Implementação do Método Bayesiano de Regionalização de Mapas Epidemiológicos

Marcelo Azevedo Costa – Departamento de Estatística, UFMG[1]

Sérgio Henrique R. Ribeiro – Departamento de Estatística, UFMG[2]

Resumo: *Neste trabalho focamos na metodologia de regionalização proposta por Knorr-Held (2000) para análise de dados de área. Implementamos a metodologia em linguagem de programação C/C++ orientada a objeto modificando a distribuição a priori do risco relativo e retiramos alguns passos do algoritmo MCMC com Saltos Reversíveis, buscando futuramente uma possível extensão espaço-tempo da metodologia. Comparamos os resultados do algoritmo implementado com o algoritmo original.*

Palavras-chave: saltos reversíveis MCMC, risco relativo, regionalização de mapas

Introdução: A análise estatística de dados geo-referenciados e suas metodologias tem sido alvo de crescente interesse e estudo. Inúmeras questões e situações geram dados que estão associados a algum tipo de componente espacial. Uma forma natural que encontramos os dados geo-referenciados são os chamados dados de área. Esses dados podem ser, por exemplo, o município de Belo Horizonte particionado por bairros contendo informações sobre a incidência de doenças ou mortalidade em cada bairro. Neste contexto uma das ferramentas da Estatística Espacial para o estudo de fatores de riscos desconhecidos estruturados espacialmente são os chamados métodos de regionalização.

Neste trabalho focamos na metodologia de regionalização proposta por Knorr-Held (2000) para análise de dados de área. Ele propõe uma metodologia não-paramétrica Bayesiana de regionalização utilizando o método MCMC com Saltos Reversíveis (Green ,1995). Basicamente, a metodologia particiona uma área em conglomerados que sejam homogêneos a um atributo de interesse, neste caso em relação ao risco relativo do conglomerado. Detalhes da metodologia de Knorr-Held (2000) serão abordados neste trabalho.

A motivação do trabalho aqui apresentado é a implementação do algoritmo proposto por Knorr-Held (2000) com algumas modificações. Temos como objetivo avaliar o desempenho das modificações propostas e para uma futura extensão espaço-temporal da mesma. Propomos uma modificação da distribuição a *priori* do risco relativo. Assim como também retiramos do algoritmo MCMC alguns passos que os próprios autores dizem ser desnecessários, em um primeiro momento.

Os autores do artigo também disponibilizaram um pequeno aplicativo, BDCD (*Bayesian Detection of Clusters and Discontinuities*), que executa a sua metodologia proposta. Finalmente

[1] Agradecimento ao CNPQ e FAPEMIG pelo apoio financeiro
contato: azevedo@est.ufmg.br
[2] Agradecimento ao CNPQ e FAPEMIG pelo apoio financeiro
contato: sergiohrr@gmail.com

comparamos os resultados obtidos pelo algoritmo implementado com os resultados do aplicativo BDCD.

Metodologia: O BDCD, método de detecção bayesiano de clusters e descontinuidades, particiona a área A em k conglomerados, de tal forma que cada conglomerado seja contínuo e constituído de uma ou mais das n regiões da área A, buscando a homogeneidade em relação ao risco relativo dentro do conglomerado e buscando a heterogeneidade em relação ao risco relativo entre os conglomerados. Dessa forma definimos o que chamaremos de *clusters,* um conjunto de regiões contínuas com um mesmo risco relativo.

Como resultado da metodologia teremos uma área A particionada em k *clusters*, tal que $A = \bigcup_{j=1}^{k} C_j$ e $C_j \cap C_l = \emptyset$ se $j \neq l$, onde C_j é o j-ésimo *cluster*. O número de *cluster* k é tratado como desconhecido, com $k = \{1, \dots, n\}$. O método BDCD modela y_i, número de casos observados na i-ésima região, com uma distribuição de Poisson com média $e_i h_j$, onde e_i, $i = 1, \dots, n$, é o número esperado de casos na i-ésima região e h_j, $j = 1, \dots, k$, é o risco relativo desconhecido do cluster *j,* sendo que a i-ésima região pertence a C_j. Assumimos que a resposta y_i é condicionalmente independente, dado o vetor de riscos relativos dos k *clusters*, $H_k = (h_1, \dots, h_k)$. Dessa forma a função de verossimilhança para a variável resposta $y = (y_1, \dots, y_n)$ é dada por: $L(y|H_k) = \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(e_i h_j)^{y_i}}{y_i!} \exp\{-e_i h_j\}$.

Utilizamos um banco de dados simulado de 245 regiões sem a existência de *clusters,* ou seja, com riscos relativos iguais a um em todo o mapa. Consideremos um mapa de uma área A que é particionada em *n* regiões $(r_i, \dots, r_n)$, ou seja, $A = \bigcup_{i=1}^{n} r_i$ e $r_i \cap r_j = \emptyset$ se $i \neq j$. Para cada região é necessário que tenhamos o número de casos observados (y_i), o número esperado de casos (e_i) e a matriz de adjacência.

Inicialmente, para particionar a área A definimos k regiões como sendo os centros dos *clusters*. Dessa forma especificamos nosso vetor de centróides desconhecidos $G_k = (g_1, \dots, g_k)$, onde cada centróide g_j representa o cluster C_j. Cada vetor G_k define uma única configuração da área A, ou seja, dado G_k só existe uma única partição possível de $A = \bigcup_{j=1}^{k} C_j$. Após definidos os centróides, para cada $(n - k)$ regiões restantes devemos atribuí-las a um dos *clusters* definidos pelos centróides G_k.

A metodologia BDCD trata o número k de clusters, o vetor de centróides G_k e o vetor de riscos relativos H_k como desconhecidos. As distribuições a priori para k, G_k e H_k são descritas a seguir. Modelamos k por uma distribuição a *priori* $P(k)$, $k = 1, \dots, n$, é proporcional a $(1 - c)^k$, com $c \in [0,1)$. Dessa forma, a distribuição do número de cluster dado n e c é dada por:

$$P(k|n, c) = \frac{(1 - c)^k}{\sum_{j=1}^{n} (1 - c)^j} \qquad (1)$$

Dado um número k de *clusters*, assumimos que cada vetor de centróides $G_k = (g_1, \dots, g_k)$ possui a mesma probabilidade de ocorrência, ou seja, assumimos uma distribuição a *priori* não informativa para $P(G_k)$: $P(G_k|k) = \frac{(n-k)!}{n!}$.

Para o vetor de riscos relativos $H_k = (h_1, \dots, h_k)$, Knorr-Held (2000) assume que h_j segue uma distribuição log-normal com hiperparâmetros μ e σ^2 desconhecidos. Assim, temos que $\log(h_j) \sim N(\mu, \sigma^2), j = 1, \dots, k$. Para μ assumimos uma distribuição a *priori* difusa sendo uma

distribuição uniforme em toda a reta real. Para σ^2 assumimos uma distribuição gama invertida própria de alta variância com parâmetros fixos a e b.

Chamaremos de BayesMaps a nossa implementação do algoritmo para diferenciar do aplicativo já existente, BDCD. Neste ponto, propomos uma nova distribuição a *priori* para os riscos relativos. Para o vetor de riscos relativos $H_k = (h_1, \dots, h_k)$ propomos para h_j uma distribuição Gama (α, β), com α e β conhecidos. Como a família de distribuição Gama é conjugada natural da família de distribuição amostral Poisson, teremos uma distribuição a *posteriori* fechada e conhecida para $h_j|y_i$.

Para coletar as amostras da distribuição a *posteriori* utilizamos o método MCMC com Saltos Reversíveis. Em cada iteração, dado um o valor de k, temos seis possíveis passos para a cadeia de Markov, são eles *Birth, Death, Shift, Switch, Height* e *Hyper*, cada possível passo tem a sua respectiva probabilidade de ocorrência $r_B(k)$, $r_D(k)$, $r_{Sh}(k)$, $r_{Sw}(k)$, $r_{He}(k)$ e $r_{Hy}(k)$. Cada passo é aceito como um novo estado da cadeia de Markov, de acordo com o cálculo da razão de probabilidade do algoritmo de Metropolis-Hastings-Green (Green, 1995).

Para o aplicativo BayesMaps utilizamos apenas os passos *Birth* e *Death*. Após cada passo realizamos um passo GIBBS para atualização dos riscos relativos. As probabilidades de ocorrência de cada passo no aplicativo BayesMaps são: $r_B(k) = r_D(k) = 0{,}5$.

Passo *Birth*: é proposto um acréscimo no número de *cluster* de k para $k+1$ adicionando um centróide g_j^* no vetor G_k. Sorteamos de forma equiprovável a região que será o novo centróide a ser inserido em G_{k+1}^* e em qual posição ele será inserido. No aplicativo BDCD o novo valor do risco relativo h_j^*, correspondente ao novo g_j^*, segue uma distribuição Gamma:

$$h_j^*|y_i \sim Gama\left(y_j + \frac{\tilde{\mu}^2}{\tilde{\sigma}^2}, e_j + \frac{\tilde{\mu}}{\tilde{\sigma}^2}\right) \qquad (2)$$

onde $e_j = \sum_{i \in C_j^*} e_i$, $y_j = \sum_{i \in C_j^*} y_i$, $\tilde{\mu} = \exp\{\mu + 0{,}5\sigma^2\}$ e $\tilde{\sigma}^2 = \exp\{\sigma^2 - 1\}\exp\{2\mu + \sigma^2\}$.

No aplicativo BayesMaps, dado que alteramos a distribuição a priori de h_j, o novo valor do risco relativo h_j^*, correspondente a g_j^*, segue uma distribuição Gama: $h_j^*|y_i \sim Gama\,(y_j + \alpha, e_j + \beta)$. O passo *Birth* é aceito com probabilidade $\phi = \min\left\{1, \frac{(1-c)p\left(h_j^*\right)}{q\left(h_j^*\right)} \frac{L(y|H_k^*)}{L(y|H_k)}\right\}$.

Passo *Death*: é proposto um decréscimo no número de cluster de $k+1$ para k excluindo um centróide g_j^*. Dessa forma, excluímos o centróide e seu respectivo risco relativo dos vetores G_{k+1} e H_{k+1}. A probabilidade ϕ de aceitação do passo *Death* é dada pelo inverso da probabilidade de aceitação do passo *Birth*.

Resultados e conclusões

Todas as etapas necessárias para realização da metodologia BDCD com as modificações propostas foram implementadas em linguagem de programação C/C++ Orientada a Objeto. Utilizamos o compilador gratuito Dev-C++ 4.9.9.2 disponível para *download* no endereço eletrônico *www.bloodshed.net/devcpp*. O programa foi dividido em três objetos principais descritos resumidamente a serguir. O arquivo *main.cpp* contém a rotina principal. Executa as

funções para leitura dos dados, clusterização, loop das iterações, criação dos arquivo de saída, etc. O arquivo *bayesm.cpp* contém a estrutura de todas as funções e o arquivo *bayesm.h* contém a declaração de todas classes e funções

Desejamos agora comparar com os dois aplicativos, BayesMaps e BDCD na regionalização de uma região simulada. Propomos a alteração na metodologia BDCD da distribuição a *priori* dos riscos relativos, que no aplicativo BDCD temos que $\log(h_j) \sim N(\mu, \sigma^2)$ com μ e σ^2 desconhecidos e no aplicativo BayesMaps temos que $h_j \sim G(\alpha, \beta)$ α e β conhecidos. Testamos a metodologia somente com dois possíveis passos da Cadeia de Markov, *Birth* e *Death,* sendo que ao final de cada passo propomos a realização de um passo Gibbs para atualização de todos os riscos relativos:

$$h_j^* | y_i \sim Gama\left(\sum_{i \in C_j} y_i + \alpha, \sum_{i \in C_j} e_i + \beta\right)$$

Para execução do aplicativo implementado BayesMaps e do aplicativo BDCD utilizamos um *burn-in* de 500.000 iterações e, em seguida, coletamos uma amostra a cada 1.000 iterações até totalizar 1.000 amostras. Totalizando 1.500.000 iterações. Para a distribuição a *priori* do número de *clusters* utilizamos a constante $c = 0{,}01$ para ambos os aplicativos.

As probabilidades de ocorrências de cada uma dos possíveis movimentos da cadeia de Markov foram: $r_B(k) = r_D(k) = 0{,}4$ e $r_{Sh}(k) = r_{Sw}(k) = r_{He}(k) = r_{Hy}(k) = 0{,}05$ com $k = \{2, \ldots, n-1\}$ para o aplicativo BDCD, e $r_B(k) = r_D(k) = 0{,}5$ com $k = \{2, \ldots, n-1\}$ para o aplicativo BayesMaps.

Para o aplicativo BayesMaps utilizamos a distribuição gama com média igual a 1 e alta variância para a distribuição a *priori* dos riscos relativos: $h_j \sim G(0{,}01; 0{,}01)$, $j = 1, \ldots, k$ iid. As especificações das distribuições a *priori* dos hiperparâmetros do aplicativo BDCD foram as recomendas no artigo Knorr-Held (2000): $\mu \propto constante$ e $\sigma^2 \sim GI(1; 0{,}01)$.

Como o banco de dados simulados possui o mesmo risco sobre toda a área era esperado que os aplicativos reconstruam a área com o menor número de *clusters* possível.

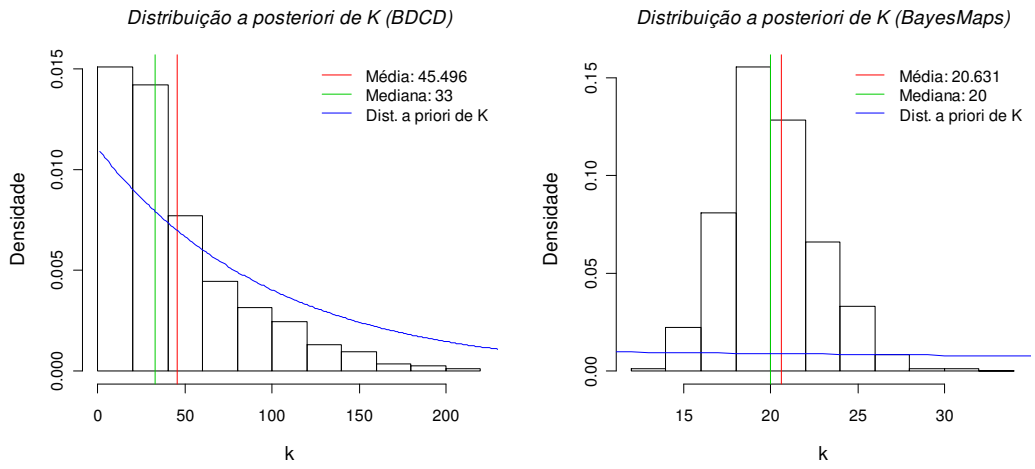

Figura 1: Histogramas do número de clusters a *posteriori*.

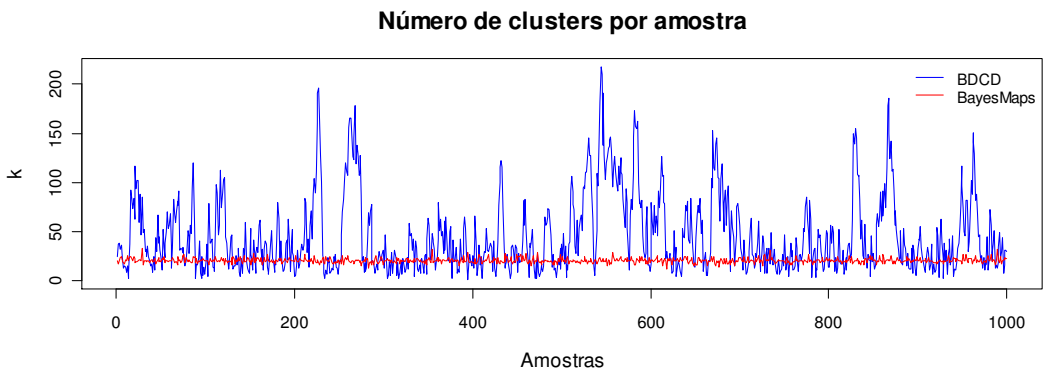

Figura 2: Número de clusters a *posteriori* por amostra.

Observamos da Figura 1 que as distribuições a *posteriori* do número de clusters são bem distintas. Vemos que a distribuição a *posteriori* do aplicativo BDCD apresentou-se semelhante a distribuição a *priori* de k. Em contraposição, temos uma distribuição a *posteriori* de k para o aplicativo BayesMaps, centrada em uma pequena faixa de valores de k.

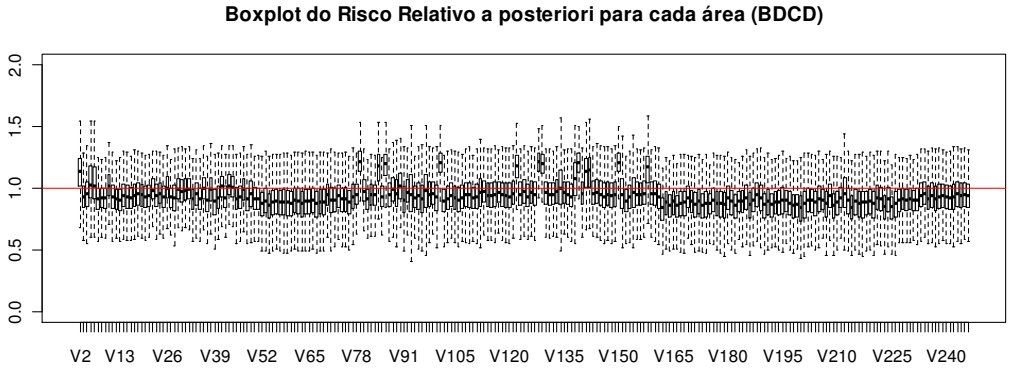

Figura 3: Boxplot - Risco relativo a *posteriori* para cada área (BDCD).

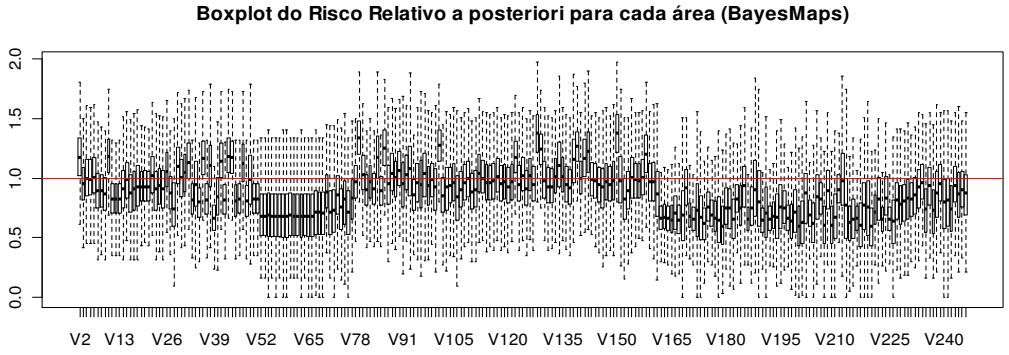

Figura 4: Boxplot - Risco relativo a *posteriori* para cada área (BayesMaps).

As Figuras 3 e 4 mostram os boxplots do risco relativo a *posteriori* para cada uma das 245 áreas enumeradas de *V1* a *V245*. Para facilitar a visualização, retiramos os *outliers* dos boxplots.

Observamos que, para as áreas *V55* a *V78* e para as últimas áreas, *V165* a *V245*, o aplicativo BayesMaps subestimou o risco relativo enquanto o BDCD apresentou resultados mais estáveis centrados no valor 1 ao longo de todas as áreas. Logo, o aplicativo BDCD para a área simulada sem a presença de *cluster* parece reconstruir melhor a superfície.

No geral, as estimativas da média e da mediana a *posteriori* para os dois aplicativos apresentaram resultados bem semelhantes. Como a região simulada possui risco relativo igual a 1 para todas as áreas, podemos concluir que, para esse cenário simulado com as modificações propostas e implementadas no aplicativo BayesMaps, o aplicativo BDCD possui melhores estimativas pontuais para o risco relativos das áreas.

Computacionalmente consideramos nesse primeiro momento como satisfatório o processo de implementação da metodologia. Ainda temos questões sobre o custo computacional para investigar e otimizar.

O aplicativo BayesMaps apresentou estimativas pontuais a *posteriori* menores para o número de *clusters*, o que consideramos como desejado para o cenário simulado. Porém o aplicativo BDCD apresentou uma distribuição a *posteriori* do risco relativo melhor que a apresentada pelo aplicativo BayesMaps. Consideramos que uma melhor estimação da distribuição a *posteriori* do risco relativo é mais útil que uma melhor estimação do número de clusters a *posteriori*.

Referências:

KNORR-HELD, L., RASSER, G. *Bayesian Detection of Clusters and Discontinuities in Disease Maps*. Biometrics v.56, p 13-21, 2000.

KNORR-HELD, L., RASSER, G. *Bayesian Detection of Clusters and Discontinuities in Disease Maps*. (Revisado em Fevereiro de 1998), Artigo de discussão 107, FSB Collaborative Research Center 386, University Munich.

KNORR-HELD, L., RASSER, G. G. *Bayesian Detection of Clusters and Discontinuities in Disease Maps: Simulations*. Artigo de discussão 142, FSB Collaborative Research Center 386, University Munich.
(Disponível em www.stat.uni-muenchen.de/sfb386/publikationen.html)

GREEN, P. J. *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.* Biometrika, v.82, p 711-732, 1995.

Castro, M. S. M. de, Vieira, V. A. e Assunção, R. M. *Padrões espaço-temporais da mortalidade por câncer de pulmão no Sul do Brasil.*Revista Brasileira de Epidemiologia, v.7 , nº 2, 2004.