Aplicación de modelos jerárquicos Bayesianos espacio-temporales a la información sobre VIH/SIDA de Costa Rica

Shu Wei Chou Chen

Escuela de Estadística

Centro Centroamericano de Población

27 de agosto de 2015

Contenidos

Introducción

- La interacción social y los factores ambientales y culturales juegan un papel importante en la propagación de las enfermedades infecciosas.
- ¿Independencia de los datos?
- La estadística espacial logra modelar tal situación.
- Tradicionalmente se analizan espacialmente los datos ignorando el componente temporal.
- Cuando la enfermedad es rara (conteos de datos escasos), si se analizan los datos por período de tiempo, la estimación del riesgo de la enfermedad es imprecisa.
- Solución: Estadística Bayesiana.

Introducción

- El virus inmunodeficiencia humana (VIH) es el virus que causa el síndrome de inmunodeficiencia adquirida (SIDA). Altera el sistema inmunológico y destruye la capacidad del cuerpo para defender otras enfermedades.
- VIH/SIDA es **la segunda causa de muerte** más importante entre los adolescentes de 10 a 19 años después de los accidentes de tránsito (Organización Mundial de la Salud, 2014).
- En Costa Rica, VIH/SIDA es una de las tres principales causas de muerte de personas de todas edades en el año 2002, junto con la cardiopatía isquémica y enfermedades cerebrovascular (Altman, 2011).

VIH/SIDA

- Organización Panamericana de Salud (2004) analiza descriptivamente demostrando las evidencias espaciales y temporales por separado.
- (Zanakis et al., 2007) analizó a nivel de los países los factores como desventajas económicas y de salud, migración (pobreza), etc. Encontraron que un país con menos incidencia de VIH posee las siguientes características:
  - densidad de población baja,
  - mejores sistemas de salud (más doctores, enfermedas y camas por hospital), y
  - mejor desarrollo en los medios de comunicación.

Distancia y vecindario

Se define la matriz de proximidad

$$\boldsymbol{W} = \{w_{ij}\}$$

$$\text{para } i = 1, ..., n \text{ y } j = 1, ..., n$$

.

- Distancia fija tipo II:

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

- Vecino que comparte la frontera:

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

- K centroides más cercanos:

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

Índice de I de Moran

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

donde w_{ij} es el elemento ij de la matriz de proximidad W.

- I es asintóticamente normal con

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

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

- (Banerjee et al., 2014) Debido al problema de convergencia asintótica es lenta, se recomienda el uso de I de Moran para análisis exploratorio espacial.

Riesgos relativos

- Sea Y_{it} el conteo del evento para el área i y el tiempo t ($i = 1, ..., n$ y $t = 1, ..., T$).

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

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

donde

  - p^* es la probabilidad global de que ocurra el evento en toda la región del estudio,
  - E_{it} es cantidad esperada global, y
  - r_{it} es riesgo relativo.

- En Estadística clásica, el estimador de máxima verosimilitud del riesgo relativo es:

$$\hat{r} = \frac{p_{it}}{p^*} = \frac{O_{it}/N_{it}}{E_{it}/N_{it}} = \frac{O_{it}}{E_{it}}$$

donde $E_{it} = \dfrac{\sum_{i,t} O_{it}}{\sum_{i,t} N_{it}} \cdot N_{it}$

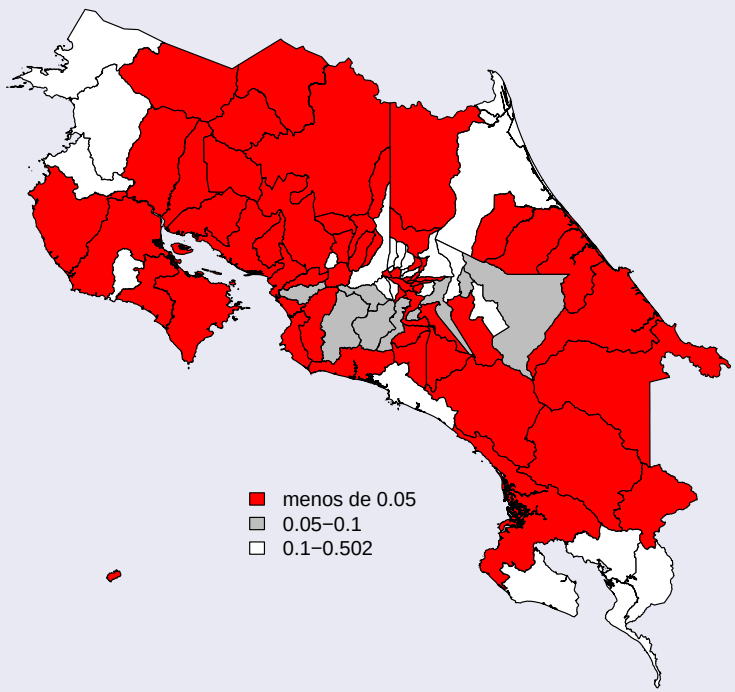
Mapeo de probabilidades

Visualización de los valores extremos.

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

donde
$\hat{E}_i$ es el conteo esperado en el área i
y_i es el conteo observado en el área i

Mapa de probabilidades de defunciones de VIH/SIDA (1998-2012)



Estimación Bayesiana

1. Problemas en la estimación de los riesgos relativos en los conteos bajos:
   - La variación del riesgo relativo es grande.
   - El valor de p siempre es pequeño en poblaciones grandes debido a que el error estándar de los riesgos relativos es pequeño.
2. Formulación del teorema de Bayes (caso continuo)

$$h(\theta|x_1,...,x_n) = \frac{f(x_1,...,x_n|\theta)g(\theta)}{\int f(x_1,...,x_n|\theta)g(\theta)d\theta},$$

3. Implementación de MCMC para el cálculo de la distribución *a posteriori*, pues generalmente las estimaciones involucran integrales de altas dimensiones y no tienen soluciones analíticas.
   - Metropolis-Hastings
   - Muestreo de Gibbs

Modelo autoregresivo condicional(CAR)

$$Y_i \sim \text{Poisson}(\mu_i = E_i r_i)$$

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

donde los s_i son los efectos espaciales estructurados y u_i son los no estructurados.
La estructura para modelar el componente espacial es

$$s_i | s_{j,j\neq i} \sim N\left(\frac{\sum_j w_{ij} s_j}{\sum_j w_{ij}}, \frac{\sigma_s^2}{\sum_j w_{ij}}\right)$$

- CAR espacial: $\boldsymbol{\theta} \sim \text{CAR}(\boldsymbol{W}, \sigma_\theta^2)$
- CAR temporal: $\boldsymbol{\alpha} \sim \text{CAR}(\boldsymbol{Q}, \sigma_\alpha^2)$

Modelos espacio-temporales

Knorr-Held (2000)

$$ln(r_{it}) = \mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it}$$

- Las distribuciones a priori de α, γ, θ y ϕ siguen una distribución multinormal con media 0 y precisión $\kappa \boldsymbol{K}$.
- Para γ y ϕ, se asume una distribución a priori intercambiable con $\boldsymbol{K} = I$ (matriz identidad).
- $\boldsymbol{\theta} \sim \text{CAR}\left(\boldsymbol{W}, \frac{1}{\kappa_\theta}\right)$
- $\boldsymbol{\alpha} \sim \text{CAR}\left(\boldsymbol{Q}, \frac{1}{\kappa_\alpha}\right)$

Modelos espacio-temporales

Para δ se tienen 4 interacciones posibles:

I γ **con** ϕ**:** el tiempo y el espacio no interactúan entre sí.

II α **con** ϕ**:** captura información a través del tiempo para cada unidad geográfica y es independiente para cada área. Por lo tanto, cuando el patrón temporal es diferente para cada área y no interactúa con el patrón espacial, este modelo es el más apropiado.

III γ **con** θ**:** el patrón espacial es diferente para cada tiempo t.

IV α **con** θ**:** una unidad geográfica es dependiente de los tiempos adyacentes, de la información de sus vecinos y de los tiempos adyacentes de los vecinos.

Modelos

Modelo 1 (Richardson et al., 2006)

Interacción tipo I

$$ln(r_{it}) = \mu + \boldsymbol{x_i}\boldsymbol{\beta} + \theta_i + \alpha_t + v_{it}, \quad (1)$$

con las distribuciones *a priori*:

$\mu,\ \beta_i \sim \mathcal{N}(0; 10^4)$,

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

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

$v_{ij} \sim \mathcal{N}(0, 1/\tau_v)$, y

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

Modelos

Modelo 2 (Waller et al., 1997)

Interacción tipo III

$$ln(r_{it}) = \mu + \boldsymbol{x_i}\boldsymbol{\beta} + \alpha_t + \phi_i^{(t)} + v_i^{(t)}, \tag{2}$$

con las distribuciones *a priori*:

$\mu,\ \beta_i \sim \mathcal{N}(0; 10^4)$,

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

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

$v_i^{(t)} \overset{iid}{\sim} N(0, 1/\tau_t)$, y

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

Modelos

Modelo 3 (Lagazio et al., 2001,0; Schmid y Held, 2004)

Interacción tipo IV

$$ln(r_{it}) = \mu + \boldsymbol{x_i}\boldsymbol{\beta} + \alpha_t + \theta_i + v_{it}, \quad (3)$$

con las distribuciones *a priori*:

$$\mu,\ \beta_i \sim \mathcal{N}(0; 10^4),$$
$$\boldsymbol{\theta} \sim \text{CAR}(\boldsymbol{W}, 1/\tau_\theta),$$
$$\boldsymbol{\alpha} \sim \text{CAR}(\boldsymbol{Q}, 1/\tau_\alpha),$$
los hiperparámetros $\tau_\theta, \tau_\alpha \sim \text{Gamma}(0,5; 0,0005)$.

Se puede notar que la especificación de las distribuciones a priori de este tipo, 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.

Modelos

Modelo 4

Regresión lineal

$$ln(r_{it}) = \mu + \boldsymbol{x_i}\beta + v_{it}, \tag{4}$$

con las distribuciones *a priori*:

$\mu,\ \beta_i \sim \mathcal{N}(0; 10^4)$,

$v_{ij} \sim \mathcal{N}(0, 1/\tau_v)$, y

los hiperparámetros $\tau_v \sim \text{Gamma}(0,5; 0,0005)$.

Criterios para seleccionar el mejor modelo

Ordenada predictiva condicional

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

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

$$NLLK_{CV} = -\sum_{i=1}^{n}\sum_{t=1}^{T} \log \mathrm{CPO}_{it}$$

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

Aplicación de datos reales

Variable dependiente

Defunciones por causa de VIH/SIDA por cantón y año en 1997-2012.

Variables independientes

- % viviendas urbanas
- % población entre 24 y 49 años.
- tasa de mortalidad infantil.

Programas

- R versión 3.1.2
- OpenBUGS versión 3.2.3

Análisis exploratorio espacial

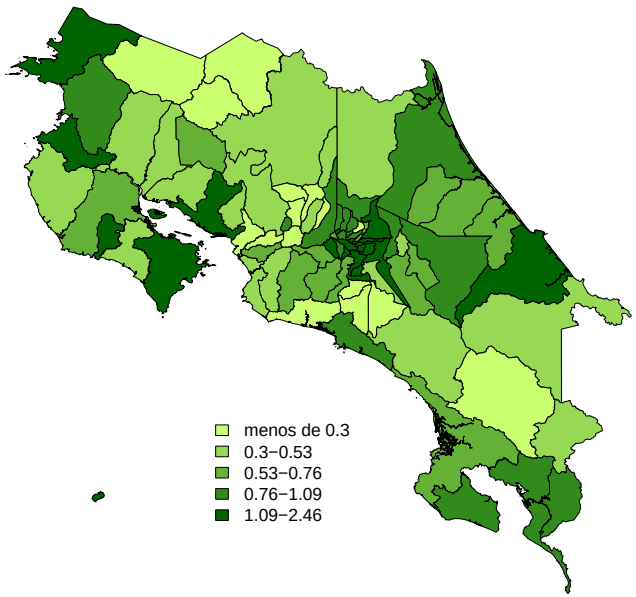
Definición de la matriz de proximidad		I de Moran	valor-p
Distancia fija tipo II[1]		0,121	0,033
Vecino que comparte la frontera		0,2985	<0,001
k centroides más cercanos	k=1	0,238	0,062
	k=2	0,317	0,001
	k=3	0,324	<0,001
	k=4	0,338	<0,001

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

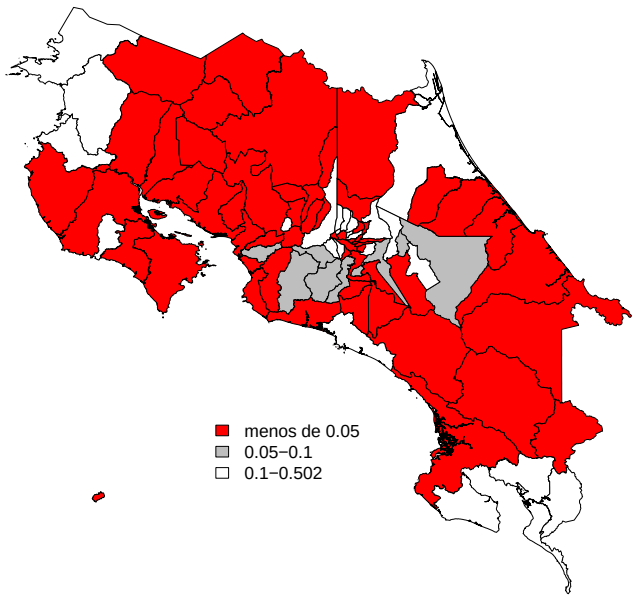
Cuadro: El índice I de Moran usando diferentes definiciones de matriz de proximidad aplicado a los datos de VIH/SIDA en Costa Rica (1997-2012)

Análisis exploratorio espacial

1. (Bailey y Gatrell, 1995) 51 de 81 cantones presentan riesgos extremadamente altos o bajos pues $p_i < 0,05$.



(a) Mapa de riesgos relativos (estimación de estadística clásica)



(b) Mapa de probabilidad

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

Análisis espacio-temporal

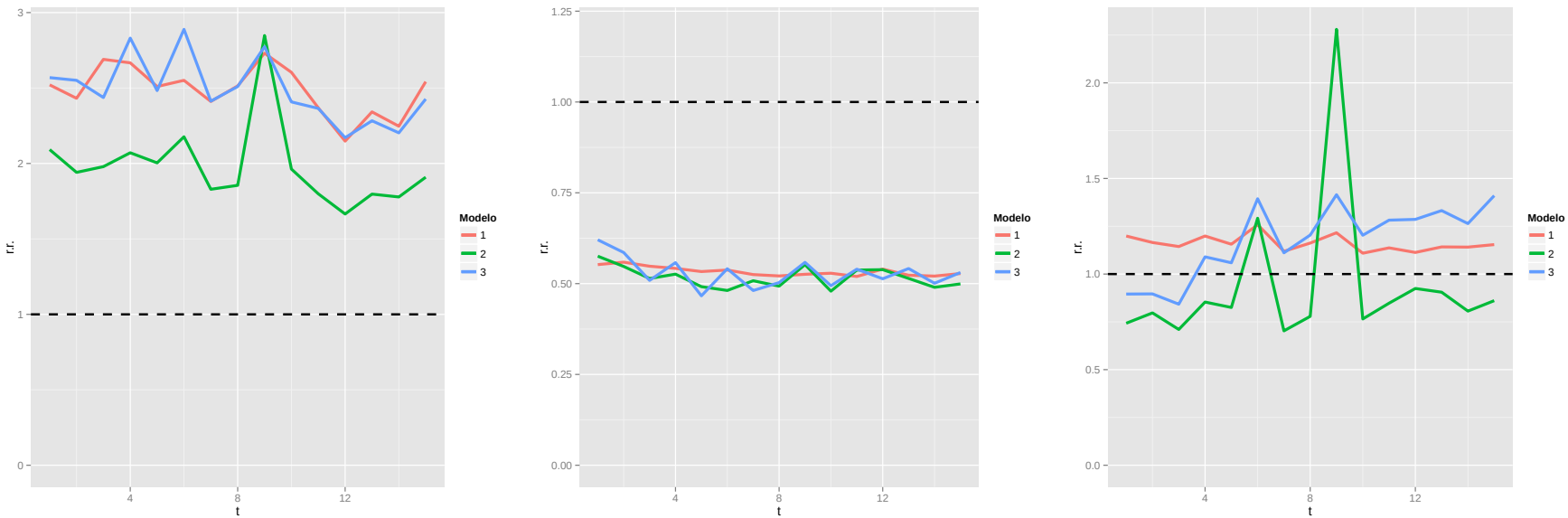
Diagnósticos

Modelo	período de calentamiento	iteraciones	*thinning*	$NLLK_{CV}$
1	10.000	10.000	10	**1448,3**
2	50.000	10.000	20	1534,1
3	4.000.000	10.000	50	**1454,8**
4	50.000	10.000	5	1512,5

Cuadro: La especificación de los parámetros del MCMC de los modelos y la bondad de ajuste

- Dichos parámetros fueron escogidos debido a que la complejidad de los modelos es distinta.
- Los parámetros de la cadena del modelo 3 son sugeridos en Lagazio et al. (2001).

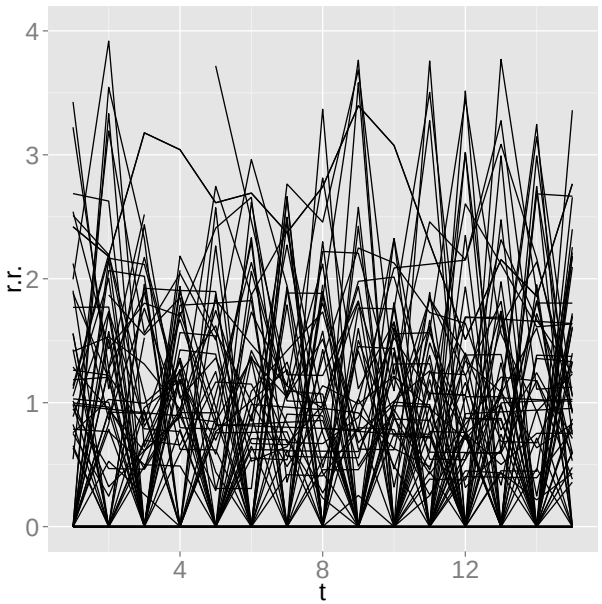
Análisis espacio-temporal



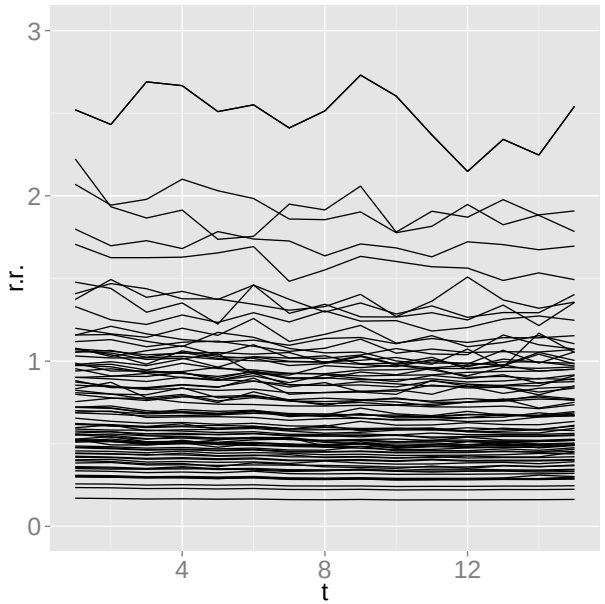
(a) San José **(b)** Orotina **(c)** Corredores

Figura: Riesgos relativos estimados con los modelos 1, 2 y 3

Análisis espacio-temporal: interpretación de los riesgos relativos



(a) los riesgos relativos clásicos



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

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

Análisis espacio-temporal: interpretación de los riesgos relativos

- Cantones con riesgos más altos: 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.
- Caracterizados por ser de alta migración, urbanización y zonas fronterizas.
- Resultado coincide con los análisis descriptivos hechos por el Ministerio de Salud (2004)

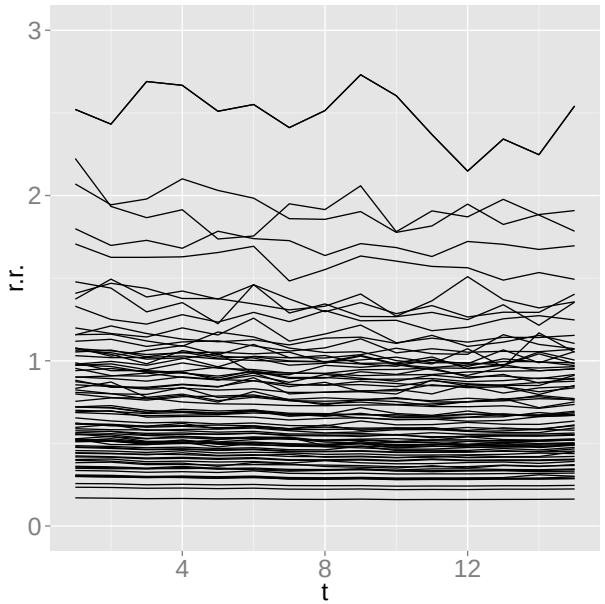
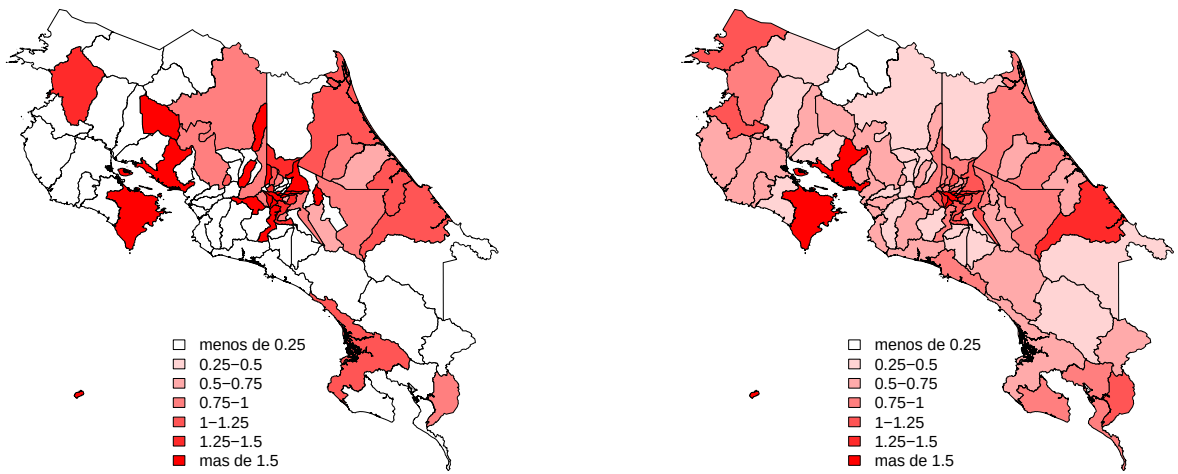


Figura: Media a posteriori de los riesgos relativos del modelo 1 por año de los 81 cantones

Análisis espacio-temporal: interpretación de los riesgos relativos

- (González-Ramírez, 2009) el VIH afecta a la psicología de los pacientes:
  - control de sus decisiones,
  - debilita su vida mental, su identidad y su autoestima.
- las personas detectadas evitan a los conocidos y emigran a lugares urbanos en donde las personas son más variadas y desconocidas. Además, estos lugares tienen mejor atención de la salud.

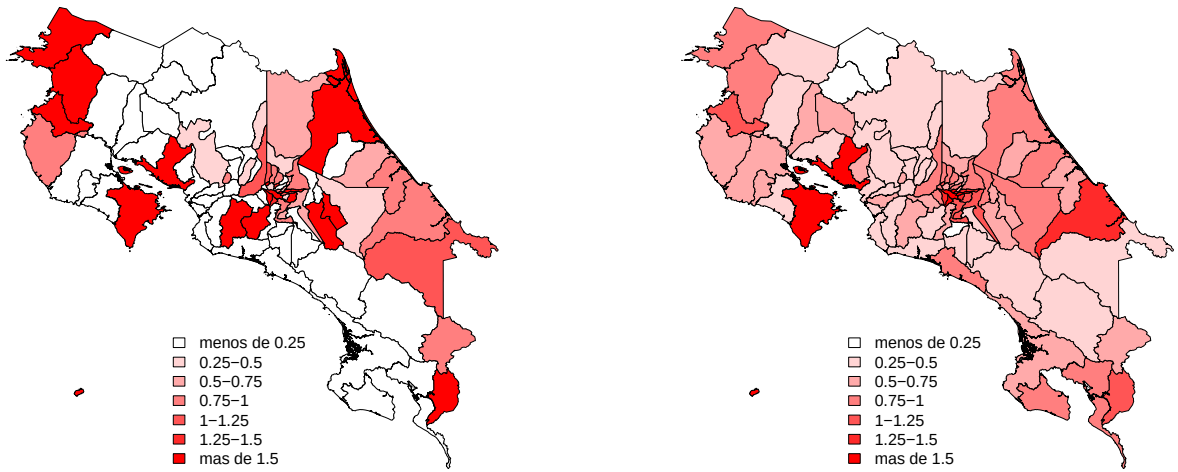
Análisis espacio-temporal: interpretación de los riesgos relativos



(a) Máxima verosimilitud **(b)** Media *a posteriori*

Figura: Estimación de riesgos relativos en el año 1998

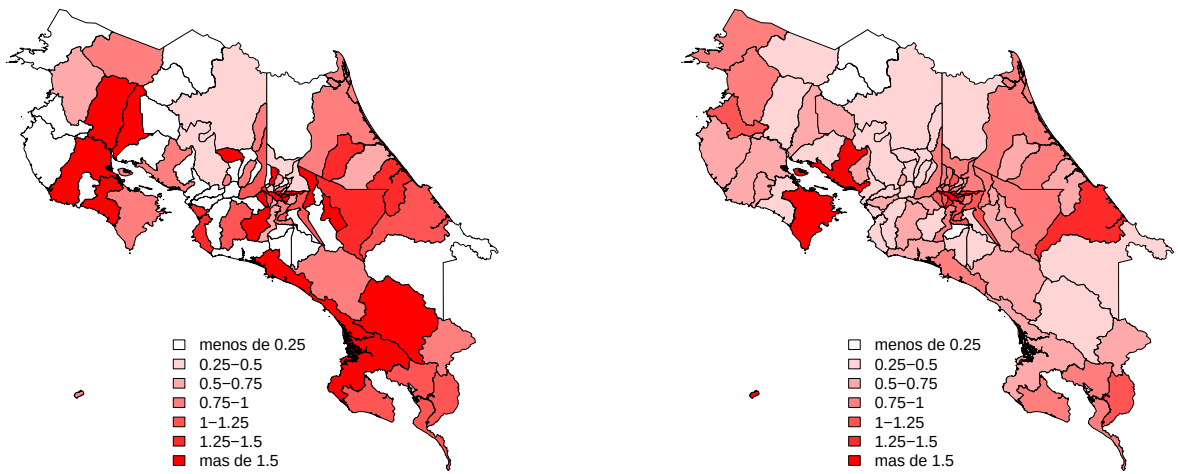
Análisis espacio-temporal: interpretación de los riesgos relativos



(a) Máxima verosimilitud **(b)** Media *a posteriori*

Figura: Estimación de riesgos relativos en el año 2005

Análisis espacio-temporal: interpretación de los riesgos relativos



(a) Máxima verosimilitud **(b)** Media *a posteriori*

Figura: Estimación de riesgos relativos en el año 2012

Conclusiones

Limitaciones

- Aspecto computacional para generar las estimaciones Bayesianas.
- La alta autocorrelación del parámetro γ del modelo 3.
- Disposición de los datos de VIH/SIDA de únicamente 16 años.
- No consideración de la información *a priori*.

Conclusiones

Discusión

- Se demostró evidencia espacial usando diferentes definiciones de matriz de proximidad.
- Se exploró con el mapa de probabilidad y se demostró la existencia de valores extremos debido a conteos bajos.
- El modelo 1 ajusta mejor a pesar de que los modelos 2 y 3 presentan estructuras espacio-temporales más complejas.
- Los cantones que presentan riesgos relativos altos están caracterizados por la alta migración, urbanización o estar ubicados en zonas fronterizas.
- Los riesgos relativos de estos cantones decrecen a lo largo de tiempo, debido a la mejora de los servicios y medicamentos en el tiempo.
- Los riesgos relativos de otros cantones se mantienen constantes y bajos a lo largo de tiempo.
- Se compararon las estimaciones Bayesianas con las de máxima verosimilitud y se obtuvo que las primeras son más suavizadas.

Conclusiones

Trabajos futuros

- Análisis espacio-temporal de casos diagnosticados de VIH y de SIDA.
- Ajuste de modelos con el período de estudio más largo.
- Incorporación de información *a priori*.

¡Gracias!

Bibliografía I

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

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

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

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

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567.

Bibliografía II

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

Lagazio, C., Biggeri, A., y Dreassi, E. (2003). Age-period-cohort models and disease mapping. *Environmetrics*, 14(5):475–490.

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 Salud (2004). La situación del VIH/SIDA en Costa Rica.

Richardson, S., Abellan, J. J., y 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.

Bibliografía III

Schmid, V. y Held, L. (2004). Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, 60(4):pp. 1034–1042.

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

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