*VI Congreso Galego de Estatística e Investigación de Operacións*
*Vigo 5-7 de Novembro de 2003*

# UN ANÁLISIS DE LOS EFECTOS FRONTERA EN EPIDEMIOLOGÍA ESPACIAL

Carmen L. Vidal Rodeiro[1], Andrew B. Lawson[1]

[1]Department of Epidemiology and Biostatistics
University of South Carolina

**RESUMEN**

La construcción y la suavización de mapas de enfermedades ha sido, durante los últimos años, objeto de muchos avances metodológicos, pero el análisis de los "***efectos frontera***" ha sido un área poco desarrollada en epidemiología espacial. Esto es lamentable ya que muchos análisis pueden verse seriamente alterados por la inclusión de los efectos frontera en diferentes formas. Este trabajo pretende averiguar como la estimación del riesgo relativo de mortalidad por una enfermedad cerca o en las fronteras de la región de estudio puede verse afectada por los efectos frontera.

**Palabras clave:** datos perdidos, efectos frontera, MCMC, modelos bayesianos.

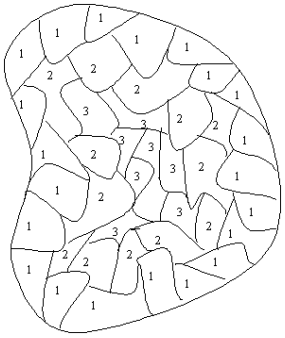
## 1. INTRODUCCIÓN

En epidemiología, los análisis espaciales se llevan a cabo en regiones finitas. Esto implica que hay un borde o frontera en la región de estudio y por tanto cualquier distribución espacial dentro de ella puede extenderse más allá de sus bordes (Griffith (1983)).
Cuando construimos mapas para estudiar la variación geográfica del riesgo de una enfermedad en áreas pequeñas, la información sobre las áreas vecinas es, muchas veces, incompleta. En particular, este es un problema que se encuentra en las áreas fronterizas.
Los efectos frontera pueden distorsionar las estimaciones del riesgo en la región de estudio y por ello deben ser acomodados en el análisis (Lawson (2001)). En la literatura, las estrategias más efectivas que han sido adoptadas para eliminar los efectos frontera son:

(a) Utilización de pesos que relacionan la posición de la observación con la frontera de la región. El peso de una observación actúa como sustituto del grado de información perdida en esa localización. Esta técnica es apropiada cuando solo una pequeña porción del área de estudio está cerca de la frontera.
(b) Construcción de un área interna anexa a la frontera. Esta área se utiliza en el proceso de estimación pero las estimaciones obtenidas en ella no se presentan como resultados lo que supone una gran perdida de información.
(c) Construcción de un área externa anexa a la frontera. A cada unidad de esta área se le asigna un valor (fijo u obtenido mediante un modelo de regresión) o se trata como un valor perdido. Esta última opción tiene grandes ventajas si se utilizan métodos de simulación como por ejemplo métodos MCMC (Gilks *et al*. (1996)).

El objetivo de este estudio es evaluar el comportamiento de una pequeña selección de métodos comúnmente empleados para el "mapeo de enfermedades" en áreas pequeñas cuando tenemos diferentes condiciones en la frontera de la región de estudio. Para eliminar los efectos frontera se considera un algoritmo MCMC basado en "data augmentation" (Tanner (1996)) que trata los datos no disponibles como parámetros del modelo que se estima.
Utilizaremos dicho algoritmo para estimar el riesgo de mortalidad en todas las áreas de la región de estudio mientras sucesivamente se tratan como datos perdidos el número de defunciones observadas en áreas concéntricas con la frontera. De esta forma sucesivos grados de censura pueden ser estudiados. La siguiente figura muestra un ejemplo de una región de estudio con tres áreas concéntricas.

Figura 1. Sucesivas áreas con datos perdidos en la región de estudio.



En este trabajo tenemos dos objetivos principales:

1) Estimación de los riesgos en las áreas donde tenemos valores perdidos en el número de defunciones.
2) Estudio de la propagación del error debido a los efectos frontera en las áreas internas de la región de estudio.

Para cumplir estos dos objetivos examinamos una base de datos sobre mortalidad por cáncer de pulmón en EEUU.

## 2. MODELOS

Supongamos que la región de estudio está dividida en $n$ áreas. Sea $\xi_i$ el riesgo relativo de mortalidad en el área $i$, $i$=1,…,$n$ y sean ($O_1$,…,$O_n$) y ($E_1$,…,$E_n$) el número de defunciones observadas y esperadas, respectivamente.
En el estudio de la variación del riesgo de mortalidad de una enfermedad hay una serie de modelos básicos que se suelen utilizar como punto de partida (Lawson (2001)). En este trabajo se consideran cuatro modelos. En primer lugar, se asume que cada $O_i$ sigue una distribución de Poisson con parámetro $E_i\xi_i$. Este es el llamado modelo clásico. Los otros modelos considerados aquí (Poisson-gamma, lognormal y Besag, York y Mollié) son extensiones de este modelo cuando se asume una distribución a priori para el riesgo relativo. Estos modelos recorren simples y complejas estructuras para el riesgo incluyendo correlación espacial. A continuación se describen en detalle.

### 2.1 EL MODELO CLÁSICO

Esta aproximación está basada en que suponiendo conocidos los casos esperados $E_i$s, los $\xi_i$s son mutuamente independientes y cada $O_i$ sigue una distribución de Poisson con media $E_i\xi_i$. Bajo estas

condiciones, el estimador de máxima verosimilitud de $\xi_i$ coincide con la Razón de Mortalidad Estandarizada (RME):

$$\hat{\xi}_i = RME_i = \frac{O_i}{E_i}$$

La RME se utiliza mucho en el mapeo de enfermedades pero tiene varios inconvenientes. Por ejemplo, cuando se calculan para áreas en las cuales los casos observados y esperados son escasos, sea debido a que el desenlace de interés es "raro", a que las áreas sobre las que se trabaja son pequeñas o, incluso, a ambas circunstancias, se suelen producir estimaciones del riesgo relativo muy extremas (muy bajas o muy altas en relación con las demás) que van ubicándose de manera caótica en el mapa hasta el punto de obstaculizar la interpretación epidemiológica.
Los métodos bayesianos ofrecen la posibilidad de "corregir" estos mapas y, dada la estabilidad que alcanzan los estimadores, propician que emerjan estructuras que no pueden apreciarse directamente con los procedimientos estadísticos clásicos (Mollié (1996)).

**2.2 EL MODELO POISSON-GAMMA**

Una distribución a priori apropiada para los riesgos relativos es la distribución gamma. Si esta distribución es gamma($\nu$,$\tau$) entonces, el riesgo relativo tiene la siguiente distribución a posteriori:

$$\xi_i \sim \text{gamma}(\nu + O_i, \tau + E_i)$$

Este modelo puede extenderse introduciendo distribuciones a priori para los parámetros $\nu$ y $\tau$.

**2.3 EL MODELO LOGNORMAL**

Una distribución gamma como distribución a priori para los riesgos relativos en conveniente pero puede ser restrictiva porque no permite añadir covariables al análisis y no es posible incorporar correlación espacial. Un modelo lognormal es más flexible:

$$\log \xi_i = \theta_i$$
$$\theta_i \sim \text{normal}(0, \sigma_\theta)$$

donde $\sigma_\theta$ es una cantidad fija. Una alternativa a este modelo consiste en dotar a $\sigma_\theta$ con una distribución a priori.

**2.4 EL MODELO DE BESAG, YORK Y MOLLIÉ (BYM)**

El modelo lognormal puede extenderse de varias formas. Clayton y Kaldor (1987) introdujeron un modelo con dos componentes para el riesgo que fue desarrollado más tarde por Besag *et al*. (1991). En este modelo el logaritmo del riesgo relativo se modeliza de la siguiente forma:

$$\log \xi_i = \theta_i + \phi_i$$

donde $\theta_i$ es la componente de heterogeneidad (UH) y $\phi_i$ es la componente de clustering o correlación espacial (CH) en el área *i*, *i*=1, …, *n*.
Es necesario especificar distribuciones a priori para estas dos componentes del riesgo.
Para la componente de heterogeneidad se considera una distribución normal centrada en 0:

$$\theta_i \sim \text{normal}(0, \sigma_\theta) \quad \forall i$$

Para la componente de clustering se utiliza una estructura de correlación espacial tal que las estimaciones en un área dependan de las áreas vecinas. Así pues $\phi_i$ sigue una distribución normal con varianza inversamente proporcional al número de unidades adyacentes a la *i*-ésima y a cierto hiperparámetro λ. Concretamente:

$$\phi_i \sim normal(\overline{\phi_i}, \frac{1}{\lambda n_i})$$

donde $n_i$ es igual al número de vecinos del área *i* y $\overline{\phi_i} = \frac{1}{n_i} \sum_{j \in V} \phi_j$ (*V* es el conjunto de los vecinos del área *i*).

## 3. IMPLEMENTACIÓN DEL ALGORITMO

### 3.1 DISEÑO

Definimos el subconjunto de áreas externas con el subíndice $b$. El número de casos observados y el riesgo relativo para esas áreas se denota $\{O_b\}$ y $\{\xi_b\}$. Asumimos que el número de casos esperados $\{E_b\}$ es conocido para todas las áreas. Denotamos los parámetros del modelo de la siguiente manera: $\{O_b\}$, $\{\xi_b\}$, $\{\xi_{\setminus b}\}$ y $\boldsymbol{\lambda}$. El subíndice $\setminus b$ denota un área interna y $\boldsymbol{\lambda}$ es el vector de los restantes parámetros en el modelo.

Definimos el algoritmo de la siguiente manera:

(a1) $[\{O_b\} \mid \{O_{\setminus b}\},\{\xi_{(b\cup\setminus b)}\},\boldsymbol{\lambda}]$

(a2) $[\{\xi\} \mid \{O_{(b\cup\setminus b)}\},\boldsymbol{\lambda}]$

(a3) $[\boldsymbol{\lambda} \mid \{O_{(b\cup\setminus b}\},\{\xi_{(b\cup\setminus b)}\}]$

Estas distribuciones son estimadas utilizando métodos MCMC.

### 3.2 APLICACIÓN

El algoritmo fue aplicado a una región con un número suficiente de áreas pequeñas que permitió extraer subconjuntos anexos a la frontera hasta el nivel $m$, donde $m$=5. Dicha región consiste en 10 estados adyacentes en los EEUU: Colorado, Iowa, Kansas, Montana, Minnesota, Missouri, Nebraska, North Dakota, South Dakota y Wyoming. Estos estados están formados por 760 condados o áreas pequeñas.

Para estudiar como la estimación del riesgo relativo cerca de las fronteras se ve afectada por la posición de la frontera y como el error debido a los efectos frontera se propaga a medida que nos adentramos en la región, se consideraron los siguientes subconjuntos de la región de estudio. El subconjunto (set) 1 consiste en los condados más externos (fronterizos). El subconjunto 2 consiste en aquellos condados situados en un segundo anillo en el interior de la región. El subconjunto 3 consiste en los condados de un tercer anillo y así sucesivamente. En total hemos considerado $J$=1,…,11 subconjuntos. El subconjunto 11 lo forman los condados más internos.

Los datos de los subconjuntos 1 a 5 se eliminan durante la estimación. En los subconjuntos 6 a 11 se estudiará el comportamiento de los estimadores.

Cada uno de los modelos se ajustará en seis situaciones diferentes que llamamos “pasos” (steps).

- Paso 0: el modelo se ajusta con todos los datos
- Paso $k$, $k$=1,…,5: el modelo se ajusta con valores perdidos para los subconjuntos $g$, $g$=1,…,$k$.

## 4. RESULTADOS

La figura 2 muestra el rango de las estimaciones del riesgo relativo en los condados que pertenecen a los subconjuntos 6 a 11. Se puede observar que el comportamiento de esas estimaciones del riesgo depende del modelo, del paso y de la distancia del condado a la frontera de la región.

A pesar de que para los modelos considerados en este estudio las estimaciones parecen estables, es importante señalar que el modelo BYM es el más robusto. Además cabría esperar un comportamiento similar entre los modelos Poisson-gamma y lognormal pero de este gráfico se desprende un comportamiento más errático en el modelo lognormal.

Con respecto a los patrones geográficos, estos no cambian mucho cuando se utilizan los tres primeros modelos. Las propiedades de suavización del método MCMC se dejan ver, sobre todo, en las regiones fronterizas.

La figura 3 muestra la distribución espacial del riesgo relativo en la región de estudio para el modelo BYM. Los patrones son muy similares en los pasos 1 y 2 y se parecen al patrón obtenido en el caso 0, aunque el grado de suavización es mayor. En los restantes pasos el patrón cambia y encontramos una clara distinción entre el noroeste y el sureste de la región. Además los valores del riesgo relativo cambian mucho en el noroeste de la región dependiendo del paso de la estimación.

Figura 2: Rango de las estimaciones del riesgo relativo en los condados del interior (subconjuntos 6 a 11).

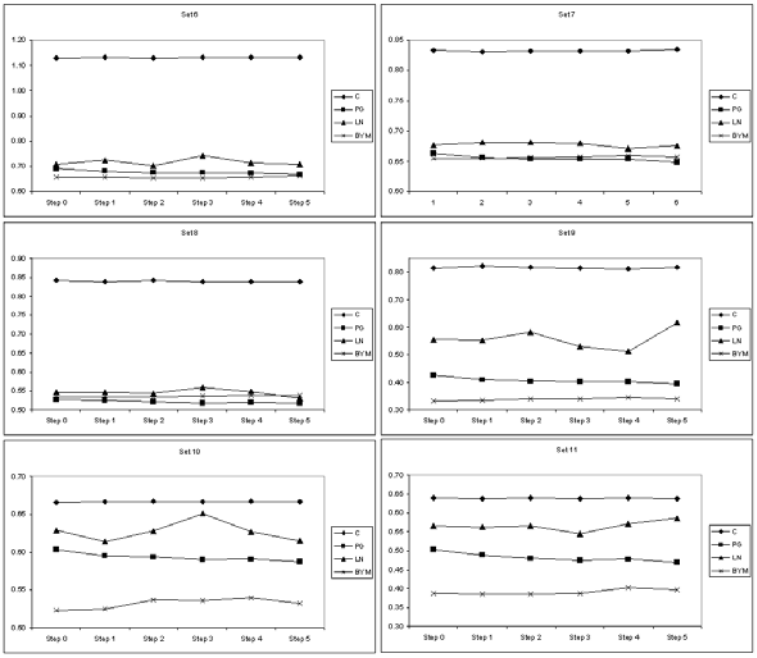
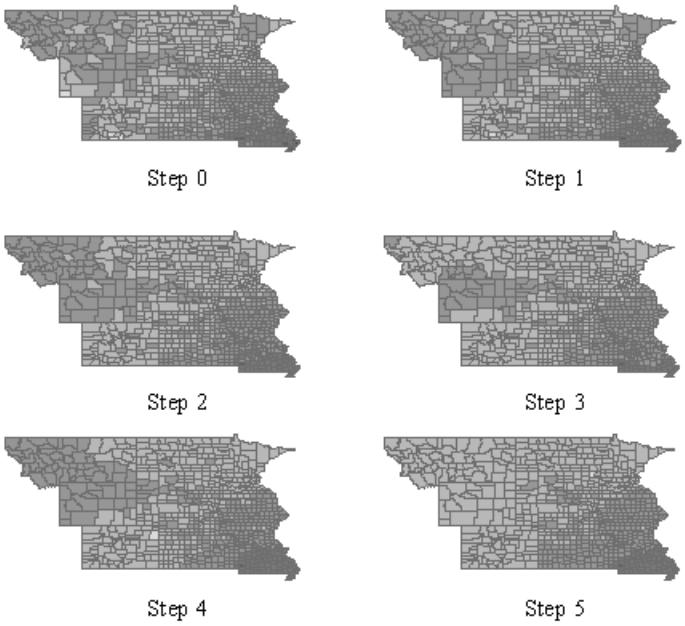



Figura 3: Distribución geográfica de la mortalidad por cáncer de pulmón. Modelo BYM, pasos 0 a 5.



Los métodos de suavización espacial, en particular el modelo BYM, utilizan datos de distintas áreas para estimar el valor en una región. Cerca de la frontera los valores del riesgo relativo son estimados utilizando información de los condados fronterizos. Entonces, si el número de casos observados es desconocido, las estimaciones estarán sesgadas. Si los condados no están cerca de la frontera los valores del riesgo relativo apenas sufren cambios.
La tabla 1 presenta las diferencias entre los riesgos relativos para el modelo BYM.

Tabla 1. Media de las diferencias entre los riesgos relativos. Modelo BYM

| | Diferencias | | | | |
|---|---|---|---|---|---|
| **Posición** | **Step (0-1)** | **Step (0-2)** | **Step (0-3)** | **Step (0-4)** | **Step (0-5)** |
| **1** | 0.1046 | 0.1091 | 0.1238 | 0.1132 | 0.1157 |
| **2** | 0.0146 | 0.0755 | 0.0912 | 0.0841 | 0.0910 |
| **3** | 0.0024 | 0.0102 | 0.0702 | 0.0687 | 0.0727 |
| **4** | 0.0015 | 0.0050 | 0.0098 | 0.0612 | 0.0586 |
| **5** | 0.0014 | 0.0040 | 0.0044 | 0.0075 | 0.0674 |
| **6** | 0.0017 | 0.0040 | 0.0043 | 0.0045 | 0.0108 |
| **7** | 0.0011 | 0.0038 | 0.0037 | 0.0041 | 0.0059 |
| **8** | 0.0012 | 0.0034 | 0.0032 | 0.0037 | 0.0033 |
| **9** | 0.0015 | 0.0043 | 0.0037 | 0.0036 | 0.0029 |
| **10** | 0.0012 | 0.0038 | 0.0033 | 0.0039 | 0.0023 |
| **11** | 0.0012 | 0.0046 | 0.0038 | 0.0043 | 0.0024 |

Cuando no se considera información de las áreas vecinas (por ejemplo en el modelo clásico), las diferencias en las estimaciones del riesgo en las áreas internas son muy pequeñas. Los modelos Poisson-gamma y lognormal son los que presentan diferencias más elevadas en las regiones internas. Como ya ocurría en el modelo BYM las diferencias en las estimaciones del riesgo decrecen a medida que nos adentramos en la región.

## 5. REFERENCIAS


Besag J., York J. and Mollié A. (1991) A Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 43, 1-59.

Clayton D. and Kaldor J. (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* 43, 671-681.

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

Griffith D.A. (1983) The boundary value problem in spatial statistical analysis. *Journal of Regional Science* 23, 377-387.

Lawson A.B. (2001) *Statistical Methods in Spatial Epidemiology*. Wiley, Chichester.

Mollié A. (1996) Bayesian Mapping of Disease. In Gilks W.R., Richardson S. and Spiegelhalter D.J. (eds). *Markov Chain Monte Carlo in Practice*. Chapman&Hall, London.

Tanner M.A. (1996) *Tools for Statistical Inference*. Springer-Verlag, New York.