Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Detecção de clusters espacias e espaço-temporais em modelos com excesso de zeros e sobredispersão

Belo Horizonte, Abril de 2015

Letícia Pereira Pinto

Detecção de clusters espacias e espaço-temporais em modelos com excesso de zeros e sobredispersão

Tese de doutorado apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como parte dos requisitos necessários para obtenção do título de Doutor em Estatística.

Orientador: Prof. Dr. Luiz H. Duczmal.

Coorientador: Prof. Dr. Max S. de Lima.

Belo Horizonte, Abril de 2015

Instituto de Ciências Exatas

Universidade Federal de Minas Gerais

"É mais fácil amar o retrato. Eu já disse que o que se ama é uma 'cena'. 'Cena' é um quadro belo e comovente que existe na alma antes de qualquer experiência amorosa. A busca amorosa é a busca da pessoa que, se achada, irá completar a cena. Antes de te conhecer eu já te amava... E então, inesperadamente, nos encontramos com o rosto que já conhecíamos antes de o conhecer. E somos então possuídos pela certeza absoluta de haver encontrado o que procurávamos. A cena está completa. Estamos apaixonados."
Retratos de Amor - Rubem Alves

Ao meu grande amor Frederico Ferreira Campos, filho.

"In the vastness of space and the immensity of time, it is my joy to share a planet and an epoch with Frederico."

Adaptação de *Cosmos* - Carl Sagan

Resumo

A Estatística Scan Espacial é um dos métodos mais importantes para a detecção e monitoramento de clusters espaciais de doenças. Geralmente, assume-se que os casos de doença seguem uma distribuição de Poisson ou binomial. Na prática, no entanto, dados de contagem de casos freqüentemente apresentam um excesso de zeros e/ ou sobredispersão, resultando na violação desses modelos comumente utilizados, aumentando a ocorrência do erro tipo I. Esta tese descreve uma modificação da Estatística Scan Espacial com o modelo Poisson Duplo inflacionado de Zeros (ZIDP) para reduzir o erro tipo I, acomodando simultaneamente o excesso de zeros e a sobredispersão. Os parâmetros do modelo nulo e alternativo, são estimados pelo algoritmo da Expectation-Maximization e o p-valor é obtido através do Fast Double Bootstrap Test. Uma aplicação é apresentada para os dados de Hanseníase na Amazônia brasileira. Uma extensão desta estatística em sistemas prospectivos de vigilância espaço-temporal foi estudada e para avaliar o seu desempenho foram utilizadas simulações de Monte Carlo.

Palavras-chave e frases: Poisson Duplo, Algoritmo EM, Sobredispersão, Estatística Scan Espacial, Análise Espaço-Temporal, Inflacionado de zeros.

Abstract

The Spatial Scan Statistic is one of the most important methods for detecting and monitoring spatial disease clusters. Usually it is assumed that disease cases follow a Poisson or Binomial distribution. In practice, however, case count datasets frequently present an excess of zeroes and/or overdispersion, resulting in the violation of those commonly used models, increasing type I error occurrence. This thesis describes a modification of the Spatial Scan Statistic with the Zero Inflated Double Poisson (ZIDP) model to reduce type I error, accommodating simultaneously an excess of zeroes and overdispersion. The null and alternative model parameters are estimated by the Expectation-Maximization algorithm and the p-value is obtained through the Fast Double Bootstrap Test. An application is presented for Hanseniasis data in the Brazilian Amazon. An extension of this statistic in prospective space-time surveillance systems has been studied and in assess their performance Monte Carlo simulations were used.

Key words and phrases: Double Poisson, EM-algorithm, overdispersion, spatial scan statistics, space-time analysis, zero inflated.

Agradecimentos

Ao meu amor Frederico, por fazer parte da minha história, pelo amor, pela amizade, pelo companheirismo, pelo incentivo desde o início desta longa caminhada, por dividir comigo o gosto pelas ciências exatas, por ter sido a grande fonte de inspiração e motivação para este trabalho, e por preencher os meus dias com alegria! Não existe uma palavra capaz de expressar o amor que sinto por você!

Meus pais (Roldão e Emília), meu amor maior... as minhas paredes, o meu alicerce. Tanto investiram, com amor, trabalho, dedicação e orações, na minha formação pessoal e profissional. É um orgulho imenso ser fruto de duas pessoas tão íntegras e de caráter extremo. Na verdade, é uma responsabilidade perpetuar tudo o que aprendi com vocês.

Minhas irmãs (Patrícia, Rita e Viviane) pelo amor, pelo apoio e pelo incentivo incondicional.

Ao meu orientador Luiz Henrique Duczmal, pelos ensinamentos, pelas preciosas discussões e contribuições, pelas sugestões, pelo aprendizado, pelo incentivo, pelos momentos de paciência em entender minhas dificuldades e por me motivar a superá-las a todo momento. Você é um grande exemplo de amor e dedicação à pesquisa e à profissão!

Ao Prof Max Sousa de Lima, pelos ensinamentos e valiosas contribuições. Com o seu jeito próprio, foi uma peça fundamental para a realização deste trabalho.

Aos amigos da pós-graduação pelos momentos compartilhados juntos em sala de aula e laboratórios.

Ao Prof. Anderson Duarte pelo precioso auxílio nos estudos de Probabilidade Avançada.

Ao Prof. Helton Matos pelo auxílio na linguagem C/C++.

Às secretárias Kate, Renata, Rogéria, Rosi e Sônia pela paciência, pela simpatia e pelo auxílio nas tarefas burocráticas.

Ao Departamento de Ciência da Computação que tornou possível o suporte financeiro cedendo uma bolsa CAPES/ Reuni ao Departamento de Estatística e pelo suporte computacional.
À CAPES pelo apoio e pelo suporte financeiro.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A Estatística Scan Espacial (Kulldorff, 1997) é um método popular para a detecção e inferência de clusters espaciais de doença. Recentemente, várias extensões foram criadas para acomodar correlação (Loh, 2007), o ajuste de covariável (Jung, 2009), modelagem log-linear (Zhang e Lin, 2009), sobredispersão (Zhang et al., 2012) e inflação de zeros (Cançado et al., 2014). Na vigilância em saúde pública, frequentemente a contagem de casos de doenças possui uma variabilidade maior do que o permitido pelo modelo de Poisson, que assume igualdade entre a média e a variância. Este excesso na variabilidade é chamado sobredispersão e tem sido amplamente discutido na literatura. Não considerar a presença de sobredispersão no modelo pode levar à superestimação de erro tipo I (decisão de rejeitar H_0 quando de fato H_0 é verdadeira) e consequente inferência errônea para os parâmetros do modelo. Na presença de sobredispersão, os modelos Poisson Generalizado (Consul e Jain, 1973) e Poisson duplo (Efron, 1986) são mais adequados. Outro problema que ocorre comumente em dados de contagem é a excessiva presença de zeros, incompatível ao modelo empregado. Nesta situação, diz-se que o conjunto de dados exibe uma inflação de zeros, ou excesso de zeros. A sobredispersão pode ser, às vezes, apenas uma consequência da inflação de zeros, e neste caso, este problema pode ser contornado usando o modelo de Poisson inflacionado de Zeros (Cançado et al., 2014) que oferece um bom ajuste para os dados. No entanto, quando sobredispersão ainda persiste, mesmo após o ajuste para a inflação de zeros, um modelo mais robusto deve ser considerado para acomodar a sobredispersão adicional em valores de contagem positivos.

Modelos inflacionados de zeros têm sido usados em muitas áreas (Hall, 2000; Cheung, 2002; Yau et al., 2004). A estimação dos parâmetros empregando o modelo de Poisson inflacionado de Zeros (**ZIP**) também pode ser severamente viciada, quando as contagens positivas são substancialmente dispersas. Nesta situação, os modelos Poisson Generalizado inflacionado Zeros (**ZIGP**), Poisson Duplo inflacionado Zeros (**ZIDP**) ou Binomial Negativo inflacionado Zeros (**ZINB**) podem ser boas alternativas para a modelagem conjunta da inflação de zeros e da sobredispersão nos dados. No contexto de detecção de cluster espacial, uma causa comum para sobredispersão é a correlação espacial (Hossain e Lawson, 2006) e, por outro lado, a inflação de zeros ocorre devido a sub-ou ausência de exposição ao risco de doença para alguns grupos de indivíduos.

Excesso de alarmes falsos pode ocorrer devido à presença simultânea de inflação de zeros e sobredispersão. Num estudo simulado, Perumean-Chaneya et al. (2012) verificaram que as estimativas do modelo base de Poisson são ineficientes, e os resultados estatisticamente significativos podem ser perdidos quando a inflação de zeros é desprezada. Da mesma forma, quando sobredispersão é ignorada, estimativas de erro tipo I são inflacionadas.

No contexto não espacial, Xiang et al. (2007) propuseram um teste escore para detectar sobredispersão baseado em um modelo misto **ZINB**. O mesmo tipo de teste escore foi utilizado através do **ZIGP** (Yang et al., 2010). Outro teste escore considerou simultaneamente a inflação de zeros e sobredispersão (Deng e Paul, 2005) em modelos de regressão **ZINB**.

No contexto espacial, Cançado et al. (2014) propuseram uma Estatística Scan Espacial para modelos inflacionados de zeros **ZIP**, enquanto Zhang et al. (2012) desenvolveram uma Estatística Scan Espacial em dados com sobredispersão baseado em um modelo de mistura Poisson-Gamma.

Os avanços tecnológicos ocorridos nos últimos anos facilitaram a coleta e análise da informação geográfica. Atualmente, os registros feitos pela maioria das agências de saúde pública trazem informações sobre o tempo e o local de ocorrência dos casos das principais doenças que atingem uma população. Uma vez que os registros são atualizados com uma frequência cada vez maior, a demanda por métodos capazes de detectar precocemente mudanças nos padrões espacial e temporal de ocorrência de eventos tem crescido. Este tipo de

método é útil não apenas no contexto de saúde pública. Métodos para detecção prospectiva de mudanças no padrão espacial ou temporal dos valores de um processo estocástico, de forma rápida e eficiente, são de grande interesse em diversas áreas do conhecimento, tais como vigilância de acidentes de trânsito, crimes em grandes cidades, dinâmica ecológica, etc. Vários métodos espaço-temporais de vigilância prospectiva têm sido propostos na literatura (Kulldorff, 2001, Kulldorff et al., 2005, Diggle et al., 2005, Rodeiro e Lawson, 2006, Takahashi et al., 2008 e Tango et al., 2011).

1.1 Motivação e Relevância da Contribuição

Neste trabalho propomos uma modificação na Estatística Scan Espacial baseada no modelo **ZIDP** (Poisson Duplo Inflacionado de Zeros), que acomoda simultaneamente o excesso de zeros e a sobredispersão. Desta forma, os modelos existentes na literatura (Poisson, Poisson inflacionado de zeros e Poisson com sobredispersão) tornam-se casos particulares do modelo proposto. Uma extensão da Estatística Scan para detecção de clusters espaço-temporais em modelos inflacionados de zeros e sobredispersos também é proposta nesta tese.

Não consta na literatura uma Estatística Scan que modele simultaneamente a sobredispersão e o excesso de zeros. Uma outra contribuição, é a elaboração de um novo método para detecção e inferência de clusters espaciais e espaço-temporais.

1.2 Objetivos

Este trabalho tem como objetivo principal desenvolver e avaliar um modelo que leve em consideração o excesso de zeros e a sobredispersão simultaneamente, compará-lo com os modelos já existentes na literatura - Estatística Scan Espacial de Kulldorff, Poisson inflacionado de zeros e Poisson com sobredispersão; e extendê-lo para situações espaço-tempo.

Inferir por máxima verossimilhança os parâmetros sobre os modelos nulo e alternativo através do algoritmo EM (Expectation-Maximization).

1.3 Organização

O conteúdo desta tese está organizado em seis capítulos e dois apêndices. No Capítulo 2, é abordada uma revisão do modelo de Poisson com Sobredispersão e Inflacionado de Zeros, da Estatística Scan Espacial e Espaço-Temporal. No Capítulo 3, é apresentada a Estatística Scan Espacial modificada com sobredispersão e inflacionada de zeros. Estudos numéricos com dados simulados são apresentados no Capítulo 4. No Capítulo 5 é mostrado um exemplo de aplicação usando dados de Hanseníase no Estado do Amazonas - Brasil. As considerações finais são resumidas no Capítulo 6. No apêndice A é mostrado o artigo publicado no Statistica Sinica e no apêndice B são apresentados alguns resultados obtidos nesta tese.

Capítulo 2

Modelos de Poisson

Este capítulo apresenta a Estatística Scan Espacial e Espaço -Temporal, e a formulação do modelo de Poisson com sobredispersão e inflacionado de zeros

2.1 Estatística Scan Espacial

Dada uma região em estudo representada por um mapa geográfico dividido em zonas, cada uma com a sua população em risco e o número de casos da doença, a Estatística Scan Espacial (Kulldorff, 1997) é um método desenvolvido para identificar um cluster (subconjunto de zonas) com elevada incidência de casos em comparação com o resto do mapa. Este emprega um teste de razão de verossimilhança e faz uso de um procedimento de varredura (o scan espacial) para procurar o cluster mais provável entre os muitos candidatos a clusters no espaço ou espaço-tempo. As versões espaciais mais simples impõem janelas de forma circular ou eliptica movendo sobre a região de estudo procurando por clusters compactos (Duczmal et al., 2006, 2011). Especificamente, seja $\mathcal{S}$ uma região de estudo projetada no plano cartesiano com L áreas identificadas por $\{s_1, \ldots, s_L\}$ e população sob risco $n(s_i) = n_i$. É usual determinar no interior de cada área s_i, um ponto (ou $centroide$) a_i no plano. Sob a hipótese de distribuição completamente aleatória dos casos (a hipótese nula H_0), seja $Y_i \sim \mathcal{P}(\theta n_i)$ para todo $s_i \in S$. Seja Z um candidato a cluster. Sob a hipótese alternativa H_1, seja $Y_i \sim \mathcal{P}(\theta_1 n_i)$ para todo $s_i \in Z$ e $Y_i \sim \mathcal{P}(\theta_2 n_i)$ para

todo $s_i \notin Z$ com $\theta_1 > \theta_2$. A função de verossimilhança para Z é dada por,

$$\mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y}) = \left(\prod_{i=1}^{L} \frac{n_i^{y_i}}{y_i!}\right) \theta_1^{\sum_{s_i \in Z} y_i} e^{-\theta_1 \sum_{s_i \in Z} n_i} \theta_2^{\left(\sum_{i=1}^{L} y_i - \sum_{s_i \in Z} y_i\right)} e^{-\theta_2 \left(\sum_{i=1}^{L} n_i - \sum_{s_i \in Z} n_i\right)}. \tag{2.1}$$

A estatística da razão de verossimilhança para $H_0 : \theta_1 = \theta_2 = \theta$ versus $H_1 : \theta_1 > \theta_2$ é

$$\Lambda_Z = \frac{\max_{\theta_1 > \theta_2} \mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y})}{\max_{\theta_1 = \theta_2} \mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y})}$$

$$= \left(\frac{\sum_{s_i \in Z} y_i / \sum_{s_i \in Z} n_i}{\sum_{i=1}^{L} y_i / \sum_{i=1}^{L} n_i}\right)^{\sum_{s_i \in Z} y_i} \left[\frac{(\sum_{i=1}^{L} y_i - \sum_{s_i \in Z} y_i)/(\sum_{i=1}^{L} n_i - \sum_{s_i \in Z} n_i)}{\sum_{i=1}^{L} y_i / \sum_{i=1}^{L} n_i}\right]^{\sum_{i=1}^{L} y_i - \sum_{s_i \in Z} y_i},$$

se $\dfrac{\sum_{s_i \in Z} y_i}{\sum_{s_i \in Z} n_i} > \left(\sum_{i=1}^{L} y_i - \sum_{s_i \in Z} y_i\right) / \left(\sum_{i=1}^{L} n_i - \sum_{s_i \in Z} n_i\right)$ e $\Lambda_Z = 1$, caso contrário.

Seja $\mathcal{Z}$ a coleção de todos os Z candidatos a cluster. Então, a estatística scan espacial é definida por

$$\Lambda = \max_{Z \in \mathcal{Z}} \Lambda_Z. \tag{2.2}$$

Como a distribuição de Λ sob H_0 é intratável analiticamente, a probabilidade de significância do teste é obtida via simulação de Monte Carlo.

2.2 Poisson inflacionado de zeros - ZIP

Suponha que existam L localizações s_i, $i = 1, 2, \cdots, L$ e seja $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))'$, onde $Y_i \equiv Y(s_i)$ é a variável aleatória que representa o número de casos de uma determinada doença na localização s_i com população em risco n_i e valor observado y_i. Modelos inflacionados de zeros (ZI) para Y_i são representados pela mistura de uma distribuição degenerada

no zero com um modelo probabilístico padrão. Esse modelos são usados quando a contagem de zeros observada excede a contagem de zeros esperada pelo modelo padrão. Um exemplo típico é o modelo de Poisson inflacionado de zeros - **ZIP** (Cançado et al., 2014), o qual assume

$$Y_i \sim \begin{cases} 0 & \text{com probabilidade } p; \\ \mathcal{P}(\mu_i) & \text{com probabilidade } 1-p; \end{cases}$$

onde $\mathcal{P}$ denota a distribuição de Poisson. A primeira parte indica que alguns zeros ocorrem com probabilidade p, enquanto a outra parte envolve uma distribuição $\mathcal{P}(\mu_i)$ com probabilidade $1-p$. Este modelo tem distribuição de probabilidade,

$$P(Y_i = y_i) = \begin{cases} p + (1-p)e^{-\mu_i} & y_i = 0; \\ (1-p)\mathcal{P}(\mu_i) & y_i = 1, 2, \ldots \end{cases}$$

Se Y_i possui distribuição inflacionada de zeros. Então,

$$\mathbb{E}(Y_i) = (1-p)\mu_i \quad \text{e} \quad \mathbb{V}(Y_i) = (1-p)\sigma_i^2 + p(1-p)\mu_i^2, \tag{2.3}$$

onde (μ_i, σ_i^2) denotam, respectivamente, a esperança e a variância do modelo padrão usado na mistura e p representa o parâmetro que mede a inflação de zeros. Os resultados em 2.3 mostram que se a inflação de zeros for ignorada no modelo, os estimadores obtidos para os parâmetros serão inadequados, independentemente da distribuição de probabilidade utilizada nos dados de contagem.

2.3 Poisson com sobredispersão e inflacionado de zeros (ZIOP) - proposto nesta tese

Um outro fator ignorado pelos modelos usuais em processos de contagem é a sobredispersão. Quando os dados possuem variação maior que a predita pelo modelo probabilístico usado, dizemos que eles são sobredispersos. A ocorrência da sobredispersão é comum quando os dados são modelados segundo as distribuições de Poisson e Binomial, pois muitas vezes os dados apresentam dispersão maior que a prevista por esses modelos. Dois mecanismos que

podem causar a sobredispersão são:

i) os dados são gerados por um processo de mistura de diferentes distribuições;

ii) as observações são positivamente correlacionadas em vez de independentes.

Para tratar a sobredispersão, os modelos Binomial Negativo (**BN**)(Winkelmann, 2003), Beta Binomial (**BB**) (Lora e Singer, 2008), Poisson Generalizado (**GP**) (Consul e Jain, 1973) e Poisson Duplo (**DP**) (Efron, 1986) podem ser usados. No contexto de inflação de zeros no modelo de Poisson, os modelos **ZIGP** (Famoye e Singh, 2006) e **ZIDP** (Lima et al., 2015) podem ser usados para acomodar a sobredispersão no **ZIP** (Cançado et al., 2014). Em particular, é considerado neste trabalho, o modelo para a sobredispersão representado por um Poisson Duplo - **DP** (Efron, 1986) que possui função de probabilidade

$$\tilde{f}_{DP}(y_i|\mu_i,\phi) = c(\mu_i,\phi) f_{DP}(y_i|\mu_i,\phi), \tag{2.4}$$

onde a constante de normalização satisfaz

$$\frac{1}{c(\mu_i,\phi)} \doteq 1 + \frac{1-\phi}{12\mu_i\phi}\left(1+\frac{1}{\mu_i\phi}\right), \ \ \mu_i > 0 \ \text{ e } \ 0 < \phi < 1,$$

e

$$f_{DP}(y_i|\mu_i,\phi) = (\phi^{1/2} e^{-\phi\mu_i})\left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right)\left(\frac{e^{\mu_i}}{y_i}\right)^{\phi y_i}, \quad y_i = 0,1,2,\cdots \tag{2.5}$$

Para garantir a existência da $f_{DP}(0|\mu_i,\phi)$, considere $0^0 = 1$ e $0log(0) = 0$. Efron (1986) mostra que,

$$\mathbb{E}(Y_i) \doteq \mu_i \ \ \text{e} \ \ \mathbb{V}(Y_i) \doteq \frac{\mu_i}{\phi}, \tag{2.6}$$

e que (2.5) é uma aproximação para (2.4) implicando que a função de probabilidade do **DP** pode ser aproximada por (2.5). Esta distribuição aproximada tem sido usada com sucesso em modelagem de séries temporais com sobredispersão (Heinen, 2003; Xu et al., 2012) e acomoda facilmente o ajuste por covariável. Em (2.6), nota-se que ϕ é o parâmetro que controla a sobredispersão quando $0 < \phi < 1$. Se $\phi = 1$, então o modelo **DP** reduz-se à distribuição de Poisson.

Para modelar conjuntamente o excesso de zeros e a sobredispersão nos dados, propomos o uso do modelo Duplo inflacionados de zeros (**ZIDP**) com função de probabilidade dada por

$$P(Y_i = y_i|p, \mu_i, \phi) = \begin{cases} p + (1-p)f_{DP}(0|\mu_i, \phi) & y_i = 0; \\ (1-p)f_{DP}(y_i|\mu_i, \phi) & y_i = 1, 2, \ldots \end{cases} \tag{2.7}$$

com $\mu_i = \theta n_i$. Combinando (2.3) e (2.6) obtém-se

$$\mathbb{E}(Y_i) = (1-p)\mu_i \quad \text{e} \quad \mathbb{V}(Y_i) = E(Y_i)\left(p\mu_i + \frac{1}{\phi}\right). \tag{2.8}$$

Note que ϕ mede a sobredispersão no modelo de Poisson inflacionado de zeros. Quando $p = 0$ e $\phi = 1$, o modelo $\mathbf{ZIDP}(\mu_i, 1, 0)$ é o modelo de Poisson padrão $\mathcal{P}(\mu_i)$. Quando $p \neq 0$ e $\phi = 1$, o modelo $\mathbf{ZIDP}(\mu_i, 1, p)$ é o modelo **ZIP**.

2.4 Estatística Scan Espaço-Temporal

Nas técnicas de detecção de clusters espaciais um período de tempo é fixado para mapear as taxas de ocorrência do evento de interesse. Este período pode ser de dias, meses ou anos, e a escolha adequada é desconhecida. Se um período curto for incluído, o poder do teste pode não ser suficiente para detectar um evento com um risco baixo a moderado, mas que ocorre há um tempo considerável. Por outro lado, incluindo-se um período longo, o poder do teste pode não ser suficiente para detectar um evento com risco elevado ocorrido num período curto (Kulldorff, 2001).

Kulldorff et al. (1998) propõem uma extensão espaço-temporal da estatística scan de Kulldorff (1997), ampliando-se a estatística de varredura de um formato circular para um formato cilíndrico. A base do cilindro representa o espaço, exatamente como na Estatística Scan Espacial usual, enquanto a altura representa o tempo. Este método é de natureza retrospectiva, projetado para testar se uma doença é aleatoriamente distribuída no espaço e no tempo para uma região e período de tempo predeterminados.

Sob H_0, assume-se que o número de casos, $c_i, i = 1, \cdots, L$, seja distribuído segundo uma distribuição de Poisson com risco constante no espaço e no tempo, e sob H_1, assume-se que

o risco seja distinto dentro e fora de pelo menos um cilindro. A estatística da razão de verossimilhança é

$$\Lambda_Z = \left(\frac{\dfrac{\sum\sum_{(t,s_i)\in Z} y_{it}}{\sum\sum_{(t,s_i)\in Z} n_{it}}}{\dfrac{\sum_{t=1}^{T}\sum_{i=1}^{L} y_{it}}{\sum_{t=1}^{T}\sum_{i=1}^{L} n_{it}}} \right)^{\sum\sum_{(t,s_i)\in Z} y_{it}} \times$$

$$\times \left[\frac{\left(\sum_{t=1}^{T}\sum_{i=1}^{L} y_{it} - \sum\sum_{(t,s_i)\in Z} y_{it} \right) \Big/ \left(\sum_{t=1}^{T}\sum_{i=1}^{L} n_{it} - \sum\sum_{(t,s_i)\in Z} n_{it} \right)}{\dfrac{\sum_{t=1}^{T}\sum_{i=1}^{L} y_{it}}{\sum_{t=1}^{T}\sum_{i=1}^{L} n_{it}}} \right]^{\sum_{t=1}^{T}\sum_{i=1}^{L} y_{it} - \sum\sum_{(t,s_i)\in Z} n_{it}},$$

$$\text{se } \frac{\sum\sum_{(t,s_i)\in Z} y_{it}}{\sum\sum_{(t,s_i)\in Z} n_{it}} > \left(\sum_{t=1}^{T}\sum_{i=1}^{L} y_{it} - \sum\sum_{(t,s_i)\in Z} y_{it} \right) \Big/ \left(\sum_{t=1}^{T}\sum_{i=1}^{L} n_{it} - \sum\sum_{(t,s_i)\in Z} n_{it} \right)$$

2.4.1 Estatística Scan Prospectiva

O desenvolvimento de métodos de vigilância espaço-tempo é uma área de pesquisa importante, recente e bastante ativa, ainda sem um método amplamente aceito, considerado superior aos demais. Algumas propostas de métodos espaço-temporais para vigilância prospectiva foram apresentadas recentemente. Kulldorff (2001) sugere o uso da estatística scan espaço-tempo baseada na estatística scan espacial (Kulldorff, 1997), para monitoramento prospectivo de doenças em dados de área. A varredura no caso espaço-tempo é feita sob todos os cilindros que atingem todo o caminho até o fim do período de estudo e busca-se aquele que maximiza a razão de verossimilhança, baseada no modelo Poisson ou Bernoulli.

Kulldorff et al. (2005) desenvolveram uma estatística scan espaço-tempo de permutação que não requer dados da população de risco, podendo ser aplicada a dados de processos pontuais.

Diggle et al. (2005) e Rodeiro e Lawson (2006) sugeriram métodos bayesianos para modelar a evolução espaço-temporal de taxas de incidência e monitorar mudanças.

Takahashi et al. (2008) propuseram uma estatística scan espaço-tempo flexível para detecção de conglomerados não circulares. A estatística scan flexível considera uma janela prismática tridimensional cuja base tem formato arbitrário.

Tango et al. (2011) argumentam que a estatística scan espaço-tempo de Kulldorff (2001) compara o número observado de casos com o número esperado condicional e sugerem uma nova estatística scan espaço-tempo que compara o número observado de casos com o número esperado não condicional.

Rogerson (2001) propõe um sistema de monitoramento prospectivo periódico no tempo, voltado para o súbito aparecimento de cluster como um fenômeno global em toda a área de estudo.

Na literatura são encontrados diferentes métodos de monitoramento que se propõem a detectar um aumento repentino no risco da doença. Porém, caso o aumento seja localizado, a detecção através destes métodos pode ser prejudicada em função do método estar sendo aplicado para o mapa inteiro. Uma solução possível é monitorar cada região separadamente, gerando, no entanto, em função das muitas áreas, multiplicidade de testes, que pode acarretar a ocorrência de alarmes falsos caso o nível de significância nominal seja utilizado.

Uma possibilidade para efetuar uma vigilância periódica é repetir a detecção puramente espacial periodicamente, para possibilitar a inclusão dos casos mais recentes. Desta forma, o período inicial é fixo e é aplicado o método levando-se em consideração todos os casos inseridos dentro do período inicial até o último período disponível. Periodicamente o limite superior é atualizado, incluindo os casos mais recentes. Esta abordagem aparentemente resolveria a questão do monitoramento de clusters emergentes. Porém, ao aplicá-la incorre-se em dois problemas. O primeiro é a redução do poder de detecção rápida do cluster. Se um risco verdadeiramente alto está presente somente no fim do período, as flutuações aleatórias nos períodos iniciais quando o risco era pequeno gera uma diluição do risco no fim do período. De forma geral, quanto maior o período considerado menor é o poder de

detecção rápida de um cluster emergente. O segundo problema é a ausência de ajuste para múltiplos testes decorrentes de muitas regiões e tamanho de clusters possíveis, devido às análises puramente espaciais repetidas em todo período de tempo.

Uma possível solução seria utilizar apenas os últimos períodos. Porém a determinação do número de períodos a ser considerado pode gerar baixo poder de detecção quando este for pequeno e desta forma insuficiente para detectar um risco pequeno a moderado, ou quando for muito grande e o risco for elevado, só que num curto período apenas, gerando assim uma diluição do risco. A solução para isto, é utilizar uma estatística espaço-temporal (Kulldorff, 2001).

O Scan Espaço-Tempo (Kulldorff, 2001) é um método de vigilância periódico no tempo. Em vez de usar uma janela em duas dimensões, este método usa uma janela cilíndrica em três dimensões. O cilindro é flexível em sua base circular, bem como na data de início, de forma independente um do outro.

Seja o intervalo de tempo $[Y_1, Y_2]$ para o qual os dados existem e sejam s e t as datas de início e fim do cilindro, respectivamente. O cilindro aqui mencionado é exatamente o mesmo relatado em Kulldorff et al. (1998). A estatística scan espaço-tempo prospectivo considera todos os cilindros para os quais, $Y_1 \leq s \leq t = Y_2$. Isto é, no contexto prospectivo, a estatística scan espaço-tempo considera apenas os clusters "ativos", ou seja, aqueles que atingem o tempo atual.

A estatística de teste da razão de verossimilhança é construída da mesma forma que para estatística scan espacial. Os casos são gerados de modo que as áreas e os intervalos de tempo possuam um número aleatório de casos, independentemente um do outro. A inferência estatística é ajustada para vários intervalos de tempo possíveis.

Capítulo 3

Estatística Scan com sobredispersão e inflacionada de zeros

Neste capítulo é apresentada a Estatística Scan Espacial modificada baseada no modelo Poisson Duplo inflacionado de zeros e sua extensão para o espaço-tempo, propostas nesta tese. Os parâmetros do modelo nulo e alternativo são estimados pelo algoritmo EM. A probabilidade de significância do teste (p-valor) é obtida através do Fast Double Bootstrap Test com o procedimento EM.

3.1 Estatística Scan Espacial para modelos ZIDP

Suponha que para acomodar simultaneamente o excesso de zeros e a sobredispersão, os dados $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))'$ são modelados pelo $\mathbf{ZIDP}(\mu_i, \phi, p)$, com distribuição de probabilidade dada em (2.7). Seguindo o modelo de cluster de Kulldorff (1997), assuma que $\mu_i = \theta_1 n_i$, se $s_i \in Z$ e $\mu_i = \theta_2 n_i$, se $s_i \notin Z$. Considere o problema de teste de hipóteses para $H_0 : \theta_1 = \theta_2 = \theta$ versus $H_1 : \theta_1 > \theta_2$. Para um dado Z, sob H_1, a função de verossimilhança é

$$\mathcal{L}_Z(p,\theta_1,\theta_2,\phi;\mathbf{y}) =$$

$$\prod_{s_i\in Z}\left[p+(1-p)f_{DP}(0|\theta_1 n_i,\phi)\right]^{1-I(y_i>0)}\left[(1-p)f_{DP}(y_i|\theta_1 n_i,\phi)\right]^{I(y_i>0)}\times$$

$$\prod_{s_i\notin Z}\left[p+(1-p)f_{DP}(0|\theta_2 n_i,\phi)\right]^{1-I(y_i>0)}\left[(1-p)f_{DP}(y_i|\theta_2 n_i,\phi)\right]^{I(y_i>0)},$$

onde $I(y_i>0)$ é a função indicadora da ocorrência de valor positivo.
Sob H_0, a função de verossimilhança é

$$\mathcal{L}_0(p,\theta,\phi;\mathbf{y}) =$$

$$\prod_{i=1}^{L}\left[p+(1-p)f_{DP}(0|\theta n_i,\phi)\right]^{1-I(y_i>0)}\left[(1-p)f_{DP}(y_i|\theta n_i,\phi)\right]^{I(y_i>0)}.$$

Sejam $(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1)$ e $(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0)$ os estimadores de máxima verossimilhança para os parâme-tros do modelo sob H_1 e H_0, respectivamente. Então a razão de verossimilhança e a Estatística Scan Espacial para o modelo ZIDP são, respectivamente

$$\hat{\Lambda}_Z = \frac{\mathcal{L}_Z\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1;\mathbf{y}}{\mathcal{L}_0\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0;\mathbf{y}}$$

e

$$\hat{\Lambda} = \max_{Z\in\mathcal{Z}}\hat{\Lambda}_Z, \tag{3.1}$$

com cluster estimado igual a $\hat{Z} = arg\left(\max_{Z\in\mathcal{Z}}\hat{\Lambda}_Z\right)$, sendo $\mathcal{Z}$ a coleção de todos os Z candidatos a cluster. Analisando $\mathcal{L}_Z(.;\mathbf{y})$ e $\mathcal{L}_0(.;\mathbf{y})$, nota-se que não há independência entre p e os demais parâmetros. Assim, a maximização da verossimilhança torna-se complicada, principalmente quando há covariáveis envolvidas. Desta forma, é necessária a inclusão de um vetor de variáveis latentes para que se possa fatorar a verossimilhança e facilitar o processo de maximização usando o algoritmo EM (Expectation-Maximization).

O algoritmo EM é uma ferramenta computacional utilizada para o cálculo do estimador de máxima verossimilhança (EMV) de forma iterativa (Dempster et al., 1977) e tem como ideia base substituir uma difícil maximização da verossimilhança por uma sequência de

maximizações mais fáceis, cujo limite é a resposta para o problema original (Casella e Berger, 2010). A demostração da convergência do algoritmo EM é apresentada em Boyles (1983) e Wu (1983).

Seja o vetor de variáveis latentes $\mathbf{U} = (U_1, \ldots, U_L)$,

$$\text{onde } U_i = \begin{cases} 1, & \text{quando } Y_i \text{ ocorre devido a um zero estrutural;} \\ 0, & \text{quando } Y_i \text{ ocorre devido a um modelo } \mathbf{DP}. \end{cases}$$

Assumindo que $U_i \sim$ Bernoulli(p), a função de verossimilhança aumentada é:

$$\begin{aligned} \mathcal{L}_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}) &= f_Z(\mathbf{y}, \mathbf{u}|p, \theta_1, \theta_2, \phi) \\ &= \prod_{s_i \in Z} f(y_i|U_i = u_i, \theta_1, \phi) P(U_i = u_i|p) \\ &\times \prod_{s_i \notin Z} f(y_i|U_i = u_i, \theta_2, \phi) P(U_i = u_i|p) \end{aligned}$$

que simplifica em

$$\mathcal{L}_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}) = \prod_{s_i \in Z} p^{u_i}[(1-p) f_{DP}(y_i|\theta_1 n_i, \phi)]^{1-u_i} \times \prod_{s_i \notin Z} p^{u_i}[(1-p) f_{DP}(y_i|\theta_2 n_i, \phi)]^{1-u_i}.$$

Marginalmente $Y_i \sim$ **ZIDP**(μ_i, ϕ, p). O logaritmo da razão de verossimilhança para o modelo ZIDP, sob H_1, é

$$\begin{aligned} l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}) &= \sum_{i=1}^{L} [u_i \log p + (1 - u_i) \log(1 - p)] \\ &+ \sum_{s_i \in Z} [(1 - u_i) \log f_{DP}(y_i|\theta_1 n_i, \phi)] \\ &+ \sum_{s_i \notin Z} [(1 - u_i) \log f_{DP}(y_i|\theta_2 n_i, \phi)] \\ &= l_Z^a(p; \mathbf{u}) + l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u}) + l_Z^a(\theta_2, \phi; \mathbf{y}, \mathbf{u}) \end{aligned} \tag{3.2}$$

e sob H_0,

$$\begin{aligned} l_Z^0(p,\theta,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}[u_i \log p + (1-u_i)\log(1-p)] \\ &+ \sum_{i=1}^{L}[(1-u_i)\log f_{DP}(y_i|\theta n_i,\phi)] \\ &= l_Z^0(p;\mathbf{u}) + l_Z^0(\theta,\phi;\mathbf{y},\mathbf{u}). \end{aligned} \tag{3.3}$$

Desta forma, a função de verossimilhança pode ser facilmente maximizada e os estimadores $(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1)$ e $(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0)$ podem ser obtidos independentemente.

3.2 Algoritmo EM

Sob H_1, o estimador para ϕ é obtido maximizando-se $l_Z^a(\phi;\mathbf{y},\mathbf{u}) = l_Z^a(\hat{\theta}_1,\phi;\mathbf{y},\mathbf{u}) + l_Z^a(\hat{\theta}_2,\phi;\mathbf{y},\mathbf{u})$ e sob H_0, é obtido maximizando-se $l_Z^0(\hat{\theta}_0,\phi;\mathbf{y},\mathbf{u})$. Para maximizar (3.2) e (3.3), usamos o algoritmo EM. Neste caso o logaritmo da função de verossimilhança é maximizado iterativamente em dois passos até a convergência. A maximização de $l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})$ é obtida da seguinte forma:

- Passo E: Inicialize o processo iterativo com $\gamma^{(0)} = (p_1^{(0)},\theta_1^{(0)},\theta_2^{(0)},\phi_1^{(0)})$ e na $(k+1)$-ésima iteração a estimativa de $u_i^{(k)}$ é a esperança condicional de $l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})$ sobre $\mathbf{y}$ e as estimativas correntes $\gamma^{(k)} = (p_1^{(k)},\theta_1^{(k)},\theta_2^{(k)},\phi_1^{(k)})$, isto é, calcule $\mathbb{E}\{l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})|\mathbf{y},\boldsymbol{\gamma}^{(k)}\}$ com respeito a distribuição condicional de $\boldsymbol{\gamma}^{(k)}$. Como $l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u})$ é linear em $\mathbf{u}$, esta quantidade é dada por $l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}^{(k)})$, onde $\mathbf{u}^{(k)} = \mathbb{E}_{H_1}(\mathbf{u}|\mathbf{y},\boldsymbol{\gamma}^{(k)})$ com o i-ésimo elemento dado por:

$$u_i^{(k)} = P_{H_1}(u_i=1|y_i,\gamma^{(k)}) = \frac{P_{H_1}(Y_i=y_i|u_i=1,\gamma^{(k)})P_{H_1}(u_i=1|p_1^{(k)})}{P_{H_1}(Y_i=y_i|u_i=1,\gamma^{(k)})P_{H_1}(u_i=1|p_1^{(k)}) + P_{H_1}(Y_i=y_i|u_i=0,\gamma^{(k)})P_{H_1}(u_i=0|p_1^{(k)})}$$

e

$$u_i^k = \begin{cases} \left(1+exp\left\{-\log\left[p_1^{(k)}/(1-p_1^{(k)})\right] - \phi_1^{(k)}\theta_1^{(k)}n_i + \frac{1}{2}\log\phi_1^{(k)}\right\}\right)^{-1}, & \text{se } y_i=0, s_i\in Z \\ \left(1+exp\left\{-\log\left[p_1^{(k)}/(1-p_1^{(k)})\right] - \phi_1^{(k)}\theta_2^{(k)}n_i + \frac{1}{2}\log\phi_1^{(k)}\right\}\right)^{-1}, & \text{se } y_i=0, s_i\notin Z \\ 0, & \text{se } y_i>0 \end{cases}$$

- Passo M: Maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$.

1. Passo M para p: Na $(k+1)$-ésima iteração maximiza-se $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ com respeito a p que é equivalente a maximizar $l_Z^a(p; \mathbf{u})$ dada em (3.2) considerando $\mathbf{u} = \mathbf{u}^{(k)}$. Analiticamente, obtém-se que $p_1^{(k+1)} = \sum_{i=1}^{L} \frac{u_i^{(k)}}{L}$ sendo L o número de localizações e $\hat{p}_1$ é o valor $p_1^{(k+1)}$ que satisfaz $|p_1^{(k+1)} - p_1^{(k)}| < \epsilon$.

2. Passo M para θ_1: Na $(k+1)$-ésima iteração maximiza-se $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ com respeito a θ_1 que é equivalente a maximizar $l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u})$ dada em (3.2) considerando $\mathbf{u} = \mathbf{u}^{(k)}$. Analiticamente, obtém-se que $\theta_1^{(k+1)} = \frac{\sum_{s_i \in Z} (1 - u_i^{(k)}) y_i}{\sum_{s_i \in Z} (1 - u_i^{(k)}) n_i}$ sendo n_i a população em risco e $\hat{\theta}_1$ é a quantidade $\theta_1^{(k+1)}$ que satisfaz $|\theta_1^{(k+1)} - \theta_1^{(k)}| < \epsilon$.

3. Passo M para θ_2: É realizado de forma similar ao passo M para θ_1 substituindo $l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u})$ por $l_Z^a(\theta_2, \phi; \mathbf{y}, \mathbf{u})$. Obtém-se que $\theta_2^{(k+1)} = \frac{\sum_{s_i \notin Z} (1 - u_i^{(k)}) y_i}{\sum_{s_i \notin Z} (1 - u_i^{(k)}) n_i}$ e $\hat{\theta}_2$ é a quantidade $\theta_2^{(k+1)}$ que satisfaz $|\theta_2^{(k+1)} - \theta_2^{(k)}| < \epsilon$.

4. Passo M para ϕ: Na $(k+1)$-ésima iteração maximiza-se

$$l_Z^a(\theta_1^{(k+1)}, \phi; \mathbf{y}, \mathbf{u}) + l_Z^a(\theta_2^{(k+1)}, \phi; \mathbf{y}, \mathbf{u})$$

com respeito a ϕ considerando $\mathbf{u} = \mathbf{u}^{(k)}$. Analiticamente, obtém-se que

$$\phi_1^{(k+1)} = \frac{\sum_{i=1}^{L} (1 - u_i^{(k)})}{2 \left\{ \sum_{s_i \in Z} (1 - u_i^{(k)}) y_i \log(\theta_i / \theta_1^{(k+1)}) + \sum_{s_i \notin Z} (1 - u_i^{(k)}) y_i \log(\theta_i / \theta_2^{(k+1)}) \right\}}$$

onde $\theta_i = y_i / n_i$ e $\hat{\phi}_1 = min\{1, \phi_1^{(k+1)}\}$ com $\phi_1^{(k+1)}$ satisfazendo $|\phi_1^{(k+1)} - \phi_1^{(k)}| < \epsilon$.

A maximização de $l_Z^0(p, \theta, \phi; \mathbf{y}, \mathbf{u})$ ocorre de maneira similar a maximização de $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$, com a seguinte modificação:

No passo E, sob H_0, tem-se que

$$u_i^k = \begin{cases} \left(1+ exp\left\{-\log\left[p_0^{(k)}/(1-p_0^{(k)})\right] - \phi_0^{(k)}\theta_0^{(k)}n_i + \frac{1}{2}\log\phi_0^{(k)}\right\}\right)^{-1}, & \text{se } y_i = 0, i = 1,\ldots,L; \\ 0, & \text{se } y_i > 0 \end{cases}$$

Maximizando $l_Z^0(p,\theta,\phi;\mathbf{y},\mathbf{u}^{(k)})$ com respeito aos parâmetros, obtém-se na $(k+1)$-ésima iteração,

$$p_0^{(k+1)} = \sum_{i=1}^{L}\frac{u_i^{(k)}}{L}, \quad \theta_0^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})y_i}{\sum_{i=1}^{L}(1-u_i^{(k)})n_i} \quad \text{e} \quad \phi_0^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\left\{\sum_{i=1}^{L}(1-u_i^{(k)})\, y_i \log\left[\frac{\theta_i}{\theta_0^{(k+1)}}\right]\right\}}$$

Após a convergência do algoritmo, denote as estimativas via algoritmo EM por $(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1)$, $(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0)$ e calcule $(\hat{\Lambda}_Z,\hat{\Lambda})$ dadas em (3.1). Agora, usando $\hat{\Lambda}$ o cluster espacial pode ser identificado em dados com excesso de zeros e sobredispersão.

3.3 Estatística Scan ZIOP Espaço-Tempo - proposta nesta tese

Nesta seção é apresentada a versão das expressões mostradas nas seções anteriores (3.1 e 3.2) incluindo a dimensão temporal. A analogia do caso espacial para o caso espaço-temporal ocorre ao pensarmos nas L regiões disponíveis nos T períodos de tempo como um mapa contendo $L \times T$ regiões. Desta forma, para acomodar simultâneamente o excesso de zeros e a sobredispersão, os dados $\mathbf{Y} = (Y(s_{11}),\cdots,Y(s_{LT}))'$ são modelados pelo $\mathbf{ZIDP}(\mu_{it},\phi_t,p_t)$, com distribuição de probabilidade dada por:

$$P(Y_{it} = y_{it}|p_t,\mu_{it},\phi_t) = \begin{cases} p_t + (1-p_t)f_{DP}(0|\mu_{it},\phi_t) & y_{it} = 0; \\ (1-p_t)f_{DP}(y_{it}|\mu_{it},\phi_t) & y_{it} = 1,2,\ldots \end{cases} \tag{3.4}$$

Considere o teste de hipótese $H_0 : \theta_{1t} = \theta_{2t} = \theta_t$ em todos os períodos estudados versus $H_1 : \theta_{1t} > \theta_{2t}$ num conjunto específico de períodos de tempo contíguos. Sob H_1, a função de verossimilhança é

$$\mathcal{L}_Z(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y_t}) =$$

$$\prod_{t=1}^{T} \prod_{s_i \in Z} [p_t + (1-p_t) f_{DP}(0|\theta_{1t} n_{it}, \phi_t)]^{1-I(y_{it}>0)} [(1-p_t) f_{DP}(y_{it}|\theta_{1t} n_{it}, \phi_t)]^{I(y_{it}>0)} \times$$

$$\prod_{t=1}^{T} \prod_{s_i \notin Z} \left[p_t + (1-p_t) f_{DP}(0|\theta_{2t} n_{it}, \phi_t)\right]^{1-I(y_{it}>0)} \left[(1-p_t) f_{DP}(y_{it}|\theta_{2t} n_{it}, \phi_t)\right]^{I(y_{it}>0)},$$

onde $I(y_{it} > 0)$ é a função indicadora da ocorrência de valor positivo.

Sob H_0, a função de verossimilhança é dada por:

$$\mathcal{L}_0(p_t, \theta_t, \phi_t; \mathbf{y_t}) =$$

$$\prod_{t=1}^{T} \prod_{i=1}^{L} \left\{ [p_t + (1-p_t) f_{DP}(0|\theta_t n_{it}, \phi_t)]^{1-I(y_{it}>0)} [(1-p_t) f_{DP}(y_{it}|\theta_t n_{it}, \phi_t)]^{I(y_{it}>0)} \right\}.$$

Seja $\mathbf{U}=(U_{1t}, \cdots, U_{Lt})$ o vetor de variáveis latentes,

$$U_{it} = \begin{cases} 1, & \text{quando } Y_{it} \text{ ocorre devido a um zero estrutural;} \\ 0, & \text{quando } Y_{it} \text{ ocorre devido a um modelo } \mathbf{DP}. \end{cases}$$

A função de verossimilhança aumentada é dada por:

$$\mathcal{L}_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y_t}, \mathbf{u_t}) =$$

$$\prod_{t=1}^{T} \prod_{s_i \in Z} \left\{ p_t^{u_{it}} [(1-p_t) f_{DP}(y_{it}|\theta_{1t} n_{it}, \phi_t)]^{1-u_{it}} \right\} \times \prod_{t=1}^{T} \prod_{s_i \notin Z} \left\{ p_t^{u_{it}} [(1-p_t) f_{DP}(y_{it}|\theta_{2t} n_{it}, \phi_t)]^{1-u_{it}} \right\}.$$

O logaritmo da razão de verossimilhança para o modelo ZIDP, sob H_1, é

$$\begin{aligned} l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y_t}, \mathbf{u_t}) &= \sum_{t=1}^{T}\sum_{i=1}^{L}\{[u_{it}\log p_t + (1-u_{it})\log(1-p_t)]\} \\ &+ \sum_{t=1}^{T}\sum_{s_i\in Z}\{[(1-u_{it})\log f_{DP}(y_{it}|\theta_{1t}n_{it}, \phi_t)]\} \\ &+ \sum_{t=1}^{T}\sum_{s_i\notin Z}\{[(1-u_{it})\log f_{DP}(y_{it}|\theta_{2t}n_{it}, \phi_t)]\} \\ &= l_Z^a(p_t; \mathbf{u_t}) + l_Z^a(\theta_{1t}, \phi_t; \mathbf{y_t}, \mathbf{u_t}) + l_Z^a(\theta_{2t}, \phi_t; \mathbf{y_t}, \mathbf{u_t}) \end{aligned} \tag{3.5}$$

e sob H_0,

$$\begin{aligned} l_Z^0(p_t, \theta_t, \phi_t; \mathbf{y_t}, \mathbf{u_t}) &= \sum_{t=1}^{T}\sum_{i=1}^{L}\{[u_{it}\log p_t + (1-u_{it})\log(1-p_t)]\} \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{L}\{[(1-u_{it})\log f_{DP}(y_{it}|\theta_t n_{it}, \phi_t)]\} \\ &= l_Z^0(p_t; \mathbf{u_t}) + l_Z^0(\theta_t, \phi_t; \mathbf{y_t}, \mathbf{u_t}). \end{aligned} \tag{3.6}$$

3.3.1 Algoritmo EM

O algoritmo EM é utilizado para maximizar o logaritmo da função de verossimilhança sob H_1 e H_0 (3.5 e 3.6, respectivamente). A maximização de $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y_t}, \mathbf{u_t})$ é obtida como segue:

- Passo E: Inicialize o processo iterativo com $\gamma_t^{(0)} = (p_{1t}^{(0)}, \theta_{1t}^{(0)}, \theta_{2t}^{(0)}, \phi_{1t}^{(0)})$ e na $(k+1)$-ésima iteração a estimativa de $u_{it}^{(k)} = \mathbb{E}\{l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y_t}, \mathbf{u_t})|\mathbf{y_t}, \boldsymbol{\gamma}_t^{(k)}\}$ com respeito a distribuição condicional de $\boldsymbol{\gamma}_t^{(k)}$. Como $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y_t}, \mathbf{u_t})$ é linear em

$\mathbf{u}_t$, esta quantidade é dada por $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t^{(k)})$, onde $\mathbf{u}_t^{(k)} = \mathbb{E}_{H_1}(\mathbf{u}_t|\mathbf{y}_t, \boldsymbol{\gamma}_t^{(k)})$ com o i-ésimo elemento dado por:

$$u_{it}^{(k)} = P_{H_1}(u_{it} = 1|y_{it}, \gamma_t^{(k)}) = \frac{P_{H_1}(Y_{it} = y_{it}|u_{it} = 1, \gamma_t^{(k)})P_{H_1}(u_{it} = 1|p_{1t}^{(k)})}{P_{H_1}(Y_{it} = y_{it}|u_{it} = 1, \gamma_t^{(k)})P_{H_1}(u_{it} = 1|p_{1t}^{(k)}) + P_{H_1}(Y_{it} = y_{it}|u_{it} = 0, \gamma_t^{(k)})P_{H_1}(u_{it} = 0|p_{1t}^{(k)})}$$

e

$$u_{it}^k = \begin{cases} \left(1 + exp\left\{-\log\left[p_{1t}^{(k)}/(1-p_{1t}^{(k)})\right] - \phi_{1t}^{(k)}\theta_{1t}^{(k)}n_{it} + \frac{1}{2}\log\phi_{1t}^{(k)}\right\}\right)^{-1}, & \text{se } y_{it} = 0, s_{it} \in Z \\ \left(1 + exp\left\{-\log\left[p_{1t}^{(k)}/(1-p_{1t}^{(k)})\right] - \phi_{1t}^{(k)}\theta_{2t}^{(k)}n_{it} + \frac{1}{2}\log\phi_{1t}^{(k)}\right\}\right)^{-1}, & \text{se } y_{it} = 0, s_{it} \notin Z \\ 0, & \text{se } y_{it} > 0 \end{cases}$$

- Passo M: Maximize $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t^{(k)})$.

1. Passo M para p_t: Na $(k+1)$-ésima iteração maximiza-se $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t^{(k)})$ com respeito a p_t que é equivalente a maximizar $l_Z^a(p_t; \mathbf{u}_t)$ dada em (3.5) considerando $\mathbf{u}_t = \mathbf{u}_t^{(k)}$. Analiticamente, obtém-se que $p_{1t}^{(k+1)} = \dfrac{\sum_{t=1}^{T}\sum_{i=1}^{L} u_{it}^{(k)}}{T \times L}$ e $\hat{p}_{1t}$ é o valor $p_{1t}^{(k+1)}$ que satisfaz $|p_{1t}^{(k+1)} - p_{1t}^{(k)}| < \epsilon$.

2. Passo M para θ_{1t}: Na $(k+1)$-ésima iteração maximiza-se $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t^{(k)})$ com respeito a θ_{1t} que é equivalente a maximizar $l_Z^a(\theta_{1t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t)$ dada em (3.5) considerando $\mathbf{u}_t = \mathbf{u}_t^{(k)}$. Analiticamente, obtém-se que $\theta_{1t}^{(k+1)} = \dfrac{\sum_{t=1}^{T}\sum_{s_i \in Z}(1 - u_{it}^{(k)})y_{it}}{\sum_{t=1}^{T}\sum_{s_i \in Z}(1 - u_{it}^{(k)})n_{it}}$, onde n_{it} é a população sob risco no tempo t e $\hat{\theta}_{1t}$ é a quantidade $\theta_{1t}^{(k+1)}$ que satisfaz $|\theta_{1t}^{(k+1)} - \theta_{1t}^{(k)}| < \epsilon$.

3. Passo M para θ_{2t}: É realizado de forma similar ao passo M para θ_{1t} substituindo $l_Z^a(\theta_{1t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t)$ por $l_Z^a(\theta_{2t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t)$. Obtém-se que $\theta_{2t}^{(k+1)} = \dfrac{\sum_{t=1}^{T}\sum_{s_i \notin Z}(1 - u_{it}^{(k)})y_{it}}{\sum_{t=1}^{T}\sum_{s_i \notin Z}(1 - u_{it}^{(k)})n_{it}}$ e

$\hat{\theta}_{2t}$ é a quantidade $\theta_{2t}^{(k+1)}$ que satisfaz $|\theta_{2t}^{(k+1)} - \theta_{2t}^{(k)}| < \epsilon$.

4. Passo M para ϕ_t: Na $(k+1)$-ésima iteração maximiza-se $l_Z^a(\theta_{1t}^{(k+1)}, \phi_t; \mathbf{y}_t, \mathbf{u}_t) + l_Z^a(\theta_{2t}^{(k+1)}, \phi_t; \mathbf{y}_t, \mathbf{u}_t)$ com respeito a ϕ_t considerando $\mathbf{u}_t = \mathbf{u}_t^{(k)}$. Analiticamente, obtém-se que

$$\phi_{1t}^{(k+1)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{L}(1-u_{it}^{(k)})}{2\left\{\sum_{t=1}^{T}\sum_{s_i\in Z}\left[(1-u_{it}^{(k)})y_{it}\log(\theta_{it}/\theta_{1t}^{(k+1)})\right] + \sum_{t=1}^{T}\sum_{s_i\notin Z}\left[(1-u_{it}^{(k)})y_{it}\log(\theta_{it}/\theta_{2t}^{(k+1)})\right]\right\}}$$

onde $\theta_{it} = y_{it}/n_{it}$ e $\hat{\phi}_{1t} = min\{1, \phi_{1t}^{(k+1)}\}$ com $\phi_{1t}^{(k+1)}$ satisfazendo $|\phi_{1t}^{(k+1)} - \phi_{1t}^{(k)}| < \epsilon$.

A maximização de $l_Z^0(p_t, \theta_t, \phi_t; \mathbf{y}_t, \mathbf{u}_t)$ ocorre de maneira similar a maximização de $l_Z^a(p_t, \theta_{1t}, \theta_{2t}, \phi_t; \mathbf{y}_t, \mathbf{u}_t^{(k)})$, com a seguinte modificação:
No passo E, sob H_0, tem-se que

$$u_{it}^k = \begin{cases} \left(1 + exp\left\{-\log\left[p_{0t}^{(k)}/(1-p_{0t}^{(k)})\right] - \phi_{0t}^{(k)}\theta_{0t}^{(k)}n_{it} + \frac{1}{2}\log\phi_{0t}^{(k)}\right\}\right)^{-1}, & \text{se } y_{it} = 0, i = 1,\ldots,L \text{ e } t = 1,\ldots,T; \\ 0, & \text{se } y_{it} > 0 \end{cases}$$

Maximizando $l_Z^0(p_t, \theta_t, \phi_t; \mathbf{y}_t, \mathbf{u}_t^{(k)})$ com respeito aos parâmetros, obtém-se na $(k+1)$-ésima iteração,

$$p_{0t}^{(k+1)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{L}u_{it}^{(k)}}{L}, \quad \theta_{0t}^{(k+1)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{L}(1-u_{it}^{(k)})y_{it}}{\sum_{t=1}^{T}\sum_{i=1}^{L}(1-u_{it}^{(k)})n_{it}} \quad \text{e}$$

$$\phi_{0t}^{(k+1)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{L}(1-u_{it}^{(k)})}{2\left\{\sum_{t=1}^{T}\sum_{i=1}^{L}\left[(1-u_{it}^{(k)})\,y_{it}\log\left(\frac{\theta_{it}}{\theta_{0t}^{(k+1)}}\right)\right]\right\}}$$

Assim, após a convergência do algoritmo EM, as estimativas obtidas são denotadas por $(\hat{p}_{1t}, \hat{\theta}_{1t}, \hat{\theta}_{2t}, \hat{\phi}_{1t})$ e $(\hat{p}_{0t}, \hat{\theta}_{0t}, \hat{\phi}_{0t})$. Calcula-se então a razão de verossimilhança e a estatística

scan como segue.

$$\hat{\Lambda}_Z = \frac{\mathcal{L}_Z \hat{p}_{1t}, \hat{\theta}_{1t}, \hat{\theta}_{2t}, \hat{\phi}_{1t}; \mathbf{y}_t}{\mathcal{L}_0 \hat{p}_{0t}, \hat{\theta}_{0t}, \hat{\phi}_{0t}; \mathbf{y}_t}$$
$$\hat{\Lambda} = \max_{Z \in \mathcal{Z}} \hat{\Lambda}_Z. \tag{3.7}$$

Usando $\hat{\Lambda}$, o cluster pode ser identificado em dados com excesso de zeros e sobredispersão no espaço-tempo.

3.4 Fast Double Bootstrap - EM para cálculo da probabilidade de significância do teste

Como a distribuição de $\hat{\Lambda}$ não pode ser expressa analiticamente, a probabilidade de significância do teste (p-valor) será obtida usando o algoritmo Fast Double Bootstrap Test (Davidson e MacKinnon, 2001) conjuntamente com a aplicação do algoritmo EM para cada novo conjunto de dados gerados sob a hipótese nula. O procedimento Fast Double Bootstrap é necessário nesta situação porque os parâmetros da distribuição de $\hat{\Lambda}$ são desconhecidos sob a hipótese nula. Detalhes sobre a convergência ver Lima et al., 2015 (Apêndice A).

Sob H_0, a distribuição de Y_i é uma mistura das distribuições Bernoulli(p) e $\mathbf{DP}(\theta n_i, \phi)$. Por Efron(1986),

$$\text{se} \quad X_i \sim \mathcal{P}(\theta n_i \times \phi) \Longrightarrow \frac{X_i}{\phi} \sim \mathbf{DP}(\theta n_i, \phi).$$

Dados (p_0, θ_0, ϕ_0), Y_i é gerado de um modelo $\mathbf{ZIDP}(\theta_0 n_i, \phi_0, p_0)$ como segue.

★ Algoritmo $\mathbf{ZIDP}(\theta_0 n_i, \phi_0, p_0)$

1. Gere $x_i \sim \mathcal{P}(\theta_0 n_i \times \phi_0)$ e $v_i \sim Uniforme(0, 1)$.
2. Se $v_i \leq p_0$ faça $y_i = 0$. Caso contrário, $y_i = \dfrac{x_i}{\phi_0}$.

Para calcular o p-valor utiliza-se o passo a seguir:

♦ Algoritmo Fast Double Bootstrap - EM para $\hat{\Lambda}$.

1. Baseado nos dados reais $\mathbf{y} = (y_1, \ldots, y_L)$, use o algoritmo EM e calcule $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$. Derive o valor observado $\hat{\Lambda}$ e denote por $\hat{\lambda}$.

2. Gere $\mathbf{y}_b^* = (y_{1,b}^*, \ldots, y_{L,b}^*)$ usando o algoritmo EM **ZIDP** com (p_0, θ_0, ϕ_0) substituídos por $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$.

3. Com base nos dados gerados $\mathbf{y}_b^*$, use o algoritmo EM e calcule os pseudos estimadores $(\hat{p}_{0,b}^*, \hat{\theta}_{0,b}^*, \hat{\phi}_{0,b}^*)$ para (p_0, θ_0, ϕ_0). Derive o "pseudo"valor de $\hat{\Lambda}_b^*$ e denote por $\hat{\lambda}_b^*$.

4. Repita os passos 2 e 3 para $b = 1, \ldots, B$, calcule o p-valor usual para $\hat{\Lambda}$ por

$$p_{valor}^* \doteq p_{valor}^*(\hat{\Lambda}) = \sum_{b=1}^{B+1} I(\hat{\lambda} \geq \hat{\lambda}_b^*)/(B+1), \text{ com } \hat{\lambda}_{B+1}^* = \hat{\lambda}.$$

5. Gere $\mathbf{y}_b^{**} = (y_{1,b}^{**}, \ldots, y_{L,b}^{**})$ usando o algoritmo **ZIDP** com (p_0, θ_0, ϕ_0) substituídos por $(\hat{p}_{0,b}^*, \hat{\theta}_{0,b}^*, \hat{\phi}_{0,b}^*)$. Usando os passos 3 e 4, derive $\hat{\Lambda}_b^{**}$ e denote por $q_{1-p_{valor}^*}^{**}$ o quantil de ordem $1-p_{valor}^*$ da distribuição empírica de $\hat{\Lambda}_b^{**}$. Este quantil é solução da equação

$$\frac{1}{B}\sum_{b=1}^{B} I(\hat{\Lambda}_b^{**} > q_{1-p_{valor}^*}^{**}) = p_{valor}^*.$$

6. Calcule o p-valor fast double bootstrap para $\hat{\Lambda}$ por

$$p_{valor}^{**} \doteq p_{valor}^{**}(\hat{\Lambda}) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\Lambda}_b^{**} > q_{1-p_{valor}^*}^{**})$$

Capítulo 4

Estudo de simulação

Os efeitos da inflação de zeros e da sobredispersão sobre os quatro modelos baseados na Estatística Scan, para análises espacial e espaço-tempo, são avaliados neste capítulo. A Tabela 4.1 mostra as siglas utilizadas para esses modelos. O **ScanZIOP** e o **ScanZIOPET** são representados pelo modelo **ZIDP**. O **ScanOP** e **ScanOPET** são obtidos do modelo **ZIDP** usando $p = 0$.

Tabela 4.1: Abreviações utilizadas para os modelos de Poisson

Modelos	**Abreviação**	
	Espacial	**Espaço-tempo**
Poisson	ScanP	ScanPET
Poisson Inflacionado de Zeros	ScanZIP	ScanZIPET
Poisson com Sobredispersão	ScanOP	ScanOPET
Poisson com Sobredispersão e Inflacionado de Zeros	ScanZIOP	ScanZIOPET

A região de estudo é o estado do Amazonas no Brasil com $L = 62$ municípios. A população sob risco n_i consiste de crianças menores de 15 anos em 2010 (Figura 4.1). Modelos sob hipótese alternativa com clusters artificiais foram simulados com o objetivo de estimar-se o poder de detecção (probabilidade de alarme verdadeiro) e sob a hipótese nula foram simulados para estimar-se o nível de significância do teste (probabilidade de alarme falso). Para cada modelo a região crítica foi construída utilizando-se o nível de significância nominal de 5%. Foram geradas 1000 replicações de Monte Carlo (sob H_0 e sob H_1) sendo contabilizada a proporção de rejeições

sob H_0, tendo-se as probabilidades estimadas do erro do tipo I e do poder de detecção.

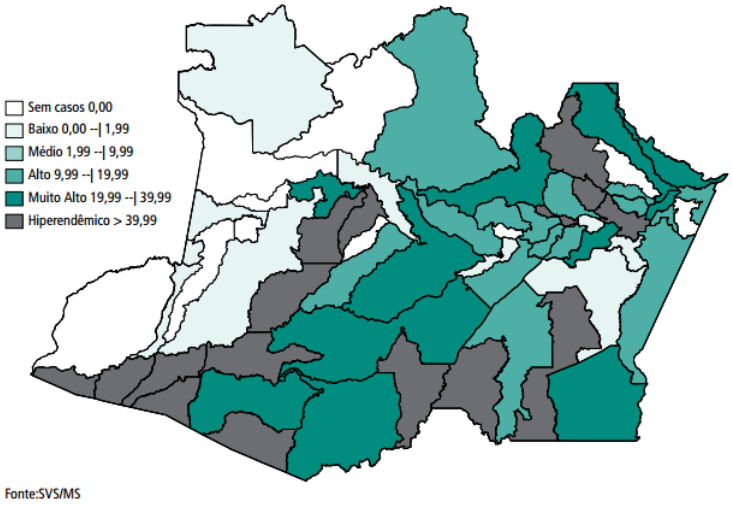


Figura 4.1: Distribuição espacial dos casos de hanseníase em 2010.

O estado do Amazonas é dividido em 13 microrregiões (Figura 4.2), cada uma constituída por um conjunto de municípios limítrofes, com a finalidade de integrar a organização, o planejamento e a execução de funções públicas de interesse comum com base em similaridades econômicas e sociais, em função da praticidade de uso pelo IBGE (Instituto Brasileiro de Geografia e Estatística).

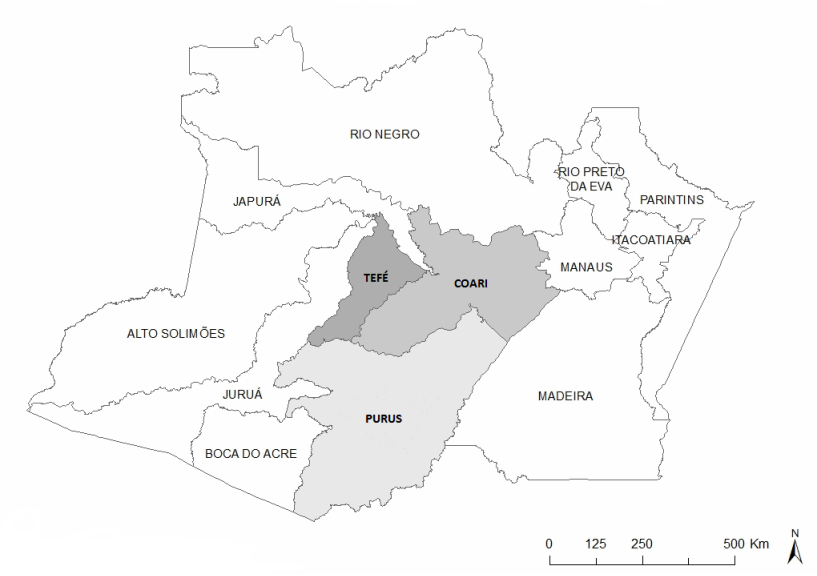
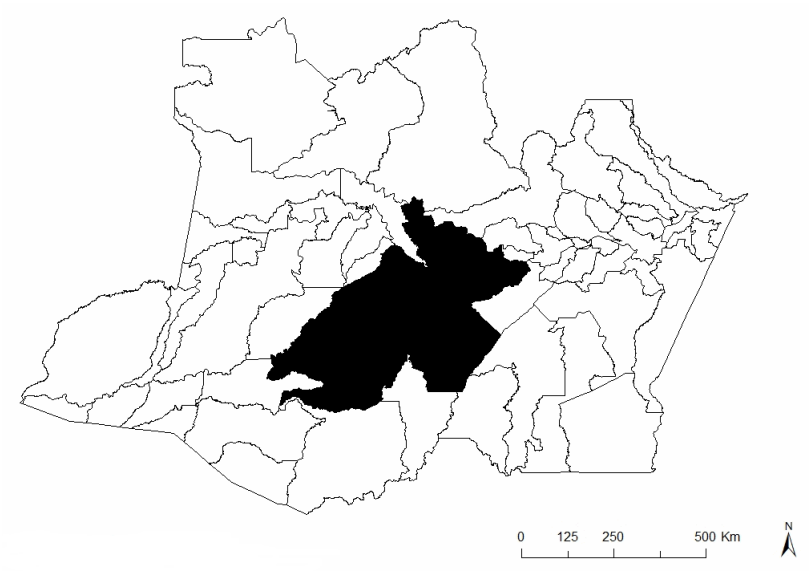
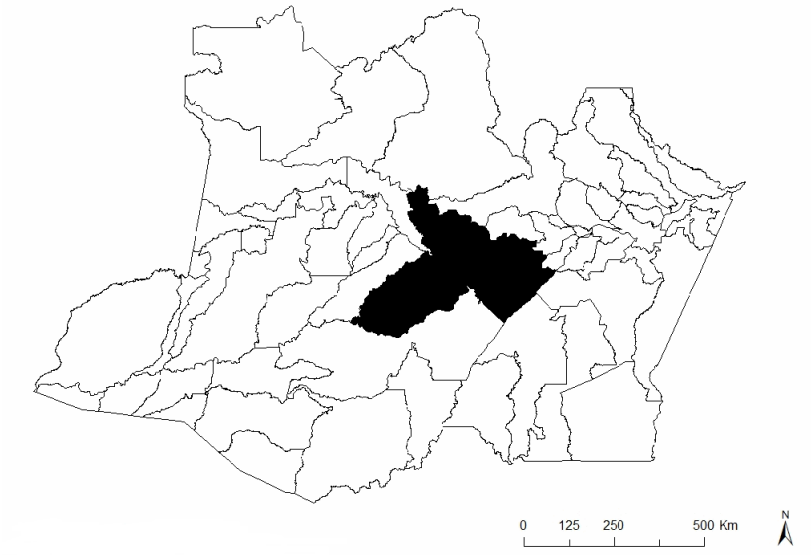


Figura 4.2: Microrregiões do estado do Amazonas.



(a) Cluster artificial - espacial



(b) Cluster artificial - espaço-tempo

Figura 4.3: Clusters artificiais utilizados nas simulações espacial e espaço-tempo.

Um cluster artificial de forma circular Z_E ={Anori, Coari, Codajás, Tapauá, Tefé} constituído por municípios de microrregiões limítrofes, localizado na parte central da região de estudo (Figura 4.3 (a)), foi utilizado para a simulação espacial e Z_{ET} ={Anamã, Anori, Beruri, Coari, Codajás} constituído por municípios de uma mesma microrregião (Figura 4.3

(b)), para a simulação espaço-tempo. Os municípios Anamã, Anori, Beruri, Coari e Codajás são pertencentes a microrregião de Coari, Tapauá pertence a microrregião do Purus e Tefé pertence a microrregião de Tefé. A microrregião de Coari tem no município de Coari, a maior reserva de petróleo e gás em área continental do país (Soares et al., 2006). Trata-se de uma região num processo muito recente de desenvolvimento no âmbito econômico. Todavia, tamanha riqueza, não é sinônimo de melhor qualidade de vida para a população local, já que, estes municípios apresentaram baixo índice de desenvolvimento humano (IDH) [1] em 2010 (IBGE, 2010).

Sob hipótese nula, $\mu_i = n_i\theta_0$, onde $\theta_0 = 0,001$ é uma taxa global de referência para a hanseníase (Ministério da Saúde, 2015). Sob a hipótese alternativa, $\mu_i = n_i\theta_0(1+\lambda)\ \forall\ s_i \in Z$ e $\mu_i = n_i\theta_0$, caso contrário, onde $\lambda > 0$ indica a intensidade do cluster. Note que $\lambda = 0$ sob o modelo nulo.

O procedimento de simulação é dado por:

(1) Gerar 1000 replicações de Monte Carlo sob H_0, com dados gerados por $\mathcal{P}(n_i\theta_0)$ e estimar o quantil superior a 5% para as quatro distribuições empíricas dos métodos **ScanP**, **ScanZIP**, **ScanOP** e **ScanZIOP**.

(2) Gerar 1000 replicações de Monte Carlo sob o modelo nulo ($\lambda = 0$) com sobredispersão $1/\phi = \{1; 1,5; 2,0; 3,0\}$ e inflacionado de zeros $p = \{0; 0,1; 0,2; 0,3\}$; estimar empiricamente o erro tipo I usando o valor crítico dado pelo quantil superior a 5% obtido em **(1)**.

(3) Gerar 1000 replicações de Monte Carlo sob o modelo alternativo ($\lambda = \{0,5; 1,0; 2,0\}$) com sobredispersão $1/\phi = \{1,5; 2,0\}$ e inflacionado de zeros $p = \{0,2; 0,3\}$; estimar empiricamente o poder de detecção usando o valor crítico dado pelo quantil superior a 5% obtido em **(1)**. Seja o cluster mais provável detectado $\hat{Z}^{(q)}$ obtido na q-ésima simulação, o estimador do cluster artificial $Z(\#\{A\}$ indica a cardinalidade do conjunto A).

[1] Índice de desenvolvimento humano - é a referência mundial para avaliar o desenvolvimento humano a longo prazo. O índice varia de 0 a 1 e é feito a partir de três variáveis: vida longa e saudável, acesso ao conhecimento e um padrão de vida decente.

A precisão para a detecção do cluster foi avaliada pelas seguintes medidas:

- Sensibilidade (**SS**) = a razão média do número de localizações corretamente detectadas por um número de localizações pertencentes ao cluster artificial:

$$\mathbf{SS} = \frac{1}{1000}\sum_{q=1}^{1000}\left(\frac{\#\{\hat{Z}^{(q)} \cap Z\}}{\#Z}\right)$$

- Valor Preditivo Positivo (**VPP**) = a razão média do número de localizações corretamente detectadas por um número de localizações pertencentes ao cluster detectado:

$$\mathbf{VPP} = \frac{1}{1000}\sum_{q=1}^{1000}\left(\frac{\#\{\hat{Z}^{(q)} \cap Z\}}{\#\{\hat{Z}^{(q)}\}}\right)$$

As medidas **SS** e **VPP** avaliam o desempenho dos métodos de acordo com suas habilidades em localizar o cluster, quando ele existe.

O procedimento de simulação espaço-tempo foi realizado da mesma forma, só que agora, além de uma varredura espacial é realizada também uma varredura temporal. Para um conjunto de dados aleatórios, os casos são gerados de modo que, as áreas e intervalos de tempo possuam um número aleatório de casos independentemente um do outro. Para facilidade nos cálculos, foi utilizada a mesma população em todos os tempos. O algoritmo calcula a função de verossimilhança de cada janela em três dimensões.

4.1 Análise dos resultados espaciais

Nesta seção, serão apresentados os resultados obtidos nas simulações de varreduras espaciais.

Na ausência de inflação de zeros ($p = 0$) e sobredispersão ($\phi = 1$), todos os modelos apresentaram a taxa de rejeição sob H_0 próxima do valor nominal (5%) - ver Tabela 4.2, exceto o **ScanP** que apresentou uma taxa um pouco abaixo porém razoável (3%). Com inflação de zeros ($p > 0$) e sem sobredispersão ($\phi = 1$), a taxa de rejeição sob H_0 para o **ScanZIP** e **ScanZIOP** permaneceu abaixo de 5%, enquanto o **ScanP** e **ScanOP** apresentaram taxas muito elevadas, mostrando-se ineficientes nesta situação. Por outro lado, na

ausência de inflação de zeros ($p = 0$) e na presença de sobredispersão ($\frac{1}{\phi} > 1$) o **ScanOP** e o **ScanZIOP** apresentaram as menores taxas de rejeição sob H_0, cujos valores são superiores a 5% devido ao fato que seus valores críticos, sob a hipótese nula, num quantil de 5%, foram obtidos assumindo que o modelo de Poisson é verdadeiro. Contudo, estas taxas decrescem à medida que a sobredispersão aumenta. Percebe-se que o **ScanP** e o **ScanZIP** tiveram taxas muito elevadas, mostrando-se inadequados para este cenário.

Tabela 4.2: Estimativas do nível de significância dos métodos (espaciais) para vários valores de p e ϕ

		Métodos			
	$1/\phi$	**ScanP**	**ScanZIP**	**ScanOP**	**ScanZIOP**
	1,00	0,609	0,035	0,393	0,035
$p = 0,3$	1,50	0,813	0,253	0,136	0,090
	2,00	0,894	0,499	0,227	0,086
	3,00	0,974	0,833	0,394	0,085
	1,00	0,390	0,038	0,371	0,040
$p = 0,2$	1,50	0,643	0,282	0,076	0,097
	2,00	0,773	0,520	0,096	0,093
	3,00	0,937	0,868	0,251	0,090
	1,00	0,383	0,042	0,211	0,043
$p = 0,1$	1,50	0,594	0,263	0,026	0,091
	2,00	0,791	0,560	0,030	0,010
	3,00	0,894	0,888	0,080	0,090
	1,00	0,026	0,048	0,056	0,047
$p = 0,0$	1,50	0,236	0,304	0,080	0,097
	2,00	0,382	0,608	0,066	0,095
	3,00	0,726	0,910	0,056	0,090

Quando a inflação de zeros e a sobredispersão ocorrem simultâneamente ($p > 0$ e $\frac{1}{\phi} > 1$), os três primeiros métodos, **ScanP**, **ScanOP** e **ScanZIP** apresentaram maiores taxas de rejeição sob H_0, somente o método **ScanZIOP** se mostrou mais adequado para este cenário, em comparação aos demais métodos, já que, obteve um nível de significância estimado em torno de 9%.

De acordo com a Tabela 4.3, o poder de detecção na presença de sobredispersão e inflação de zeros é maior para o **ScanP** e o **ScanZIP**, como era esperado. Devido ao fato que estes métodos obtiveram altas taxas de rejeição sob H_0, causando assim, uma superestimação do poder de detecção. Somente o **ScanZIOP** apresentou uma estimativa de poder mais confiável para este cenário. Nas simulações, observa-se que o poder do **ScanZIOP** aumenta rapidamente com pequenos aumentos na intensidade de casos no cluster ($\lambda > 0$).

Tabela 4.3: Estimativas do poder dos métodos (espaciais) para vários valores de (λ, p, ϕ)

		Métodos			
	$(p, 1/\phi)$	**ScanP**	**ScanZIP**	**ScanOP**	**ScanZIOP**
$\lambda = 0,5$	(0,2; 1,5)	0,994	0,800	0,195	0,518
	(0,2; 2,0)	0,998	0,830	0,245	0,366
	(0,3; 1,5)	0,998	0,716	0,323	0,442
	(0,3; 2,0)	1,000	0,768	0,355	0,346
$\lambda = 1,0$	(0,2; 1,5)	1,000	0,984	0,452	0,944
	(0,2; 2,0)	1,000	0,982	0,493	0,898
	(0,3; 1,5)	1,000	0,950	0,432	0,882
	(0,3; 2,0)	1,000	0,956	0,478	0,830
$\lambda = 2,0$	(0,2; 1,5)	1,000	1,000	0,894	1,000
	(0,2; 2,0)	1,000	1,000	0,890	0,992
	(0,3; 1,5)	1,000	0,998	0,830	1,000
	(0,3; 2,0)	1,000	0,996	0,821	0,978

Quando a intensidade do cluster e a inflação de zeros permanecem fixadas, o poder decresce a medida que a sobredispersão aumenta. O mesmo efeito é observado quando a intensidade do cluster e sobredispersão permanecem fixadas e a inflação de zeros aumenta. Esta é uma evidência que o **ScanZIOP** é mais adequado para detectar clusters espaciais com pequenos valores de inflação de zeros e sobredispersão.

A partir dos resultados apresentados na Tabela 4.4, a precisão para detectar o verdadeiro cluster quando ele existe (**SS** e **VPP**) é menor para o **ScanOP** sob inflação de zeros e sobredispersão; esta precisão aumenta à medida que aumenta a intensidade do cluster.

O **ScanP** apresenta valores baixos de **VPP** e a sensibilidade diminui com o aumento

Tabela 4.4: Estimativas da sensibilidade (**SS**) e do valor preditivo positivo (**VPP**)

			Métodos			
		$(p, 1/\phi)$	**ScanP**	**ScanZIP**	**ScanOP**	**ScanZIOP**
		(0,2; 1,5)	0,723	0,600	0,176	0,415
	$\lambda = 0,5$	(0,2; 2,0)	0,687	0,568	0,218	0,281
		(0,3; 1,5)	0,651	0,533	0,251	0,351
		(0,3; 2,0)	0,645	0,513	0,288	0,260
		(0,2; 1,5)	0,784	0,909	0,427	0,882
SS	$\lambda = 1,0$	(0,2; 2,0)	0,781	0,846	0,454	0,794
		(0,3; 1,5)	0,742	0,838	0,377	0,798
		(0,3; 2,0)	0,729	0,791	0,415	0,719
		(0,2; 1,5)	0,405	0,950	0,782	0,951
	$\lambda = 2,0$	(0,2; 2,0)	0,423	0,946	0,775	0,944
		(0,3; 1,5)	0,473	0,900	0,675	0,900
		(0,3; 2,0)	0,494	0,890	0,677	0,881
		(0,2; 1,5)	0,277	0,482	0,044	0,335
	$\lambda = 0,5$	(0,2; 2,0)	0,279	0,395	0,049	0,199
		(0,3; 1,5)	0,177	0,424	0,053	0,272
		(0,3; 2,0)	0,188	0,344	0,058	0,164
		(0,2; 1,5)	0,581	0,823	0,287	0,787
VPP	$\lambda = 1,0$	(0,2; 2,0)	0,557	0,749	0,257	0,715
		(0,3; 1,5)	0,375	0,781	0,154	0,732
		(0,3; 2,0)	0,387	0,680	0,171	0,620
		(0,2; 1,5)	0,262	0,977	0,723	0,979
	$\lambda = 2,0$	(0,2; 2,0)	0,268	0,951	0,741	0,949
		(0,3; 1,5)	0,254	0,962	0,618	0,960
		(0,3; 2,0)	0,265	0,920	0,590	0,915

da intensidade do cluster. Esta é uma indicação de que o **ScanP** tende a detectar clusters maiores do que o verdadeiro cluster. Os métodos **ScanZIP** e **ScanZIOP** comportam-se de forma semelhante, em termos de precisão. As medidas de **SS** e de **VPP** aumentam com o aumento da intensidade do cluster. Quando a intensidade do cluster é pequena ($\lambda = 0,5$), **ScanZIP** tem maior precisão do que **ScanZIOP**. No entanto, com o aumento da intensidade

do cluster, as diferenças são insignificantes.

4.2 Análise dos resultados espaço-tempo

Os resultados obtidos nas simulações das varreduras espaço-tempo sob a hipótese nula e alternativa são exibidos nas Tabelas 4.5 e 4.6.

Percebe-se que na ausência de inflação de zeros ($p = 0$) e de sobredispersão ($\phi = 1$) todos os modelos apresentaram a taxa de rejeição sob H_0 próxima do valor nominal (5%), exceto o **ScanZIPET** que teve uma taxa um pouco abaixo porém razoável (3%).

Quando se tem apenas a sobredispersão presente ($\frac{1}{\phi} > 1$ e $p = 0$), o **ScanOPET** e o **ScanZIOPET** apresentaram as menores taxas de rejeição sob H_0, mas superiores a 5%. Porém, estas taxas decrescem à medida que ϕ aumenta. O **ScanPET** e o **ScanZIPET** apresentaram taxas muito elevadas, mostrando-se inadequados para este cenário, já que, o aumento do nível de significância tende a aumentar o poder.

Na presença somente de inflação de zeros ($p > 0$ e $\phi = 1$), o **ScanZIPET** e o **ScanZIOPET** tiveram taxas próximas do valor nominal, enquanto o **ScanPET** e o **ScanOPET** apresentaram taxas muito acima prejudicando a comparação destes com os demais métodos.

Na ocorrência simultânea da inflação de zeros e da sobredispersão ($p > 0$ e $\frac{1}{\phi} > 1$) os métodos **ScanPET** e **ScanZIPET** apresentaram taxas muito elevadas. O **ScanOPET** apresentou taxas próximas de 5% para $p = 0{,}1$ e $0{,}2$. À medida que ϕ decresce as taxas ficam acima do valor nominal (11% e 15%, respectivamente). Para $p = 0,3$, as taxas foram muito elevadas para todos os valores de ϕ . Somente o **ScanZIOPET** obteve um nível de significância estimado próximo a 5% (7 e 9% para $\frac{1}{\phi} = 1{,}5$ e $2{,}0$, respectivamente), exceto para $\frac{1}{\phi} = 3$ em todos os valores de p ($0{,}1; 0{,}2$ e $0{,}3$), onde as taxas foram mais elevadas (11%), mas inferiores as obtidas pelos demais métodos.

Pode-se observar também que na presença e na ausência de inflação de zeros ($p \geq 0$) à

medida que ϕ decresce, as taxas de rejeição sob H_0 aumenta para todos os métodos.

Tabela 4.5: Estimativas do nível de significância dos métodos (espaço-tempo) para vários valores de p e ϕ

		Métodos			
	$1/\phi$	**ScanPET**	**ScanZIPET**	**ScanOPET**	**ScanZIOPET**
	1,00	0,687	0,041	0,352	0,026
$p = 0,3$	1,50	0,786	0,270	0,369	0,071
	2,00	0,889	0,474	0,437	0,092
	3,00	0,987	0,889	0,561	0,109
	1,00	0,551	0,043	0,297	0,032
$p = 0,2$	1,50	0,697	0,329	0,037	0,070
	2,00	0,863	0,451	0,070	0,087
	3,00	0,980	0,863	0,154	0,109
	1,00	0,382	0,054	0,106	0,035
$p = 0,1$	1,50	0,627	0,204	0,030	0,068
	2,00	0,866	0,338	0,060	0,089
	3,00	0,994	0,866	0,112	0,113
	1,00	0,059	0,031	0,050	0,037
$p = 0,0$	1,50	0,215	0,191	0,052	0,055
	2,00	0,476	0,501	0,082	0,082
	3,00	0,796	0,823	0,138	0,113

Apesar do **ScanPET** e do **ScanZIPET**, na presença de inflação de zeros e de sobredispersão, apresentarem poderes de detecção maiores que os obtidos pelos outros métodos (ver Tabela 4.6), estas estimativas não são confiáveis, já que, apresentaram taxas de rejeição sob H_0 muito superiores a 5% para este cenário. Neste caso, somente as estimativas obtidas pelo **ScanZIOPET** são mais confiáveis. Nota-se que com o aumento da inflação de zeros, mantendo a intensidade do cluster (λ) e a sobredispersão ($\frac{1}{\phi}$) fixos, o poder diminui. Analogamente, o poder diminui com o aumento da sobredispersão, mantendo-se fixos a intensidade do cluster (λ) e a inflação de zeros (p). O poder do **ScanZIOPET** apresentou um rápido aumento com pequenos aumentos na intensidade do cluster.

À medida que a intensidade do cluster aumenta o poder dos métodos também aumenta, como esperado.

Tabela 4.6: Estimativas do poder dos métodos (espaço-tempo) para vários valores de (λ, p, ϕ)

		Métodos			
	$(p, 1/\phi)$	**ScanPET**	**ScanZIPET**	**ScanOPET**	**ScanZIOPET**
$\lambda = 0,5$	(0,2; 1,5)	0,996	0,875	0,201	0,510
	(0,2; 2,0)	0,997	0,906	0,278	0,402
	(0,3; 1,5)	0,990	0,828	0,321	0,409
	(0,3; 2,0)	1,000	0,864	0,391	0,368
$\lambda = 1,0$	(0,2; 1,5)	1,000	0,985	0,568	0,951
	(0,2; 2,0)	1,000	0,969	0,601	0,834
	(0,3; 1,5)	1,000	0,957	0,501	0,745
	(0,3; 2,0)	1,000	0,959	0,570	0,738
$\lambda = 2,0$	(0,2; 1,5)	1,000	1,000	0,855	0,997
	(0,2; 2,0)	1,000	1,000	0,849	0,988
	(0,3; 1,5)	1,000	0,998	0,816	0,994
	(0,3; 2,0)	1,000	0,983	0,802	0,980

A Tabela 4.7 mostra os resultados obtidos para a sensibilidade(**SS**) e o valor preditivo positivo (**VPP**) para os métodos **ScanPET**, **ScanOPET**, **ScanZIPET** e **ScanZIOPET**.

Percebe-se que na presença de inflação de zeros e de sobredispersão o **ScanOPET** apresentou a menor precisão em detectar o cluster verdadeiro quando ele existe, para λ = 0,5 e 1,0. À medida que a intensidade do cluster (λ) aumenta, a precisão aumenta para todos os métodos, exceto o **ScanPET** que apresentou baixos valores de **VPP** e a sensibilidade diminui com o aumento da intensidade do cluster. Desta forma, o **ScanPET** possui uma pior capacidade em rejeitar a hipótese nula e tende a detectar clusters maiores que o cluster real. Os métodos **ScanZIPET** e **ScanZIOPET** apresentaram uma precisão bem semelhante. A sensibilidde e o valor preditivo positivo aumentam com o aumento da intensidade do cluster. O **ScanZIPET** mostrou uma precisão maior que o **ScanZIOPET** para pequena intensidade do cluster, λ = 0,5. À medida que a intensidade do cluster aumenta, a precisão destes métodos torna-se cada vez mais semelhante.

Tabela 4.7: Estimativas da sensibilidade (**SS**) e do valor preditivo positivo (**VPP**)

			Métodos			
		$(p, 1/\phi)$	**ScanPET**	**ScanZIPET**	**ScanOPET**	**ScanZIOPET**
SS	$\lambda = 0,5$	(0,2; 1,5)	0,720	0,656	0,175	0,374
		(0,2; 2,0)	0,682	0,536	0,219	0,273
		(0,3; 1,5)	0,636	0,508	0,241	0,368
		(0,3; 2,0)	0,631	0,471	0,206	0,210
	$\lambda = 1,0$	(0,2; 1,5)	0,775	0,902	0,371	0,807
		(0,2; 2,0)	0,761	0,854	0,405	0,798
		(0,3; 1,5)	0,760	0,816	0,396	0,799
		(0,3; 2,0)	0,737	0,705	0,443	0,708
	$\lambda = 2,0$	(0,2; 1,5)	0,411	0,939	0,703	0,979
		(0,2; 2,0)	0,433	0,911	0,706	0,978
		(0,3; 1,5)	0,466	0,897	0,672	0,907
		(0,3; 2,0)	0,498	0,874	0,686	0,840
VPP	$\lambda = 0,5$	(0,2; 1,5)	0,270	0,468	0,083	0,328
		(0,2; 2,0)	0,274	0,379	0,094	0,222
		(0,3; 1,5)	0,171	0,419	0,123	0,266
		(0,3; 2,0)	0,199	0,324	0,138	0,161
	$\lambda = 1,0$	(0,2; 1,5)	0,583	0,814	0,283	0,780
		(0,2; 2,0)	0,516	0,768	0,226	0,726
		(0,3; 1,5)	0,412	0,791	0,135	0,742
		(0,3; 2,0)	0,450	0,705	0,177	0,633
	$\lambda = 2,0$	(0,2; 1,5)	0,251	0,976	0,734	0,979
		(0,2; 2,0)	0,257	0,941	0,764	0,962
		(0,3; 1,5)	0,221	0,952	0,729	0,970
		(0,3; 2,0)	0,255	0,912	0,715	0,911

4.3 Conclusões gerais

Através dos resultados obtidos nas simulações espaço-tempo percebe-se que a inclusão do tempo não modificou as características apresentadas pelos métodos na varredura espacial.

O **ScanP** e o **ScanPET** apresentaram níveis de significância estimados muito altos exceto quando tem-se ausência de inflação de zeros e sobredispersão. O **ScanOP** e o **ScanOPET** nos casos onde não havia sobredispersão apresentou uma taxa de rejeição sob H_0 muito elevada, já nos casos onde a sobredispersão estava presente, as taxas foram mais próximas do valor nominal (5%), exceto quando a inflação de zeros é elevada ($p = 0{,}3$). O nível de significância estimado pelos **ScanZIP** e **ScanZIPET** ficou próximo de 5% nos casos onde não havia sobredispersão e muito acima na presença de sobredispersão. Em geral, o **ScanZIOP** e o **ScanZIOPET** apresentaram as menores taxas de rejeição sob H_0 em todos os cenários. Assim, na presença de inflação de zeros e de sobredispersão apenas os métodos **ScanZIOP** e **ScanZIOPET** mostraram uma estimativa de poder mais confiável.

Capítulo 5

Aplicação: Clusters de Hanseníase

Este estudo utiliza os dados de novos casos de hanseníase em menores de 15 anos no estado do Amazonas, Brasil, de 2008 a 2010, para cada um dos seus 62 municípios . O conjunto de dados foi dividido em dois períodos: 2008/2009 (207 novos casos em dois anos, $\frac{C}{N} = \frac{207}{1245595} = 0{,}0001662$ casos por crianças por 2 anos, ou equivalentemente 0,0000831 casos por criança por ano) e 2010 (190 novos casos, $\frac{C}{N} = \frac{190}{1245595} = 0{,}0001525$ casos por crianças por ano), ver Figuras 5.1 e 5.2.

A hanseníase é uma doença endêmica infecciosa, crônica, causada pelo $Mycobacterium\ leprae$ ($M.\ leprae$) relacionada à pobreza extrema. Afeta, em geral, a pele e os nervos periféricos, embora possua um amplo espectro de manifestações clínicas. Nos países desenvolvidos o controle da hanseníase foi possível, mesmo antes da descoberta de uma droga bactericida, em virtude das garantias de bens sociais que influíram significativamente para o decréscimo da endemia. Enquanto nos países subdesenvolvidos, o quadro epidemiológico da hanseníase ainda é grave. Além dos fatores individuais, as condições sociais desfavoráveis, sem a garantia de moradia digna, de educação, de alimentação adequada, aliado a baixa qualidade da prestação de serviços de saúde estão associados ao aparecimento da doença. Sem nenhuma perspectiva de mudanças de paradigma, a permanência da endemia assume caráter secular. O Brasil apresenta uma das situações mais graves na América Latina, é o principal responsável pela endemia no continente Americano, com 92% dos 36.178 casos novos detectados nas Américas (WHO, 2013).

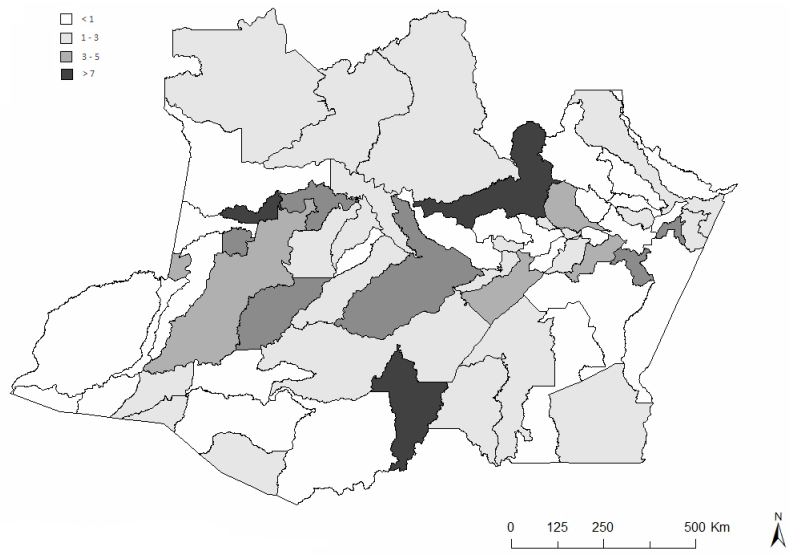


Figura 5.1: Distribuição espacial dos casos de hanseníase em 2008/2009.

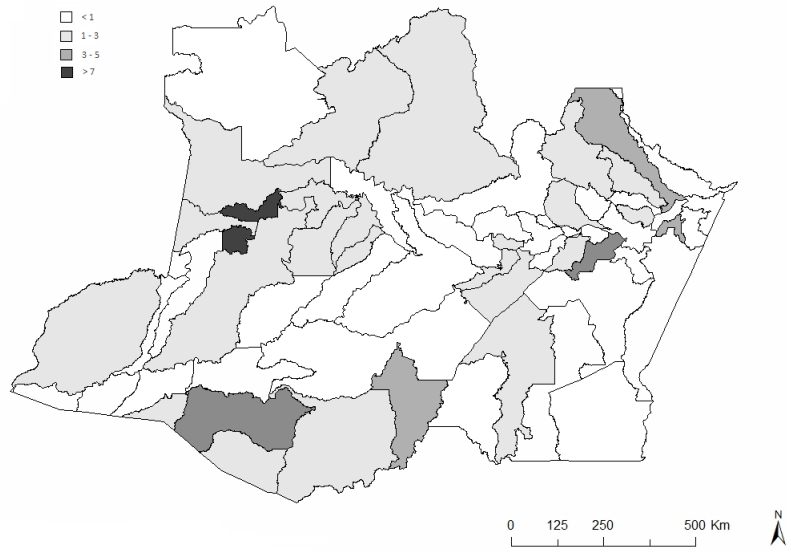


Figura 5.2: Distribuição espacial dos casos de hanseníase em 2010.

Segundo o Ministério da Saúde, foram obtidos resultados importantes no combate a hanseníase nos últimos 10 anos. Apesar da importante redução do coeficiente de prevalência e de detecção de casos novos de hanseníase no Brasil, algumas regiões, como norte, nordeste e centro-oeste, demandam intensificação de ações para eliminação da doença, já que,

apresentam um padrão de alta endemicidade, sendo consideradas como áreas de importante manutenção da transmissão. Um importante obstáculo a ser superado é o abandono ao tratamento.

A detecção de casos de hanseníase em menores de 15 anos foi adotada como principal indicador de monitoramento da endemia, com meta de redução estabelecida em 26% até 2015 (Ministério da saúde, 2012a). Esse indicador expressa a força de transmissão recente, a tendência da endemia e a capacidade dos serviços de saúde na identificação dos sinais e sintomas da doença. A ocorrência da hanseníase nessa faixa etária revela uma exposição precoce ao bacilo, sugerindo presença de casos bacilíferos entre as populações, ou seja, reflete circuitos de transmissão ativos (Ministério da saúde, 2010).

A Organização Mundial da Saúde (OMS) recomenda que a doença seja classificada de acordo com a carga bacilar, paucibacilar (PB) forma menos grave ou multibacilar (MB) com acometimento de certos nervos periféricos, que pode resultar em padrões característicos de incapacidade. A hanseníase pode atingir pessoas de ambos os sexos, de todas as idades. Porém, a incidência da doença é maior nos homens que nas mulheres. Observa-se ainda que crianças, menores de quinze anos, adoecem mais quando há uma maior endemicidade da doença (Ministério da saúde, 2010).

Entre as doenças transmissíveis, a hanseníase é uma das principais causas de incapacidade física permanente. O diagnóstico e o tratamento precoces dos casos, antes que ocorra a lesão neural, são as medidas mais eficazes para se prevenir as incapacidades decorrentes da doença. A abordagem das complicações da hanseníase – incluindo reações e neurites – pode prevenir ou minimizar o desenvolvimento de incapacidades adicionais. A doença e as deformidades a ela associadas são responsáveis pelo estigma social e pela discriminação contra os pacientes e suas famílias em muitas sociedades.

O modo de transmissão do bacilo da hanseníase permanece indeterminado, mas a maioria dos investigadores acredita que isto se dê de pessoa a pessoa, principalmente através de infecção por gotículas nasais. O período de incubação é excepcionalmente longo para uma doença bacteriana: geralmente, de cinco a sete anos. Em geral, a fase na qual ocorrem as primeiras manifestações da hanseníase é o início da vida adulta, entre os 20-30 anos de idade;

raramente a doença é vista em crianças de menos de cinco anos. Embora os seres humanos sejam considerados o principal hospedeiro e reservatório do *M. leprae*, outras fontes animais como o tatu, o macaco mangabei e o chimpanzé, têm sido identificados como reservatórios da infecção. Não se conhece a importância epidemiológica desses achados, mas é provável que seja muito limitada – exceto, talvez, na América do Norte. Sabe-se que a vacinação com o BCG tem algum efeito protetor contra a doença.

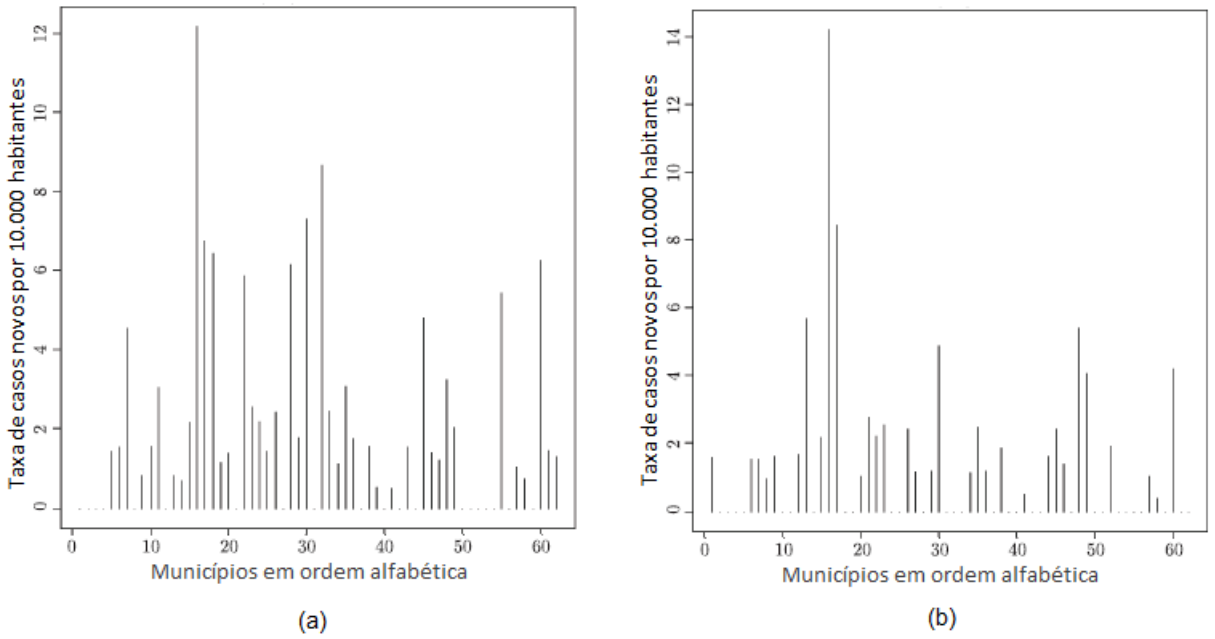


Figura 5.3: Novos casos de hanseníase na Amazônia brasileira, 2008/2009 (a) e 2010 (b).

No período 2008/2009, 20 municípios (32%) não registraram novos casos, em comparação com 30 municípios (48%) que não registraram casos novos só em 2010. No período 2008/2009, a média das taxas de novos casos dos 62 dos municípios para 10 mil pessoas foi de 2,944, com variância igual a 6,776 (no período de dois anos). No período de 2010, os correspondentes valores da média e da variância foram, respectivamente, 2,706 e 7,421 no período de um ano. As Figuras 5.3 (a) e (b) mostram as taxas para os 62 municípios. Como a variância é substancialmente maior do que a média para os estes dois cenários, há uma indicação para o uso do modelo **ZIDP**.

Neste estudo, o Scan Circular (ver seção 2.1) utiliza a coleção de clusters circulares com tamanho máximo $S = 15$ (25% dos municípios), para os quatro modelos do capítulo 3: **ScanP**, **ScanZIP**, **ScanOP** e **ScanZIOP**, a nível de 5% de significância. Os resultados

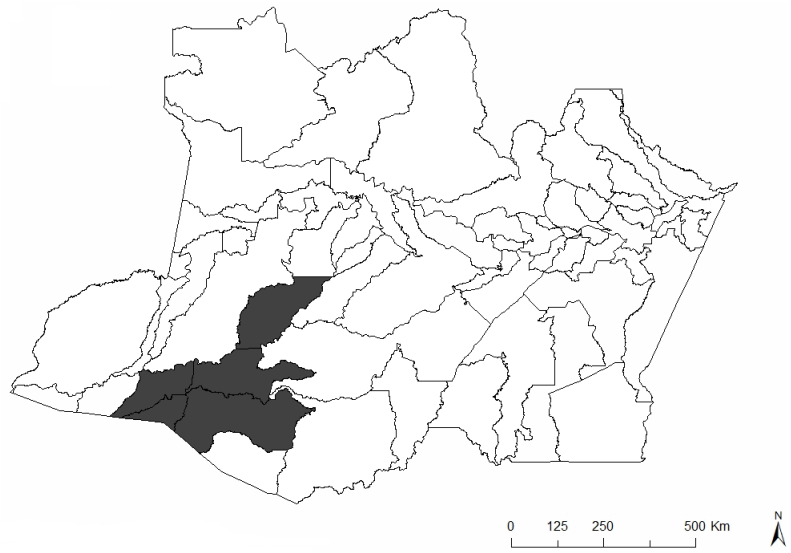
são apresentados na Tabela 5.1.

Tabela 5.1: Cluster espacial de novos casos de Hanseníase, 2008/2009.

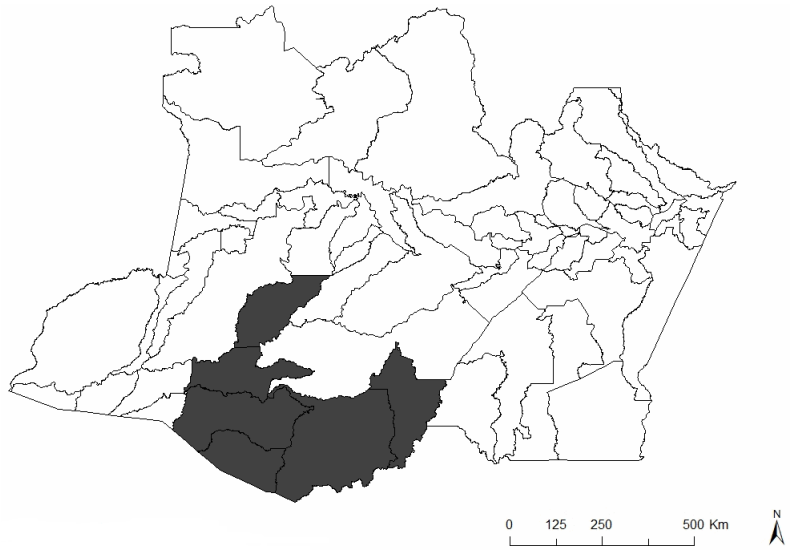
Ano	Scan Espacial	$log\hat{\Lambda}$	p-valor	$(\hat{p}_0, \hat{\phi}_0, \hat{\theta}_0)$	$(\hat{p}_1, \hat{\phi}_1, \hat{\theta}_1, \hat{\theta}_2)$
2008/ 2009	**ScanP**	12,510	0,001	(0,000; 1,000; 0,000166)	(0,000; 1,000; 0,000427; 0,000148)
	ScanZIP	11,074	0,001	(0,010; 1,000; 0,000167)	(0,013; 1,000; 0,000427; 0,000149)
	ScanOP	5,886	0,122	(0,000; 0,428; 0,000166)	(0,000; 0,518; 0,000427; 0,000148)
	ScanZIOP	5,882	0,114	(0,009; 0,430; 0,000167)	(0,013; 0,521; 0,000427; 0,000149)
2010	**ScanP**	16,785	0,001	(0,000; 1,000; 0,000152)	(0,000; 1,000; 0,000518; 0,000134)
	ScanZIP	15,955	0,001	(0,224; 1,000; 0,000171)	(0,013; 1,000; 0,000597; 0,000154)
	ScanOP	7,696	0,034	(0,000; 0,406; 0,000152)	(0,000; 0,520; 0,000517; 0,000134)
	ScanZIOP	8,849	0,006	(0,224; 0,442; 0,000171)	(0,258; 0,689; 0,000597; 0,000154)

No período 2008/2009, **ScanZIOP** e **ScanOP** não detectaram clusters significativos (p-valor = 0,114 e 0,122, respectivamente). A sobredispersão estimada pelo **ScanZIOP** foi $\frac{1}{\hat{\phi}_0} = \frac{1}{0{,}430} = 2{,}325$, a inflação de zeros foi inferior a 1% ($\hat{p}_0 = 0{,}009$) e a taxa de casos foi 1,67 por 10.000 pessoas ($\hat{\theta}_0 = 0{,}000167$). No entanto, o **ScanZIP** e o **ScanP** detectaram um cluster significativo (ambos com p-valor = 0,001) (Carauari, Eirunepé, Envira, Itamarati e Pauini - Figura 5.4 (a)). A inflação de zeros estimada pelo **ScanZIP** foi $\hat{p}_1 = 0{,}013$, com taxa estimada dentro e fora do cluster dada por $\hat{\theta}_1 = 0{,}000427$, $\hat{\theta}_2 = 0{,}000149$, respectivamente (o risco relativo estimado foi 2,866). Levando em conta que **ScanZIP** não acomoda sobredispersão nas contagens positivas, este valor de significância do cluster é duvidoso.

No período de 2010, os quatro métodos detectaram o mesmo cluster (Boca do Acre, Canutama, Carauari, Itamarati, Lábrea e Pauini - Figura 5.4 (b)), com 30 casos novos quando o número esperado era $(1 - \hat{p}_1)n_{\hat{Z}}\hat{\theta}_1 = (1 - 0{,}258) \times 57950 \times 0{,}000597 = 25{,}67$. A inflação de zeros e a sobredispersão estimadas pelo **ScanZIOP** foi $\hat{p}_1 = 0{,}258$ e $\frac{1}{\hat{\phi}_1} = \frac{1}{0{,}689} = 1{,}471$, respectivamente. O cluster está situado em uma região bem conhecida por seu alto índice de vulnerabilidade social.



(a)



(b)

Figura 5.4: Cluster detectado em 2008/2009 (a). Cluster detectado em 2010 (b).

Capítulo 6

Considerações finais

6.1 Comentários gerais sobre os resultados

Uma modificação da Estatística Scan, o Scan Poisson Duplo Inflacionado de Zeros (**ZIPD**), foi proposta para acomodar simultaneamente o excesso de zeros e a sobredispersão. Ela pode ser útil na vigilância de doenças, onde a variação excessiva para contagens positivas é frequente. Às vezes, quando a estatística scan usual é usada sob a hipótese nula de taxa constante, um p-valor pequeno pode ocorrer devido à grande variabilidade no número de casos entre um reduzido número de áreas, ou alternativamente, uma pequena variabilidade entre muitas áreas. Isso pode causar, por sua vez, a existência de um cluster falso positivo; esta anomalia pode ser evitada pela mudança do modelo de Poisson usual por um modelo com sobredispersão. Este tipo de problema ficou evidente no Capítulo 4 (Estudo de simulação) e no Capítulo 5 (Aplicação: Clusters de Hanseníase). As simulações mostraram que na presença de sobredispersão e inflação de zeros o modelo **ZIDP** reduziu substancialmente a probabilidade de erro do tipo I, em comparação com os modelos de Poisson, Poisson com sobredispersão e Poisson inflacionado de zeros, que se mostraram inadequados nestes cenários. Isso significa que, quando um cluster não é detectado pelo **ZIDP**, e é detectado por outros métodos, então ele deve ser analisado cuidadosamente antes de ser reconhecido como um cluster legítimo.

Na presença de sobredispersão para valores de contagem positivos, a detecção de clusters

espaciais baseada no modelo inflacionado de zeros pode não ser a melhor opção e nesta situação, foi mostrado que o Scan Espacial **ZIDP** é uma abordagem mais flexível para este problema, mas não a única. A Binomial Negativa (**NB**), Beta-Binomial (**BB**), Poisson Generalizada (**GP**) pode tratar também a sobredispersão e similarmente ao **ZIPD** também é possível detectar e avaliar clusters espaciais com base nos modelos **ScanZINB**, **ScanZIBB** e **ScanZIGP**. A significância dos clusters encontrados utilizando esses métodos, pode ser calculada também usando a mesma estratégia baseada no Fast Double Bootstrap utilizado neste trabalho.

6.2 Proposta de trabalhos futuros

Dentre as várias possibilidades de trabalhos futuros relacionados ao modelo proposto, destacamos:

- Estudar o efeito da incorporação de covariáveis no modelo de regressão **ZIDP** para a vigilância no espaço-tempo, já que, a hanseníase é uma doença que pode estar ligada à fatores socioeconômicos, tais como a desigualdade de renda, o nível educacional, o crescimento relativo da população e outros.
- A adaptação do modelo **ZIDP** para detecção de conglomerados de forma irregular, diferente das formas circular e cilíndrica, uma vez que com o uso de janelas destas formas é difícil detectar corretamente alguns clusters não circulares, tais como, aqueles ao longo de um rio.
- Implantar o modelo **ZIDP** com abordagem Bayesiana.
- Comparar o modelo **ZIDP** com os modelos **ScanZINB**, **ScanZIBB** e **ScanZIGP**.

Apêndice A

Artigo publicado no Statistica Sinica

Statistica Sinica **25** (2015), 000-000
doi:http://dx.doi.org/10.5705/ss.2013.220w

SPATIAL SCAN STATISTICS FOR MODELS WITH OVERDISPERSION AND INFLATED ZEROS

Max S. de Lima[1], Luiz H. Duczmal[2], José C. Neto[1] and Letícia P. Pinto[2]

[1]*Federal University of Amazonas and* [2]*Federal University of Minas Gerais*

Abstract: The Spatial Scan Statistic is one of the most important methods for detecting and monitoring spatial disease clusters. Usually it is assumed that disease cases follow a Poisson or Binomial distribution. In practice, however, case count datasets frequently present an excess of zeroes and/or overdispersion, resulting in the violation of those commonly used models, increasing type I error occurrence. This paper describes a modification of the Spatial Scan Statistic with the Zero Inflated Double Poisson (ZIDP) model to reduce type I error, accommodating simultaneously an excess of zeroes and overdispersion. The null and alternative model parameters are estimated by the Expectation-Maximization algorithm and the p-value is obtained through the Fast Double Bootstrap Test. An application is presented for Hanseniasis data in the Brazilian Amazon.

Key words and phrases: Double Poisson, EM-algorithm, overdispersion, spatial scan statistics, zero inflated.

1. Introduction

The Spatial Scan Statistics (Kulldorff (1997)) is a popular method for the detection and inference of spatial disease clusters. Recently, several extensions have been devised to accommodate correlation (Loh and Zhu (2007)), covariate adjustment (Jung (2009)), log-linear modeling (Zhang and Lin (2009)), overdispersion (Zhang, Zhang, and Lin (2012)) and zero inflation (Cançado, da-Silva, and da Silva (2011, 2014)). In public health surveillance, the disease count variability is often greater than allowed by the Poisson model, which assumes that the mean and variance have the same value. This variability excess is called overdispersion and has been widely discussed in the literature. Disregarding the presence of overdispersion in the model may lead to the inflation of type I error and consequent erroneous inference for the model parameters. In the presence of overdispersion, the Generalized Poisson (Consul and Jain (1973)) and the Double Poisson (Efron (1986)) are more adequate data models. Another commonly occurring problem in count data, unexpected from the employed model, is that the dataset exhibits an excess of zeroes, or zero inflation. Overdispersion may sometimes occur as a consequence of zero inflation; in this case the Zero-Inflated

Poisson (**ZIP**) model offers a good adjustment to data. However, when overdispersion still persists, after adjusting for zero inflation modeling, a more robust model must be considered to accommodate additional overdispersion in positive count values.

Zero inflated models have been used in many areas (Hall (2000); Cheung (2002); Yau, Lee, and Carrivick (2004)). The estimation of parameters employing **ZIP** may also be severely biased when the positive counts exhibit significantly larger variability than expected. Then, good alternatives, modeling simultaneously zero inflation and overdispersion, are the Zero-Inflated Generalized Poisson (**ZIGP**), Double Poisson (**ZIDP**), or Negative Binomial (**ZINB**) models. In the context of spatial cluster detection, a common cause for overdispersion is spatial correlation (Houssian and Lawson (2006)); on the other hand, zero inflation occurs due to underreporting or absence of disease risk exposure for some groups of individuals.

Excessive false alarm may occur due to the simultaneous presence of zero inflation and overdispersion. In a simulated study, Perumean-Chaneya et al. (2012) verified that the Poisson based model estimates are inefficient, and statistically significant results may be lost when zero inflation is neglected. Likewise, when overdispersion is ignored, type I error estimates are inflated.

In the non-spatial context, a score test was proposed (Xiang et al. (2007)) to detect overdispersion based on a mixed **ZINB** model. The same type of score test was used through **ZIGP** (Yang, Harding, and Addyb (2010)). Another score test considered zero inflation and overdispersion simultaneously (Deng and Paul (2005)) in regression models (**ZINB**).

In the spatial context, a Spatial Scan Statistic for zero-inflated models **ZIP** was proposed (Cançado, da-Silva, and da Silva (2011, 2014)). Further, a Spatial Scan Statistic developed for overdispersion models was presented (Zhang, Zhang, and Lin (2012)), based on a Poisson-Gamma mixture.

In this paper, a modified Spatial Scan Statistics is developed, based on the **ZIDP** model, incorporating simultaneously zero inflation and overdispersion. The null and alternative model parameters are estimated by the EM (Expectation-Maximization) algorithm and the p-value is obtained through the Fast Double Bootstrap Test (Davidson and MacKinnon (2001)).

The paper is organized as follows. Section 2 reviews the Zero-Inflated Overdispersed Poisson model and the Spatial Scan Statistics. Section 3 presents the modified Spatial Scan Statistic with overdispersion and inflated zeros. Numerical studies with simulated data are reported in Section 4. Section 5 shows an application for Hanseniasis data in the Brazilian Amazon. Final remarks are in Section 6.

2. Background

2.1. Zero inflated overdispersed Poisson-ZIOP

Consider L locations with counts given by $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))'$, where $Y_i \equiv Y(s_i)$ is a random variable representing the number of disease cases at location s_i, with population at risk n_i and observed count value y_i. Zero-inflated models for Y_i are employed when the observed zero counts exceed the zero counts expected by the standard model. A typical example is given by the **ZIP** model, which assumes

$$Y_i \sim \begin{cases} 0 & \text{with probability} \quad p, \\ \mathcal{P}(\mu_i) & \text{with probability} \quad 1-p, \end{cases}$$

where $\mathcal{P}$ denotes the Poisson distribution. The resulting distribution is

$$P(Y_i = y_i) = \begin{cases} p + (1-p)e^{-\mu_i} & y_i = 0, \\ (1-p)\mathcal{P}(\mu_i) & y_i = 1, 2, \ldots. \end{cases}$$

It can be shown generally that

$$\mathbb{E}(Y_i) = (1-p)\mu_i \quad \text{and} \quad \mathbb{V}(Y_i) = (1-p)\sigma_i^2 + p(1-p)\mu_i^2, \tag{2.1}$$

where (μ_i, σ_i^2) denotes, respectively, the mean and variance of the standard model and p is the zero inflated parameter. If the zero inflation is ignored in the model, estimators will be inconsistent with the parameters.

Overdispersion appears when data variance is greater than predicted by the probabilistic model. Two mechanisms can cause overdispersion: data is generated by a process consisting of a mixture of two or more distributions; the observed data are not independent, but positively correlated. To treat overdispersion, Negative Binomial (**BB**), Generalized Poisson (**GP**) and Double Poisson (**DP**) models are utilized. Within the zero inflation context, **ZIGP** and **ZIDP** can be used to accommodate overdispersion in the **ZIP** model. Consider here the overdispersion **DP** model, with probability function

$$\tilde{f}_{DP}(y_i|\mu_i, \phi) = c(\mu_i, \phi) f_{DP}(y_i|\mu_i, \phi), \tag{2.2}$$

where the normalization constant satisfies the relation

$$\frac{1}{c(\mu_i, \phi)} = 1 + \frac{1-\phi}{12\mu_i\phi}\left(1 + \frac{1}{\mu_i\phi}\right),$$

and

$$f_{DP}(y_i|\mu_i, \phi) = (\phi^{1/2} e^{-\phi\mu_i}) \left(\frac{e^{-y_i} y_i^{y_i}}{y_i!}\right) \left(\frac{e\mu_i}{y_i}\right)^{\phi y_i}. \tag{2.3}$$

(By convention, $0^0 = 1$ and $0\log(0) = 0$). Efron (1986) shows that

$$\mathbb{E}(Y_i) \doteq \mu_i \quad , \quad \mathbb{V}(Y_i) \doteq \frac{\mu_i}{\phi}, \tag{2.4}$$

and (2.3) is an approximation for (2.2). The approximate distribution has been used with success in temporal series modeling under overdispersion (Heinen (2003); Xu et al. (2012)) and easily accommodates covariate adjustment. In (2.4), it can be seen that ϕ is the parameter controlling overdispersion when $0 < \phi < 1$. If $\phi = 1$, then **DP** is the Poisson distribution.

To model simultaneously the zeroes excess and overdispersion in data, we propose the use of $\mathbf{ZIDP}(\mu_i, \phi, p)$ with probability function

$$P(Y_i = y_i|p, \mu_i, \phi) = \begin{cases} p + (1-p)f_{DP}(0|\mu_i, \phi) & y_i = 0, \\ (1-p)f_{DP}(y_i|\mu_i, \phi) & y_i = 1, 2, \ldots. \end{cases} \tag{2.5}$$

with $\mu_i = \theta n_i$. Combining (2.1) with (2.4),

$$\mathbb{E}(Y_i) = (1-p)\mu_i \quad \text{e} \quad \mathbb{V}(Y_i) = \mathbb{E}(Y_i)\left(p\mu_i + \frac{1}{\phi}\right). \tag{2.6}$$

Clearly, ϕ measures the overdispersion in the Zero-Inflated Poisson model. When $p = 0$ and $\phi = 1$, the model $\mathbf{ZIDP}(\mu_i, 1, 0)$ is the standard Poisson $\mathcal{P}(\mu_i)$; when $p \neq 0$ and $\phi = 1$, the model $\mathbf{ZIDP}(\mu_i, 1, p)$ is the **ZIP** model.

2.2 Spatial scan statistics

Given a study region represented by a geographic map divided into areas, each with an assigned population at risk and number of disease cases, the Spatial Scan Statistic (Kulldorff (1997)) is a test devised to identify a cluster (subset of the study area) with elevated incidence of cases compared to the rest of the map. This is a likelihood ratio test and makes use of a scanning procedure (the spatial scan) to search for the most likely cluster among the many candidate clusters in space or space-time. The simplest spatial version imposes circularly or elliptically shaped moving windows over the study region looking for compact clusters (Duczmal, Kulldorff, and Huang (2006), Duczmal et al. (2011)).

Specifically, let $\mathcal{S}$ be a study region projected in the Cartesian plane with L areas $\{s_1, \ldots, s_L\}$, population at risk $n(s_i) = n_i$. It is usual to determine, in the interior of each area s_i, a point (or *centroid*) a_i in the plane. Under the assumption of completely random distribution of cases (the null hypothesis H_0), let $Y_i \sim \mathcal{P}(\theta n_i)$ for every $s_i \in \mathcal{S}$. Let Z be a candidate cluster. Under the alternative hypothesis H_1, let $Y_i \sim \mathcal{P}(\theta_1 n_i)$ for every $s_i \in Z$ and $Y_i \sim \mathcal{P}(\theta_2 n_i)$ for every $s_i \notin Z$ with $\theta_1 > \theta_2$. The likelihood function for Z is given by

$$\mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y}) = \Big(\prod_{i=1}^{L} \frac{n_i^{y_i}}{y_i!}\Big)\theta_1^{y_z} e^{-\theta_1 n_z} \theta_2^{(y_+ - y_z)} e^{-\theta_2(n_+ - n_z)}, \tag{2.7}$$

where

$$y_+ = \sum_{i=1}^{L} y_i, \quad y_z = \sum_{s_i \in Z} y_i, \quad n_+ = \sum_{i=1}^{L} n_i \quad \text{e} \quad n_z = \sum_{s_i \in Z} n_i.$$

The likelihood ratio function for $H_0 : \theta_1 = \theta_2 = \theta$ versus $H_1 : \theta_1 > \theta_2$ is (Kulldorff (1997)):

$$\Lambda_Z = \frac{\max\limits_{\theta_1 > \theta_2} \mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y})}{\max\limits_{\theta_1 = \theta_2} \mathcal{L}_Z(\theta_1, \theta_2; \mathbf{y})} = \left(\frac{y_z/n_z}{y_+/n_+}\right)^{y_z} \left(\frac{(y_+ - y_z)/(n_+ - n_z)}{y_+/n_+}\right)^{(y_+ - y_z)},$$

if $y_z/n_z > (y_+ - y_z)/(n_+ - n_z)$ and $\Lambda_Z = 1$ otherwise. With $\mathcal{Z}$ the collection of all cluster candidates Z, the Spatial Scan Statistics is defined as

$$\Lambda = \max_{Z \in \mathcal{Z}} \Lambda_Z, \tag{2.8}$$

and the *most likely cluster* is $\hat{Z} = \arg\big(\max\limits_{Z \in \mathcal{Z}} \Lambda_Z\big)$. A Monte Carlo procedure is usually employed to obtain the test p-value. The Circular Scan is the most popular variant of the Spatial Scan Statistic (Kulldorff (1999)): given the area $s_{i_1} = s_i$ with centroid $a_{i_1} = a_i$, consider the L areas $(s_{i_1}, \ldots, s_{i_L})$ with the respective centroids $(a_{i_1}, \ldots, a_{i_L})$ sorted by their increasing order of distance from the centroid a_i. The candidate clusters $z_{i_m} = \{s_{i_1}, \ldots, s_{i_m}\}$, $i = 1, \ldots, L$, $m = 1, \ldots, S$ (not all distinct) form the collection of circular clusters of maximum size S, $S = 1, \ldots, L$.

3. Spatial Scan Statistics with Overdispersion and Inflated Zeros

3.1. Spatial scan statistics for ZIDP models

In order to accommodate simultaneously an excess of zeroes and overdispersion, suppose that the data $\mathbf{Y} = (Y(s_1), \ldots, Y(s_L))'$ are modeled by the $\mathbf{ZIDP}(\mu_i, \phi, p)$ model, with distribution given by (2.5). Following Kulldorff's (1997) cluster model, assume that $\mu_i = \theta_1 n_i$ when $s_i \in Z$, and $\mu_i = \theta_2 n_i$ when $s_i \notin Z$. Consider testing $H_0 : \theta_1 = \theta_2 = \theta$ against $H_1 : \theta_1 > \theta_2$. For a given Z, under H_1, the likelihood function is

$$\begin{aligned}
&\mathcal{L}_Z(p, \theta_1, \theta_2, \phi; \mathbf{y}) \\
&= \prod_{s_i \in Z} \left(p + (1-p) f_{DP}(0|\theta_1 n_i, \phi)\right)^{1 - I(y_i > 0)} \left((1-p) f_{DP}(y_i|\theta_1 n_i, \phi)\right)^{I(y_i > 0)} \\
&\quad \times \prod_{s_i \notin Z} \left(p + (1-p) f_{DP}(0|\theta_2 n_i, \phi)\right)^{1 - I(y_i > 0)} \left((1-p) f_{DP}(y_i|\theta_2 n_i, \phi)\right)^{I(y_i > 0)},
\end{aligned}$$

where $I(y_i > 0)$ is the indicator function of positive value occurrence. Under H_0 the likelihood function is

$$\mathcal{L}_0(p,\theta,\phi;\mathbf{y}) = \prod_{i=1}^{L}\left(p+(1-p)f_{DP}(0|\theta n_i,\phi)\right)^{1-I(y_i>0)}\left((1-p)f_{DP}(y_i|\theta n_i,\phi)\right)^{I(y_i>0)}.$$

Let $(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1)$ and $(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0)$ be respectively the maximum likelihood estimators for the parameters of the model under H_1 and H_0. Then the likelihood ratio statistic and the Spatial Scan Statistics for the ZIDP model are, respectively,

$$\hat{\Lambda}_Z = \frac{\mathcal{L}_Z(\hat{p}_1,\hat{\theta}_1,\hat{\theta}_2,\hat{\phi}_1;\mathbf{y})}{\mathcal{L}_0(\hat{p}_0,\hat{\theta}_0,\hat{\phi}_0;\mathbf{y})} \quad \text{and} \quad \hat{\Lambda} = \max_{Z\in\mathcal{Z}}\hat{\Lambda}_Z, \tag{3.1}$$

with estimated cluster $\hat{Z} = \arg\big(\max_{Z\in\mathcal{Z}}\hat{\Lambda}_Z\big)$. By inspecting $\mathcal{L}_Z(.;\mathbf{y})$ and $\mathcal{L}_0(.;\mathbf{y})$ it may be noted that there is no independence between the parameter p and the remaining parameters. This fact complicates the maximization of the likelihood function, especially when there are covariates involved. Thus, the inclusion of a latent vector of variables is necessary to factorize the likelihood to facilitate the maximization process, making use of the EM (Expectation-Maximization) algorithm. Let $\mathbf{U} = (U_1,\ldots,U_L)$, where $U_i = 1$ when Y_i occurs due to a zero state, and $U_i = 0$ when Y_i occurs due to a **DP** model. Assume that $U_i \sim$ Bernoulli(p). Then the augmented likelihood is

$$\mathcal{L}_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) = \prod_{s_i\in Z} p^{u_i}\left[(1-p)f_{DP}(y_i|\theta_1 n_i,\phi)\right]^{1-u_i} \times \prod_{s_i\notin Z} p^{u_i}\left[(1-p)f_{DP}(y_i|\theta_2 n_i,\phi)\right]^{1-u_i}.$$

Marginally, $Y_i \sim \mathbf{ZIDP}(\mu_i,\phi,p)$. The logarithm of the likelihood ratio for the ZIDP model under H_1 is

$$\begin{aligned} l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}\left(u_i\log p+(1-u_i)\log(1-p)\right)+\sum_{s_i\in Z}(1-u_i)\log f_{DP}(y_i|\theta_1 n_i,\phi) \\ &\quad+\sum_{s_i\notin Z}(1-u_i)\log f_{DP}(y_i|\theta_2 n_i,\phi) \\ &= l_Z^a(p;\mathbf{u})+l_Z^a(\theta_1,\phi;\mathbf{y},\mathbf{u})+l_Z^a(\theta_2,\phi;\mathbf{y},\mathbf{u}), \end{aligned} \tag{3.2}$$

and under H_0 is

$$\begin{aligned} l_0^a(p,\theta,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}\left(u_i log p+(1-u_i)\log(1-p)\right)+\sum_{i=1}^{L}(1-u_i)\log f_{DP}(y_i|\theta n_i,\phi) \\ &= l_0^a(p;\mathbf{u})+l_0^a(\theta,\phi;\mathbf{y},\mathbf{u}). \end{aligned} \tag{3.3}$$

Here the likelihood is easily maximized and the estimators $(\hat{p}_1, \hat{\theta}_1, \hat{\theta}_2, \hat{\phi}_1)$ and $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$ may be independently obtained. The estimator for ϕ in H_1 is obtained by maximizing $l_Z^a(\phi; \mathbf{y}, \mathbf{u}) = l_Z^a(\hat{\theta_1}, \phi; \mathbf{y}, \mathbf{u}) + l_Z^a(\hat{\theta_2}, \phi; \mathbf{y}, \mathbf{u})$, and for H_0 it is obtained by maximizing $l_0^a(\hat{\theta_0}, \phi; \mathbf{y}, \mathbf{u})$. To maximize (3.2) and (3.3) the EM algorithm is used. In this case the logarithm of the likelihood function is maximized iteratively in two steps until convergence. The maximization of $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u})$ is obtained as follows.

- Step E: Initialize the iterative process with $\boldsymbol{\gamma}^{(0)} = (p_1^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}, \phi_1^{(0)})$. At the $(k+1)$th iteration the estimate of $u_i^{(k)}$ is the conditional mean over $\mathbf{y}$ and the current estimates $\boldsymbol{\gamma}^{(k)} = (p_1^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}, \phi_1^{(k)})$. Thus compute $\mathbb{E}\{l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u})|\mathbf{y}, \boldsymbol{\gamma}^{(k)}\}$ with respect to the conditional distribution of $\mathbf{u}$. As $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u})$ is linear in $\mathbf{u}$, this is $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$, where $\mathbf{u}^{(k)} = \mathbb{E}_{H_1}(\mathbf{u}|\mathbf{y}, \boldsymbol{\gamma}^{(k)})$, with the ith element

$$u_i^{(k)} = P_{H_1}(u_i = 1|y_i, \boldsymbol{\gamma}^{(k)})$$
$$= \frac{P_{H_1}(Y_i = y_i|u_i{=}1, \boldsymbol{\gamma}^{(k)}) P_{H_1}(u_i{=}1|p_1^{(k)})}{P_{H_1}(Y_i{=}y_i|u_i{=}1, \boldsymbol{\gamma}^{(k)}) P_{H_1}(u_i{=}1|p_1^{(k)}) + P_{H_1}(Y_i{=}y_i|u_i{=}0, \boldsymbol{\gamma}^{(k)}) P_{H_1}(u_i{=}0|p_1^{(k)})}$$

and

$$u_i^k = \begin{cases} \left(1 + \exp\{-\log(\frac{p_1^{(k)}}{1-p_1^{(k)}}) - \phi_1^{(k)}\theta_1^{(k)} n_i + \frac{1}{2}\log\phi_1^{(k)}\}\right)^{-1} & \text{if} \quad y_i = 0, s_i \in Z, \\ \left(1 + \exp\{-\log(\frac{p_1^{(k)}}{1-p_1^{(k)}}) - \phi_1^{(k)}\theta_2^{(k)} n_i + \frac{1}{2}\log\phi_1^{(k)}\}\right)^{-1} & \text{if} \quad y_i = 0, s_i \notin Z, \\ 0 & \text{if} \quad y_i > 0. \end{cases}$$

- Step M: Maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$.
  1. Step M for p : In the $(k+1)$th iteration maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with respect to p, equivalently maximize $l_Z^a(p; \mathbf{u})$ as (3.3) considering $\mathbf{u} = \mathbf{u}^{(k)}$. Analytically, $p_1^{(k+1)} = \sum_{i=1}^{L} u_i^{(k)}/L$ and $\hat{p}_1$ is the value $p_1^{(k+1)}$ satisfying $|p_1^{(k+1)} - p_1^{(k)}| < \epsilon$.
  2. Step M for θ_1: In the $(k+1)$th iteration maximize $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with respect to θ_1, equivalently to maximize $l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u})$ as (3.3) considering $\mathbf{u} = \mathbf{u}^{(k)}$. Analytically, ${\theta_1}^{(k+1)} = \sum_{s_i \in Z}(1 - u_i^{(k)})y_i / \sum_{s_i \in Z}(1 - u_i^{(k)})n_i$ and $\hat{\theta}_1$ is the quantity ${\theta_1}^{(k+1)}$ satisfying $|{\theta_1}^{(k+1)} - {\theta_1}^{(k)}| < \epsilon$.
  3. Step M for θ_2: Similar to Step M for θ_1 subtitute $l_Z^a(\theta_1, \phi; \mathbf{y}, \mathbf{u})$ by $l_Z^a(\theta_2, \phi; \mathbf{y}, \mathbf{u})$. Then ${\theta_2}^{(k+1)} = \sum_{s_i \notin Z}(1 - u_i^{(k)})y_i / \sum_{s_i \notin Z}(1 - u_i^{(k)})n_i$ and $\hat{\theta}_2$ is the quantity ${\theta_2}^{(k+1)}$ satisfying $|{\theta_2}^{(k+1)} - {\theta_2}^{(k)}| < \epsilon$.
  4. Step M for ϕ: In the $(k+1)$th iteration maximize $l_Z^a(\theta_1^{(k+1)}, \phi; \mathbf{y}, \mathbf{u}) +$

$l_Z^a(\theta_2^{(k+1)}, \phi; \mathbf{y}, \mathbf{u})$ with respect to ϕ considering $\mathbf{u} = \mathbf{u}^{(k)}$. Analytically,

$$\phi_1^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\left\{\sum_{s_i\in Z}(1-u_i^{(k)})y_i\log(\theta_i/\theta_1^{(k+1)}) + \sum_{s_i\notin Z}(1-u_i^{(k)})y_i\log(\theta_i/\theta_2^{(k+1)})\right\}},$$

where $\theta_i = y_i/n_i$ and $\hat{\phi}_1 = min\{1, \phi_1^{(k+1)}\}$ with $\phi_1^{(k+1)}$ satisfying $|\phi_1^{(k+1)} - \phi_1{}^{(k)}| < \epsilon$.

The maximization of $l_0^a(p, \theta, \phi; \mathbf{y}, \mathbf{u})$ is processed similarly to the maximization of $l_Z^a(p, \theta_1, \theta_2, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with the following modification. At step E, under H_0, use

$$u_i^k = \begin{cases} \left(1+\exp\{-\log(\frac{p_0^{(k)}}{1-p_0^{(k)}}) - \phi_0^{(k)}\theta_0^{(k)}n_i + \frac{1}{2}\log\phi_0^{(k)}\}\right)^{-1} & \text{if } y_i = 0, i = 1, \ldots, L, \\ 0 & \text{if } y_i > 0. \end{cases}$$

Now maximize $l_0^a(p, \theta, \phi; \mathbf{y}, \mathbf{u}^{(k)})$ with respect to the parameters, obtaining at the $(k+1)$th iteration,

$$p_0^{(k+1)} = \frac{\sum_{i=1}^{L} u_i^{(k)}}{L}, \quad \theta_0^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})y_i}{\sum_{i=1}^{L}(1-u_i^{(k)})n_i},$$

$$\phi_0^{(k+1)} = \frac{\sum_{i=1}^{L}(1-u_i^{(k)})}{2\left\{\sum_{i=1}^{L}(1-u_i^{(k)})y_i\log(\theta_i/\theta_0^{k+1})\right\}}.$$

After the convergence of the algorithm, denote the estimates via the EM algorithm by $(\hat{p}_1, \hat{\theta}_1, \hat{\theta}_2, \hat{\phi}_1)$, $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$ and compute $(\hat{\Lambda}_Z, \hat{\Lambda})$ given in (3.1). Now, using $\hat{\Lambda}$, the spatial cluster may be identified under an excess of zeroes and overdispersion.

3.2. Fast Double Bootstrap-EM for the p-value computation

As the distribution of $\hat{\Lambda}$ is not available analytically, the statistic p-value is computed using the Fast Double Bootstrap Test (Davidson and MacKinnon (2001)), jointly with the application of the EM algorithm for each new dataset generated under the null hypothesis. The Fast Double Bootstrap procedure is necessary in this situation because the parameters of the $\hat{\Lambda}$ distribution are unknown under the null hypothesis.

Under H_0, Y_i is a Bernoulli(p)-$\mathbf{DP}(\theta n_i, \phi)$ mixture. By Efron (1986),

$$X_i \sim \mathcal{P}(\theta n_i \times \phi) \Longrightarrow \left(\frac{X_i}{\phi}\right) \dot{\sim} \mathbf{DP}(\theta n_i, \phi).$$

Given (p_0, θ_0, ϕ_0), Y_i is generated from the **ZIDP**$(n_i\theta_0, \phi_0, p_0)$ model as follows.

- Algorithm **ZIDP**$(n_i\theta_0, \phi_0, p_0)$
  1. Generate $x_i \sim \mathcal{P}(\theta_0 n_i \times \phi_0)$ and $v_i \sim Uniform(0,1)$.
  2. If $v_i \leq p_0$ let $y_i = 0$. Else $y_i = x_i/\phi_0$.

The p-value is computed as follows.

- Fast Double Bootstrap-EM algorithm for $\hat{\Lambda}$.
  1. Based on data $\mathbf{y} = (y_1, \ldots, y_L)$, use the EM algorithm and compute $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$. Derive the observed value $\hat{\Lambda}$ and denote it by $\hat{\lambda}$.
  2. Generate $\mathbf{y}_b^* = (y_{1,b}^*, \ldots, y_{L,b}^*)$ using the EM-algorithm **ZIDP** with (p_0, θ_0, ϕ_0) substituted by $(\hat{p}_0, \hat{\theta}_0, \hat{\phi}_0)$.
  3. Based on generated data $\mathbf{y}_b^*$, use the EM algorithm and compute the pseudo-estimators $(\hat{p}_{0,b}^*, \hat{\theta}_{0,b}^*, \hat{\phi}_{0,b}^*)$ for (p_0, θ_0, ϕ_0). Derive the pseudo-value of $\hat{\Lambda}_b^*$ and denote it by $\hat{\lambda}_b^*$.
  4. Repeat Steps 2 and 3 for $b = 1, \ldots, B$, compute the usual p-value for $\hat{\Lambda}$ as
$$p_{value}^* \doteq p_{value}^*(\hat{\Lambda}) = \sum_{b=1}^{B+1} \frac{I(\hat{\lambda} \geq \hat{\lambda}_b^*)}{(B+1)}, \quad \text{with} \quad \hat{\lambda}_{B+1}^* = \hat{\lambda}.$$
  5. Generate $\mathbf{y}_b^{**} = (y_{1,b}^{**}, \ldots, y_{L,b}^{**})$ using the **ZIDP** algorithm with (p_0, θ_0, ϕ_0) substituted by $(\hat{p}_{0,b}^*, \hat{\theta}_{0,b}^*, \hat{\phi}_{0,b}^*)$. Using Steps 3 and 4, derive $\hat{\Lambda}_b^{**}$ and denote it by $q_{1-p_{value}^*}^{**}$, the $1 - p_{value}^*$-quantile of the empirical distribution of $\hat{\Lambda}_b^{**}$. This quantile is the solution of the equation
$$\frac{1}{B}\sum_{b=1}^{B} I(\hat{\Lambda}_b^{**} > q_{1-p_{value}^*}^{**}) = p_{value}^*.$$
  6. Compute the fast double bootstrap $p - value$ for $\hat{\Lambda}$ by
$$p_{value}^{**} \doteq p_{value}^{**}(\hat{\Lambda}) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\Lambda}_b^* > q_{1-p_{value}^*}^{**}).$$

The convergence of the ZIDP EM algorithm is studied through simulations, and a proof of the convergence is also given (see the Supplementary Materials Section). A program implementing the ZIOP algorithm was written in C language, and can be requested from the corresponding author.

4. A Simulation Study

The zero inflation and overdispersion effects on type I error probability and power of detection for the four Poisson based Spatial Scan Statistic models are

evaluated in this section, namely the Poisson (**ScanP**), Zero Inflated Poisson (**ScanZIP**), Overdispersed Poisson (**ScanOP**), and Zero Inflated Overdispersed Poisson (**ScanZIOP**). The **ScanZIOP** is represented by the **ZIDP** model and the **ScanOP** is obtained from the **ZIDP** model by using $p = 0$.

The study region is the Amazonas state in Brazil with $L = 62$ municipalities (Figure 1). The populations at risk consist of children under 15 years living in 2010. Alternative hypotheses models with artificial clusters were simulated to evaluate the power of detection, and null hypothesis model maps were simulated to evaluate the type I error. For each model, 1,000 Monte Carlo replications were generated. An artificial circularly shaped (Kulldorff (1999)) spatial cluster $Z = \{\text{Anori, Coari, Codajás, Tefé, Tapauá}\}$ is located in the central part of the study region (Figure 1(D)).

Under null hypothesis, $\mu_i = n_i\lambda_0$, where $\lambda_0 = 0.001$ is a global rate reference for the disease; under the alternative model, $\mu_i = n_i\lambda_0(1+\theta)$ for every $s_i \in Z$ and $\mu_i = n_i\lambda_0$ otherwise, where $\theta > 0$ indicates the cluster intensity. Note that $\theta = 0$ under the null model.

The simulation procedure was given by

- (1) Generate 1,000 Monte Carlo replications under H_0, with data generated by $\mathcal{P}(n_i\lambda_0)$ and estimate the upper 5% quantile for each one of the four empirical distributions of the methods **ScanP**, **ScanZIP**, **ScanOP**, and **ScanZIOP**.
- (2) Generate 1,000 Monte Carlo replications under the null ($\theta = 0$) and alternative ($\theta = \{0.5, 1.0, 2.0\}$) models with overdispersion $1/\phi = \{1, 1.5, 2.0, 3.0\}$, zero inflation $p = \{0, 0.1, 0.2, 0.3\}$; estimate empirically the type I error and power of detection using the critical value given by the previously obtained upper 5% quantile.
  Let the detected most likely cluster $\hat{Z}^{(q)}$ obtained in the qth simulation be the estimator of the artificial cluster Z ($\#\{\text{A}\}$ indicates the cardinality of the set A).
- The precision for the cluster detection was evaluated by the following measures:
  - Sensitivity-(**SS**)= the average ratio of the number of locations correctly detected by the number of locations belonging to the artificial cluster:

$$\mathbf{SS} = \frac{1}{1,000}\sum_{q=1}^{1,000}\left(\frac{\#\{\hat{Z}^{(q)} \cap Z\}}{\#Z}\right),$$

  - Positive Predicted Value-(**PPV**)= the average ratio of the number of locations correctly detected by the number of locations belonging to the detected cluster:

$$\mathbf{PPV} = \frac{1}{1,000}\sum_{q=1}^{1,000}\left(\frac{\#\{\hat{Z}^{(q)} \cap Z\}}{\#\{\hat{Z}^{(q)}\}}\right),$$

The measures **SS** and **PPV** evaluate the performance of the methods according to their ability to locate the cluster, when it exists.

The simulation results are summarized on Tables 2, 3 and 4 in the Supplementary Materials Section.

In the absence of zero inflation ($p = 0$) and overdispersion ($\phi = 1$), type I error probability is adequate for all four methods (see Table 2 in the Supplementary Materials Section). With zero inflation ($p > 0$) but no overdispersion ($\phi = 1$), the type I error probability for the **ScanZIP** and **ScanZIOP** stay below 5%, whereas the corresponding values for **ScanP** and **ScanOP** are elevated, showing their inefficiency in this situation. In the absence of zero inflation ($p = 0$) and in the presence of overdispersion ($1/\phi > 1$), the **ScanOP** and **ScanZIOP** attain the lowest type I error probability; those values are somewhat larger than 5% due to the fact that their null hypothesis critical values 5% quantiles were obtained under the assumption that the true model is Poisson. However, these probabilities decrease when the overdispersion increases. The **ScanP** and **ScanZIP** attain large type I error probability values, making both of them inadequate for this scenario. When zero inflation and overdispersion occur simultaneously ($p > 0$ and $1/\phi > 1$), the three first methods, **ScanP**, **ScanOP** and **ScanZIP**, exhibit large values of type I error probability; only the **ScanZIOP** method presents an adequate performance.

According to Table 3 of the Supplementary Materials Section, the power of detection is greater in the presence of overdispersion and zero inflation for the **ScanP** and **ScanZIP**, as expected, as these methods attained high values of probability of type I error. The only reliable power estimate in this scenario is the one for the **ScanZIOP**. In the simulations, it was also observed that **ScanZIOP**'s power increases rapidly with small increases in cluster cases intensity ($\theta > 0$). When the cluster intensity and zero inflation remain fixed, power decreases. The same effect is observed when the cluster intensity and overdispersion remain fixed. This is evidence that the **ScanZIOP** is better suited to detect spatial clusters for small values of zero inflation and overdispersion.

From the results in Table 4 of the Supplementary Materials Section, **SS** and **PPV** are low for the **ScanOP** under zero inflation and overdispersion but increase as the cluster intensity increases. The **ScanP** attains low **PPV** values and sensitivity decreases when the cluster intensity increases, an indication that the **ScanP** tends to detect larger clusters than the true cluster. The methods **ScanZIP** and **ScanZIOP** behave similarly in terms of precision: the **SS** and

PPV measures increase when the cluster intensity increases. When cluster intensity is small ($\theta = 0.5$) the **ScanZIP** has more precision than the **ScanZIOP**. However, as the cluster intensity increases, the differences are negligible.

The artificial cluster $Z_1 = \{$Anori, Coari, Codajás, Tefé, Tapauá$\}$ of Figure 1(D) is located in the central part of the map, including about 8% of the total population. On the other hand, the small population artificial cluster $Z_2 = \{$Fonte Boa, Japurá Jutaí Maraã Tonantins$\}$ to the west contains only 3.5% of the total population. The power of detection of **ScanZIOP** was compared for those two population clusters. The results for those alternative model sets, with 1,000 simulations each, are presented in Tables 5 and 6 of the Supplementary Materials Section. The power is almost the same, except for $\theta = 0.5$, when there is a slight reduction of power for Z_2, compared to the Z_1 cluster.

5. Application: Hanseniasis Clusters

This study uses data for new Hanseniasis cases in children under 15 years old in the Amazonas state, Brazil, from 2008 to 2010 for each of their 62 municipalities. The dataset was divided into two periods: 2008/2009 (207 new cases in two years, 0.0000831 cases per child per year) and 2010 (190 new cases, 0.0001525 cases per child per year), see Figure 1 (A and B). Hanseniasis is an endemic contageous disease related to extreme poverty. In the 2008/2009 period, 20 municipalities (32%) registered zero new cases, compared with 30 municipalities (48%) that registered zero new cases in 2010 alone. In the 2008/2009 period, the average of the 62 municipalities' rates of new cases for 10,000 persons was 2.944, with variance equal to 6.776 (in the two years period). In the 2010 period, the corresponding mean and variance values were respectively 2.706 and 7.421 in the one year period. Figure 2 A and B displays the rates for the 62 municipalities. As the variance is substantially greater than the mean for those two scenarios, the **ZIDP** model seems quite plausible.

In this application, the Circular Scan employs the collection of circular clusters with maximum size $S = 15$ (25% of the municipalities), for the four models of Section 3: **ScanP**, **ScanZIP**, **ScanOP** and **ScanZIOP**.

The results are shown in Table 1.

In the 2008/2009 period, **ScanZIOP** and **ScanOP** did not detect significant clusters (p-value=0.114 and 0.112, respectively). The estimated overdispersion by **ScanZIOP** was $1/\hat{\phi}_0 = 2.325$, the zero inflation was below 1% ($\hat{p}_0 = 0.009$), and the cases rate was 1.67 per 10,000 persons ($\hat{\theta}_0 = 0.000167$). However, **ScanZIP** and **ScanP** detected a significant cluster (both with p-value=0.001). The zero inflation estimated by **ScanZIP** was $\hat{p}_1 = 0.013$, with estimated rates inside and outside the cluster given by $\hat{\theta}_1 = 0.000427$, and $\hat{\theta}_2 = 0.000149$, respectively (the estimated relative risk was 2.866). Taking into account that **ScanZIP**

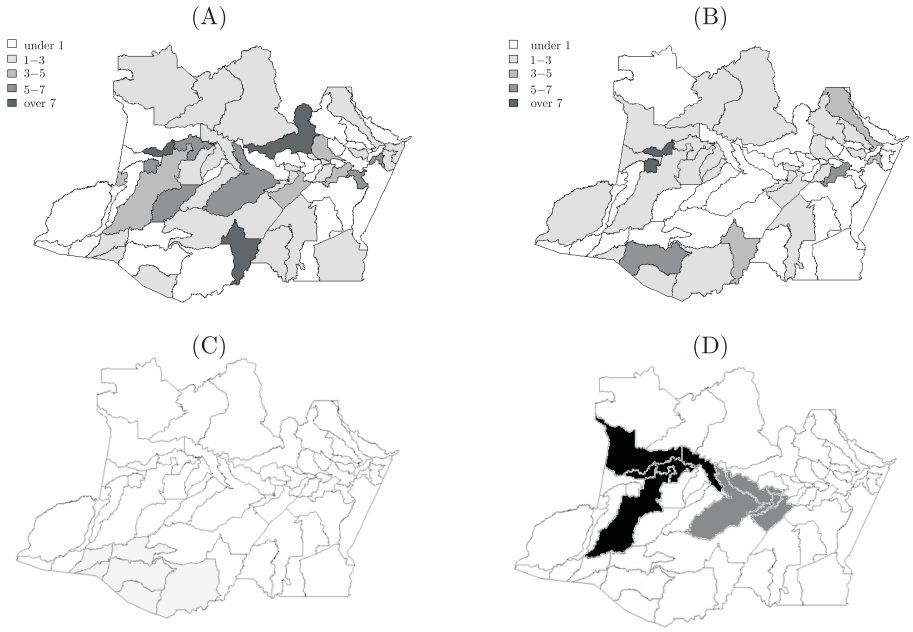


Figure 1. Spatial distribution of Hanseniasis cases: 2008/2009 (A) and 2010 (B). Detected Cluster in 2010 (C). Artificial cluster used in the simulations (section 4) (D).

Table 1. Spatial Clusters of new cases of Hanseniasis for 2008/09 and 2010.

Year	Scan	$\log \hat{\Lambda}$	p-value	$(\hat{p}_0, \hat{\phi}_0, 1000 \times \hat{\theta}_0)$	$(\hat{p}_1, \hat{\phi}_1, 1000 \times \hat{\theta}_1, 1000 \times \hat{\theta}_2)$
2008/09	**ScanP**	12.510	0.001	(0.000, 1.000, 0.166)	(0.000, 1.000, 0.427, 0.148)
	ScanZIP	11.074	0.001	(0.010, 1.000, 0.167)	(0.013, 1.000, 0.427, 0.149)
	ScanOP	5.886	0.122	(0.000, 0.428, 0.166)	(0.000, 0.518, 0.427, 0.148)
	ScanZIOP	5.882	0.114	(0.009, 0.430, 0.167)	(0.013, 0.521, 0.427, 0.149)
2010	**ScanP**	16.785	0.001	(0.000, 1.000, 0.152)	(0.000, 1.000, 0.518, 0.134)
	ScanZIP	15.955	0.001	(0.224, 1.000, 0.171)	(0.013, 1.000, 0.597, 0.154)
	ScanOP	7.696	0.034	(0.000, 0.406, 0.152)	(0.000, 0.520, 0.517, 0.134)
	ScanZIOP	8.849	0.006	(0.224, 0.442, 0.171)	(0.258, 0.689, 0.597, 0.154)

does not accommodate overdispersion in the positive counts, this cluster significance value is doubtful.

In the 2010 period, the four methods detected the same cluster (Figure 1 (C)), with 30 new cases when the expected number was $(1 - \hat{p}_1) n_{\hat{Z}} \hat{\theta}_1 = 25.67$. The zero inflation and overdispersion estimated by **ScanZIOP** was $\hat{p}_1 = 0.258$ and $1/\hat{\phi}_1 = 1.471$ respectively. The cluster is situated in a region well known for

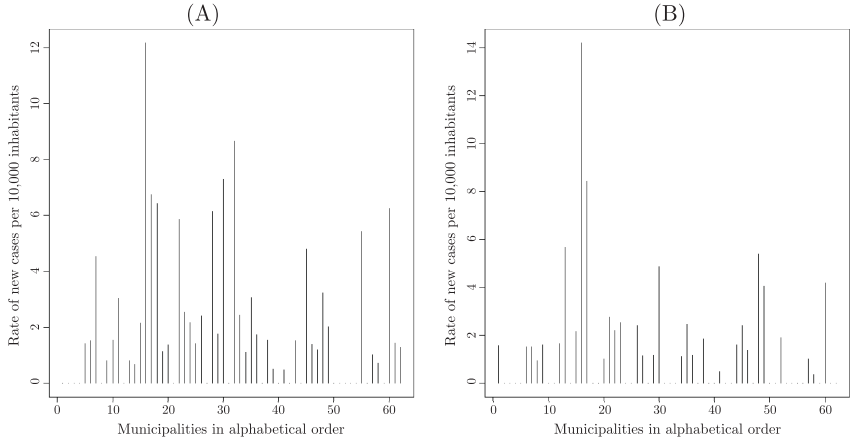


Figure 2. New cases of Hanseniasis in the Brazilian Amazon, 2008/2009 (A) and 2010 (B).

its high social vulnerability index.

6. Final Remarks

A modification of the Spatial Scan Statistics, the Zero Inflated Double Poisson Scan (**ZIPD**), is proposed to accommodate simultaneously an excess of zeroes and overdispersion. It might also be useful in disease surveillance, where the excessive variation for positive counts is frequent.

Sometimes, when the usual scan statistic is used under the null hypothesis of constant rate, a small p-value may result due to the high variability in the number of cases among a reduced number of areas or, alternatively, a small variability among many areas. This may cause in turn the existence of a false positive cluster; this anomaly could be avoided by changing the usual Poisson model by an overdispersed model. This kind of problem was evident in Section 4 (simulations) and Section 5 (applications). The simulations show that accounting for the presence of overdispersion and zero inflation in the **ZIDP** model reduces substantially the probability of type I error, compared to the Poisson, overdispersed Poisson, and zero inflated Poisson, shown here to be inadequate in those scenarios. That means that when a cluster is not detected by the **ZIDP**, and detected by the other methods, it should be carefully analyzed before being recognized as a legitimate cluster.

In the presence of overdispersion for positive count values, the detection of spatial clusters based on the zero-inflated model may be not the best option. In this situation the **ZIDP** Spatial Scan is a more flexible approach, but not the only one. The Binomial Negative (**NB**), Beta-Binomial (**BB**), and Generalized Poisson (**GP**) can also treat overdispersion and, similarly to the **ZIPD**,

it is possible to detect and evaluate spatial clusters based on the **ScanZINB**, **ScanZIBB** and **ScanZIGP** models. The significance of clusters found using those methods may be also assessed using the same strategy based on the Fast Double Bootstrap employed in this paper.

Spatial correlations could also be modeled with the proposed approach. These may be present due to the contagious nature of the disease, heterogeneous distribution of phenotypic traits, environmental causes, or to some latent variables that are related to the disease but not included in the data collection or in the model (Loh and Zhu (2007)). In fact, the objective of the cluster detection process is to see whether the counts from different locations are spatially correlated or not. The existence of a spatial cluster is an indication of the presence of spatial correlation, it signals the presence of a subregion with anomalous counts compared to the rest of the study region. Two approches can be used to tackle this problem, depending on how easily one can identify the spatial correlation factors.

Spatial correlation can be added to the model in order to include some known specific feature related to the population. As example, female population age is known to be strongly related to the occurrence of breast cancer, and a covariate may be added to the model in order to take into account this feature: the usual procedure is to stratify the population of each area by age and recompute the spatial counts, thus reducing the case counts for locations with older than average population. If eventually some breast cancer cluster is found in the modified study region, then it is not due to the age effect (supposing that the stratification was carefully done!). If the study region is not corrected for the age covariate, a cluster may be found that is simply consequence of the concentration of older people in some part of the study region. The ZIOP model allows the introduction of covariates in a straightforward manner, similarly to the other models compared in our work.

When the factors causing the spatial correlation cannot be easily identified, the algorithm of Section 2.3 of (Loh and Zhu (2007)) is a good option. In this case, the number of expected cases in the area i can be rewritten as

$$\mu_i = \exp(\log(n_i) + \theta_i I_{\{s_i \in Z\}} + v_i),$$

where $\log(n_i)$ is the populational adjustment, θ_i is the parameter measuring the intensity of cases in the cluster Z compared to the exterior of Z, and v_i is the random effect used to capture the spatial dependence. The ZIOP model could be adapted to use this modification without additional problems, similarly to the other models compared in our work.

Acknowledgements

The authors thank the Editors and two anonymous reviewers for their comments, which helped to improve the manuscript. The authors were funded with grants from the Brazilian agencies CAPES, UFAM, CNPq and FAPEMIG.

References

Cançado, A. L. F., da-Silva, C. Q. and da Silva, M. F. (2011). A zero-inflated Poisson-based spatial scan statistic. *Emerging Health Threats J.* 4.

Cançado, A. L. F., da-Silva, C. Q. and da Silva, M. F. (2014). A spatial scan statistic for zero-inflated Poisson process. *Envir. Ecol. Stat.* (doi: 10.1007/s10651-013-0272-1).

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statist. Medicine* **21**, 1461-1469.

Consul, P. C. and Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics* **15**, 791-799.

Davidson, R. and MacKinnon, J. G. (2001). Improving the reliability of bootstrap tests. Queen's University Institute for Economic Research Discussion Paper, No. 995, revised.

Deng, D. and Paul, S. R. (2005). Score test for zero-inflated and over-dispersion in generalized linear models. *Statist. Sinica* **15**, 257-276

Duczmal, L., Kulldorff, M. and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters. *J. Comput. Graph. Statist.* **15**, 428-442.

Duczmal, L. H., Moreira, G. J. P., Burgarelli, D., Takahashi, R. H. C., Magalhães, F. C. O. and Bodevan, E. C. (2011). Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast Brazilian town. *Int. J. Health Geogr.* **10**, 29.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81**, 709-721.

Hall, D. B. (2000). Zero inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030-1039.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional Poisson Model. CORE Discussion Paper, No. 2003-63. University of Louvain. Belgium.

Houssian, M. M. and Lawson, A. B. (2006). Cluster detection diagnostics for small area health data, with reference to evaluation of local likelihood models. *Statist. Medicine* **25**, 771-786.

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

Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26**, 1481-1496.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications. *Scan Statistics and Applications.* (Edited by Glaz and Balakrishnan), 303-322. Birkhauser, Boston.

Loh, J. M. and Zhu, Z. (2007). Accounting for spatial correlation in the scan statistic. *Ann. Appl. Statist.* **1**, 560-584.

Perumean-Chaneya, S. E., Morganb., C., McDowallc., D. and Aband., I. (2012). Zero-inflated and overdispersion: what's one to do? *J. Statist. Comput. Simulation*, 1-13.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statist. Sinica* **15**, 831-840.

Xiang, L., Lee, A.H., Yau, K. K. W. and McLachlan, G. J. (2007). A score test for overdispersion in zero-inflated Poisson mixed regression model. *Statist. Medicine* **26** , 1608-1622.

Xu, H. Y., Xiea, M., Goha, T. N. and Fub, X. (2012). A model for interger-valued time series with conditional overdispersion. *Comput. Statist. Data Anal.* **56**, 4229-4242.

Yang, Z., Harding, J. W. and Addyb, C. L. (2010) . Testing overdispersion in the zero inflated Poisson model. *J. Statist. Plann. Inference* **139**, 3340-3353.

Yau, K. K. W., Lee, A. H. and Carrivick, P. J. W. (2004). Modeling zero-inflated count series with application to occupational health. *Comput. Methods Programs Biomed.* **74**, 47-52.

Zhang, T. and Lin, G. (2009). Spatial scan statistics in loglinear models. *Comput. Statist. Data Anal.* **53**, 2851-2858.

Zhang, T., Zhang, Z. and Lin, G. (2012). Spatial scan statistics with overdispersion. *Statist. Medicine* **2**, 762-774.

Federal University of Amazonas, Manaus, Amazonas, Brizil.

E-mail: maxlima@ufam.edu.br

Federal University of Minas Gerais, Avenida Presidente Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brizil.

E-mail: duczmal@est.ufmg.br

Federal University of Amazonas, Manaus, Amazonas, Brizil.

E-mail: jcardoso@ufam.edu.br

Federal University of Minas Gerais, Avenida Presidente Antônio Carlos, 6627 - Pampulha, Belo Horizonte - MG, 31270-901, Brizil.

E-mail: leticia@dcc.ufmg.br

(Received August 2013; accepted March 2014)

Apêndice B

Resultados obtidos

$$f_{DP}(y_i|\mu_i,\phi) = (\phi^{1/2}e^{-\phi\mu_i})\left(\frac{e^{-y_i}y_i^{y_i}}{y_i!}\right)\left(\frac{e^{\mu_i}}{y_i}\right)^{\phi y_i}$$

$$\mu_i = \begin{cases} \theta_1 n_i, & \text{se } s_i \in Z \\ \theta_2 n_i, & \text{se } s_i \notin Z. \end{cases}$$

O logaritmo da razão de verossimilhança para o modelo ZIDP, sob H_1, é

$$\begin{aligned} l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) &= \underbrace{\sum_{i=1}^{L}[u_i\log p + (1-u_i)\log(1-p)]}_{(\mathbf{1})} \\ &+ \sum_{s_i\in Z}[(1-u_i)\log f_{DP}(y_i|\theta_1 n_i,\phi)] \\ &+ \sum_{s_i\notin Z}[(1-u_i)\log f_{DP}(y_i|\theta_2 n_i,\phi)] \end{aligned}$$

$$\begin{aligned} (\mathbf{1}) &= \sum_{i=1}^{L}[u_i\log p + (1-u_i)\log(1-p)] \\ &= \sum_{i=1}^{L}[u_i\log p + \log(1-p) - u_i\log(1-p)] = \sum_{i=1}^{L}\{u_i[\log p - \log(1-p)] + \log(1-p)\} \\ &= \sum_{i=1}^{L}\left[u_i\log\left(\frac{p}{1-p}\right) + \log(1-p)\right] \end{aligned}$$

$$
\begin{aligned}
l_Z^a(p,\theta_1,\theta_2,\phi;\mathbf{y},\mathbf{u}) &= \sum_{i=1}^{L}\left[u_i \log\left(\frac{p}{1-p}\right)+\log(1-p)\right] \\
&+ \sum_{s_i\in Z}\left\{(1-u_i)\log\left[\phi^{1/2}e^{-\phi\theta_1 n_i}\left(\frac{e^{-y_i}y_i^{y_i}}{y_i!}\right)\left(\frac{e^{\theta_1 n_i}}{y_i}\right)^{\phi y_i}\right]\right\} \\
&+ \sum_{s_i\notin Z}\left\{(1-u_i)\log\left[\phi^{1/2}e^{-\phi\theta_1 n_i}\left(\frac{e^{-y_i}y_i^{y_i}}{y_i!}\right)\left(\frac{e^{\theta_2 n_i}}{y_i}\right)^{\phi y_i}\right]\right\}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l_Z^a}{\partial p} &= \frac{\partial}{\partial p}\left\{\sum_{i=1}^{L}\left[u_i\left(\log p-\log(1-p)\right)+\log(1-p)\right]\right\}+0+0 \\
&= \sum_{i=1}^{L}\left[u_i\left(\frac{1}{p}\right)+u_i\left(\frac{1}{1-p}\right)-\frac{1}{1-p}\right] \\
&= \sum_{i=1}^{L}\left[\frac{u_i}{p}-(1-u_i)\left(\frac{1}{1-p}\right)\right] \\
&= \frac{\sum_{i=1}^{L}u_i}{p}-\frac{\sum_{i=1}^{L}(1-u_i)}{1-p}
\end{aligned}
$$

Fazendo $\frac{\partial l_Z^a}{\partial p}=0$, temos

$$
\frac{\sum_{i=1}^{L}u_i}{\hat{p}}=\frac{\sum_{i=1}^{L}(1-u_i)}{1-\hat{p}}
$$

$$
\sum_{i=1}^{L}u_i-\hat{p}\sum_{i=1}^{L}u_i=\hat{p}\sum_{i=1}^{L}(1-u_i)
$$

$$\sum_{i=1}^{L} u_i - \hat{p}\sum_{i=1}^{L} u_i = \hat{p}L - \hat{p}\sum_{i=1}^{L} u_i$$

$$\text{Logo,} \quad \hat{p} = \frac{\sum_{i=1}^{L} u_i}{L}$$

$$\begin{aligned}
\frac{\partial l_Z^a}{\partial \theta_1} &= 0 + \frac{\partial}{\partial \theta_1}\left\{\sum_{s_i \in Z}\left[(1-u_i)\left(\frac{1}{2}\log\phi - \phi\theta_1 n_i + \frac{y_i \log y_i - y_i}{\log(y_i!)} + \underbrace{\phi y_i\left(\frac{1+\log(\theta_1 n_i)}{\log y_i}\right)}_{(\mathbf{2})}\right)\right]\right\} + 0 \\
&= \sum_{s_i \in Z}\left\{(1-u_i)\left[-\phi n_i + \phi y_i\left(\frac{n_i}{\theta_1 n_i}\right)\right]\right\} \\
&= \sum_{s_i \in Z}\left\{(1-u_i)\left[-\phi n_i + \frac{\phi y_i}{\theta_1}\right]\right\}
\end{aligned}$$

Fazendo $\dfrac{\partial l_Z^a}{\partial \theta_1} = 0$, temos

$$-\phi\sum_{s_i \in Z}(1-u_i)n_i + \frac{\phi\sum_{s_i \in Z}(1-u_i)y_i}{\hat{\theta}_1} = 0$$

$$\text{Logo,} \quad \hat{\theta}_1 = \frac{\sum_{s_i \in Z}(1-u_i)y_i}{\sum_{s_i \in Z}(1-u_i)n_i}$$

$$\begin{aligned}
(\mathbf{2}) &= \phi y_i\left(\frac{1+\log(\theta_1 n_i)}{\log y_i}\right) \\
&= \phi y_i\left[\frac{1}{\log y_i} + \frac{\log(\theta_1 n_i)}{\log y_i}\right] \\
&= \frac{\phi y_i}{\log y_i} + \log(\theta_1 n_i) - \log y_i
\end{aligned}$$

$$\frac{\partial l_Z^a}{\partial \theta_2} = 0 + 0 + \frac{\partial}{\partial \theta_2}\left\{\sum_{s_i \notin Z}\left[(1-u_i)\left(\frac{1}{2}\log\phi - \phi\theta_2 n_i + \frac{y_i \log y_i - y_i}{\log(y_i!)} + \phi y_i\left(\frac{1+\log(\theta_2 n_i)}{\log y_i}\right)\right)\right]\right\}$$

$$= \sum_{s_i \notin Z}\left\{(1-u_i)\left[-\phi n_i + \frac{\phi y_i}{\theta_2}\right]\right\}$$

Fazendo $\frac{\partial l_Z^a}{\partial \theta_2} = 0$, temos

$$-\phi\sum_{s_i \notin Z}(1-u_i)n_i = \frac{\phi\sum_{s_i \notin Z}(1-u_i)y_i}{\hat{\theta}_2}$$

$$\text{Logo, } \hat{\theta}_2 = \frac{\sum_{s_i \notin Z}(1-u_i)y_i}{\sum_{s_i \notin Z}(1-u_i)n_i}$$

$$\frac{\partial l_Z^a}{\partial \phi} = 0 + \frac{\partial}{\partial \phi}\left\{\sum_{s_i \in Z}\left[(1-u_i)\log f_{DP}(y_i/\phi, \theta_1)\right]\right\} + \frac{\partial}{\partial \phi}\left\{\sum_{s_i \notin Z}\left[(1-u_i)\log f_{DP}(y_i/\phi, \theta_2)\right]\right\}$$

Fazendo $\frac{\partial l_Z^a}{\partial \phi} = 0$, temos

$$\underbrace{\sum_{s_i \in Z}\left\{(1-u_i)\frac{\partial}{\partial \phi}\left[\log f_{DP}(y_i/\phi, \hat{\theta}_1)\right]\right\}}_{(\mathbf{1})} + \underbrace{\sum_{s_i \notin Z}\left\{(1-u_i)\frac{\partial}{\partial \phi}\left[\log f_{DP}(y_i/\phi, \hat{\theta}_2)\right]\right\}}_{(\mathbf{2})} = 0$$

$$
\begin{aligned}
(\mathbf{1}) &= \sum_{s_i \in Z} \left\{ (1-u_i) \left[\frac{1}{2\phi} - \hat{\theta}_1 n_i + y_i + y_i \log \hat{\theta}_1 + y_i \log \left(\frac{n_i}{y_i} \right) \right] \right\} \\
&= \frac{\sum_{s_i \in Z} (1-u_i)}{2\phi} - \hat{\theta}_1 \sum_{s_i \in Z} \left[(1-u_i)\, n_i \right] + \sum_{s_i \in Z} \left[(1-u_i)\, y_i \right] + \sum_{s_i \in Z} \left[(1-u_i)\, y_i \log \left(\frac{\hat{\theta}_1}{\theta_i} \right) \right] \\
&= \frac{\sum_{s_i \in Z} (1-u_i)}{2\phi} - \frac{\sum_{s_i \in Z} \left[(1-u_i)\, y_i \right]}{\sum_{s_i \in Z} \left[(1-u_i)\, n_i \right]} \sum_{s_i \in Z} \left[(1-u_i)\, n_i \right] + \sum_{s_i \in Z} \left[(1-u_i)\, y_i \right] + \\
&\quad + \sum_{s_i \in Z} \left[(1-u_i)\, y_i \log \left(\frac{\hat{\theta}_1}{\theta_i} \right) \right] \\
&= \frac{\sum_{s_i \in Z} (1-u_i)}{2\phi} + \sum_{s_i \in Z} \left[(1-u_i)\, y_i \log \left(\frac{\hat{\theta}_1}{\theta_i} \right) \right]
\end{aligned}
$$

Analogamente,

$$
\begin{aligned}
(\mathbf{2}) &= \sum_{s_i \notin Z} \left\{ (1-u_i) \left[\frac{1}{2\phi} - \hat{\theta}_2 n_i + y_i + y_i \log \hat{\theta}_2 + y_i \log \left(\frac{n_i}{y_i} \right) \right] \right\} \\
&= \frac{\sum_{s_i \notin Z} (1-u_i)}{2\phi} + \sum_{s_i \notin Z} \left[(1-u_i)\, y_i \log \left(\frac{\hat{\theta}_2}{\theta_i} \right) \right]
\end{aligned}
$$

Assim,

$$
\frac{\partial l_Z^a}{\partial \phi} = \frac{\sum_{s_i \in L} (1-u_i)}{2\phi} - \left\{ \sum_{s_i \in Z} \left[(1-u_i)\, y_i \log \left(\frac{\theta_i}{\hat{\theta}_1} \right) \right] + \sum_{s_i \notin Z} \left[(1-u_i)\, y_i \log \left(\frac{\theta_i}{\hat{\theta}_2} \right) \right] \right\}
$$

Logo, $\hat{\phi} = \dfrac{\sum_{s_i \in L} (1-u_i)}{2\left\{ \sum_{s_i \in Z} \left[(1-u_i)\, y_i \log \left(\frac{\theta_i}{\hat{\theta}_1} \right) \right] + \sum_{s_i \notin Z} \left[(1-u_i)\, y_i \log \left(\frac{\theta_i}{\hat{\theta}_2} \right) \right] \right\}}$

Referências

Boyles, R. A. (1983). On the convergence of the em algorithm, *J. Roy. Statist. Soc. Ser. B,* **45**: 47–50.

Cançado, A. L. F., Silva, C. Q. & Silva, M. F. (2014). A spatial scan statistic for zero-inflated poisson process, *Environmental and Ecological Statistics,* **21**: 627–650.

Casella, G. & Berger, R. L. (2010). *Inferência estatística*, tradução 2a ed. norte americana. Páginas 147–151, 291–294, 329.

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of cout data: a study of growth and development, *Statistics in Medicine,* **21**: 1461–1469.

Consul, P. & Jain, G. (1973). A generalization of the poisson distribution, *Technometrics,* **15**(4): 791–799.

Davidson, R. & MacKinnon, J. G. (2001). Improving the reliability of bootstrap tests, *Queen's University Institute for Economic Research Discussion Paper,* **995**. Revised.

Dempster, A. P., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society, B,* **39**: 1–22.

Deng, D. & Paul, S. R. (2005). Score test for zero-inflated and over-dispersion in generalized linear models, *Statistica Sinica,* **15**: 257–276.

Diggle, P., Rowlingson, B. & Su, T. (2005). Point process methodology for on-line spatio-temporal disease surveillance, *Environmetrics,* **16**: 423–434.

Duczmal, L. H., Moreira, G. J. P., Burgarelli, D., Takahashi, R. H. C., Magalhães, F. C. O. & Bodevan, E. C. (2011). Voronoi distance based prospective space-time scans for point data sets: a dengue fever cluster analysis in a southeast brazilian town, *Int. J. Health Geogr.,* **10**: 29.

Duczmal, L., Kulldorff, M. & Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters, *J. Comput. Graph. Statist.,* **15**: 428–442.

Efron, B. (1986). Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association,* **81**: 709–721.

Famoye, F. & Singh, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data, *Journal of Data Science,* **4**: 117–130.

Hall, D. B. (2000). Zero inflated poisson and binomial regression with random effects: a case study, *Biometrics,* **56**: 1030–1039.

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional poisson model, *CORE Discussion Paper,* **2003-63**. University of Louvain. Belgium.

Houssian, M. M. & Lawson, A. B. (2006). Cluster detection diagnostics for small area health data. with reference to evaluation of local likelihood models, *Statist. Med.,* **25**: 771–786.

Instituto Brasileiro de Geografia e Estatística. (2010). Índice de desenvolvimento humano. IBGE. Disponível em: www.ibge.gov.br. Acessado em: 12 de janeiro de 2015.

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

Kulldorff, M. (1997). A spatial scan statistic, *Communs Statist. Theory Meth,* **26**: 1481–1496.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications, *Scan Statistics and Applications. Edited by: Glaz, Balakrishnan. Boston: Birkhauser,* pp. 303–322.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *J. R. Statist Soc. A,* **164**: 61–72.

Kulldorff, M., Athas, W., Feuer, E. J., Miller, B. A. & Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in los amos, *American Journal of Public Health,* **88**: 1377–1380.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R. & Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection, *PLoS Medicine,* **2**: e59.

Lima, M. S., Duczmal, L. H., Neto, J. C. & Pinto, L. P. (2015). Spatial scan statistics for models with overdispersion and inflated zeros, *Statistica Sinica,* **25**: 225–241.

Lora, M. I. & Singer, J. M. (2008). Beta-binomial/ poisson models for repeated bivariate counts, *Statistics in Medicine,* **27**: 3366–3381.

Meng, B. J. & Zhu, Z. (2007). Accounting for spatial correlation in the scan statitics, *The Ann Appl Stat,* **2**: 560–584.

Ministério da Saúde. Secretaria de Vigilância em Saúde, Departamento de Vigilância e Doenças Transmissíveis (2015). *Exercício de monitoramento da eliminação da hanseníase no Brasil LEM–2012*, 1a ed. Brasília: Ministério da Saúde, 72p.

Ministério da Saúde. Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica (2010). *Estratégia global aprimorada para redução adicional da carga da hanseníase: período do plano: 2011-2015*, 1a ed. Brasília: Ministério da Saúde, 34p.

Ministério da Saúde. Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica (2013). *Plano integrado de ações estratégicas de eliminação da hanseníase, filariose, esquistossomose e oncocercose como problemas de saúde pública, tracoma como causa de cegueira e controle das geohelmintíases: plano de ação: 2011-2015*, 1a ed. Brasília: Ministério da Saúde, 13p.

Perumean-Chaneya, S., Morganb, C., McDowallc, D. & Aband, I. (2012). Zero-inflated and overdispersion: what's one to do?, *J Stat Comput Simul,* pp. 1–13.

Rodeiro, C. & Lawson, A. (2006). Monitoring changes in spatio-temporal maps of disease, *Biometrical Journal,* **48**: 463–480.

Rogerson, P. A. (2001). Monitoring point patterns for the development of space-time clusters, *Journal of the Royal Statistical Society. Series A,* **164**: 87–96.

Soares, S., Strauch, J. C. M. & Ajara, C. (2006). Análise espaço-temporal dos índices de sustentabilidade na microrregião de coari - estado do amazonas, *Anais XV Encontro Nacional de Estudos Populacionais*, Caxambú, 18-22 Setembro 2006. ABEP.

Takahashi, K., Kulldorff, M., Tango, T. & Yih, K. (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring, *International Journal of Health Geographics,* **7**: 14.

Tango, T., Takahashi, K. & Kohriyamma, K. (2011). A space-time scan statistic for detecting emerging outbreaks, *Biometrics,* **67**: 106–115.

WHO (2013). Global leprosy situation, *Weekly epidemiological record,* **88**(35): 365–380. Disponível em: www.who.int/wer.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*, 4th Edition. Berlin: Springer-Verlag.

Wu, C. F. J. (1983). On the convergence of the em algorithm, *Ann. Statist.,* **11**: 95–113.

Xiang, L., Lee, A. H., Yau, K. K. W. & McLachlan, G. J. (2007). A score test for overdispersion in zero-inflated poisson mixed regression model, *Statistics in Medicine,* **26**: 1608–1622.

Xu, H. Y., Xiea, M., Goha, T. N. & Fub, X. (2012). A model for interger-valued time series with conditional overdispersion, *Comput Stat Data Anal.,* **56**(12): 4229–4242.

Yang, Z., Harding, J. & Addyb, C. (2010). Testing overdispersion in the zero inflated poisson model, *Journal of Stat Plann Inf,* **139**: 3340–3353.

Yau, K. K. W., Lee, A. H. & Carrivick, P. J. W. (2004). Modeling zero-inflated count series with application to occupational health, *QComput. Methods. Programs. Biomed,* **74**: 47–52.

Zhang, T. & Lin, G. (2009). Spatial scan statistics in loglinear models, *Comput Stat Data Anal,* **53**: 2851–2858.

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