Métodos estadísticos para analizar el riesgo con patrones de distribución espacial

Rosa M. Abellana[a] y Carlos Ascaso[a,b]

[a]*Bioestadística. Departamento de Salud Pública. Universitat de Barcelona. Barcelona. España.*
[b]*Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS). Barcelona. España.*

El desarrollo de los sistemas de información geográficos ha impulsado el interés por analizar las características de las poblaciones teniendo en cuenta el lugar donde viven.
La salud de estas poblaciones, medidas como riesgos de morbilidad o mortalidad, está condicionada por una gran variedad de factores de riesgo característicos de su región, como los medioambientales, que en ocasiones no se pueden medir. Los modelos de análisis espacial recogen la dependencia de los indicadores de salud entre regiones próximas provocada por la existencia de los factores de riesgo que éstas comparten pero que a veces no están medidos, de tal manera que la distribución espacial de los indicadores depende del patrón geográfico que siguen los factores de riesgo.
En este trabajo se comentan algunas de las limitaciones que presentan los métodos de estandarización y la regresión de Poisson para modelar la espacialidad y se muestran las ventajas que aportan los modelos espaciales autorregresivos. Esta metodología se ilustra mediante los datos de incidencia de diabetes tipo 1 en Cataluña durante el período 1989-1998.

Palabras clave: Análisis espacial. Modelos espaciales autorregresivos. Regresión de Poisson. Razón estandarizada de mortalidad.

Statistical methods to analyze risk with spatial distribution patterns

The development of the geographic information systems has improved the interest in analyzing the characteristics of the population, taking into account the place where they live. The health of the population, measured as the risk of morbidity or mortality, is conditioned by the variety of risk factors characteristic of the region which cannot be measured. Analysis of spatial models considers the dependence of the health indicators between close regions. This dependence is due to the existence of the risk factors that are not measured but are shared by the region. Thus, the spatial distribution of these indicators depends on the geographic pattern of these risk factors.
In this study, some limitations of the standardized methods and the Poisson regression used to model the spatiality are discussed and the advantages of the spatial models are shown. The methodology is illustrated by the insulin-dependent diabetes type 1 data from Catalonia during 1989 and 1998.

Key words: Spatial analysis. Autoregressive spatial model. Poisson regression. Standarized mortality rate.

Introducción

En la actualidad, la mayoría de las instituciones y la Administración recogen sistemáticamente información de las regiones de las áreas geográficas que gestionan. Estas bases de datos almacenan sus características ambientales, poblacionales, epidemiológicas, etc. y dan lugar a distintos sistemas de información que pueden estar conectados entre sí por el común denominador de la región geográfica donde se han realizado las medidas. Si las regiones tienen medida su posición geográfica y vinculada a ella su información, los datos están referenciados geográficamente y forman un sistema de información geográfica (SIG).

Los sistemas con información referenciada geográficamente que almacenan datos relacionados con la salud de las comunidades se denominan sistemas de información geográficos sanitarios (SIGS) y su análisis numérico tiene como principales objetivos[1,2]: *a)* facilitar la identificación de áreas y/o poblaciones con mayores necesidades insatisfechas de salud, de manera que permita tomar decisiones de una forma ágil y focalizar hacia esos grupos prioritarios las intervenciones; *b)* facilitar la identificación de los patrones espaciales del riesgo de morbilidad o mortalidad con el propósito de buscar hipótesis sobre sus causas, y *c)* evaluar la relación entre los niveles de exposición promedio de un factor de riesgo y la carga de morbilidad o mortalidad.

Una parte importante de los datos que almacenan los SIGS son recuentos de la presencia de un episodio en función de las regiones en que se divide el área geográfica de estudio. Pero cuando se analiza esta información, lo más usual es expresar los recuentos de casos en función de la población en que se detectan. En epidemiología esta razón recibe el nombre de tasa y se puede interpretar como una medida de riesgo.

Los procedimientos estadísticos que se deben aplicar para modelar este tipo de datos deben tener en cuenta que, si las tasas siguen patrones de distribución espacial, la información de las regiones próximas estará correlacionada y por tanto se violará la asunción de independencia que exigen los métodos estadísticos clásicos. Las herramientas de análisis numérico que se usan para construir y analizar los modelos de información geográfica se aglutinan con el nombre de estadística espacial, y en este entorno el uso de mapas, particularmente si son computarizados, es el método más efectivo para la transmisión de los resultados de los análisis.

El punto de partida del uso de los modelos de información geográfica en medicina podemos situarlo en 1986, a partir de un trabajo de Gesler[3] donde se lleva a cabo una revisión acerca de los usos del análisis espacial en la geografía médica. Desde entonces, y especialmente en los últimos 10 años, se han realizado experiencias que han tenido y tienen como objetivo integrar un conjunto de herramientas en un sistema automatizado capaz de recoger, almacenar, manejar, analizar y visualizar información referenciada geográficamente[4]. Mientras los sistemas de información se pueden crear usando los paquetes de gestión de bases de datos habituales, por ejemplo Access, FoxPro, etc., los SIG que permiten visualizar la información usando mapas necesitan aplicaciones informáticas como el SIGEpi[5] y el ArcGis[6].

Correspondencia: Dra. R.M. Abellana.
Bioestadística. Departamento de Salud Pública. Universitat de Barcelona.
Casanova, 143.
08036 Barcelona. España.
Correo electrónico: sangra@medicina.ub.es

Este artículo tiene como objetivo presentar y comentar algunos de los métodos estadísticos que se utilizan para analizar datos agrupados (tasas) con correlación espacial. Primero, se enumeran las limitaciones de los procedimientos clásicos y, posteriormente, se presentan las extensiones de éstos para poder realizar análisis geográficos que tengan en cuenta la correlación espacial. Por último, se aplica la metodología expuesta a los datos de incidencia de diabetes tipo 1 en Cataluña para mostrar las diferencias de los resultados. El desarrollo de este ejemplo se ha realizado siguiendo el proceso de análisis necesario para comparar tasas de mortalidad o morbilidad de las regiones de un área y evaluar si presentan un patrón de distribución espacial.

Metodología estadística

Estandarización

Cuando se quieren comparar las tasas de 2 o más poblaciones, el uso de las tasas brutas lleva a resultados incorrectos ya que las diferencias que se pueden observar entre regiones pueden ser imputables no sólo a la intensidad de la característica que se está estudiando, sino también a la estructura de la población respecto a una o más variables, como por ejemplo la edad y el sexo[7]. Este tipo de variables se conoce como «variables de confusión», debido a que están distorsionando la verdadera intensidad del fenómeno de estudio. La estandarización es un método que permite obtener estimaciones de las tasas eliminando el efecto de las variables de confusión, utilizando para ello las tasas específicas de cada uno de los estratos en que se dividen estas variables.

Existen diferentes tipos de estandarización[8], de entre los que destacaremos aquí la estandarización indirecta. Una de las situaciones en la que se utiliza este tipo de estandarización es cuando el número de personas en riesgo es insuficiente para obtener unas tasas específicas representativas, de forma que éstas se pueden obtener internamente utilizando como población de riesgo la de toda el área de estudio o externamente mediante una población de referencia.

Cuando se utiliza la estandarización indirecta, habitualmente se trabaja con la razón estandarizada de mortalidad o morbilidad (SMR) como medida de riesgo. Los SMR se obtienen como el cociente del número de casos observados y el número de casos esperados obtenidos a partir de la estandarización indirecta. De forma que un SMR mayor a 1 indica que existe un riesgo superior en la región de estudio que en la población de referencia; si el SMR es inferior a 1, existe menos riesgo, y si es igual a 1, el riesgo es el mismo. No obstante los SMR de regiones con poca población en riesgo, normalmente rurales, tienden a presentar valores extremos debido a que el número de casos esperados es pequeño. Este hecho provoca que las estimaciones sean muy variables y poco representativas del SMR real de la región.

Un procedimiento que se utiliza para que las estimaciones de los SMR tengan menos variabilidad, es decir sean más estables y así evitar las consecuencias de tener pocos casos esperados, es ajustar los SMR mediante un modelo de regresión.

Modelo de regresión

El modelo de regresión, además de estabilizar los SMR, también permite controlar las estimaciones por posibles variables explicativas, tanto de confusión como factores de riesgo, y por las interacciones que pueden existir entre estas variables.

Cuando se trabaja con tasas o SMR, la variable de estudio corresponde a recuentos, como por ejemplo el número de casos nuevos de una enfermedad en una región y en un período de tiempo determinado. Se asume que los recuentos de cada región se distribuyen bajo una Poisson, por lo que el modelo regresión clásico con residuos distribuidos según una normal no será adecuado, y será preferible utilizar modelos lineales generalizados[9], y en particular la regresión de Poisson[10].

Si definimos $Y = (Y_1, \ldots Y_N)$ como el vector de los recuentos en cada una de las N regiones de estudio, y se considera que Y sigue una Poisson con media $\mu = E \times \theta$, donde E es el vector del número de casos esperados obtenidos de la estandarización y θ es el vector del riesgo relativo para las N regiones, la regresión de Poisson relaciona la media de los recuentos con las variables explicativas mediante la función logarítmica. Este modelo se expresa mediante la siguiente ecuación:

$$\log(\mu) = \log(E) + X\beta$$

donde X es la matriz de diseño de las variables explicativas y β el vector de los coeficientes de regresión de estas variables.

El problema que puede surgir cuando se trabaja con la regresión de Poisson es la presencia del fenómeno conocido como «sobredispersión»[11], que se presenta cuando la variabilidad de los datos es superior a la variabilidad asumida por el modelo de Poisson; es decir, cuando la varianza de los recuentos observados es mayor que su media. La sobredipersión puede aparecer por no haber tenido en cuenta variables explicativas relevantes, o por la presencia de correlación espacial entre los SMR de las regiones. En el primer caso se denomina «sobredispersión no estructurada», mientras que en el segundo se habla de «sobredispersión estructurada espacialmente».

La correlación espacial aparece por el hecho de que regiones próximas comparten factores de riesgo desconocidos que no comparten con regiones más alejadas. Estos factores de riesgo pueden asociarse a factores medioambientales, sociales o culturales. Por tanto, esta correlación espacial implica que el riesgo de una región está condicionado por el riesgo de las regiones vecinas, por tanto estas regiones tenderán a presentar riesgos similares.

La principal consecuencia que tiene la presencia de sobredispersión es que las estimaciones, tanto de los SMR como de su error estándar, no serán correctas. Por lo tanto, es necesario utilizar un procedimiento que sea capaz de tener en cuenta esta sobredispersión. Este procedimiento es el modelo lineal generalizado mixto (GLMM)[12].

Modelos lineales generalizados mixtos

Los GLMM son una extensión de los modelos lineales generalizados en los que se incorporan efectos aleatorios. En este caso particular, se añade el efecto aleatorio región en la regresión de Poisson, obteniéndose la siguiente ecuación:

$$\log(\mu) = \log(E) + X\beta + Zb,$$

siendo b el vector de los coeficientes de los efectos aleatorios y Z la matriz de diseño de estos efectos aleatorios. El efecto aleatorio región se añade con la finalidad de capturar la sobredispersión observada. En función del tipo de sobredispersión que se quiere controlar, se pueden especificar 3 modelos para los efectos aleatorios: el de heterogeneidad, el autorregresivo condicional intrínseco (CAR intrínseco)[13,14] y

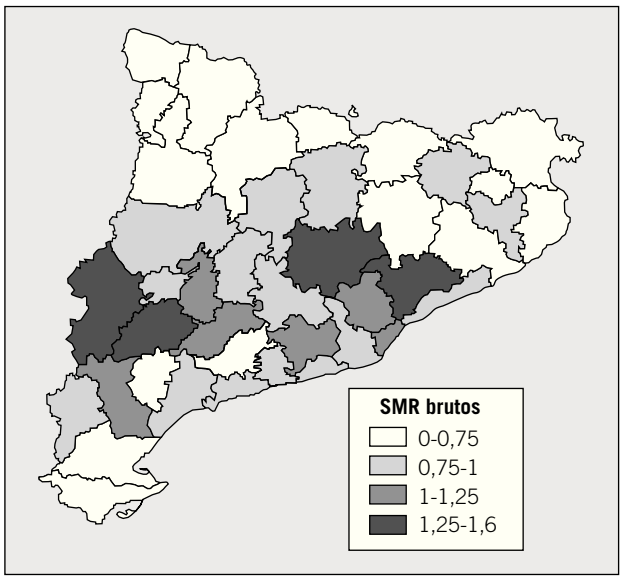

Fig. 1. Distribución espacial de las razones estandarizadas de morbilidad brutas (SMR brutos) de incidencia de diabetes tipo 1 de la población de menores de 30 años durante el período 1989-1998 en las comarcas de Cataluña.

el autorregresivo condicional no intrínseco (CAR no intrínseco)[15,16].

El modelo de heterogeneidad permite tener en cuenta la sobredispersión no estructurada, el modelo CAR intrínseco, la sobredispersión estructurada espacialmente y el modelo CAR no intrínseco es una conjunción de los 2 modelos anteriores y permite modelar la sobredispersión no estructurada y la sobredispersión estructurada.

Las técnicas de estimación de los parámetros de los GLMM se pueden llevar a cabo bajo dos perspectivas: la bayesiana y la frecuentista.

La estimación de los parámetros mediante la estadística bayesiana se realiza a partir de la distribución conocida como «distribución posterior de los parámetros». Esta distribución se obtiene de combinar la información de la muestra, recogida en la función de verosimilitud, con los conocimientos previos que tiene el investigador de los parámetros desconocidos que se recogen en las distribuciones *a priori*.

La obtención de la distribución posterior, cuando las integrales a resolver no se pueden obtener de una forma analítica sencilla, se realiza mediante técnicas de simulación, como el Gibbs Sampling[17,18]. Mediante esta técnica se obtiene una muestra de valores de la distribución posterior de los parámetros sin la necesidad de resolver las integrales y la estimación de los parámetros se realiza a partir de estadísticos descriptivos como, por ejemplo, la media de los valores simulados.

Las estimaciones con la perspectiva frecuentista se obtienen maximizando la función de verosimilitud. Esta maximización no siempre se puede resolver analíticamente. En estos casos se han descrito diferentes métodos para obtener estimaciones de los parámetros entre los que se puede destacar la casi verosimilitud penalizada[19].

En la actualidad, para estimar los parámetros de los modelos usando la estadística bayesiana se puede utilizar el programa estadístico de libre distribución WinBUGS. Este programa está disponible para uso público en la página web: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. En cambio, si deseamos utilizar métodos frecuentistas, es necesario que el propio investigador realice sus programas para obtener las estimaciones.

Para seleccionar el modelo que mejor se ajusta a los datos analizados, se utilizan medidas de bondad de ajuste como por ejemplo el estadístico DIC (*Deviance Information Criterion*[20]). Mediante este estadístico el modelo que ajustará mejor a los datos será el que tenga un DIC menor.

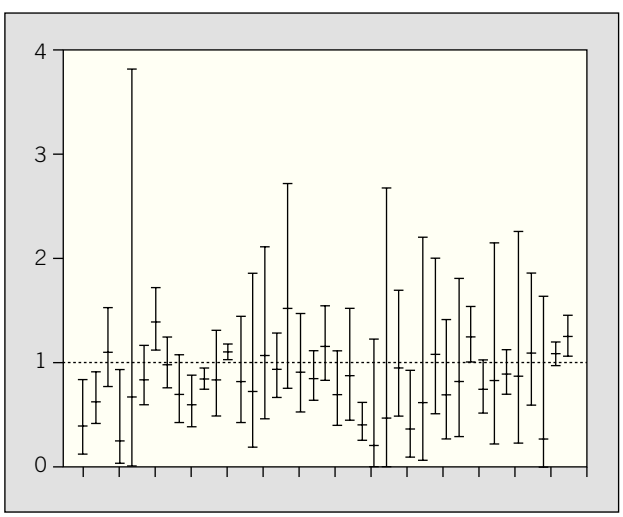

Fig. 2. Intervalos de confianza del 95% de las razones estandarizadas de morbilidad brutas (SMR brutos) de la incidencia de diabetes tipo 1 de la población de menores de 30 años durante el período 1989-1998 en las comarcas de Cataluña.

Ejemplo: datos de la incidencia de diabetes tipo 1 en la población de menores de 30 años de Cataluña durante el período 1989-1998

En el siguiente ejemplo se analizan los datos de diabetes de tipo 1 de Cataluña. Estos datos corresponden a casos declarados y confirmados en el registro de diabetes entre los años 1989 y 1998 de personas con una edad inferior a 30 años. El objetivo del estudio es analizar la incidencia de la diabetes de tipo 1 a lo largo del territorio. Para ello los datos se han agregado para las 41 comarcas de Cataluña y la población en riesgo para cada comarca se ha obtenido del padrón poblacional del año 1996. Estos datos se han estandarizado para poder eliminar el efecto confusor de las variables género y edad.

En las figuras 1 y 2 se representan las razones estandarizadas de incidencia para cada comarca y los intervalos de confianza del 95%, que se han construido utilizando la distribución de Poisson[21]. En relación con la figura 1, se puede observar que aproximadamente la mitad de las comarcas (21) presentan SMR próximos a 1, entre 0,75 y 1,25. Además, se aprecia que la mayoría de las comarcas con menor riesgo (SMR < 0,75) están situadas en el norte de Cataluña y que sólo 4 presentan riesgos mayores de 1,25.

En la figura 2 se observa una gran amplitud en algunos de los intervalos de confianza, lo que indica que estas estimaciones de los SMR brutos presentan mucha variabilidad. Este hecho, tal como se ha explicado en «Metodología estadística», se debe a la diferencia de población en riesgo entre las comarcas. Así, para poder estabilizar y suavizar estas estimaciones es necesario ajustar un modelo de regresión de Poisson.

A partir del modelo se obtiene una sobredispersión de 3,2, superior a 1, lo que implica que la variabilidad de los datos es superior a la asumida por el modelo. Para tener en cuen-

TABLA 1

Estimación del *deviance information criterion* (DIC) para los modelos espaciales de heterogeneidad, autorregresivo condicional intrínseco (CAR intrínseco) y autorregresivo condicional no intrínseco (CAR no intrínseco)

Modelo	DIC
Heterogeneidad	61,3
CAR intrínseco	61,38
CAR no intrínseco	59,24

ta esta sobredispersión se considera un modelo lineal generalizado mixto con un efecto aleatorio comarca. La estimación de este modelo se realiza mediante métodos bayesianos utilizando el programa WinBUGS.

Se consideran los 3 modelos propuestos en «Modelos lineales generalizados mixtos»: heterogeneidad, CAR intrínseco y CAR no intrínseco. Utilizando el DIC como criterio, se llega a la conclusión de que el modelo CAR no intrínseco es el que mejor ajusta los datos (tabla 1). Esto significa que el modelo considera que existe una sobredispersión tanto estructurada espacialmente como no estructurada.

En las figuras 3 y 4 se representan los SMR estimados a partir del modelo CAR no intrínseco y sus intervalos de confianza del 95%. Como se puede observar, los SMR se han suavizado y tan sólo uno de ellos se encuentra en el intervalo de más de 1,25 (fig. 3). Además, también se han obtenido unos intervalos de confianza más precisos, es decir, al haber disminuido la variabilidad de las estimaciones se ha conseguido que éstas sean más representativas del riesgo real.

En la figura 5 se representa el mapa de Cataluña en función de si el SMR de cada comarca es significativamente distinto de 1, es decir, que tienen un riesgo significativamente superior o inferior al riesgo general. Mediante esta representación se puede observar que 3 comarcas presentan un SMR significativamente superior a 1, mientras que 8 tienen un SMR significativamente inferior a 1.

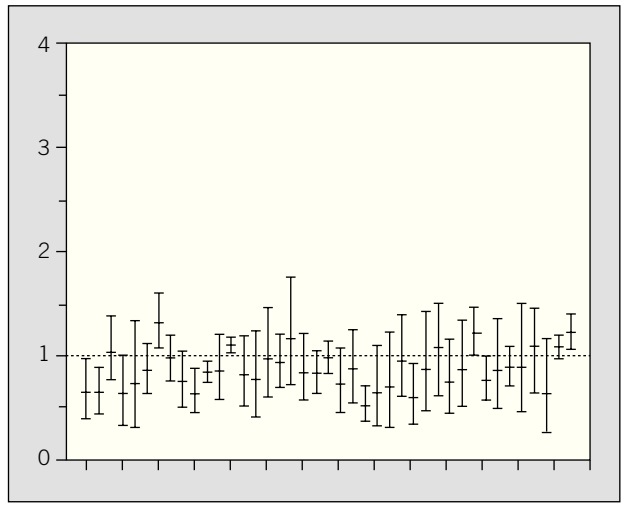

Fig. 4. Intervalos de confianza del 95% de las razones estandarizadas de morbilidad (SMR) de la incidencia de diabetes tipo 1 de la población de menores de 30 años durante el período 1989-1998 en las comarcas de Cataluña estimadas con el modelo autorregresivo condicional no intrínseco.

La representación de los SMR estimados no muestra claramente un patrón de distribución espacial, debido a que también existe un efecto de heterogeneidad. No obstante, se identifica un agrupamiento de las comarcas con más riesgo.

Discusión y conclusiones

El objetivo de este trabajo ha sido mostrar las técnicas que se emplean para modelar la distribución espacial del riesgo en un área geográfica. En estos casos la estimación bruta del riesgo se ha mostrado muy variable debido a la heterogeneidad de la población en riesgo, provocando que las

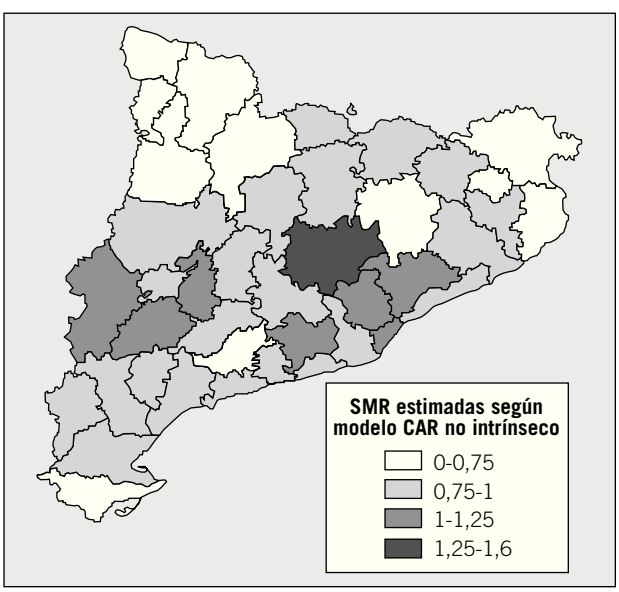

Fig. 3. Distribución espacial de las razones estandarizadas de morbilidad (SMR) de la incidencia de diabetes tipo 1 de la población de menores de 30 años durante el período 1989-1998 en las comarcas de Cataluña estimadas con el modelo autorregresivo condicional no intrínseco (CAR no intrínseco).

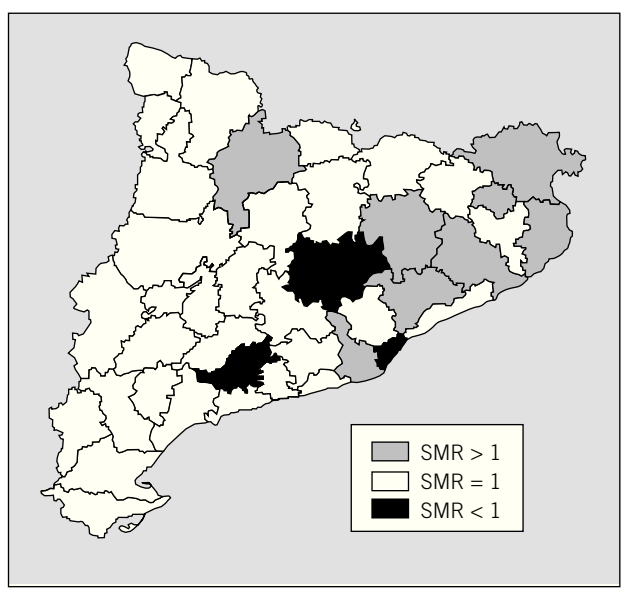

Fig. 5. Distribución espacial de las razones estandarizadas de morbilidad de la incidencia de diabetes tipo 1 de la población de menores de 30 años durante el período 1989-1998 en las comarcas de Cataluña estimadas con el modelo autorregresivo condicional no intrínseco, clasificadas en tres categorías: SMR significativamente mayores de 1, iguales a 1 y menores de 1.

estimaciones sean poco representativas del riesgo real. Mediante un GLMM, y concretamente una regresión de Poisson con la región como efecto aleatorio, se consigue suavizar las estimaciones, controlar por variables de confusión y tener en cuenta la posible sobredispersión de los datos debida tanto a la heterogeneidad del riesgo en una región como a la presencia de correlación espacial. Así, utilizando este procedimiento se obtienen unas mejores estimaciones del riesgo y de su error estándar.

El riesgo de las regiones sólo presenta un patrón de distribución espacial cuando se incluye en el modelo la sobredispersión estructurada espacialmente. En este caso este componente espacial se puede interpretar como que existen agrupaciones de regiones con un riesgo común. Este hecho viene provocado por la existencia de factores de riesgo que comparten estas regiones.

La representación de las medidas de riesgo mediante un mapa ayuda a la interpretación y a la búsqueda de estas agrupaciones de regiones o la identificación de regiones con mayor riesgo, facilitando la toma de decisiones e intervenciones sanitarias a nivel de región.

Agradecimientos

Nuestro agradecimiento a la Dra. C. Castell y al Dr. R. Tresserras por facilitarnos los datos del Registre de Diabetis mellitus tipo 1 de Catalunya del Consell Assessor sobre la Diabetis a Catalunya del Departament de Sanitat i Seguretat Social y a l'Associació Catalana de Diabetis.

REFERENCIAS BIBLIOGRÁFICAS

1. Organización Panamericana de la Salud. Uso de sistemas de información geográfica en epidemiología. Boletín Epidemiológico de la Organización Panamericana de la Salud 1996; 17(1). Disponible en: http://www.paho.org/Spanish/SHA/epibul_95-98/bs961sig.htm
2. Borja-Aburto VH. Estudios ecológicos. Salud Pública de México 2000; 42:533-8.
3. Gesler W. The uses of spatial analysis in medical geography: a review. Social Science & Medicine 1986;23:963-73.
4. Salazar A, Guilar S, Melchor I, Castaño B, Gil J, Sanz M, et al. Sistema de información geográfica en salud pública. Una herramienta para la vigilancia. Boletín Epidemiológico de España 1999;7:181-8.
5. Martínez P, Vidaurre M, Nájera P, Loyola E, Castillo-Salgado C. SIGEpi: Sistema de Información Geográfica en Epidemiología y Salud. Boletín Epidemiológico de la Organización Panamericana de la Salud [revista electrónica]. 2001;22. Disponible en: http://www.paho.org/Spanish/SHA/be_v22n3-cover.htm.
6. ArcGis [Página oficial]. Disponible en: http://www.esri.com/soffware/arcgis
7. Livi-Bacci M. Introducción a la demografía. Barcelona: Arial, 1993.
8. Kahn HA, Sempos CT. Statistical methods in epidemiology. Monographs in Epidemiology and Bioestatistics. Vol. 12. Oxford: Oxford University Press, 1989.
9. Nelder JA, Wedderburn RWM. Generalized linear models. Journal Royal Statistical Society 1972;135:370-84.
10. Enciclopedia of Statistical Sciences. Vol. 7. New York: Board, 1981.
11. Breslow NE. Extra-Poisson variation in log-linear models. Applied Statistics 1984;33:38-44.
12. McCulloch CE, Searle SR. Generalized, linear, and mixed models. New York: Wiley, 2001.
13. Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society 1974;B36:192-236.
14. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987;43:671-81.
15. Besag J, York J, Mollié A. Bayesian image restoration, with applications in spatial statistics. Annals of the Institute of Statistical Mathematics 1991;43:1-59.
16. Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. En: Halloran ME, Berry D, editors. Statistical models in epidemiology, the environment, and clinical trials. Springer, 1999.
17. Clayton D, Bernardinelli L. Bayesian methods for mapping disease risks. En: Elliot P, Wakefield JC, Bert NG, Briggs DJ, editors. Small area studies in geographical and environmental epidemiology. Oxford: Oxford University Press, 1992; p. 205-20.
18. Bernardinelli L, Montomoli C. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risks. Statistics in Medicine 1992;11:983-1007.
19. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993;88:9-25.
20. Spiegelhalter DJ, Best NG, Carlin BP. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Bioestatistics, University of Minnesota. Minneapolis: 1998.
21. Documenta Geigy Scientific. Tables. 7 ed. Basel: Ciga-Geigy, 1970.