Ugarte, M. D.
Cuesta, C.
Isern, G.
Barbona I.
Lupachini, E.
Fantasía Fiorela
Instituto de Investigaciones Teóricas y Aplicadas en Estadística (IITAE)

AJUSTE DE MODELOS "CAR" PARA LA ESTIMACION ESPACIO-TEMPORAL DE EVENTOS

Introducción

La representación cartográfica de enfermedades (conocida en ingles como "disease mapping") tiene como finalidad conocer la distribución geográfica de enfermedades e identificar los factores de riesgo que podrían explicarlas.

Estos métodos en general se utilizan para describir tasas estandarizadas que implican la comparación de los casos observados en un área con aquellos casos que esperaríamos observar si el riesgo de morir o enfermar en esa zona fuese el mismo que en cierta población de referencia. Estas tasas pueden ser muy inestables cuando se estudian enfermedades raras o las áreas están poco pobladas. Esto puede distorsionar el patrón real de la enfermedad cuando dichas tasas se presentan en un mapa.

Las técnicas de modelización espacial han jugado un papel importante en la literatura epidemiológica (por ejemplo para detectar áreas con riesgos extremos). En este sentido, los modelos condicionales autorregresivos (CAR) han sido y siguen siendo una herramienta útil que permiten suavizar los riesgos teniendo en cuenta una posible sobredispersión de los datos.

El estudio de la evolución geográfica de patrones de mortalidad o incidencias provee mucha información a los epidemiólogos y puede llevarse a cabo como una extensión del estudio a nivel espacial. Cuando se estudia la evolución de los patrones espaciales de una enfermedad a través del tiempo puede ocurrir que el patrón geográfico de la enfermedad no cambie a través del tiempo (es decir que la evolución de las tasas es similar en todas las áreas) o que cada área presente diferentes cambios en las tasas a través del tiempo (en este caso, se dice que hay interacción espacio-tiempo). Un modelo con interacción espacio-tiempo es mas complejo que uno que incluye las componentes de tiempo y espacio en forma aditiva.

El modelo CAR espacial puede extenderse al caso espacio-temporal incluso agregando un termino de interacción espacio-temporal (aunque esto complejice el modelo y sea computacionalmente mas "costoso" que un modelo sin interacción).

En este trabajo se presentará la metodología de modelos CAR y se propondrá su uso para la aplicación a datos de mortalidad infantil en la República Argentina.

Metodología

Tasas crudas vs tasas ajustadas

La representación de las tasas crudas no permite la comparación entre áreas dado que las diferencias entre la población de áreas distintas pueden ocurrir debido a otros factores de riesgo que no han sido tenidos en cuenta (por ejemplo la edad). Para poder comparar las tasas se recurre a la estandarización la cual puede hacerse mediante dos métodos: directo e indirecto. En general para la representación de las enfermedades se usa el método indirecto. Esto involucra la comparación entre los casos observados en un área con aquellos esperados si el riesgo (para cada grupo de edad) fuera el mismo que en una población de referencia. El cociente observado/esperado se denomina tasa de mortalidad estandarizada (o tasa de incidencia estandarizada) y es una estimación del riesgo relativo de muerte (o incidencia) en un área en relación a una población de referencia. Si el riesgo relativo es mayor a 1 en una cierta área, indica que el área es de mayor riesgo que la región completa (para ello es necesario construir intervalos de confianza, si el limite inferior es mayor que 1, significa que el área es de alto riesgo).

Esta medida de riesgo relativo puede ser inestable si la enfermedad es poco frecuente (rara) o si hay áreas que contienen un número pequeño de personas (problema de estimación en "áreas pequeñas") y pueden distorsionar el patrón real de la enfermedad cuando se representa en un mapa.

Modelo espacio-temporal con distribución condicional autorregresiva (CAR)

Supongamos que el área bajo estudio está dividida en n regiones contiguas $(i = 1, \dots, n)$ y los datos están disponibles para T períodos de tiempo $(t = 1, \dots, T)$ y que para un área dada, O_{it} y E_{it} son el número de casos observados y esperados respectivamente en el área i, período t.

En este contexto el interés recae en estimar el riesgo relativo r_{it} (a menudo denominado SMR por "Standardized Mortality Ratio") para cada área i y período t.

El estimador máximo verosímil de r_{it} es: $\hat{r}_{it} = \frac{O_{it}}{E_{it}}$ y su variancia es $var(\hat{r}_{it}) = \frac{Var(O_{it})}{E_{it}^{\,2}} = \frac{r_{it}}{E_{it}}$ y entonces $\widehat{var}(\hat{r}_{it}) = \frac{O_{it}}{E_{it}^{\,2}}$.

Asumiendo que $O_{it} \sim Poisson(\mu_{it} = E_{it}\, r_{it})$, $\log(\mu_{it}) = \log(E_{it}) + \log(r_{it})$.

La especificación de $\log(r_{it})$ da origen a los distintos modelos de mapeo de enfermedades. Entre ellos los modelos CAR (modelos condicionales autorregresivos) . Estos modelos fueron introducidos por Besag (1974) y se difundieron mucho en su aplicación al mapeo de enfermedades y en particular a datos de conteo. Se trata de modelos jerárquicos que tienen en cuenta la estructura espacial (y/o temporal).

Se asume que O_{it} tiene una distribución Poisson condicional al efecto aleatorio de la región, el número de casos observados en cada área y período de tiempo. Entonces, el logaritmo

del riesgo se modela como:

$$u_{it} = \log(r_{it}) = \beta + \emptyset_i + \gamma_t + \delta_{it}$$

donde β es un nivel de riesgo global, $\emptyset_i$representa los efectos espaciales, $\gamma_t$representa los efectos temporales y δ_{it} representa los efectos de interacción espacio-tiempo. Matricialmente, $\boldsymbol{u} = \mathbf{X}\beta + \mathbf{Z}\boldsymbol{\alpha}$, donde $\mathbf{X}$ es una columna de unos, $\mathbf{Z}$ es una matriz de dimensión nt x $(n + t + nt)$, $\boldsymbol{\alpha} = (\boldsymbol{\phi}', \boldsymbol{\gamma}', \boldsymbol{\delta}') \sim N(\mathbf{0}, \boldsymbol{G})$**.** Donde $\boldsymbol{G} = \text{diag}(\sigma_s^2 \boldsymbol{D}_s, \sigma_s^2 \boldsymbol{D}_t, \sigma_{st}^2 \boldsymbol{D}_{st})$ **;** siendo,

$$\boldsymbol{D}_s = (\lambda_s \boldsymbol{Q}_s + (1 - \lambda_s)\, \boldsymbol{I}_s)^-$$

$$\boldsymbol{D}_t = \boldsymbol{Q}_t^-$$

$$\boldsymbol{D}_{st} = \boldsymbol{Q}_t^- \otimes \boldsymbol{Q}_s^-,$$

donde,

$\boldsymbol{Q}_s$ está determinado por la estructura de vecinos en el espacio, los elementos diagonales de la matriz son iguales al número de vecinos en la región i-ésima para $i \neq j$, y los elementos fuera de la diagonal principal toman el valor -1 si i es vecino de j y toman el valor cero si i no es vecino de j. Es necesario notar que cuando λ_s toma valores entre 0 y 1, cuando λ_s=0 no hay variabilidad espacial y cuando λ_s=1 toda la variabilidad es espacial.

$\boldsymbol{I}_s$ es una matriz identidad de orden nxn

$\boldsymbol{Q}_t$ está determinada por una estructura de vecinos en el tiempo, los elementos de la diagonal principal toman el valor 2 y el resto de los elementos toman el valor -1 si se trata de dos períodos contiguos o cero en otro caso. Esta matriz indica que la distribución del efecto temporal es un camino aleatorio de primer orden.

A los efectos de estimación se utiliza el algoritmo de cuasi-verosimilitud penalizada (PQL) que es un procedimiento aproximado para algunos modelos lineales generalizados. El método de estimación PQL se basa en una serie de aproximaciones al modelo mixto normal usando series de Taylor de la función de enlace. La característica principal de PQL es que provee estimaciones puntuales adecuadas, es computacionalmente simple y tiene muy pocos problemas de convergencia. El algoritmo consiste en definir un vector de trabajo en correspondencia con un modelo lineal mixto gausiano. Las componentes del vector de trabajo son:

$$Y_{it} = \eta_{it} + (O_{it} - \mu_{it})\, g'(\mu_{it}) - \log(E_{it}),$$

donde $\mu_{it} = E_{it} \exp(r_{it})$, $\eta_{it} = g(\mu_{it}) = \log(E_{it}) + \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{z}_{it}\boldsymbol{\alpha}$; $g'(\mu_{it}) = {}^1/_{\mu_{it}}$;

El modelo mixto asociado es $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Con $\boldsymbol{X}$ y $\boldsymbol{Z}$ definidas mas arriba y $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{W}^{-1})$ y $\boldsymbol{W} = \text{diag}(\mu_{it})$.

Los efectos fijos se estiman a partir de $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}$ con variancia asintótica $(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})$ y $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$**.** Los efectos aleatorios se predicen como $\widehat{\boldsymbol{\alpha}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ (teniendo en cuenta la distribución condicional de α/Y). Las componentes aleatorias de la

variancia se estiman por Máxima Verosimilitud Restringida (REML).

Dados estimaciones iniciales de los parámetros, el procedimiento PQL estima $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$, Luego se actualiza las estimaciones de los parámetros de las matrices de variancia a través de las ecuaciones REML. El proceso se repite hasta la convergencia. Finalmente se estima el riesgo relativo como:

$$\hat{r}_{it} = \exp(\boldsymbol{x}_{it}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_{it}\widehat{\boldsymbol{\alpha}})$$

Aplicación

La metodología propuesta es evaluada con datos de Mortalidad Infantil Anual, cuya tasa se define como:

$$TMI = \frac{\begin{array}{c}número\ de\ muertes\ de\ menores\ de\ 1\ año\ de\ edad\ acaecidas\ en\ la\ población\ de\ un \\ área\ \ geográfica\ dada\ durante\ un\ año\ dao\end{array}}{\begin{array}{c}número\ de\ nacidos\ vivos\ registrados\ en\ la\ población\ del \\ área\ geográfica\ dada\ durante\ el\ mismo\ año\end{array}}$$

Los datos a analizar corresponden a las 24 provincias de la República Argentina y están disponibles para los años 1980 a 2011.

Esta tasa es considerada como uno de los indicadores más importantes para la planificación y programación de actividades en salud, por lo tanto, un estudio minucioso es de gran utilidad para los hacedores de políticas públicas en salud. Por otro lado, su estudio contribuye a la evaluación del cumplimiento de los Objetivos de Desarrollo del Milenio planteados por las Naciones Unidas (y adoptado por la República Argentina) cuyo vencimiento del plazo está previsto para el año 2015. El cuarto de dichos objetivos tiene como meta general reducir en dos terceras partes (entre 1990 y 2015) la mortalidad en niños menores de cinco años (y en particular, en Argentina se agrega como meta, reducir el 10% la desigualdad entre provincias). Uno de los indicadores propuesto para la evolución del cumplimiento de estos objetivos es la tasa de mortalidad infantil que será el objeto de nuestro estudio.

La estandarización de esta tasa se lleva a cabo considerando como referencia a la población completa de la República Argentina. La cantidad de casos esperados para la Rep Arg en la provincia “i” tiempo “t” se obtiene como el número de muertes de menores de 1 año de edad acaecidas en la población de un área geográfica durante un año dado multiplicada por la TMI global de la Rep Arg para ese año.

Antes de comenzar con el análisis de evaluó el *software* a utilizar. Se compararon las ventajas y desventajas del programa libre R y del programa SAS. A los efectos de las representaciones gráficas, el primero resultó más versátil y simple. Aún cuando la versión 9.3 de SAS ha incluido mapas en sus registros y mejorado su capacidad gráfica, no logra presentar las ventajas de R tales como permitir construir mapas comparativos (con las mismas escalas de color), tamaño reducido y buena definición de los mapas. Por otro lado, en los mapas predefinidos se SAS, no se cuenta con Ciudad Autónoma de Buenos Aires (CABA) como un territorio separado (tal como lo hace el ministerio de Salud de la Nación). Otro problema lo presenta el territorio de Islas Malvinas que se incluye como un territorio independiente de Tierra del Fuego. Por ello se opta para las construcciones gráficas utilizar el programa R (ver programa en Anexo).

Como un primer análisis descriptivo, se construyeron mapas con las tasas originales (a mo-

do de ejemplo en la Figura 1 se muestran los mapas para los años 2006 a 2011). En los mismos se ven cómo, en líneas generales, la TMI aumenta de sur a norte. Sin embargo, por tratarse de valores observados, los mismos presentan una variabilidad entre provincias que podría ser suavizada a partir de un modelo estadístico.

También se observan provincias con mayor variabilidad a través del tiempo como por ejemplo Formosa, Chaco, La Pampa y otras con mayor estabilidad como Buenos Aires, Neuquén, Córdoba. Santa Fe. En la Figura 2 se muestran las series históricas correspondientes a algunas provincias (Chubut, CABA, Formosa, San Luis). En todas ellas se ve un leve decrecimiento, sin embargo, es imperioso el ajuste de un modelo que permita estimar estas tendencias y probar si las mismas son iguales en todos los territorios de nuestro país.

Figura 1. Tasa **d**e mortalidad infantil, período 2006-2011 (x1000)

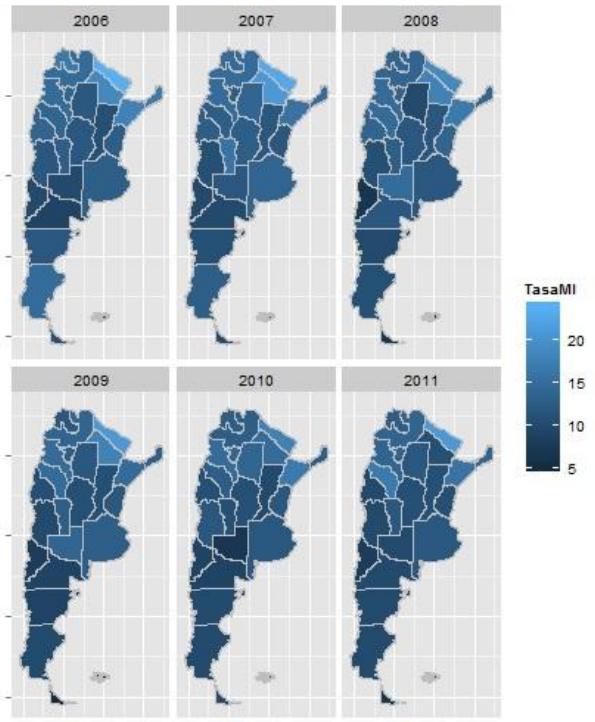
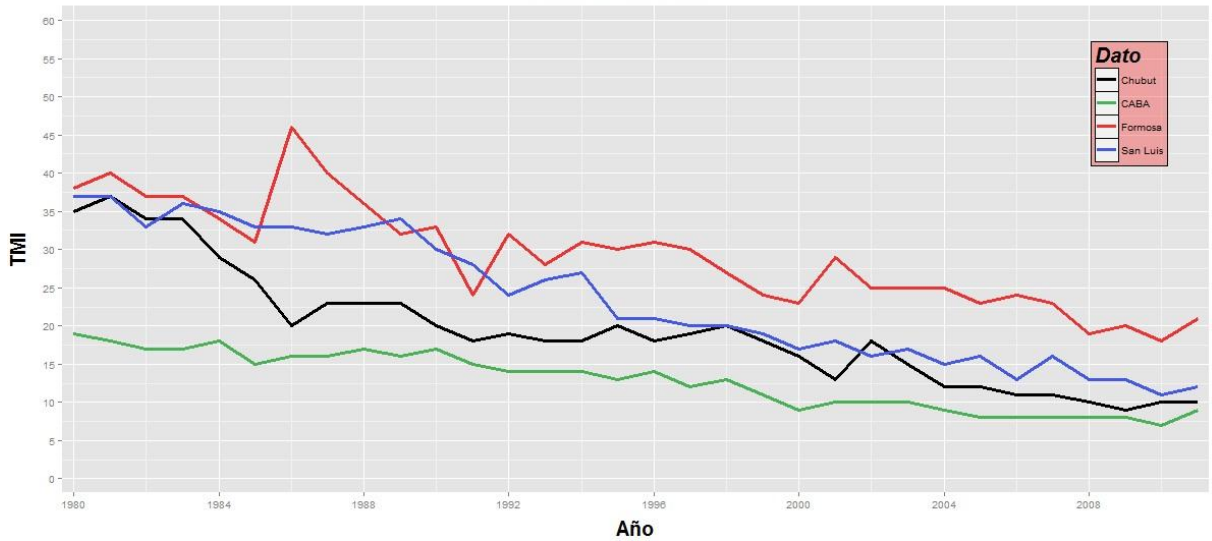


Figura 2. Serie histórica (1980-2011) de la TMI para las provincias Chubut, CABA, Formosa y San Luis



Al momento de presentación de este escrito, se está trabajando en el ajuste de los modelos CAR, sin embargo aún no se tienen resultados.

Discusión

Al momento de presentar este escrito aún no se han realizado los ajustes del modelo CAR, sin embargo se espera que el mismo sea capaz de suavizar algunas observaciones espurias así como proveer herramientas para realizar predicciones confiables.

Se propone, luego del ajuste del modelo CAR, explorar otros modelos alternativos para el tratamiento de este tipo de información, por ejemplo hacer uso de modelos mixtos generalizados aditivos (usando P-*spline)*. Asimismo, otros métodos de estimación podrían proponerse como por ejemplo los métodos bayesianos (INLA).

ANEXO

Sentencias utilizadas para la construcción de los mapas en R

```
library(sp)
library(maptools)
library(ggplot2)
library(gpclib)
library(plyr)
library(reshape)
setwd("C:/ ")
arg<- readShapePoly("paisprov2012.shp")
data<-read.csv2("C:/MI.csv",header=T,sep=";")
y1<-data
y1<-rename(y1,c(X2006="2006"))
y1<-rename(y1,c(X2007="2007"))
y1<-rename(y1,c(X2008="2008"))
y1<-rename(y1,c(X2009="2009"))
y1<-rename(y1,c(X2010="2010"))
y1<-rename(y1,c(X2011="2011"))
gpclibPermit()
tmi<- fortify(arg, region="NOMPROV")
tmi<- merge(tmi, y1, by.x="id", by.y="NOMPROV")
argentina.data.melt <- melt(y1, id=c("NOMPROV"))
plot.serie <- merge(tmi, argentina.data.melt,
                    by.x="id", by.y="NOMPROV")
plot.serie<-rename(plot.serie,c(value="TasaMI"))
plot.serie <- plot.serie[order(plot.serie$order), ]
mapita<-ggplot(data = plot.serie, aes(x = long, y = lat, fill =
                                  TasaMI, group = group))+ geom_polygon()+
  geom_path(colour="grey", lwd=0.1)+ ggtitle ("Tasa de mortalidad infantil período 2006-2011 (x1000)")
+
  coord_equal()+facet_wrap(~variable)
```

REFERENCIAS BIBLIOGRÁFICAS

Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussions). Journal of the Royal Statistical Society, B, 36, 192-236

Congdon P. (2006) . A spatio-temporal forecasting approach for helth indicator. Journal of Data Science, 4(4), 413-424.

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

Goicoa T, Ugarte MD, Etxeberria J, Militino AF (2012) Comparing CAR and P-spline models in spatial disease mapping. Environmental and Ecological Statistics: DOI 10.1007/s10651-012-0201-8

Lawson AB (2006) Statistical Methods in spatial epidemiology. 2nd ed. New Jersey: John Wiley &sons

Schroedle, B. and Held, L. (2011a). A primer on disease mapping and ecological regression using.Computational statistics, 26(2):241.258.

Singh BB, Shukla GK, Kundu D (2005) Spatio-temporal models in small área estimation. Survey Methodology, 31, 183-195

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

Ugarte MD, Goicoa T, Militino AF (2010a) Spatio-temporal modeling of mortality risks using penalized splines. Environmetrics,21,270-289

Ugarte MD, Goicoa T, Etxeberria J, Militino AF (2012b). A P-spline ANOVA-type model in space-time disease mapping. Stochastic Environmental Research and Risk Assessment,26,835-845

FUENTES

Secretaría de políticas, regulación e Institutos. Dirección de Estadísticas e Información de la Salud. Ministerio de Salud de la República Argentina. Buenos Aires, Argentina. 2013.