ANÁLISE DE DADOS DE SOBREVIVÊNCIA ESPACIALMENTE CORRELACIONADOS

Calisto Manuel MÁQUINA[1]
Fábio Nogueira DEMARQUI[2]
Marcos Oliveria PRATES[2]
Nerilson Terra SANTOS[1]

■ RESUMO: O ajuste de modelos de fragilidade para dados de sobrevivência espacialmente correlacionados é discutido. Considerando uma abordagem Bayesiana completamente paramétrica, o modelo ICAR (*intrinsically conditional autoregressive*) é assumido para a distribuição das fragilidades, enquanto a distribuição de base dos tempos de sobrevivência é modelada pela distribuição Weibull. Um critério para a seleção de modelos Bayesianos é discutido. Uma descrição detalhada de como utilizar o *software* `R` em conjunto com o *software* `OpenBUGS` para realizar inferência em modelos de sobreviência espacial é fornecida. A análise de um banco de dados reais envolvendo pacientes diagnosticados com câncer de bexiga no estado de Connecticut, EUA, é apresentada.

■ PALAVRAS-CHAVE: Fragilidade espacial; modelo ICAR, inferência Bayesiana; OpenBUGS.

1 Introdução

O avanço computacional observado nas últimas décadas, aliado ao desenvolvimento de novas metodologias estatísticas, tem possibilitado a modelagem de dados com estruturas cada vez mais complexas. Neste contexto, modelos estatísticos que incorporam efeitos aleatórios, isto é, variáveis não observáveis que são inseridas na modelagem a fim de acomodar diferentes tipos de estruturas de dados, têm se tornado cada vez mais frequentes.

[1]Universidade Federal de Viçosa - UFV, Departamento de Estatística, CEP: 36570-977, Viçosa, MG, Brasil. E-mail: *calisto.calmaqna.maquina@gmail.com; nsantos@ufv.br*
[2]Universidade Federal de Minas Gerais - UFMG, Departamento de Estatística, CEP: 31270-901, Belo Horizonte, MG, Brasil. E-mail: *fndemarqui@est.ufmg.br; marcosop@est.ufmg.br*

Em análise de sobrevivência os modelos com efeitos aleatórios, também conhecidos como modelos de fragilidades (HOUGAARD, 2000; DUCHATEAU; JANSSEN, 2010; WIENKE, 2010; HANAGAL, 2011), https://start.fedoraproject.org/, têm se tornado bastante populares para a modelagem dados de sobrevivência correlacionados. Exemplos de dados de sobrevivência correlacionados compreendem tempos de sobrevivência associados a: i) membros de uma mesma família (presença de grupos ou conglomerados); ii) indivíduos sujeitos a múltiplos eventos do mesmo tipo (eventos recorrentes); ou ainda, iii) dados de sobrevivência com padrão espacial (espacialmente correlacionados). Em qualquer desses casos, a suposição de independência entre tempos associados a indivíduos não é razoável, e a negligência de tal dependência pode comprometer a análise estatística dos dados.

O modelo de fragilidade mais comum é o chamado modelo de fragilidade compartilhada, que corresponde a uma generalização do modelo de taxas de falhas proporcionais proposto por (COX, 1972). Este modelo é definido da seguinte forma. Seja t_{il} o tempo de sobrevivência associado ao i-ésimo elemento pertencente ao l-ésimo grupo, $i = 1, \ldots, n_l$ e $l = 1, \ldots, L$, em que n_l representa o número de elementos pertencentes ao l-ésimo conglomerado, e denote por z_l a fragilidade (não observável) associada ao l-ésimo grupo. Então, dado z_l, a função de taxa de falha condicional é dada por:

$$\begin{aligned} h(t_{li}|x_{li}, z_l) &= h_0(t_{li}) z_l \exp\{\mathbf{x}_{li}\boldsymbol{\beta}\} \\ &= h_0(t_{li}) \exp\{\mathbf{x}_{li}\boldsymbol{\beta} + \omega_l\}, \end{aligned} \tag{1}$$

em que $h_0(\cdot)$ denota a função de taxa de falha de base, comum a todos os elementos, e que pode ser modelada de maneira paramétrica ou não paramétrica, $\mathbf{x}_{il}$ é um vetor $1 \times p$ de covariáveis associadas ao i-ésimo elemento pertencente ao l-ésimo grupo, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de coeficientes não conhecidos da regressão e $\omega_l = \log(z_l)$, $l = 1, \ldots, L$.

Como pode ser observado a partir da Equação (1), a fragilidade atua de maneira multiplicativa na função taxa de falha. Desta forma, quanto maior o valor da fragilidade, mais frágil (ou suscetível ao evento de interesse) é o indivíduo ou elemento. Além disso, a variância da fragilidade fornece um indicativo da heterogeneidade associada aos grupos ou conglomerados (quanto maior/menor a variância, maior/menor a heterogeneidade entre os grupos).

As escolhas mais comuns para a fragilidade são as distribuições gama e log-normal (THERNEAU; GRAMBSCH, 2000). As versões semiparamétricas dos modelos de fragilidade gama e lognormal, isto é, quando $h_0(\cdot)$ é modelada de forma não-paramétrica, estão implementadas em diversos *softwares* estatísticos, tais como o `R` (R CORE TEAM, 2018) e `SAS` (⟨www.sas.com⟩). Por esse motivo, estes modelos têm se tornado bastante populares em análise de sobrevivência.

Os modelos de fragilidade compartilhada gama e log-normal tem sido largamente utilizados para modelar dados de sobrevivência em situações como as descritas em (i) e (ii). Entretanto, esses modelos não são adequados para a modelagem de dados de sobrevivência espacialmente correlacionados, pois não

são capazes de acomodar o padrão de dependência espacial presente nos dados, uma vez que é razoável esperarmos que elementos pertencentes a áreas adjacentes sejam mais similares entre si do que aqueles elementos provenientes de áreas não adjacentes. Neste contexto, o modelo ICAR (*intrisically conditional autoregressive*; veja Banerjee, Carlin e Gelfand (2003) para uma excelente revisão deste modelo) fornece uma maneira bastante atrativa para modelarmos efeitos espaciais associados a dados de área como o do exemplo (iii).

Neste trabalho discutiremos uma abordagem Bayesiana completamente paramétrica para a modelagem de dados de sobrevivência espacialmente distribuídos em áreas de um determinado mapa. Especificamente, será considerada a distribuição Weibull para a modelagem da função taxa de falha de base, e o modelo ICAR será utilizado para acomodar a dependência espacial entre os tempos de sobrevivência observados. Um banco de dados reais envolvendo pacientes diagnosticados com câncer de bexiga será utilizado para ilustrar o modelo estudado, e uma discussão sobre como ajustar o modelo utilizando o *software* `OpenBUGS` (LUNN et al., 2009) utilizando-se o *software* `R` através do pacote `rbugs` (YAN; PRATES, 2013) é apresentada.

Este trabalho está organizado da seguinte forma. Na Seção 2 é apresentado o arcabouço teórico para o ajuste do modelo de fragilidade ICAR. Especificamente, uma revisão sobre o modelo ICAR é fornecida na Subseção 2.1. O modelo de fragilidade compartilhada é apresentado na Subseção 2.2. Na Subseção 2.3 a abordagem Bayesiana para o ajuste do modelo é discutida. Um método para a seleção de modelos Bayesianos é descrito na Subseção 2.4. A análise dos tempos de vida de paciêntes diagnosticados com câncer de bexiga é explorada na Seção 3. Finalmente, na Seção 4 são apresentadas as conclusões sobre o trabalho.

2 Formulação do modelo

Na sequência apresentaremos a estrutura básica para a modelagem paramétrica de sobrevivência espacialmente correlacionados, considerando a abordagem Bayesiana para situações nas quais os elementos/indivíduos estão espacialmente distribuídos em áreas associadas a um determinado mapa.

2.1 O modelo ICAR

A literatura relacionada a dados espacialmente correlacionados é bastante rica. Importantes referências incluem, por exemplo, os livros de Cressie (1993) e Banerjee, Carlin e Gelfand (2003), entre outros.

Quando os grupos de elementos/indivíduos estão espacialmente distribuídos em áreas associadas a um determinado mapa, é comum usarmos o termo fragilidade espacial para nos referimos ao efeito aleatório utilizado para modelar a associação entre elementos pertencentes a uma mesma unidade de área. O termo fragilidade espacial também pode ser empregado para efeitos aleatórios associados a dados de

sobrevivência geo-referenciados. Entretanto, este tipo de abordagem está fora do escopo deste trabalho, e será deixada para trabalhos futuros.

Seja $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_L)$ o vetor não observável de fragilidades, em que ω_l é a fragilidade associada à l-ésima área do mapa. O modelo ICAR pode ser utilizado para introduzir uma estrutura de dependência espacial entre as componentes do vetor $\boldsymbol{\omega}$. Este modelo é construído com base em uma matriz $L \times L$ de proximidade (ou vizinhanças) V, cujas entradas v_{lr} assumem valor 1 se as áreas l e r possuem fronteiras comuns (evento denotado por $l \sim r$), e zero caso contrário. De acordo com Banerjee, Carlin e Gelfand (2003), a distribuição condicional completa (a *priori*) associada a cada unidade de área é dada por:

$$(\omega_l | \boldsymbol{\omega}_{(-l)}) \sim N(\bar{\omega}_l; 1/(\phi w_{l.})),\ l = 1, \ldots, L, \tag{2}$$

em que $\boldsymbol{\omega}_{(-l)} = (\omega_r : r \neq l)$, $w_{l.} = \sum_{r=1}^{L} w_{lr}$, de tal forma que $w_{l.}$ corresponde ao número de vizinhos da área l, e ϕ é um parâmetro de precisão que controla a variabilidade comum associada ao conjunto de todas as áreas sob investigação.

É importante notar a partir de (2) que a fragilidade espacial ω_l segue uma distribuição normal cuja média é dada pela média das fragilidades associadas a todas as áreas vizinhas à área l, e cuja variância é proporcional ao número de vizinhos da referida área.

O modelo ICAR definido em (2) têm sido extensivamente utilizado na literatura para a análise de dados de área. Na próxima seção mostramos como esse modelo pode ser empregado no contexto de análise de sobrevivência.

2.2 O modelo de fragilidade

Para a construção da função de verossimilhança assumiremos que os tempos de falha estão sujeitos a censura à direita, isto é, quando uma censura é observada, sabemos que o tempo até a falha está à direita do tempo observado. Também assumiremos aqui que o mecanismo gerador das censuras é não informativo, ou seja, que os tempos de falha e censura são independentes. A censura à direita é o tipo de censura que ocorre com mais frequência na prática. Além disso, a grande maioria dos modelos de sobrevivência encontrados na literatura são baseados na suposição de que a censura é não informativa.

Seja $Y = \min(T, C)$ o tempo de sobrevivência observável, em que T e C são variáveis aleatórias que representam, respectivamente, o tempo até a ocorrência do evento de interesse (tempo de falha) e o tempo até a ocorrência de uma censura. Denote por $\delta = I\{Y = T\}$ a função indicadora de falha ou censura, e seja $\mathbf{x}$ um vector de variáveis exploratórias de dimensão $1 \times p$. Considere agora um mapa com L áreas, e defina n_l como sendo o número de elementos pertencentes à l-ésima área. Finalmente, denote por $D = \{(y_{il}, \delta_{il}, \mathbf{x}_{il}, \omega_l) : i = 1, \ldots, n_l; l = 1, \ldots, L\}$ o conjunto de dados observados e não observados (fragilidades). Então, a função de verossimilhança aumentada (isto é, condicional nos dados observados e não observados) associada a um modelo de fragilidade compartilhada pode ser expressa da seguinte forma:

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}|D) = \prod_{l=1}^{L}\prod_{i=1}^{n_l} \left[h_0(y_{li}|\boldsymbol{\psi})\exp(\mathbf{x}_{li}\boldsymbol{\beta}+\omega_l)\right]^{\delta_{li}} \exp\left\{-H_0(y_{li}|\boldsymbol{\psi})\exp(\mathbf{x}_{li}\boldsymbol{\beta}+\omega_l)\right\}, \quad (3)$$

em que $\boldsymbol{\psi}$ é um vetor de parâmetros associado à função taxa de falha de base $h(\cdot|\boldsymbol{\psi})$, e $H(t|\boldsymbol{\psi}) = \displaystyle\int_0^t h(u|\boldsymbol{\psi})du$ é a função taxa de falha acumulada de base.

Conforme pode ser observado através da função de verossimilhança dada em (3), do modelo de fragilidade compartilhada é especificado de maneira hierárquica. Com o objetivo de acomodar a estrutura de dependência espacial dos dados de sobrevivência, assumiremos aqui que cada componente do vetor $\boldsymbol{\omega}$ segue a distribuição dada em (2). A distribuição Weibull será considerada para modelar a distribuição de base $h_0(\cdot|\boldsymbol{\psi})$. O modelo Weibull é um modelo bi-paramétrico bastante simples e flexível, capaz de acomodar funções taxas de falha monótonas crescentes, decrescentes e constantes. Por esses motivos, este modelo tem sido largamente utilizado para a modelagem de dados de sobrevivência na literatura Ibrahim, Chen e Sinha (2001).

As funções taxa de falha e taxa de falha acumulada associadas ao modelo Weibull são, respectivamente, dadas por:

$$h_0(t|\boldsymbol{\psi}) = \alpha\gamma t^{\alpha-1} \text{ e } H_0(t|\boldsymbol{\psi}) = \gamma t^{\alpha}, \quad (4)$$

com $\boldsymbol{\psi} = (\alpha, \gamma)$, em que $\alpha > 0$ e $\gamma > 0$ são, respectivamente, os parâmetros de forma e escala da distribuição Weibull. Funções taxa de falha decrescentes são obtidas quando $0 < \alpha < 1$. Se $\alpha = 1$, então a função taxa de falha é constante, e o modelo exponencial surge como um caso particular da distribuição Weibull. Para os demais valores de α a função taxa de falha é crescente, apresentando uma forma côncava para $1 < \alpha < 2$, linear para $\alpha = 2$, e convexa para $\alpha > 2$.

Na próxima seção discutiremos como inferências podem ser realizadas com base no modelo estudado.

2.3 Inferência Bayesiana

Na abordagem Bayesiana a especificação completa do modelo requer a elicitação de distribuições *a priori* para os parâmetros do modelo. Com esse objetivo, assumiremos as seguintes distribuições *a priori* para os parâmetros: $\alpha \sim \Gamma(a_\alpha, b_\alpha)$, $\phi \sim \Gamma(a_\phi, b_\phi)$e $\beta_j \overset{iid}{\sim} N(0, \sigma^2_\beta)$, $j = 0, \ldots, k$. A distribuição gama, por ser rica em formas e possuir suporte nos números reais positivos, é uma escolha bastante comum tanto para o parâmetro de forma α da distribuição Weibull, quanto para o parâmetro de precisão ϕ do modelo ICAR. Finalmente, a distribuição normal tem sido amplamente utilizada para representar a informação subjetiva associada aos coeficientes da regressão.

Esquematicamente, nosso modelo completo pode ser especificado da seguinte

forma:

$$\begin{array}{l}\beta_j \sim N(\mu_\beta, \sigma^2_\beta), j = 0, \ldots, k \\ \alpha \sim \Gamma(a_\alpha, b_\alpha) \\ \phi \sim \Gamma(a_\phi, b_\phi) \\ (\omega_l | \boldsymbol{\omega}_{(-l)}) \sim N(\bar{\omega}_l; 1/\phi v_{l.}), l = 1, \ldots L \\ (T_{li} | \alpha, \boldsymbol{\beta}, \omega_l) \sim Weibull(\alpha, \gamma_{li})\end{array} , \quad (5)$$

em que $\gamma_{li} = \exp(\beta_0 + \beta_1 x_{li1} + \ldots + \beta_k x_{lik})$, $l = 1, \ldots, L$ e $i = 1, \ldots n_l$.

No paradigma Bayesiano inferências sobre as quantidades de interesse, no nosso caso $(\alpha, \boldsymbol{\beta}, \phi, \boldsymbol{\omega})$, devem ser feitas com base na sua respectiva distribuição conjunta *a posteriori*. Infelizmente, tal distribuição *a posteriori* não pode ser obtida analiticamente para o nosso modelo. Uma maneira de contornarmos este problema consiste em utilizarmos métodos computacionais para aproximarmos distribuições *a posteriori* de interesse.

Os chamados métodos MCMC (Monte Carlo via Cadeias de Markov) têm sido historicamente utilizados para aproximarmos numericamente a distribuição *a posteriori* de interesse. Neste contexto, o algoritmo conhecido como amostrador de Gibbs é muito provavelmente o algoritmo mais popular da literatura. O amostrador de Gibbs envolve a obtenção das chamadas distribuições condicionais completas, que correspondem às distribuições de uma quantidade de interesse, condicional nos dados observados e nas demais quantidades de interesse. Pode ser mostrado que amostras da distribuição conjunta *a posteriori* podem ser obtidas amostrando-se sucessivamente das distribuições condicionais completas. Uma discussão mais aprofundada sobre métodos MCMC, bem como sobre o amostrador de Gibbs, estão fora do escopo deste trabalho. Para maiores detalhes sobre o assunto, recomendamos ao leitor interessado o livro dos autores Gamerman e Lopes (2006).

Tanto o cálculo das distribuições condicionais completas, quanto a implementação do amostrador de Gibbs, podem se tornar tarefas bastante árduas, principalmente para aqueles analistas de dados menos experientes em cálculos matemáticos e programação. Felizmente, existem atualmente *softwares* que permitem ao analista especificar apenas a estrutura hierárquica do modelo, conforme apresentado em (5).

Neste trabalho ilustramos como o *software* `R` em cojunto com o *software* `OpenBUGS` (LUNN et al., 2009), através do pacote `rbugs` (YAN; PRATES, 2013), pode ser utilizado para o ajuste do modelo de fragilidade espacial. Para o ajuste do modelo o `OpenBUGS` requer um arquivo de modelo como apresentado a seguir:

```
model{
        # Verosimilhanca
        for (i in 1:N){
                t[i] ~ dweib(alpha, mu[i])C(c[i],)
                log(gamma[i]) <- inprod(X[i,], beta[]) + w[area[i]]
        }
```

```
        # Distribuicao para os efeitos aleatorios (ICAR)
        w[1:L] ~ car.normal(adj[], weights[], n[], tau)

        # Pesos para matrix de adjacencia
        for(k in 1:lengthAdj){  weights[k] <- 1 }

        # Distribuicoes a priori
        for(k in 1:p){ beta[k] ~ dnorm(0.0, 0.001) }
        alpha ~ dgamma(0.001, 0.001)
        phi ~ dgamma(0.001, 0.001)
}
```

Como podemos observar a partir do código apresentado acima, a especificação do modelo de fragilidade ICAR é bastante simples, e de bastante útil para aqueles analistas de dados mais preocupados com a aplicação do que com aspectos matemáticos/computacionais. No Apêndice apresentamos o *script* completo para rodar o `R` em conjunto com o `OpenBUGS`.

2.4 Seleção de modelos

O Critério de Informação do Desvio (DIC) (SPIEGELHALTER et al., 2002) é consolidado na literatura como um alternativa Bayesiana para medida de qualidade de ajuste, e portanto, seleção de modelos. O DIC é obtido como uma gernalização do Critério de Informação Akaike

Assim o DIC é dado por

$$\text{DIC} = \overline{\text{K}}(\theta) + \rho_{\text{K}} = 2\overline{\text{K}}(\theta) - \text{K}(\tilde{\theta}),$$

em que $\theta = (\alpha, \beta, \phi)$ são os parâmetros do modelo, $\tilde{\theta} = \text{E}[\theta|D]$, $\overline{\text{K}}(\theta)$ é a média a posteriori do desvio e ρ_{K} é a medida efetiva do número de parâmetros no modelo.

O número de parâmetros no modelo, ρ_{K}, é dado por

$$\rho_{\text{K}} = \overline{\text{K}}(\theta) - \text{K}(\tilde{\theta}),$$

com

$$\overline{\text{K}}(\theta) = -2\text{E}[\log f(D|\theta)|D].$$

A resolução da integral $\overline{\text{K}}(\theta)$ não é simples, uma possível solução para seu cálculo é a utilização da amostra *a posteiori* obtida no MCMC de $\{\theta_1, \ldots, \theta_M\}$. Dessa forma podemos aproximar o DIC inicialmente calculando a média a posteriori dos desvios

$$\overline{\text{K}} = -2\frac{1}{M}\sum_{m=1}^{M} \log f(D|\theta_m)$$

e em seguinda

$$\widehat{\text{DIC}} = 2\overline{\text{K}} - \text{K}(\tilde{\theta}).$$

Dessa forma, menor o valor da estatística melhor a qualidade de ajuste do modelo.

3 Aplicação

O câncer é uma doença que afeta toda humanidade sendo a maior causa de de mortes pelo mundo com uma quantidade estimada de 8 milhões de mortes somente em 2012 (STEWART; WILD, 2014). Pesquisas mostram que países com melhor qualidade de vida são aqueles acometidos com uma maior incidência da doença, porém possui uma maior capacidade de diagnosticar, tratar e curar seus pacientes (STEWART; WILD, 2014).

O programa de Epidemiologia Vigilância e Resultados Finais do Instituto Nacional do Câncer Americano ⟨https://seer.cancer.gov/⟩ oferece, de forma gratuita, dados sobre a incidência de diversos tipos de câncer para Estados Unidos da América. Nesse trabalho utilizamos uma amostra aleatorizada de 1500 indivíduos com o tempo até a morte de pacientes que desenvolveram câncer de bexiga para o estado de Connecticut. Além da localização de residência de cada paciente, outras variáveis sócio-demográficas disponíveis são: sexo (1=masculino; 0=feminino); raça (1=branca; 0=outras) e idade (em anos).

Utilizando a metodologia apresentada na Seção 2, foram ajustados dois modelos: 1) modelo sem o termo de efeito aleatório (modelo de taxas proporcionais Weibull); 2) modelo de fragilidade Weibull-ICAR. Para cada um dos modelos foi gerada uma cadeia única de 600000 iterações, as quais as primeiras 100000 iterações foram descartadas para o aquecimento da cadeia. Para diminuir autocorrelação das amostras geradas, foi utilizado um *lag* de 100, resultando em uma amostra *a posteriori* de 5000 observações. Após a obtenção da amostra *a posteriori* para as quantidades de interesse, o método de Geweke (1992) foi utilizado para verificar a convergência da cadeia.

Distribuições *a priori* pouco informativas foram escolhidas para $\alpha \sim \Gamma(10^{-3}, 10^{-3})$, ou seja, $E(\alpha) = 1$ e $V(\alpha) = 1000$, $\beta_j \sim N(0, 1000)$ para $j = 0, \ldots, 3$ e para o hiperparâmetro de fragilidade ICAR, $\phi \sim \Gamma(10^{-3}, 10^{-3})$. O modelo sem fragilidade obteve o DIC=$13800,00$ enquanto o modelo de fragilidade espacial obteve um DIC=$13740,00$. Assim, como seu DIC foi maior o modelo sem fragilidade mostrou-se menos adequado para análise dos dados de câncer de bexiga no estado de Connecticut, o que atesta a importância e necessidade da incorporação de efeitos espaciais na modelagem deste tipo de dados.

Assim sendo, apresentaremos os resultados apenas para o modelo com fragilidade espacial. A Tabela 1 apresenta a média *a posteriori* das estimativas dos parâmetros assim como seus intervalos de credibilidade (IC) de 95%. Temos que o parâmetro forma da distribuição Weibull é significamente menor do que 1, o que implica em uma taxa de falha decrescente, ou seja, a medida que o tempo passa a chance de morte pelo câncer de bexiga se torna menor. A covariável sexo apresentou um valor positivo, implicando que homens tem maior risco de morte pelo câncer de bexiga do que mulheres. Esse fato é validado na literatura por autores como Parkin et al. (2005) e Ferlay et al. (2015). Enquanto isso raça branca e idades

maiores refletem um menor risco proporcional no tempo até a morte.

Tabela 1 - Estimativas pontual (média) e intervalo de credibilidade de 95% das amostras das distribuições a posteriori dos parâmetross do modelo de fragilidade ICAR

Parâmetros	Média	95% CI
α	0,72	(0,69;0,75)
intercepto	-2,61	(-2,85;-2,37)
sexo	0,39	(0,26;0,52)
raça	-0,32	(-0,51;-0,14)
idade	-0,11	(-0,17;-0,06)
ϕ	120,83	(0,61;480,80)

A Figura 1 apresenta um padrão de distribuição das taxas medianas de risco proporcional bem estruturado, com aparições de cluster entre regiões. Dessa forma, podemos ver que as regiões mais a sudoeste (perto a Nova Iorque) possuem um menor risco em relação as demais, servindo portanto como refêrencia para as demais regiões do estado. Além disso, o risco cresce de oeste para leste.

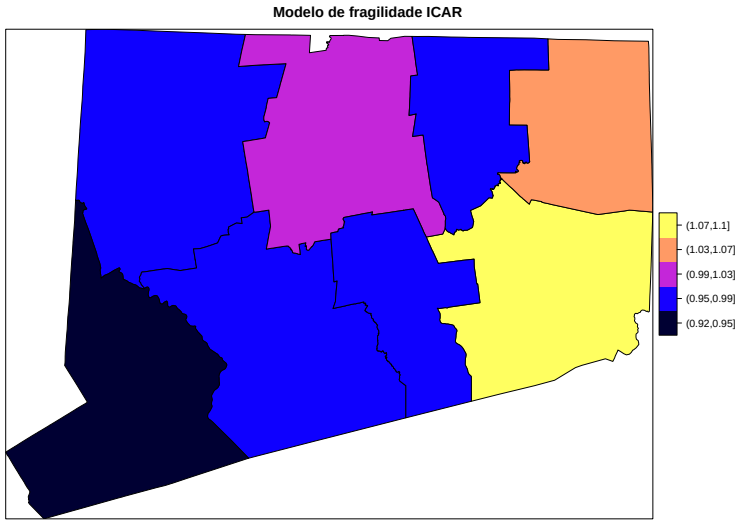

Figura 1 - Mapa de fragilidades das medianas a posteriori para o modelo de fragilidade ICAR.

4 Conclusões

Nesse trabalho apresentamos como utilizar o *software* `R` em conjunto com o *software* `OpenBUGS` para realizar inferência em modelos de sobreviência espacial através do pacote `rbugs`. A junção desses softwares estatísticos torna acessível a uma ampla comunidade essa família de modelos que tem sido amplamente estudada obtendo resultados significativos em um melhor entendimento do tempo até a morte de paciêntes de cancer (BANERJEE; CARLIN; GELFAND, 2003; OSEI, 2010; LI; HANSON; ZHANG, 2015), entre outros.

Para o estudo de câncer de bexiga no estado de Connecticut o modelo de fragilidade espacial apresentou melhor resultado do que o modelo base sem fragilidade. A utilização da fragilidade espacial permite ao pesquisador visualizar os efeitos das fragilidades espacialmente e interpretá-los. Além disso, as covariáveis utilizadas no modelo foram significativas com homens tendo um maior risco em relação ao sexo feminino enquanto aumento na idade e a raça branca apresentaram uma redução risco.

Agradecimentos

Os autores agradecem os revisores e editores pelos comentários e sugestões que possibilitaram uma melhoria na qualidade do artigo.

MÁQUINA, C. M.; DEMARQUI, F. N.; PRATES, M. O; SANTOS, N. T. Analysis of spatially correlated survival data. *Rev. Bras. Biom.,* Lavras, v.37, n.4, p.419-434, 2019.

- *ABSTRACT: A survival frailty model for spatially correlated lifetime data is discussed. By considering a completely parametric Bayesian approach, the intrinsically conditional autoregressive (ICAR) model is assumed for the frailty distribution, whereas the base distribution of lifetimes is modeled by the Weibull distribution. A Bayesian model selection criterion is discussed. A detailed description of how to use the* `R` *software along with* `OpenBUGS` *software to carry out inferences on the survival frailty model is provided. The analysis of a real data set involving patients diagnosed with bladder cancer in the state of Connecticut, USA, is presented.*

- *KEYWORDS: Spatial frailty; ICAR model; Bayesian inference; OpenBUGS.*

Referências

BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. *Hierarchical modeling and analysis for spatial data*. London: Chapman & Hall/CRC, 2003.

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, v. 34, p. 187–220, 1972. With discussion.

CRESSIE, N. *Statistics for spatial data*. New York: John Wiley, 1993.

DUCHATEAU, L.; JANSSEN, P. *The frailty model*. 1.ed. New York: Springer.

FERLAY, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, v. 136, n. 5, p.E359–E386, 2015.

GAMERMAN, D.; LOPES, H. F. *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference*. London: Chapman & Hall/CRC, 2006.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*. London: Clarendon Press, v. 4, p. 641–649, 1992.

HANAGAL, D. *Modeling survival data using frailty models*. 2011.

HOUGAARD, P. *Analysis of multivariate survival data*. New York: Springer, 2000.

IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. *Bayesian survival analysis*. New York: Springer, 2001.

LI, L.; HANSON, T.; ZHANG, J. Spatial extended hazard model with application to prostate cancer survival. *Biometrics*, v. 71, n. 2, p. 313–322, 2015.

LUNN, D. et al. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, v. 28, n. 25, p. 3049–3067, 2009.

OSEI, P. P. *Statistical methods for disease mapping*. Tese (Doutorado) — African Institute for Mathematical Sciences, 2010.

PARKIN, D. M. et al. Global cancer statistics, 2002. *CA: A Cancer Journal for Clinicians*, v. 55, n. 2, p. 74–108, 2005.

R CORE TEAM. *R: A language and environment for statistical computing*. Vienna, Austria, 2018. Disponível em: ⟨https://www.R-project.org/⟩.

SPIEGELHALTER, D. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, v. 64, p. 583–639, 2002.

STEWART, B. W.; WILD, C. P. *World Cancer Report 2014*. Lyon, France: International Agency for Research on Cancer, 2014.

THERNEAU, T. M.; GRAMBSCH, P. *Modeling survival data: Extending the Cox model*. v.48, 2000.

WIENKE, A. *Frailty models in survival analysis*. 2010.

YAN, J.; PRATES, M. *rbugs: Fusing R and OpenBugs and Beyond*. [S.l.], 2013. R package version 0.5-9. Disponível em: ⟨https://CRAN.R-project.org/package=rbugs⟩.

Recebido em 16.07.2018.

Aprovado após revisão em 21.01.2019.

Apêndice

A seguir, apresentamos o *script* do `R` para ajustar o modelo de fragilidade espacial utilizando o *software* `OpenBUGS` através do pacote `rbugs`. Estamos assumindo aqui que o `R` está sendo executado em um computador com sistema operacional `Windows`, e que tanto o banco de dados quanto o arquivo.shp estão estão no diretório de trabalho corrente do `R`.

Como os dados analisados neste trabalho não são de domínio público, o *script* foi adaptado para dados fictícios. Especificamente, o `data.frame` dados contém as seguintes variáveis:

- t: tempo de falha;
- delta: variável indicadora de falha ou censura;
- x_1, x_3, x_3: variáveis explicativas;
- area: variável indicadora de área;

```
# limpando a memoria do R:
rm(list=ls(all=TRUE))

# carregando os pacotes necessarios:
library(rbugs)
library(maptools)
library(spdep)

# obtendo o diretorio de trabalho corrente:
wd <- getwd()

# lendo o banco de dados:
dados <- read.table("dados.txt", header=TRUE)

# excluindo observacoes com t=0
dados <- subset(dados, dados$t !=0)

# organizando/preparando os dados:
t <- dados$t
c <- rep(0,N)
for(i in 1:N){
  if(dados$e[i]==0){
    c[i]=dados$t[i]
    t[i]=NA
  }
}
```

```
area <- dados$area
N <- length(dados$t)
X <-  with(dados, cbind(1, x1, x2, x3))
p <- ncol(X)

# lendo o arquivo.shp:
mapa <- readShapePoly("arquivo.shp", IDvar=NULL,
        proj4string=CRS(as.character(NA)), verbose=FALSE,
        repair=FALSE, force_ring=FALSE, delete_null_obj=FALSE,
        retrieve_ABS_null=FALSE)

# desenhando o mapa:
nomes <- names(mapa[7])
plot(mapa, border="blue", axes=TRUE, las=1)
text(coordinates(mapa), labels=row.names(mapa), cex=0.6)

# obtendo a matriz de vizinhancas:
matriz <- poly2nb(mapa)
W <- nb2mat(matriz, glist=NULL, style="B", zero.policy=NULL)

## Gera a estrutura de vizinhanca do mapa:
vizinhanca <- poly2nb(mapa)

## total de areas:
L <- length(vizinhanca)

# Encontrando a posicao de ilhas no mapa:
position <- which(!is.na(match(vizinhanca, 0)))

## numero de vizinhos de cada area
n <- sapply(vizinhanca, length)
n[position] <- 0

## lista de adjacencias para cada area
adj <- unlist(vizinhanca)
adj <- adj[which(adj != 0)]

## Soma do numero vizinhancas do mapa
lengthAdj <- length(adj)

# Lendo/visualizando o modelo do OpenBUGS:
modelo <- file.path(wd, "modelo.R")
file.show(modelo)
```

```
# lista com os dados que dever ser passados para o OpenBUGS:
data <- list(t=t, c=c, X=X, N=N, p=p, L=L, areas=areas,
             adj=adj, lengthAdj=lengthAdj, n=n)

# chute inicial:
inits <- function(){list(alpha = rexp(1,1), tau = rexp(1,1),
                         beta = rnorm(p), w = rep(0,L))}

# parametros a serem salvos:
parameters <- c("alpha", "tau", "beta", "w")

# configuracao do MCMC:
n.chains <- 1
n.burnin <- 100000
n.thin <- 100
n.post <- 5000
n.iter <- n.burnin + n.thin*n.post

# chamando o OpenBUGS do R usando o pacote rbugs:
ajuste <- rbugs(data, inits, parameters, modelo, n.chains=nchains,
          n.iter=n.iter, n.burnin=n.burnin, n.thin = lag,
          bugs= "C:/Program Files (x86)/OpenBUGS/OpenBUGS321/OpenBugs.exe",
          bugsWorkingDir = wd, dic=TRUE, verbose=TRUE)

# salvando tudo que esta na memoria do R
save.image(file="modelo.RData")

# extraindo a amostra a posteriori:
amostra <- rbugs2coda(ajuste)

# examinando a amostra a posteriori:
plot(amostra, ask=TRUE)
summary(amostra)

# testes de convergencia:
HPDinterval(amostra)
heidel.diag(amostra)
geweke.diag(amostra)

# gera os nomes das colunas dos efeitos aleatorios
nome.w <- paste("w[",1:L,"]",sep="")

# calcula os efeitos aleatorios na escala original
z <- exp(as.matrix(amostra[,nome.w]))
```

```
# recupera os valoria medianos da posteriori para cada regiao
z.median <- apply(z,2,median)

# cria o grid com o alcance dos feitos aleatorios e depois os cortes
grid <- round(seq(min(z.median)-0.02, max(z.median)+0.02,length.out=6),2)
z <- cut(z.median, breaks=grid)

# grava os dados no objeto do mapa
mapa@data$z <- z

# plota os mapa
spplot(mapa,"z", main="Modelo de fragilidade ICAR",names=mapa@data$NAME10)
```