UFAM

DETECÇÃO DE CLUSTERS ESPACIAIS EM MODELOS DE REGRESSÃO BETA

Vanessa Souza dos Santos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Matemática

Orientador: Max Sousa Lima

Manaus

Abril de 2015

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Santos, Vanessa Souza dos
S237d Detecção de clusters espaciais em modelos de regressão beta / Vanessa Souza dos Santos. 2015
61 f.: il.; 31 cm.

Orientador: Max Sousa Lima
Dissertação (Mestrado em Matemática - Estatística) - Universidade Federal do Amazonas.

1. Detecção de Clusters . 2. Estatística Scan Espacial. 3. Modelos de Regressão Beta. 4. Valor p Bootstrap. I. Lima, Max Sousa II. Universidade Federal do Amazonas III. Título

DETECÇÃO DE CLUSTERS ESPACIAIS EM MODELOS DE REGRESSÃO BETA

Vanessa Souza dos Santos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DA UNIVERSIDADE FEDERAL DO AMAZONAS, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM MATEMÁTICA.

Examinada por:

Prof. Max Sousa de Lima, D.Sc.

Prof. Luiz Henrique Duczmal, D.Sc

Prof. James Dean Oliveira Santos Junior, D.Sc.

MANAUS, AM – BRASIL
ABRIL DE 2015

Este trabalho dedido à minha mãe

Walterina (in memoriam).

Agradecimentos

A Deus, por todas as bençãos recebidas, pelo ar que respiro, pelos dons que me deste e pelos relacionamentos que possibilitam que eu cresça a cada dia.

Ao meu Orientador, professor Max Lima, pelo incentivo, confiança e por suas valiosas contribuições para a elaboração deste trabalho.

Ao professor Diego Souza pela sua significativa contribuição na elaboração do pacote no R do modelo proposto neste trabalho.

Ao professor Luiz Henrique Duczmal e ao professor James Dean Oliveira por fazerem parte da banca examinadora e por suas contribuições.

A todos os professores do curso de Estatística da Universidade Federal do Amazonas por me proporcionar o conhecimento. Ao coordenador da pós graduação em Matemática, professor Roberto Cristóvão. Em especial, quero muito agradecer ao professor Celso Rômulo Cabral, por ter sempre acreditado em mim , desde sua orientação de PIBIC.

Agradeço especialmente minha amiga de mestrado, Márcia Brandão, que desde o início do curso, sempre enfrentamos unidas todas as barreiras das dificuldades e no final sempre vencemos. Ao seu companheirismo, descontrações e conselhos.

À minha amiga Carina, pela mão amiga sempre estendida nos momentos de dificuldade; pela generosidade, pela motivação constante e pelo exemplo de humildade.

Agradeço a todos meus colegas do curso de mestrado em matemática, que fizeram parte dessa etapa parcial em minha vida. Às minhas amigas Regina, Carla Zeline, Camila Pinheiro que sempre acreditaram em mim.

Ao meu irmão André, que está sempre comigo.

Aos meus falecidos pais Nilton e Walterina. À minha eterna Mãe, apesar que ela não está comigo e sim com o Pai, mas estará sempre no meu coração. Todos seus sonhos dela, estão sendo concretizados a partir desse momento em diante.

À CAPES, pelo apoio financeiro nesses 2 anos de estudos.

“ O reino dos céus é semelhante a um grão de mostarda, que um homem tomou e plantou no seu campo; o qual grão é, na verdade, a menor de todas as sementes, mas depois de crescido, é a maior das hortaliças e faz-se árvore, de tal modo que as aves do céu vêm pousar nos seus ramos.” Jesus Cristo (Mateus 13:31-32).

Resumo da Dissertação apresentada ao Programa de Pós-Graduação em Matemática, da Universidade Federal do Amazonas, como parte dos requisitos necessários para a obtenção do grau de Mestre em Matemática. (M.Sc.)

DETECÇÃO DE CLUSTERS ESPACIAIS EM MODELOS DE REGRESSÃO BETA

Vanessa Souza dos Santos

Abril/2015

Orientador: Max Sousa Lima

Linha de Pesquisa: Estatística

A Estatística Scan Espacial tem sido desenvolvida para detecção de cluster espacial em diferentes tipos de modelos, como por exemplo, Bernoulli, Multinomial, Poisson, Exponencial, Weibull e Normal. Entretanto, alguns dados são contínuos no intervalo $(0,1)$, tais como as taxas e proporções, ou são limitados no intervalo (a,b), $a < b$. Portanto, neste trabalho, vamos propor uma estatística scan espacial baseada em modelos de regressão Beta. A estatística de teste é baseada na razão de verossimilhança e avaliada usando o método de Bootstrap para o valor p. O método proposto é aplicado usando a taxa de mortalidade infantil no Estado do Amazonas, Brasil. A função poder, a sensibilidade e o valor predito positivo do teste são analisadas através de um estudo de simulação.

Abstract of Dissertation presented to Postgraduate in Mathematics, of the Federal University of Amazonas, as a partial fulfillment of the requirements for the degree of Master of Mathematics. (M.Sc.)

SPATIAL CLUSTER DETECTION FOR BETA REGRESSION

Vanessa Souza dos Santos

April/2015

Advisor: Max Sousa Lima

Research lines: Statistics

Spatial Scan Statistics has been developed for geographical cluster detection in different types of models, for example, Bernoulli, Multinomial, Poisson, Exponential, Weibul and Normal. However, some data are continuos in the interval $(0,1)$ such as rates and proportions or are limited in the interval (a,b), $a < b$. Therefore, in this work, we propose a spatial scan statistic for Beta regression model. The test statistics is based on a likelihood ratio test and evaluated using Bootstrap p-value. The proposed method is illustrated using infant mortality in the Amazonas State, Brazil. The Statistical power, sensitivity and positive predicted value of the test are examined through a simulation study.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A distribuição geográfica ou espacial de incidência de algum fenômeno de interesse, como doenças, homicídios, desmatamento, é de extrema importância para a implementação e planejamento de políticas públicas em uma área, por exemplo, município, estado ou país.

Diversos trabalhos, principalmente nas áreas de epidemiologia e saúde pública, vem sendo desenvolvidos avançadas técnicas computacionais relacionadas à detecção de conglomerados espaciais. Neste texto, conglomerados espaciais serão tratados pela palavra em inglês, clusters, como já é bastante usual nesta área de pesquisa.

Um cluster espacial é uma parte de um mapa, uma determinada área em que a ocorrência de casos de um fenômeno de interesse é discrepante do restante do mapa, isto é, alta demais ou baixa demais, com grande potencial de risco à população monitorada. No jargão epidemiológico, um cluster é uma inesperada aglomeração de eventos relacionados à saúde. Além de epidemiologia e saúde pública, a detecção de cluster é comumente usada em outras áreas como engenharia, astronomia, biologia, genética, veja Glaz (2009). Essa detecção e localização tem sido abordada através de teste de hipóteses, ou seja queremos responder as seguintes questões: Os casos estão distribuídos de forma aleatória nestas áreas? Existe uma região do mapa em que há algum valor discrepante dos demais? O nosso objeto de interesse é testar:

$$\begin{cases} H_0: & \text{não existe cluster no mapa} \\ H_1: & \text{existe cluster no mapa} \end{cases}$$

Existem na literatura vários métodos para detecção de clusters espaciais. Nesse

contexto, o Scan Espacial (Kulldorff, 1997) é atualmente usado em vários departamentos de saúde para detecção de clusters circulares (Kulldorff & Nagarwalla, 1995) e tem sido desenvolvido para dados discretos e contínuos usando os modelos Bernoulli (Kulldorff & Nagarwalla, 1995), Poisson (Kulldorff, 1997), Multinomial (Jung *et al.*, 2010), Exponencial (Huang *et al.*, 2007), Normal (Huang *et al.*, 2010) e Weibull (Bhatt & Tiwari, 2014).

Extensões desse método foram propostas para acomodar correlação espacial (Loh & Zhu, 2007), ajuste por Covariáveis (Jung, 2009), modelos log-lineares (Zhang & Lin, 2009), dados multivariados (Kulldorff *et al.*, 2007; Neill *et al.*, 2013), eventos repetidos (Rosychuk & Chang, 2013), sobredispersão e inflação de zeros (Zhang *et al.*, 2012; Cançado *et al.*, 2014; Lima *et al.*, 2015). Comparação de poder e aproximações da Estatística Scan são descritos em (Kulldorff *et al.*, 2003; Lima, 2004; Read *et al.*, 2013) e, recentemente (Prates *et al.*, 2014) tem discutido o vício deste método na estimação dos riscos relativos.

O Scan Espacial é utilizado para detectar e testar a significância de clusters localizados, sem o conhecimento a priori da localização e tamanho do cluster ajustando para o problema de testes múltiplos. Na terminologia computacional, dizemos que o método varre o mapa em estudo impondo sobre ele uma janela que pode apresentar qualquer forma geométrica (Duczmal & Assunção, 2004; Duczmal *et al.*, 2006; Assunção *et al.*, 2006), porém, neste trabalho, usamos o Scan Circular (Kulldorff & Nagarwalla, 1995). Neste caso, o método utiliza uma janela circular centrada nos cetróides das áreas avaliadas e os círculos contruídos contém diferentes conjuntos de áreas incluindo áreas vizinhas. Para cada ponto onde o círculo é centrado, o raio varia continuamente de zero até um limite superior, que usualmente não ultrapassa a 50% do total da população em risco. Esta janela circular é flexível tanto em tamanho como em localizações. No total o método cria uma classe Z de círculos distintos e cada diferente conjunto de áreas que pertencem a um determinado círculo é chamado de zona z , e cada zona $z \in Z$ é um possível candidato a cluster. A significância estatística de z é avaliada através do teste da razão de verossimilhança.

Em alguns dos modelos probabilísticos citados anteriormente para construção do Scan Espacial, supõe-se que o suporte da variável aleatória é ilimitado. No entanto existem situações na qual a variável de interesse é continuamente limitada no intervalo (a,b),

onde a e b são escalares conhecidos com $a < b$. Uma particular situação ocorre quando $a = 0$ e $b = 1$ de modo que a variável aleatória assume valores em $(0,1)$, como é o caso de taxas, proporções e números índices. Para esse tipo de dados, uma modelagem via distribuição Beta é mais adequada. Usando uma nova parametrização da distribuição Beta, Ferrari & Cribari-Neto (2004) desenvolvem um modelo de regressão o qual em muitos aspectos é similar a classe de modelos lineares generalizados, mas esta distribuição não pertence a esta classe. Desta forma, neste trabalho, uma Regressão Beta é utilizada para construir uma nova Estatística Scan Espacial para dados limitados no intervalo (a,b), onde o valor esperado do modelo é ajustado por covariáveis. A estimação dos parâmetros é realizada via Método de Newton-Raphson.

1.1 Objetivos

O presente trabalho tem como objetivo principal apresentar um novo método de detecção de clusters espacias através da Estatística de Scan de Kulldorff para o modelo de regressão Beta. Analisamos a performance da Scan Espacial através de um estudo de simulação do poder do teste e a precisão na detecção do cluster, medida através da Sensibilidade e do Valor Predito Positivo. Outro objetivo, é aplicar os dados da taxa de mortalidade infantil nos municípios do Estado do Amazonas no modelo proposto. O desempenho da scan é avaliado usando o valor-p bootstrap.

1.2 Indicadores Quantitativos

Durante a elaboração deste trabalho, foram realizadas as seguintes produções:

- Apresentação de resumo do trabalho no 21º SINAPE - Simpósio Nacional de Probabilidade e Estatística em 2014;
- Criação de um pacote em R do modelo proposto neste trabalho;
- Artigo submetido em um periódico internacional.

1.3 Estrutura do Trabalho

O presente trabalho está estruturado em 6 capítulos, cujos conteúdos são descritos abaixo.

O Capítulo 1 consiste na introdução, descrevendo o tema a ser estudado , além de apresentar os seus objetivos. São relacionadas ainda as publicações decorrentes do trabalho desenvolvido durante o curso.

No Capítulo 2 apresentamos uma breve introdução dos principais métodos para a detecção de clusters espaciais, descrevendo com mais detalhes a Estatística Scan Circular proposta por Kulldorff & Nagarwalla (1995) .

No Capitulo 3, é realizado um breve resumo do modelo de regressão Beta proposto por Ferrari & Cribari-Neto (2004), e em seguida apresentamos a metodologia proposta, o modelo Scan Espacial para o Modelo de Regressão Beta . Um estudo de simulação para verificar a performance do Scan Circular proposto é realizado no Capitulo 4.

No Capítulo 5 aplica-se o método em um conjuntos de dados reais e analisamos os resultados obtidos e pacote `betaScan` implementado no software `R`.

O Capítulo 6 apresenta as considerações finais e as propostas de continuidade desse trabalho.

Capítulo 2

Detecção de Clusters Espaciais

Do ponto de vista epidemiológico, denomina-se conglomerado ou cluster a um excesso de casos ou taxas de ocorrências de eventos relacionados à saúde em uma determinada área geográfica (conglomerado espacial), em um período de tempo limitado (conglomerado temporal), ou ainda considerando o monitoramento simultâneo do espaço e tempo (conglomerado espaço-temporal). Neste capítulo, é apresentada uma breve introdução sobre os principais métodos para detectar clusters Espaciais.

2.1 Tipos de Dados

A análise de agrupamento espacial (*Clusters Espaciais*) desempenha um papel importante na quantificação dos padrões de variação geográfica. Normalmente, é usado em vigilância de doenças, epidemiologia espacial, genética de populações, astronomia, análise criminal e muitos outros campos, mas os princípios são os mesmos.

Os dados utilizados em estatística espacial possuem um índice que faz referência a uma área geográfica, geralmente representada em um mapa bidimensional. Essa referencia é representada pela coordenada geográfica do local estudado. Como exemplo, suponha que há interesse em estudar os casos de assaltos em uma cidade. Neste caso, se a ocorrência do assalto for a saída de um banco, é necessário a informação exata da localização de cada ocorrência, ou se há informação do número de ocorrência nos bairros dessa cidade, uma alternativa será usar o centróide da coordenada geográfica para cada bairro.

Os diferentes tipos de dados espaciais são tradicionalmente classificados de acordo

como:

1. Dados de Processos Pontuais;

2. Dados de Área;

3. Dados de Superfícies Aleatórias .

No caso de dados pontuais, um par (s_{1l}, s_{2l}) indica a coordenada geográfica de ocorrência do evento de interesse de um dado mapa S particionado em L localizações s_l, para $l = 1, 2, \ldots, L$. Dados de área são obtidos quando não estão disponíveis as coordenadas de cada ocorrência de um evento, mas apenas o número total de ocorrências em cada região, por exemplo, a ocorrência de assaltos nos bairros da cidade de Manaus. Já os dados de superfície são obtidos, ao se realizar medições em determinadas localizações do mapa, sendo então cada elemento do conjunto de dados formado por (s_{1l}, s_{2l}, s_{3l}) que corresponde à coordenada geográfica aliada à medição feita naquela localização (por exemplo, temperatura, umidade ou pressão). Ao se analisar dados pontuais e dados de área, deve-se considerar se a ocorrência dos eventos se dá de forma aleatória. Nesse caso, é importante conhecer a natureza dos dados, a fim de encontrar o modelo estatístico mais adequado, se for o caso.

Um processo pontual pode ser transformado em dados de área (Lima, 2011), pois algumas técnicas requerem um ponto de referencia da área limitada, em geral a observação é representada pelo centróide dessa área. Por isso, nosso foco são processos espaciais modelados como processos medidos em áreas (ou dados de área). Nesse caso, supõe-se que existe um processo estocástico $\mathbf{Y}(s) = \{Y(s_l), l = 1, 2, \ldots, L\}$, onde $Y(s_l)$ é a variável aleatória do processo em uma determinada área A_l, identificada por um ponto $s_l \in S = \{s_1, \ldots, s_L\}$ que corresponde ao centro do polígono limitado por A_l.

2.2 Testes para Detecção de Clusters

O estudo de clusters espaciais são abordados através de testes de hipóteses, e por isso muitos métodos estatísticos são desenvolvidos para detectar clusters incorporando a variação espacial da população em estudo. Esses testes tem como objetivo averiguar quando um padrão observado de eventos em uma ou mais áreas pode ser completamente distribuídos ao acaso. Por exemplo, considere um mapa S particionado em L localizações

s_l e seja $Y(s_l)$ o número de eventos ocorridos em uma área A_l delimitada. Geralmente, sob hipótese nula

$$H_0 : Y(s_l) \sim Poisson(\lambda N(s_l))$$

onde λ é a taxa global dos eventos e $N(s_l)$ o total da população em risco na área A_l. A hipótese nula do teste representa a completa aleatoriedade espacial dos eventos, implicando que λ é a mesma em todas as áreas, ou seja, o número esperado de eventos em um local é proporcional à sua população.

2.2.1 Classificação dos Testes para Detecção de Clusters

Besag & Newell (1991) classificaram os testes como: Teste Geral e Teste Focado. No entanto, (Lawson & Kulldorff, 1999) subdividiram o Teste Geral em teste global e localizado. Teste global é útil para investigar se uma doença é ou não infecciosa . Os testes para cluster localizado, são usados para estimar a localização de pequenas áreas com elevado risco e avaliar sua significância estatística. Em contrapartida os testes focados concentram o estudo em uma ou mais áreas pré-selecionadas. Geralmente, esses testes utilizam técnicas computacionais intensivas como permutação aleatória, como o Monte Carlo, Bootstrap, etc. Em uma exaustiva revisão, Kulldorff destaca a existência de mais de 100 métodos diferentes. A seguir, vamos apresentar alguns métodos baseados em teste geral.

2.2.2 Métodos para a detecção de clusters espaciais

Dentre os métodos para a detecção de *clusters* espaciais mais conhecidos estão: (a) o Método GAM (Geographical Analysis Machine) (Openshaw *et al.*, 1988); (b) Método de Besag e Newel (Besag & Newell, 1991); (c) Método de Cuzick e Edward (Cuzick & Edwards, 1990); (d) Método Scan Espacial (Kulldorff, 1997). Esses métodos serão descritos a seguir, exceto o método Scan Espacial, que será discutido na seção 2.3.

Método GAM (Geographical Analysis Machine)

Openshaw *et al.* (1988) propôs um método intensivo computacionalmente e com grande apelo visual, conhecido como Geographical Analysis Machine e abreviado por

GAM. O método GAM se baseia em construir múltiplos círculos sobrepostos e de tamanhos variáveis, observar a contagem do número de casos e do número de pessoas em risco dentro do círculo, calcular uma proporção (taxa) de incidência local e apresentar aqueles círculos com taxas excedendo algum limiar pré-estabelecido. O objetivo para os círculos sobrepostos era combinar informações de áreas vizinhas a fim de estabilizar estimativas locais. Considere $Y(s_l)$ o número de eventos em uma área A_l do mapa, com valor esperado dado por $\lambda N(s_l)$, onde $N(s_l)$ é o número total da população na área A_l. Associe os valores de cada área ao seu centróide (centro do polígono da área A_l) denotado por s_l. O procedimento GAM utiliza o seguinte algorítimo:

1. Selecione um raio r (por exemplo, $r = 1,3$ ou 4 km);

2. Em cada centróide s_l fixe um circulo C_{l_r} de raio r;

3. Calcular

$$Y(s_l)_r = \sum_{l=1}^{L} Y(s_l)\mathtt{I}_{s_l \in C_{l_r}} \qquad \text{e} \qquad N(s_l)_r = \sum_{l=1}^{L} N(s_l)\mathtt{I}_{s_l \in C_{l_r}}$$

   o número de eventos e o número total da população em risco habitando o círculo C_{l_r} de raio r, onde $\mathtt{I}$ é a função indicadora;

4. Calcule o valor p, p_{l_r} do teste associado a $Y(s_l)$, sob hipótese nula, considere

$$H_0 : Y(s_l) \sim Poisson(\lambda N(s_l));$$

5. Desenhe o círculo C_{l_r} se $p_{l_r} \geq 0.002$;

6. Repita o procedimento acima aumentando (ou escolhendo) outro raio para o círculo.

O resultado final é a identificação de clusters de áreas por emaranhados de círculos como mostrado na Figura (2.1).

As vantagens e desvantagens do método GAM são: é simples de entender, é um método exploratório e não inferencial devido ao problema de muitos testes simultâneos e dependentes; é intensivo computacionalmente; os círculos não são inteiramente comparáveis entre sí, pois as variáveis aleatórias envolvidas possuem diferentes distribuições.

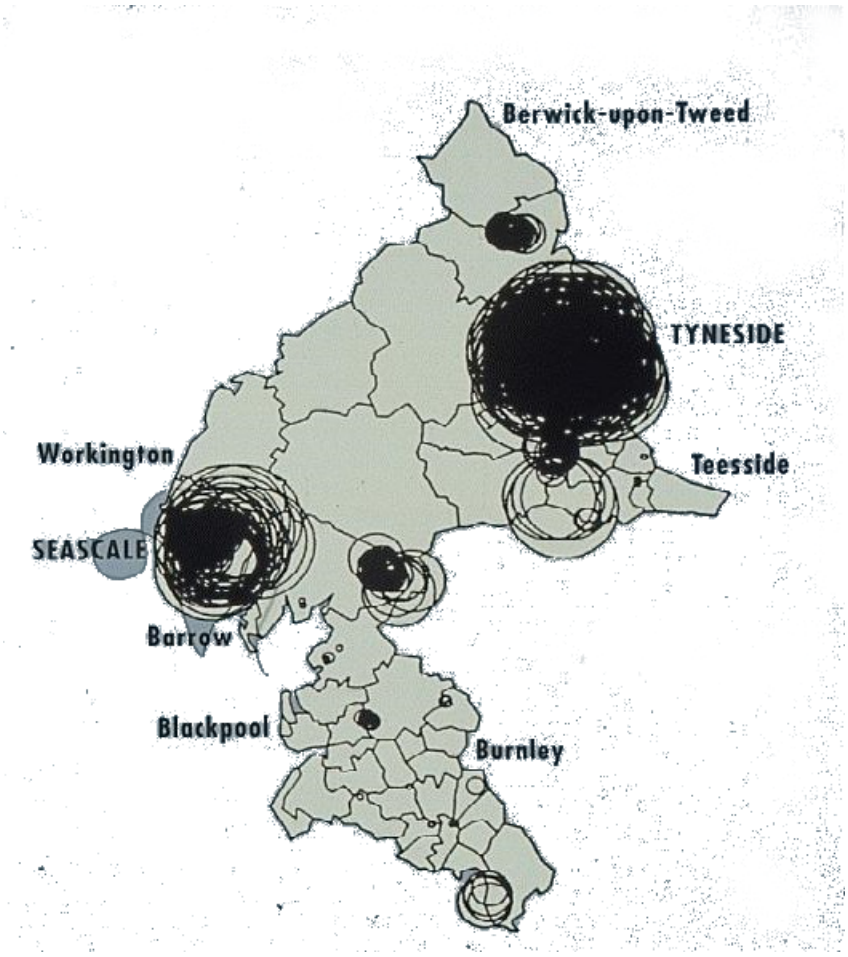


Figura 2.1: Exemplo visual do uso do método GAM mostrando clusters de área por emaranhado de círculos

Método de Besag e Newel

O método proposto por Besag & Newell (1991) é um método visual, semelhante ao método GAM , que procura identificar conglomerados verossímeis de formato circular. A área de risco é identificada por um emaranhado de círculos significativos, sobrepostos. Cada círculo contém em seu interior um número fixo de k eventos que devem ser buscados e calcula-se o raio necessário para englobá-los. No circulo resultante, calcula-se o valor p e, como procedendo o método GAM, desenha apenas os círculos significativos (valor $p \leq 0.002$). Em seguida, varia k para verificar a estabilidade dos resultados.

Para computar o valor p, seja $X = \sum_{l=1}^{L} Y(s_l)$ e $N = \sum_{l=1}^{L} N(s_l)$. Centrado em s_l, assuma que a área A_l possui pelo menos um caso. Seja $\widetilde{L}$ a variável aletória que conta o número de outras áreas (ou centróides) necessárias para acumular os k primeiros casos mais próximos de s_l. Seja $\widetilde{l}$ o valor observado de $\widetilde{L}$ e $N_{\widetilde{l}}$ o total da população nessas $\widehat{l}$ áreas. Sob hipótese nula (a mesma do método GAM), $y_{\widetilde{l}}$ o número de eventos nessas $\widetilde{l}$

áreas segue distribuição Poisson com valor esperado dado por $N_{\widetilde{l}}Y/N$. Agora, notando que $P(\widetilde{L} \leq \widetilde{l}) = 1 - P(\widetilde{L} > \widetilde{l}+1)$ representa 1 menos a probabilidade de as $\widetilde{l}$ primeiras áreas possuam menos que k eventos. Então o valor p para um círculo C_{l_k} centrado em s_l contendo k eventos é dado por,

$$p_{l_k} = 1 - \sum_{j=1}^{k-1} P(X_{\widetilde{l}} = j) = 1 - \sum_{j=1}^{k-1} \frac{\left(N_{\widetilde{l}}Y/N\right)^j}{j!} e^{N_{\widetilde{l}}Y/N}.$$

Método de Cuzick e Edwards

Cuzick & Edwards (1990) desenvolveram uma proposta que representa uma pequena variação em relação aos métodos de Besag & Newell (1991). Como em Besag & Newell (1991), inicia-se fixando o número de eventos k. A seguir, em torno do centróide de cada área A_l que possui pelo menos um evento, traça-se um círculo que vai aumentar, de acordo com a variação do seu respectivo raio até que contenha uma população para qual espera-se observar k eventos. Depois, verifica-se quantos eventos Y_l foram de fato observados e calcula-se a estatística

$$U_k = \sum_{l=1}^{L} (Y_l - k)\mathtt{I}_{\{y(s_l)>0\}}.$$

Cuzick & Edwards (1990) derivaram as fórmulas dos momentos dessa estatística sob hipótese nula e mostraram que ela possui distribuição assintoticamente normal possibilitando assim calcular o valor p para o teste.

2.3 A Estatística Scan Circular de Kulldorff

O método baseado na Estatística Scan (Spatial Scan) foi desenvolvido para detectar e testar a significância de cluster local, sem o conhecimento a priori de sua localização e tamanho.

A estatística Scan espacial foi pela primeira vez estudada por Naus (1965) para detecção de clusters na escala temporal. Porém, Kulldorff & Nagarwalla (1995) e Kulldorff (1997) estenderam essa metodologia para o caso espacial para detectar áreas com elevada taxa de incidência. Vários métodos têm sido desenvolvidos para detecção de clusters espaciais, mas o Scan Espacial de Kulldorff tem se mostrado mais eficiente que os

demais métodos citados na seção 2.2.2 , pois esse método soluciona problemas de ajustes em testes múltiplos.

Considere um mapa S dividido em L localizações $s_1, s_2, \ldots, s_L$. Definimos zona, denotada por z, ao conjunto de quaisquer regiões conectadas entre si. Seja Z o conjunto das áreas z candidatas à cluster. Um primeiro objetivo é encontrar dentro de um mapa todas as possíveis zonas. Esta tarefa pode se tornar computacionalmente impossível para mapas com um número grande de regiões. Por exemplo, suponha um mapa com 800 regiões, portanto existem $2^{800} - 1$ subconjuntos não-vazios que podem vir a formar possíveis zonas. Para contornar este problema, alguns autores propuseram o uso de janelas circulares para verificação de conexidade.

As zonas z canditadas a cluster são polígonos centrados em cada região s_l de coordenada conhecida (s_{1l}, s_{2l}). Em um mapa, a zona z pode assumir diversas formas geométricas, como elipses, círculos, quadrados e assim por diante. Em Kulldorff (1997) utiliza uma janela circular de raio r limitado, que varia de zero até um r_{max} estabelecido. Geralmente, esses raios variam até que o percentual máximo especificado da população total esteja contido no circulo, no geral pode assumir valor em até 50% do tamanho da população total. O máximo de zonas circulares a serem avaliadas é L^2, que do ponto de vista computacional é relativamente simples. Na Figura 2.2 mostra uma possível zona obtida.

2.3.1 Estatística de Teste

Dada uma variável aleatória de interesse $Y(s_l)$ definida na região s_l com função densidade (ou probabilidade) $f(y_l;\theta)$, onde $Y(s_l)$ irá assumir distribuição P_0, se não existir um cluster no mapa S. Caso contrário, $Y(s_l)$ segue distribuição P_1. Ou seja,

$$\begin{cases} H_0 : Y(s_l) & \sim P_0 \quad \forall s_l \in S, \\ H_1 : Y(s_l) & \sim P_1 \quad \forall s_l \in z. \end{cases}$$

A hipótese nula H_0 do teste representa a completa aleatoriedade espacial dos eventos, implicando que a ocorrência de $Y(s_l)$ é a mesma para todas as áreas do mapa S.

A função de verossimilhança é definida como

$$\mathscr{L}(\theta) = f(\mathbf{y};\theta) = \prod_{l=1}^{L} f(y_l;\theta) \tag{2.1}$$

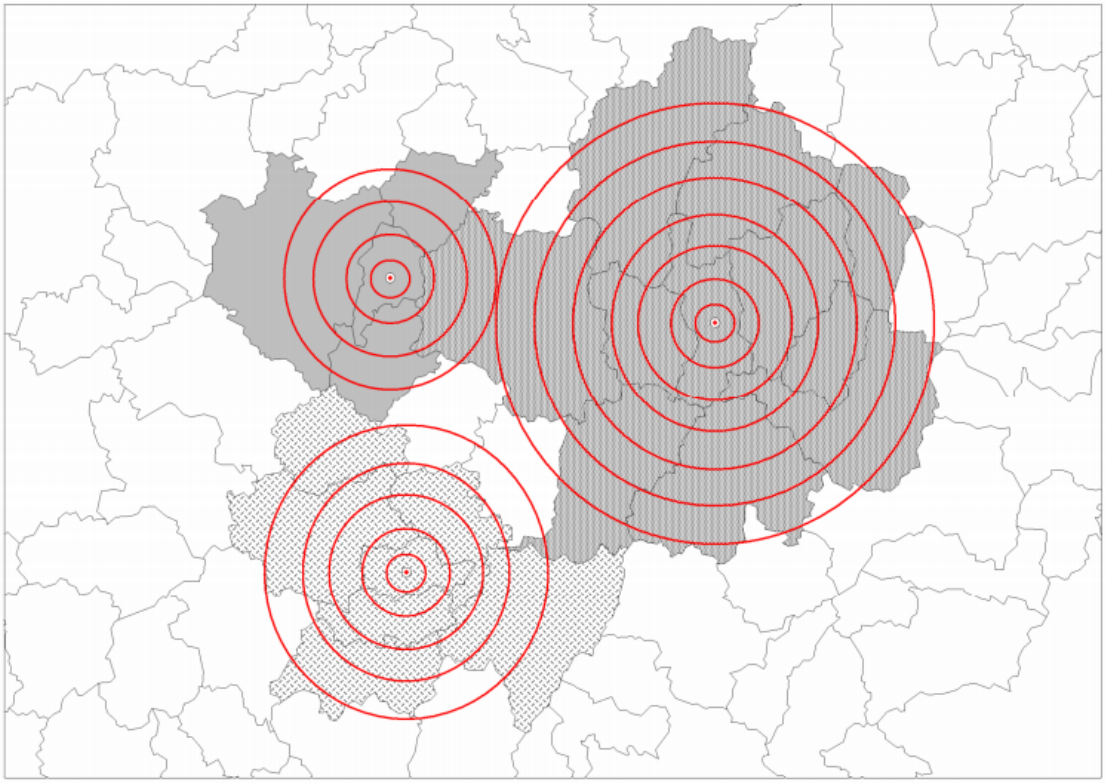


Figura 2.2: Varredura espacial de três regiões. Os círculos são centrados no centróide de cada sub-área e seus raios crescem continuamente, formando zonas candidatas à composição de clusters.

em que $\theta \in \Theta$ é o parâmetro desconhecido do modelo $P(.)$ e Θ denota todo o espaço paramétrico, ver Casella (2002).

Seja $z \in Z$ uma zona, então defina-se $\mathscr{L}(z)$ a função de verossimilhança sob a hipótese alternativa H_1 de que exista uma zona z^* que é um cluster , e $\mathscr{L}(0)$ a verossimilhança sob a hipótese nula H_0 de que não exista um cluster, ou seja

$$\mathscr{L}(\theta) = \prod_{s_l \in S} f(y_l;\theta) = \prod_{s_l \in Z} f(y_l;\theta) \prod_{s_l \notin Z} f(y_l;\theta)\,. \tag{2.2}$$

Para identificar a zona mais provável de ser o cluster z^*, dentre todas as possíveis, o teste proposto por Kulldorff & Nagarwalla (1995) usa o Teste da Razão de Verossimilhança

$$\Lambda^*(z) = \frac{\sup_{H_1} \mathscr{L}(z)}{\sup_{H_0} \mathscr{L}(0)}\,. \tag{2.3}$$

A zona z mais verossímil é aquela que maximiza a função $\Lambda^*(z)$ com respeito ao conjunto Z . Desta forma, a estatística de teste fica definida por $\Lambda = \max_{z \in Z} \Lambda(z)$. Em geral, a

função $\Lambda(z)$ assume valores muito grandes. Para amenizar esse problema, utiliza-se o logaritmo da razão de verossimilhança para $\Lambda(z)$. Dado que a função logaritmo é monotonicamente crescente, temos

$$\Lambda(z) = \{\ell(z) - \ell(0)\} . \tag{2.4}$$

É importante salientar que identificar a zona mais verossímil não constitui em identificar um cluster. Precisa-se ainda verificar a sua significância estatística para que a zona detectada seja considerada como cluster. Visto que a distribuição de $\Lambda(z)$ é intratável analiticamente, a significância estatística da zona mais verossímil identificada nos dados observados é calculada através de simulação Monte Carlo, de acordo com o procedimento descrito em Dwass (1957), ou simulação Bootstrap Efron (1979) . Sob a hipótese nula, casos simulados são distribuídos sobre a região em estudo e a estatística de teste é calculada. Este procedimento é repetido uma grande quantidade de vezes, com o objetivo de produzir uma distribuição empírica para a estatística de teste Λ, sob a hipótese nula. O valor da estatística de teste nos dados observados é então comparado com essa distribuição empírica afim de determinar seu nível de significância (o valor p).

A Estatística Scan Espacial de Kulldorff é mais indicada para detecção de um único cluster bem definido, pois apresenta grande poder de teste, ou seja, o teste baseado na Estatística de Kulldorff é *uniformemente mais poderoso* [1] para detecção de clusters como mostra Kulldorff (1997). Esse poder diminui no caso do mapa em estudo apresentar mais de um cluster ou cluster de formato muito irregular como descrito em Kulldorff *et al.* (2003) e Duczmal *et al.* (2006). A redução do poder do teste está quase sempre associada à superestimação (cluster detectado maior do que o cluster real), ou à subestimação (cluster detectado menor do que o cluster real), como mostra a Figura 2.3.

[1] Um teste uniformemente mais poderoso é um teste de hipótese que tem o maior poder (probabilidade do teste rejeitar corretamente a hipótese nula) entre todos os possíveis testes de um dado tamanho. Mais detalhes em Casella (2002).

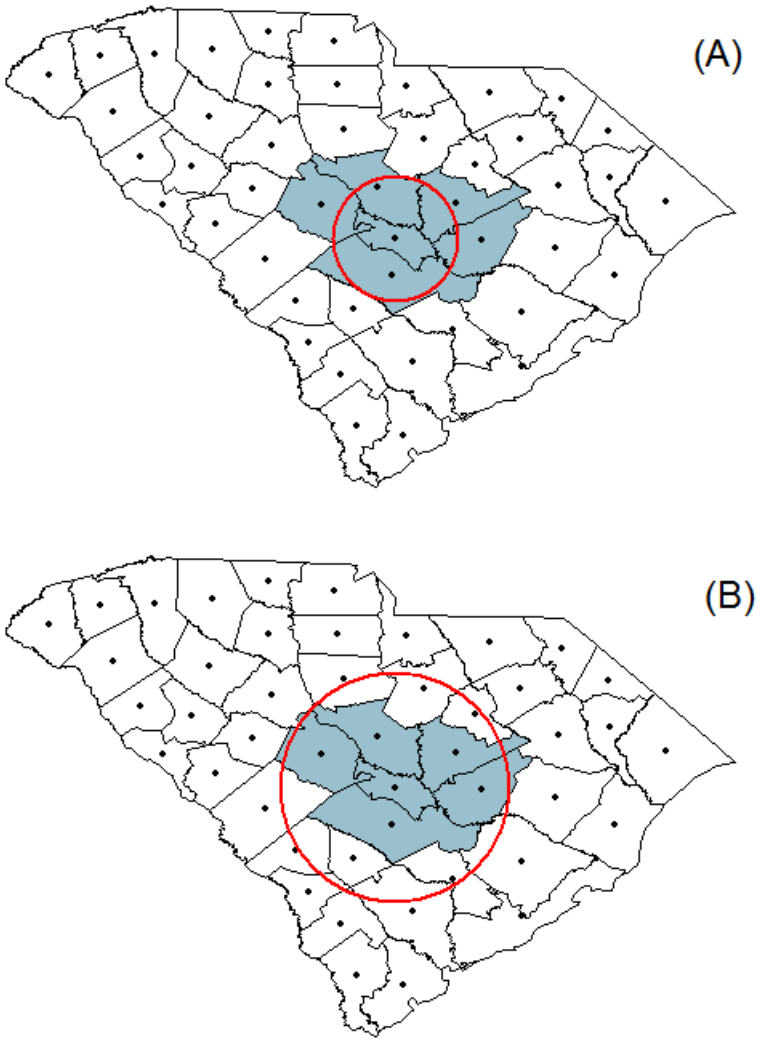


Figura 2.3: Subestimação de cluster (A). Superestimação de Cluster (B)

2.3.2 Representação espacial dos clusters

Para um mapa $S = \{s_1, \ldots, s_L\}$ particionado em L localizações, em que $s_l = (s_{1l}, s_{2l})$ corresponde à coordenada geográfica do centróide da l-ésima área. A distância entre dois centróides quaisquer é dada através da seguinte expressão:

$$D_{l,m} = \sqrt{(s_{1l} - s_{1m})^2 + (s_{2l} - s_{2m})^2} \tag{2.5}$$

onde $D_{l,m}$ representa o elemento da matriz quadrada D de ordem $L \times L$ na l-ésima linha e na m-ésima coluna. Para $l = m$ temos $D_{l,m} = 0$. O próximo passo é ordenar as distâncias encontradas em D, guardando o respectivo índice l do centróide de s_l e o índice c do centróide mais próximo s_c em uma matriz de adjacência I, ou seja

$$I_{l,m} = \begin{cases} l, & \text{se} \quad m = 1 \\ c, & \text{se} \quad s_c \text{ é o } m\text{-ésimo centróide mais próximo de } s_l \end{cases} \tag{2.6}$$

Para um exemplo fácil e ilustrativo, vamos supor um mapa com $L = 5$. Para a linha 1, temos o vetor $(1,5,4,2,3)$. Isto implica que o centróide s_5 é o segundo mais próximo de s_1 , s_4 é o terceiro centróide mais próximo de s_1 e assim sucessivamente. Fixando o centróide s_l, identifique um cluster por um vetor $z_{li} = \left(l_{[i,1]}, l_{[i,2]}, \ldots, l_{[i,L]}\right)$ construído da seguinte forma:

1. Defina $l_{[i,m]} = 1$ se $l = m$, $i = 1,2,\ldots,L$ e $m = 1,2,\ldots,L$;

2. Defina $l_{[i,I_{l,m}]} = 1$, se $s_{I_{l,m}}$ é um dos m centróides mais próximo de s_l e $m \leq i$. Caso contrário, faça $l_{[i,I_{l,m}]} = 0$.

Como exemplo, suponha um mapa $S = \{s_1, s_2, s_3, s_4, s_5\}$ (veja a representação gráfica na seção 3.2.3). Para a linha 1, após ordenar a matriz de distâncias, obteve o vetor $\{s_1, s_5, s_4, s_3, s_2\}$ de centróides. Então para z_{l_i} , obtemos $z_{1_1} = (1,0,0,0,0)$, $z_{1_2} = (1,0,0,0,1)$, $z_{1_3} = (1,0,0,1,1)$, $z_{1_4} = (1,0,1,1,1)$, $z_{1_5} = (1,1,1,1,1)$. Para cada valor de m, z_{l_i} recebe o valor 1 no índice do vizinho mais próximo de s_l em sua posição original no espaço. Esta representação é única a menos do cluster z_{l_L} que surge L vezes diferenciado apenas pelo seu centróide. Para verificarmos que de fato esta é a representação dos clusters, exemplificamos a formação da representação de z_{l_2}. Neste caso $l = 1$, $i = 2$ e $m = 1,2,\ldots,L$.

1. Quando $m = 1$, $l = j$ e portanto $1_{[1,1]} = 1$;

2. Quando $m = 2$, $I_{1,2} = 5$, e s_5 é o segundo $(i = 2)$ centróide mais próximo de s_l e também, $m = 2 \leq 2 = i$. Portanto, $1_{[1,5]}$=1. Agora note que para todo $j \geq 3$ temos que $j > i$. Assim não satisfaz a condição $j \leq i$ implicando que as outras coordenadas de z_{1_2} são nulas. Portanto $z_{1_2} = (1,0,0,0,1)$.

Repetindo o procedimento descrito acima para L aréas, obtemos todos os possíveis candidatos a cluster, $\widetilde{Z} = \{z_{l_i} : l, i = 1,2,\ldots,L\}$. O número total de clusters em $\widetilde{Z}$ é L^2. Assim, podemos construir a seguinte definição:

$$Z = \left\{ z_{l_i} \in \widetilde{Z} : (< z_{l_i}, z_{l_i} >) \leq a \right\} \tag{2.7}$$

onde $(<,>)$ denota o produto interno entre dois vetores e a é um valore fixo, que representa a restrição da quantidade (tamanho do cluster) de localizações espaciais em z_{l_i}. Por exemplo, em muitos casos há interesse em detectar clusters com o número de áreas menor que $a = L/2$, então o maior número de localizações espaciais em z_{l_i} é $L^2/2$ raio dos círculos em z_{l_i}. O raio máximo do círculo z_{l_i} é encontrado através da seguinte expressão:

$$r_{l_i} = \max_{s_i \in z_{l_i}} [D_{l,i}] \tag{2.8}$$

onde $D_{l,i}$ é a distancia euclidiana, conforme encontrado na Equação (2.5).

Finalmente, encontradas todas as possíveis zonas $z_{l_i} \in Z$, calcule a estatística de teste $\Lambda(z_{l_i})$. Então, como visto anteriormente, o valor da estatística de teste será:

$$\Lambda = \max\{\Lambda(z_{l_i}) : i = 1,2,\ldots,a; l = 1,2,\ldots,L\} \tag{2.9}$$

onde $\widehat{z} = \text{argmax}(\widehat{\Lambda}(z_{l_i})$.

2.3.3 Algoritmo Scan Circular

O algoritmo Scan Circular proposto por Kulldorff (1997) apresenta baixa complexidade computacional, facilmente implementável e, por estes motivos, é amplamente utilizado. Este método é similar ao apresentado por Besag & Newell (1991), porém, utiliza-se da estatística para maximizar a função de log verossimilhança para encontrar o *cluster* mais verossímil.

Este método se baseia em uma janela de forma, tamanho e localização que varia sobre uma área geográfica. Para cada janela é calculada a verossimilhança com base no número esperado de eventos dentro e fora desta janela. As regiões contidas na janela de maior verossimilhança definem o cluster mais provável. A significância do teste é feita pelo método de Monte Carlo ou Bootstrap, sob a hipótese nula de que não há existência do cluster, sobre a distribuição da máxima verossimilhança dos dados aleatórios gerados. A hipótese alternativa é de existência do cluster. Uma escolha natural para a forma da janela é a circular Kulldorff (1997), a qual será usada no algoritmo a seguir. O Algoritmo Scan Circular pode ser resumido nos seguintes passos:

INÍCIO

1. Escolher uma região s_l no mapa em estudo;

2. Calcular as distâncias até as outras regiões, ordenando-as em ordem crescente, e guardando-as em um vetor;

3. Criar um círculo centrado de raio limitado por r_{max} na região escolhida no passo 1 e continuamente aumentar o seu raio de acordo com as distâncias encontradas no passo 2. Para cada região s_l que entrar no círculo, atualizar $Y(s_l)$ dentro do círculo Z. Calcular Λ_z para cada $Y(s_l)$. O cluster mais verossímil é aquele de maior Λ_z;

4. Repetir os passos 1, 2 e 3 para cada região do mapa;

5. Utilizar simulações de Bootstrap ou Monte Carlo para avaliar a significância do teste;

6. Se a hipótese nula for rejeitada, então a zona $\widehat{Z}$ associada com a maximização de é o cluster mais plausível e deve ser armazenada para que se faça o mapa destacando o cluster encontrado.

   FIM.

2.3.4 Medidas de eficiência

Espera-se que um bom método de detecção de cluster seja sensível o suficiente para detectar um cluster quando este realmente existe. A eficiência do algoritmo será avaliada calculando-se seu poder de detecção, sua sensibilidade (SS) e seu valor de predito positivo (VPP).

O poder de um teste de hipóteses é definido como a probabilidade de que a hipótese nula seja rejeitada quando esta é, de fato, falsa. O poder do método é, então, a probabilidade de que o método detecte um cluster quando este realmente existe. O poder será estimado através de simulações (Monte Carlo, Bootstrap,..), executando o algoritmo N vezes em cenários artificiais, construídos de forma que sabe-se que neles há a presença de um cluster. Assim, deve-se fazer a contagem da quantidade m de vezes em que um cluster foi detectado no mapa em estudo, visando estimar a probabilidade desejada. Desta

forma, o poder sera dado pela proporção, m/N, de detecções em relação ao número total de execuções.

Além do poder, outras medidas bastante utilizadas para avaliação da eficiência do algoritmo de detecção de cluster são a sensibilidade (SS) e o valor predito positivo (VPP). Considere N o total de simulações no estudo. A sensibilidade é definida como a proporção de indivíduos do cluster verdadeiro "capturados" pelo cluster detectado, tal como

$$\mathbf{SS} = \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\{\text{Cluster Detectado}\}^{(q)} \cap \{\text{Cluster Verdadeiro}\}}{\{\text{Cluster Verdadeiro}\}}\right)$$

O valor de predição positiva avalia a proporção de indivíduos do cluster detectado pertencentes ao cluster verdadeiro:

$$\mathbf{VPP} = \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\{\text{Cluster Detectado}\}^{(q)} \cap \{\text{Cluster Verdadeiro}\}}{\{\text{Cluster Detectado}\}^{(q)}}\right)$$

No Capítulo 4 serão apresentados resultados numéricos que atestem a eficiência do algoritmo para o modelo proposto nesse trabalho.

2.3.5 Estatistica Scan baseado em Modelos Lineares Generalizados

Diferentes tipos de dados discretos podem ser analisados por meio de estatística espacial scan de Bernoulli, de Poisson proposto por Kulldorff (1997). Após a publicação de Kulldorff, diversas outras distribuições de probabilidade foram incorporadas ao estudo espacial, tais como a distribuição ordinal Jung *et al.* (2007), exponencial Huang *et al.* (2007) e normal Kulldorff *et al.* (2009). Modelos de Bernoulli e de Poisson estão entre os modelos mais populares para dados discretos em vigilância geográfica de doenças tais como a prevalência, a incidência da doença ou mortalidade. O modelo ordinal é usado para dados categóricos com informações de ordem intrínseca, como por exemplo, o estágio ou de grau do câncer. Os modelos exponencial e weibull foram desenvolvidos para dados de sobrevivência (com ou sem censura) , e o modelo normal para resultado contínuo, como peso dos bebês ao nascerem.

Existem muitas situações que há necessidade de incorporar covariáveis no estudo espacial (Jung, 2009) . Por exemplo, casos de hanseníase em uma região, que é uma doença que pode estar ligada aos fatores socioeconômicos, como a desigualdade de renda,

o crescimento relativo da população, o nível educacional e etc.

Nelder & Wedderburn (1972) propuseram os Modelos Lineares Generalizados (MLG) que são uma extensão dos Modelos Lineares Normais. A distribuição de probabilidade associada à uma variável aleatória Y já não se restringe à Normal, podendo ser qualquer distribuição numa classe designada família exponencial de distribuições.

Dado $\mathbf{Y}(s) = (Y(s_1), Y(s_2), \ldots, Y(s_L))^\top$ um vetor aleatório $L \times 1$ de respostas independentes e $\mathbf{X}(s_l) = (X(s_{l1}), X(s_{l2}), \ldots, X(s_{lk}))^\top$ uma matriz $L \times k$ de valores de covariáveis para a l-ésima localização s_l do mapa S. Denote $Y(s_l) \equiv Y_l$ e $X(s_l) = X_l$. Vamos assumir que a densidade marginal de Y_l pertence à família da exponencial, isto é, sua densidade (ou função de probabilidade) é dada por

$$f(y_l; \boldsymbol{\theta}, \phi) = \exp\{\phi\,[y_l\theta_l - b(\theta_l)] + c(y_l, \phi)\} \tag{2.10}$$

onde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_L)^\top$ é o vetor de parâmetro canônico; ϕ é o parâmetro de precisão, ou de forma equivalente, ϕ^{-1} é o parâmetro de dispersão; $b(.)$ e $c(.)$ são funções específicas que definem a distribuição. Considerando $\ell(\boldsymbol{\theta}) = \log(y_l, \theta_l, \phi)$ a função de log-verossimilhança e as condições usuais de regularidade definidas por

$$\begin{aligned} \mathbb{E}\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_l}\right) &= \mathbf{0} \quad \text{e} \\ \mathbb{E}\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_l^2}\right) &= -\mathbb{E}\left[\left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_l}\right)^2\right], \end{aligned} \tag{2.11}$$

para $\forall l$, obtém a média e a variância de Y_l pelos seguintes resultados

$$\mathbb{E}(Y_l) = \mu_l = b'(\theta_l), \qquad \text{Var}(Y_l) = \phi^{-1} V(\mu_l), \tag{2.12}$$

onde $V(\mu_l) = \partial \mu_l / \partial \theta_l$ é a função de variância. Esse resultado é bastante importante para definir a característica da classe de distribuição no qual a função pertence. Com isso, é possível realizar comparações através das funções de variâncias das distribuições.

O MLG é composto por três elementos:

1. A distribuição de probabilidade a partir da família exponencial.

2. Um indicador linear $\eta_l = X_l \boldsymbol{\gamma}$, onde $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$, $k < L$ é um vetor de parâmetros desconhecidos a serem estimados.

3. A função de ligação g tal que $\mathbb{E}(Y) = \mu = g^{-1}(\eta)$, em que g é uma função monótona e diferenciável.

A função de ligação estabelece a relação entre o preditor linear $\eta_l = X_l\boldsymbol{\gamma}$, que é função das variáveis explicativas e a média μ_l. Para os modelos lineares normais, esta ligação sempre é a identidade, ou seja $\eta_l = \mu_l$. Entretanto, para os modelos lineares generalizados algumas distribuições demandam que a média das observações seja sempre um valor positivo, tornando esta ligação inviável pois pode resultar em valores negativos para o preditor da média.

Entre as funções de ligação, está a função de ligação canônica. Esta consiste na ligação natural entre o preditor e a média, sendo encontrado através de um vetor de *estatísticas suficientes* [2] para o vetor de parâmetros $\boldsymbol{\gamma}$, ambos de mesma dimensão. Uma das vantagens de usarmos ligações canônicas é que as mesmas garantem a concavidade de $\ell(\boldsymbol{\theta})$ e consequentemente muitos resultados assintóticos são obtidos mais facilmente. Por exemplo, a concavidade de $\ell(\boldsymbol{\theta})$ garante a unicidade da estimativa de máxima verossimilhança de $\boldsymbol{\gamma}$, quando essa existe. As ligações canônicas mais comuns são dadas abaixo.

Distribuições	Normal	Poisson	Binomial	Gama
Ligação η	μ	$\log\mu$	$\log\left(\frac{\mu}{1-\mu}\right)$	μ^{-1}

Quando as observações da variável resposta Y é limitada no intervalo (0,1), uma alternativa é modelar através do modelo de regressão Beta. Segundo Ferrari & Cribari-Neto (2004) a função de densidade e probabilidade Beta, não pertence à família exponencial, pois sua função de densidade não pode ser escrita na forma canônica e apresentar um parâmetro de localização μ. Para solucionar tal problema, Ferrari & Cribari-Neto (2004) propôs uma reparametrização para esse modelo, que será apresentada no Capítulo 3.

Usando MLG, Jung (2009) propôs uma estatística scan espacial para diferentes modelos de probabilidade, tais como Bernoulli, Poisson, Normal e Gama que podem ser formulados em uma única estrutura. O modelo geral proposto Jung pode ser escrito como

$$g(\mu_l) = X_l\boldsymbol{\gamma} + \tau \mathtt{I}_{\{s_l \in z\}}\,, \tag{2.13}$$

[2] Uma estatística suficiente para um parâmetro θ é uma estatística que , de certa maneira, capta todas as informações sobre θ contidas na amostra. Mais detalhes em Casella (2002).

onde g é a função de ligação, que é escolhida dependendo da relação entre Y_l e as covariáveis X_{lk}, μ_l é a média da variável resposta, τ é um valor escalar desconhecido e $\boldsymbol{\gamma}$ é o vetor de parâmetros desconhecidos. Este modelo permite-nos comparar a média da variável resposta dos eventos que estão dentro da zona z contra os eventos que acorrem fora deste zona através do parâmetro τ. O parâmetro τ também calcula um risco relativo para indivíduos dentro da zona z em comparação com aqueles que estão fora da zona obtida. Dessa forma, as hipóteses a serem testadas são

$$H_0 : \tau = 0 \tag{2.14}$$

sob hipótese nula, sendo $\mu_l \equiv \mu_{0,l}$ que não depende de τ. E a alternativa de que existe o cluster, no qual a μ_l é maior (ou menor) do que as restantes regiões, que é expresso como

$$H_1 : \tau > 0 \,(\text{ou } \tau < 0)\,. \tag{2.15}$$

em que $\mu_l \equiv \mu_{z,l}$ depende de τ.

A estatística de teste é baseado na razão de verossimilhança, conforme a equação (2.4), e aqui será denotada por

$$\Lambda(z) = \left\{\ell_z(y_l;\mu_{z,l},\phi) - \ell_0(y_l;\mu_{0,l},\phi)\right\} \tag{2.16}$$

onde $\mu_{z,l}$ é a média de Y_l no qual pertence a zona z, $\mu_{0,l}$ representa a média em que Y_l está fora da zona z . No Capítulo 3, será apresentada uma proposta de detecção de cluster espaciais quando a variável resposta segue distribuição Beta, baseado no método Scan Circular.

O vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \phi, \tau)^\top$do modelo apresentado na equação (2.13), é estimado pelo método de máxima verossimilhança (EMV). Conforme Casella (2002), as vantagens do uso deste tipo de estimador são suas propriedades de suficiência, invariância e não ser viesado assintoticamente, entre outras.

O procedimento consiste em encontrar a solução de $U_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \mathbf{0}$ e $U_\phi(\boldsymbol{\theta}) = 0$, que são as funções escores e são obtidas através da equação (2.17). No primeiro passo, serão estimados apenas os parâmetros do vetor $\boldsymbol{\theta}_0 = (\boldsymbol{\gamma}, \phi)^\top$, sob hipótese nula, ou seja quando $\tau = 0$ para o modelo descrito em (2.13). As funções escore do vetor $\boldsymbol{\theta}_0 = (\boldsymbol{\gamma}, \phi)^\top$ é dada

por $U(\boldsymbol{\theta}_0) = (U_{\boldsymbol{\gamma}}, U_{\phi})^{\top}$ e são obtidas como

$$U_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0) = \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma}} \quad \text{e} \quad U_{\phi}(\boldsymbol{\theta}_0) = \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \phi} \tag{2.17}$$

que são as derivadas de ordem 1 da função de log-verossimilhança em que $\ell(\boldsymbol{\theta}_0) = \ell_z(y_l; \mu_{0,l}, \phi)$. A matriz de Informação de Fisher Esperada é

$$\mathfrak{I} = \begin{pmatrix} \mathfrak{I}_{\gamma\gamma} & \mathfrak{I}_{\gamma\phi} \\ \mathfrak{I}_{\phi\gamma} & \mathfrak{I}_{\phi\phi} \end{pmatrix} \tag{2.18}$$

onde

$$\mathfrak{I}_{\gamma\gamma} = -\mathbb{E}\left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^{\top}} \right\}, \quad \mathfrak{I}_{\gamma\phi} = -\mathbb{E}\left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma} \partial \phi} \right\} \quad \text{e} \quad \mathfrak{I}_{\phi\phi} = -\mathbb{E}\left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \phi^2} \right\}.$$

Se $\boldsymbol{\gamma}$ e ϕ são ortogonais, então

$$\mathfrak{I}_{\gamma\phi} = -\mathbb{E}\left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\gamma} \partial \phi} \right\} = \mathbf{0}$$

Isso resulta em uma matriz de informação de Fisher bloco diagonal dada por $\mathfrak{I}_{\theta\theta} = \text{diag}\left\{ \mathfrak{I}_{\gamma\gamma}, \mathfrak{I}_{\phi\phi} \right\}$.

Através dessas quantidades, podemos demonstrar que $\widehat{\boldsymbol{\gamma}}$ e $\widehat{\phi}$ são assintoticamente distribuídos, ou seja

$$\widehat{\boldsymbol{\gamma}} \sim N_k\left(\boldsymbol{\gamma}, \mathfrak{I}_{\gamma\gamma}^{-1}\right), \quad \widehat{\phi} \sim N\left(0, \mathfrak{I}_{\phi\phi}^{-1}\right) \tag{2.19}$$

em que $\mathfrak{I}_{\gamma\gamma}$ é uma matriz não singular.

Normalmente, os EMVs para os parâmetros da regressão em MLG não possuem forma fechada para o vetor de parâmetros. Portanto, é necessária a utilização de métodos interativos para encontrar as estimativas dos parâmetros. Dentre vários, está o método de otimização iterativo de Newton-Raphson para a obtenção dos EMVs. A iteração começa com um valor inicial $\boldsymbol{\theta}^{(0)}$. Para a r-ésima interação, temos

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \mathfrak{I}(\boldsymbol{\theta}^{(r)})^{-1} U(\boldsymbol{\theta}^{(r)}). \tag{2.20}$$

onde $r \in \mathbb{N}$.

O critério de parada será: se $|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}| < \varepsilon$ $(\varepsilon > 0)$, então $\boldsymbol{\theta}^{(r+1)}$ é o valor

desejado $\widehat{\boldsymbol{\theta}}$. Normalmente na literatura o valor de ε é pequeno, por exemplo $\varepsilon = 0.001$. Essa escolha, é para garantir que os valores das estimativas se aproximem de forma precisa dos parâmetros verdadeiros, ou seja $\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}$.

A estimativa para o parâmetro τ é obtida pelo mesmo procedimento descrito anteriormente, caso $U_\tau = 0$ não tiver forma fechada. Nesse passo, é considerando as estimativas $\widehat{\boldsymbol{\gamma}}$ e $\widehat{\phi}$ fixas, obtidas através do modelo sob hipótese nula (Jung, 2009) . Portanto, tomando $\ell(\boldsymbol{\theta}_1) = \ell_z(y_l; \mu_{z,l}, \phi)$, obtemos a função escore e a informação de Fisher, respectivamente

$$U_\tau = \frac{\partial \ell(\boldsymbol{\theta}_1)}{\partial \tau}, \qquad \mathfrak{I}_{\tau\tau} = -\mathbb{E}\left\{ \frac{\partial^2 \ell(\boldsymbol{\theta}_1)}{\partial \tau^2} \right\} \tag{2.21}$$

onde $\boldsymbol{\theta}_1 = (\widehat{\boldsymbol{\gamma}}^\top, \widehat{\phi}, \tau)^\top$. Desse modo, encontrada a esperança e a informação de fisher, logo para a *r*-ésima interação obtemos a estimativa para τ através do método de NR, como descrito na equação (2.20).

Capítulo 3

A Estatística Scan para Modelos de Regressão Beta

No Capítulo 2 foram discutidos alguns métodos para detecção de clusters espaciais, dentre esses, a Estatística Scan Circular de Kulldorff é a que está sendo mais utilizada atualmente, principalmente no ramo da epidemiologia. No entanto, ao decorrer dos anos diversos Modelos Probabilísticos, tanto discretos, como contínuos, foram incorporados a esse método. Dessa forma, também surgiu a necessidade de incorporar variáveis explicativas (covariáveis) ao estudo de clusters espaciais. Nesses estudos estão inclusos os modelos lineares generalizados. Nesse Capítulo , o Modelo de Regressão Beta será desenvolvido para solucionar problemas de detecção de cluster espaciais em proporções, taxas ou números índices.

3.1 O modelo de regressão Beta

Em alguns dos modelos probabilísticos citados na Seção 2.3 para construção do Scan Espacial, supõe-se que o suporte da variável aleatória é ilimitado. No entanto existem situações na qual a variável de interesse é continuamente limitada no intervalo (a,b), onde a e b são escalares conhecidos com $a < b$. Uma particular situação ocorre quando a = 0 e b = 1 de modo que a variável aleatória assume valores em (0, 1), como é o caso de taxas, proporções e números índices. Para esse tipo de dados, uma modelagem via distribuição Beta é mais adequada.

Definição 1. *A distribuição Beta é uma distribuição de probabilidade contínua, com dois*

parâmetros p e q cuja função de densidade para valores $0 < y < 1$ *é*

$$f_Y(y;p,q) = \frac{1}{B(p,q)} y^{p-1}(1-y)^{q-1}, \quad p,q > 0, \tag{3.1}$$

onde $B(p,q)$ *denota a função Beta,*

$$B(p,q) = \int_0^1 y^{p-1}(1-y)^{q-1} dy.$$

A função Beta está relacionada à função gama por meio da seguinte identidade:

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{3.2}$$

O cálculo dos momentos da distribuição Beta é bastante simples, devido a forma da função densidade definida em (3.1) e utilizando algumas propriedades da função gama. Assim, para o *n*-ésimo momento, temos

$$\mathbb{E}(Y^n) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^1 y^{p+n-1}(1-y)^{q-1} dy = \frac{\Gamma(p+q)\Gamma(p+n)}{\Gamma(p+q+n)\Gamma(p)}.$$

Com a função de momentos em mãos podemos encontrar o valor esperado e a variância.

$$\mathbb{E}(Y) = \frac{\Gamma(p+q)\Gamma(p+1)}{\Gamma(p+q+1)\Gamma(p)} = \frac{\Gamma(p+q)p\Gamma(p)}{(p+q)\Gamma(p+q)\Gamma(p)} = \frac{p}{p+q}$$

Para o cálculo da variância necessitamos de $\mathbb{E}\left(Y^2\right)$, que é dado por

$$\mathbb{E}\left(Y^2\right) = \frac{\Gamma(p+q)\Gamma(p+2)}{\Gamma(p+q+2)\Gamma(p)} = \frac{\Gamma(p+q)(p+1)p\Gamma(p)}{(p+q+1)(p+q)\Gamma(p+q)\Gamma(p)} = \frac{(p+1)p}{(p+q+1)(p+q)}.$$

Portanto, temos que

$$\mathrm{Var}(Y) = \mathbb{E}\left(Y^2\right) - \mathbb{E}^2(Y) = \frac{(p+1)p}{(p+q+1)(p+q)} - \frac{p^2}{(p+q)^2} = \frac{pq}{(p+q+1)(p+q)^2}.$$

Ferrari & Cribari-Neto (2004) propuseram a classe de modelos de regressão Beta baseada na suposição de que a variável de interesse (y) segue distribuição Beta e consideraram uma parametrização alternativa para a função de densidade Beta que permite a modelagem do parâmetro de locação e escala. Estes parâmetros podem ser interpretados em termos da média das observações, que é modelada usando um preditor linear que rela-

ciona a resposta média a covariáveis e parâmetros desconhecidos através de uma função de ligação, como acontece nos modelos lineares generalizados. Fazendo $\mu = p/p+q$ e $\phi = p+q$, onde $0 < \mu < 1$ e $\phi > 0$. Assim, como segue o resultado em (2.12), obtemos a esperança e a variância reparametrizada como sendo

$$\mathbb{E}(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \frac{\text{Var}(\mu)}{1+\phi}$$

onde $\text{Var}(\mu) = \mu(1-\mu)$, de modo que μ é a média da variável resposta e ϕ pode ser interpretado como um parâmetro de precisão, no sentido de que, para μ fixo, quanto maior for o valor de ϕ menor é a variância de Y.

A função densidade para a variável resposta Y pode ser escrita como

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad y \in (0,1). \tag{3.3}$$

onde $0 < \mu < 1$ e $\phi > 0$. Através da escolha de diferentes valores para os parâmetros (μ,ϕ), podem ser obtidas diferentes formas para a densidade (3.3) no intervalo unitário padrão. A Figura 3.1 apresenta algumas densidades Beta juntamente com os valores de μ para cada ϕ correspondentes. Quando $\mu = 1/2$ a curva da densidade assume uma forma simétrica e para $\mu > 1/2$, há assimetria à direita e, de forma análoga, quando $\mu < 1/2$, existe assimetria à esquerda. Ainda se pode notar que quando aumenta o valor de ϕ, diminui a variância de Y para cada valor de μ.

Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes com distribuição Beta, cuja densidade assume conforme a equação (3.3). O modelo de regressão Beta é definido supondo que a média de Y_i satisfaz uma relação funcional da forma

$$g(\mu_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\gamma_j = x_i^\top \boldsymbol{\gamma} \quad \text{para} \quad i = 1,2,\ldots,n \quad \text{e} \quad k < n \tag{3.4}$$

em que $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)^\top$ é um vetor de parâmetros de regressão desconhecidos ($\boldsymbol{\gamma} \in \mathbb{R}^k$); as covariáveis $x_{i1}, x_{i2}, \ldots, x_{ik}$ são assumidas fixas e conhecidas; η_i é o preditor linear do modelo (*i.e.*, $\eta_i = \gamma_1 x_{i1} + \gamma_2 x_{i2} + \ldots + \gamma_k x_{ik}$ usualmente $x_{i1} = 1$ para todo i de modo que o modelo tenha um intercepto). A função de ligação $g : (0,1) \mapsto \mathbb{R}$ é estritamente monótona e duas vezes diferenciável . Entre as funções de ligação mais utilizadas

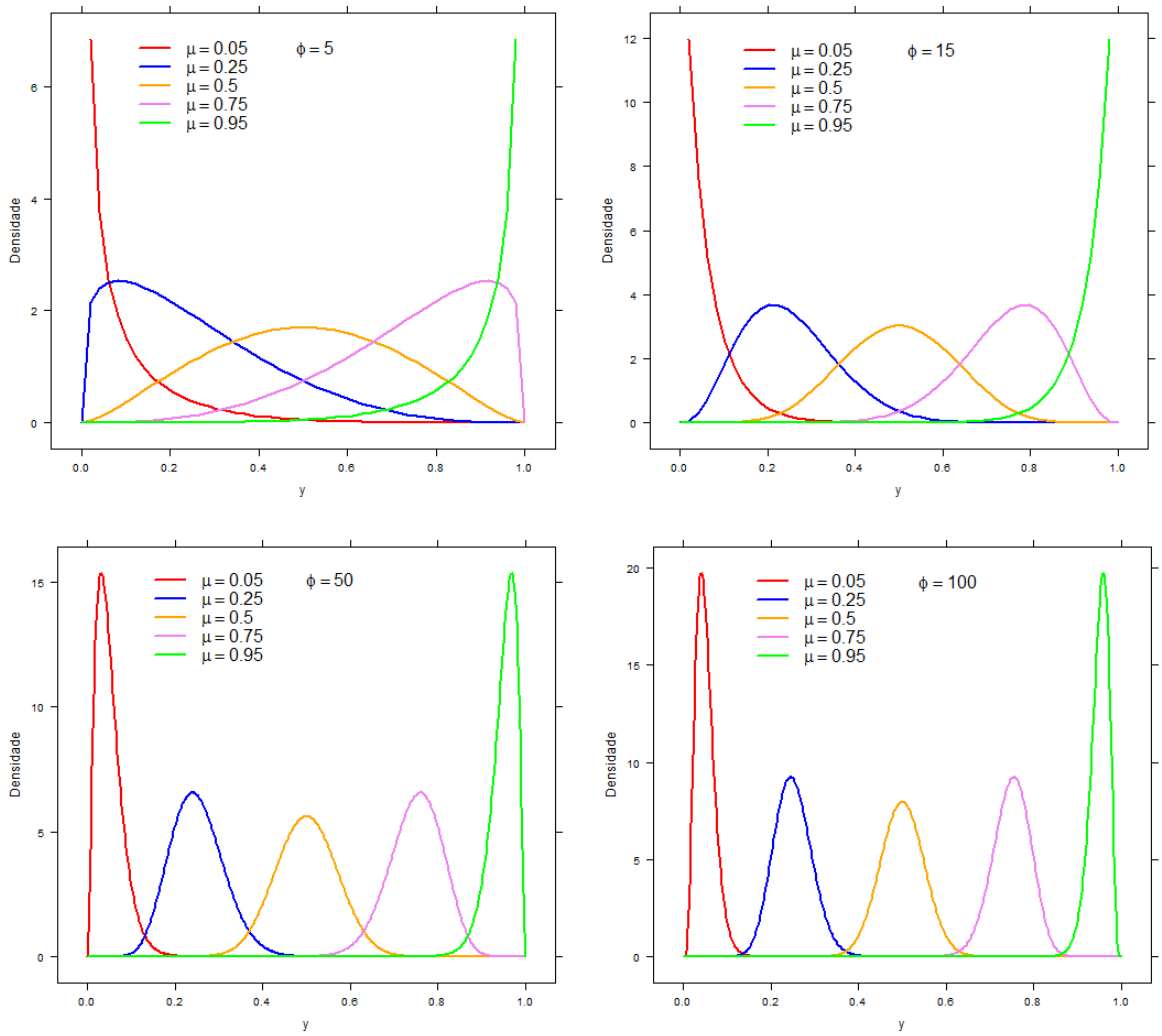


Figura 3.1: Densidades Beta para diferentes valores de (μ, ϕ).

no modelo de regressão Beta estão a *logit* $g(\mu) = \log(\mu/1-\mu)$, a *probit* $g(\mu) = \Phi^{-1}(\mu)$, sendo Φ^{-1} a função da distribuição acumulada da normal padrão, e a *log-log* complementar $g(\mu) = \log(-\log(\mu))$. Uma discussão detalhada sobre essas e outras funções de ligação pode ser encontrada em McCullagh & Nelder (1989).

Daqui por diante, será usada a função de ligação *logit* para o modelo proposto. Geralmente, no modelo *logit* a média pode ser escrita como

$$\mu_i = \frac{e^{x_i^\top \gamma}}{1 + e^{x_i^\top \gamma}} .$$

3.2 O Modelo de Regressão β-Scan

Suponha que existam L localizações s_l e seja $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))^\top$, onde $Y_l \equiv Y(s_l)$ é a variável aleatória continua no intervalo (0,1). Especificamente assumimos que Y_l segue uma Distribuição Beta, denotada por $Y_l \sim Beta(\mu_{.l}, \phi)$, com função densidade dada pela equação (3.3) em termos da média $\mu_{.l}$ e um parâmetro de precisão ϕ.

Seja Z um cluster potencial, em nossa proposta, o processo espacial $Y_1, \ldots, Y_L$ é modelado por $\boldsymbol{\beta}$-SCAN$(\mu_{.l}, \phi_l, \tau.)$, $l = 1, 2, \ldots, L$, o qual assume

$$\log\left(\frac{\mu_l}{1-\mu_l}\right) = x_l\boldsymbol{\gamma} + \tau \mathtt{I}_{\{s_l \in z\}}$$

onde $x_l = (x_{l1}, \ldots, x_{lk})^\top$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^\top$ e $\mathtt{I}$ é a função indicadora. Então,

$$\mu_l \equiv \mu_{0,l} = \frac{\exp\{x_l\boldsymbol{\gamma}\}}{1+\exp\{x_l\boldsymbol{\gamma}\}} \quad \text{se} \quad s_l \notin z \tag{3.5}$$

$$\mu_l \equiv \mu_{z,l} = \frac{\exp\{x_l\boldsymbol{\gamma}+\tau\}}{1+\exp\{x_l\boldsymbol{\gamma}+\tau\}} \quad \text{se} \quad s_l \in z. \tag{3.6}$$

De modo que $g(\mu) = \log(\mu/1-\mu)$ é a função de ligação do modelo, $\boldsymbol{\gamma}$ é o vetor de parâmetros fixos desconhecidos. Nota-se que

$$e^\tau = \frac{\mu_{0,l}(1-\mu_{z,l})}{\mu_{z,l}(1-\mu_{0,l})}. \tag{3.7}$$

Então, na escala logarítmica, podemos interpretar τ como uma medida de razão de chance ajustada por covariáveis para as observações $y_l \in z$ em comparação com $y_l's$ que não pertencem a z.

3.2.1 Estimação dos parâmetros

A estimação conjunta dos parâmetros no modelo de regressão Beta é realizada por máxima verossimilhança. Para tanto, utiliza-se o logaritmo da função de verossimilhança dada por

$$\ell_0(\boldsymbol{\gamma}, \phi) = \sum_{s_l \in S} \ell(\mu_{0,l}, \phi)$$

onde

$$
\begin{aligned}
&\ell_0(\boldsymbol{\gamma},\phi,0)\\
&=\sum_{s_l\notin z}\ell(\mu_{0,l},\phi)+\sum_{s_l\in z}\ell(\mu_{0,l},\phi)\\
&=\sum_{s_l\notin z}\big\{\log\Gamma(\phi)-\log\Gamma(\mu_{0,l}\phi)-\log\Gamma((1-\mu_{0,l})\phi)+(\mu_{0,l}\phi-1)\log y_i\\
&\qquad+[(1-\mu_{0,l})\phi-1]\log(1-y_i)\big\}+\sum_{s_l\in z}\big\{\log\Gamma(\phi)-\log\Gamma(\mu_{0,l}\phi)\\
&\qquad-\log\Gamma((1-\mu_{0,l})\phi)+(\mu_{0,l}\phi-1)\log y_i+[(1-\mu_{0,l})\phi-1]\log(1-y_i)\big\}
\end{aligned}
\tag{3.8}
$$

com $\mu_{0,l}$ definido segundo a equação(3.5). Para alguma área $s_l \in Z$, logo a função de verossimilhança será da forma

$$
\begin{aligned}
&\ell_Z(\boldsymbol{\gamma},\phi,\tau)\\
&=\sum_{s_l\notin z}\ell(\mu_{0,l},\phi)+\sum_{s_l\in z}\ell(\mu_{z,l},\phi)\\
&=\sum_{s_l\notin z}\big\{\log\Gamma(\phi)-\log\Gamma(\mu_{0,l}\phi)-\log\Gamma((1-\mu_{0,l})\phi)+(\mu_{0,l}\phi-1)\log y_i\\
&\qquad+[(1-\mu_{0,l})\phi-1]\log(1-y_i)\big\}+\sum_{s_l\in z}\big\{\log\Gamma(\phi)-\log\Gamma(\mu_{z,l}\phi)\\
&\qquad-\log\Gamma((1-\mu_{z,l})\phi)+(\mu_{z,l}\phi-1)\log y_i+[(1-\mu_{z,l})\phi-1]\log(1-y_i)\big\}
\end{aligned}
\tag{3.9}
$$

com $\mu_{z,l}$ dado pela equação (3.6) .

O vetor escore, obtido a partir das primeiras derivadas do logaritmo da função de verossimilhança (Equação 3.8) com relação aos parâmetros $\boldsymbol{\theta} = (\boldsymbol{\gamma},\phi)$, é dado por $\big(U_{\boldsymbol{\gamma}}(\boldsymbol{\theta}),U_{\phi}(\boldsymbol{\theta})\big)^{\top}$ onde

$$
\begin{aligned}
U_{\boldsymbol{\gamma}}(\boldsymbol{\theta})=\frac{\partial\ell_0(\boldsymbol{\gamma},\phi,0)}{\partial\gamma_j} &= \sum_{s_l\in S}\frac{\partial\ell_0(\boldsymbol{\gamma},\phi,0)}{\partial\mu_{0,l}}\frac{\partial\mu_{0,l}}{\partial\eta_{0,l}}\frac{\partial\eta_{0,l}}{\partial\gamma_j}\\
&= \phi\sum_{s_l\in S}\left(y_l^*-\mu_{0,l}^*\right)\frac{1}{g'(\mu_{0,l})}x_{lj}
\end{aligned}
$$

sendo que

$$\frac{\partial \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \mu_{0,l}} = \phi\left(y_l^* - \mu_{0,l}^*\right), \qquad \frac{\partial \mu_{0,l}}{\partial \eta_{0,l}} = \frac{1}{g'(\mu_{0,l})} \quad \text{e} \quad \frac{\partial \eta_{0,l}}{\partial \gamma_j} = x_{ij}$$

tal que $y_l^* = \log(y_l/1-y_l)$ e $\mu_{0,l}^* = \psi(\mu_{0,l}\phi) - \psi((1-\mu_{0,l})\phi)$, ψ é derivada da função $\log\Gamma(.)$.

Seja $\mathbf{y}^* = (y_1^*,\ldots,y_L^*)'$, $\boldsymbol{\mu}_0^* = (\mu_{01}^*,\ldots,\mu_{0L}^*)'$, $T = \text{diag}(1/g'(\mu_{01}),\ldots,1/g'(\mu_{0L}))$ e a matriz X constituída pelos elementos x_{lj}, para $j = 1,\ldots,k$ e $l = 1,\ldots,L$. Portanto, na forma matricial, para cada elemento γ_j temos

$$U\boldsymbol{\gamma}(\boldsymbol{\theta}) = \phi X^\top T(\mathbf{y}^* - \boldsymbol{\mu}_0^*). \tag{3.10}$$

De forma similar, pode-se mostrar que a função escore para o parâmetro ϕ pode ser escrita como

$$U_\phi(\boldsymbol{\theta}) = \frac{\partial \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \phi} = \sum_{s_l \in S}\left[\mu_{0,l}(y_l^* - \mu_{0,l}^*) + \log(1-y_l) - \psi((1-\mu_{0,l})\phi) + \psi(\phi)\right]. \tag{3.11}$$

Através do sistema

$$\begin{cases} U\boldsymbol{\gamma}(\boldsymbol{\theta}) = 0 \\ U_\phi(\boldsymbol{\theta}) = 0 \end{cases}$$

obtêm-se os Estimadores de Máxima Verossimilhança $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}},\widehat{\phi})^\top$. Como esse sistema linear não possui forma fechada, $\widehat{\boldsymbol{\theta}}$ deve ser calculado iterativamente através do Método de Newton-Raphson. Para mais informações, ver Nocedal & Wright (1999).

Para determinar a variabilidade das estimativas dos parâmetros do modelo de regressão Beta, Ferrari & Cribari-Neto (2004) obtiveram uma expressão para a matriz de informação de Fisher. As segundas derivadas da função de log verossimilhança com relação aos parâmetros desconhecidos resultam na matriz de informação de Fisher observada, que é definida por

$$J = \begin{pmatrix} J_{\gamma\gamma} & J_{\gamma\phi} \\ J_{\phi\gamma} & J_{\phi\phi} \end{pmatrix} \tag{3.12}$$

com cada elemento obtido através da segunda derivada da função de log verossimilhança (3.8), ou seja

$$\begin{aligned} J_{\gamma\gamma} &= \frac{\partial^2 \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \gamma_j^2} \\ &= \sum_{s_l \in S} \left[\frac{\partial \ell_0^2(\boldsymbol{\gamma},\phi,0)}{\partial \mu_{0,l}^2} \frac{\partial \mu_{0,l}}{\partial \eta_{0,l}} \frac{\partial \eta_{0,l}}{\partial \gamma_j} + \frac{\partial \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \mu_{0,l}} \frac{\partial}{\partial \mu_{0,l}} \left(\frac{\partial \mu_{0,l}}{\partial \eta_{0,l}} \frac{\partial \eta_{0,l}}{\partial \gamma_j} \right) \right] \frac{\partial \mu_{0,l}}{\partial \eta_{0,l}} \frac{\partial \eta_{0,l}}{\partial \gamma_j} \\ &= \phi \sum_{s_l \in S} \left[-w_l x_{lj}^2 - \left(y_l^* - \mu_{0,l}^* \right) \frac{g''(\mu_{0,l})}{(g'(\mu_{0,l}))^2} x_{lj} \right] \end{aligned}$$

onde,

$$\frac{\partial \ell_0^2(\boldsymbol{\gamma},\phi,0)}{\partial \mu_{0,l}^2} = -\phi[\psi'(\mu_{0,l}\phi) + \psi'((1-\mu_{0,l})\phi)]$$

em que $w_l = \phi[\psi'(\mu_{0,l}\phi) + \psi'((1-\mu_{0,l})\phi)]\frac{1}{g'(\mu_{0,l})^2}$ e $\psi'(.)$ é a derivada da função $\psi(.)$.

Podemos obter ainda,

$$\begin{aligned} J_{\phi\gamma} = \frac{\partial^2 \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \gamma_j \phi} &= \sum_{s_l \in S} \frac{\partial}{\partial \phi} \left\{ \frac{\partial \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \mu_{0,l}} \frac{\partial \mu_{0,l}}{\partial \eta_{0,l}} \frac{\partial \eta_{0,l}}{\partial \gamma_j} \right\} \\ &= \sum_{s_l \in S} \left\{ \left[\left(y_i^* - \mu_{0,l}^* \right) - c_l \right] \frac{1}{g'(\mu_{0,l})} x_{ij} \right\}, \end{aligned}$$

onde $c_l = \phi[\psi'(\mu_{0,l}\phi)\mu_{0,l} - \psi'((1-\mu_{0,l})\phi)(1-\mu_{0,l})]$.

E finalmente obtemos

$$\begin{aligned} J_{\phi\phi} = \frac{\partial^2 \ell_0(\boldsymbol{\gamma},\phi,0)}{\partial \phi^2} &= \sum_{s_l \in S} \frac{\partial}{\partial \phi} \left[\mu_{0,l}(y_i^* - \mu_{0,l}^*) + \log(1-y_i) - \psi((1-\mu_{0,l})\phi) + \psi(\phi) \right] \\ &= \sum_{s_l \in S} \left[-\mu_{0,l}^2 \psi'(\mu_{0,l}\phi) - (1-\mu_{0,l})^2 \psi'((1-\mu_{0,l})\phi) + \psi'(\phi) \right] \\ &= \sum_{s_l \in S} d_l \,. \end{aligned}$$

A Matriz de Informação de Fisher para cada elemento da matriz J (3.12), é obtida através da seguinte expressão

$$\mathfrak{I}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{\partial^2 \ell_0(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \middle| \boldsymbol{\theta} \right]. \tag{3.13}$$

com a restrição $\mathbb{E}(y^*) = \mu_{0,l}^*$ (Ver Ferrari & Cribari-Neto (2004)).

Assim, fazendo $W = \text{diag}(w_1, \ldots, w_L)$; $C = (c_1, \ldots, c_L)$ e $D = \text{diag}(d_1, \ldots, d_L)$.

Tem-se que a matriz de informação de Fisher Esperada (3.13) dada por

$$\mathfrak{I} = \begin{pmatrix} \mathfrak{I}_{\gamma\gamma} & \mathfrak{I}_{\gamma\phi} \\ \mathfrak{I}_{\phi\gamma} & \mathfrak{I}_{\phi\phi} \end{pmatrix} = \begin{pmatrix} \phi X^\top W X & X^\top T C \\ C^\top T^\top X & \mathrm{tr}(D) \end{pmatrix} \tag{3.14}$$

O resultado (3.14) é similar ao obtido em Ferrari & Cribari-Neto (2004) após uma simples manipulação de índices. Agora, a obtenção de $\widehat{\boldsymbol{\theta}}$ via o algoritmo Newton-Rapshon (**NR**) para iteração $r+1$, $r \in \mathbb{N}$, é

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \mathfrak{I}(\boldsymbol{\theta}^{(r)})^{-1} U(\boldsymbol{\theta}^{(r)})$$

onde $U(\boldsymbol{\theta})$ é a função escore e $\mathfrak{I}(\boldsymbol{\theta})$ é a matriz de Informação de Fisher.

Ferrari & Cribari-Neto (2004) ressaltam que, ao contrário do caso dos modelos lineares generalizados (McCullagh & Nelder, 1989), os parâmetros γ e ϕ não são ortogonais, pois $\mathfrak{I}_{\gamma\phi} = \mathfrak{I}_{\phi\gamma} = X^\top T C \neq 0$. A matriz de variância assintótica dos estimadores de máxima verossimilhança dos parâmetros do modelo de regressão Beta é dada pela inversa de $\mathfrak{I}$ (i.e $\mathfrak{I}^{-1}$). A significância estatística da regressão é avaliada usando a distribuição assintótica (Ferrari & Cribari-Neto, 2004), $\widehat{\theta} \sim \mathcal{N}(\theta, \mathfrak{I}^{-1})$.

Encontrado o estimador $\widehat{\boldsymbol{\theta}}$, o passo a seguir é estimar τ. A função escore é obtida similarmente como para $\boldsymbol{\theta}$. Porém, agora utilizaremos a função de log verossimilhança da equação (3.9). Então:

$$\begin{aligned} U_\tau = \frac{\partial \ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \tau)}{\partial \tau} &= \sum_{s_l \in z} \frac{\partial \ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \tau)}{\partial \mu_{z,l}} \frac{\partial \mu_{z,l}}{\partial \eta_{z,l}} \frac{\partial \eta_{z,l}}{\partial \tau} \\ &= \phi \sum_{s_l \in z} \left(y_i^* - \mu_{z,l}^*\right) \frac{1}{g'(\mu_{z,l})}, \end{aligned} \tag{3.15}$$

e a Informação de Fisher observada:

$$
\begin{aligned}
J_\tau &= \frac{\partial^2 \ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \tau)}{\partial \tau^2} \\
&= \sum_{s_l \in z} \left[\frac{\partial \ell_z^2(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \tau)}{\partial \mu_{z,l}^2} \frac{\partial \mu_{z,l}}{\partial \eta_{z,l}} \frac{\partial \eta_{z,l}}{\partial \tau} + \frac{\partial \ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \tau)}{\partial \mu_{z,l}} \frac{\partial}{\partial \mu_{z,l}} \left(\frac{\partial \mu_{z,l}}{\partial \eta_{z,l}} \frac{\partial \eta_{z,l}}{\partial \tau} \right) \right] \frac{\partial \mu_{z,l}}{\partial \eta_{z,l}} \frac{\partial \eta_{z,l}}{\partial \tau} \\
&= -\widehat{\phi}^2 \sum_{s_l \in z} \left[\widehat{\phi} \left(\psi'(\mu_{z,l}\widehat{\phi}) + \psi'((1-\mu_{z,l})\widehat{\phi}) \right) d_{z,l}^2 - (y_l^* - \mu_{z,l}^*)(1-2\mu_{z,l}) d_{z,l} / \widehat{\phi} \right]
\end{aligned}
\tag{3.16}
$$

onde $\mu_{z,l}^* = \psi(\mu_{z,l}\widehat{\phi}) - \psi((1-\mu_{z,l})\widehat{\phi})$, $\mu_{z,l} = \exp\{x_l\widehat{\boldsymbol{\gamma}} + \tau\} / \left(1 + \exp\{x_l\widehat{\boldsymbol{\gamma}} + \tau\}\right)$ e $d_{z,l} = \mu_{z,l}(1-\mu_{z,l})$.

O algoritmo de Newton-Rapshon para a k-ésima iteração

$$
\tau^{(k+1)} = \tau^{(k)} + J_\tau(\tau^{(k)})^{-1} U_\tau(\tau^{(k)}),
$$

onde U_τ é a função score e $\mathfrak{I}_\tau$ a Informação de Fisher, respectivamente.

3.2.2 Estatística de Teste e Estimação do Cluster

Para detecção e significância do cluster usamos o teste de hipóteses

$$
H_0 : \tau = 0 \quad \text{versus} \quad H_1 : \tau > 0
$$

para alguma zona $Z \in S$, e o logaritmo da razão de verossimilhança como estatística de teste . Agora, seja $\widehat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}, \hat{\phi})$ o Estimador de Máxima Verossimilhança para os parâmetros da regressão sob a hipótese nula e $\hat{\tau}$ o estimador de Máxima Verossimilhança de τ sob a hipótese alternativa. Em nosso teste, sob H_1, o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \phi)$ é fixado usando as estimativas sob o modelo nulo. Neste caso os coeficientes das covariáveis que são utilizadas para o ajuste da média permanecem iguais mesmo para conjuntos Z diferentes. A estatística $\boldsymbol{\beta}$-SCAN utilizada é

$$
\Lambda = \max_{z \in Z} \Lambda_z .
$$

Onde,

$$
\widehat{\Lambda_z} = \left\{ \ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \widehat{\tau}) - \ell_0(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, 0) \right\}
$$

com $\ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \widehat{\tau})$ representando a Função de Log- verossimilhança (equação 3.9) sob H_1 para um particular conjunto de localizações espaciais z e $\ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \widehat{\tau})$ é a Função de Log-verossimilhança (equação 3.8) sob H_0. Se $\widehat{\tau} > 0$, podemos mostrar que

$$\begin{aligned}
\widehat{\Lambda_z} &= \left\{ \ell_z(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, \widehat{\tau}) - \ell_0(\widehat{\boldsymbol{\gamma}}, \widehat{\phi}, 0) \right\} \\
&= \sum_{s_l \notin z} \ell(\mu_{0,l}, \phi) + \sum_{s_l \in z} \ell(\mu_{z,l}, \phi) - \sum_{s_l \notin z} \ell(\mu_{0,l}, \phi) - \sum_{s_l \in z} \ell(\mu_{0,l}, \phi) \\
&= \sum_{s_l \in z} \left\{ \ell(\mu_{z,l}, \phi) - \ell(\mu_{0,l}, \phi) \right\} .
\end{aligned}$$

Substituindo as expressões (3.8) e (3.9), obtemos

$$\widehat{\Lambda_z} = \begin{cases} \Sigma_{s_l \in z} \left[\log \left(\frac{\Gamma(\mu_{0,l}\phi)\Gamma((1-\mu_{0,l})\phi)}{\Gamma(\mu_{z,l}\phi)\Gamma((1-\mu_{z,l})\phi)} \right) + (\mu_{z,l} - \mu_{0,l})\phi \log \left(\frac{y_l}{1-y_l} \right) \right] \\ 0, \text{ caso contrário.} \end{cases} \tag{3.17}$$

Portanto, um estimador para o cluster z é

$$\widehat{z} = \arg \left(\max(\widehat{\Lambda_z}) \right) . \tag{3.18}$$

Na seção a seguir, será mostrado como encontrar o estimador $\widehat{z}$.

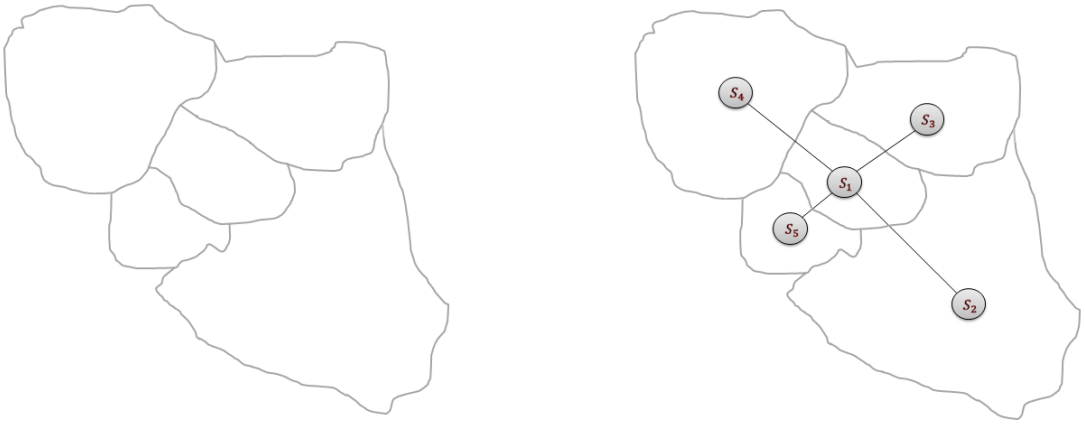
3.2.3 Ilustrando a estatística Scan Circular

No Capítulo 2, foi mostrado o funcionamento interno do Algoritmo Scan Circular de Kulldorff. Nessa seção vamos ilustrar esse funcionamento através de um exemplo ilustrativo.

Vamos supor, como ilustração, um mapa $S = \{s_1, s_2, s_3, s_4, s_5\}$ com 5 regiões, como na Figura 3.2. Cada região s_l ($l = 1, \ldots, 5$) temos $y_l \sim beta(\mu_{0,l}, \phi)$. O funcionamento do interno da Scan Circular para esse exemplo é mostrado na Figura 3.3.

Suponha que a matriz de distâncias foi ordenada, e fixando o centróide s_1 obteve o vetor $\{s_1, s_5, s_4, s_3, s_2\}$, ou seja, s_5 é o primeiro centróide mais próximo de s_1 e s_4 é o segundo mais próximo e assim sucessivamente como mostra a Figura3.3 (a).

Portanto, obtemos os vetores das zonas candidatas a cluster:



(a) (b)

Figura 3.2: Exemplo: (a) Mapa divido em 5 regiões; (b) Centroides de cada região

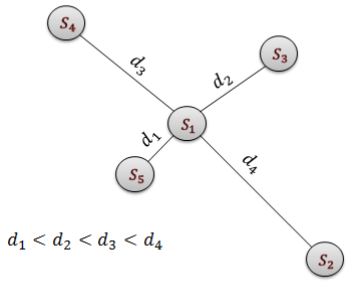
$$\begin{cases} z_{1_1} = & (1,0,0,0,0); \\ z_{1_2} = & (1,0,0,0,1); \\ z_{1_3} = & (1,0,0,1,1); \\ z_{1_4} = & (1,0,1,1,1); \\ z_{1_5} = & (1,1,1,1,1). \end{cases}$$

Observe que para cada vetor z_{l_i} as posições recebe o valor 1 no índice do vizinho mais próximo de s_l em sua posição original no espaço. Esta representação é única a menos do cluster z_{l_L} que surge L vezes diferenciado apenas pelo seu centróide. Para cada $z_{1_i}, i = 1,2,\ldots,5$ calcula-se a estatística $\Lambda_{z_{1_i}}$ (Figura 3.3) obtida através da equação(3.2.2). Dessa forma, ao terminar o processo de varredura do mapa com referência ao centróide s_1, o próximo passo é fixar outro centróide, por exemplo s_5, e realizar o mesmo processo descrito anteriormente, e assim por diante.

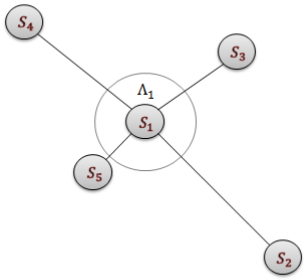
A classe $\widetilde{Z} = \{\{1\};\{1,5\};\{1,5,4\};\{1,5,4,3\};\{1,5,4,3,2\};\ \{2\};\{2,4\},\ldots\}$ é o conjunto de todas as zonas circulares z_{l_i} . Suponha que

$$\Lambda_{z_{1_3}} = \max\{\Lambda_{z_{1_1}}, \Lambda_{z_{1_2}}, \Lambda_{z_{1_3}}, \Lambda_{z_{1_4}}, \Lambda_{z_{1_5}}, \Lambda_{z_{2_1}}, \Lambda_{z_{2_2}}, \Lambda_{z_{2_3}}, \ldots\}$$

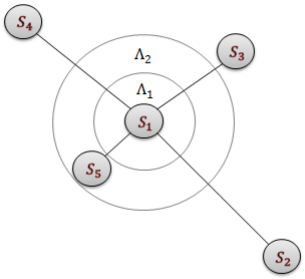
é a estatística mais verossímil. Logo, $\{1\}$ é o centro do cluster detectado $\widehat{z} = \{1,3,5\}$ que corresponde as regiões s_1, s_3 e s_5, respectivamente, como mostra a Figura 3.4.



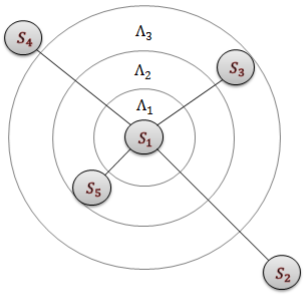
(a) Ordenando as distâncias



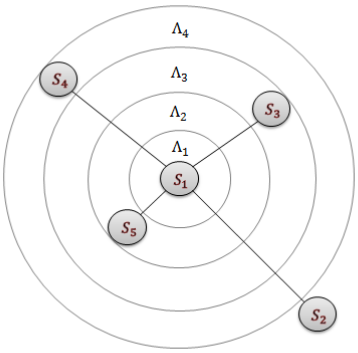
(b) Candidato a cluster z_{1_1}



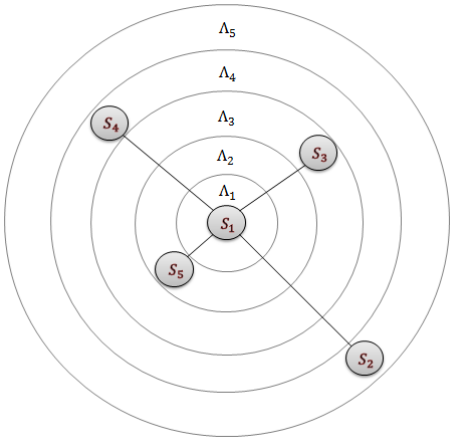
(c) Candidato a cluster z_{1_2}



(d) Candidato a cluster z_{1_3}

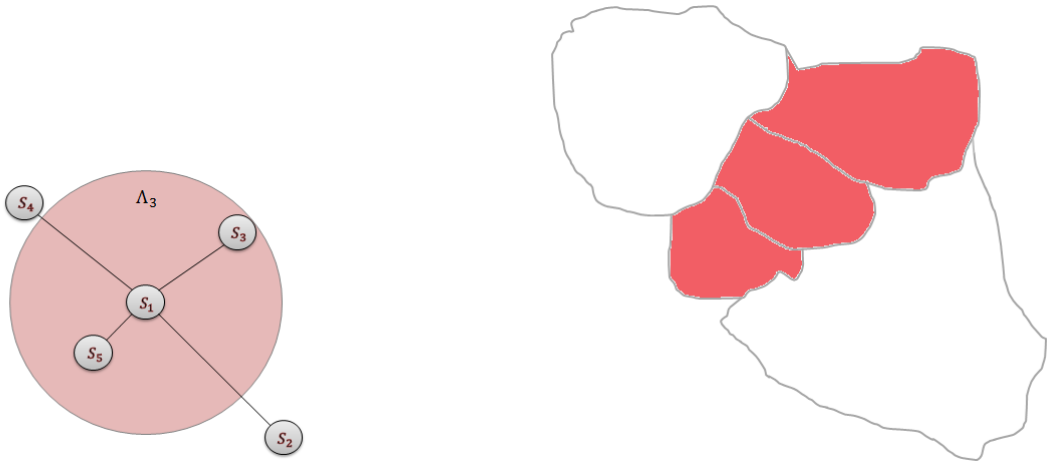


(e) Candidato a cluster z_{1_4}



(f) Candidato a cluster z_{1_5}

Figura 3.3: Funcionamento da Estatística Scan Circular de Kulldorff



(a) (b)

Figura 3.4: Detecção do Cluster correspondente às regiões s_1,s_3 e s_5.

3.2.4 Bootstrap para o valor-p da Estatística Espacial β-SCAN

Existem diversas técnicas de reamostragem que visam estimar parâmetros de uma distribuição de interesse, dentre os mais usados, está o Método de *Bootstrap* proposto por Efron (1979). É um método de reamostragem que se baseia na construção de amostragem empírica de uma estatística de interesse. A distribuição empírica de uma estatística gerada pelo *Bootstrap* tem aproximadamente a mesma forma e amplitude da distribuição amostral que estatística.

A amostra original representa a população da qual foi retirada. Portanto, tratando a amostra como se ela fosse a população, realizando sucessivas amostragens com reposição. A partir daí, torna-se possível estimar características da população, tais como a média, variância, percentis e etc.

O Algoritmo Bootstrap Newton-Rapshon para avaliar a Estatística Λ obtida na equação (3.17), é conforme os passos seguir:

- Algoritmo Bootstrap-Newton-Rapshon para Λ.

  INÍCIO

  1. Baseado nos dado reais $\mathbf{y} = (y_1, ..., y_L)$ e matriz de covariável $\mathbf{X}$, use o algoritmo Newton-Rapshon e compute $\hat{\boldsymbol{\theta}}$ e $\widehat{\tau}$. Derive o valor observado de Λ e denote por $\hat{\lambda}_b$.
  2. Gere amostras bootstrap $\mathbf{y}_b^* = (y_{1,b}^*, ..., y_{L,b}^*)$ de $\boldsymbol{\beta}$-**SCAN**$(\mu_{0,l}(\widehat{\boldsymbol{\gamma}}), \widehat{\phi}_l, 0)$, $l =$

$1, 2, ..., L$.

3. Com base nos dados gerados em 2, use o algoritmo Newton-Rapshon e compute os pseudos estimadores $\hat{\boldsymbol{\theta}}^*$. Derive o pseudo valor de Λ_b^* e denote por $\hat{\lambda}_b^*$.

4. Repetindo os passos 2 e 3 para $q = 1, ..., B-1$ compute o valor-p para Λ por $p_{valor} \stackrel{\circ}{=} p^*_{valor}(\Lambda) = \frac{1}{B}\sum_{q=1}^{B} I(\hat{\lambda} \geq \hat{\lambda}_b^*)$.

FIM.

Capítulo 4

Estudo de Simulação

Neste capítulo, avaliamos a performance do Scan Circular proposto através de um conjunto de dados simulados. A região de estudo é o mapa do estado do Amazonas no Brasil com $L = 62$ municípios (Figura 4.1). O poder do teste para detectar o cluster depende de alguns fatores, como dados gerados sobre vários cenários usando diferentes valores para os parâmetros de precisão, da regressão e da razão de chance.

Primeiramente, sob a hipótese nula foram gerados 6 cenários, 3 cenários com um cluster com 4 regiões, e 3 cenários para um cluster de 8 regiões. Foram executados 1000 mapas para obter o valor crítico do teste ao nível de significância $\alpha = 0.05$ usando o modelo $\boldsymbol{\beta}$-**SCAN**$(\mu_{0,l}, \phi, 0)$, $l = 1, 2 \ldots, L$, de modo que

$$\mu_{0,l} = \frac{\exp\{-2 - 3,9x_l\}}{1 + \exp\{-2 - 3,9x_l\}}$$

para $\phi = 50, 100, 250$ e $x_l \sim Uniforme(0,1)$. O vetor de parâmetros fixos é $\gamma = (-2, 3.9)$ e o valor esperado de Y_l é aproximadamente 0,02.

Sob a hipótese alternativa foram gerados 60 cenários com 1000 mapas para estimar empiricamente o poder, a sensibilidade(SS) e o valor predito positivo(VPP) do teste. Sendo 30 cenários com um cluster artificial de 4 áreas e os demais 30 com um cluster artificial de 8 áreas (Figura 4.1). Os cenários sob a hipótese alternativa foram gerados com $\tau = log(i), i = 1, 2, \ldots, 10$ de modo que a razão de chance varia de 1 a 10. O poder é estimado pela proporção de vezes que o método rejeitou a hipótese nula ao nível $\alpha = 0.05$.

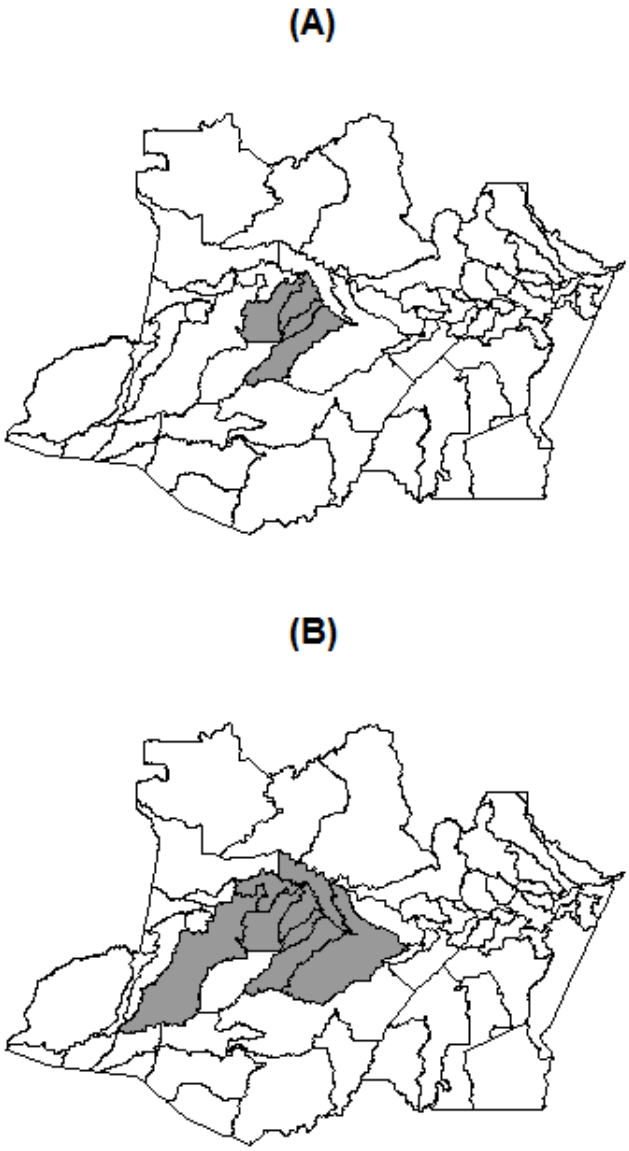


Figura 4.1: Cluster Artificial alocado no mapa: (A) com 4 áreas e (B) com 8 áreas

A precisão na detecção do cluster é medida por

✓ Sensibilidade (**SS**) - a razão média entre população em risco corretamente detectada pela verdadeira população em risco

$$\mathbf{SS} = \frac{1}{1000} \sum_{q=1}^{1000} \left(\frac{\text{pop}\{\hat{z}^{(q)} \cap z\}}{\text{pop}\{z\}} \right)$$

✓ Valor Predito Positivo (**VPP**) - a razão média entre população em risco corretamente detectada pela população em risco detectada

$$\textbf{VPP} = \frac{1}{1000}\sum_{q=1}^{1000}\left(\frac{\text{pop}\{\widehat{z}^{(q)} \cap z\}}{\text{pop}\{\widehat{z}^{(q)}\}}\right)$$

onde $\widehat{z}^{(q)}$ é o cluster estimado na q-ésima simulação, z é o cluster artificial alocado no mapa e pop$\{A\}$ é a população em risco do conjunto de localizações espaciais A. As medidas **SS** e **VPP** avaliam a habilidade do método para localizar o cluster, quando este existe.

4.1 Análise dos resultados

Na Figura 4.2 é apresentado a distribuição da Estatística de teste Λ sob a hipótese nula para $\phi = 50, 100, 250$. Observa-se que a distribuição de Λ depende do valor de ϕ e que o ponto crítico ao nível de significância de 5% decresce com aumento de ϕ. Isso deve ocorrer pelo fato que quando aumentamos o valor do parâmetro ϕ a variação nos valores observados $y's$ tende diminuir e esse efeito pode está sendo replicado para a estatística de teste.

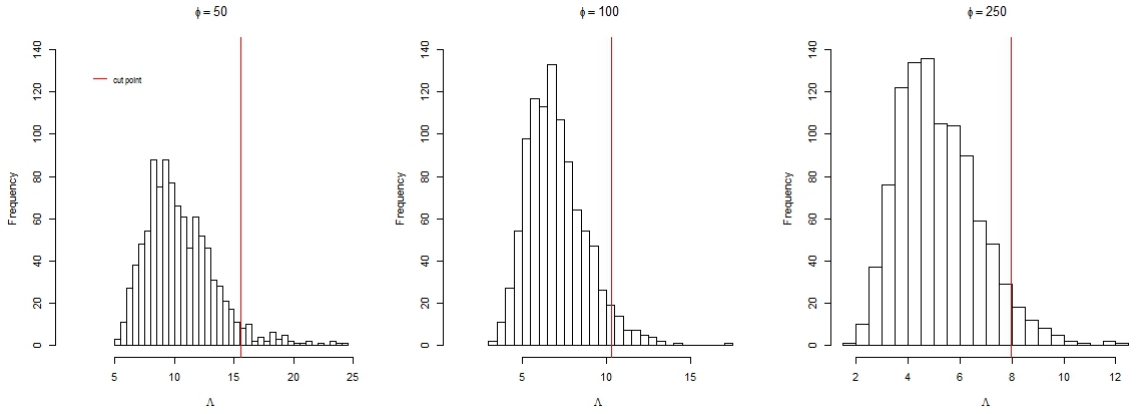


Figura 4.2: Distribuição da Estatística de teste Λ sob a hipótese nula para $\phi = 50, 100, 250$

Para o cluster artificial com 4 áreas alocado no mapa, os resultados de poder do teste, **VPP** e **SS** são mostrados na Figura 4.3. Observamos que essas medidas crescem com o aumento do parâmetro de clusterização τ . O poder e o **VPP** crescem com o aumento de ϕ mas, a sensibilidade **SS** decresce. Isso indica que o método tende a subestimar a população em risco (**VPP** $>$ **SS**). O efeito da variação de ϕ na performance do método para detecção do verdadeiro cluster é mais evidente no **VPP**. O vício na estimação do parâmetro τ também é avaliado na Figura 4.3 onde observamos que o método sobre-estima

o verdadeiro valor do parâmetro. No entanto, quando aumentamos simultaneamente os valores de ϕ e τ , essa sobre-estimação torna-se negligenciável. Estes resultados para o vício na estimação parecem estar em consonância com os obtidos em Prates *et al.* (2014) mesmo com modelos diferentes. Esse resultado é bem plausível, pois a medida que aumentamos o valor teórico de τ o VPP e SS tendem simultaneamente para 1, ou seja, o cluster detectado é praticamente igual ao verdadeiro de modo que $\widehat{\tau} \approx \tau$.

Os valores de poder, SS e VPP para o cluster com 4 regiões (Tabela 4.1) mostram-se sensivelmente menores, se comparados com o cluster para 8 regiões (Tabela 4.2). Ou seja, o modelo detecta clusters envolvendo mais regiões.

Tabela 4.1: Estimativas para o Poder, Sensibilidade (**SS**) e Valor Predito Positivo (**VPP**) para os diferentes valores de ϕ, $\tau = \log(i), i = 1,2,\ldots,10$ e $\{\#z\} = 4$.

Est.	ϕ	log(1)	log(2)	log(3)	log(4)	log(5)	log(6)	log(7)	log(8)	log(9)	log(10)
	50	1.083	1.193	1.422	1.687	1.868	2.123	2.231	2.408	2.537	2.627
$\widehat{\tau}$	100	0.921	1.176	1.454	1.685	1.863	2.008	2.187	2.286	2.380	2.474
	250	0.740	1.165	1.440	1.623	1.796	1.897	2.005	2.107	2.200	2.285
	50	0.050	0.082	0.166	0.320	0.442	0.610	0.727	0.783	0.876	0.921
Poder	100	0.054	0.214	0.517	0.766	0.919	0.946	0.985	0.992	0.995	1.000
	250	0.050	0.534	0.882	0.966	0.993	0.998	0.999	1.000	1.000	0.994
	50	0.349	0.612	0.765	0.827	0.856	0.868	0.897	0.900	0.903	0.925
SS	100	0.262	0.589	0.730	0.819	0.860	0.865	0.888	0.890	0.910	0.916
	250	0.203	0.541	0.712	0.781	0.832	0.872	0.891	0.912	0.924	0.947
	50	0.088	0.238	0.401	0.533	0.625	0.712	0.747	0.798	0.823	0.848
VPP	100	0.126	0.433	0.683	0.822	0.877	0.907	0.941	0.949	0.960	0.970
	250	0.273	0.769	0.933	0.967	0.987	0.988	0.991	0.994	0.994	0.992

Tabela 4.2: Estimativas o Poder, Sensibilidade (**SS**) e Valor Predito Positivo (**VPP**) para os diferentes valores de ϕ, $\tau = \log(i), i = 1,2,\ldots,10$ e $\{\#z\} = 4$.

Est.	ϕ	log(1)	log(2)	log(3)	log(4)	log(5)	log(6)	log(7)	log(8)	log(9)	log(10)
	$\phi = 50$	1.047	1.282	1.555	1.692	1.884	1.996	2.134	2.284	2.340	2.453
$\widehat{\tau}$	$\phi = 100$	0.903	1.266	1.429	1.612	1.724	1.854	1.948	2.046	2.140	2.212
	$\phi = 250$	0.842	1.237	1.364	1.462	1.579	1.704	1.792	1.900	1.995	2.091
	$\phi = 50$	0.093	0.151	0.364	0.633	0.789	0.861	0.940	0.968	0.979	0.984
Poder	$\phi = 100$	0.097	0.393	0.784	0.933	0.986	0.999	1.000	1.000	1.000	1.000
	$\phi = 250$	0.112	0.768	0.981	0.999	1.000	0.999	1.000	1.000	1.000	1.000
	$\phi = 50$	0.337	0.591	0.691	0.765	0.807	0.828	0.849	0.856	0.869	0.885
SS	$\phi = 100$	0.255	0.518	0.679	0.739	0.805	0.813	0.853	0.878	0.886	0.910
	$\phi = 250$	0.202	0.447	0.621	0.743	0.811	0.828	0.869	0.888	0.896	0.909
	$\phi = 50$	0.174	0.460	0.661	0.757	0.834	0.870	0.890	0.919	0.929	0.948
VPP	$\phi = 100$	0.245	0.720	0.872	0.940	0.963	0.975	0.979	0.986	0.987	0.992
	$\phi = 250$	0.491	0.932	0.982	0.990	0.997	0.997	0.998	0.997	0.998	0.999

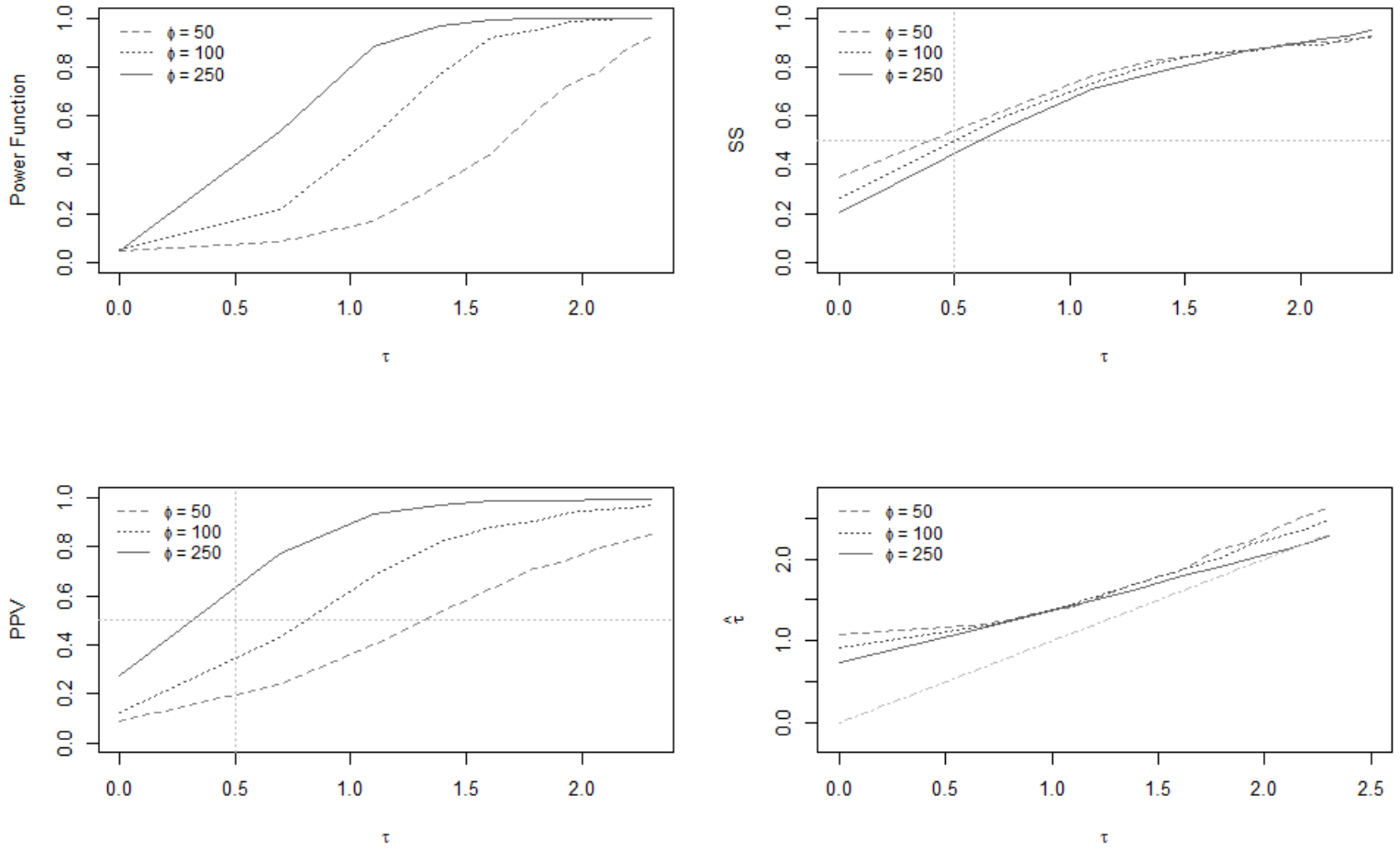


Figura 4.3: Estimativas o Poder, Sensibilidade (**SS**) e Valor Predito Positivo (**VPP**) para os diferentes valores de ϕ, $\tau = \log(i), i = 1,2,\ldots,10$ e $\{\#z\} = 4$.

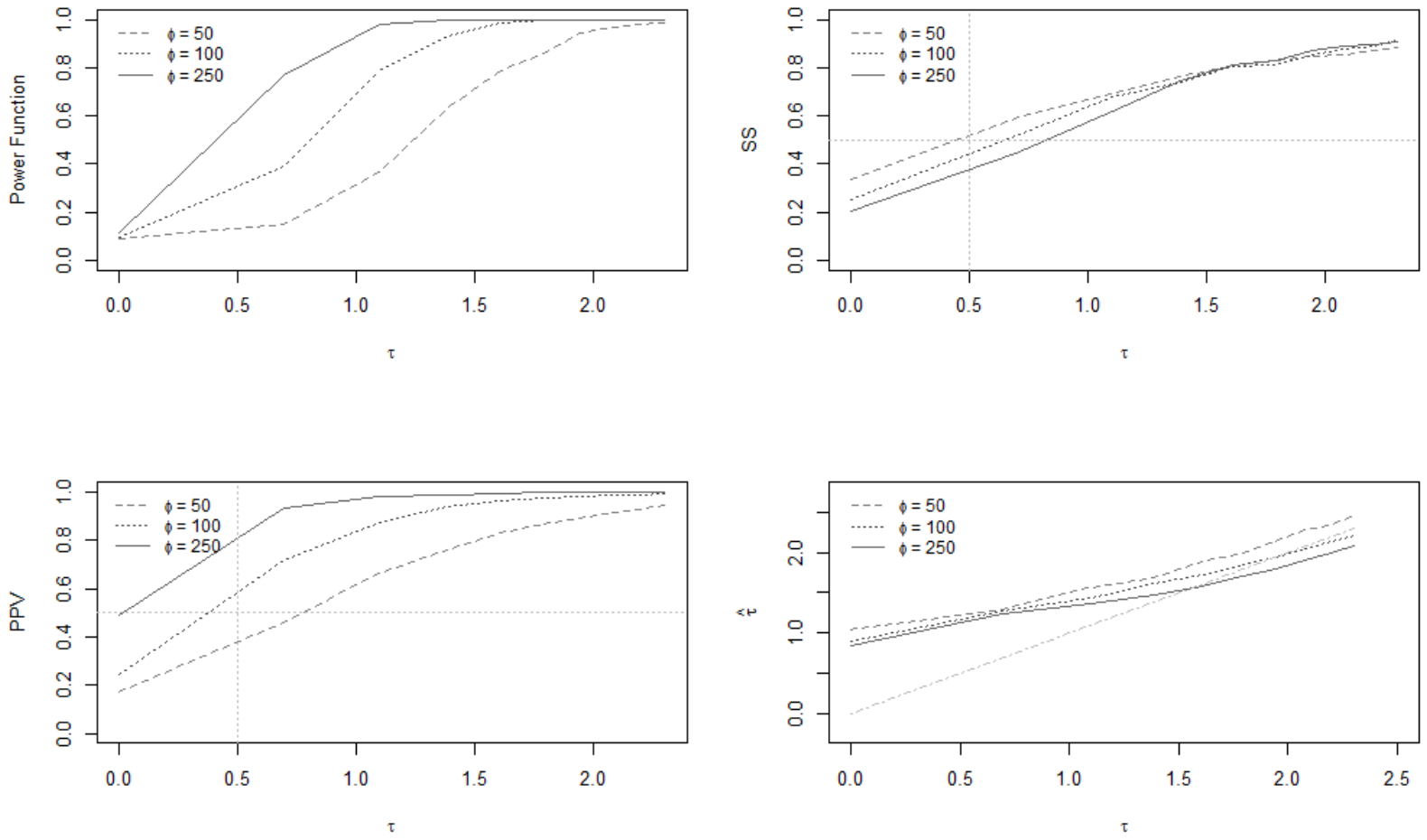


Figura 4.4: Estimativas o Poder, Sensibilidade (**SS**) e Valor Predito Positivo (**VPP**) para os diferentes valores de ϕ, $\tau = \log(i), i = 1,2,\ldots,10$ e $\{\#z\} = 8$.

Capítulo 5

Aplicação

5.1 Estudo de Caso : Taxa de Mortalidade Infantil no Estado do Amazonas

5.1.1 Dados de Mortalidade Infantil

Os dados utilizados nesta aplicação são referentes a taxa de mortalidade infantil ocorridas no Estado do Amazonas no período de 2004 a 2009 em cada um dos seus 62 municípios. Estes dados foram obtidos nos Cadernos de Informações de Saúde Amazonas e podem ser acessados no endereço `http://tabnet.datasus.gov.br/tabdata`. Foram observadas, no total de 7.731 mortes nesse período. A taxa média de mortalidade infantil para cada mil nascidos vivos foi de $17,44$. Embora, trabalhos anteriores haviam usado modelos Poisson com valor esperado proporcional à população de risco, a análise de Regressão Poisson apresentou elevada sobredispersão (desvio / graus de liberdade $= 577,20/59 \approx 9,78$), sendo que essa bordagem torna-se inadequada, veja Lima *et al.* (2015) para uma discussão sobre o efeito da sobredispersão no problema de detecção de clusters espacias.

Outro fator importante é o fato que a capital do Amazonas (cidade de Manaus) concentra mais de 50% de toda população do Estado e a população n_l de nascido vivos nos demais $l-$municípios é pequena de forma que a modelagem $0 < y_l = y_l^*/n_l < 1$ é mais adequada para remover o efeito populacional, onde $0 < y_l^* < n_l$ representa o número de casos de mortalidade infantil. A distribuição espacial dessa taxa de mortalidade é mostrada na Figura 5.1(a). No Brasil, essa mortalidade é considerada um dos mais importantes in-

dicadores para medir a qualidade de vida da população. Alguns pesquisadores defendem que a partir do momento que houver uma preocupação em melhorar as condições socioeconômicas da população de baixa renda, o acesso à educação e ao saneamento básico, as taxas de mortalidade poderão diminuir consideravelmente no Brasil (Scalo *et al.*, 2012). Por isso, utilizamos como covariável regressora x_{l1} o Indice de Desenvolvimento Humano Municipal (IDHM) do ano de 2010 e outra covariável regressora x_{l2} referente ao índice de aleitamento materno IAM (percentual de crianças com aleitamento materno exclusivo (IAM) do período de 2004 a 2009 para os municípios do Estado.

5.1.2 Análise dos resultados para Detecção de Cluster

Aplicando os resultados da seção 3.2 no modelo via regressão Beta proposto, verificamos através da Tabela 5.1 que a mortalidade infantil é significativamente relacionada com o IDHM e o IAM.

Tabela 5.1: Estimativas dos parâmetros para o Modelo de Regressão Beta

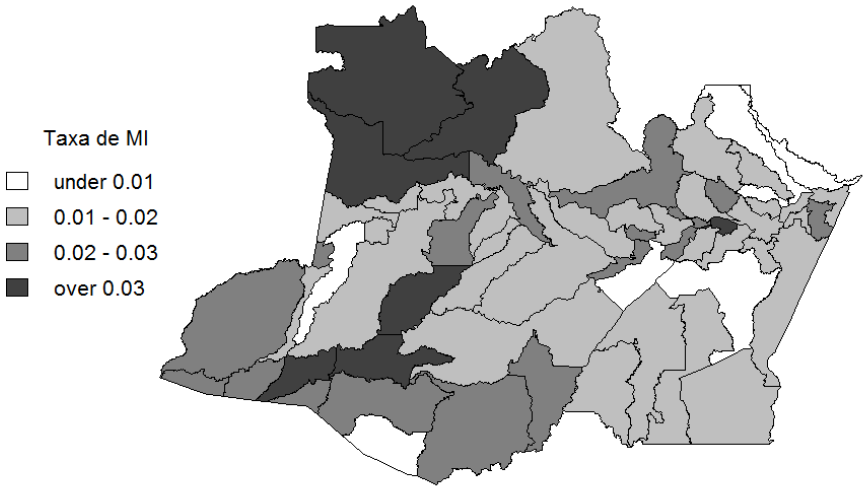
Parâmetro	Estimativa	Desvio	valor p
Intercepto	$-2,0257$	$0,5821$	$0,000502$
IDHM	$-1,9028$	$0,9360$	$0,042066$
IAM	$-0,0107$	$0,0044$	$0,015475$
ϕ	$332,59$	$60,46$	$3,78\times10^{-8}$

O valor $\widehat{\phi}=332,59$ indica uma alta precisão e pequena variância na distribuição das taxas. Usando o valor estimado $\widehat{\theta}_0=(-2,0257;-1,9028;-0,0107;332,59)$ o Scan Circular foi aplicado com raio máximo que agregue até 50% das áreas do mapa. O valor obtido para a estatística de teste foi $\widehat{\Lambda}=6,923$, para 1000 bootstrap obtivemos p-valor$=0,026$ com cluster espacial estimado $\widehat{z}$ formado pelos municípios $\widehat{z}=$ { Japurá, São Gabriel, Santa Izabel}.

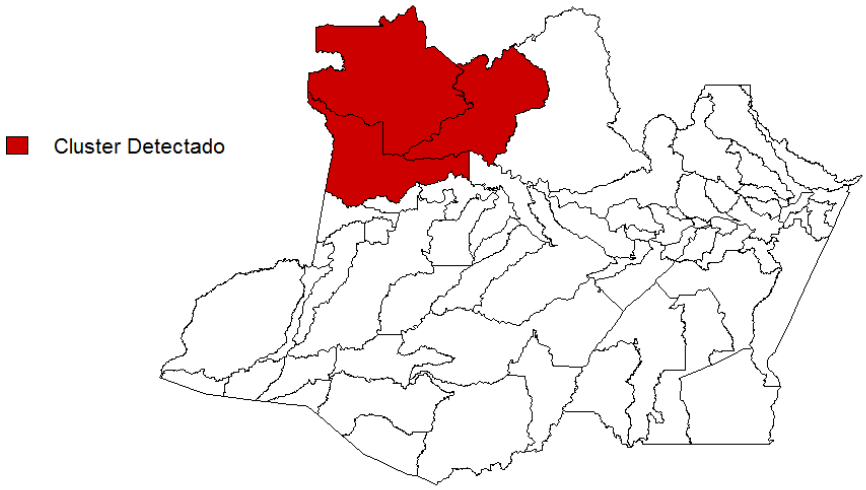
O parâmetro de clusterização estimado foi $\widehat{\tau}=0,71353$, o qual pode ser interpretado como uma razão de chance e $e^{\widehat{\tau}}=2,0412$, ou seja, a chance de ocorrência de mortalidade infantil na região detectada $\hat{z}$ é duas mais provável que em qualquer outro município do estado escolhido aleatoriamente no mapa. O modelo ajustado é

$$g(\widehat{\mu_{z,l}})=\log\left(\frac{\widehat{\mu_{z,l}}}{1-\widehat{\mu_{z,l}}}\right)=-2,0257-1,9028x_{l1}-0,0107x_{l2}+0,71353\mathtt{I}_{\{s_l\in\hat{z}\}}\,.$$

A Figura 5.1(a) mostra a distribuição espacial das taxas de mortalidade infantil no



(a)



(b)

Figura 5.1: (a) Distribuição Espacial da Taxa de Mortalidade Infantil; (b) Cluster Espacial Detectado.

período de 2004 a 2009. A localização de $\widehat{z}$ está na Figura 5.1(b), onde se pode observar que essa região pertence a uma região onde existe a maior concentração de população indígena do estado e isso, pode justificar a presença desse cluster, visto que segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), cerca de 40% das mortes indígenas registrada é de crianças com até 4 anos. Esse número é aproximadamente 9 vezes maior que o de crianças não indígenas na mesma faixa etária. Esses resultados são importantes, pois podem direcionar ações básicas de saúde nas comunidades indígenas.

5.2 O pacote betaScan

Durante o processo de elaboração dos algoritmos deste trabalho, o Scan espacial circular (Capítulo 2), o algoritmo de estimação do modelo $\boldsymbol{\beta}$-SCAN, a significância do teste (Capítulo 3) e as medidas de eficiências (Poder, SS, VPP) (Capítulo 4) foram construídas através de procedimentos implementados na linguagem de programação `OX` em sua versão 7.0 (distribuída gratuitamente para uso acadêmico e disponível no site `http://www.doornik.com`). Apesar da programação em `OX` ser mais rápida nas execuções de tarefas, surgiu a necessidade de transpor os códigos para o software `R`, por esse apresentar maior uso na área da estatística.

No decorrer deste trabalho, foi desenvolvido um pacote chamado `betaScan` [1] no software `R` para detectar cluster espaciais através do $\boldsymbol{\beta}$-SCAN. O objetivo deste capítulo é fornecer uma introdução geral ao pacote `betaScan`. Ao longo do capítulo, os mesmos dados de Mortalidade Infantil apresentados na Seção 5.1 será utilizado repetidamente como um exemplo. A estrutura do presente capítulo é como se segue. Na Seção 5.2.1 introduz o funcionamento do pacote. E na Seção 5.2.2 é apresentada o valor *p* Bootstrap para avaliar a significância da estatística de teste.

5.2.1 Descrição do Pacote

A função `betaScan()` é responsável pela execução dos procedimentos de inferência para a estatística $\boldsymbol{\beta}$-SCAN , no qual tem a seguinte forma

```
betaScan(formula, data, geo, alpha = 0.05, imax = 100)
```

[1] O download da primeira versão do pacote betaScan está disponível no site http://icede.ufam.edu.br/index.php/corpo-docente/9-de-departamento-de-estatistica/corpo-docente/29-max-sousa

em que `formula = data[,1] ~ data[,2] + ...` , onde `data` é a matriz de dados composto pelo vetor de variáveis respostas (primeira coluna) no intervalo $(0,1)$ e a matriz de covariáveis (nas demais colunas); o argumento `geo` representa a matriz bidimensional das coordenadas geográficas (latitude e longitude); `alpha` é o nível de significância dos coeficientes do modelo; e `imax` é o limite superior do intervalo $(0, imax)$ para controlar a estimação de τ.

Para um exemplo ilustrativo, temos os dados da taxa de mortalidade infantil, como organizado a seguir para as seis primeiras observações

```
> head(data)
    taxa     IDHM   IAM
1 0.019090  0.5275 77.97132
2 0.014118  0.5600 94.25496
3 0.023830  0.5940 69.75146
4 0.025563  0.5610 80.15112
5 0.018304  0.6370 83.08643
6 0.024275  0.4500 76.49954
```

onde `taxa` representa a variável resposta; `IDHM` é o índice de desenvolvimento humano municipal do ano de 2010; `IAM` é a taxa de crianças com aleitamento materno exclusivo. Esses dados são para os 62 municípios do Estado do Amazonas. Os resultados de saída são mostrados a seguir.

```
> formula = data[,1] ~ data[,2] + data[, 3]
> betaScan(formula, data, geo, alpha = 0.05, imax = 100)
$formula
data[, 1] ~ data[, 2] + data[, 3]


$betareg


Call:
betareg::betareg(formula = form, data = as.data.frame(dat))


Standardized weighted residuals 2:
```

```
Min      1Q  Median      3Q     Max
-2.9133 -0.5385  0.1186  0.6356  2.6667

Coefficients (mean model with logit link):
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.025682   0.582129  -3.480 0.000502 ***
data[, 2]   -1.902817   0.936028  -2.033 0.042066 *
data[, 3]   -0.010704   0.004421  -2.421 0.015475 *

Phi coefficients (precision model with identity link):
Estimate Std. Error z value Pr(>|z|)
(phi)   332.59      60.46   5.501 3.78e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 218.2 on 4 Df
Pseudo R-squared: 0.1319
Number of iterations: 55 (BFGS) + 6 (Fisher scoring)

$beta
(Intercept)   data[, 2]   data[, 3]
-2.0256821  -1.9028167  -0.0107037

$tau
[1] 0.7135266

$phi
[1] 332.5896

$likH0
[1] 218.216
```

```
$cluster
[1] 32 50 52

$max.likH1
[1] 6.923357

$odd.ratio
[1] 2.041177
```

O argumento `betareg` é a saída dos resultados retornados da função `betareg` (Pacote provido do modelo de regressão Beta padrão); `beta` é o vetor das estimativas dos parâmetros $\boldsymbol{\gamma}$ da regressão Beta, derivados do pacote `betareg`; `tau` é a estimativa do parâmetro τ; `phi` é a estimativa do parâmetro de precisão ϕ; `likH0` é a estimativa da função de log verossimilhança sob a hipótese nula; `cluster` é o vetor de índices do cluster detectado; `max.likH1` é a estimativa da estatística de teste Λ do modelo $\boldsymbol{\beta}$-SCAN; e `odd.ration` é a razão de chance e^{τ}. Mais detalhes sobre o pacote `betaScan`, veja o apêndice.

5.2.2 Estimação do valor *p* Bootstrap

A verificação da significância da estatística de teste do modelo $\boldsymbol{\beta}$-SCAN, é realizada através do comando `betaScan.boot`, que retorna a estimativa do valor *p*. Portanto, o comando a ser usado segue a estrutura:

```
betaScan.boot(B, formula, data, geo, alpha = 0.05, imax = 50)
```

onde `B` é o número de replicas de Bootstrap. Usando o banco de dados da seção anterior, obtemos a estimativa do valor *p* para `B=1000` réplicas, como apresentado abaixo.

```
$p_value

[1] 0.026
```

onde `p_value` retorna o valor *p*.

Capítulo 6

Considerações Finais

6.1 Conclusões

Este trabalho teve como proposta estudar a estatística scan espacial baseada em modelos de regressão Beta, a $\boldsymbol{\beta}$-SCAN para detecção de clusters geográficos em dados contínuos distribuídos no intervalo $(0,1)$ ou limitados em (a,b), $a < b$. A detecção de clusters geográficos de taxas e proporções são situações práticas onde o método proposto pode ser aplicado. A suposição do método é que a resposta (taxa ou proporção) segue uma distribuição Beta. A Estatística $\boldsymbol{\beta}$-SCAN é muito flexível para detecção de clusters de taxas, pois a distribuição Beta pode assumir diferentes formas dependendo dos valores dos parâmetros que indexam a distribuição. Sob a hipótese nula de completa aleatoriedade espacial das taxas, nós usamos uma reparametrização na qual a taxa média é uma função de um preditor linear definido por parâmetros da regressão e variáveis explicativas. Sob a hipótese alternativa, acrescentamos no preditor linear um parâmetro de clusterização que pode ser interpretado em termos da razão de chance de ocorrêcnia de eventos no cluster comparado com as demais áreas do mapa.

A estimação dos parâmetros foi obtida por máxima verossimilhança e a significância estatística do cluster foi realizada através do valor p bootstrap. Nossos estudos simulados mostraram que a metodologia proposta possui um alto poder, uma boa sensibilidade e um bom valor predito positivo para localizar corretamente o cluster, ou parte dele. Essas medidas de performance do método crescem com o aumento dos parâmetros de precisão e clusterização do modelo. Os resultados mostraram que quando a população em risco e o número de ocorrências de eventos são conhecidos, a modelagem via

$\boldsymbol{\beta}$-SCAN é mais eficiente e adequada.

6.2 Sugestões para trabalhos futuros

1. **Realizar vigilância epidemiológica no tempo**. Para utilizar esse método, realizamos um monitoramento estatístico de um processo estocástico $\{Y_t : t = 1,2,\ldots\}$ com o objeivo de detectar uma mudança importante no processo em um tempo desconhecido κ, tão rápida e precisa quanto possível. Suponha $Y_t \sim Beta(\mu_t, \phi)$, para $t \leq \kappa$. Se $t > \kappa$, então $Y_t \sim \boldsymbol{\beta}ARMA(\mu_t, \phi, \tau)$. Nesse caso, um modelo autorregressivo e de média móvel seria da seguinte forma:

$$g(\mu_t) = \mathbf{x}_t\boldsymbol{\gamma} + \delta_t + \tau \mathtt{I}_{t>\kappa}$$

onde $\mathbf{x}_t$ é a matriz de covariáveis; $\boldsymbol{\gamma}$ é o vetor de parâmetros fixos; τ é o parâmetro de mudança no tempo; e δ_t assume

$$\delta_t = \alpha + \sum_{i=1}^{p} \varphi_i\{g(y_{t-i}) - \mathbf{x}'_{t-i}\boldsymbol{\gamma}\} + \sum_{j=1}^{q} \sigma_j \omega_{l,t-j}$$

onde $\alpha \in \mathbb{R}$ é uma constante; $p, q \in \mathbb{R}$ representam, respectivamente, a ordem autorregressivo e média móvel; $\varphi's$ e $\sigma's$ são os parâmetros autorregressivo e de média móvel; ω_t é o erro aleatório. Mais detalhes do modelo $\boldsymbol{\beta}$ARMA padrão, ver em Rocha & Cribari-Neto (2009) .

2. **Realizar vigilância epidemiológica no espaço-tempo**. Há situação que é necessário realizar por um longo período de tempo um vigilância espacial. Considere $S = \{s_1, s_2, \ldots, s_L\}$ um mapa particionado em L áreas de polígono A_l . Suponha que em S realizamos um monitoramento estatístico de um processo estocástico $\mathbf{Y} = \{Y_t(s_l), t = 1,2,\ldots \text{ e } l = 1,2,\ldots,L\}$. A cada tempo $t \geq 1$, observamos um vetor L-variado $Y_t = (Y_t(s_1), Y_t(s_2), \ldots, Y_t(s_L))^\top$. Onde, na nossa proposta, $Y_t(s_L) \sim Beta(\mu_{l,t}, \phi)$ quando $s_l \in S$. Seja z um cluster potencial, então o processo espaço-temporal modelado por $\boldsymbol{\beta}ARMASCAN(\mu_{l,t}, \phi, \tau)$, no qual assume

$$g(\mu_{l,t}) = \mathbf{x}_{l,t}\boldsymbol{\gamma} + \delta_{l,t} + \tau \mathtt{I}_{s_l \in z; t>\kappa}$$

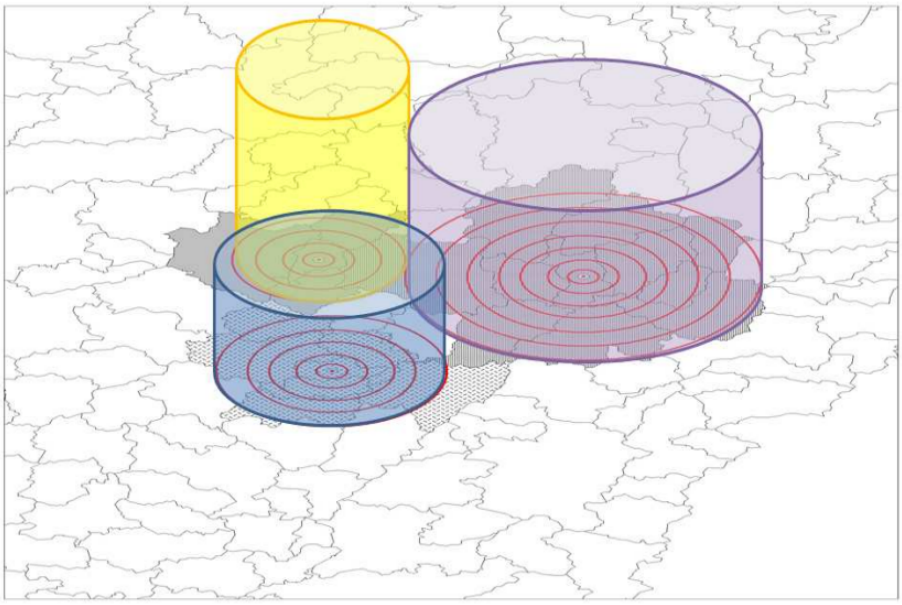


Figura 6.1: .Exemplos de cilindros encontrados mediante varredura espaço temporal de uma região. O centro dos cilindros é localizado no centróide de cada sub-área. Para cada centróide o raio e a altura crescem independentemente, constituindo zonas candidatas à composição de conglomerados.

onde $\mathbf{x}_t$ é a matriz de covariáveis; $\boldsymbol{\gamma}$ é o vetor de parâmetros fixos; τ é o parâmetro de mudança no espaço-tempo; e δ_t assume

$$\delta_{l,t} = \alpha + \sum_{i=1}^{p} \varphi_i \{ g(y_{l,t-i}) - \mathbf{x}'_{l,t-i}\boldsymbol{\gamma} \} + \sum_{j=1}^{q} \sigma_j \omega_{l,t-j}$$

onde $\alpha \in \mathbb{R}$ é uma constante; $p, q \in \mathbb{R}$ representam, respectivamente, a ordem autorregressivo e média móvel; $\varphi's$ e $\sigma's$ são os parâmetros autorregressivo e de média móvel; ω_l, t é o erro aleatório. A Figura 6.1 mostra o funcionamento da estatística scan circular no espaço-tempo.

3. **Propor uma estatística scan espacial, temporal ou espaço-temporal baseado em modelos regressão Beta inflacionados**: Existem situações que, além da variável $y_l \in s_l$ está no intervalo $(0,1)$, também há casos que essas observações vem assumir valores nos intervalos $[0,1]$, $[0,1)$ ou $(0,1]$. Para mais detalhes sobre esses modelos, ver em Ospina & Ferrari (2012). Ainda, podemos propor uma estatística scan espacial, temporal, ou espaço temporal baseado em modelos regressão Beta com formatos irregulares.

Apêndice

Package 'betaScan'

February 22, 2015

Title A Spatial Scan Statistics for Beta Regression.

Date 2015-01-07

Version 1.0-0

Author Max Sousa de Lima, Luiz Henrique Duczmal and Vanessa Sousa Santos

Description Provide functions for Cluster Detection using Spatial Scan Statistics for Beta Regression.

Imports betareg, rootSolve

Depends R (>= 3.0.1)

Maintainer Max Sousa de Lima <max.lima@ufam.edu.br> and Diego da Silva Souza <souzadiegosilva@gmail.com>

License GPL (>=2)

Repository CRAN

Date/Publication 2015-01-07 00:00:00 GMT-4

NeedsCompilation no

R topics documented:

`betaScan-package` *A Spatial Scan Statistics for Beta Regression.*

Description

Provide functions for Cluster Detection using Spatial Scan Statistics for Beta Regression.

Details

Package:	betaScan
Type:	Package
Version:	1.0
Date:	2015-01-07
License:	What license is it under?

Author(s)

Max Sousa de Lima, Luiz Henrique Duczmal and Vanessa Sousa Santos.

Maintainer: Max Sousa de Lima <max.lima@ufam.edu.br> and Diego da Silva Souza <souzadiegosilva@gmail.com>

`betaScan` *A Spatial Scan Statistics for Beta Regression.*

Description

Provide a function for Cluster Detection using Spatial Scan Statistics for Beta Regression.

Usage

```
betaScan(formula, data, geo, alpha = 0.05, imax = 100)
```

Arguments

`formula`	Symbolic description of the model.
`data`	A matrix or data.frame of observations of variables in *formula*.
`geo`	A matrix with geographic coordinates.
`alpha`	The level of signification of coefficients from the respective models.
`imax`	The upper end point of the interval to be searched the clustering parameter tau.

Value

`formula`	The formula of the model.
`betareg`	The summary of results returned from betareg function.
`beta`	A vector of beta parameters estimatives.
`tau`	The estimative of tau parameter.
`phi`	The estimative of phi parameter.
`likH0`	The estimative of log likelihood under the null hypothesis.
`cluster`	A vector of indices of the detected cluster.
`max.likH1`	The maximum estimative of log likelihood under the alternative hypothesis.
`odd.ration`	The maximum odd ratio.

Author(s)

Max Sousa de Lima, Luiz Henrique Duczmal and Vanessa Sousa Santos.

betaScan-Internal *Internal betaScan Functions*

Description

Internal maxRV functions

Details

These functions are not to be called by the user, it is for "internal" use only.

Author(s)

Max Sousa de Lima, Luiz Henrique Duczmal and Vanessa Sousa Santos.

betaScan.boot *Estimate the Bootstrap p-value of the detected cluster.*

Description

This function provides a estimative of p-value using the Bootstrap method to evaluate the detected cluster significance.

Usage

```
betaScan.boot(B, formula, data, geo, alpha = 0.05, imax = 50)
```

Arguments

B	The number of bootstrap replicates
formula	Symbolic description of the model.
data	A matrix or data.frame of observations of variables in *formula*.
geo	A matrix with geographic coordinates.
alpha	The level of signification of coefficients from the respective models.
imax	The upper end point of the interval to be searched the clustering parameter tau.

Value

p_value	the p-value estimated via Bootstrap.

Author(s)

Max Sousa de Lima, Luiz Henrique Duczmal and Vanessa Sousa Santos

Referências Bibliográficas

Assunção, R., Costa, M., Tavares, A. & Ferreira, S. (2006). Fast detection of arbitrarily shape disease clusters. *Statistics in Medicine*, **25**, 723–742.

Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **154**(1), pp. 143–155.

Bhatt, V. & Tiwari, N. (2014). A spatial scan statistic for survival data based on weibull distribution. *Statistics in Medicine*, **33**(11), 1867–1876.

Cançado, A., da Silva, C. & da Silva, M. (2014). A zero-inflated poisson-based spatial scan statistic. *Environmental and Ecological Statistical*, **to appear**.

Casella, G. (2002). *Statistical Inference*. Duxbury Advanced Series. Duxbury Thomson Learning. ISBN 9780495391876.

Cuzick, J. & Edwards, R. (1990). Spatial clustering for inhomogeneos populations. *Journal of Royal Statistical Society*, **52**, 73–104.

Duczmal, L. & Assunção, R. (2004). A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, **45**, 269–286.

Duczmal, L., Kulldorff, M. & Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics*, **15**(2).

Dwass, M. (1957). On the distribution of ranks and of certain rank order statistics. *Ann. Math. Statist*, **28**(2), 424–431.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), pp. 1–26.

Ferrari, S. L. P. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Glaz, J. (2009). *Applications of Spatial Scan Statistics: A Review*. Statistics for Industry and Technology. Springer/BirkhÃ$\frac{1}{4}$se, Boston, MA.

Huang, L., Kulldorff, M. & Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, **63**(1), 109–118.

Huang, L., Tiwari, R. C., Pickle, L. W. & Zou, Z. (2010). Covariate adjusted weighted normal spatial scan statistics with applications to study geographic clustering of obesity and lung cancer mortality in the united states. *Statistics in Medicine*, **29**(23), 2410–2422.

Jung, I. (2009). A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statistics in Medicine*, **28**, 1131–1143.

Jung, I., Kulldorff, M. & Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine*, **26**(7), 1594–1607.

Jung, I., Kulldorff, M. & Richard, O. J. (2010). A spatial scan statistic for multinomial data. *Statistics in Medicine*, **29**(18), 1910–1918.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods*, **26**(6), 1481–1496.

Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, **14**(8), 799–810.

Kulldorff, M., Tango, T. & Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, **42**(4), 665–684.

Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. & Platt, R. (2007). Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, **26**(8), 1824–1833.

Kulldorff, M., Huang, L. & Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, **8**(1), 58.

Lawson, A. & Kulldorff, M. (1999). A review of cluster detection methods. *Disease Mapping and Risk Assessment for Public Health*, pages 99–110.

Lima, M., Duczmal, L., Neto, J. & Pinto, L. (2015). Spatial scan statistics for models with overdispersion and inflated zeros. *Statistica Sinica*, page to appear.

Lima, M. S. (2004). Avaliação do poder do teste da estatística scan para múltiplos clusters.

Lima, M. S. (2011). Método adaptativo para detecção de clusters no espaço-tempo.

Loh, J. M. & Zhu, Z. (2007). Accounting for spatial correlation in the scan statistic. *Ann. Appl. Stat.*, **1**(2), 560–584.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models (Second edition)*. London: Chapman & Hall.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. **60**(??), 532–538.

Neill, D. B., McFowland, E. & Zheng, H. (2013). Fast subset scan for multivariate event detection. *Statistics in Medicine*, **32**(13), 2185–2208.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, **135**, 370–384.

Nocedal, J. & Wright, S. (1999). *Numerical Optimization*. Springer series in operations research and financial engineering. Springer. ISBN 9780387987934.

Openshaw, S., Craft, A. W. & Birch, J. (1988). Investigation of leukaemia cluster by use of a geographical analysis machine. *Lancet*, **1**.

Ospina, R. & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, **56**(6), 1609 – 1623.

Prates, M. O., Kulldorff, M. & Assunção, R. (2014). Relative risk estimates from spatial and space-time scan statistics: are they biased? *Statistics in Medicine*, **33**, 2634–2644.

Read, S., Bath, P., Willet, P. & Maheswaran, R. (2013). Study on the use of gumbel approximation with the bernoulli spatial scan statistics. *Statistics in Medicine*, **32**, 3300–3313.

Rocha, A. & Cribari-Neto, F. (2009). Beta autoregressive moving average models. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, **18**(3), 529–545.

Rosychuk, R. & Chang, H. (2013). A spatial scan statistics for compound poisson data. *Statistics in Medicine*, **32**, 5106–5118.

Scalo, J., Jardim, S., Santos, G. & Nogueira., D. (2012). Analysis of spatial patter of infant mortality using geostatistics. *Revista Univap*, **18**(32), 149–160.

Zhang, T. & Lin, G. (2009). Spatial scan statistics in loglinear models. *Computational Statistics and Data Analysis*, **53**(8), 2851–2858.

Zhang, T., Zhang, Z. & Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in Medicine*, **31**(8), 762–774.