}(t) z_j(t)) + h_i(z_i(t))\} \quad (8)$$

en la expresión (8) $\boldsymbol{z}(t)^*$ y $\boldsymbol{x}(t)^*$ recogen el pasado de $\boldsymbol{z}(t)$ e $\boldsymbol{x}(t)$ respectivamente. La influencia de las covariables asi como la influencia del propio pasado y del pasado de las covariables queda nuevamente recogido el parámetro $\alpha_i(t) = \alpha_i((\boldsymbol{z}(t)^*, \boldsymbol{x}(t)), \boldsymbol{x}(t)^*)$ particularizando al caso Gaussiano:

$$\begin{aligned} &\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{z}(t)^*, \mathbf{x}(t)^* \ \sim \ \mathrm{N}(\boldsymbol{\mu}(t), (I - B(t))^{-1} diag(\sigma_i^2(t))) \\ &\boldsymbol{\mu}(t) = \boldsymbol{\mu}(\mathbf{x}(t), \mathbf{z}^*(t), \mathbf{x}^*(t)) \end{aligned} \quad (9)$$

en el caso de los modelos jerárquicos podemos citar como un ejemplo de su extensión al análisis espacio-temporal [15] que presentan un modelo de suavización espacio-temporal auto-regresiva. Este modelo extiende el modelo Gaussiano Auto-regresivo intrínseco al caso espacio-temporal considerando un proceso auto-regresivo de primer orden ($AR(1)$) para introducir la dependencia temporal de las observaciones, en el segundo nivel de jerarquía. Para la estimación del modelo se usan técnicas MCMC. Este modelo ha sido aplicado recientemente para estudiar la evolución espacio-temporal de varias causas de mortalidad para el periodo 1987-2006, en la comunidad Valenciana (ver [20]).

Referencias

[1] J. Besag, Spatial interaction and the statistical analysis of lattice systems, Journal of the Royal Statistical Society B, vol 36, pp 192-225, (1974).

[2] J. Besag, J. York y A. Mollié, Bayesian image restoration with two applications in spatial statistics, Annals of the Institute of Statistical Mathematics, vol 43, pp 1–59, (1991).

[3] E. Castillo, J.M. Gutierrez, y A.S. Hadi, Expert Systems and Probabilistic Networks Models, Springer, New York, (1997).

[4] D. Clayton y J. Kaldor, Empirical Bayes estimates of age-standarized relative risks for use in disease mapping. Biometrics, vol 43 , pp 671–681 (1987).

[5] N. Cresssie, Statistics for Spatial Data. Wiley, New York. (1991).

[6] J. Ferrándiz, E. Castillo y P. Sanmartín, Temporal Aggregation in Chain Graphs Models, Journal of Statistical Planning and Inference, vol 133, pp. 69-93, (2005).

[7] S. Lauritzen, Graphical models, Oxford Statistical Science Series , Vol 17, Oxford University Press, (1996).

[8] G. Winkler, Image Analysis, Random Fields and Dynamic Monte Carlo Methods, A matematical introduction, Springer (1995).

[9] C.J. Geyer. Estimation and Optimization of Functions. En Markov Chain Monte Carlo in Practice. Eds. W.R. Gilks, S. Richardson y D. J. Spiegelhalter. Chapman and Hall, pp 241-258. (1996).

[10] Markov Chain Monte Carlo in Practice. Eds. W.R. Gilks, S. Richardson y D. J. Spiegelhalter. Chapman and Hall, (1996).

[11] J. Ferrándiz, A. López, M. Morales, y M. Tejerizo. Spatial interaction between neighbouring counties: cancer data in Valencia (Spain). Biometrics, vol 51, pp. 665-678. (1995).

[12] J. Ferrándiz, A. López, y P. Sanmartín. Spatial regression models in epidemiological studies. En *Disease Mapping and Risk Assessment for Public Health Decision Making* (Lawson, A., Boehning, D., Lesaffre, E., Biggeri, A., Viel, J. F., and Bertollini, R., editors), pp 203–214. Wiley, Chichester. (1999).

[13] J. Ferrándiz, F. Martínez Navarro, P. Sanmartín, Concatenación temporal de modelos espaciales y su aplicación al estudio de la meningitis en España, Questiió: Quaderns d'Estadística, Sistemes, Informatica i Investigació Operativa, Vol. 25, pp. 47-68 (2001)

[14] Disease Mapping and Risk Assessment for Public Health Decision Making, A. Lawson, D. Boehning, E. Lesaffre, A. Biggeri, J. Viel, R. and Bertollini, R., editors, . Wiley, Chichester (1999).

[15] M.A., Martínez-Beneito, A., López Quílez y P. Botella Rocamora. An autoregressive approach to spatio-temporal disease mapping. Statistics in Medicine. vol 27 pp 2874-2889. (2008)

[16] R Development Core Team (2009). R: A Language and Environment for Statistical Computing. R foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07. URL:www.R-project.com.

[17] S, Richardson , C. Guihenneuc, and V. Laserre. Spatial linear models with autocorrelated error estructure. The Statistician. vol 41, pp 539-557. (1992)

[18] D. Spiegelhalter, A. Thomas, N., Best y D. Lunn. WinBUGS. User Manual, version 1.4. Technical report, Cambridge, UK:MRC Biostatistics Unit.

[19] http://www.R-project.com/

[20] http://www.geeitema.org/AtlasET/