1. ARTÍCULOS DE ESTADÍSTICA

A BRIEF INTRODUCTION TO SPATIO-TEMPORAL MODELLING

María Dolores Ugarte*

Departamento de Estadística e Investigación Operativa
Universidad Pública de Navarra, Campus de Arrosadía, 31006 Pamplona, Spain

Abstract

In recent years, spatio-temporal modelling is becoming a crucial area of research related to the statistical analysis of data arising from a wide variety of real applications in different fields such as ecology, biology, geology, epidemiology, and environmental health, to name a few. All of these fields treat with spatio-temporal data. From a statistical point of view, the data can be regarded as realizations of random variables spread out in space and evolving in time. In this paper we provide a brief introduction to the different types of spatio-temporal data.

Keywords: space-time geostatistics, space-time lattice data, space-time point processes

1. Introduction

El análisis y la modelización de datos espacio-temporales es un área de investigación con enorme proyección en estadística. Numerosas reuniones científicas celebradas recientemente se han dedicado en exclusividad a la modelización espacio-temporal. Algunas de ellas nacieron en España como las ya conocidas ediciones del METMA, cuyo acrónimo se refiere a la Modelización Espacio-Temporal de Procesos Medioambientales. En concreto, la última edición se celebró en Pamplona (www.unavarra.es/metma3/). El lector interesado puede leer las contribuciones de los ponentes invitados en un volumen especial dedicado al evento (ver Ugarte, 2007). En este volumen y también en las actas de la reunión (ver, Militino et al., 2006) se recogen además las interesantes aportaciones de los grupos españoles trabajando en este campo. Una monografía muy reciente sobre modelización espacio-temporal es la editada por Finkenstädt, Held e Isham (2007).

La modelización espacio-temporal se ha entendido como una extensión del caso espacial. Por ello es interesante recordar que en estadística espacial se distinguen tres tipos de datos (ver, por ejemplo, Cressie, 1993): datos geoestadísticos o georreferenciados (*geostatistical data*), datos en un enrejado o datos en un área (*lattice data*), y datos de procesos puntuales (*point processes data*). En geoestadística los datos se observan en $\mathbf{s}$ localizaciones de un conjunto incontable $D \in R^d$ donde d especifica la dimensionalidad del espacio. Típicamente las localizaciones se expresan en dos o tres coordenadas espaciales. Por ejemplo, longitud, latitud, y/o altitud. Las observaciones se toman en cada localización y se consideran como una realización de un proceso estocástico espacial denotado generalmente por $Z(\mathbf{s})$. El objetivo de la geoestadística en aplicaciones reales es generalmente predecir el proceso en nuevas localizaciones mediante técnicas de krigeado (*kriging*). Este método permite realizar predicciones óptimas en nuevas localizaciones a partir de los datos observados en las localizaciones muestreadas. El método precisa del conocimiento de la función de covarianza o del conocido semivariograma. Para datos en un enrejado, la región de estudio D es de nuevo un subconjunto de R^d pero contable. Es quizás más fácil pensar en un número finito de áreas con vecindad bien definida. Las áreas pueden estar regular o irregularmente espaciadas. Las aplicaciones más importantes en este tipo de datos son de tipo epidemiológico, aunque también son comunes las aplicaciones en estadística oficial. En los procesos puntuales espaciales se considera al conjunto D como una colección de sucesos aleatorios cuya realización se llama proceso puntual. En otras palabras, un proceso puntual espacial es una colección

*Corresponding Author. E-mail: lola@unavarra.es

de sucesos aleatorios donde cada suceso indica la localización de un evento de interés. A veces en cada localización se observa una variable de interés o marca. En este caso se habla de procesos puntuales marcados.

Los datos espacio-temporales son, como su propio nombre indica, datos recogidos en el espacio y que evolucionan en el tiempo. Un ejemplo ya clásico en la literatura es el recogido en el artículo de Haslett y Raftery (1989). Se trata de un conjunto de series temporales de velocidades medias del viento en once estaciones meteorológicas de Irlanda en el periodo 1961-1978. El objetivo final del estudio es predecir la capacidad de producción de energía eólica de Irlanda. Un ejemplo de naturaleza distinta es el análisis de la evolución espacio-temporal de la mortalidad por cáncer en España a nivel provincial en los últimos diez años. En este caso, la variable de interés es el número de casos de mortalidad anuales por provincia y el objetivo es analizar la evolución en el tiempo del patrón geográfico de mortalidad por cáncer en España. Un tercer ejemplo que también difiere de los dos anteriores se refiere al estudio de contagios masivos como en el caso de la enfermedad de las vacas locas en el Reino Unido. Aquí es de vital importancia analizar la distribución espacial de la epidemia conjuntamente con su evolución temporal (ver Diggle, 2007). Los tres ejemplos citados representan, en este orden, a tres tipos de datos espacio-temporales: datos geoestadísticos o georreferenciados, datos en un enrejado o datos de área, y datos de procesos puntuales. Observemos que los tipos de datos considerados conservan los nombres surgidos del campo de la estadística espacial, con la salvedad de que ahora se considera además una nueva variable: el tiempo, y por tanto, es conveniente añadir a los tipos de datos anteriores el calificativo de espacio-temporales.

Los casos reales citados son tan sólo tres ejemplos de un campo de aplicaciones cada vez más abundante. Desde el punto de vista estadístico es claro que disponemos de observaciones que no son, en general, independientes. En el Reino Unido se infectaban más fácilmente las granjas próximas a otras ya infectadas, si bien es cierto también que en determinadas ocasiones las infecciones se producían en granjas más lejanas entre sí pero con características similares. La posible dependencia de las observaciones en tiempos cercanos debe también contemplarse en los posibles modelos.

Desafortunadamente, las herramientas estadísticas necesarias para el análisis de datos espacio-temporales no están tan desarrolladas como los métodos de análisis de datos espaciales y de series temporales por separado. A continuación describimos algunos avances realizados, sin ánimo de ser exhaustivos, tarea que resultaría demasiado compleja dada la enorme actividad reciente en este campo.

2. Geoestadística espacio-temporal

La extensión de los datos geoestadísticos espaciales al caso espacio-temporal no es inmediata. Se pueden destacar dos aproximaciones posibles: la primera aproximación se refiere a la utilización de modelos inicialmente diseñados para el análisis de datos espaciales o temporales. Esta aproximación no permite modelizar completamente la posible dependencia espacio-temporal de los datos. La segunda aproximación utiliza una función aleatoria espacio-temporal que permite considerar interacciones entre el espacio y el tiempo. Es claro que esta segunda aproximación es más adecuada y es la que discutiremos aquí brevemente. En este enfoque se considera que las observaciones son una realización parcial de una función aleatoria espacio-temporal típicamente Gaussiana del tipo

$$Z(\mathbf{s},t), \;\; (\mathbf{s},t) \in R^d \times R$$

Aquí, de nuevo $\mathbf{s}$ se refiere a la localización espacial y t al tiempo. Normalmente se asume que existen los momentos de segundo orden de la función aleatoria y que son finitos.

En numerosos estudios el objetivo fundamental es la predicción espacio-temporal. La utilización del método del krigeado en este contexto descansa en definir apropiadamente una estructura de covarianza espacio-temporal. Gran parte de la investigación desarrollada en geoestadística espacio-temporal tiene que ver con la construcción de modelos de covarianza válidos y lo suficientemente flexibles (ver, por ejemplo, Cressie y Huang, 1999; Gneiting, 2002; Stein, 2005; Gneiting, Genton y Guttorp, 2007). La covarianza entre $Z(\mathbf{s}_1,t_1)$ y $Z(\mathbf{s}_2,t_2)$ depende únicamente de las coordenadas espacio-temporales si no se realizan hipótesis adicionales. En ocasiones es interesante que las funciones de covarianza satisfagan ciertas hipótesis como las de separabilidad,

simetría completa, y estacionariedad. Estas hipótesis facilitan, en general, el cálculo de las funciones de covarianza.

Se dice que Z tiene una función de covarianza separable si existe una función de covarianza estrictamente espacial, que denotaremos por C_S y una función de covarianza sólo temporal, C_T de modo que se verifica lo siguiente

$$\text{cov}\left[Z(\mathbf{s}_1,t_1),Z(\mathbf{s}_2,t_2)\right]=C_S(\mathbf{s}_1,\mathbf{s}_2)C_T(t_1,t_2)$$

$$\forall\,(\mathbf{s}_1,t_1)\text{ y }(\mathbf{s}_2,t_2)\in R^d\times R$$

En términos computacionales esta propiedad es muy interesante aunque tiene el inconveniente de que en muchos problemas prácticos es difícilmente justificable. Existen numerosos trabajos que proponen tests que contrastan la hipótesis de separabilidad. Tres de los más recientes son: Mitchell et al. (2005), Scaccia y Martin (2005) y Fuentes (2006).

Una función de covarianza es completamente simétrica si

$$\text{cov}\left[Z(\mathbf{s}_1,t_1),Z(\mathbf{s}_2,t_2)\right]=\text{cov}\left[Z(\mathbf{s}_1,t_2),Z(\mathbf{s}_2,t_1)\right]$$

$$\forall\,(\mathbf{s}_1,t_1)\text{ y }(\mathbf{s}_2,t_2)\in R^d\times R$$

La separabilidad es un caso particular de la simetría completa, de modo que aquellas estructuras de covarianza que no sean completamente simétricas no pueden ser separables.

Por último, una función de covarianza espacio-temporal es espacialmente estacionaria si depende de las localizaciones $\mathbf{s}_1$ y $\mathbf{s}_2$ sólo a través de su diferencia $\mathbf{s}_1-\mathbf{s}_2$, y es temporalmente estacionaria si depende de los tiempos de observación sólo a través de su diferencia t_1-t_2. Si un proceso espacio-temporal tiene función de covarianza espacial y temporalmente estacionaria se dice que el proceso tiene función de covarianza estacionaria. Cuando esta hipótesis se verifica, existe una función C definida en $R^d\times R$ que verifica

$$\text{cov}\left[Z(\mathbf{s}_1,t_1),Z(\mathbf{s}_2,t_2)\right]=C(\mathbf{s}_1-\mathbf{s}_2,t_1-t_2)$$

Se pueden encontrar contrastes sobre estacionariedad en el trabajo de Fuentes (2005).

En resumen y como hemos comentado, las predicciones con procesos espacio-temporales descansan en definir funciones de covarianza adecuadas y lo suficientemente flexibles que permitan contemplar formas complicadas de interacción entre el espacio y el tiempo. En la práctica se suele requerir que se satisfagan algunas hipótesis como separabilidad, simetría completa o estacionariedad que facilitan los cálculos desde un punto de vista computacional.

3. Modelos de área espacio-temporales

Los modelos a nivel de área se aplican con éxito en el campo de la representación cartográfica de enfermedades (*disease mapping*). Aquí vamos a presentar brevemente los modelos más utilizados en este campo. En el caso estrictamente espacial, uno de los modelos más utilizados es el de Besag, York y Mollié (1991), que llamaremos BYM. Existen en la literatura algunos modelos alternativos y reparametrizaciones de este mismo modelo que permiten contrastar la presencia de dependencia espacial de modo sencillo. Ver, por ejemplo, Dean, Ugarte y Militino (2001) y Ugarte, Ibáñez y Militino (2005, 2006).

Los modelos espacio-temporales que se presentan como una extensión a BYM se construyen de forma jerárquica. Supongamos que disponemos de una región subdividida en S unidades geográficas (la región puede ser una provincia y las subdivisiones zonas básicas de salud) y con datos de mortalidad por cáncer en los años $t=1,\dots T$.

El primer nivel de la jerarquía es común a todos los modelos y asume que, condicionando al riesgo relativo subyacente denotado por r_{st}, el número de casos de mortalidad Z_{st} en el área s y en el periodo t sigue una distribución de Poisson de media $\mu_{st}=e_{st}r_{st}$. Es decir,

$$Z_{st}|r_{st}\sim\mathcal{P}(\mu_{st}=e_{st}r_{st}),\ \log(\mu_{st})=\log(e_{st})+b_{st},\qquad(3.1)$$
$$s=1,\dots,S,\quad t=1,\dots T,$$

donde e_{st} es el número esperado de casos de mortalidad en el área s y en el periodo t y $b_{st}=\log(r_{st})$. Los modelos difieren en el segundo nivel de la jerarquía, que permite modelizar la dependencia espacio-temporal de diversas maneras. Vamos a presentar aquí algunas ideas posibles.

Modelo 1. La parte temporal es lineal, y la evolución temporal de los riesgos es la misma para todas las áreas. El segundo nivel de la jerarquía viene dado por

$$b_{st} = \alpha + u_s + v_s + \beta t, \tag{3.2}$$

donde α es la tasa global y u_s y v_s son efectos aleatorios que recogen la dependencia espacial y la variabilidad no espacial (heterogeneidad no estructurada), respectivamente. β es la tendencia media temporal de todas las áreas. Para los efectos aleatorios espaciales se asume una distribución de tipo CAR (*Conditional Autoregressive*) cuya distribución conjunta es $\mathbf{u} = (u_1, \ldots, u_S)' \sim N(\mathbf{0}, \sigma_u^2 \mathbf{Q}^-)$. Aquí $\mathbf{Q}$ es una matriz de vecinos y $\mathbf{Q}^-$ indica la matriz inversa generalizada Moore-Penrose. El elemento diagonal s-ésimo de $\mathbf{Q}$ viene dado por el número de vecinos del área s-ésima y para $s \neq j$ $\mathbf{Q}_{sj} = -1$ si s y j son áreas vecinas, y toma el valor cero en otro caso. Los efectos aleatorios v_s que recogen la variabilidad no espacial siguen una distribución aleatoria del tipo $v_s \sim N(0, \sigma_v^2)$.

Modelo 2. En este modelo se permiten distintas ordenadas en el origen en los distintos periodos temporales (ver Knorr-Held y Besag, 1998). Este modelo puede ser apropriado cuando el número de periodos temporales es pequeño. El segundo nivel de la jerarquía se especifica así

$$b_{st} = \alpha_t + u_s + v_s, \tag{3.3}$$

donde α_t es el efecto del periodo t, y u_s y v_s se definen como en el modelo anterior.

Modelo 3. Este modelo se debe a Bernardinelli et al. (1995). La componente temporal es lineal pero el modelo incluye interacción espacio-temporal. El segundo nivel de la jerarquía es

$$b_{st} = \alpha + u_s + v_s + (\beta + \delta_s)t, \tag{3.4}$$

y ahora, δ_s es la diferencia entre la tendencia específica del área y la tendencia media β. Además se supone que $\delta_s \sim N(0, \sigma_\delta^2)$.

Modelo 4. Este modelo es más sofisticado. El efecto temporal se incluye también como un efecto aleatorio. El segundo nivel de la jerarquía es

$$b_{st} = \alpha + u_s + v_s + \beta_t + \phi_{st}, \tag{3.5}$$

donde α es la tasa global, u_s y v_s se modelizan como en el modelo 1, y β_t es el efecto aleatorio de la tendencia temporal. Aquí se considera un paseo aleatorio de primer orden como distribución a priori de los efectos temporales (Knorr-Held, 2000; Richardson et al., 2006). Es decir, una versión unidimensional del modelo CAR espacial con matriz de adyacencias $\mathbf{W}$ definida de forma análoga a $\mathbf{Q}$ y que define como vecinos temporales del periodo t a los periodos $t-1$ y $t+1$. Obviamente el primer y último periodo sólo tienen un vecino. Así, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_T)' \sim N(\mathbf{0}, \sigma_\beta^2 \mathbf{W}^-)$. Los efectos aleatorios ϕ_{st} representan la interacción espacio-temporal y su distribución viene dada por $\phi_{st} \sim N(0, \sigma_\phi^2)$.

Si los modelos anteriores se ajustan desde un punto de vista completamente Bayesiano, enfoque muy común en el campo de la representación cartográfica de enfermedades, hay que especificar las distribuciones a priori de los parámetros del modelo. Así, siguiendo las indicaciones de Wakefield et al. (2000), es común utilizar distribuciones gamma del tipo $\Gamma(0{,}5, 0{,}0005)$ para los parámetros de precisión $\tau_u = 1/\sigma_u^2$, $\tau_v = 1/\sigma_v^2$, $\tau_\delta = 1/\sigma_\delta^2$, $\tau_\beta = 1/\sigma_\beta^2$ y $\tau_\phi = 1/\sigma_\phi^2$. Para los efectos fijos suele asumirse α, $\alpha_t \propto 1$ y $\beta \sim N(0, 10^5)$. De todos modos, es importante realizar estudios de sensibilidad a las distribuciones a priori. Los modelos se ajustan utilizando métodos MCMC (*Markov Chain Monte Carlo*). Ver, por ejemplo, Gilks, Richardson y Spiegelhalter (1996).

Si los modelos se tratan desde un punto de vista empírico-Bayesiano es común utilizar técnicas de verosimilitud aproximadas como PQL (*Penalized Quasi-likelihood*), que popularizaron Breslow y Clayton (1993). Un análisis del comportamiento de esta técnica en datos con dependencia espacial puede verse en Dean, Ugarte y Militino (2004).

4. Procesos puntuales espacio-temporales

Esta sección trata del análisis de datos del tipo (s_i, t_i), $i = 1, \ldots, n$, donde cada s_i denota la localización geográfica del dato y t_i el tiempo de ocurrencia del suceso de interés. Se asume que se dispone de todos los sucesos ocurridos en una determinada región espacial D y en un intervalo de tiempo especificado $(0, T)$. Un conjunto de datos de este tipo se llama patrón puntual espacio-temporal y al modelo estocástico subyacente se le denomina proceso puntual espacio-temporal.

En el análisis de este tipo de datos es importante distinguir si los sucesos individuales (s_i, t_i) ocurren en un espacio-tiempo continuo o se considera que

la escala temporal es o bien naturalmente discreta o se discretiza considerando sólo los sucesos del patrón espacial agregado sobre una secuencia de periodos de tiempo discretos. Los métodos utilizados para el análisis difieren según los tres casos anteriores. Es decir, no existen por el momento métodos genéricos de análisis de procesos puntuales espacio-temporales. Una interesante revisión sobre métodos y aplicaciones de este tipo de procesos puede verse en Diggle (2007) y en las referencias ahí citadas.

Agradecimientos: Quiero agradecer a M. Carmen Pardo, editora del boletín, que me haya invitado a escribir este pequeño artículo, y que se lo haya leído con dedicación, ayudándome a mejorar una versión previa. Este trabajo ha sido realizado en el marco del proyecto MTM2005-00511 del Ministerio de Educación y Ciencia.

Referencias

[1] Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. y Songini, M. 1995. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* **14**: 2433-2443.

[2] Besag, J., York, J. y Mollié, A., 1991. Bayesian image restoration with two Applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**: 1-59.

[3] Breslow, N. E. y Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**: 9-25.

[4] Cressie, N. (1993). *Statistics for Spatial Data*. Second Edition. New York: Wiley.

[5] Cressie, N. y Huang, H. (1999). Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association* **94**: 1330-1340.

[6] Dean C. B., Ugarte M. D. y Militino A. F. (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics* **57**: 197-202.

[7] Dean C. B., Ugarte M. D. y Militino A. F. (2004). Penalized Quasi-Likelihood with Spatially Correlated Data. *Computational Statistics and Data Analysis* **45**: 235-248.

[8] Diggle, P. (2007). "Spatio-temporal point processes: methods and applications,"in Finkenstaedt, B., Held, L. and Isham, V. (eds.), Statistics of Spatio-Temporal Systems, Chapman and Hall/CRC Press, Monographs in Statistics and Applied Probability, 2-23.

[9] Finkenstädt, B., Held, L. e Isham, V. (Editors) (2007). *Statistical Methods for Spatio-Temporal Systems*. Boca Raton: Chapman and Hall/CRC.

[10] Fuentes, M. (2005). A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis* **96**: 30-55

[11] Fuentes, M. (2006). Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference* **136**: 447-466.

[12] Gilks W. R., Richardson, S. y Spiegelhalter, D. J. (eds) (1996) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

[13] Gneiting, T. (2002). Stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97**: 590-600.

[14] Gneiting, T., Genton, M. G. y Guttorp, P. (2007), "Geostatistical space-time models, stationarity, separability and full symmetry,"in Finkenstaedt, B., Held, L. and Isham, V. (eds.), Statistics of Spatio-Temporal Systems, Chapman and Hall/CRC Press, Monographs in Statistics and Applied Probability, 151-175.

[15] Haslett, J. y Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource (with Discussion). *Journal of the Royal Statistical Society, Series C - Applied Statistics* **38**, 1-50.

[16] Knorr-Held, L. y Besag, J. (1998). Modelling risks from a disease in time and space. *Statistics in Medicine* **17**: 2045-2060.

[17] Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* **19**: 2555-2567.

[18] Militino, A. F., Ugarte, M. D., González-Ramajo, B. y Goicoa, T. (Editors) (2006). *International Workshop on Spatio-Temporal Modelling (METMA3)*. Pamplona: Nova Text.

[19] Mitchell, M., Genton, M. G. y Gumpertz, M. (2005). Testing for separability of space-time covariances. *Environmetrics* **16**: 819-831.

[20] Richardson, S., Abellán, J. J., Best, N. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire. *Statistical Methods in Medical Research* **15**: 385-407.

[21] Scaccia, L. y Martin, R. J. (2005). Testing axial symmetry and separability of lattice processes. *Journal of Statistical Planning and Inference* **131**: 19-39.

[22] Stein, M. L. (2005). Space-time covariance functions. *Journal of the American Statistical Association* **100**: 310-321.

[23] Ugarte, M. D., Ibáñez, B. y Militino, A. F. (2005). Detection of spatial variation in risk when using CAR models for smoothing relative risks. *Stochastic Environmental Research and Risk Assessment*, **19**, 33-40.

[24] Ugarte, M. D., Ibáñez, B. y Militino A. F. (2006) Modelling Risks in Disease Mapping. *Statistical Methods in Medical Research*, **15**, 21-35.

[25] Ugarte, M. D. (2007) Guest Editorial: METMA3 Workshop 2006. *Environmetrics* **18**: 663-664.

[26] Wakefield, J. C., Best, N. G. y Waller, L. (2000). Bayesian approaches to disease mapping. In Elliot P, Wakefield JC, Best NG, Bridges DJ, eds. *Spatial Epidemiology*, Oxpord University Press, 104-127.

M. Dolores Ugarte es profesora titular y actual directora del Departamento de Estadística e Investigación Operativa de la Universidad Pública de Navarra. Es además vocal de estadística en el Consejo Ejecutivo de la SEIO y representante del grupo de comunicaciones del Consejo Ejecutivo de la *Statistical Modelling Society*. Sus líneas de investigación más destacadas son la modelización espacial y espacio-temporal, y la estimación en áreas pequeñas. Desde enero de 2007 es editora asociada del Journal of the Royal Statistical Society (Series A). Sus publicaciones más relevantes, proyectos recientes y libros docentes pueden consultarse en el sitio web: http://www.unavarra.es/personal/lugarte.