UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2018, Vol. 11, No. 1, pp. 9–35

Un modelo jerárquico bayesiano espacio-temporal con variable de conteos: aplicación de VIH/SIDA en Costa Rica

A bayesian hierarchical spatio-temporal model with count data: application to the HIV/AIDS in Costa Rica

Shu Wei Chou-Chen[a]
shuwei.chou@ucr.ac.cr

Ricardo Alvarado-Barrantes[b]
ricardo.alvarado@ucr.ac.cr

Resumen

Los modelos espaciales que suavizan las tasas de mortalidad estandarizada o los riesgos relativos son utilizados ampliamente en el mapeo de enfermedades, lo anterior, con el objetivo de explorar y describir patrones espaciales de un evento de interés; generalmente, la estimación de estos riesgos relativos es imprecisa cuando los eventos son raros. Ahora bien, al momento de incluir la tendencia temporal, el problema es aún más grave, pues el conteo de las defunciones en el período dado se divide en varios años, lo que resulta en que los conteos sean más bajos. En este trabajo, se analizan los modelos bayesianos espacio-temporales que toman en cuenta la información geográfica y temporal, además de algunas covariables como el porcentaje de viviendas urbanas, porcentaje de personas entre 24 y 49 años y la tasa de mortalidad infantil de cada cantón en el 2011; se concluyó que estos modelos producen mejores estimaciones de riesgos relativos por cantón y año, además de que el modelo que asume una interacción espacio-temporal más simple ajusta mejor. Finalmente, se comparan los riesgos relativos estimados con el modelo seleccionado, contra la estimación obtenida vía máxima verosimlitud, y resulta que el método propuesto es más eficiente y preciso.

***Palabras clave*:** Epidemiología, mapeo de enfermedades, modelos espacio-temporales, modelos jerárquicos bayesianos, SIDA, VIH.

Abstract

Spatial models that smooth standardized mortality ratios are widely used in disease mapping. Usually, estimation is imprecise when events are rare. In situations

[a]Escuela de Estadística y Centro Centroamericano de Población, Universidad de Costa Rica
[b]Escuela de Estadística, Universidad de Costa Rica

where each areal count splits into different time periods, this problem is more evident because of the presence of even lower counts for the areal units for each time period. In this work, we analyze models that include geographic and temporal information and some covariates such as percentage of urban household, percentage of people between 24 and 49 years old, and infant mortality ratio of each county in 2011. As a result, these models produce better estimations, especially for the model with the simplest space-time interaction. Finally, HIV/AIDS mortality data in Costa Rica (1998-2012) are used as an illustration to compare classic standardized mortality ratios and posterior means of relative risk. The proposed method is more efficient and more precise than the maximum likelihood.

***Keywords*:** AIDS, Bayesian hierarchical models, disease mapping, epidemiology, HIV, space-time model.

1. Introducción

La interacción social de los individuos es un factor importante en la propagación de enfermedades infecciosas. Uno de los intereses de la investigación en el campo de la epidemiología es el mapeo de enfermedades (*Disease Mapping*), el cual surgió hace pocas décadas y se ha popularizado rápidamente; esta rama se interesa en conocer la distribución geográfica de una enfermedad en la población de cierta región geográfica. Generalmente, la dirección exacta de los pacientes es confidencial o desconocida, por lo tanto, el estudio por zonas geográficas, como los cantones o distritos, es más utilizado en el campo (Lawson & Williams 2001).

La distribución o comportamiento de una enfermedad en la región de estudio es influenciada por la estructura social y económica de la región, así como las costumbres y accesibilidad de servicios de salud. La estadística espacial permite modelar dicha estructura. La información descrita generalmente está disponible a nivel de áreas geográficas, por ejemplo, la prevalencia de una enfermedad o la mortalidad por cantones. El hecho de que una región tenga un alto riesgo de presentar una enfermedad puede estar relacionado a factores sociodemográficos a una acumulación de individuos con dicha enfermedad en el transcurso del tiempo. En muchas ocasiones, mediante la agregación de datos de todo el período de investigación, se pueden producir resultados engañosos, además, la enfermedad puede tener algún patrón a lo largo del tiempo, pero si se tienen enfermedades raras, el conteo por zonas es escaso, lo cual puede dificultar el evidenciar estos patrones. De esta manera, al tener conteos de datos escasos por zonas geográficas y, más aún, desagregados por el tiempo, la estimación del riesgo del evento por zona y tiempo es imprecisa debido a que hay zonas con conteos muy bajos o iguales a cero. La estadística bayesiana es una solución que permite suavizar los riesgos usando el patrón espacial y temporal, pues se utilizan estos patrones como información *a priori.* (Langford 1994).

El virus de inmunodeficiencia humana (VIH) es el virus causante del síndrome de inmunodeficiencia adquirida (SIDA), el cual altera el sistema inmunológico y destruye la capacidad del cuerpo para defenderse de otras infecciones o cánceres

(American Cancer Society 2014). Los principales mecanismos de transmisión del VIH son vía sexual, parenteral y vertical (perinatal). Los primeros casos fueron encontrados a principios de la década de 1980 y dicha enfermedad se expandió a nivel epidémico.

El estudio del VIH/SIDA es de gran importancia ya que esta es una de las principales causas de muerte entre los adolescentes de diez a diecinueve años a nivel mundial. De acuerdo con la (Organización Mundial de la Salud 2014), el VIH/SIDA es la segunda causa de muerte más importante entre los adolescentes después de los accidentes de tránsito. En Costa Rica, el VIH/SIDA fue una de las tres principales causas de muerte entre personas de todas edades en el 2002, junto con la cardiopatía isquémica y enfermedades cerebrovasculares (Altman 2011).

En la literatura sobre el VIH/SIDA existen criterios acerca de la distribución geográfica de la enfermedad (Poundstone et al. 2004, Zanakis et al. 2007, Hunter et al. 2008). De acuerdo con la clasificación de (Poundstone et al. 2004), los factores de la prevalencia del VIH/SIDA se clasifican en factores individuales, sociales y estructurales; los factores individuales son los riesgos biológicos, demográficos y de comportamiento de las personas (género, etnia, edad, ingreso, educación, costumbres, etc.), los factores sociales son producto de la interacción de los individuos (aspectos culturales, factores socioeconómicos y de residencia), por último, los factores estructurales son los relacionados con las leyes y políticas.

La (Agencia de los Estados Unidos para el Desarrollo Internacional 2011) promueve el uso de las herramientas espaciales descriptivas para analizar datos de VIH/SIDA. Específicamente en Costa Rica, el (Ministerio de Salud 2004) analiza la morbilidad y la mortalidad por año, grupo quinquenal de edad y por cantón separadamente. En los análisis muestran evidencias del patrón espacial y temporal; sin embargo, en dicho análisis no toman en cuenta las covariables que afectan la distribución del VIH/SIDA, por lo tanto, se requieren herramientas para unir los diferentes componentes, pues al agregar otros componentes para analizar año, cantón y grupo quinquenal de edad por separado pueden haber patrones ocultos.

Por otro lado, los análisis espaciales cantonales de VIH en Costa Rica se han realizado solamente utilizando los métodos descriptivos, por ejemplo, la (Organización Panamericana de la Salud 2004) muestra evidencias espaciales y temporales analizando descriptivamente los dos componentes por separado. Por consiguiente, se llevan a cabo análisis exploratorios sobre el patrón espacial y temporal de los datos de VIH/SIDA.

En la presente investigación se utilizan datos a nivel de cantón debido a que los resultados obtenidos serían más fáciles y prácticos de implementar a este nivel, sumado a esto, con una división geográfica más pequeña como los distritos, se obtendrían más ceros por unidad, lo cual reduciría la utilidad del estudio.

El objetivo de este trabajo es analizar los modelos bayesianos espacio-temporales que toman en cuenta covariables regionales para modelar datos de defunciones de VIH/SIDA en Costa Rica por cantón, para el periodo 1998-2012.
En la segunda sección se exponen los fundamentos teóricos y las herramientas

metodológicas para cumplir con el objetivo; en la tercera sección se presentan los resultados obtenidos del análisis; finalmente, se presentan la discusión y conclusiones de la investigación realizada.

2. Métodos y datos

2.1. Datos analizados y fuente de información

Los datos de defunciones por VIH/SIDA corresponden al periodo 1998-2012 a nivel cantonal. Se utilizan como covariables el porcentaje de viviendas urbanas, porcentaje de personas entre 24 y 49 años y la tasa de mortalidad infantil de cada cantón, las cuales son tomadas de las bases de datos del censo 2011. Estos datos, junto con las proyecciones poblacionales por cantón y año, fueron obtenidas de las bases de datos en línea del (Centro Centroamericano de Población 2014) de la Universidad de Costa Rica.

Una de las limitaciones del registro de defunciones es el lugar de ocurrencia del deceso, pues aunque sucediese en un hospital no se registra el cantón del hospital como lugar de ocurrencia, sino la dirección de residencia de la persona. Con lo cual, el sistema solo refleja los datos de vivienda de la persona. Sin embargo, en caso del VIH/SIDA, al ser una enfermedad crónica, existe la posibilidad de que los pacientes con diagnóstico positivo se trasladen a vivir cerca de hospitales y clínicas para acceder con facilidad a los servicios médicos debido a la cercanía geográfica.

Esta investigación se centra en el análisis de datos de área, pues los conteos del evento raro están agregados en unidades geográficas y comprenden las defunciones a causa de VIH/SIDA por cantón.

2.2. Incorporación de la información geográfica

Cuando se desea incluir la localización de las unidades estadísticas en el estudio, es necesario cuantificar de cierta manera la ubicación de ellas. Se espera que la correlación de la variable de interés entre dos unidades geográficas cercanas sea alta, mientras que la correlación entre unidades geográficas lejanas sea baja.

Para datos puntuales (distribución espacial de la ubicación de los eventos como robo de domicilios de un país) y datos continuos (datos que pueden ser medidos en un espacio continuo como temperatura), la distancia es una medida natural para cuantificar la ubicación geográfica. Se define una *matriz de proximidad* W $(n \times n)$, donde cada entrada w_{ij} representa la proximidad entre los pares ordenados i y j, donde $i = 1, ..., n$; $j = 1, ..., n$ y n es la cantidad de unidades geográficas. Una medida natural de proximidad es el inverso de la distancia entre dos localidades i y j.

En el caso de datos de área, por la naturaleza de los mismos, es necesario definir el concepto de proximidad para la matriz W, donde cada entrada w_{ij} representa

la proximidad entre las áreas A_i y A_j. (Bailey & Gatrell 1995, p. 261) proponen las siguientes definiciones de proximidad entre áreas.

- *K centroides más cercanos:*

$$w_{ij} = \begin{cases} 1 & \text{si el centroide de } A_j \text{ es uno de los k centroides} \\ & \text{más cercanos a } A_i \\ 0 & \text{otros casos} \end{cases} \quad (1)$$

- *Distancia fija:*

$$w_{ij} = \begin{cases} d_{ij}^{\gamma} & \text{si } d_{ij} < \delta, \text{ siendo } d_{ij} \text{ la distancia entre los centroides} \\ & \text{de } A_i \text{ y } A_j, \text{ con } d_{ij} > 0, \delta > 0 \text{ y } \gamma < 0 \\ 0 & \text{otros casos} \end{cases} \quad (2)$$

- *Vecino que comparte la frontera:*

$$w_{ij} = \begin{cases} 1 & \text{si } A_j \text{ comparte frontera con } A_i \\ 0 & \text{otros casos} \end{cases} \quad (3)$$

Estas definiciones de proximidad son usadas para analizar la dependencia espacial en la sección 3.1, y, específicamente, la definición de proximidad (3) es usada para definir los modelos espaciales descritos en la sección 2.6.

Para explorar la dependencia espacial, se usó el índice I de Moran, el cual se estima como:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_{i=1}^{n}(y_i - \bar{y})^2)(\sum \sum_{i \neq j} w_{ij})},$$

Donde y_i es la variable del interés, w_{ij} es el elemento ij de la matriz de proximidad W y n es la cantidad de áreas.

Este índice está relacionado con el covariograma, el cual es un instrumento que sirve para la exploración inicial gráfica de la dependencia espacial. El covariograma es la función de covarianza entre dos puntos en el espacio del estudio (Bailey & Gatrell 1995); si dos puntos están cercanos, se espera que el covariograma sea mayor que el de dos puntos lejanos. Este índice tiene problemas de interpretación, ya que no varían entre -1 y 1 como sucede con el coeficiente de correlación de Pearson (Bailey & Gatrell 1995, p. 270). Si la autocorrelación no existe, el índice I sigue una distribución normal cuando $n \to \infty$ con:

$$E(I) = -\frac{1}{(n-1)},$$

$$\text{VAR}(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2},$$

Donde $S_0 = \sum\sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2}\sum\sum_{i \neq j}(w_{ij} + w_{ij})^2$, $S_2 = \sum_k(\sum_j w_{kj} + \sum_i w_{ik})^2$,y n es la cantidad de observaciones.

De acuerdo con (Ripley 2004), la aproximación normal para este índice es apropiada para $n > 10$. (Banerjee et al. 2014) recomiendan el uso del I de Moran para un análisis exploratorio espacial. Debido a que esta medida requiere del supuesto de normalidad y, dado que la variable respuesta es Poisson y el conteo por zonas geográficas es muy escaso, resulta inadecuado hacer un análisis espacial para un año específico. Sin embargo, al agregar los datos del período 1998-2012 se muestra posteriormente que al hacer una transformación logarítmica, se logra cumplir el supuesto de normalidad y el patrón espacial resulta ser evidente. Esta medida será calculada posteriormente en la sección 3.

2.3. Riesgos relativos

Para contemplar el tamaño de la población se usan los riesgos relativos. Sea Y_{it} el conteo del evento para el área i y el tiempo t ($i = 1, ..., n$ y $t = 1, ..., T$), se supone que Y_{it} se distribuye independientemente como

$$Y_{it} \sim \text{Poisson}(\lambda_{it}),$$

$$\lambda_{it} = N_{it}p_{it}$$

Donde N_{it} es la población en riesgo, y p_{it} es la probabilidad de que ocurra el evento (ambos en el área i y el tiempo t). También se puede reescribir el parámetro de la distribución de Poisson como el producto de E_{it} (cantidad esperada si en todas las áreas y tiempos hay una probabilidad homogénea de ocurrencia del evento) por r_{it} (riesgo relativo), donde $E_{it} = N_{it}p^*$, $r_{it} = p_{it}/p^*$ y p^* es la probabilidad global de que ocurra el evento. Entonces

$$\lambda_{it} = N_{it} \cdot p_{it} = N_{it} \cdot p^* \left(\frac{p_{it}}{p^*}\right) = E_{it} \cdot r_{it},$$

El parámetro de interés en la estimación en los modelos de este artículo es r_{it}. Aunque el parámetro que se puede interpretar con facilidad es λ_{it}, este depende principalmente de la población en riesgo N_{it}, por lo que no se puede comparar el verdadero riesgo entre las áreas geográficas ya que está condicionado al tamaño de la población del cantón i en el año t. La ventaja de trabajar con r_{it} es que proporciona una visualización del fenómeno sin que sea afectada por la población del área i. Si $r_{it} > 1$, entonces la ocurrencia del evento es mayor que el conteo esperado E_{it}; mientras que si $r_{it} < 1$, la ocurrencia de los casos es menor que el conteo esperado (Richardson et al. 2004, Waller et al. 1997).

Cuando se trata de eventos raros en el área i y tiempo t, el estimador clásico de máxima verosimilitud, $\hat{r}_{it} = Y_{it}/E_{it}$ no posee propiedades asintóticas apropiadas

que garanticen estimaciones confiables y estables (Bailey & Gatrell 1995); este estimador recibe el nombre de razón de mortalidad estandarizada. Así, en la literatura se recomienda el suavizamiento de las estimaciones por medio del uso de la información que da el patrón espacial y temporal de los riesgos relativos para eliminar los ruidos que tienen los datos, tal método incorpora la información *a priori* y los datos recolectados para la estimación *a posteriori* de los riesgos relativos, además, modela la sobredispersión y la dependencia de los conteos entre áreas cercanas (Richardson et al. 2004, Bailey & Gatrell 1995).

2.4. Mapeo de probabilidades

(Richardson et al. 2004) y (Bailey & Gatrell 1995) comentan sobre la deficiencia al utilizar frecuencias absolutas de cada área en los mapas debido a que causa problemas de interpretación, pues la población de referencia de cada área es distinta.

Una forma de visualizar el patrón espacial es construir el *mapa de probabilidades* (Bailey & Gatrell 1995). El gráfico muestra las zonas geográficas con valores extremos, los cuales implican variabilidad grande a través de la región. Se utiliza la siguiente fórmula:

$$p_i = \begin{cases} \sum\limits_{x \geq y_i} \frac{\hat{E}_i^x e^{-\hat{E}_i}}{x!} & \text{si } y_i \geq \hat{E}_i \\ \sum\limits_{x \leq y_i} \frac{\hat{E}_i^x e^{-\hat{E}_i}}{x!} & \text{si } y_i < \hat{E}_i \end{cases} \tag{4}$$

Donde $\hat{E}_i$ es el conteo esperado en el área i y y_i es el conteo observado en el área i. Valores pequeños de p_i implican que la tasa del área es muy alta o muy baja. Este instrumento tiene como propósito de ayudar a localizar áreas con riesgos extremadamente altos o bajos. (Bailey & Gatrell 1995) recomiendan establecer el valor crítico para p_i en $0,05$ para considerar áreas con tasa alta o baja. Esta metodología es utilizada como análisis exploratório en la sección 3.

2.5. Estimación bayesiana

La estimación clásica de los riesgos relativos no garantiza estimaciones confiables y estables, por lo que (Langford 1994) recomienda el uso de la estimación bayesiana cuando se tienen eventos raros. El autor utiliza el ejemplo de la leucemia infantil para mostrar los problemas que presentan los riesgos relativos cuando se calcula con eventos raros y poblaciones pequeñas; en el estudio, el objetivo era detectar la posibilidad de agrupamiento de casos alrededor de la planta de reprocesamiento de material nuclear Sellafield. (Langford 1994) comenta dos problemas que tiene la estimación de los riesgos relativos cuando se tienen eventos raros. En primer lugar, cuando el valor esperado de un evento raro en el área i es pequeño, el riesgo relativo del área i cambia drásticamente debido a la variación que puede tener el valor observado en el área i. De esta forma, si se tiene un valor esperado de

2000, el riesgo relativo no va a resultar muy diferente cuando el observado varía poco; por ejemplo, dos valores observados de 2002 y 2000 producen los riesgos relativos $1,001(2002/2000)$ y $1(2000/2000)$, respectivamente. Sumado a esto, si el valor esperado del área i es 1, un valor observado de 2 en el área i produce un riesgo relativo de 2, mientras que si el valor observado es 0, el riesgo relativo sería nulo.

En segundo lugar, el valor-p de la prueba cuya hipótesis nula es $r_i = 1$ presenta problemas en áreas con poblaciones grandes, como puede suceder cuando la concentración de la población en las zonas urbanas es alta; en tal situación, el error estándar del riesgo relativo es muy pequeño y por ende produce valores de p extremadamente pequeños. Este comportamiento lleva a rechazar con alta probabilidad la hipótesis de que los riesgos relativos sean iguales a 1.

El autor considera que el uso de la estimación bayesiana es una buena recomendación para resolver los dos problemas mencionados anteriormente. Con el uso de la estimación bayesiana, el mapeo de los riesgos relativos es ajustado por los criterios estadísticos y, por consiguiente, elimina los ruidos que tienen los datos al suavizar el patrón espacial de los riesgos relativos en el espacio y el tiempo.

Debido a la posible complejidad de las integrales de altas dimensiones en la obtención de las distribuciones *a posteriori* de los parámetros de interés, se utilizan los métodos Monte Carlo vía cadenas de Markov. La convergencia de las estimaciones es evaluada con el diagnóstico de secuencias múltiples de Gelman y Rubin (1992), por (Brooks & Gelman 1998). Se utilizan tres cadenas con diferentes valores iniciales y se calcula el factor de reducción de escala potencial multivariado (MPSRF) para evaluar la convergencia de dichas 3 cadenas.

2.6. Modelos analizados

Si el conteo es pequeño, se puede modelar con la distribución de Poisson o binomial por medio de la función de enlace de logaritmo o logit, respectivamente. La distribución binomial se utiliza cuando la población de riesgo de cada cantón es pequeña. Esta investigación se centra en el caso de Poisson y usa la aproximación a la binomial debido a que el interés es modelar conteo de eventos raros, donde la población de riesgo de cada unidad geográfica es grande.

La modelización de datos de conteo de dimensión pequeña utiliza generalmente modelos bayesianos para suavizar los riesgos estimados en cada área. Debido a que los eventos son escasos en cada área, los riesgos estimados son generalmente imprecisos. (Richardson et al. 2004) muestran que los modelos bayesianos de mapeo de enfermedades son conservadores con las distribuciones *a priori* especificadas y el grado de suavización de los riesgos; por tanto, estas herramientas logran un balance entre la información espacial y temporal con los datos recolectados.

(Besag et al. 1991) introducen el *modelo autoregresivo condicional* (CAR, por sus siglas en inglés) para modelizar espacialmente variables de conteo. La mayor ventaja del CAR es que utiliza el estimador de contracción para suavizar los riesgos

relativos estimados vía máxima verosimilitud que se calculan para cada área por separado. Este estimador de contracción resuelve el problema de la estimación de los riesgos relativos cuando el evento es raro y el tamaño de la población de cada área es diferente.

Sea y_i el conteo del evento en el área i $(i = 1, ..., n)$, el cual se distribuye como Poisson$(E_i r_i)$, entonces se modeliza el riesgo relativo r_i mediante:

$$ln(r_i) = \boldsymbol{x_i\beta} + s_i + u_i,$$

Donde $\boldsymbol{x\beta}$ es el componente de las covariables, s_i es el componente espacial estructurado y u_i es el componente no estructurado o aleatorio.

La estructura para modelar el componente espacial es

$$s_i|s_j \sim \mathcal{N}\left(\frac{\sum_j w_{ij}s_j}{\sum_j w_{ij}}, \frac{\sigma_s^2}{\sum_j w_{ij}}\right).$$

De esta forma, s_i condicionado a los demás s_j sigue una distribución normal con media igual al promedio ponderado de los s_j y varianza inversamente proporcional a la suma de los pesos. Más específicamente, si la matriz W es definida como $w_{ij} = 1$, si el área j es vecino de i y 0 otros casos, entonces

$$s_i|s_j \sim N\left(\frac{\sum_{j\in\delta_i} s_j}{|\delta_i|}, \frac{\sigma_s^2}{|\delta_i|}\right), \tag{5}$$

Donde δ_i es el conjunto de las áreas j que son vecinos del área i. Se puede notar que la varianza es inversamente proporcional a la cantidad de vecinos del área i. Si el área tiene relativamente muchos vecinos, entonces la varianza es pequeña.

Con la especificación de (5), se tiene la distribución *a priori* conjunta de los s_i como sigue

$$f(s_1, ..., s_n) \propto \exp\left\{-\frac{1}{2\sigma_s^2}\boldsymbol{s}'(D_w - \boldsymbol{W})\boldsymbol{s}\right\},$$

Donde D_w es la matriz diagonal con $\{D_w\}_{ii} = \sum_j w_{ij}s_j$. Note que la matriz de varianzas y covarianzas, $\boldsymbol{\Sigma}_s$, no existe, pues $\boldsymbol{\Sigma}_s^{-1} = (D_w - \boldsymbol{W})$ es una matriz singular; como consecuencia, la distribución es impropia.

Para facilitar la notación, si $\boldsymbol{s} = (s_1, ..., s_n)'$ es el componente espacial, se denota $\boldsymbol{s} \sim \text{CAR}(\boldsymbol{W}, \sigma_s^2)$ para indicar que $\boldsymbol{s}$ se modeliza como el CAR. Se recomienda utilizar la distribución gamma inversa para la distribución *a priori* del hiperparámetro σ_s^2 (Richardson et al. 2004, Abellan et al. 2008, Fortunato et al. 2011).

En este trabajo, la estructura espacial es especificada con W, (definición 3). Para la estructura temporal, se puede considerar el tiempo como una recta y, por tanto, los vecinos del tiempo t como los tiempos $t-1$ y $t+1$; de esta manera, se puede especificar una matriz de proximidad para el componente temporal, es decir, que Q es la matriz definida como $q_{ij} = 1$ si el tiempo j es adyacente al tiempo i, y

$q_{ij} = 0$ si no lo es. (Richardson et al. 2006) y (Abellan et al. 2008) utilizan esta especificación para modelizar el componente temporal.

Antes de continuar con la descripción de los cuatro modelos analizados, se definen las siguientes notaciones:

r_{it} es el riesgo relativo del área i y tiempo t,

μ es el efecto global de la región,

$\boldsymbol{x_i}$ es el vector de los valores de las covariables del cantón i, y

$\boldsymbol{\beta}$ son los coeficientes de las covariables.

A su vez, se definen las distribuciones *a priori*: μ, $\beta_i \sim \mathcal{N}(0; 10^4)$.

Modelo 1

En (Richardson et al. 2006) se propone un modelo multivariado para modelizar la mortalidad por cáncer de pulmón en hombres y mujeres. Para esta investigación se usa el mismo modelo pero se restringe el caso univariado al agregar el componente de las covariables. Este modelo asume que los errores no dependen de la estructura espacial y temporal, está dado por

$$ln(r_{it}) = \mu + x_i\beta + \theta_i + \alpha_t + v_{it}. \tag{6}$$

Donde:

θ_i y α_t son los parámetros relacionados al efecto espacial y temporal, respectivamente.

v_{it} son efectos aleatorios independientes entre sí.

Por otro lado, se definen las distribuciones *a priori*:

$\boldsymbol{\theta} \sim \text{CAR}(\boldsymbol{W}, 1/\tau_\theta)$.

$\boldsymbol{\alpha} \sim \text{CAR}(\boldsymbol{Q}, 1/\tau_\alpha)$.

$v_{it} \sim \mathcal{N}(0, 1/\tau_v)$.

Los hiperparámetros $\tau_\theta, \tau_\alpha, \tau_v \sim \text{Gamma}(0,5; 0,0005)$ son tomados como recomendación de (Richardson et al. 2006) con varianza grande.

Modelo 2

(Waller et al. 1997) proponen el modelo con interacción, donde el componente espacial está anidado dentro de cada período. El modelo se escribe como

$$ln(r_{it}) = \mu + \boldsymbol{x_i\beta} + \alpha_t + \phi_{it} + v_{it}, \tag{7}$$

Donde:

α_t es el efecto temporal.

ϕ_{it} es el efecto espacial anidado en cada tiempo t ($t = 1, ..., 15$).

v_{it} es el efecto aleatorio del área i anidado en el tiempo t ($t = 1, ..., 15$).

Por otro lado, se definen las distribuciones *a priori*:

$\boldsymbol{\phi_t} \sim \text{CAR}(\boldsymbol{W}, 1/\tau_\phi^t)$ para $t = 1, ..., 15$, es decir que es un modelo CAR condicional al tiempo t,

$v_{it} \overset{iid}{\sim} N(0, 1/\tau_t)$, y

los hiperparámetros $\tau_\theta^t, \tau_t \sim \text{Gamma}(0,5; 0,0005)$.

Modelo 3

Este modelo asume una interacción compleja entre el espacio y tiempo, es utilizado por (Lagazio et al. 2001), (Lagazio et al. 2003), (Schmid & Held 2004). En esta investigación, se extiende el modelo al agregar linealmente el componente de las covariables como sigue

$$ln(r_{it}) = \mu + \boldsymbol{x_i}\boldsymbol{\beta} + \theta_i + \alpha_t + v_{it}, \tag{8}$$

Donde:

θ_i y α_t son los parámetros relacionados al efecto espacial y temporal, respectivamente.

v_{it} es la interacción entre el tiempo y el espacio.

Por otro lado, se definen las distribuciones *a priori*:

$\boldsymbol{\theta} \sim \text{CAR}(\boldsymbol{W}, 1/\tau_\theta)$.

$\boldsymbol{\alpha} \sim \text{CAR}(\boldsymbol{Q}, 1/\tau_\alpha)$.

Los hiperparámetros $\tau_\theta, \tau_\alpha \sim \text{Gamma}(0,5; 0,0005)$.

La distribución de v_{it} dado las otras zonas es multinormal con media

$$\mu_{it} = \begin{cases} v_{i,t+1} + \frac{1}{m_i}\sum_{j\sim i} v_{jt} - \frac{1}{m_i}\sum_{j\sim i} v_{j,t+1} & t = 1 \ , \\ v_{i,t-1} + \frac{1}{m_i}\sum_{j\sim i} v_{jt} - \frac{1}{m_i}\sum_{j\sim i} v_{j,t-1} & t = T \ , \\ \frac{1}{2}(v_{i,t-1} + v_{i,t+1}) + \frac{1}{m_i}\sum_{j\sim i} v_{jt} - \frac{1}{2m_i}\sum_{j\sim i}(v_{j,t-1} + v_{j,t+1}) & t = 2, ..., T-1 \ , \end{cases}$$

La matriz de precisión es

$$\tau_{it} = \begin{cases} m_i\kappa_v & t = 1 \text{ ó } t = T \ , \\ 2m_i\kappa_v & t = 2, ..., T-1 \ . \end{cases}$$

Se puede notar que la especificación de las distribuciones *a priori* de este tipo implica que la distribución de v_{it} depende de:

1. $v_{i,t-1}$ y/o $v_{i,t+1}$, el efecto temporal de primer orden.
2. v_{jt} con $j \sim i$, el efecto espacial de los vecinos.
3. $v_{j,t-1}$ y/o $v_{j,t+1}$ con $j \sim i$, el efecto temporal de los vecinos.

Modelo 4

Por último, para evaluar la inclusión del patrón espacio-temporal de los datos reales de VIH/SIDA, se comparan los modelos anteriores con un modelo de regresión lineal para el logaritmo de los riesgos relativos, es decir, un modelo que supone la independencia espacial y temporal de los cantones de la forma:

$$ln(r_{it}) = \mu + \boldsymbol{x_i}\boldsymbol{\beta} + v_{it}, \tag{9}$$

Donde v_{it} son errores aleatorios independientes entre sí definidos como el Modelo 1.

2.7. Criterio de selección de modelos:

En el contexto bayesiano, el criterio de información de deviancia (DIC, por sus siglas en inglés) propuesto por (Spiegelhalter et al. 2002) es el más utilizado para la comparación de modelos debido a su facilidad computacional producto del MCMC y por su utilidad (Banerjee et al. 2014); no obstante, (Plummer 2008) señala los problemas que tiene el DIC en el contexto de mapeo de enfermedades. Debido a que la cantidad de parámetros en los modelos jerárquicos espacio-temporales es grande, comparado con el tamaño de observaciones, no se cumplen las propiedades asintóticas de estos criterios. Por lo tanto, como recomienda (Cai et al. 2012), se considera el método de comparación de modelos propuesto por (Gelfand et al. 1992) basado en la Ordenada Predictiva Condicional (CPO por sus siglas del inglés, *conditional predictive ordinate*).

El CPO para el área i y tiempo t es definido como la densidad *a posteriori* de la validación cruzada marginal:

$$\begin{aligned} \mathrm{CPO_{it}} &= f(y_{it}|y_{(it)}) \\ &= \int f(y_{it}|\theta, x_{it}) f(\theta|y_{(it)}, x_{(it)})\, \mathrm{d}\theta \\ &= \left(\int \frac{1}{f(y_{it}|\theta, x_i)} f(\theta|y, x) \mathrm{d}\theta \right)^{-1} \end{aligned}$$

Donde $y_{(it)}$ es el vector de todas las observaciones excepto la observación it y θ es el vector de parámetros. Para el cálculo de CPO_{it} no se tiene una expresión cerrada, por lo que usualmente para calcularlo se recurre al MCMC para extraer una muestra de la distribución *a posteriori* de $f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x})$ como sigue:

$$\widehat{\mathrm{CPO}}_{it} = \left(\frac{1}{S} \sum_{s=1}^{S} \frac{1}{f(y_{it}|\theta^{(s)}, x_i)} \right)^{-1},$$

Donde S es la cantidad de iteraciones después del período de calentamiento (*burn-in*). Por último, para resumir la bondad de ajuste de cada modelo, se calcula la pseudo-verosimilitud de la validación cruzada, la cual se estima con

$$L_{CV} = \prod_{i=1}^{N} \prod_{t=1}^{T} \text{CPO}_{it}.$$

Debido a que L_{CV} es cercano a cero, se calcula el negativo del logaritmo natural de L_{CV}

$$LPMP_{CV} = -\sum_{i=1}^{N} \sum_{t=1}^{T} \log \text{CPO}_{it}.$$

Esta medida es llamada por *log pseudo-marginal likelihood* en la literatura. El modelo del mejor ajuste es aquel que tiene el menor $LPMP_{CV}$, pues esto correspondería al mayor L_{CV}.

2.8. Programas y técnicas a emplear

Los análisis fueron realizados utilizando los programas R versión 3.1.2 y OpenBUGS versión 3.2.3, ambos *software* libres. El programa OpenBUGS es muy conveniente para análisis bayesiano. El programa R es de gran utilidad y de uso global para todo tipo de análisis estadístico; además, R tiene una interfaz amigable que permite interactuar con OpenBUGS para facilitar otros análisis y la visualización de los patrones espaciales y temporales a partir de los resultados obtenidos con OpenBUGS. Los paquetes requeridos en R para los análisis son: sp (Pebesma 2005), ape (Paradis et al. 2004), maptools (Bivand & Lewin-Koh 2014), ggplot2 (Wickham 2009), lmtest (Zeileis & Hothorn 2002), forecast (Hyndman & Khandakar 2008), fBasics (Team et al. 2014) y BRugs (Thomas et al. 2006).

3. Resultados

3.1. Análisis espacial exploratorio

Para fines de una exploración de patrones espaciales, se siguen la recomendación de (Banerjee et al. 2014) de usar el índice I de Moran. Debido a que la variable de defunciones sigue una distribución asimétrica, se utliza una transformación con el logarítmo natural para corregir la asimetría de los datos y poder explorar el comportamiento espacial usando el I de Moran.

Bajo la hipótesis nula de que las defunciones de los 81 cantones son independientes e idénticamente distribuidas, la I es asintóticamente normal con valor esperado $E(I) = -1/(n-1) = -0,0125$. En la Tabla 1 se muestran los valores del índice I de Moran calculados con las tres definiciones de matriz de proximidad. Con la definición de k centroides más cercanos se utilizan $k = 1, 2, 3$ *y* 4. Todos los

índices muestran que se puede concluir una dependencia espacial (con un nivel de significancia de 0,05), excepto para el caso de $k = 1$ centroide más cercano donde no se rechaza la hipótesis nula de que las defunciones transformadas de los 81 cantones son espacialmente independientes. Sin embargo, cabe destacar que (Banerjee et al. 2014) recomiendan el uso del índice I de Moran como una medida exploratoria de patrones espaciales en vez de una "prueba de significancia".

Tabla 1: *El índice I de Moran usando diferentes definiciones de matriz de proximidad aplicado a los datos de VIH/SIDA en Costa Rica (1997-2012)*

Definición de la matriz de proximidad		I de Moran	valor-p
Distancia fija[1]		0,121	0,033
Vecino que comparte la frontera[2]		0,2985	<0,001
k centroides más cercanos	k=1	0,238	0,062
	k=2	0,317	0,001
	k=3	0,324	<0,001
	k=4	0,338	<0,001

[1] definición 2 con $\delta = Q_1 = 32528$ el primer cuartil y $\gamma = -1$.

[2] definición 3.

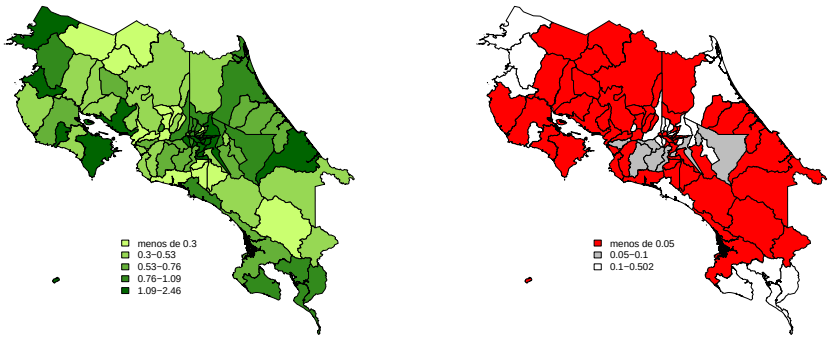
En la Figura 1a se pueden observar los riesgos relativos clásicos por cantón al acumular datos para todo el período de estudio. La estimación de este mapa es menos afectada por conteos bajos, pues al agregar los datos de todo el período se obtienen conteos no tan bajos. Los cantones con riesgos relativos altos se encuentran principalmente en la región central, provincia de Limón y la costa pacífica cerca de las fronteras con Nicaragua y Panamá, por tanto, los riesgos relativos son heterogéneos a nivel de cantón.

El mapa de probabilidad (Figura 1b) muestra los resultados de p_i obtenidos con la fórmula 4 para cada cantón. Los cantones con un color rojo más oscuro (51 de 81 cantones) son los que tienen riesgos extremadamente altos o bajos según el criterio de $p_i < 0,05$ recomendado por (Bailey & Gatrell 1995). Estos resultados muestran que los riesgos relativos cantonales no son uniformes en la región, es decir, existe evidencia de dependencia espacial.

3.2. Análisis espacio-temporal de las defunciones por VIH/SIDA (1997-2012)

3.2.1. Diagnósticos y comparación de los modelos

Las estimaciones del modelo 1 (ecuación 6) fueron obtenidas con facilidad computacional debido a la rápida convergencia en comparación con los modelos 2 (ecuación 7) y 3 (ecuación 8). El modelo 1 tiene relativamente pocos parámetros y es menos complejo comparado con los modelos 2 y 3. Los parámetros del modelo 1 son



(a) Mapa de riesgos relativos (b) Mapa de probabilidad

Figura 1: *Mapas de defunciones por VIH/SIDA por cantón (1998-2012)*

simulados utilizando un período de calentamiento de 10.000 y luego se seleccionan 10.000 muestras del total de las 100.000 iteraciones, tomando una muestra cada 10. Los parámetros del modelo 2 son simulados utilizando un periodo de calentamiento de 50.000 iteraciones y luego se seleccionan 10.000 muestras del total de 200.000 iteraciones, tomando una muestra cada 20. Las cadenas de MCMC del modelo 3 convergen lentamente. Las estimaciones fueron obtenidas seleccionando 10.000 muestras del total de 500.000 iteraciones, tomando una muestra cada 50, después de un periodo de calentamiento de 400.000. Finalmente, las cadenas del MCMC del modelo 4 fueron obtenidas seleccionando 10.000 muestras del total de 50.000 iteraciones, tomando una muestra de 5 después de un período de calentamiento de 50.000.

La razón principal para haber seleccionado estos parámetros de adelgazamiento es porque la complejidad de los modelos es distinta. El modelo 3, el más complejo, se estimó con un adelgazamiento de 50, al seguir el ejemplo del mismo modelo que aplica (Lagazio et al. 2001). No obstante, la selección de estos parámetros de adelgazamiento para los otros tres modelos se basa en que existe una alta autocorrelación en las cadenas y se considera que con estas especificaciones se reducen las autocorrelaciones y se obtienen estimaciones aceptables.

Se realiza el diagnóstico de la convergencia de los parámetros de los cuatro modelos. La media ergódica muestra que la media *a posteriori* de cada parámetro converge y la traza de los parámetros da una indicación de que las simulaciones han alcanzado una distribución estacionaria. Las autocorrelaciones de las cadenas no son altas, excepto en el modelo 3. La complejidad de los parámetros de este modelo hace que los parámetros del patrón temporal presenten autocorrelaciones altas, mientras que los demás parámetros convergen apropiadamente.

Al analizar los valores del factor de reducción de escala potencial multivariado ($MPSRF$) para los cuatro modelos en el cuadro 2, los resultados muestran convergencia aceptable. Cabe destacar que, para todos los modelos, al principio se

generaron las cadenas de los parámetros, de tal manera que muestren convergencia aceptable. Posteriormente, se realiza el diagnóstico de Gelman y Rubin con tres valores iniciales para cada parámetro de los modelos de acuerdo con la distribución *a posteriori* (el primer valor cerca del mínimo, el segundo en el centro y el último cerca del máximo de la distribución *a posteriori*). El valor de MPSRF debe converger a 1 si el uso de diferentes valores iniciales de los parámetros logra converger a la misma media *a posteriori*. Se utilizan tres cadenas para cada parámetro y se obtienen resultados aceptables para los 4 modelos ($MPSRF < 1,2$). El modelo 3 es el único modelo con un valor de MPSRF muy cercano a 1,2 debido a la convergencia lenta mencionada anteriormente.

Tabla 2: *Bondad de ajuste ($LPMP_{CV}$) y la evaluación de convergencia de MCMC basada en $MPRSF$ de los 4 modelos*

Modelo	$LPMP_{CV}$	MPSRF
1	1448,3	1,01
2	1534,1	1,02
3	1454,8	1,18
4	1512,5	1,00

El ajuste de los modelos se evalúa utilizando el valor negativo del logaritmo natural de la verosimilitud de la validación cruzada ($LPMP_{CV}$). Un valor de $LPMP_{CV}$ bajo corresponde a un mejor ajuste. En la Tabla 2 se observa que al no considerar el patrón espacio-temporal en el análisis (modelo 4), el $LPMP_{CV}$ es $1512,5$, mientras que con los modelos 1 y 3 se obtienen valores más bajos, lo cual hace necesaria la inclusión del patrón espacial y temporal.

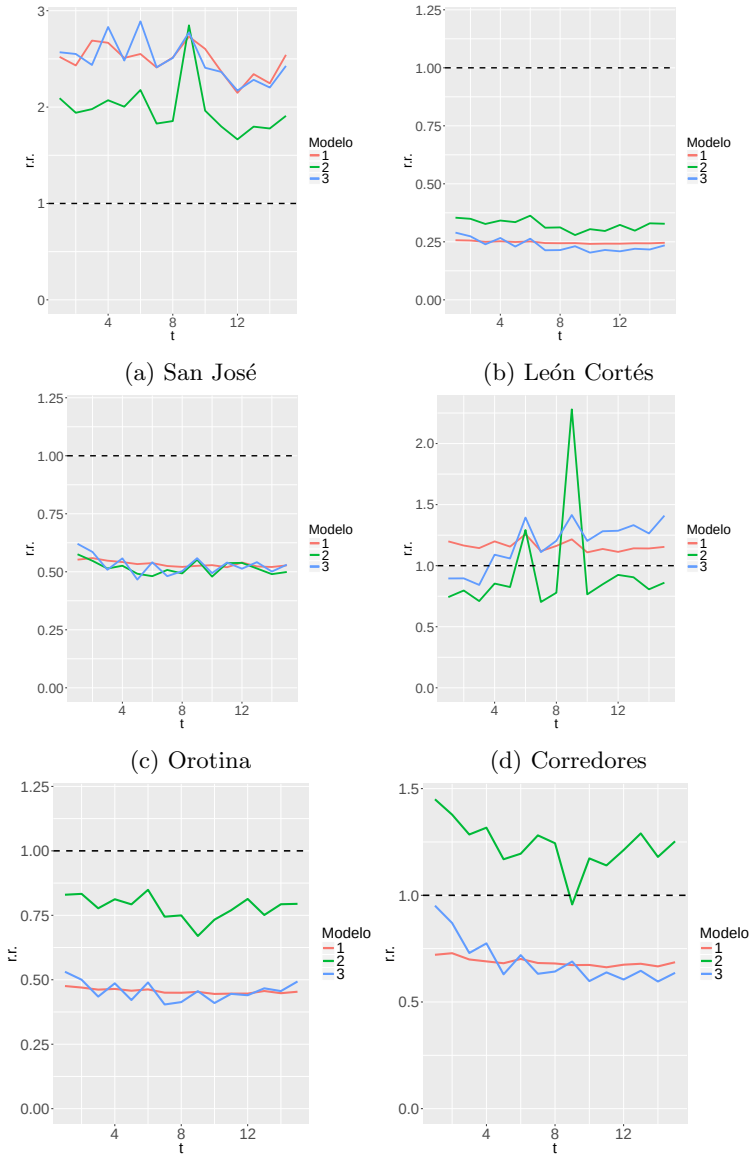
El valor del $LPMP_{CV}$ para el modelo 2 es aún más alto que para el modelo 4, el cual no considera el patrón espacio-temporal. Con este resultado se puede excluir el modelo 2, ya que el considerar el patrón espacio-temporal implica la inclusión de una gran cantidad de parámetros que no están produciendo una mejora en la medida de bondad de ajuste.

De esta forma, se puede concluir que la estructura del modelo 2 (interacción tipo II) no logra capturar eficientemente la estructura de los datos de VIH/SIDA en Costa Rica; además, la estimación de los parámetros del modelo 3 es un proceso lento por la complejidad estructural del modelo.

Finalmente, aunque el $LPMP_{CK}$ del modelo 1 es más bajo que el del modelo 3 ($1448,28$ contra $1454,82$), no se puede concluir que el modelo 1 es el que se ajuste mejor, dado que la diferencia no es muy grande. Sin embargo, la complejidad de la estructura de los parámetros del modelo 3 hace que la convergencia sea lenta y que presente autocorrelación en los parámetros de $\boldsymbol{\alpha}$ (patrón temporal). Por lo tanto, se elige el modelo 1 como el más apropiado para ajustar los datos de VIH/SIDA en Costa Rica.

En la Figura 2 se presenta la media *a porteriori* de los riesgos relativos a través del

tiempo para seis cantones: San José, León Cortés, Orotina, Corredores, Tarrazú y Palmares. Se seleccionaron estos 6 cantones con fines ilustrativos, ya que presentan diferentes comportamientos encontrados en los 81 cantones que tiene Costa Rica.



(a) San José (b) León Cortés

(c) Orotina (d) Corredores

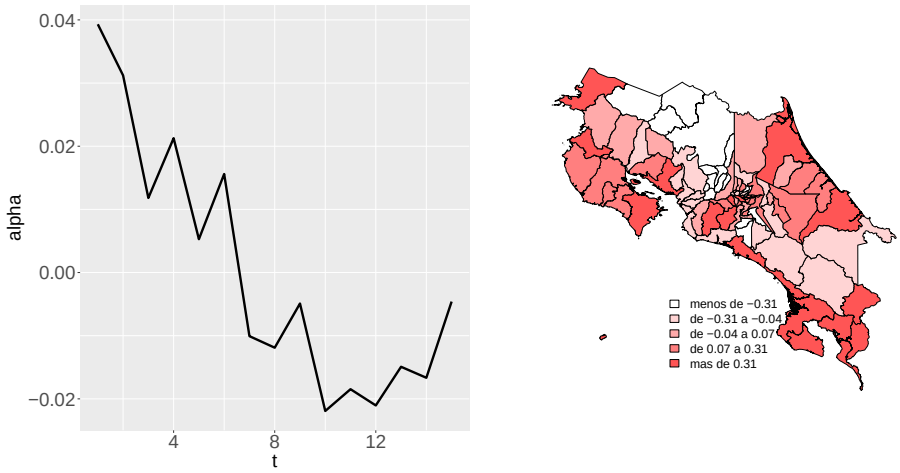
(e) Tarrazú (f) Palmares

Figura 2: *Riesgos relativos estimados con los modelos 1, 3 y 4 para los cantones San José, León Cortés, Orotina, Corredores, Tarrazú y Palmares*

En las figuras 2b y 2c se observa que los cantones de León Cortés y Orotina tienen riesgos relativos de defunciones similares al usar los tres modelos, mientras que para San José, Corredores, Tarrazú y Palmares (figuras 2a, 2d, 2e y 2f) los resultados no son siempre tan parecidos. En estos cuatro cantones los riesgos relativos estimados de defunciones con los modelos 1 y 3 son similares, mientras que los riesgos relativos estimados con el modelo 2 son muy distintos. Además de esto, los riesgos relativos estimados del modelo 2 son más variables a través del tiempo. Esto concuerda con el análisis de validación cruzada ($LPMP_{CV}$) donde se observa que los modelos 1 y 3 tienen mejores ajustes.

Al comparar los riesgos relativos estimados con los modelos 1 y 3 en la figura 2, se nota que los riesgos relativos estimados con el modelo 1 son más suavizados a través de tiempo, mientras que el modelo 3 produce riesgos relativos estimados más variables a través de los años, es decir, las curvas tienen "más picos".

Como uno de los objetivos de los modelos Bayesianos espacio-temporales es suavizar los riesgos relativos con el fin de descubrir patrones espaciales y temporales que están ocultos por los eventos raros, se concluye que el modelo 1 es el que mejor ajusta los datos. Tal decisión concuerda con lo obtenido mediante el $LPMP_{CV}$.



(a) α_t (b) θ_t

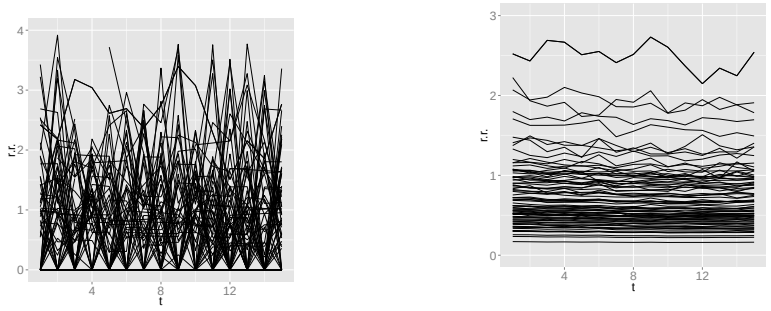
Figura 3: *La media a posteriori de α_t (efecto temporal) y θ_t (efecto espacial).*

3.2.2. Análisis e interpretación de los riesgos relativos

A continuación, se analizan e interpretan los parámetros estimados con el modelo 1. Primer lugar, como el modelo 1 no asume interacción del tiempo y espacio, se pueden interpretar los parámetros estimados por separado. Se puede observar que la media *a posteriori* del efecto temporal presenta una tendencia decreciente pero con una leve tendencia creciente en los últimos años (Figura 3a). Por otro

lado, en la Figura 3b se presenta la media *a posteriori* del efecto espacial para cada cantón. Se puede observar un patrón espacial, cuya la interpretación se hará con los riesgos relativos estimados en la sección 3.2.2. Finalmente, únicamente el intervalo de credibilidad de 95 % $(0,011; 0,028)$ de la covariable porcentaje de viviendas urbanas no contiene a cero. Esto implica que el porcentaje de viviendas urbanas de un cantón está relacionado positivamente con el logarítmo natural de su riesgo relativo. Cabe destacar aquí que aunque los intervalos de credibilidad de las otras dos covariables incluyen cero, se decide incluirlas en el modelo para estimar los riesgos relativos debido a la importancia socioeconómica de estos factores (Poundstone et al. 2004, Zanakis et al. 2007).

Una vez que se decide utilizar el modelo 1, se procede a analizar e interpretar específicamente los riesgos relativos de defunciones *a posteriori* de ese modelo. En la Figura 4b se observan las medias *a posteriori* de los riesgos relativos de defunciones de los 81 cantones a través de tiempo. En este gráfico se observan estimaciones suavizadas que permiten ver los patrones de las defunciones, en cambio en el gráfico de las estimaciones clásicas (Figura 4a), se puede observar una gran inestabilidad ya que las estimaciones se ven afectadas por los conteos bajos de cada año y cantón.



(a) Riesgos relativos clásicos

(b) Media a posteriori de los riesgos relativos (modelo 1)

Figura 4: *Comparación de los riesgos relativos estimados por año de los 81 cantones*

Las curvas son aproximadamente planas para los 81 cantones a través del tiempo. Sin embargo, es posible observar una tendencia muy leve hacia abajo para los cantones que tienen riesgos relativos altos (mayores que 1), mientras que los cantones que tienen riesgos relativos bajos (menores que 1) parecen tener riesgos relativos constantes a través del tiempo. Esto implica que el riesgo de mortalidad del VIH/SIDA de aquellos cantones con riesgos relativos altos ha bajado. Sin embargo, los cantones con riesgo de mortalidad bajo para VIH/SIDA mantienen este índice de manera constante a lo largo del tiempo.

Cuando se habla del VIH/SIDA el periodo de prevalencia es largo y las personas detectadas como positivas emigran a lugares urbanos donde encuentran mejor atención médica (González-Ramírez 2009). Es por eso que los riesgos relativos

en los cantones con riesgos relativos bajos se mantienen constantes en el tiempo. No obstante, una buena señal para los cantones con riesgos relativos altos es que muestran una tendencia decreciente, es decir, que el riesgo de morir por VIH/SIDA es más bajo conforme avanza el tiempo.

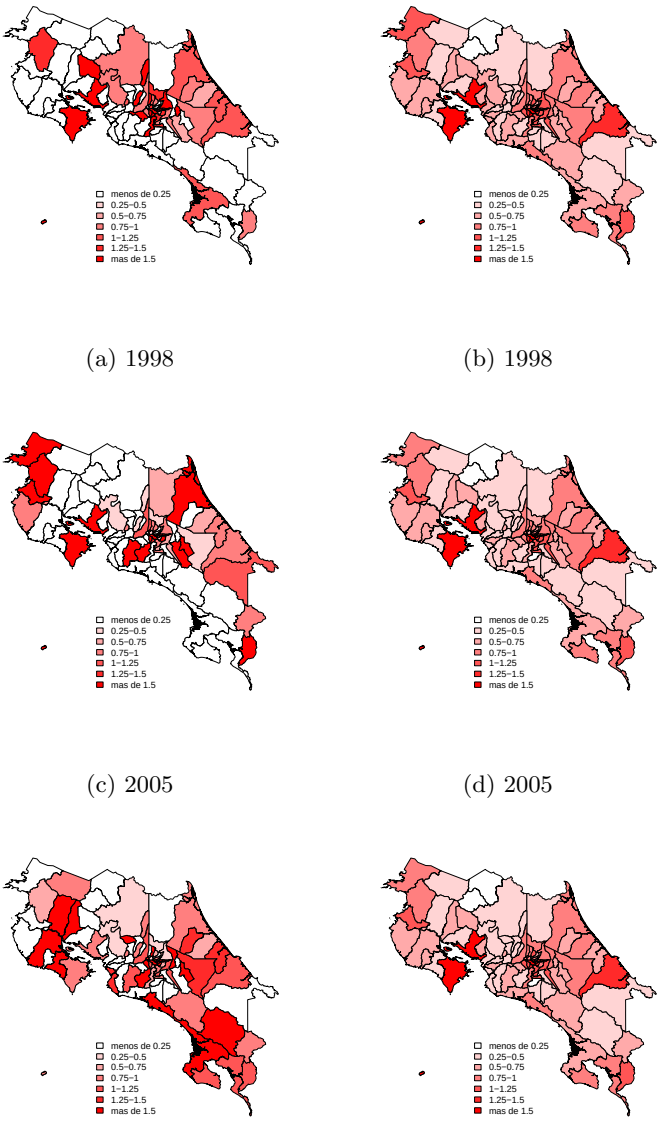
Específicamente, los cantones que tienen riesgos relativos altos son: San José, Puntarenas, Montes de Oca, Tibás, Alajuelita, Goicoechea, Limón, Curridabat, Escazú Corredores, Desamparados, Carrillo, Flores, Cartago, Santa Ana, Vázquez de Coronado, La Unión, San Rafael y La Cruz (en orden descendente). Estos cantones presentan características particulares tales como ser lugares con alta migración, alta urbanización o estar ubicados en zonas fronterizas. Estos resultados coinciden con los análisis descriptivos hechos por el (Ministerio de Salud 2004).

De acuerdo con (González-Ramírez 2009), el VIH afecta a la psicología de los pacientes como el control de sus decisiones, debilita su vida mental, su identidad y su autoestima; por esta razón, las personas detectadas evitan a los conocidos y emigran a lugares urbanos en donde las personas son más variadas y desconocidas. Por otro lado, Corredores y La Cruz corresponden a los cantones que están en la frontera con Nicaragua y Panamá, y Carrillo queda cerca de la frontera, por lo que estos cantones están relacionados con la migración. Estas características de la mortalidad del VIH/SIDA son estudiadas y analizadas por (Poundstone et al. 2004) y (Zanakis et al. 2007) entre otros.

Dado que los datos de VIH/SIDA están disponibles únicamente por un periodo de quince años, gracias a los modelos bayesianos espacio-temporales, con la incorporación de los patrones espaciales y temporales se hace más eficiente la estimación de los riesgos relativos por año y cantón.

La Figura 5 muestra los mapas coropléticos de los riesgos relativos estimados por máxima verosimilitud (razón de mortalidad estandarizada) y la media *a posteriori* de los riesgos relativos de los 81 cantones durante los años 1998, 2005 y 2012. En las Figuras 5b, 5d y 5f los riesgos relativos son más suavizados en comparación con las Figuras 5a, 5c y 5e. Esto se debe a que, como el evento es raro, al analizar los datos por año, algunos cantones en ciertos años tienen una cantidad de defunciones nula ($Y_{it} = 0$), y eso implica que el cálculo de $\hat{r}_{it} = Y_{it}/E_{it}$ va a ser 0. De la misma manera, los riesgos relativos son sobreestimados en los casos en que el conteo de defunciones es relativamente alto. Por lo tanto, muestran más valores extremos que la realidad, por ende, los mapas muestran un cambio de forma más brusca.

En la práctica, si se utilizan las estimaciones de riesgos relativos de defunciones por máxima verosimilitud, los tomadores de decisiones no considerarán en riesgo por VIH/SIDA aquellos cantones que tengan riesgos relativos iguales a cero. Sin embargo, los riesgos de morir por VIH/SIDA en estos cantones no son nulos, existen riesgos de morir por VIH/SIDA aunque sean bajos. El suavizamiento de las estimaciones de riesgos relativos de defunciones *a posteriori* resuelve el problema en que las estimaciones clásicas se ven afectadas por los valores extremos debido a eventos raros.



(a) 1998 (b) 1998

(c) 2005 (d) 2005

(e) 2012 (f) 2012

Figura 5: *Estimación vía máxima verosimilitud (izquierda) y media a posteriori (derecha) de los riesgos relativos de los años 1998, 2005 y 2012*

4. Discusión

Los modelos bayesianos espacio-temporales han sido trabajados desde los años noventa. Gracias a los avances computacionales, las estimaciones de estos modelos son posibles por medio de los algoritmos MCMC; sin embargo, la incorporación conjunta del componente espacial y temporal hace que los modelos sean mucho más complicados y la cantidad de parámetros sea muy grande con respecto a la cantidad de observaciones.

Se mostró la evidencia del patrón espacial de los datos de VIH/SIDA de Costa Rica por medio del índice I de Moran al considerar diferentes definiciones de proximidad; además, se exploró con el mapa de probabilidad, el cual demostró la existencia de valores extremos debido a los conteos bajos. En cuanto al aspecto espacio-temporal se encontró que el modelo 1 es el que mejor ajusta los datos reales y la complejidad de los parámetros de los modelos 2 y 3 no mejora la bondad del ajuste basada en el $LPMP_{CV}$.

Finalmente, se analizaron los riesgos relativos de defunciones estimados por medio del modelo 1. Los cantones que presentan riesgos relativos altos son aquellos caracterizados por ser lugares con alta migración, alta urbanización o estar ubicados en zonas fronterizas, estos son los cantones con puertos, ubicados en la zona fronteriza con Panamá y Nicaragua, o que están localizados en el centro del país donde se encuentran las clínicas y los hospitales especializados en VIH/SIDA.

Los cantones que tienen riesgos relativos bajos mantienen estos riesgos relativos relativamente constantes a través de los años, mientras que los que presentan riesgos relativos altos muestran una leve tendencia decreciente en sus riesgos relativos. Esto también es de esperar ya que las personas emigran a los lugares con mejor atención a los pacientes de VIH/SIDA y los riesgos relativos decrecen debido a la mejora de los servicios y medicamentos en el tiempo.

También se hizo una comparación de los riesgos relativos estimados *a posteriori* con los riesgos relativos vía máxima verosímilitud. El resultado muestra que los estimados usando el modelo 1 son más suavizados y no presentan valores extremos como los de máxima verosímilitud.

La ventaja de los métodos bayesianos estudiados en esta investigación es el suavizamiento de los riesgos relativos en el espacio, se refleja en la interpretación de estos riesgos relativos en la práctica. Las autoridades toman decisiones de salud pública con base en los riesgos calculados a partir de los datos. El suavizamiento de los riesgos relativos de defunciones resuelve el problema en que las estimaciones clásicas de riesgos se ven afectadas por los valores extremos debido a eventos raros; por lo tanto, podría darse una interpretación errónea de que dichos cantones presentan riesgos nulos.

5. Conclusiones

Se llevó a cabo un análisis espacial y luego se hizo otro análisis espacio-temporal sobre las defunciones de VIH/SIDA en el período 1998-2012. Se analizaron los riesgos relativos estimados con el modelo y, Finalmente, se hizo una comparación con la estimación clásica.

La principal limitación que tuvo este trabajo es el aspecto computacional para generar estimaciones de los datos de VIH/SIDA. El uso del programa OpenBUGS limita los algoritmos del MCMC en únicamente el muestreo de Gibbs y Metropolis-Hastings. Existen otras variaciones o técnicas para acelerar la convergencia del MCMC, las cuales requieren programación y son más eficientes en cierto tipo de datos y modelos (Press 2003). Sin embargo, por parte del investigador, esto requiere una mayor habilidad de programación y una dedicación de tiempo más extensa a la que se disponía en este período de realización de la investigación.

Otra limitación fue la convergencia del parámetro γ del modelo 3. Las cadenas de este parámetro presentan cierto grado de autocorrelación, pero se dejó el modelo 3, pues la media ergódica de γ mostró convergencia aceptable.

Sumado a lo anterior, los datos de VIH/SIDA están disponibles solamente por un período de quince años, se espera tener un período de tiempo más largo con el fin de ajustar modelos con un periodo más extenso. Por otro lado, debido a la falta de información previa, se suponen distribuciones *a priori* objetivas, como consecuencia, se utilizan densidades con varianza grande para asumir información previa objetiva debido a la facilidad computacional que tiene el programa; por lo tanto, para el análisis de sensibilidad no es viable variar formas de distribución *a priori* de los modelos ya que se desconoce dicha información. A pesar de esta inconveniencia, se hizo el diagnóstico de secuencias múltiples de Gelman y Rubin al considerar diferentes valores iniciales de las cadenas, como resultado se obtuvo que las diferentes cadenas convergen a la misma media *a posteriori*.

Como trabajos futuros, al usar los datos de VIH/SIDA en Costa Rica, se pueden proponer varias líneas de investigación. En primer lugar, cuando se tenga disponibilidad de datos de más años se podrían ajustar modelos que permitan una mejor visualización del patrón temporal para mejorar las estimaciones espacio-temporales. Posteriormente, se puede incorporar la información *a priori* y analizar la sensibilidad de los cambios de las distribuciones *a priori*. Por otro lado, cuando se habla de las enfermedades, interesa conocer también la morbilidad, incidencia, prevalencia y mortalidad (Lawson & Williams 2001). En un futuro se podrían modelar estos 3 aspectos para entender mejor la distribución del VIH/SIDA en Costa Rica.

En segundo lugar, cabe destacar que para modelizar la tendencia temporal, los modelos analizados utilizan un modelo CAR que imponen una dependencia de los vecinos temporales. Esta estructura es comúnmente utilizada en la práctica; sin embargo, cabe la posibilidad de utilizar modelos dinámicos para estimar la evolución temporal de una forma más natural. Como trabajo futuro, el uso de

otros tipos de dependencia espacial y temporal se puede incorporar para mejorar la estimación de la distribución del VIH/SIDA. Finalmente, este trabajo sirve como base en la modelización espacio-temporal de otros eventos raros, pues dichos eventos presentan la misma dificultad en la estimación de los riesgos relativos.

Recibido: 29 de mayo de 2017
Aceptado: 5 de marzo de 2018

Referencias

Abellan, J. J., Richardson, S. & Best, N. (2008), 'Use of space-time models to investigate the stability of patterns of disease', *Environmental Health Perspectives* **116**(8), pp. 1111–1119.
*http://www.jstor.org/stable/25071151

Agencia de los Estados Unidos para el Desarrollo Internacional (2011), 'An overview to spatial data protocols for HIV/AIDS activities: why and how to include the " where" in your data'.

Altman, B. (2011), Continuing promise 2011: Host nation health brief, Technical report, National Center for Disaster Medicine and Public Health. Recuperado de `http://ncdmph.usuhs.edu/Documents/2011-CRC.pdf` el 17 de marzo del 2015.

American Cancer Society (2014), 'Infección con VIH, SIDA y cáncer', Recuperado de de `http://www.cancer.org/acs/groups/cid/documents/webcontent/002296-pdf.pdf` el 13 de marzo del 2015.

Bailey, T. & Gatrell, A. (1995), *Interactive Spatial Analysis*, Prentice Hall, Harlow.

Banerjee, S., Carlin, B. & Gelfand, A. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC, Boca Raton.

Besag, J., York, J. & Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics* **43**(1), 1–20.
*http://dx.doi.org/10.1007/BF00116466

Bivand, R. & Lewin-Koh, N. (2014), *maptools: Tools for reading and handling spatial objects*. R package version 0.8-30.
*http://CRAN.R-project.org/package=maptools

Brooks, S. P. & Gelman, A. (1998), 'General methods for monitoring convergence of iterative simulations', *Journal of Computational and Graphical Statistics* **7**(4), 434–455. `http://www.stat.columbia.edu/~gelman/research/published/brooksgelman2.pdf`.

Cai, B., Lawson, A. B., Hossain, M. M. & Choi, J. (2012), 'Bayesian latent structure models with space-time dependent covariates', *Statistical Modelling* **12**(2), 145–164.
*http://smj.sagepub.com/content/12/2/145.abstract

Centro Centroamericano de Población (2014), Consulta del 10 de octubre del 2014, de la base de datos del Censo del 2011 y de defunciones de Centro Centroamericano de Población: `http://ccp.ucr.ac.cr`.

Fortunato, L., Abellan, J., Beale, L., LeFevre, S. & Richardson, S. (2011), 'Spatio-temporal patterns of bladder cancer incidence in utah (1973-2004) and their association with the presence of toxic release inventory sites', *International Journal of Health Geographics* **10**(1), 16.
*http://www.ij-healthgeographics.com/content/10/1/16

Gelfand, A., Dey, D. & Chang, H. (1992), 'Model determination using predictive distributions with implementation via sampling based methods (with discussion)', *Bayesian Statistics* **4**, 147–67.

González-Ramírez, V. (2009), 'Intervención psicológica en VIH/SIDA', *UARICHA Revista de Psicología* (13), pp. 49–63.

Hunter, L. M., Souza, R.-M. D. & Twine, W. (2008), 'The environmental dimensions of the HIV/AIDS pandemic: A call for scholarship and evidence-based intervention', *Population and Environment* **29**(3/5), pp. 103–107.
*http://www.jstor.org/stable/40212350

Hyndman, R. & Khandakar, Y. (2008), 'Automatic time series forecasting: the forecast package for R', *Journal of Statistical Software* **26**(3), 1–22.
*http://ideas.repec.org/a/jss/jstsof/27i03.html

Lagazio, C., Biggeri, A. & Dreassi, E. (2001), 'A hierarchical bayesian model for space-time variation of disease risk', *Statistical Modelling* **1**, 17–29.

Lagazio, C., Biggeri, A. & Dreassi, E. (2003), 'Age-period-cohort models and disease mapping', *Environmetrics* **14**(5), 475–490.
*http://dx.doi.org/10.1002/env.600

Langford, I. H. (1994), 'Using empirical bayes estimates in the geographical analysis of disease risk', *Area* **26**(2), pp. 142–149.
*http://www.jstor.org/stable/20003399

Lawson, A. & Williams, F. (2001), *An Introductory Guide to Disease Mapping*, John Willy and Sons, New York.

Ministerio de Salud (2004), 'La situación del VIH/SIDA en Costa Rica'.

Organización Mundial de la Salud (2014), 'Health for the world's adolescents: A second chance in the second decade'.

Organización Panamericana de la Salud (2004), 'La situación del VIH/SIDA en Costa Rica'.

Paradis, E., Claude, J. & Strimmer, K. (2004), 'APE: analyses of phylogenetics and evolution in R language', *Bioinformatics* **20**, 289–290.

Pebesma, E.J., R. B. (2005), 'Classes and methods for spatial data in r', *R News* **5**(2).
*http://cran.r-project.org/doc/Rnews/

Plummer, M. (2008), 'Penalized loss functions for bayesian model comparison', *Biostatistics* **9**(3), 523–539.
*http://biostatistics.oxfordjournals.org/content/9/3/523.abstract

Poundstone, K. E., Strathdee, S. A. & Celentano, D. D. (2004), 'The social epidemiology of human immunodeficiency virus/acquired immunodeficiency syndrome', *Epidemiologic Rev* **26**(1), pp. 22–35.

Press, J. (2003), *Subjective and Objective Bayesian Statistics*, John Wiley and Son, New Jersey.

Richardson, S., Abellan, J. J. & Best, N. (2006), 'Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in yorkshire (uk)', *Statistical Methods in Medical Research* **15**(4), pp. 385–407.

Richardson, S., Thomson, A., Best, N. & Elliott, P. (2004), 'Interpreting posterior relative risk estimates in disease-mapping studies', *Environmental Health Perspectives* **112**(9), pp. 1016–1025.
*http://www.jstor.org/stable/3838103

Ripley, B. (2004), *Spatial Statistics*, John Wiley and Sons, Ltd., New Jersey.

Schmid, V. & Held, L. (2004), 'Bayesian extrapolation of space-time trends in cancer registry data', *Biometrics* **60**(4), pp. 1034–1042.
*http://www.jstor.org/stable/3695483

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.
*http://dx.doi.org/10.1111/1467-9868.00353

Team, R. C., Wuertz, D., Setz, T. & Chalabi, Y. (2014), *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3011.87.
*http://CRAN.R-project.org/package=fBasics

Thomas, A., O'Hara, B., Ligges, U. & Sturtz, S. (2006), 'Making bugs open', *R News* **6**(1), 12–17.
*http://cran.r-project.org/doc/Rnews/

Waller, L. A., Carlin, B. P., Xia, H. & Gerfand, A. E. (1997), 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association* **92**, 607–17.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Springer New York.
*http://had.co.nz/ggplot2/book

Zanakis, S. H., Alvarez, C. & Li, V. (2007), 'Socio-economic determinants of HIV/AIDS pandemic and nations efficiencies', *European Journal of Operational Research* **176**, pp. 1811–1838.

Zeileis, A. & Hothorn, T. (2002), 'Diagnostic checking in regression relationships', *R News* **2**(3), 7–10.
*http://CRAN.R-project.org/doc/Rnews/