

Sobrevivência relativa Bayesiana com efeitos espaciais

Uma aplicação utilizando R/RStan

Jony Arrais Pinto Junior, Victor Hugo Soares Ney

Universidade Federal Fluminense – Graduação em Estatística

Motivação

- Principal problema de saúde pública no mundo e está entre as **quatro principais causas de morte prematura** em grande maioria dos países (BRAY et al., 2018);
- Instituto Nacional de Câncer (INCA) em 2019 projetou a incidência de **625 mil novos casos** de câncer no Brasil para cada ano do triênio 2020/2022;
- Entender melhor as características dessa doença, tais como o tempo de sobrevida, é de extrema importância para a tomada de políticas públicas eficientes.

Análise de Sobrevida

- Conjunto de métodos que permitem analisar estatisticamente dados de sobrevida;
- Variável aleatória T registra tempo até o evento de interesse;
- Evento de interesse: evento definido no início da pesquisa (morte por x doença, falha de um sistema por y motivo, recuperação total de um paciente, etc);
- Censura: ocorre quando o indivíduo chegou ao fim estudo e não se observou o evento ou quando se observa qualquer evento que não tenha a ver com o motivo em estudo (por exemplo, observou-se óbito mas não por conta da doença x) – tem-se uma **informação parcial** sobre o tempo de sobrevida do indivíduo;
- Período de *follow-up*: tempo pelo qual será feito acompanhamento dos pacientes a cada dado intervalo de tempo, definido no início da pesquisa;

Funções na Análise de Sobrevida

Função de Sobrevida

- T variável aleatória não negativa, tempo de sobrevida;
- Função de sobrevida $S(t)$: probabilidade da variável T ultrapassar o tempo t

$$S(t) = P(T > t), t \geq 0 \quad (1)$$

Função de Risco

- Taxa instantânea da ocorrência do acontecimento de interesse, dado que o evento ainda não aconteceu;
- Função de risco é inversamente proporcional à função de sobrevida;
- Função $\lambda(t)$ dada por

$$\lambda(t) = \lim_{\epsilon \rightarrow 0^+} \frac{P[(t \leq T < t + \epsilon) | T > t]}{\epsilon}, \quad t \geq 0 \quad (2)$$

Por que sobrevivência relativa?

Sobrevivência relativa

- Não necessita de informação sobre a causa de morte do indivíduo;
- Variável aleatória T representa tempo de sobrevivência;
- Sobrevivência relativa é dada por $S_R(t)$, razão entre a sobrevivência observada $S_O(t)$ e a sobrevivência esperada $S_E(t)$ de um grupo;

$$S_R(t) = \frac{S_O(t)}{S_E(t)} \quad (3)$$

- Estimativa do tempo de vida dos indivíduos caso a doença em estudo seja a única causa de morte possível (**sobrevivência líquida**);
- Permite analisar o impacto das mortes causadas pela doença em estudo em relação à mortalidade da população geral.

Por que sobrevivência relativa?

- Mortes por câncer declaradas muitas vezes de forma incorreta pelas complicações da doença;
- **Exemplo:** paciente com câncer de mama sofre metástase para o pulmão e se observa o óbito devido por pneumonia certo tempo depois. Em um estudo de câncer de mama:

Análise de sobrevivência usual → observação censurada

Análise de sobrevivência relativa → ocorrência de um evento

Modelos de regressão para sobrevivência relativa

- Proposto em Breslow e Day (1987).
- O número de mortes no estrato k em um intervalo de tempo t , assumindo riscos constantes para esse grupo, segue distribuição *Poisson*,

$$d_{kt} \sim \text{Poisson}(\mu_{kt}) , \quad (4)$$
$$\mu_{kt} = \lambda_{kt} \cdot y_{kt} ,$$

em que λ_{kt} é função de risco para o t -ésimo intervalo e y_{kt} o tempo em risco dos indivíduos durante o t -ésimo intervalo.

Modelo de regressão aditivo

- Proposta em Dickman et al. (2004);
- A função de risco para um vetor de covariáveis \mathbf{x} é modelada como

$$\lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) + \nu(\mathbf{x}), \quad (5)$$

em que $\lambda^*(\mathbf{x})$ é o risco esperado e $\nu(\mathbf{x})$ é o excesso de risco devido ao diagnóstico de câncer;

- $\lambda^*(\mathbf{x})$ é estimada de dados externos – por exemplo, dados populacionais de mortalidade.

Modelo de regressão aditivo

- O excesso de risco $\nu(\mathbf{x})$ é assumido ser função multiplicativa das covariáveis, $\nu(\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})$, (Dickman et al. (2004)), em que $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos associados com \mathbf{x} ;
- Reescreve-se a Equação 5 como

$$\lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) + \exp(\mathbf{x}\boldsymbol{\beta}) \quad (6)$$

Modelo de regressão aditivo

- Utilizando das Equações 4 e 6, podemos reescrever o modelo de regressão como:

$$\begin{aligned} d_{kt} &\sim \text{Poisson}(\mu_{kt}) , \\ \ln(\mu_{kt} - d_{kt}^*) &= \ln(y_{kt}) + \mathbf{x}\boldsymbol{\beta} , \end{aligned} \tag{7}$$

d_{kt}^* é o número esperado de mortes devido a outras causas além do câncer (estimado das tabelas populacionais de mortalidade);

- O modelo acima é um modelo linear generalizado com função *link* $\ln(\mu_{kt} - d_{kt}^*)$ e *offset* $\ln(y_{kt})$.

Modelo de regressão aditivo espacial para sobrevivência relativa

- Inclusão de um efeito aleatório ϕ_i , representando excessos de variabilidade da área i ;
- Pode-se incluir, por meio do efeito, uma estrutura de dependência espacial.
- Inclusão de um intercepto α_t para representar excesso de risco por pertencer ao t -ésimo intervalo de tempo após o diagnóstico de câncer (Cox (1972)).

Modelo Espacial para Sobrevivência relativa

$$d_{ikt} \sim \text{Poisson}(\mu_{ikt}) , \quad (8)$$

$$\ln (\mu_{ikt} - d_{ikt}^*) = \ln (y_{kt}) + \alpha_t + \mathbf{x}_{ik}\boldsymbol{\beta}_k + \phi_i , \quad (9)$$

$$\alpha_t \sim N(0, 1000) , \quad (10)$$

$$\boldsymbol{\beta}_k \sim N(0, 1000) \quad (11)$$

Sobre as distribuições *a priori* de ϕ_i

- Pode-se assumir $\phi = (\phi_1, \phi_2, \dots, \phi_N)^T$ sem uma estrutura de dependência espacial, resultando no modelo a seguir

Modelo sem estrutura de dependência espacial

$$\phi_i | \tau \sim N(0, \tau^{-1}) , \quad (12)$$

$$\tau \sim \text{Gama}(1, 1) , \quad (13)$$

em que τ representa a precisão geral.

- Pode-se utilizar modelos condicionais autorregressivos – conhecidos como modelos CAR – para representar estrutura de dependência espacial;
- Em Fairley et al. (2008), destaca-se a importância de uma estrutura de dependência espacial nesses modelos.

Modelo Intrínseco

- Proposto em Besag, York e Mollié (1991);
- Modelo CAR mais simples para representar estrutura de dependência espacial, dado por

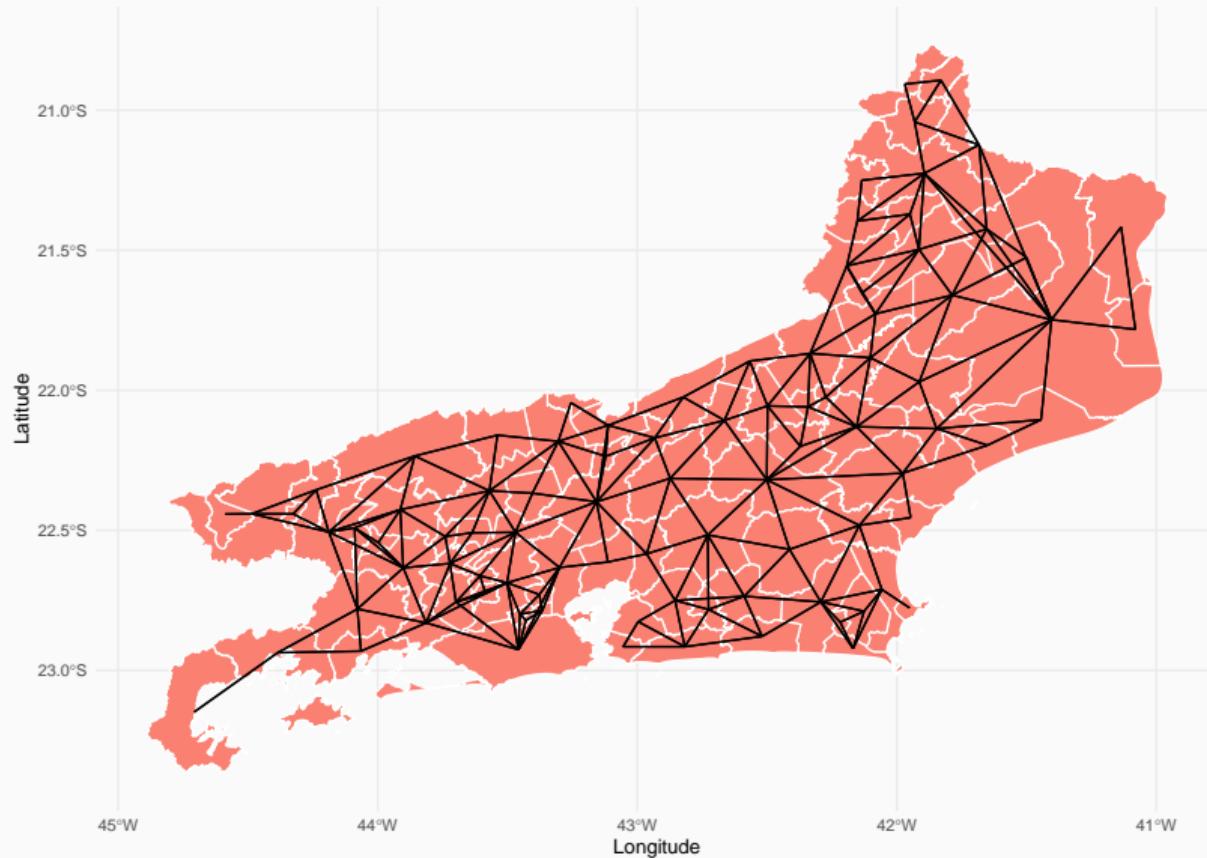
$$\phi | \tau, \mathbf{W} \sim \mathcal{N}(\mathbf{0}, [\tau \mathbf{W}]^{-1}), \quad (14)$$

$$\tau \sim \text{Gama}(1, 1), \quad (15)$$

em que em que \mathbf{W} representa uma matriz $n \times n$ de vizinhança tal que

$$w_{ji} = \begin{cases} n_i, & \text{se } j = i \\ -1, & \text{se } j \sim i \\ 0, & \text{caso contrário} \end{cases}$$

ggplot2 – Mapa de vizinhança com base em contiguidade



Modelo de Convolução (BYM)

- Proposto em Besag, York e Mollié (1991);
- Combina o modelo intrínseco com um conjunto adicional de efeitos aleatórios $\theta = (\theta_1, \dots, \theta_N)^T$ para representar efeitos heterogêneos das áreas;
- O modelo é dado por

$$\phi_i = \theta_i + \psi_i , \quad (16)$$

$$\theta_i | \tau \sim N(0, \tau^{-1}) , \quad (17)$$

$$\psi | \mathbf{W}, \tau \sim \text{IAR}(\mathbf{W}, \tau^{-1}) , \quad (18)$$

$$\tau \sim \text{Gama}(1, 1) , \quad (19)$$

em que $\psi = (\psi_1, \psi_2, \dots, \psi_N)^T$ segue o modelo intrínseco.

Modelo de Cressie

- Proposto em Stern e Cressie (2000);
- Inclui um parâmetro de correlação espacial, $0 \leq \rho \leq 1$, com $\rho = 0$ sendo um cenário de independência espacial e $\rho = 1$ completa dependência espacial (modelo intrínseco);
- O modelo é dado por

$$\phi | \mathbf{D}, \mathbf{A}, \tau, \rho \sim \mathcal{N}(\mathbf{0}, [\tau(\mathbf{D} - \rho\mathbf{A})]^{-1}) , \quad (20)$$

$$\tau \sim \text{Gama}(1, 1) , \quad (21)$$

$$\rho \sim \text{Beta}(1, 1) , \quad (22)$$

em que \mathbf{D} é uma matriz diagonal e \mathbf{A} é uma matriz de adjacência tal que

$$d_{ji} = \begin{cases} n_i, & \text{se } j = i \\ 0, & \text{caso contrário} \end{cases} , \quad a_{ji} = \begin{cases} 1, & \text{se } j \sim i \\ 0, & \text{caso contrário} \end{cases}$$

Modelo de Leroux

- Proposto em Leroux, Lei e Breslow (2000);
- Utiliza a mesma matriz de vizinhança \mathbf{W} e inclui o mesmo parâmetro ρ de correlação espacial, sendo $\rho = 1$ o modelo intrínseco.
- O modelo é dado por

$$\phi | \mathbf{W}, \tau, \rho \sim \mathcal{N} \left(\mathbf{0}, \frac{[\rho \mathbf{W} + (1 - \rho) \mathbf{I}_n]^{-1}}{\tau} \right), \quad (23)$$

$$\tau \sim \text{Gama}(1, 1), \quad (24)$$

$$\rho \sim \text{Beta}(1, 1), \quad (25)$$

em que \mathbf{I}_n é uma matriz identidade $n \times n$.

Modelo BYM2

- Proposto em Riebler et al. (2016);
- Dois efeitos aleatórios assim como no modelo BYM e um parâmetro ρ para explicar o quanto da variância ocorre devido aos efeitos espacialmente estruturados (ψ) e o quanto ocorre devido ao efeitos heterogêneos independentes (θ).
- O modelo é dado por

$$\phi = \frac{1}{\sqrt{\tau}} \left[\left(\sqrt{\frac{\rho}{s}} \right) \psi^* + \left(\sqrt{1 - \rho} \right) \theta^* \right] , \quad (26)$$

$$\theta^* \sim N(0, 1) , \quad (27)$$

$$\psi^* | \mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1}) , \quad (28)$$

$$\tau \sim \text{Gama}(1, 1) , \quad (29)$$

$$\rho \sim \text{Beta}(0.5, 0.5) \quad (30)$$

Aplicação no R

Cenário

- Indivíduos com residência permanente no estado do Rio de Janeiro diagnosticados com câncer de pulmão entre 2010 e 2019;
- Período de *follow-up* de 3 anos;
- Covariáveis: sexo e faixa etária (25 até 54 anos, 55 até 59 anos, 60 até 64 anos, 65 até 69 anos e 70 anos ou mais).

Proposta

- Utilizar o R para ajustar diferentes modelos Bayesianos para os dados sob a óptica da sobrevivência relativa e comparar a qualidade os ajustes;
- Todos ajustes serão feitos com 4 cadeias e 8.000 iterações, 2.000 de *warm-up* e 3 de *thin*, resultando em 4 cadeias de 2.000 iterações cada.

- **Dados populacionais de câncer:** Registros Hospitalares de Câncer (RHC), disponibilizado pelo INCA;
- **Dados populacionais de mortalidade:** Sistema de Informações sobre Mortalidade (SIM), disponibilizado por meio do TabNet;
- **Dados de população residente:** Estudo de Estimativas Populacionais por Município, Sexo e Idade, disponibilizado por meio do TabNet;

Pacotes Utilizados no R

Manipulação de Dados

tidyverse

lubridate

splitstackshape

reshape2

Leitura de dados

readxl

read.dbc

foreign

Criação de Gráficos

ggplot2

tmap

colorspace

Dados espaciais

sdpd

Análise de Sobrevida

Epi

popEpi

Modelos Bayesianos

rstan

INLA

loo

Por que R/RStan?

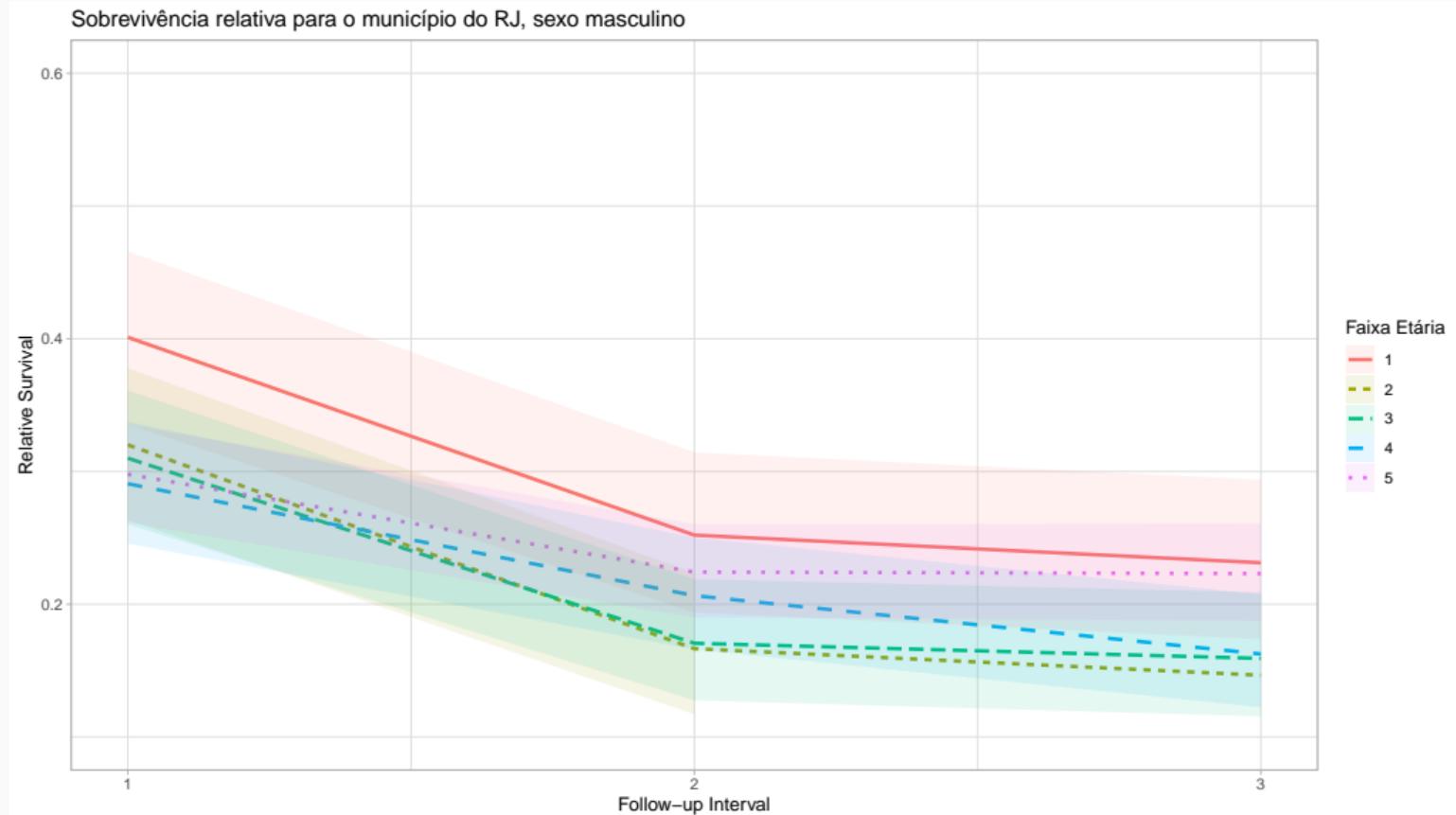
Stan

- Plataforma para modelagem e computação estatística (CARPENTER et al., 2017);
- Amostrador MCMC com dinâmica Hamiltoniana (*Hamiltonian Monte Carlo – HMC*);
- Amostrador mais eficiente e com convergência mais rápida para modelos complexos.

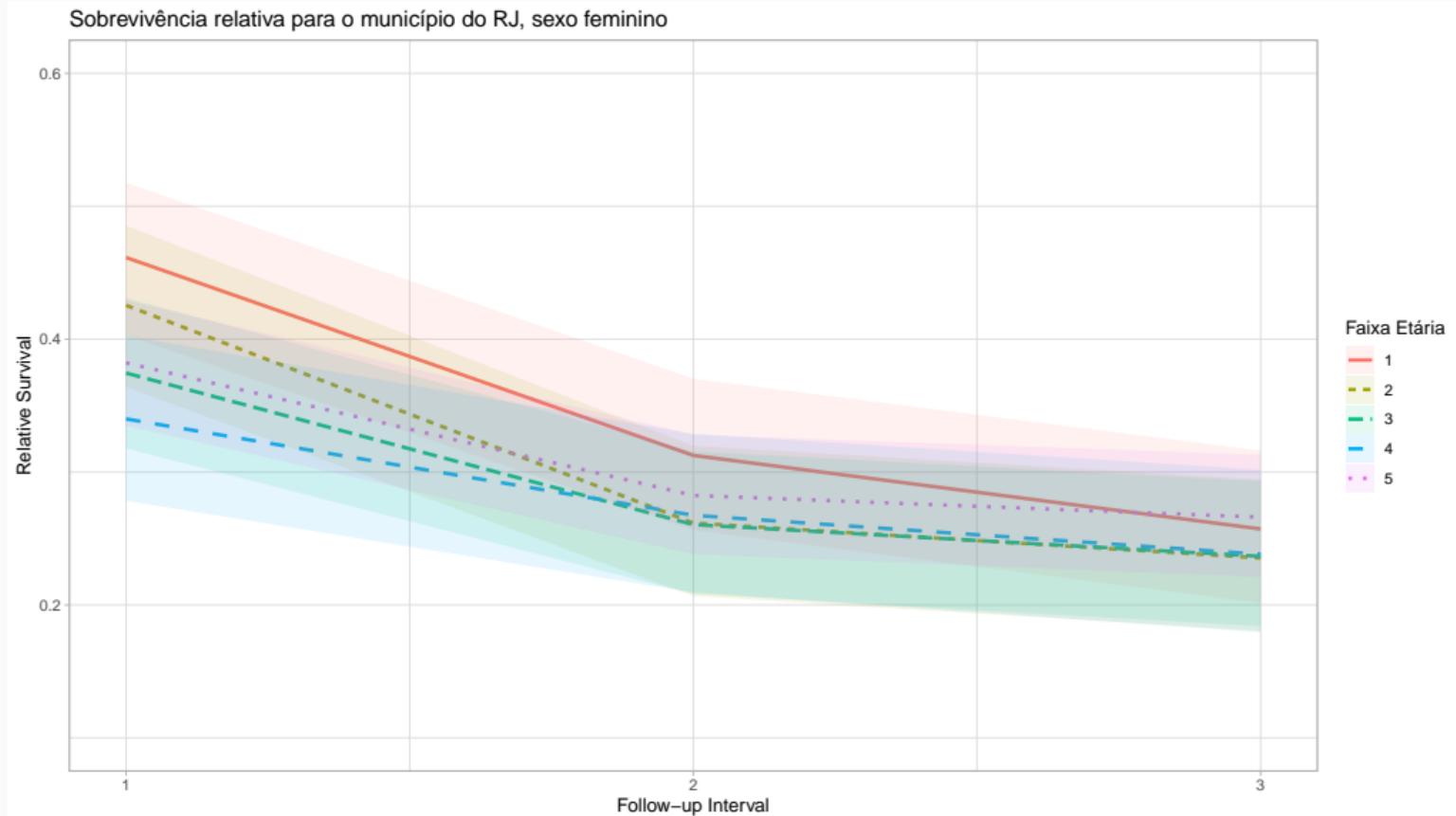
R

- Alta manipulação dos dados;
- Possui pacotes para análise de sobrevivência relativa;
- Integração com o Stan por meio do pacote RStan (Stan Development Team, 2018).

ggplot2 – Sobrevivência relativa



ggplot2 – Sobrevivência relativa



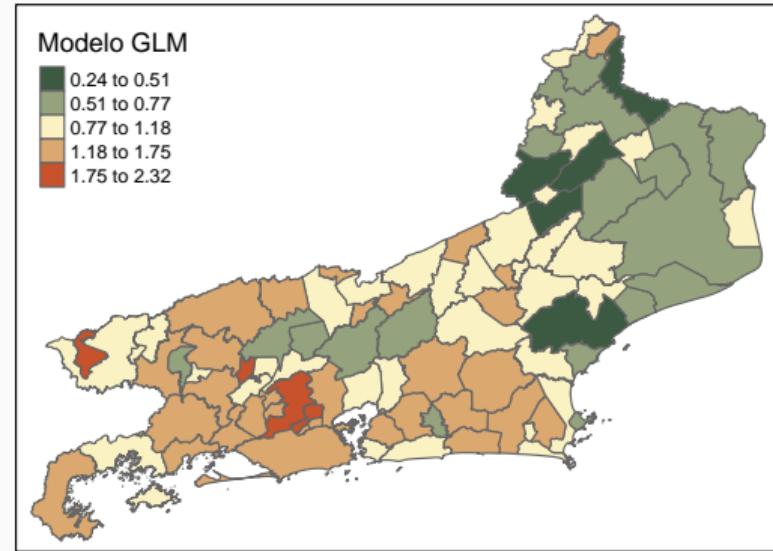
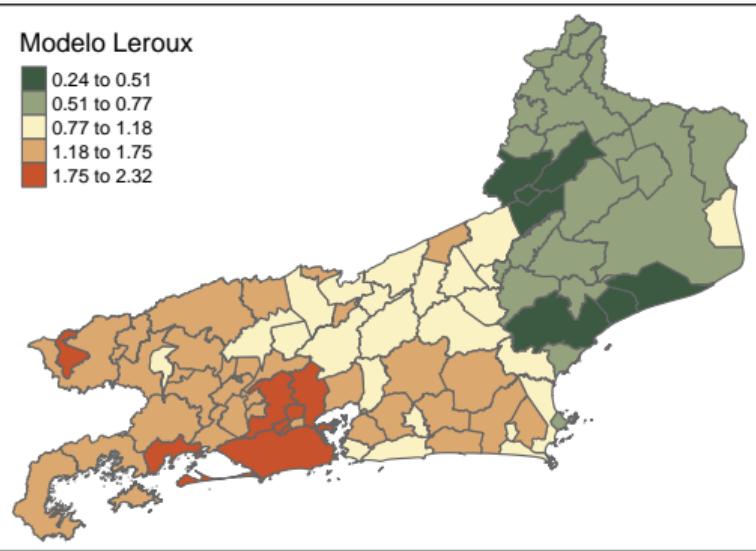
Comparação dos ajustes utilizando o pacote *loo*

Pacote *loo*

- É possível utilizar o pacote *loo* para a comparação de modelos Bayesianos;
- WAIC (*Widely Applicable Information Criterion*) é uma das métricas possíveis para calcular utilizando *loo*;
- Menor WAIC indica melhor qualidade do modelo.

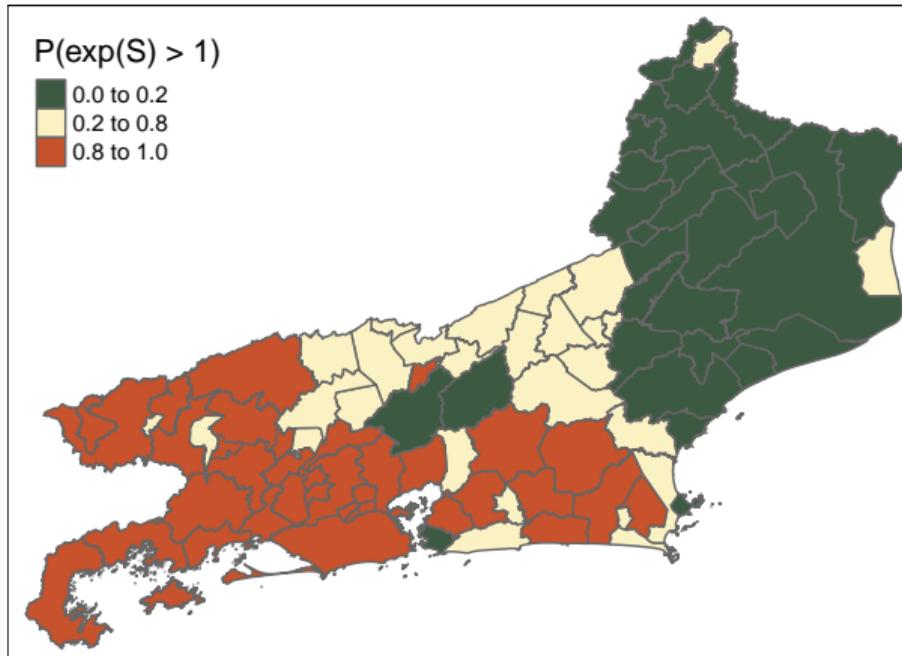
Modelo	WAIC
Intrínseco	3.648, 31
Leroux	3.648, 63
BYM2	3.648, 67
Cressie	3.649, 36
BYM	3.652, 46
GLM	3.662, 76

Comparação do Modelo Leroux com o GLM (sem estrutura de dependência espacial)



Probabilidade posterior de excesso de risco na área i , $P(\exp(\phi_i) > 1)$

Modelo Leroux



Estimativas dos Parâmetros – Leroux

<i>Estimativas dos parâmetros</i>	<i>Média e I.C. 95%</i>	<i>Estimativas dos parâmetros</i>	<i>Média e I.C. 95%</i>
55 até 59 anos (β_1)	0.06 (-0.45, 0.16)	α_1	-0.51 (-0.62, -0.40)
60 até 64 anos (β_2)	0.10 (0.00, 0.20)	α_2	-1.62 (-1.75, -1.48)
65 até 69 anos (β_3)	0.12 (0.02, 0.22)	α_3	-3.13 (-3.41, -2.87)
70 anos ou mais (β_4)	0.06 (-0.03, 0.16)	ρ	0.86 (0.66, 0.95)
Sexo feminino (β_5)	-0.16 (-0.21, -0.09)	τ	2.52 (1.44, 4.14)

Algumas métricas

- Propostas em Saez et al. (2012);
- Sobrevivência relativa de k -anos

$$\text{relative survival}_k = \exp \left[- \sum_{i=1}^k \exp(\alpha_i) \right] \quad (31)$$

- Excesso de risco no ano k em comparação com ano 1

$$\text{rel.h}_k = \exp(\alpha_k - \alpha_1), \quad k = 1, 2, 3 \quad (32)$$

- Excesso de risco associado com as covariáveis – indicador de pertencer à faixa etária i , $\exp(\beta_i)$; indicador de pertencer ao sexo feminino, $\exp(\beta_{\text{fem}})$

Algumas métricas – Leroux

<i>Estimativas de excesso de risco</i>	<i>Média e I.C. 95%</i>
25 até 54 anos	
55 até 59 anos	1.06 (0.96, 1.17)
60 até 64 anos	1.11 (1, 1.22)
65 até 69 anos	1.13 (1.02, 1.25)
70 anos ou mais	1.07 (0.97, 1.17)
Sexo feminino	0.86 (0.81, 0.91)

<i>Sobrevida Relativa (k-anos)</i>	<i>Média e I.C. 95%</i>
1 ano	0.55 (0.51, 0.58)
2 anos	0.45 (0.41, 0.49)
3 anos	0.43 (0.39, 0.48)
<i>Excesso de risco no ano k</i>	
2 anos	0.33 (0.32, 0.34)
3 anos	0.073 (0.06, 0.09)

Bibliografia

-  BESAG, J.; YORK, J.; MOLLIÉ, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, Springer, v. 43, n. 1, p. 1–20, 1991.
-  BRAY, F. et al. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 68, n. 6, 2018. Disponível em: <<https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21492>>.
-  BRESLOW, N. E.; DAY, N. E. Statistical methods in cancer research. volume ii. International agency for research on cancer Lyon, 1987.
-  CARPENTER, B. et al. Stan: A probabilistic programming language. *Journal of statistical software*, Columbia Univ., New York, NY (United States); Harvard Univ., Cambridge, MA (United States), v. 76, n. 1, 2017.

Bibliografia

-  COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, 1972. Disponível em: <<https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x>>.
-  DICKMAN, P. W. et al. Regression models for relative survival. *Statistics in Medicine*, v. 23, n. 1, 2004. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1597>>.
-  FAIRLEY, L. et al. Spatial variation in prostate cancer survival in the northern and yorkshire region of england using bayesian relative survival smoothing. *British journal of cancer*, Nature Publishing Group, v. 99, n. 11, p. 1786–1793, 2008.
-  LEROUX, B. G.; LEI, X.; BRESLOW, N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: *Statistical models in epidemiology, the environment, and clinical trials*. [S.l.]: Springer, 2000. p. 179–191.
-  RIEBLER, A. et al. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 25, n. 4, p. 1145–1165, 2016.

Bibliografia

-  SAEZ, M. et al. Spatial variability in relative survival from female breast cancer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Wiley Online Library, v. 175, n. 1, p. 107–134, 2012.
-  Stan Development Team. *RStan: the R interface to Stan*. 2018. R package version 2.17.3. Disponível em: <<http://mc-stan.org/7>>.
-  STERN, H. S.; CRESSIE, N. Posterior predictive model checks for disease mapping models. *Statistics in medicine*, Wiley Online Library, v. 19, n. 17-18, p. 2377–2397, 2000.

Agradecimentos



victor_ney@id.uff.br

[linkedin.com/in/victorsney](https://www.linkedin.com/in/victorsney)