Algunas aplicaciones de los modelos gráficos en epidemiología

P. Sanmartín[1] A. López-Quílez[2] y E. Castillo[3]
[1]Departamento de Matemática Aplicada y Estadística.
Universidad Politécnica de Cartagena
[2]Departament d'Estadística i Investigació Operativa. Universitat de València
[3]Departamento de Estadística e Investigación Operativa. Universidad de Cantabria

Antecedentes y objetivos

En el análisis espacio-temporal de datos medioambientales, las observaciones vienen dadas frecuentemente mediante la agregación temporal de valores no observados directamente. Esta situación es bastante usual en el contexto de los estudios epidemiológicos en los que los datos de mortalidad y morbilidad en áreas geográficas son el resultado de una agregación previa de otros valores no observados de forma directa. Por ejemplo, en enfermedades contagiosas, se parte de tasas anuales como resultado de la agregación de datos observados mensual, semanal o diariamente. Lo que se observa es la suma de los valores que fueron observados en una escala temporal menor. La importancia o no de este hecho viene marcada por la naturaleza de la enfermedad y su velocidad de transmisión y propagación. En muchos casos es natural suponer que no existe contagio instantáneo y que por tanto los datos originales sólo presentan una estructura de dependencia temporal. Sin embargo, cuando se observan las variables agregadas, puede que sí aparezca algún tipo de dependencia espacial motivada por la agregación. Tal como puntualiza Cressie (1991), *"el cambio de la escala temporal es crucial para determinar si un modelo espacio-temporal debe tener una componente puramente espacial"*. Ferrándiz et al. (2003) comparan distintos modelos espacio-temporales en diferentes niveles de agregación temporal con el objetivo de poder captar el efecto de covariables en estudios medioambientales.

En el presente trabajo se aborda esta problemática, es decir, el análisis de la estructura de dependencia de un proceso inducido por agregación de un fenómeno espacial que evoluciona en el tiempo. Se parte de los resultados obtenidos en Ferrándiz et al. (2004) y se discuten sus posibles aplicaciones en el ámbito de la epidemiología.

Se describe la agregación temporal de datos espacio-temporales mediante modelos gráficos. La estructura de dependencia de los datos se representa a través de grafos cadena, (ver Lauritzen, 1996; Lauritzen y Richardson, 2002). Éstos surgen de manera natural cuando se pretende extender al caso espacio-temporal las estructuras markovianas espaciales descritas a través de los grafos no dirigidos (ver por ejemplo Besag, 1974; Darroch et al., 1980). Extensiones al caso temporal de este tipo de estructuras

pueden encontrarse en Guyon (1995); Lacruz et al. (2000); Dahlhaus y Eichler (2003).

Métodos

Se considera un vector de variables aleatorias correspondientes a un conjunto de L localizaciones espaciales a lo largo de T instantes temporales, que puede representarse como una serie temporal multivariante. En cada instante, los valores observados en cada localización espacial se suponen condicionalmente independientes del resto de valores "contemporáneos" dados los valores observados en los instantes anteriores. Se supone además que cada variable depende del pasado únicamente a través de un conjunto reducido de localizaciones vecinas. Esta estructura de independencia se puede modelizar de forma muy sencilla y clara mediante el empleo de grafos de independencia no dirigidos acíclicos. Cada vértice del grafo representa una variable y las relaciones de dependencia se expresan mediante flechas dirigidas de "padres" a "hijos" (siempre en la dirección natural de la secuencia temporal), no existiendo flechas entre vértices correspondientes a variables de un mismo instante temporal.

El proceso espacio-temporal de interés no se observa directamente. Sólo se dispone de alguna clase de agregación de este proceso oculto y es ese proceso agregado resultante el que constituye el proceso espacio-temporal directamente observable. Una situación de este tipo puede aparecer cuando limitaciones de tipo administrativo impiden un seguimiento detallado de la difusión de una enfermedad infecciosa. Pueden distinguirse las localizaciones geográficas correspondientes a las unidades administrativas en que se recogen los datos pero no puede evitarse que los datos estén agregados temporalmente debido a la dificultad y complejidad que acompaña a la recogida de información, en este caso la agregación correspondería a sumas de valores de una variable observada en una misma localización y sucesivos instantes temporales.

La relación entre el proceso original y el proceso oculto se puede modelizar asimismo a través de grafos dirigidos acíclicos. Basta añadir al grafo anteriormente descrito un vértice asociado a cada una de las variables resultantes y una flecha desde cada uno de los vértices de las variables originales que se agregan dirigida al vértice de la variable agregada obtenida.

Al considerar únicamente el proceso directamente observado y pretender describir la estructura de dependencia de las variables agregadas no se puede garantizar que las variables "contemporáneas" sean condicionalmente independientes dado el pasado y se debe recurrir a otras representaciones gráficas de independencia que consideren la dependencia espacial instantánea junto con la dependencia del pasado, para ello se usan grafos cadena. Estas estructuras describen la independencia de las variables a través de bloques ordenados secuencialmente (que pueden asociarse a las variables espaciales en un mismo instante temporal), de forma que sólo existen "flechas" entre vértices de bloques distintos (respetando el orden secuencial) y dentro de cada bloque

las relaciones de dependencia se consideran "simétricas" entre "vecinos" y se representan mediante aristas (corresponderían a las relaciones de dependencia instantáneas entre diferentes localizaciones geográficas).

De la lectura del grafo cadena se pueden reconocer propiedades de independencia de las distribuciones de probabilidad definidas sobre el conjunto de variables asociadas a los vértices, bien para pares de variables, para un subconjunto en un sentido local o respecto de todas las variables globalmente, todo ello a partir de criterios de separación aplicados al grafo. Asimismo esta estructura gráfica se puede interpretar en términos de propiedades de factorización de las correspondientes densidades de probabilidad (respecto a la media adecuada en cada caso). Esta última caracterización es la que se usa para estudiar el proceso de agregación a fin de evitar problemas con la llamada "condición de positividad" (ver Lauritzen, 1996).

Resultados

Partiendo de un grafo dirigido acíclico que refleja la estructura temporal de los datos originales, se plantea la obtención del grafo cadena que refleje la estructura espacio-temporal resultante de los datos agregados temporalmente. Para ello se introduce un algoritmo, que se llama algoritmo de agregación (ver Ferrándiz et al., 2004) que construye dicho grafo en varias pasos, y de forma secuencial añadiendo sucesivamente nuevos instantes temporales a los ya incorporados en la etapa anterior. Para cada instante temporal en la escala de las variables agregadas se construye primero el grafo dirigido acíclico conjunto de las variables originales junto con las agregadas hasta ese instante. Posteriormente se considera un grafo marginal no dirigido (ver Castillo et al., 1998) asociado a las variables agregadas (proceso observable) y se transforma en un grafo cadena incorporando la información de la etapa previa, bien eliminando las aristas entre variables de distintos bloques temporales o sustituyéndolas por flechas.

El grafo cadena obtenido mediante esta construcción es, dentro de la clase de grafos cadena, el que mejor refleja la estructura de dependencia del proceso resultante. Para representar de forma completa todas las independencias existentes en el proceso agregado se necesitan otras estructuras gráficas tales como grafos ancestrales (ver Richardson y Spirtes, 2002) o grafos MC (Koster, 2002). Sin embargo todas ellas pueden conducir a representaciones del modelo agregado que no respete la representación secuencial en bloques temporales del proceso bajo estudio.

Aplicación

Se ilustra el anterior algoritmo y su posible aplicación en el estudio de datos epidemiológicos mediante una serie de ejemplos simulados. Para ello se plantean distintas estructuras hipotéticas de dependencia espacial y temporal junto con su correspondiente grafo asociado y se analiza el grafo resultante al considerar las variables agre-

gadas. Esto permite observar el funcionamiento del algoritmo y evaluar qué tipo de dependencias capta correctamente. Se comparan estos resultados con los grafos ancestrales y MC, que también reflejan la estructura de agregación del proceso resultante y se ve si en cada caso se respeta o no la secuencia temporal natural del proceso. Se acompaña este estudio cualitativo mediante un estudio cuantitativo a través de la simulación de datos correspondientes a los modelos originales, suponiendo las distribuciones de probabilidad asociadas al modelo gráfico dentro de los modelos lineales generalizados (ver por ejemplo Ferrándiz et al., 1995, 1999).

Conclusiones

El algoritmo de agregación y el grafo cadena agregado resultante proporciona una descripción natural de las propiedades de independencia tras la agregación temporal de datos espacio-temporales muchas veces impuesto por restricciones de tipo administrativo. Mediante la metodología anterior se comprueba que nuevas dependencias espaciales y temporales pueden aparecer como resultado del proceso de agregación. El algoritmo puede aplicarse en el sentido inverso, dado el proceso agregado resultante directamente observado se puede aplicar el algoritmo a diferentes grafos de independencia iniciales e intentar identificar las dependencias existentes en el proceso oculto (ver Arnold et al., 2004). Todos estos resultados se ilustran mediante ejemplos que permiten ver sus posibilidades de aplicación en el contexto de la epidemiología.

Referencias

Arnold, B., Castillo, E. y Sarabia, J. M. (2004). Compatibility of partial or complete conditional probability specifications. *Journal of Statistical planning and Inference* 123:133–159.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B* 36:192–225.

Castillo, E., Ferrándiz, J. y Sanmartín, P. (1998). Marginalizing in undirected graph and hypergraph models. G. F. Cooper y S. Moral (eds.), *Uncertainty in Artificial Intelligence*, capítulo 14, 69–78. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Cressie, N. (1991). *Statistics for spatial data*. Wiley, New York.

Dahlhaus, R. y Eichler, M. (2003). Causality and graphical models in time series analysis. P. Green, N. Hjort y S. Richardson (eds.), *Highly Structured Stochastic Systems*, 115–134. Oxford University Press.

Darroch, J., Lauritzen, S. L. y Speed, T. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics* 8:522–539.

Ferrándiz, J., Castillo, E. y Sanmartín, P. (2004). Temporal aggregation in chain graph models. *Journal of Statistical Planning and Inference* En prensa.

Ferrándiz, J., López, A., Gómez-Rubio, V., Sanmartín, P., Martínez-Beneito, M. A., Melchor, I., Vanaclocha, H., Zurriaga, O., Ballester, F., Gil, J. M., Pérez-Hoyos, S. y Abellán, J. J. (2003). Statistical relationship between hardness of drinking water and cerebrovascular mortality in Valencia: a comparison of spatiotemporal models. *Environmetrics* 14:491–510.

Ferrándiz, J., López, A., Llopis, A., Morales, M. y Tejerizo, M. (1995). Spatial interaction between neighbouring counties: Cancer mortality data in Valencia (Spain). *Biometrics* 51:665–678.

Ferrándiz, J., López, A. y Sanmartín, P. (1999). Spatial regression models in epidemiological studies. A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J. F. Viel y R. Bertolini (eds.), *Disease Mapping and Risk Assessment for Public Health*, 203–215.

Guyon, X. (1995). Random fields on a network. modelling, statistics and applications. K. F. P. y W. R. J. (eds.), *Stochastic Networks*. Springer Verlag, New York.

Koster, J. T. A. (2002). Marginalizing and conditioning in graphical models. *Bernoulli* 8:817–840.

Lacruz, B., Lasala, P. y Lekuona, A. (2000). Dynamic graphical models and nonhomogeneous hidden Markov models. *Statistics and Probability Letters* 49:377–385.

Lauritzen, S. L. (1996). *Graphical models*, volumen 17 de *Oxford Statistical Science Series*. Oxford University Press.

Lauritzen, S. L. y Richardson, T. S. (2002). Chain graph models and their causal interpretations (with discussion). *Journal of the Royal Statistical Society Series B* 64:321–361.

Richardson, T. S. y Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics* 30:962–1030.