**UNIVERSIDADE DE BRASÍLIA**

DEPARTAMENTO DE ESTATÍSTICA

# Extensão da estatística Scan para detecção de conglomerados espaço-temporais em dados com excesso de zeros.

**Thiago Costa Araújo**

**Brasília - DF**

**2012**

**THIAGO COSTA ARAÚJO**

# Extensão da estatística Scan para detecção de conglomerados espaço-temporais em dados com excesso de zeros.

Dissertação apresentada ao Departamento de Estatística do Instituto de Exatas da Universidade de Brasília, como parte dos requisitos necessários para o grau de Mestre em Estatística.

Orientador: *Prof.º André Luiz Fernandes Cançado*

**Brasília - DF**

**2012**

# Agradecimentos

Agradeço primeiramente a Deus, por tudo o que nos foi dado.

Aos meus pais, pela educação, carinho e suporte.

Ao meu orientador, Professor Dr. André Luiz Fernandes Cançado, por toda disponibilidade e paciência.

À minha namorada Izabela Siqueira Cavalcante, por todo carinho, apoio e compreensão.

Ao meu amigo Lucas Rocha Soares de Assis, pelo auxílio e sugestões para resolver algumas questões da dissertação.

Aos pesquisadores do Instituto de Pesquisa Econômica Aplicada: Aguinaldo Nogueira Maciente, Paulo Augusto Meyer M. Nascimento e Rafael Henrique Moraes Pereira, por disponibilizar mais tempo para que eu pudesse me dedicar exclusivamente à construção desta dissertação e por todos os trabalhos que viemos desenvolvendo nos últimos anos.

A todos os meus amigos e familiares, pelos conselhos, conversas e bons momentos.

Aos professores do Departamento de Estatística, pelo esforço e constante preocupação em fazer os cursos de graduação e pós-graduação em estatística cada vez melhores.

# Resumo

A detecção de *clusters* espaciais ou espaço-temporais tem papel importante para a decisão das instituições competentes. Nas aplicações em que os dados apresentam grande concentração de zeros visualizam-se as distorções que tal ocorrência pode gerar, caso se utilize o modelo Poisson. Para lidar com o excesso de zeros na detecção de *clusters* espaciais propôs-se a utilização do modelo ZIP em conjunto com a estatística scan de Kulldorff (1997)(Cançado et al., 2012). Neste trabalho propomos uma extensão da estatística Scan-ZIP para o caso espaço-temporal (estatística Scan-ZIPET). A estatística Scan-ZIPET é aplicada em simulações numéricas e o seu desempenho é comparado com a versão Poisson. Realiza-se a comparação entre as duas estatísticas para os dados de tuberculose no estado da Geórgia, nos Estados Unidos da América. Obtém-se resultados positivos para a nova estatística, mas condicionados ao conceito de zero estrutural que permeia a aplicação.

*Palavras-chaves: detecção de clusters espaciais, estatística scan espaço-temporal, modelo Poisson inflacionada de zeros, zero estrutural.*

# Abstract

Detection of space-time clusters plays an important role in helping health officials in the decision making process. The assumption of Poisson distribution for count data may lead to distortions in the cluster detection process when the number of zero counts is greater than the expected. To handle this situation, Cançado et al. (2012) proposed the combination of the ZIP distribution with Kulldorff's scan statistic (Kulldorff, 1997). In this work we propose an extension of the ZIP spatial scan statistic to the space-time context, the Scan ZIPET. The Scan ZIPET statistic is compared to Kulldorff's space-time scan statistic (Kulldorff et al., 1998) through numerical simulations. An application to death cases due to tuberculosis in the state of Georgia, USA, is presented. The proposed space-time scan statistic presents better results in the presence of structural zeros.

*Keywords: Spatial cluster detection, space-time scan statistic, zero-inflated Poisson model, structural zero.*

# Sumário

# 1 Introdução

Dentro do cenário atual da sociedade, a rápida aprendizagem dos mecanismos associados a determinada doença é insumo para que medidas de proteção da população possam ser tomadas para evitar uma possível epidemia.

As agências de vigilância de epidemias, responsáveis por detectar, monitorar e prevenir epidemias, investigam os fatores de risco associados às doenças. Porém, devido à escassez de recursos, existe a necessidade de priorizar algumas regiões em detrimento de outras. As técnicas de detecção de clusters espaciais são um importante recurso no auxílio da identificação das áreas que necessitam de maior atenção.

O termo *cluster*, que será frequentemente utilizado ao longo do texto, é proveniente do inglês e significa um grupo composto por observações que se assemelham segundo características pertinentes ao contexto da análise. A questão do que define similaridade entre observações é largamente discutida e algumas medidas mais comuns são citadas na seção 2.2.

A análise de *cluster* é um vasto ramo da estatística e fornece diversas técnicas que possuem como objetivo fornecer grupos de objetos semelhantes. Dentro do contexto de clusters espaciais, estamos interessados na detecção de uma zona, conjunto de uma ou mais regiões contíguas no mapa, que contenha alguma característica em comum. Na epidemiologia, por exemplo, podemos pensar em um conjunto de municípios que apresente uma incidência, de uma determinada doença, significativamente maior do que no restante do estado. O tópico de técnicas de detecção de clusters espacias será abordado em maiores detalhes na seção 2.3.

Dentro da seção 2.4 será apresentada a estatística Scan de Kulldorff, uma das ferramentas mais recentes na detecção, identificação e inferência sobre clusters espacias. O método de Kulldorff utiliza de uma varredura em toda a área, buscando identificar o *cluster* onde o logaritmo da razão de verossimilhança é mais alto, para então aplicar o método de Monte Carlo para a avaliação da significância.

O caso espaço-temporal da estatística Scan de Kulldorff é apresentado na seção 2.5, revelando as potencialidades da estatística Scan ao se acrescentar a dimensão temporal na detecção de clusters.

O capítulo 3 fornecerá uma introdução à distribuição de Poisson inflacionada de zeros (Zero Inflated Poisson - ZIP), modelo que procura adaptar a distribuição de Poisson para a presença de excesso de zeros. Aplicando a distribuição de Poisson inflacionada de zeros para a estatística Scan Espacial de Kulldorff e para a estatística Scan Espaço-Temporal, chegamos a versões destas estatísticas capazes de lidar com dados de contagem que possuem excesso de zeros. A descrição destas estatísticas encontra-se nas seções 3.3 e 3.5 respectivamente.

O excesso de zeros é lidado nas aplicações através do zero estrutural. As ideias por trás do conceito de zero estrutural e algumas aplicações são apresentadas na seção 3.1. A seção 3.6 busca estender a definição de zero estrutural para o caso espaço temporal, apresentando algumas possibilidades e suas consequências.

No capítulo 4 é feita uma comparação entre o modelos Scan-ET, proposto por Kulldorff et al. (1998), e o Scan-ZIPET, extensão do Scan-ZIP para o caso espaço-temporal. Essa comparação é feita em termos de poder, sensibilidade e valor preditivo em três cenários construídos artificialmente e avaliados através de extensas simulações.

O capítulo 5 apresenta a comparação entre o modelo Scan ZIPET, segundo duas variações de definição de zero estrutural, e o Scan-ET, para casos de morte por tuberculose no estado da Geórgia.

Ao final apresentamos as conclusões obtidas com o trabalho, bem como propostas de trabalhos futuros.

# 2 Detecção de Clusters Espaciais e Espaço- Temporais

Em estudos nas mais diversas áreas frequentemente tem-se o interesse em buscar conhecer melhor as informações disponíveis e, como estas informações estão conectadas entre si, estabelecer relações e verificar hipóteses. Diversos métodos existem nas mais variadas áreas de conhecimento nesse sentido. Dentro da estatística podemos citar a análise de agrupamentos, também conhecida como análise de clusters. Esta classe de métodos busca fornecer grupos "naturais" das observações (ou das variáveis) de tal forma a fornecer indícios sobre o tipo de relação entre as observações(variáveis) que estão sendo estudadas.

Na primeira parte do capítulo serão apresentados alguns conceitos referentes ao contexto epidemiológico de tal forma a fornecer uma base que será útil ao longo do presente trabalho.

Na segunda parte deste capítulo será apresentado o problema de *clustering* mais usual, que busca estabelecer os grupos para que possam fornecer indícios de relações entre as observações ou variáveis. Em seguida apresentaremos o problema de detecção de clusters espaciais e espaço-temporais, que é o foco do presente trabalho e tem por intuito a delimitação de uma área com uma característica espacial diferenciada para determinada variável.

## 2.1 Conceitos Preliminares

A seguir fornecemos de forma simples as ideias que permeiam os conceitos de incidência e prevalência. A incidência de uma doença está associada à quantidade de casos novos dentro de uma janela de tempo pequena, enquanto a prevalência revela o número de pessoas infectadas em um intervalo temporal, determinando o impacto da doença sobre determinada população. Prevalência e incidência diferem quando uma pessoa infectada tem elevado tempo de sobrevivência. À medida que o tempo de sobrevivência se reduz, a incidência e a prevalência se aproximam.

O termo 'risco' dentro do contexto epidemiológico está associado a probabilidade de uma pessoa vir a contrair determinada doença. Risco é uma variável determinada e modificada por variáveis inerentes ao indivíduo, como idade, predisposição genética, ocupação, gênero e também por variáveis do meio, como saneamento urbano, coleta de lixo, presença de determinados animais e etc.

Como estimativa para o risco médio de uma pessoa contrair uma doença, utilizaremos a proporção de incidência da doença, que é calculada como a divisão do número total de pessoas que contraíram a doença pelo número total de pessoas em risco, ambos os valores relativos a um intervalo de tempo. A definição do intervalo de tempo a ser utilizado pode gerar críticas, uma vez que o numerador pode aumentar enquanto o denominador pode variar pouco ou até mesmo ser considerado constante, sendo o segundo caso mais comum nos estudos.

Um dos objetivos principais em um estudo epidemiológico é determinar o risco de uma doença para determinada população em determinado intervalo de tempo. Desta maneira, a identificação dos fatores que influem no risco tem papel importante no processo do estudo.

Para introduzir o conceito de diferença demográfica, podemos pensar na questão de duas cidades que possuem a mesma população, só que uma possuindo 70% de idosos enquanto a outra só possui 15% de idosos em sua composição populacional. Desta forma, se o estudo for focado em uma doença que tem maior

incidência em idosos, nosso cluster será consequência da composição demográfica e não necessariamente de um risco mais elevado.

## 2.2 Detecção de Clusters

No intuito de obter grupos, a análise de agrupamentos fornece uma partição do conjunto de observações/variáveis, que define grupos que sejam homogêneos internamente e heterogêneos entre si. A maneira de definir a existência de homogeneidade ou não dentro de um grupo é dada pela escolha da medida de similitude que será utilizada. A maioria das técnicas não exige a fixação do número de grupos a serem encontrados, apesar de que, existindo o interesse de obter um determinado número de grupos, pode-se fixar esse valor para a maioria das técnicas.

A similaridade ou dissimilaridade existente entre as observações pode ser definida de várias formas, dependendo de que medida seja escolhida. Entre as medidas mais utilizadas temos: distância euclidiana, distância de Minkowsky, distância generalizada, coeficiente de concordância e coeficiente de Czekanowk. A questão de que medida utilizar é algo importante e deve ser definida de acordo com a situação em que cada problema está inserido.

A classificação das técnicas de agrupamento depende dos critérios utilizados. A que utilizaremos é a que divide em agrupamentos hierárquicos e não hierárquicos. Outra forma possível seria em métodos paramétricos e não paramétricos.

As técnicas hierárquicas iniciam através de uniões ou divisões sucessivas, começando de cada item ou de um só grande grupo, respectivamente, e a partir daí efetua-se a operação (união ou divisão) dos itens até chegar a situação inversa da de partida, um só grupo para o processo de união e um número de grupos igual ao número de itens para o processo de divisão. Algumas técnicas de visualização, como o dendrograma, revelam a quantidade de grupos por nível da medida de similaridade. Portanto, para um determinado nível de similaridade temos os itens

que integram os grupos e a quantidade de grupos. Dentre as técnicas hierárquicas podemos citar: método da média das distâncias, método do centroide, método da ligação completa, método da ligação simples e método de Ward.

As técnicas não hierárquicas são iniciadas através de uma partição inicial dos grupos ou um ponto inicial que definirá o cenário inicial. A ausência do cálculo da matriz de distâncias e a não necessidade de se guardar a base para que o processo seja executado permite aplicar esse grupo de técnicas em bancos maiores dos que os possíveis de serem utilizados com as técnicas hierárquicas. Um exemplo bem conhecido de técnica não hierárquica é o método das K-médias.

Para mais detalhes sobre as técnicas de agrupamento, assim como as medidas de similaridade, ver Johnson e Wichern (2007).

Apesar de apresentar brevemente estes conceitos sobre a análise de cluster tradicional, nosso objetivo não é partir deles para a construção das técnicas de agrupamento espacial e sim apresentá-los para mostrar que, apesar de as técnicas de detecção de clusters espaciais apresentarem um objetivo semelhante, as técnicas em si são bastante diversas das que foram aqui mencionadas resumidamente.

## 2.3 Detecção de Clusters Espaciais

Os mais diversos eventos espaciais ocorrem e em alguns deles existe o interesse em determinar se esta ocorrência faz parte de um padrão ou se é possível atribui-la de forma razoável ao acaso. Um exemplo comum na literatura são os casos de câncer que ocorreram nas proximidades de instalações que lidavam com energia nuclear. Desta forma, era importante para as entidades responsáveis ter conhecimento se estes casos, de fato, formavam um cluster real.

Diversas técnicas foram empregadas ao longo do tempo para responder a esta questão: taxas que comparam casos observados e esperados, probabilidades através da suposição de distribuições para o número de casos, desenho de círculos em torno do que seria o foco, utilização de incidência, utilização da mortalidade, entre muitas outras (Openshaw et al., 1988).

Na análise de clusters tradicional buscam-se agrupar observações das mais diversas, como por exemplo, pessoas, animais, empresas ou imóveis. Já a detecção de clusters espaciais procura agrupar entidades georreferenciadas, que podem ser pontos ou áreas, buscando estabelecer relação entre regiões do plano de acordo com alguma variável sob estudo.

Como colocado em Kulldorff et al. (2003), apesar dos testes de detecção de clusters serem comumente utilizados na área epidemiológica, sua aplicação se estende às mais diversas áreas, como arqueologia, botânica, criminologia, demografia, ecologia, economia, engenharia, genética, geografia, história, neurologia, sociologia, zoologia, etc. No presente trabalho focaremos na aplicação destas técnicas na área epidemiológica.

A detecção de doenças pode ser feita em diferentes dimensões: espacial, temporal, ocupacional e suas combinações. O aumento do número de dimensões leva a um aumento na complexidade do teste para avaliar a significância de um *cluster* detectado.

Dentre os testes desenvolvidos, a maioria procura lidar com a detecção espacial, havendo casos também em que a detecção se dá puramente no tempo e outros que atuam na combinação tempo e espaço. Alguns inserem a dimensão do tempo como forma de flexibilizar o teste, dispensando, assim, a necessidade de se determinar o intervalo de tempo para uma aplicação puramente espacial.

Para a maioria dos testes, existem correções que podem ser aplicadas no intuito de levar em consideração a heterogeneidade demográfica existente entre as regiões. Uma região que possui mais pessoas idosas, por exemplo, pode ser considerada uma região mais suscetível à proliferação de certas doenças. Desta forma, é necessário levar em conta determinadas classes populacionais, para evitar que um número elevado de casos de certa doença seja associado a um *cluster*, quando é consequência de uma maior presença de determinado perfil populacional.

O conhecimento dos fatores associados a uma doença é fundamental para um controle efetivo de uma possível epidemia. Neste sentido, a utilização de técnicas de detecção de clusters espaciais fornece um ferramental para a verificação

de quais regiões, em uma ampla área, possuem um risco mais elevado de um indivíduo que ali se encontre vir a contrair determinada doença, e uma vez que uma região de alto risco é identificada, pode-se efetuar estudos no intuito de detectar fatores associados a esta elevação.

No processo de detecção de padrões espaciais, clusters podem ser encobertos pela distribuição de ocorrências de casos não relacionados à causa que gerou o *cluster*, assim como clusters podem ser gerados por fatores não relacionados ao processo da doença, como variações na distribuição populacional ou variação na distribuição de subgrupos demográficos com maior risco de infecção.

No estudo da detecção espacial de uma concentração de determinada doença acima do esperado, é necessário estabelecer o intervalo de tempo que será levado em conta e a delimitação da região geográfica. Para a região ainda é necessário estabelecer qual será o menor nível de desagregação possível (região censitária, município, microrregião,...).

Os primeiros esforços para detectar a existência de uma região com risco elevado foram através da apresentação espacial de valores absolutos ou relativos para cada área, no intuito de detectar uma região onde a doença sob estudo tivesse uma presença mais forte do que nas demais localidades. Atualmente, as técnicas de detecção de *cluster* são aplicadas em diversas áreas e foram desenvolvidas técnicas mais sofisticadas, buscando detectar um *cluster* que exibe uma característica especial.

De forma geral, ao utilizar-se os métodos estatísticos de detecção de clusters espaciais ou espaço-temporais, deseja-se verificar se o padrão espacial (espaço-temporal) observado, ocorrência de algum evento em uma determinada região geográfica, vem a ser um *cluster* ou se o padrão pode ser atribuído simplesmente ao acaso. Os diferentes métodos de detecção de clusters espaciais que foram propostos ao longo do tempo diferem em relação a seus potenciais. Dentre estes podemos citar os seguintes:

- Detecção: Procura responder à pergunta sobre a existência de *cluster*.
- Identificação: Fornece a localização aproximada do *cluster* detectado.

- Significância: Fornece a estatística do *cluster* encontrado e sua significância.

O intuito desta seção é dar ao leitor uma perspectiva da evolução das técnicas de detecção de *cluster* espaciais. Não se tem a pretensão de incluir todas. Para uma revisão extensiva deve-se remeter a trabalhos que já se preocuparam em efetuar essa revisão, como em Marshall (1991b)

### 2.3.1 Mapas Baseados em Probabilidades - Choynowski (1959)

Choynowski (1959) fornece uma metodologia que busca algo além do cálculo das frequências absolutas e relativas de cada região para a visualização espacial. A situação em que não existe uma região com risco superior de contrair a doença é caracterizada por uma incidência homogênea da doença em toda a área de estudo. É natural esperar que, sob esta condição, quanto maior a população de uma área, mais casos espera-se observar ali, de forma que o número esperado de casos será proporcional à população da região. Assim, se foram observados $C$ casos em todas as áreas, cada área terá um número esperado de casos igual a $Cn_i/N$, onde $n_i$ é a população da $i$-ésima área, $N$ a população total e $i = 1,\ldots,R$, onde $R$ é o número de regiões. Uma vez que o número de casos $C_i$ em cada região é uma contagem, pode-se modelá-lo utilizando uma distribuição de Poisson com média igual ao valor esperado, $\mu_i = \frac{Cn_i}{N}$, ou seja:

$$C_i \sim Poisson(\mu_i).$$

Com a distribuição de $C_i$ e a utilização do valor esperado como média, pode-se calcular a probabilidade de ocorrer um número igual ou superior ao valor observado de casos na área $i$. No caso do valor observado ser inferior ao valor esperado, calcula-se a probabilidade de ocorrer um número igual ou inferior ao observado:

$$\begin{aligned} &P(C_i \geq c_i) \quad \text{se } c_i \geq \mu_i \quad \text{ou} \\ &P(C_i \leq c_i) \quad \text{se } c_i < \mu_i. \end{aligned} \tag{2.1}$$

No artigo de Choynowski (1959) esta metodologia é aplicada à distribuição de tumores cerebrais em parte da Polônia. Uma observação feita no artigo é que entre os municípios que possuíam altas taxas e não exibiam razões aparentes (qualidade do serviço médico, composição etária etc.) todos possuíam uma população pequena, levando o autor a crer que isso pode ser consequência de variações amostrais.

Este método fornece uma probabilidade para cada área onde foi aplicado, não fornecendo uma informação de significância global em relação a presença de *cluster*. Nas situações em que existe o interesse em responder a questão sobre a existência ou não de um *cluster* em toda a área através de uma informação final única, seria necessária a adaptação do teste para lidar com a multiplicidade das probabilidades encontradas. Uma possibilidade é a utilização do método de Bonferroni.

### 2.3.2 Distribuição do Tamanho do Máximo de Clusters em uma linha - Naus (1965b)

Seguindo o rumo proposto por Choynowski (1959), da modelagem através da probabilidade, Naus (1965b) procura fornecer qual a probabilidade de, sorteados $N$ pontos independentes de uma distribuição uniforme $(0,1)$, existir um subintervalo de $(0,1)$ de tamanho $p$ que contém pelo menos $n$ dos $N$ pontos observados, onde $n > N/2$.

Em função do cálculo da probabilidade ser realizado em apenas uma dimensão, a aplicabilidade da técnica no contexto espacial fica comprometida. Porém, uma situação possível seria somar os casos por longitude ou latitude, transformando os dados para apenas uma dimensão. Desta forma, seríamos capazes de fornecer qual a probabilidade da ocorrência da distribuição observada a partir da suposição de que os dados se distribuem uniformemente e assim julgar se parece ter vindo de uma distribuição aleatória uniforme. Claro que este tipo de procedimento gera alguns inconvenientes, tornando a sua aplicação limitada.

No artigo de Naus (1965b) são apresentadas duas aplicações. A primeira delas procura resolver a questão: Começa-se a discar 15 telefones diferentes em tempos que se distribuem aleatoriamente em um período de um minuto. O tempo de discagem para uma chamada é de 10 segundos. Qual a probabilidade de oito ou mais chamadas estarem sendo discadas ao mesmo tempo?

A segunda aplicação consiste em um problema de contagem. Um gerador de impulsos emite de acordo com um processo de Poisson com média λ. Estes impulsos são recebidos por um contador. É realizado um registro pelo contador toda vez que $n$ impulsos ocorrem em um intervalo de tamanho menor do que $t$. Encontrar a distribuição de tempo de espera até o contador realize o primeiro registro.

### 2.3.3 Agrupamento de pontos aleatórios em duas dimensões - Naus (1965a)

Este artigo, publicado ainda no mesmo ano do anterior, traz a situação para as duas dimensões e, consequentemente, aumenta a complexidades do problema. Naus (1965a) não fornece o análogo do que foi calculado para uma dimensão, e sim uma forma de cálculo do limite superior e inferior da probabilidade de ocorrer o evento: existe um sub-retângulo do quadrado unitário, com lados de tamanho $u$ e $v$ orientados paralelamente aos eixos $x$ e $y$ respectivamente, que contenha pelo menos $n$ dos $N$ pontos.

Na situação especial em que $n = N$, é possível encontrar a probabilidade através da seguinte equação:

$$P(N|N;u,v) = N^2A^{N-1} + (N-1)^2A^N - N(N-1)A^{N-1}(u+v), \tag{2.2}$$

onde $A = uv$.

Dentre as limitações da técnica temos, além da ausência do cálculo exato da probabilidade, a necessidade de que os lados sejam paralelos aos eixos e o formato

fixo de um retângulo. O artigo fornece também a convergência da probabilidade de ocorrência do evento em questão quando $u$ e $v$ se aproximam de zero.

Na primeira aplicação encontrada no artigo, dado um valor fixo para a área, questiona-se sobre qual o formato da janela de detecção fornece a maior probabilidade de se encontrar um *cluster* maior e se o formato do *cluster* tem algum efeito sobre o número esperado de *cluster*. Ambos os questionamentos são respondidos de acordo com as equações fornecidas no artigo, colocando que dentre as janelas de detecção retangular, a que possui a maior probabilidade é aquela que é quadrada.

Na segunda aplicação calculam-se os limites superior e inferior da probabilidade de entre $10^\circ$ de latitude e longitude encontrarmos pelo menos 4, de 5 navios, dentro do quadrado correspondente a $2^\circ$ de longitude e $3^\circ$ de latitude. Os limites são apresentados e o autor conclui dizendo que o teorema de convergência apresentado indica que a melhor estimativa é obtida utilizando o limite inferior.

### 2.3.4 Estimador de Bayes Empírico para o Risco Relativo - Clayton e Kaldor (1987)

Clayton e Kaldor (1987) ressaltam que, no intuito de estimar um conjunto de riscos relativos, ou seja, estimar $\{\hat{\theta}_i, i = 1, \ldots, R\}$, a utilização do estimador de máxima verossimilhança para cada $\{\theta_i\}$ não é, necessariamente, a melhor escolha. Com suporte em alguns artigos que revelam estimadores com erros quadrados inferiores, busca-se então a construção de um estimador de Bayes empírico para o risco relativo.

Supondo $\theta_i$ como sendo o risco relativo para a i-ésima região, supõe-se que a distribuição de $C_i$, número de casos observados dado $\theta_i$, seja uma variável que segue uma distribuição de Poisson com média $\theta_i\mu_i$, onde $\mu_i$ é o número esperado de casos, e desta forma a distribuição marginal dos $C_i$ permite estimar os parâmetros de $f(\theta)$, distribuição conjunta dos $\theta_i$. A esperança *a posteriori* de $\theta_i$ dado $C_i$ fornece então as estimativas do risco relativo. No estudo, são considerados três modelos de mistura para $f(\theta)$: distribuição Gama, Log-Normal e a utilização de

um método não paramétrico para a estimação de $f(\theta)$. Para o modelo Log-Normal é considerado ainda o caso em que se supõe que os logaritmos dos riscos relativos são correlacionados através de uma dependência que leva em conta a proximidade geográfica.

O estimador empírico de Bayes é aplicado em casos de câncer de lábio na Escócia. É apresentado então uma tabela que possibilita a comparação entre a razão de mortalidade padronizada(SMR, na sigla em inglês), $C_i/E_i$, e os modelos propostos pelo artigo.

O modelo Gama e o Log-Normal apresentaram valores próximos. Ambos apresentaram uma amplitude menor do que o apresentado pelo SMR. Comparando-se com a ordenação do SMR, que representa o nível de incidência, as ordenações através do modelo gama e do modelo log-normal diferem pouco.

O modelo espacial de autocorrelação (CAR, na sigla em inglês) revela que os dados possuem um alto grau de relação espacial. Em relação ao caso Log-Normal não correlacionado, não leva em consideração a correlação espacial. A diferença entre as estimações dos riscos relativos é pequena, com exceção para as regiões com poucos casos e com SMR que difere substancialmente dos SMR's das regiões vizinhas.

A estimação de Bayes empírica não paramétrica dos riscos relativos traz informação sobre a quantidade de pessoas que está submetida a diferentes médias de risco relativo de vir a ter câncer de lábio. Mais uma vez a diferença na ordenação varia pouco quando comparado aos modelos Gama e Log-Normal.

O artigo conclui que as estimações propostas funcionam como métodos de alisamento do SMR puro e que, ao contrário de outros métodos de alisamento, o grau de alisamento é totalmente determinado pelos dados.

### 2.3.5 Um Teste para a detecção de clusters de Doenças - Whittemore et al. (1987)

No artigo de Whittemore et al. (1987) é proposto um método para responder à questão sobre a existência de um *cluster* espacial, ao invés de estabelecer uma

probabilidade para cada área como em Choynowski (1959). Apesar de responder à questão sobre a existência do *cluster*, o método não identifica sua localização. A hipótese nula tratada no artigo é a de que qualquer membro da população possui a mesma probabilidade de contrair a doença, pensando-se no contexto epidemiológico. A estatística do teste é construída utilizando-se a distância média entre todos os pares de observações:

$$\delta = \frac{\sum_{i<j} \Delta(i,j)}{\binom{n}{2}}, \tag{2.3}$$

onde $\Delta(i,j)$ representa a distância entre o $i$-ésimo e o $j$-ésimo caso. Em Whittemore et al. (1987) é fornecida a média e a variância de $\delta$ sob a hipótese nula e sua normalidade assintótica, portanto, é possível calcular $\{\delta - E(\delta)\}/\sqrt{\delta}$ e obter o p-valor.

Em Whittemore et al. (1987) também encontra-se uma versão da estatística que leva em consideração a composição demográfica da população, possibilitando corrigir possíveis conclusões equivocadas quando não se leva em conta os fatores demográficos associados a doença em estudo.

Utilizou-se o teste proposto para verificar a existência de *cluster* nos setores censitários de São Francisco em relação ao carcinoma de células escamosas do reto e do ânus ocorridas de 1973 até 1981, totalizando 63 casos. Utilizando-se os resultados assintóticos chega-se ao resultado de que o valor encontrado para $\delta$ nesta situação é significativo($p < 0,001$). Para efeito de constatação o presente artigo ainda realiza 1000 simulações e obtém valores para a média e variância bem próximos dos utilizados para avaliar a significância.

A utilização de testes para a existência global de clusters é útil quando a localização do *cluster* não é de interesse, por exemplo, quando se deseja saber se uma determinada doença é infecciosa (Kulldorff e Nagarwalla, 1995).

### 2.3.6 Máquina de Análise Geográfica - Openshaw et al. (1988)

No artigo de Openshaw et al. (1988) encontramos um procedimento denominado Máquina de Análise Geográfica (GAM, na sigla em inglês), para identificar a localização espacial de um possível *cluster*, objetivo a que o artigo anterior não se ateve.

Na construção do teste é necessário sobrepor uma malha quadriculada sobre o mapa. A distância entre cada um dos vértices é de $r/5$, onde $r$ é um valor que deve ser pré-definido e representará o raio do círculo que será utilizado na busca do *cluster*. Ao longo da análise varia-se o valor de $r$ entre alguns valores pré-definidos.

Cada um dos vértices será o centro de um círculo de raio $r$ e para cada um destes círculos calcula-se a quantidade de casos que ocorreram no seu interior e compara-se esse valor com o 99,8º percentil da distribuição do número de casos sob a hipótese nula. Este valor é encontrado no artigo com a utilização do método de Monte Carlo. Porém, devido ao alto custo computacional, outros artigos propuseram o cálculo através da suposição de que o número de casos segue uma distribuição de Poisson e então se calcula o quantil de ordem 99,8 para uma distribuição de Poisson com valor esperado igual $Cn_i/N$. Quando o 99,8º percentil é superado, desenha-se no mapa o círculo correspondente.

O valor $r$ do raio do círculo varia de acordo com valores pré-fixados, repetindo-se toda a análise para cada tamanho de raio selecionado. O resultado final do procedimento é um conjunto de círculos de diferentes tamanhos individualmente significativos desenhados no mapa.

A escolha do 99,8º percentil ocorre em função do grande número de comparações, o que eleva a probabilidade do erro tipo I. Assim como no método proposto por Choynowski (1959), aqui não é fornecido uma medida final para responder sobre a existência de *cluster* de forma geral e, devido ao grande número de testes, a utilização do método de Bonferroni levaria a um teste bastante conservador. A aplicação da técnica tem também um apelo visual, uma vez que fornece uma região sombreada pelos círculos que foram desenhados, variando-se a intensidade

do sombreamento de acordo com a quantidade de círculos desenhados em cada área.

A aplicação do teste foi feita nos 853 casos de leucemia em crianças com menos de 15 anos, ocorridos de 1968 a 1985 em algumas regiões específicas da Inglaterra. De todos os 812.993 círculos examinados, 1792 foram considerados significantes. Estes se distribuem basicamente em volta de 5 áreas.

### 2.3.7 Monitorando clusters de doenças - Turnbull et al. (1989)

Sob a hipótese de ausência de um *cluster* espacial, duas regiões distintas com a mesma quantidade de pessoas esperariam observar a mesma quantidade de ocorrências de determinada doença. Baseando-se então neste fato, Turnbull et al. (1989) procuram redefinir as áreas para que cada nova área contenha o mesmo número $P$ de pessoas, fazendo com que o número esperado de casos em cada nova área seja a mesma e aumentando a comparabilidade entre as diferentes novas áreas.

O valor da população que cada nova área deverá conter é especificado *a priori*. Utilizando-se do centroide da região, busca-se englobar as regiões vizinhas até alcançar a população fixada, podendo assim ocorrer sobreposição das novas áreas. Com a redefinição da unidade espacial, calcula-se então para cada uma das novas áreas o número de casos observados, $T_i$, diferindo, portanto, do número de casos observados nas regiões originais, $C_i$.

Sob a hipótese nula considera-se $\{T_i; i = 1,2,\ldots,R\}$, onde $R$ é o número de regiões, como sendo variáveis aleatórias identicamente distribuídas sob a hipótese nula, porém não independentes, e portanto pode-se testar desvios da hipótese nula. O artigo sugere dentre as várias opções para a construção da estatística de teste, a escolha do máximo $(M_R = Max(T_1, T_2, \ldots, T_R))$, como sendo a escolha mais natural. Em função do valor $P$ fixado para a população a hipótese alternativa se torna: existe um *cluster* entre as novas áreas definidas de população igual a $P$.

Para obter a distribuição de $M_R$ seria necessário computar o seu valor para todas as possíveis distribuições dos $C$ casos para uma população de tamanho $N$,

o que equivaleria a $N!/(C!(N-C)!)$ permutações. Devido ao custo computacional associado à obtenção da distribuição exata, opta-se por utilizar uma simulação de Monte Carlo e então obter uma amostra do conjunto de todas as permutações. Desta forma, calculando o $M_R$ para cada uma das amostras e comparando-se com o valor observado para o caso real, obtemos a significância para o *cluster* encontrado.

O método indica que se testem alguns valores para $P$ de tal forma a fornecer certa flexibilidade para o raio populacional do *cluster*. Para que se possa gerar uma significância geral é necessário se levar em consideração estes vários valores estabelecidos para $P$, o que acarreta em lidar com múltiplos testes. Além disso, deve-se levar em consideração que existe correlação entre os testes realizados para cada $P$.

A aplicação do método proposto se dá comparativamente aos procedimentos encontrados em Whittemore et al. (1987) e em Openshaw et al. (1988) para os casos de leucemia ocorridos em parte do norte do estado de Nova Iorque. A escolha da leucemia ocorreu em função da sua distribuição ser preponderantemente uniforme, apesar do conhecimento de que alguns clusters aparentes ocorreram. Na aplicação de cada um dos métodos são realçadas as respectivas vantagens e desvantagens comparativamente aos outros métodos.

### 2.3.8 A detecção de Clusters de Doenças Raras - Besag e Newell (1991)

No artigo de Besag e Newell (1991) encontramos um mecanismo semelhante ao visto em Turnbull et al. (1989), com círculos sobrepostos que procuram para cada caso acumular as regiões vizinhas para redefinir a área. Porém ao invés de se atingir um valor pré-determinado da população, atinge-se um valor pré-determinado de casos. Desta forma, para cada caso, teremos uma zona com um número de casos igual a $k+1$, onde $k+1$ é o valor de casos a ser atingido incluindo-se um caso já presente no centro. Para cada uma das zonas efetua-se um teste de significância para verificar se a zona consiste em um *cluster*.

A hipótese nula $H_0$ é a de que o número total de casos se distribui de forma aleatória pela população. Considerando-se um caso específico, atribui-se o rótulo $A_0$ para a região em que o caso ocorre e $A_1, A_2 \ldots$, para as outras regiões à medida que a distância em relação ao ponto considerado aumenta. Desta forma, define-se $D_i = \left(\sum_{j=0}^{i} c_j\right) - 1$ e $u_i = \left(\sum_{j=0}^{i} n_j\right) - 1$, de tal forma que $D_0 \leq D_1 \leq \ldots$ são os casos acumulados em $A_0, A_1, \ldots$ e $u_0 \leq u_1 \leq \ldots$ são as populações acumuladas correspondentes. Define-se então $M = min\{i : D_i \geq k\}$, e portanto, um valor observado pequeno de $M$ é indicativo de um *cluster*. Isto ocorre em função de o número estabelecido de casos ser atingido em um número pequeno de regiões, gerando assim um possível *cluster* em volta de $A_0$, cuja avaliação da intensidade desta concentração ocorrerá através do cálculo da significância. Desta forma, sendo $m$ o valor observado de $M$, o nível de significância é dado por $P(M \leq m)$, calculada sob $H_0$.

Utilizando então uma aproximação da distribuição hipergeométrica pela distribuição de Poisson para a probabilidade de observar $k$ indivíduos entre $u_m$ com a doença, tem-se que:

$$P(M \leq m) = 1 - P(M > m) \tag{2.4}$$

$$= 1 - \sum_{s=0}^{k-1} \exp\{-u_m p\}(u_m p)^s / s!. \tag{2.5}$$

Desta forma, selecionando-se um nível de significância igual a $\alpha$ o artigo sugere desenhar todos os clusters que atingiram valor de significância menor ou igual a $\alpha$ e portanto teremos um resultado similar ao de Openshaw et al. (1988). No artigo ainda é proposta uma forma de avaliar se a quantidade de clusters encontrados é significativa, algo próximo de um teste como o de Whittemore et al. (1987), ou seja, um teste global sobre a existência de clusters.

A aplicação da técnica utiliza o mesmo banco de dados encontrado em Openshaw et al. (1988), alterando apenas o intervalo de tempo de 1968–1985, para 1975–1985. Para a aplicação foram utilizados os seguintes valores para $k = 2, 4, 6, 8$, no entanto, o artigo exibe somente o resultado para $k = 4$. Dentre

as análises encontradas na exibição dos resultados, temos a comparação do número de clusters encontrados com o número esperado, a separação dos clusters encontrados em grupos sem sobreposição e a utilização da simulação de Monte Carlo para verificar significância de cada um dos clusters e para a região como um todo.

## 2.4 Estatística Scan

Antes de entrarmos na descrição da estatística scan de Kulldorff, cabe uma reflexão sobre o que os métodos anteriores propuseram, de tal forma a perceber a continuidade existente entre estes trabalhos e o desenvolvido em Kulldorff e Nagarwalla (1995).

Todos eles possuem uma hipótese nula comum, a de que não existindo *cluster* espacial todas as pessoas têm probabilidades iguais de se tornarem um caso, ou seja, contrair a doença em estudo. Para o número de casos também é comum imagina-los como uma Poisson, uma vez que esse tipo de distribuição é bastante utilizada na modelagem de contagens.

Uma diferença entre eles ocorre na definição do processo de varredura, na definição do que será o centro de cada círculo, que pode ser feita de duas formas, através de uma malha como em Openshaw et al. (1988) ou utilizando dos centroides das regiões. Uma vez definido o centro, pode-se proceder de forma mais geral das seguintes formas:

- Como em Openshaw et al. (1988) que fixa valores para o raio e varre a área.
- Como em Turnbull et al. (1989) que fixa um valor populacional e gera novas áreas que possuem tal população;
- Como em Besag e Newell (1991) que fixa o número de casos.

Percebemos que todos os três métodos fixaram o seu raio de busca, apesar de que em cada um deles encontramos uma definição diferente para o que seria o raio de varredura.

Em relação ao teste, Besag e Newell (1991) e Openshaw et al. (1988) constroem seus métodos através de múltiplos testes. Já Besag e Newell (1991) fornece um teste de tal forma a responder acerca da existência global de um *cluster*. Turnbull et al. (1989) apesar de apresentar construção semelhante, resume seu método em apenas uma estatística que terá sua significância testada por meio do método de Monte Carlo.

### 2.4.1 Kulldorff e Nagarwalla (1995)

O artigo de Kulldorff (1997), que apresenta o método sob o qual se apoia a técnica que será apresentada nesta dissertação, teve sua base no artigo de Kulldorff e Nagarwalla (1995), quando são estabelecidas as conexões com os artigos de Openshaw et al. (1988) e Turnbull et al. (1989).

Como nos outros métodos, é necessário o conhecimento, para cada área, do centroide geográfico ou populacional, da população e do número de casos. A construção da malha de pontos que será utilizada como centro dos círculos, pode ser como em Openshaw et al. (1988), de forma igualmente espaçada, ou como em Turnbull et al. (1989) que utiliza os próprios centroides. Uma vez estabelecida a malha dos centros, vem a questão do raio. Ao invés de pré fixar um valor para o raio como nos outros métodos (‘raio de população’ em Turnbull et al., 1989, e ‘raio de casos’ em Besag e Newell, 1991), o método permite que o raio varie de 0 em diante, permitindo assim que clusters de diferentes tamanhos, populações e números de casos possam ser detectados.

Apesar de possibilitar ao raio variar, não é de interesse que o raio seja muito grande, uma vez que se incluindo quase toda a população, pode-se argumentar que na verdade o *cluster* seja a menor parte, sendo assim um *cluster* que reúne regiões de risco baixo, o que foge dos objetivos propostos. Kulldorff e Nagarwalla (1995) observam que um valor aparentemente razoável para limitar o raio seja 50% da população. Outras possibilidades incluem 50% do número de regiões ou 50% do número de casos. O valor de 50% também pode ser alterado, desde que seja realizado a priori e não por tentativa e erro.

Para cada centro do círculo e variando-se o raio, o procedimento procura criar zonas que são definidas por todos os indivíduos pertencentes às regiões cujos centroides foram incluídos dentro do círculo. Para cada zona calcula-se o valor da razão de verossimilhança de acordo com o modelo proposto no artigo e seleciona-se o maior valor.

Dentre as novidades propostas pelo método, está a utilização da razão de verossimilhança, que equilibra em seu cálculo o quão grande é o risco dentro e fora do *cluster* comparativamente ao risco calculado na situação em que não há *cluster*. Os detalhes acerca da construção da razão de verossimilhança assim como detalhes adicionais e aperfeiçoamentos que ocorreram em Kulldorff (1997), serão descritos na próxima seção.

Para avaliar a estatística, Kulldorff e Nagarwalla (1995) utilizam a bases de dados encontrada em Turnbull et al. (1989) e, desta forma, compara o seu procedimento não só com o de Turnbull et al. (1989), mas também com os procedimento de Whittemore et al. (1987) e Openshaw et al. (1988).

Calcula-se a razão de verossimilhança para todas as zonas e obtém-se a significância através da simulação de Monte Carlo, onde o número de casos é distribuído aleatoriamente para a população. Para cada uma das regiões encontradas pelo método de Openshaw et al. (1988) apresentado em Turnbull et al. (1989), com exceção de uma, seleciona-se a zona que obteve a maior razão de verossimilhança e exibe-se a sua posição de acordo com o seu posicionamento em relação as razões de verossimilhança calculadas sob a hipótese nula.

Chega-se a conclusão de que o *cluster* mais provável difere do encontrado em Turnbull et al. (1989). Porém, o *cluster* encontrado nesse artigo também mostra-se significativo segundo a estatística da razão de verossimilhança. Vale lembrar que apesar de ambos os métodos possuírem o mesmo objetivo, suas hipóteses diferem em função do método presente em Turnbull et al. (1989) fixar a população, alterando a hipótese alternativa para clusters de um tamanho fixo.

### 2.4.2 A Estatística Scan espacial de Kulldorff

Kulldorff (1997) apresenta uma estatística de varredura já amadurecida, quando comparada com a de Kulldorff e Nagarwalla (1995), colocando Naus (1965a), Turnbull et al. (1989) e Kulldorff e Nagarwalla (1995) como casos especiais.

Em Kulldorff (1997) mantem-se grande parte do que foi apresentado em Kulldorff e Nagarwalla (1995), uma estatística de varredura espacial capaz de verificar a existência de clusters espaciais e determinar sua localização aproximada. Enquanto o artigo anterior apresentava apenas a modelagem através da distribuição de Poisson, o presente já inclui também o modelo Binomial. A diferença no resultado de acordo com a escolha entre as duas distribuições é reduzida quando o número de eventos é pequeno comparado ao tamanho da população. Em outras circunstâncias a escolha será determinada pela aplicação. Para mais detalhes ver Kulldorff (1997).

A estatística de varredura proposta pode ser empregada tanto em situações em que as ocorrências encontram-se agregadas em um determinado nível geográfico, quanto quando são conhecidas as coordenadas exatas da ocorrência do evento.

A hipótese nula considerada no artigo é a de que não existe *cluster* espacial e, desta forma, os casos são distribuídos de maneira uniforme pela população. Sendo assim o número de casos observados em cada área é proporcional ao tamanho da população na respectiva área.

Uma das propriedades da estatística proposta, que é provada no presente artigo, é a de que se a hipótese nula for rejeitada, de acordo com a distribuição atual dos pontos, então fixando-se os pontos presentes dentro do *cluster* mais provável, não importa de que forma os pontos fora do *cluster* estejam distribuídos, continua-se a rejeitar a hipótese nula. Kulldorff (1997) ressalta que, embora isso pareça evidente, muitas estatísticas não têm essa propriedade.

A aplicação da estatística de razão de verossimilhança utiliza as ocorrências de síndrome da morte súbita infantil na Carolina do Norte. A população conside-

rada foi o número de nascimentos. Duas zonas são consideradas significativas segundo os dois modelos. Ao controlar-se por raça, um terceiro *cluster* significativo é revelado.

A seguir é apresentada a dedução da verossimilhança para os modelos Poisson e Bernoulli e os comentários pertinentes a cada um deles.

#### Modelos Bernoulli e Poisson

Utilizaremos a seguinte notação em ambos os modelos:

- $N$ é a população total;
- $C$ representa o número total de casos;
- $Z$ é o conjunto de todas as combinações de regiões possíveis;
- $z$ é definido como um conjunto específico de regiões (zona);
- $n_z$ a população na zona $z$;
- $c_z$ é o número de casos na zona $z$;
- $\mu_z$ é o número esperado de casos em $z$;
- $p$ é a probabilidade de uma observação ser um caso dentro da zona $z$;
- $q$ é a probabilidade de ocorrer um caso fora da zona $z$.

#### Modelo Bernoulli

Sob $H_0$ não existe nenhuma área que represente um *cluster* espacial e sob $H_a$ existe um conjunto de regiões que possuem a probabilidade de uma observação se tornar um caso superior a essa mesma probabilidade no conjunto das demais regiões. Portanto, definimos as seguintes hipóteses:

$$H_0 \quad : \quad p = q, \tag{2.6}$$

$$H_a \quad : \quad p > q. \tag{2.7}$$

Desta forma, a verossimilhança sob $H_0$ será dada por:

$$L_0(p) = p^C(1-p)^{N-C}, \tag{2.8}$$

e portanto, teremos:

$$l_0(p) = C\log p + (N-C)\log(1-p). \tag{2.9}$$

Chegando então ao estimador de p sob $H_0$:

$$\hat{p} = \frac{C}{N}. \tag{2.10}$$

Sob a hipótese alternativa temos que existe uma localização espacial $\hat{z}$ onde existe um risco maior de uma pessoa vir a se tornar um caso, ou seja, onde $p > q$. Desta forma a verossimilhança toma o seguinte formato:

$$L(z,p,q) = p^{c_z}(1-p)^{n_z-c_z}q^{C-c_z}(1-q)^{(N-C)-(n_z-c_z)}, \tag{2.11}$$

e a log-verossimilhança será dada por:

$$l(z,p,q) = c_z\log p + (n_z-c_z)\log(1-p) + (C-c_z)\log q + [(N-C)-(n_z-c_z)]\log(1-q). \tag{2.12}$$

E portanto chegamos aos estimadores de $p$ e $q$ sob a hipótese alternativa:

$$\hat{p} = \frac{c_z}{n_z} \quad ; \quad \hat{q} = \frac{C-c_z}{N-n_z}. \tag{2.13}$$

Substituindo então a estimativa do parâmetro $p$ sob $H_0$ em $L_0(p)$ teremos:

$$L_0(p,q) = \left(\frac{C}{N}\right)^C \left(1-\frac{C}{N}\right)^{N-C}. \tag{2.14}$$

Substituindo a estimativa do parâmetro $p$ e $q$ sob $H_a$ em $L(z,p,q)$ teremos:

$$L(z,p,q) = \left(\frac{c_z}{n_z}\right)^{c_z} \left(1-\frac{c_z}{n_z}\right)^{n_z-c_z} \left(\frac{C-c_z}{N-n_z}\right)^{C-c_z} \left(1-\frac{C-c_z}{N-n_z}\right)^{(N-C)-(n_z-c_z)}. \quad (2.15)$$

Desta forma podemos definir a razão de verossimilhança dada pela fórmula:

$$\lambda = \frac{\sup_{z\in Z, p>q} L(z,p,q)}{\sup_{p=q} L(z,p,q)} = \frac{L(\hat{z})}{L_0}; \quad (2.16)$$

$$= \frac{\left(\frac{c_z}{n_z}\right)^{c_z} \left(1-\frac{c_z}{n_z}\right)^{n_z-c_z} \left(\frac{C-c_z}{N-n_z}\right)^{C-c_z} \left(1-\frac{C-c_z}{N-n_z}\right)^{(N-C)-(n_z-c_z)}}{\left(\frac{C}{N}\right)^{C} (1-\frac{C}{N})^{N-C}}. \quad (2.17)$$

Observando a equação 2.17 percebemos que o denominador não depende da disposição espacial dos dados, e somente do número de casos e da população total. Já no numerador a dependência da disposição espacial dos dados é realizada através da inserção do conjunto de regiões $z$, que é construído através de regiões inseridas em um círculo de raio definido.

#### Modelo Poisson

No caso do modelo Poisson, definiremos $\mu_{\bar{z}}$ como sendo o número de casos esperados fora da zona $z$. Desta forma, sob $H_0$ teremos:

$$\mu_z = pn_z \;\; ; \;\; \mu_{\bar{z}} = p(N-n_z). \quad (2.18)$$

Desta forma, nossa verossimilhança será dada por:

$$L_0(z,p) = \frac{\mu_z^{c_z} e^{-\mu_z}}{c_z!} \times \frac{\mu_{\bar{z}}^{c_{\bar{z}}} e^{-\mu_{\bar{z}}}}{c_{\bar{z}}!}; \quad (2.19)$$

$$= \frac{(pn_z)_z^c e^{-pn_z}}{c_z!} \times \frac{[p(N-n_z)]^{(C-c_z)} e^{-p(N-n_z)}}{(C-c_z)!}. \quad (2.20)$$

obtendo assim a log-verossimilhança exibida abaixo:

$$\begin{aligned} l_0(z,p) &= c_z[\log p + \log n_z] - pn_z - \log c_z! + (C-c_z)[\log p + \log(N-n_z)] \\ &\quad -p(N-n_z) - \log[(C-c_z)!]. \end{aligned} \tag{2.21}$$

Que fornecerá estimador igual ao apresentado para o modelo Bernoulli, equação (2.10), isto é, $p = C/N$.

No caso da hipótese alternativa, onde teremos uma probabilidade maior de ocorrência de um caso em determinada região do mapa, teremos o seguinte:

$$\mu_z = pn_z \;\; ; \;\; \mu_{\bar{z}} = q(N-n_z). \tag{2.22}$$

Desta forma, nossa verossimilhança será dada por:

$$L_0(z,p,q) = \frac{(pn_z)^{c_z} e^{-pn_z}}{c_z!} \times \frac{[q(N-n_z)]^{(C-c_z)} e^{-q(N-n_z)}}{(C-c_z)!}, \tag{2.23}$$

e a log-verossimilhança apresentará o seguinte formato

$$\begin{aligned} l_0(z,p,q) &= c_z[\log p + \log n_z] - pn_z - \log c_z! + (C-c_z)[\log q + \log(N-n_z)] \\ &\quad -q(N-n_z) - \log[(C-c_z)!]. \end{aligned} \tag{2.24}$$

Derivando e igualando a zero obteremos estimadores para $p$ e $q$ iguais aos fornecidos para o modelo Bernoulli, equação (2.13).

Substituindo $p = c_z/n_z$ e $q = (C-c_z)/(N-n_z)$ em $L(z;p;q)$ teremos:

$$L(z) = \frac{c_z^{c_z}(C-c_z)^{C-c_z} e^{-C}}{c_z!(C-c_z)!}. \tag{2.25}$$

Substituindo $p = C/N$ em $L_0(z;p)$:

$$L_0(z) = \frac{\mu_z^{c_z}(C-\mu_z)^{C-c_z} e^{-C}}{c_z!(C-c_z)!}, \tag{2.26}$$

onde $\mu_z = \frac{Cn_z}{N}$.

Desta forma, obtemos a razão de verossimilhança:

$$\lambda = \frac{L}{L_0} = \begin{cases} \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z} & , \ \frac{c_z}{\mu_z} > \frac{C-c_z}{C-\mu_z}; \\ 1 & , \ \text{caso contrário.} \end{cases} \tag{2.27}$$

A definição de $\lambda = 1$ para os casos em que $\frac{c_z}{\mu_z} \leq \frac{C-c_z}{C-\mu_z}$ ocorre em função de não estarmos interessados em clusters negativos, quando a ocorrência de casos é inferior ao número esperado de casos. No entanto, caso se tenha interesse pela busca de clusters negativos, basta realizar esta alteração de sinal na condição da equação (2.27) e a busca de clusters se concentrará em conjuntos de regiões em que o valor esperado supera o observado.

Observamos através da equação (2.27) que o cálculo da razão de verossimilhança se concentra na comparação entre o valor observado de casos com o valor esperado de casos dentro da região analisada em contraste com o cálculo deste mesmo valor só que para fora da região estudada.

A tabela 2.1 exemplifica o cálculo, utilizando o logaritmo natural da razão de verossimilhança, para algumas combinação de valores observados $c_z$ e esperados $\mu_z$ em um determinado conjunto de regiões $z$ supondo-se um número total de casos $C$ igual a 1000.

Tabela 2.1: Cálculo do ln da razão de verossimilhança.

| $\mu_z \backslash c_z$ | 1 | 5 | 10 | 50 | 100 | 500 |
|---|---|---|---|---|---|---|
| 1 | 1 | 4,06 | 14,07 | 147,82 | 366,59 | 2.761,23 |
| 5 | 1 | 1 | 1,94 | 71,16 | 209,26 | 1.958,52 |
| 10 | 1 | 1 | 1 | 41,29 | 144,48 | 1.614,46 |
| 50 | 1 | 1 | 1 | 1 | 20,65 | 830,37 |
| 100 | 1 | 1 | 1 | 1 | 1 | 510,83 |

Percebemos que à medida que a proporção do número total de casos se concentra todo em um conjunto de regiões, o ln da razão de verossimilhança cresce, representando a elevação da concentração de casos (vide equação (2.27)). O valor esperado reflete o tamanho da população do conjunto de regiões sob análise em relação ao total da população. Desta forma, quanto maior a população dentro

do conjunto de regiões $z$, maior a quantidade de casos esperados. Utilizando a equação $\mu_z = Cn_z/N$, para um valor de $\mu_z = 100$ e $C = 1000$, chegamos a conclusão que a soma das populações das regiões presentes na zona $z$ equivale a 10% da população total. Assim, ao observar-se 50% dos casos em um conjunto de regiões que tem população equivalente a 10% da população total, encontramos o ln da razão de verossimilhança representado na última linha da última coluna da tabela 2.1.

De forma simples, o que a razão de verossimilhança faz é comparar a quantidade de casos em $z$ com o tamanho da populacão em $z$, penalizando por esta mesma comparação realizada fora de $z$.

### 2.4.3 Oliveira et al. (2011)

Como ressaltado em Kulldorff e Nagarwalla (1995), na aplicação dos métodos de detecção de clusters é difícil que o *cluster* mais provável corresponda exatamente ao *cluster* 'real'. Variando-se um pouco o círculo que definiu o *cluster* mais provável será possível encontrar outros conjuntos de regiões muito semelhantes ao que forneceu o *cluster* mais provável e que possuem razão de verossimilhança quase tão alta quanto a que foi definida como o máximo. Desta forma, pode-se encarar a estatística scan de Kulldorff (1997) como sendo uma estimativa para a posição e raio do *cluster* real.

É dentro deste contexto que Oliveira et al. (2011) buscam sair da estimação da posição e raio do verdadeiro *cluster*, fornecendo mais informação sobre de que forma cada região estaria ligada ao *cluster* 'real' e propondo um método que busca fornecer mais informação para cada região do que o usual pertence ou não pertence ao *cluster*.

O método proposto define uma função de intensidade de tal forma a medir a plausibilidade de cada região ser parte de uma anomalia no mapa. Para construção da função de intensidade, altera-se o mecanismo de detecção de *cluster*. Ao invés de se detectar o *cluster* mais provável no mapa original, procura-se detectar

o *cluster* em um mapa cujos valores observados sejam replicações de um vetor de variáveis aleatórias, cujas médias são baseadas nos valores observados.

A replicação deste mecanismo $m$ vezes, utilizando a estatística scan circular para detectar o *cluster* mais provável em cada repetição, fornecerá um conjunto de $m$ razões de verossimilhança $\{LLR_1,\ldots,LLR_m\}$ correspondentes aos clusters mais prováveis $\{MLC_1,\ldots,MLC_m\}$. As razões de verossimilhança são então ordenadas em ordem crescente $\{LLR_{(1)},\ldots,LLR_{(m)}\}$ para os correspondentes *cluster* mais prováveis $\{MLC_{(1)},\ldots,MLC_{(m)}\}$. Desta forma, define-se a função de intensidade $f:1,\ldots,m\to\Re$ por $f(j)=LLR_{(j)}, j=1,\ldots,m$. Para cada área calcula-se então:

$$q(a_i)=\frac{1}{m}\arg\max_{1\leq j\leq m, a_i\in MLC_{(j)}} f(j), \quad i=1,\ldots,R. \tag{2.28}$$

Se a área $a_i$ não pertencer a nenhum dos clusters $MLC_1,\ldots,MLC_m$, faz-se $q(a_i)=0$. O valor $q(a_i)$ pode ser interpretado como a importância relativa da área $a_i$ como parte da anomalia no mapa.

O método é aplicado tanto em simulações numéricas quanto em dados reais. Para as simulações numéricas foram considerados cinco cenários: um único *cluster* de formato circular com dois níveis de risco relativo, um *cluster* com formato em L e por último, dois clusters com formato circular também com dois níveis de risco relativo.

Para os dados reais foram utilizados informações do número de homicídios no estado de Minas Gerais, do número de casos de câncer de mama no nordeste dos Estados Unidos da América e nos casos de doença de chagas em mulheres no estado puerperal em Minas Gerais.

Em todas as aplicações compara-se o desempenho do método de Oliveira et al. (2011) com o de Kulldorff (1997).

## 2.5 Estatística Scan Espaço-Temporal

As técnicas de detecção de clusters espaciais necessitam da fixação de um período de tempo para a agregação dos casos que ocorreram dentro deste pe-

ríodo. Este período pode ser de dias até anos, e a escolha do período utilizado pode ser questionável. A especificação desse valor pode gerar dois tipos de problemas. Incluindo-se poucos períodos, o teste pode não ter poder suficiente para detectar uma doença de risco baixo a moderado que ocorre há um tempo considerável. Caso se inclua muitos períodos, o teste pode não ter poder suficiente para detectar um risco elevado que ocorreu em um período curto. Esta situação é exemplificada em Kulldorff (2001) para o caso da detecção de novos clusters que estão aparecendo.

Entre os métodos iniciais de detecção espaço-temporal podemos citar o de Knox e Bartlett (1964), em que eram estabelecidas distâncias de corte para o espaço e para o tempo de tal forma a julgar se pares de observações seriam consideradas próximas ou distantes para cada uma das dimensões, obtendo-se então uma tabela $2 \times 2$. Utilizando-se dos totais marginais é possível calcular o valor esperado para cada uma das células e em especial para aquela que contém as observações próximas em relação ao espaço e ao tempo. Assim, utilizando a distribuição de Poisson com média igual ao valor esperado para esta última célula, pode-se calcular a probabilidade de se observar um número de casos igual ou superior ao observado nos dados, fornecendo assim a significância.

Seguindo a mesma linha de raciocínio utilizada por Knox e Bartlett (1964), diversas outras técnicas foram definidas alterando-se a medida utilizada para avaliar a associação das distâncias no tempo e espaço e mudando-se o critério utilizado para determinar a significância (Mantel, 1967).

Estes métodos são construídos sob o princípio de que, se existe um *cluster*, as observações estarão próximas tanto no espaço quanto no tempo, enquanto que se as observações não são relacionadas estas possuirão uma separação maior no tempo e no espaço. Desta forma, busca-se avaliar se há uma relação positiva entre distância espacial e distância temporal. A princípio, exceto em algumas situações, o método de medidas de distância a partir das observações pareadas é insensível a clusters puramente espaciais ou puramente temporais. Em relação a esta associação positiva encontramos o seguinte trecho em Whittemore et al. (1987): ‘No

entanto, casos de doenças crônicas causados por um agente espacialmente localizado devem estar próximos no espaço, mas é improvável estarem próximos no tempo, em função de o tempo entre a exposição e o diagnóstico ser variável e longo.'

A estatística de detecção de *cluster* espaço-temporal que será apresentada a seguir não tem sua base no pareamento de observações, como nos métodos citados anteriormente, e sim em uma extensão da estatística scan encontrada em Kulldorff (1997).

### 2.5.1 Uma estatística scan espaço temporal - Kulldorff et al. (1998)

Como foi mencionado anteriormente, a extensão da estatística scan de Kulldorff (1997) do espaço para o espaço-tempo ocorreu através da ampliação da estatística de varredura com formato circular para um formato cilíndrico. A base circular corresponde à dimensão geográfica e a altura, ao intervalo de tempo.

Sob $H_0$ assume-se que o número de casos, $C_i$, $i = 1,\ldots,R$, seja distribuído segundo uma Poisson com risco constante no espaço e no tempo e sob $H_a$ assume-se que o risco seja distinto dentro e fora de pelo menos um cilindro.

Os cálculos realizados em Kulldorff (1997) são replicados em Kulldorff et al. (1998), porém, onde havia uma janela circular, agora obtém-se uma janela de formato cilíndrico, que irá varrer a região de estudo no espaço e no tempo. Desta forma, a equação 2.27 tem sua versão espaço-temporal abaixo:

$$\lambda = \sup_{z \in Z} \left( \frac{\sum\limits_{(i,t) \in z} \sum c_{it}}{\sum\limits_{(i,t) \in z} \sum \mu_{it}} \right)^{\sum\limits_{(i,t) \in z} \sum c_{it}} \left( \frac{\sum\limits_{(i,t) \notin z} \sum c_{it}}{\sum\limits_{(i,t) \notin z} \sum \mu_{it}} \right)^{\sum\limits_{(i,t) \notin z} \sum c_{it}} \tag{2.29}$$

se

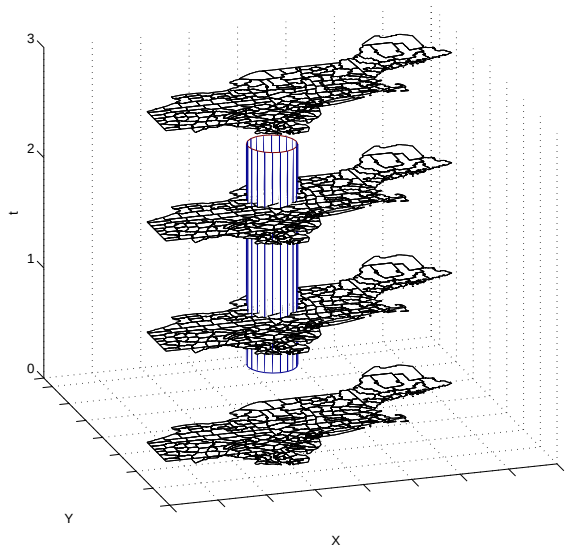
$$\frac{\sum\limits_{(i,t) \in z} \sum c_{it}}{\sum\limits_{(i,t) \in z} \sum \mu_{it}} > \frac{\sum\limits_{(i,t) \notin z} \sum c_{it}}{\sum\limits_{(i,t) \notin z} \sum \mu_{it}},$$

onde $\mu_i$ represente o número de casos esperados para a $i$-ésima região.

Em Kulldorff et al. (1998) não fica claro se o denominador da razão de verossimilhança é constituído pela população em um momento fixo do tempo ou se leva em consideração as populações para os diferentes períodos de tempo. De toda forma, apesar disso aumentar/reduzir a verossimilhança, a princípio não há alteração na significância em função de a mesma fórmula ser aplicada tanto sob $H_0$ quanto sob $H_a$. Optamos em incluir na equação (2.29) a situação em que a população nos diferentes períodos de tempo entra no cálculo em função de levar em consideração variações de tamanhos populacionais ao longo do tempo.

Se ao tamanho da janela fosse permitido variar de tal forma a praticamente encobrir todo o mapa, a verossimilhança não mais refletiria um *cluster* de alto risco dentro do cilindro, e sim um *cluster* de risco reduzido fora do cilindro. Desta forma, restringe-se a janela a ter no máximo metade do máximo de casos esperados e o intervalo de tempo a ser no máximo igual a metade do período total. A inclusão do cilindro cobrindo todo o período de tempo visa possibilitar a detecção de um *cluster* puramente espacial. A figura 2.1 fornece um exemplo da janela espaço-temporal.

Figura 2.1: Exemplo da utilização da estatística scan espaço-temporal.




Esta metodologia foi aplicada a um conjunto de dados referentes a casos de neoplasia maligna no cérebro de 1973 até 1991 no estado americano do Novo

México. Em 1991 houve grande preocupação no condado de Los Alamos, parte do Novo México, se um incremento observado de casos de câncer no cérebro nesta região não seria consequência da implantação do projeto Manhattan no período da guerra fria, responsável por desenvolver e testar a primeira bomba nuclear. As autoridades responsáveis não encontraram taxas alarmantes, colocando o aumento na categoria das flutuações aleatórias e concluindo que não havia assim razões para alarme. A aplicação da estatística scan espaço-temporal de Kulldorff et al. (1998) apresenta resultados semelhantes.

### 2.5.2 Estatística scan prospectiva - Kulldorff (2001)

Em Kulldorff (2001) coloca-se a importância de efetuar uma detecção rápida de clusters que estão emergindo, possibilitando assim, detectar doenças previamente desconhecidas, detectar fatores de risco previamente desconhecidos para doenças conhecidas e detectar locais previamente desconhecidos de fatores de risco conhecidos.

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 na ocorrência de falsos alarmes 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.

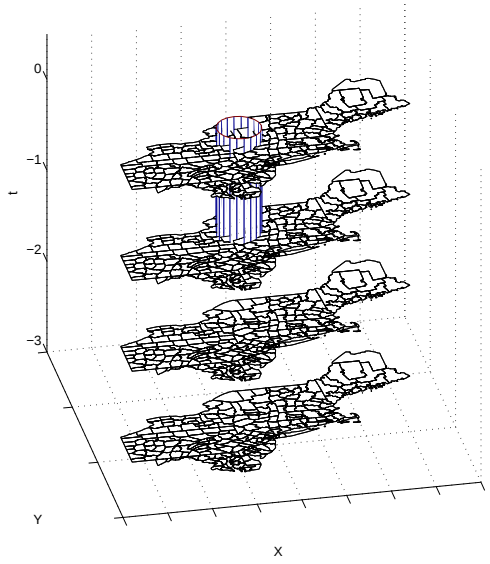
O primeiro problema é a redução do poder de detecção rápida do *cluster*. Se um risco verdadeiramente alto está presente somente no final do período, as flutuações aleatórias nos períodos iniciais quando o risco era pequeno gera uma diluição do risco no final do período. De forma geral, quanto maior o período considerado menor é o poder de detecção rápida de um *cluster* emergente. O segundo problema é a ausência de ajuste na significância devido à repetição periódica da detecção puramente espacial.

Uma possível solução seria utilizar apenas os últimos anos. Porém a determinação da quantidade de anos que será considerada pode gerar baixo poder de detecção quando a quantidade de anos for pequena e desta forma insuficiente para detectar um risco no período, de pequeno a moderado, ou quando o período for muito grande e o risco for elevado, só que em apenas em um curto período, gerando assim uma diluição do risco. A solução para isto, como já foi exposto nas seções anteriores, é utilizar uma estatística espaço-temporal (Kulldorff, 2001).

Como o interesse consiste no monitoramento de clusters emergentes, o método proposto em Kulldorff (2001) terá interesse apenas nos clusters que ainda existam no último período disponível. Desta forma, define-se $Y_1$ e $Y_2$ como sendo o tempo inicial e final, respectivamente, no qual os dados estão disponíveis, consideram-se apenas os cilindros que satisfaçam $Y_1 \leq s \leq t = Y_2$, em que $s$ e $t$ são os períodos de tempo em que o cilindro inicia e termina, respectivamente. O cilindro aqui mencionado é exatamente o mesmo relatado em Kulldorff et al. (1998).

À medida que novos casos são adicionados é possível repetir o procedimento, porém é necessário realizar um ajuste no p-valor calculado, de tal forma a permitir a comparabilidade entre os clusters. Esta correção é realizada através da maximização da verossimilhança sob $H_0$ sob todos os cilindros utilizados nas análises anteriores e na presente, ou seja, serão considerados todos os cilindro que satisfazem $Y_1 \leq s \leq t = Y_2$ e $t > Y_m$, em que $m$ é o primeiro período em que se iniciou o monitoramento. A figura 2.2 exibe um exemplo do funcionamento da estatística scan espaço-temporal prospectiva.

Figura 2.2: Exemplo da utilização da estatística scan espaço-temporal prospectiva.




O método é aplicado para os casos de incidência de câncer de tireoide no estado do Novo México no período de 1973 até 1992. Para estes dados são aplicadas a análise puramente espacial, a análise espaço-temporal sem correção na significância e a com correção, possibilitando assim comparar as diferenças ao se utilizar cada uma das abordagens.

Neste capítulo foi possível observar métodos de detecção anteriores à estatística scan de Kulldorff (1997) e que influenciaram na sua construção. Pudemos observar também, a expansão da estatística scan proposta em 1997 para o caso espaço-temporal (Kulldorff et al., 1998) e uma estatística scan que fornece uma medida que pode ser interpretada como a importância relativa da respectiva área em relação ao *cluster* (Oliveira et al., 2011).

Todos os métodos apresentados assumiram que o número de casos se distribui de acordo com um processo de Poisson, gerando assim um modelo Poisson para a detecção de clusters geográficos. Dependendo da aplicação alguns outros modelos podem ser utilizados, como o modelo Bernoulli (Kulldorff, 1997), ordinal (Jung et al., 2007), exponencial (Huang et al., 2007), normal (Kulldorff et al., 2009) e multinomial (Jung et al., 2010).

No caso especial em que se usa a distribuição de Poisson, pode ocorrer do número de regiões com casos iguais a zero ser mais alto do que o esperado, gerando uma distorção no processo de detecção do *cluster*. Uma alternativa para a resolução desta questão será apresentada no capítulo a seguir.

# 3 Distribuição Poisson Inflacionada de Zeros

É usual a suposição de que o número de casos segue uma distribuição de Poisson, como podemos constatar nos trabalhos citados no capítulo precedente. Uma das características da distribuição de Poisson é ter sua média igual à sua variância. Quando esta suposição é violada pode-se gerar um viés no cálculo da significância para o cluster detectado.

Dentre os fatores que podem gerar a desigualdade entre a média e a variância, tem-se a alta presença de valores observados iguais a 0. Ao retirar-se uma amostra de tamanho grande de uma distribuição de Poisson com média $\lambda$, espera-se observar um número de valores iguais a zero próximo de $ne^{-\lambda}$. Porém, pode ocorrer de o número observado ser bem maior. Uma forma de abordar esta situação é através do modelo de Poisson Inflacionado de Zeros (ZIP, na sigla em inglês).

## 3.1 Aplicações

Em Lambert (1992) encontra-se a modelagem por ZIP em um contexto não ligado à estatística espacial. Na situação em análise uma linha de montagem oscila entre dois estados. O primeiro deles é aquele em que todas as condições estão perfeitas. Desta forma, a presença de produtos defeituosos ocorre com uma frequência baixíssima. O outro estado é aquele em que ocorre uma má calibração no sistema, podendo-se observar um número de defeitos variável. Pode-se definir $\pi$ como sendo a probabilidade do processo estar no *estado perfeito*, linha de montagem incapaz de gerar produtos defeituosos, e o número de defeitos como sendo

distribuído de acordo com uma Poisson de parâmetro $\lambda$, número médio de defeitos no processo.

Uma segunda aplicação pode ser encontrada na descrição dos fatores que levam determinado animal a escolher a sua moradia. Dependendo da espécie e da região, a presença em determinadas áreas pode ser praticamente impossível. Agarwal et al. (2002) propõem um modelo espacial que utiliza do ZIP para lidar com o número excessivo de zeros que ocorrem na contagem de ninhos de isópodes, *Hemilepistus reaumuri*. Busca-se explicar a presença e a quantidade de ninhos em quadrados de tamanho $5 \times 5$, através de fatores como o tempo médio de disponibilidade de orvalho, porcentagem do solo coberto por arbustos e porcentagem do solo coberto por pedras.

Em um contexto de desenho experimental com medidas repetidas, encontramos uma aplicação do modelo ZIP em Hall (2000), que aplica o modelo para o número de moscas brancas, *Trialeurodes vaporariorum*, sobreviventes após a aplicação de 5 tipos de tratamentos e de 1 controle. Os tratamentos foram aplicados com um desenho experimental utilizando blocos completos aleatorizados com medições repetidas ao longo de 12 semanas. A unidade experimental considerada foi de três plantas. Desta forma, foram necessárias 54 plantas no total, para os seis tratamentos e três blocos.

O excesso de zeros em dados de contagem pode ter origem em diversos fatores. Na aplicação utilizada por Lambert (1992) o excesso de zeros é decorrência de um estado do sistema de produção em que a apresentação de defeitos no produto é praticamente impossível. Em Agarwal et al. (2002) este excesso de zeros pode ocorrer em função de determinados quadrantes apresentarem uma combinação de fatores que impossibilitam a instalação dos isópodes na área. Em Hall (2000), os tratamentos aplicados às folhas das plantas podem impossibilitar a capacidade de sobrevivência das moscas brancas.

Percebemos que em todos os exemplos de aplicação do ZIP descritos aqui, existia por trás do excesso de zeros uma situação ou combinação de fatores em que a observação de ocorrências era praticamente impossível. Este tipo de zero, que é

consequência da impossibilidade de ocorrência de outros valores, é denominado zero estrutural. Por outro lado, quando observamos um zero proveniente de uma situação em que as condições eram propícias à ocorrência de casos, denomina-se este zero como zero amostral. Geralmente, nas aplicações não se tem conhecimento suficiente para distinguir se um determinado zero é um zero amostral ou um zero estrutural.

A modelagem através do ZIP também pode ocorrer para casos de falso negativo. Em pesquisas de contagem de animais por regiões pode ocorrer de registrar-se zero, quando na verdade havia presença da espécie analisada. Outro exemplo é quando a disponibilidade de determinado equipamento, um mamógrafo por exemplo, não está disponível na região. Desta forma o zero ali encontrado não representa a ausência do câncer de mama, e sim a ausência das condições favoráveis para o diagnóstico.

Apesar do ZIP poder ser aplicado nestas duas circunstâncias, cada uma com interpretações diferentes, nos concentraremos geralmente em exemplos que envolvem o zero estrutural.

## 3.2 Especificações do Modelo ZIP

O modelo ZIP pode ser visualizado como uma mistura entre uma distribuição de Poisson e uma distribuição degenerada no ponto zero, com probabilidade de mistura igual a $\pi$. O modelo ZIP propõe que com probabilidade $\pi$ observa-se o valor zero e que com probabilidade $(1-\pi)$ observa-se uma distribuição de Poisson($\lambda$). Lambert (1992) apresenta a situação onde tanto $\pi$ quanto $\lambda$ podem estar associados a covariáveis, podendo estas covariáveis ser as mesmas para os dois parâmetros. A distribuição ZIP é apresentada a seguir:

$$\begin{aligned} P(Y=0|\pi,\lambda) &= \pi+(1-\pi)e^{-\lambda}, \\ P(Y=y|\pi,\lambda) &= (1-\pi)\frac{e^{-\lambda}\lambda^{y}}{y!},\ y>0. \end{aligned}$$

Esta distribuição pode ser visualizada como se uma proporção $\pi$ fosse retirada da distribuição de Poisson($\lambda$) e atribuída ao evento $\{y = 0\}$. Este incremento da probabilidade de ocorrência do zero busca refletir na distribuição a existência dos zeros estruturais. O valor esperado da distribuição ZIP é dado por:

$$
\begin{aligned}
E(Y|\pi,\lambda) &= 0.P(Y=0|\pi,\lambda) + \sum_{y>0} yP(Y=y|\pi,\lambda) && (3.1)\\
&= \sum_{y>0} y(1-\pi)\frac{e^{-\lambda}\lambda^{y}}{y!} && (3.2)\\
&= (1-\pi)e^{-\lambda}\sum_{y>0} y\frac{\lambda^{y}}{y!} && (3.3)\\
&= (1-\pi)e^{-\lambda}\lambda e^{\lambda} && (3.4)\\
&= (1-\pi)\lambda && (3.5)
\end{aligned}
$$

Pode-se mostrar também que $Var(Y|\pi,\lambda) = (1-\pi)\lambda(1+\pi\lambda)$. A tabela 3.1 revela os valores da $E(Y|\pi,\lambda)$ e de $Var(Y|\pi,\lambda)$ para alguns valores fixos de $\pi$ e $\lambda$.

Tabela 3.1: Cálculo da $E(Y|\pi,\lambda)$ e em parênteses da $Var(Y|\pi,\lambda)$

| $\pi\backslash\lambda$ | 1 | 10 | 50 |
|---|---|---|---|
| 0,05 | 0,95 (0,9975) | 9,5 (14,25) | 47,5 (166,25) |
| 0,25 | 0,75 (0,9375) | 7,5 (26,25) | 37,5 (506,25) |
| 0,90 | 0,1 (0,19) | 1 (10) | 5 (230) |

A tabela 3.1 nos possibilita visualizar a comparação entre a média e a variância para $\pi = \{0,05; 0,25; 0,90\}$ e $\lambda = \{1; 10; 50\}$. Possibilita assim, ver como média e variância se diferenciam de acordo com a combinação dos valores das duas variáveis. Percebemos que à medida que $\lambda$ aumenta a diferença tende a ser maior, revelando assim, que uma presença de zeros estruturais pequena já poderia trazer prejuízos para a modelagem caso se utilizasse o modelo Poisson.

Na prática o modelo ZIP terá uma vantagem maior em relação ao modelo Poisson quando valores altos de $\lambda$ encontram valores elevados de $\pi$, desde que $\pi < 1$. No caso em que $\pi = 1$ encontra-se uma distribuição degenerada no zero e portanto não há interesse em utilizar da distribuição de Poisson. Apesar do modelo ZIP ser uma generalização do modelo Poisson, apresentando resultados

iguais quando $\pi = 0$, e ajustando o excesso de zeros quando $0 < \pi < 1$, a comparação dos resultados provenientes dos dois modelos não é tão direta quanto se desejaria. Cada modelo apresenta concepções diferentes e seus resultados devem ser interpretados atentando-se a isso.

## 3.3 Estatística Scan ZIP Espacial

A causa de um excesso de zeros dentro do contexto epidemiológico pode ser explicada por fatores presentes em algumas regiões que impossibilitem a ocorrência de casos. Dentre estes fatores, podemos citar, por exemplo, campanhas de saúde pública realizadas pelas entidades responsáveis para imunizar a população. Desta forma, não se espera encontrar casos de uma determinada doença infecto-contagiosa em uma área onde toda a população esteja imunizada. Já nas outras regiões, as populações estarão sujeitas a uma probabilidade de vir a contrair a doença.

A aplicação destes programas pode ser mais ou menos bem sucedida de acordo com as características de cada uma das regiões. Enquanto uma região pode alcançar total sucesso na implementação de um programa, suas regiões vizinhas podem ter problemas maiores. Pode ocorrer assim a presença de um zero estrutural nas proximidades de, ou até mesmo inserido em, um conjunto de regiões que possui probabilidade elevada de apresentar casos.

Apesar de conseguirmos especular sobre alguns fatores que possam gerar zeros estruturais, sua identificação nas aplicações não é simples. A concentração destes fatores e suas localizações podem ser as mais diversas, podendo gerar em algumas regiões zeros estruturais enquanto em outras apenas reduz a probabilidade de incidência.

Uma forma de incorporar o zero estrutural na detecção de clusters espaciais de doenças é substituir a distribuição de Poisson pela distribuição ZIP na estatística Scan de Kulldorff (1997), ou de algum outro método de detecção de clusters que utilize a distribuição de Poisson como distribuição para o número de casos (Can-

çado et al., 2012). Desta forma seremos capazes de abordar o grande número de observações iguais a zero de forma mais adequada. No caso de um mapa dividido em regiões, podemos considerar que o número de casos na $i$-ésima região $C_i$ siga uma distribuição $ZIP(\pi; n_i\theta_i)$, isto é:

$$\begin{aligned} P(C_i = 0|\pi, n_i\theta_i) &= \pi + (1-\pi)e^{-n_i\theta_i} \\ P(C_i = c_i|\pi, n_i\theta_i) &= (1-\pi)\frac{e^{-n_i\theta_i}(n_i\theta_i)^{c_i}}{c_i!},\ c_i > 0 \end{aligned} \tag{3.6}$$

onde $\pi$ é a probabilidade associada à ocorrência de zero estrutural, $n_i$ é a população da $i$-ésima região e $\theta_i$ é a probabilidade de observar um caso na $i$-ésima região.

Portanto, a modelagem pelo ZIP nos fornece a probabilidade de ocorrência de um zero estrutural, $\pi$, e a probabilidade de ocorrência de um zero amostral $(1-\pi)e^{-n_i\theta_i}$. Para a identificação das regiões que possuem zero estrutural definiu-se a variável $\delta_i$, exibida a seguir:

$$\delta_i = \begin{cases} 1, & \text{se há zero estrutural na região i;} \\ 0, & \text{caso contrário.} \end{cases} \tag{3.7}$$

Pela definição da variável $\delta_i$, para $i = 1, \ldots, R$ temos que $P(\delta_i = 1) = \pi$. A distribuição conjunta de $C_i$ e $\delta_i$ é dada por:

$$P(C_i = c_i, \delta_i = d_i|\pi, n_i\theta_i) = \begin{cases} (1-\pi)e^{-n_i\theta_i}, & c_i = 0 \quad \text{e} \quad d_i = 0 \\ \pi, & c_i = 0 \quad \text{e} \quad d_i = 1 \\ \frac{(1-\pi)(n_i\theta_i)^{c_i}e^{-n_i\theta_i}}{c_i!}, & c_i > 0 \quad \text{e} \quad d_i = 0 \\ 0, & \text{Caso contrário .} \end{cases} \tag{3.8}$$

que pode ser escrita de forma mais sucinta como:

$$P(C_i = c_i, \delta_i = d_i|\pi, n_i\theta_i) = \begin{cases} \pi^{d_i}\left((1-\pi)\frac{e^{-n_i\theta_i}(n_i\theta)^{c_i}}{c_i!}\right)^{1-d_i}, & c_i = 0 \text{ ou } d_i = 0 \\ 0, & \text{Caso contrário.} \end{cases} \tag{3.9}$$

A verossimilhança então para as $R$ regiões sob o pressuposto de que $C_i \sim ZIP(\pi; n_i\theta_i)$, $i = 1,\ldots,R$, é dada por:

$$L(\pi,\theta;\mathbf{C},\delta) = \prod_{i=1}^{R} P(C_i = c_i, \delta_i = d_i) \tag{3.10}$$

$$= \prod_{i=1}^{R} \pi^{d_i} \left( (1-\pi) \frac{e^{-n_i\theta}(n_i\theta)^{c_i}}{c_i!} \right)^{1-d_i}, \tag{3.11}$$

onde $\mathbf{C} = (C_1,\ldots,C_R)$ e $\delta = (\delta_1,\ldots,\delta_R)$. Desta forma, a menos de uma constante, teremos:

$$L(\pi,\theta;\mathbf{C},\delta) = \pi^{\sum_{i=1}^{R} d_i} (1-\pi)^{R-\sum_{i=1}^{R} d_i} e^{-\theta\sum_{i=1}^{R} n_i(1-d_i)} \theta^{\sum_{i=1}^{R} c_i(1-d_i)} \tag{3.12}$$

Este resultado é referente ao caso em que não se distingue entre os riscos relativos das regiões, supondo desta forma que o risco $\theta$ seja o mesmo para todo o mapa, ou seja, estamos lidando com a situação sob $H_0$. No entanto, sob a hipótese de que existe um conjunto de regiões $z$ no qual se observa um risco elevado, $\theta$ deixa de ser um valor e passa a ser um vetor $\theta = (\theta_0, \theta_z)$, onde $\theta_z$ é o risco presente em um conjunto de regiões $z$ e $\theta_0$ o risco nas demais regiões. Desta forma teremos:

$$L(\pi,\theta,z;\mathbf{C},\delta) = \left[ \prod_{i\in z} P(C_i = c_i, \delta_i = d_i) \right] \times \left[ \prod_{j\notin z} P(C_j = c_j, \delta_j = d_j) \right] \tag{3.13}$$

$$= \left[ \prod_{i\in z} \pi^{d_i} \left[ (1-\pi) \frac{e^{-n_i\theta_z}(n_i\theta_z)^{c_i}}{c_i!} \right]^{(1-d_i)} \right] \times$$

$$\times \left[ \prod_{j\notin z} \pi^{d_j} \left[ (1-\pi) \frac{e^{-n_j\theta_0}(n_j\theta_0)^{c_j}}{c_j!} \right]^{(1-d_j)} \right] \tag{3.14}$$

Portanto, a menos de uma constante, temos a seguinte log-verossimilhança:

$$l(\pi,\theta,z;\mathbf{C},\delta) = \pi^{\sum_{i=1}^{R} d_i} (1-\pi)^{R-\sum_{i=1}^{R} d_i} e^{-\theta_z \sum_{i\in z} n_i(1-d_i)} \theta_z^{\sum_{i\in z} c_i(1-d_i)}$$

$$\times e^{-\theta_0 \sum_{j\notin z} n_j(1-d_j)} \theta_0^{\sum_{j\notin z} c_j(1-d_j)} \tag{3.15}$$

Derivando a log-verossimilhança encontrada sob $H_0$ e sob $H_a$, onde:

$$H_0: \quad \theta \text{ é o mesmo para toda a área, e} \tag{3.16}$$

$$H_a: \quad \theta_0 \neq \theta_z \text{ para pelo menos um conjunto de regiões } z, \tag{3.17}$$

encontramos os estimadores abaixo:

$$\text{Sob } H_0 \quad \hat{\theta}_0 = \frac{\sum_{i=1}^{R} c_i(1-d_i)}{\sum_{i=1}^{R} n_i(1-d_i)} \quad \text{e} \quad \hat{\pi} = \frac{\sum_{i=1}^{R} d_i}{R} \tag{3.18}$$

$$\text{Sob } H_a \quad \begin{array}{c} \hat{\theta}_z = \dfrac{\sum_{i\in z} c_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)} \\ \hat{\theta}_0 = \dfrac{\sum_{j\notin z} c_j(1-d_j)}{\sum_{j\notin z} n_j(1-d_j)} \end{array} \quad \text{e} \quad \hat{\pi} = \frac{\sum_{i=1}^{R} d_i}{R} \tag{3.19}$$

Com os estimadores encontrados para ambas as hipóteses, podemos calcular a razão de verossimilhança, dada por:

$$\lambda = \sup_{z\in Z} \left( \frac{\sup_{\theta_z>\theta_0} L(\pi,\theta,z;\mathbf{C},\delta)}{\sup_{\theta_z=\theta_0} L(\pi,\theta,z;\mathbf{C},\delta)} \right) \tag{3.20}$$

O $\sup_{\theta_z>\theta_0} L(\pi,\theta,z;\mathbf{C},\delta)$ é obtido substituindo os valores encontrados em (3.18) em (3.12) e o $\sup_{\theta_z=\theta_0} L(\pi,\theta,z;\mathbf{C},\delta)$ é obtido substituindo (3.19) em (3.15).

Desta forma:

$$\sup_{\theta_z>\theta_0} L(\pi,\theta,z;\mathbf{C},\delta) = \exp\left\{ -\sum_{i=1}^{R} c_i(1-d_i) \right\} \left( \frac{\sum_{i=1}^{R} c_i(1-d_i)}{\sum_{i=1}^{R} n_i(1-d_i)} \right)^{\sum_{i=1}^{R} c_i(1-d_i)} ; \tag{3.21}$$

$$\begin{aligned}\sup_{\theta_z=\theta_0} L(\pi,\theta,z;\mathbf{C},\delta) &= \exp\left\{-\sum_{i\in z} c_i(1-d_i)-\sum_{j\notin z} c_j(1-d_j)\right\} \\ &\times\left(\frac{\sum_{i\in z} c_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)}\right)^{\sum_{i\in z} c_i(1-d_i)} \\ &\times\left(\frac{\sum_{j\notin z} c_j(1-d_j)}{\sum_{j\notin z} n_j(1-d_j)}\right)^{\sum_{j\notin z} c_j(1-d_j)} . \end{aligned} \tag{3.22}$$

Temos então:

$$\lambda=\sup_{z\in Z}\frac{\left(\frac{\sum_{i\in z} c_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)}\right)^{\sum_{i\in z} c_i(1-d_i)}\left(\frac{\sum_{j\notin z} c_j(1-d_j)}{\sum_{j\notin z} n_j(1-d_j)}\right)^{\sum_{j\notin z} c_j(1-d_j)}}{\left(\frac{\sum_{i=1}^{R} c_i(1-d_i)}{\sum_{i=1}^{R} n_i(1-d_i)}\right)^{\sum_{i=1}^{R} c_i(1-d_i)}} \tag{3.23}$$

se

$$\frac{\sum_{i\in z} c_i(1-d_i)}{\sum_{i\in z} n_i(1-d_i)} > \frac{\sum_{j\notin z} c_j(1-d_j)}{\sum_{j\notin z} n_j(1-d_j)}.$$

A razão de verossimilhança encontrada através da utilização da distribuição ZIP, equação (3.23), é análoga a razão de verossimilhança encontrada em (2.27), com a diferença de incorporar no cálculo a presença de zeros estruturais. Essa incorporação ocorre através de uma ponderação na população e no número de casos para cada uma das regiões que possuem zero estrutural. Na equação (3.23) não foi utilizada a notação de $\mu_i$ como nas equações (2.27) e (2.29) no intuito de tornar clara a interação entre $\delta_i$ e $n_i$, para $i = 1,\ldots,R$.

Para visualizarmos a diferença entre a estatística Scan-ZIP e a estatística de Kulldorff (1997) na presença de zeros estruturais, basta verificar o que ocorre

ao incorporar a região com zero estrutural ao conjunto de regiões que está sendo analisado. Segundo o conceito de zero estrutural, esta região não apresenta as condições favoráveis a apresentar casos e desta forma o número esperado de casos deveria ser igual a zero. Na estatística scan de Kulldorff (1997) a inserção desta região levará a um aumento da população e proporcionalmente a uma elevação no número esperado de casos, no entanto, este aumento não é condizente com a situação de zero estrutural da região. Esta elevação do valor esperado leva a uma redução no valor da razão de verossimilhança, obscurecendo assim um possível cluster.

Para a estatística Scan-ZIP, cada região inserida é ponderada por $(1-d_i)$, onde $d_i$ assume valor 1 quando a região apresenta zero estrutural e 0 caso contrário. Desta forma a inserção de uma região com zero estrutural não levará a nenhum aumento do valor esperado de casos na região, o que condiz com a situação de zero estrutural da região.

Os valores assumidos por $d_i$,1 e 0, só ocorrem nos casos em que existe informação suficiente para determinar quais regiões apresentam zero estrutural, desta forma, $d_i$ funciona como uma variável indicadora da presença de zero estrutural. Porém, o caso mais comum é a ausência de informação suficiente para distinguir entre um zero amostral e um zero estrutural, impossibilitando a construção de tal variável indicadora. Veremos então na seção seguinte, em especial na equação (3.26), que através do algoritmo EM pode-se associar $d_i$ à probabilidade de ocorrer um zero estrutural dado o valor observado para $C_i$.

A partir da estimação de $d_i$ através do algoritmo EM pondera-se a população de acordo com a probabilidade desta região ser um zero estrutural. Assim, uma região com alta probabilidade terá uma porção pequena da sua população contabilizada no cálculo da razão de verossimilhança, enquanto que uma região de baixa probabilidade terá quase que a totalidade de sua população contabilizada.

Apresenta-se na próxima seção a descrição dos cálculos para a estimação do vetor $\delta = (\delta_i, \ldots, \delta_R)$, dos $\theta$'s para dentro e fora da região pesquisada e da probabilidade $\pi$ de ocorrência do zero estrutural utilizando o algoritmo EM.

## 3.4 Algoritmo EM

O algoritmo EM foi proposto em Dempster et al. (1977) para o cálculo da função de máxima verossimilhança para os casos de dados faltantes. Pode ser aplicado nas áreas de valores *missing*, dados truncados ou censurados, modelos de mistura finita, estimação das componentes da variância, estimação de hiper-parâmetros, atualização do peso de forma iterativa nos mínimos quadrados e na análise fatorial.

Em cada iteração do algoritmo EM, passa-se por duas etapas: etapa de expectância e etapa de maximização. As duas etapas são alternadamente executadas até que se atinja um nível de convergência previamente definido. A demonstração do cálculo de cada uma das etapas será mostrada para o caso específico que se apresenta no presente trabalho.

Informações adicionais, assim como exemplos da aplicação do algoritmo EM, podem ser encontradas em Dempster et al. (1977). Na próxima seção serão fornecidos os cálculos dos parâmetros utilizando-se do algoritmo EM seguindo os mesmos passos encontrados em Cançado et al. (2012).

### 3.4.1 Passo E: Estimar $\delta_i$ através de $E(\delta_i|C_i = 0)$

O primeiro passo do algoritmo EM consiste na estimação do vetor desconhecido a partir do seu valor esperado. Pela definição de $\delta_i$ temos que a distribuição de $(\delta_i|C_i = 0)$ é uma Bernoulli. Desta forma, temos que $E(\delta_i|C_i = 0) = P(\delta_i = 1|C_i = 0)$ e portanto, a $k$-ésima iteração será dada por:

$$\begin{aligned}
\delta_i^{(k)} &= E(\delta_i|C_i = 0) && (3.24)\\
&= P(\delta_i = 1|C_i = 0) && (3.25)\\
&= \frac{P(C_i = 0|\delta_i = 1)P(\delta_i = 1)}{P(C_i = 0|\delta_i = 1)P(\delta_i = 1) + P(C_i = 0|\delta_i = 0)P(\delta_i = 0)} && (3.26)
\end{aligned}$$

Consequentemente

$$\delta_i^{(k)} = \begin{cases} \frac{\pi^{(k)}}{\pi^{(k)}+(1-\pi^{(k)})e^{-n_i\theta^{(k)}}} & , \text{ se } c_i = 0 \\ 0 & , \text{ caso contrário,} \end{cases} \tag{3.27}$$

para $H_0$. Sob $H_a$ teremos:

$$\delta_i^{(k)} = \begin{cases} \frac{\hat{\pi}^{(k)}}{\hat{\pi}^{(k)}+(1-\hat{\pi}^{(k)})e^{-n_i\hat{\theta}_z^{(k)}}} & , \text{ se } c_i = 0 \text{ e } i \in z \\ \frac{\hat{\pi}^{(k)}}{\hat{\pi}^{(k)}+(1-\hat{\pi}^{(k)})e^{-n_i\hat{\theta}_0^{(k)}}} & , \text{ se } c_i = 0 \text{ e } i \notin z \\ 0 & , \text{ se } c_i > 0. \end{cases} \tag{3.28}$$

Obtém-se assim a $k$-ésima estimação para a probabilidade da região $i$ conter um zero estrutural sob cada uma das hipóteses. Com os vetores obtidos, podemos calcular a verossimilhança para os dados completos em ambas as hipóteses.

Sob $H_0$:

$$L_C(\theta,\pi|\delta^{(k)}) = \prod_{i=1}^{R} \pi^{\delta_i^{(k)}} \left((1-\pi)\frac{e^{-n_i\theta}(n_i\theta)^{c_i}}{c_i!}\right)^{1-\delta_i^{(k)}} \tag{3.29}$$

Sob $H_a$:

$$\begin{aligned} L_C(\theta_z,\theta_0,\pi|\delta^{(k)}) = & \left[\prod_{i\in z} \pi^{\delta_i^{(k)}} \left[(1-\pi)\frac{e^{-n_i\theta_z}(n_i\theta_z)^{c_i}}{c_i!}\right]^{(1-\delta_i^{(k)})}\right] \times \\ & \times \left[\prod_{j\notin z} \pi^{\delta_j^{(k)}} \left[(1-\pi)\frac{e^{-n_j\theta_0}(n_j\theta_0)^{c_j}}{c_j!}\right]^{(1-\delta_j^{(k)})}\right] \end{aligned} \tag{3.30}$$

### 3.4.2 Passo M: Maximizar $L_C(\theta_z,\theta_0,\pi|\delta^{(k)})$

Para cada uma das máximas verossimilhanças exibidas nas equações (3.29) e (3.30) podemos calcular os estimadores que maximizem cada uma das funções, que serão dados por:

Sob $H_0$:

$$\hat{\theta}_0^{(k+1)} = \frac{\sum_{i=1}^{R} c_i(1-\delta_i^{(k)})}{\sum_{i=1}^{R} n_i(1-\delta_i^{(k)})} \tag{3.31}$$

$$\hat{\pi}^{(k+1)} = \frac{\sum_{i=1}^{n} \delta_i^{(k)}}{R} \tag{3.32}$$

Sob $H_a$

$$\hat{\theta}_z^{(k+1)} = \frac{\sum_{i\in z} c_i(1-\delta_i^{(k)})}{\sum_{i\in z} n_i(1-\delta_i^{(k)})} \tag{3.33}$$

$$\hat{\theta}_0^{(k+1)} = \frac{\sum_{j\notin z} c_j(1-\delta_j^{(k)})}{\sum_{j\notin z} n_j(1-\delta_j^{(k)})} \tag{3.34}$$

$$\hat{\pi}^{(k+1)} = \frac{\sum_{i=1}^{n} \delta_i^{(k)}}{R} \tag{3.35}$$

Em posse destas estimativas, repete-se o passo E, calculando-se $\delta_i^{(k+1)}$, e depois recalcula-se o passo M e assim sucessivamente até que o critério estipulado para a convergência seja atingido. Em posse do valor de $\delta_i^{(k)}$ que satisfez os critérios estipulados para cada uma das hipóteses, substitui-se esse valor na equação (3.23) onde se encontra $d_i$. Desta forma, obtém-se o valor para a razão de verossimilhança para os casos em que o vetor $\delta = (\delta_i, \ldots, \delta_R)$ é desconhecido.

Chega-se então a uma estatística que engloba em seus cálculos a presença de zeros estruturais tanto nas situações em que se conhecem as suas localizações, quanto quando estas são completamente alheias ao pesquisador.

### 3.4.3 Inicialização do Algoritmo EM

Para a inicialização dos valores $\delta_i$, para $i = 1,\ldots,R$, utiliza-se a seguinte equação:

$$\delta_i^0 \;=\; \begin{cases} 0{,}5 & , \text{ se } c_i = 0 \\ 0 & , \text{ se } c_i > 0 \end{cases} . \tag{3.36}$$

Desta forma, estima-se valores para $\delta_i$ tanto para $i \in z$ quanto para a situação oposta.

Na aplicação do algoritmo EM na estimação dos $\delta_i$ no presente trabalho percebeu-se uma rápida convergência com 4 iterações e decidiu-se então pela fixação de 9 iterações para a estimação dos parâmetros.

## 3.5 Estatística Scan ZIP Espaço-Tempo (Scan ZIPET)

Como mencionado na seção 2.5, a agregação dos casos para um determinado período de tempo pode trazer distorções no poder da estatística. Pode tanto diluir quanto acentuar um determinado nível de agrupamento, modificando assim sua aparência real. Por isso, possibilita-se à estatística scan realizar uma varredura não só na dimensão espacial, mas também na temporal.

Nesta seção apresentaremos a versão das expressões mostradas na seção anterior ao se incluir a dimensão temporal. Iniciamos pela reescrita das hipóteses:

$H_0$ : $\theta$ é o mesmo para toda a área em todos os períodos estudados, e

$H_a$ : $\theta_0 \neq \theta_z$ para pelo menos um conjunto de regiões $z$ em um conjunto específico de períodos de tempo.

Daqui em diante será utilizado $z$ para representar um conjunto de regiões em um conjunto de períodos de tempo. Desta forma, $z$ representa uma limitação horizontal e vertical se pensarmos no cilindro representado na figura 2.1. O

conjunto de períodos contidos em $z$ engloba necessariamente períodos de tempo contíguos.

A analogia do caso espacial para o caso espaço-temporal surge quando pensamos que as $R$ regiões disponíveis nos $T$ períodos de tempo compõem um mapa com $R \times T$ regiões. Portanto, as contas nos levariam aos mesmos estimadores no espaço-tempo que aqueles encontrados no caso espacial se assim procedêssemos.

Sob $H_0$

$$\hat{\theta}_0 = \frac{\sum_{t=1}^{T}\sum_{i=1}^{R} c_{it}(1-d_{it})}{\sum_{t=1}^{T}\sum_{i=1}^{R} n_{it}(1-d_{it})} \tag{3.37}$$

$$\hat{\pi} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{R} d_{it}}{RT} \tag{3.38}$$

Sob $H_a$

$$\hat{\theta}_z = \frac{\sum_{(i,t)\in z}\sum c_{it}(1-d_{it})}{\sum_{(i,t)\in z}\sum n_{it}(1-d_{it})} \tag{3.39}$$

$$\hat{\theta}_0 = \frac{\sum_{(i,t)\notin z}\sum c_{it}(1-d_{it})}{\sum_{(i,t)\notin z}\sum n_{it}(1-d_{it})} \tag{3.40}$$

$$\hat{\pi} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{R} d_{it}}{RT} \tag{3.41}$$

onde $c_{it}$ representa o número de casos na $i$-ésima região no período $t$, $n_{it}$ é a população da $i$-ésima região no período $t$ e $d_{it}$ é o fator de ponderação na $i$-ésima região no período $t$ e suas características serão discutidas ulteriormente.

Substituindo os estimadores em suas respectivas máximo-verossimilhanças e calculando a razão entre elas chegamos à equação abaixo:

$$\lambda = \sup_{z \in Z} \frac{\left( \frac{\sum_{(i,t) \in z} c_{it}(1-d_{it})}{\sum_{(i,t) \in z} n_{it}(1-d_{it})} \right)^{\sum_{(i,t) \in z} c_{it}(1-d_{it})} \left( \frac{\sum_{(i,t) \notin z} c_{it}(1-d_{it})}{\sum_{(i,t) \notin z} n_{it}(1-d_{it})} \right)^{\sum_{(i,t) \notin z} c_{it}(1-d_{it})}}{\left( \frac{\sum_{t=1}^{T} \sum_{i=1}^{R} c_{it}(1-d_{it})}{\sum_{t=1}^{T} \sum_{i=1}^{R} n_{it}(1-d_{it})} \right)^{\sum_{t=1}^{T} \sum_{i=1}^{R} c_{it}(1-d_{it})}} \tag{3.42}$$

se

$$\frac{\sum_{(i,t) \in z} c_{it}(1-d_{it})}{\sum_{(i,t) \in z} n_{it}(1-d_{it})} > \frac{\sum_{(i,t) \notin z} c_{it}(1-d_{it})}{\sum_{(i,t) \notin z} n_{it}(1-d_{it})}.$$

Ressaltamos que a condição colocada acima é consequência do objetivo de se encontrar clusters onde o número de casos observados supere o número de casos esperados.

A equação (3.42) é bastante próxima da equação encontrada em (3.23), só que ao invés de somarem-se todos os casos presentes em um conjunto de períodos de tempo, acrescenta-se uma janela temporal e permite-se que esta varie ao longo do tempo agregando as regiões em diferentes conjuntos de períodos de tempo contíguos.

A equação (3.42) apresenta o cálculo quando as regiões com zero estrutural são conhecidas. Na situação em que isto se mostra impossível, aplica-se o estimador EM para obtenção do vetor $\delta$. As equações provenientes do algoritmo EM para o caso espaço-temporal são exibidas abaixo:

Sob $H_0$:

$$\hat{\theta}_0^{(k)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{R} c_{it}(1-\delta_{it}^{(k)})}{\sum_{t=1}^{T}\sum_{i=1}^{R} n_{it}(1-\delta_{it}^{(k)})} \tag{3.43}$$

$$\hat{\pi}^{(k)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{R} \delta_{it}^{(k)}}{RT} \tag{3.44}$$

$$\delta_{it}^{(k)} = \begin{cases} \frac{\pi^{(k)}}{\pi^{(k)}+(1-\pi^{(k)})e^{-n_{it}\theta^{(k)}}} & , \text{ se candidato a zero estrutural} \\ 0 & , \text{ caso contrário.} \end{cases} \tag{3.45}$$

Sob $H_a$

$$\hat{\theta}_z^{(k)} = \frac{\sum_{(i,t)\in z}\sum c_{it}(1-\delta_{it}^{(k)})}{\sum_{(i,t)\in z}\sum n_{it}(1-\delta_{it}^{(k)})} \tag{3.46}$$

$$\hat{\theta}_0^{(k)} = \frac{\sum_{(i,t)\notin z}\sum c_{it}(1-\delta_{it}^{(k)})}{\sum_{(i,t)\notin z}\sum n_{it}(1-\delta_{it}^{(k)})} \tag{3.47}$$

$$\hat{\pi}^{(k)} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{R} \delta_{it}^{(k)}}{RT} \tag{3.48}$$

$$\delta_{it}^{(k)} = \begin{cases} \frac{\hat{\pi}^{(k)}}{\hat{\pi}^{(k)}+(1-\hat{\pi}^{(k)})e^{-n_{it}\hat{\theta}_z^{(k)}}}, & \text{se candidato a zero estrutural e} \\ & (i,t)\in z, \\ \frac{\hat{\pi}^{(k)}}{\hat{\pi}^{(k)}+(1-\hat{\pi}^{(k)})e^{-n_{it}\hat{\theta}_0^{(k)}}}, & \text{se candidato a zero estrutural e} \\ & (i,t)\notin z, \\ 0, & \text{se } c_{it}>0. \end{cases} \tag{3.49}$$

Na seção seguinte abordaremos a questão de quais fatores podem ser levados em consideração para que uma região seja candidata a ter o seu zero considerado como um zero estrutural. Estas condições podem variar bastante. Em função disso, nas fórmulas apresentadas acima optou-se por apresentar uma con-

dição genérica e na próxima seção aprofundar nas condições que poderiam ser ali colocadas.

## 3.6 Definição de Zero-Estrutural Espaço-Temporal

Na seção 3.3 apresentou-se a estatística Scan-ZIP espacial. Na estatística Scan-ZIP, e em aplicações puramente espaciais de forma geral, as informações temporais são somadas e obtém-se um vetor de casos $C$ que representa o número de casos no período considerado para cada uma das $R$ regiões. Neste exemplo o zero estrutural surge naturalmente e a região que tiver sua soma no período igual a zero terá sua probabilidade de ser um zero estrutural calculada através do algoritmo EM.

A passagem do zero estrutural do caso puramente espacial para o caso espaço-temporal não é tão direta. O primeiro fator a ser considerado é que a população será contabilizada período a período, desta forma é necessário que $\delta$ seja calculado para todas as regiões em cada um dos períodos de tempo considerados, ou seja, calcula-se $\delta_{it}$, para $i = 1,\ldots,R$ e $t = 1,\ldots,T$.

Uma primeira opção é 'herdar' a definição proveniente do caso puramente espacial. Nesta circunstância são consideradas regiões com possível zero estrutural aquelas que apresentaram zero casos em todos os períodos de tempo sob análise. Esta definição inicial parece refletir bem o conceito de zero estrutural: regiões que apresentaram zero em todos os períodos de tempo são fortes candidatas a terem condições estruturais que as impossibilitem de apresentar casos. Desta forma, para cada uma destas regiões será calculada a sua probabilidade de conter um zero estrutural. Uma consequência desta definição é que regiões que apresentaram zero em alguns períodos de tempo e casos em outros períodos têm os seus zeros classificados automaticamente como zeros amostrais.

Observando esta consequência da primeira definição, percebemos que não existe flexibilidade quando se supõe que a apresentação de zero estrutural pode variar ao longo do tempo para as regiões. Os fatores responsáveis pela inibição da

ocorrência de casos nem sempre estiveram presentes nas regiões. Suponhamos, por exemplo, que uma determinada região apresentava casos até um determinado período $x$ e, paralelamente, atuava para estabelecer os mecanismos de completa supressão da doença. A partir do momento $x+1$ a supressão é completa e não mais se encontram as condições favoráveis para a ocorrência de casos, apresentando então valor zero nos anos maiores ou iguais a $x+1$.

Utilizando-se da primeira definição e escolhendo como período inicial da aplicação um valor inferior a $x$, a região do exemplo estaria fora daquelas que terão $\delta$ estimado, quando na verdade apresenta zero estrutural em um subconjunto do período de tempo considerado. Nesta situação então, estaria sendo calculado o valor esperado de casos para esta região, quando na verdade não existe possibilidade de ocorrência. Desta forma, a detecção do cluster estaria sendo prejudicada.

A situação inversa, em que uma região pode deixar de apresentar zeros estruturais, também pode ser imaginada. Os níveis de fatores alcançados para a apresentação de zero estrutural podem ter sido perdidos, ou então um novo nível é necessário para isto. Um exemplo surge quando pensamos na ocorrência de uma mutação em determinado vírus. Assim, imunizações podem não ser tão eficazes quanto para a versão inicial do vírus. Os problemas da utilização da primeira definição se repetem como no outro caso em que se conquistou o patamar de não apresentação de zeros estruturais.

Podemos ver então que esta definição é rígida em relação ao tempo e quanto maior a quantidade de períodos de tempo maior a possibilidade de ocorrerem as situações descritas acima. Uma alternativa é a flexibilização completa da definição de zero estrutural. Esta definição também apresenta semelhança com o caso espacial. Lá todos os zeros tinham sua probabilidade de ser um zero estrutural calculada, o que não ocorre na definição de zero estrutural espaço-temporal apresentada acima.

Esta nova definição estabelece então que cada zero apresentado por uma região ao longo do tempo pode ser um zero estrutural. Desta forma, todos os zeros encontrados terão sua probabilidade calculada. Esta aplicação desconsidera

o histórico de apresentação de casos da região. Desta forma, uma região que sempre apresenta casos e por fatores diversos em um determinado período não apresentou casos terá a contribuição da sua população reduzida neste período para o cálculo da razão de verossimilhança. É improvável que uma região atinja e perca o estado de zero estrutural em apenas um período de tempo, distanciando assim esta definição do conceito de zero estrutural.

Em aplicações onde existe apresentação de zeros de forma intermitente e raramente consecutiva, imagina-se que haverá grande diferença entre as duas definições de zero estrutural, sendo que se espera que a primeira definição se aproxime mais do resultado que seria obtido utilizando-se a estatística scan de Kulldorff (1997).

Percebemos que é difícil estabelecer uma definição de zero estrutural sem conhecer muito bem a estrutura de zero dos dados. Outras definições podem ser imaginadas, como apresentação consecutiva de zero em 2 períodos de tempo, ou 3 períodos. Pode-se pensar também em mecanismos que busquem definir se a região apresenta zero estrutural segundo algumas covariáveis.

No presente trabalho nos concentraremos em verificar o desempenho da estatística Scan ZIPET nos dois casos extremos. O caso em que se calcula a probabilidade de zero estrutural apenas para as regiões com zero em todos os períodos, chamaremos de zero estrutural rígido, e em contrapartida a outro caso será denominado zero estrutural flexível.

# 4 Simulação numérica e análise de desempenho

Os mais diversos métodos têm sido propostos para atuar nas mais diversas situações em que se pretende obter uma resposta sobre a existência de um cluster espacial. O Scan ZIPET, desenvolvido nesta dissertação, é um método de detecção de grupos no espaço-tempo, que esperamos venha a ser bastante utilizado na detecção de cluster espaço-temporais.

Para avaliar o desempenho da estatística scan zip espaço temporal, abreviada como Scan ZIPET, utilizaremos alguns cenários com diferentes características e compararemos o seu desempenho com o da estatística scan espaço tempo de Kulldorff et al. (1998), segundo três critérios:

- **Poder**: Calculando-se a proporção de vezes em que $H_0$ foi rejeitada ao longo das simulações;
- **Sensibilidade**: Através da média (ao longo das simulações) da proporção da população da interseção entre o cluster detectado e o cluster real em relação à população do cluster real;
- **Valor preditivo positivo**: Obtendo a média (ao longo das simulações) da proporção da população da interseção entre o cluster detectado e o cluster real em relação à população do cluster detectado; isso mede o quanto do cluster detectado faz parte do cluster real.

Vale ressaltar que, os cenários gerados, contêm regiões com zeros estruturais. A possibilidade dos dados conterem zeros estruturais não foi prevista no

desenvolvimento da estatística scan espaço tempo de Kulldorff et al. (1998). O presente trabalho se ateve em resolver tal problema como uma extensão do artigo de Cançado et al. (2012) na detecção de *clusters* incluindo-se a dimensão espacial.

## 4.1 Cenários

A construção dos cenários artificiais foi realizada com base em uma malha de 203 hexágonos, onde cada hexágono representa uma região. A população foi distribuída uniformemente no espaço, tendo cada hexágono uma população igual a 1000 indivíduos. Ao longo do tempo as populações foram mantidas fixas, de forma que todas as regiões têem a mesma população do primeiro ao último período de tempo, gerando, assim, uma população total de 203.000 indivíduos em um período de tempo e de 2.030.000 nos 10 períodos considerados. A soma das populações das regiões que possuem zero estrutural é igual a 150.000 em todos os períodos de tempo. O número de casos foi fixado como sendo igual a 0,025% da População total (tempo e espaço) das regiões que não possuem zero estrutural, totalizando 470 casos que foram distribuídos para os hexágonos ao longo dos períodos de tempo. Das 203 regiões, 15 foram selecionadas para terem zero estrutural em todo o período de tempo.

O primeiro cenário, denominado cenário A, contém um cluster com formato circular com diâmetro de 5 regiões, tendo uma diagonal, da direita superior para a esquerda inferior, de zeros estruturais. A presença desta diagonal funciona como um divisor do cluster, separando cada lado através de uma linha de regiões com número de casos igual a zero. Esta estrutura pode ser visualizada na figura 4.1.

O cenário B possui as mesmas características de tamanho e formato do cenário A, diferindo, no entanto, em relação à distribuição dos zeros estruturais. A primeira característica é que no cenário B possui um zero estrutural a menos, totalizando assim 4 regiões com esta característica. O segundo fator é que os zeros estruturais não estão distribuídos com algum padrão bem definido, estando

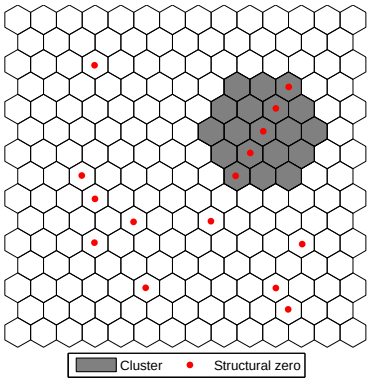



Figura 4.1: Localização do cluster no cenário A.

3 destes próximos do núcleo e um deles na extremidade, como exibido na figura 4.2.

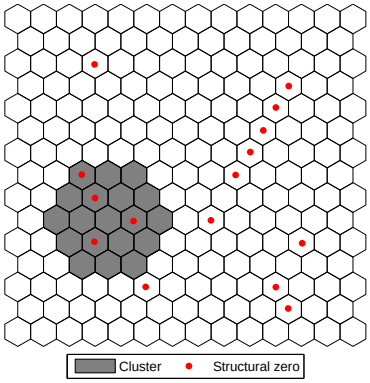



Figura 4.2: Localização do cluster no cenário B.

O cenário C continua tendo o formato circular, porém, agora, com um tamanho reduzido de 3 regiões de diâmetro. Acerca dos zeros estruturais e suas posições, este cenário possui apenas um e está localizado na região central do cluster. A imagem que representa este cenário é exibida na figura 4.3.

O cenário D não possui formato circular, sendo sua forma aproximadamente a de um L deitado. Possui 3 zeros estruturais e estes não apresentam uma distribuição espacial com um formato bem definido, estando 2 deles na parte inferior e próximos um do outro e o terceiro na parte superior, como exibido na figura 4.1.

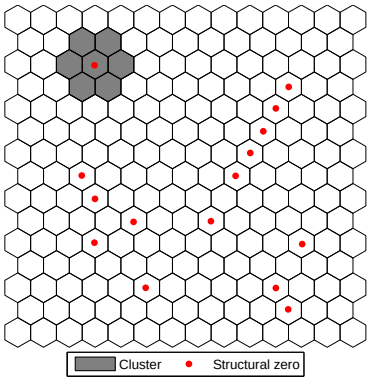



Figura 4.3: Localização do cluster no cenário C.

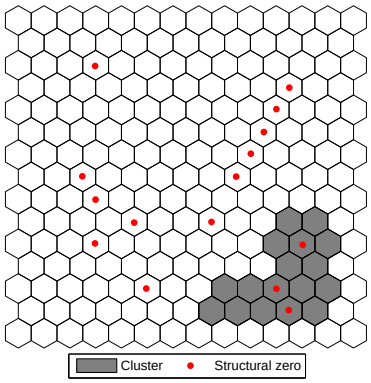



Figura 4.4: Localização do cluster no cenário D.

Estes 4 cenários buscam combinar fatores como tamanho, formato e distribuição espacial dos zeros estruturais de tal forma a verificar a capacidade dos algoritmos sob determinadas combinações. Pode-se observar que a localização dos zeros estruturais não foi alterada ao longo dos cenários, sendo suas posições preservadas de tal forma que a distribuição sob a hipótese nula seja comum a todos os cenários.

Os cenários que foram descritos trouxeram somente a estrutura espacial. A estrutura espaço-temporal do cluster ocorre quando se utiliza o caso puramente espacial de cada cenário nos períodos de tempo 3, 4 e 5. Para os demais períodos não existe nenhuma região definida com probabilidade maior de concentrar casos.

O cluster espaço-temporal surge quando o risco relativo supera 1 ao se comparar a probabilidade dentro do cluster espaço-temporal e fora do cluster espaço-temporal. A construção do risco relativo será apresentada na próxima seção.

Foram utilizados 10 períodos de tempo, no presente estudo não se procurou identificar as consequências na detecção em função do tamanho do intervalo de tempo em que o cluster está presente.

## 4.2 Simulação

Seguindo o mesmo método encontrado em Kulldorff et al. (2003) para simular o número de casos em cada região, para cada um dos cenários estabeleceu-se um risco relativo de tal forma a termos um $\alpha$ de $0,05$ e um poder de $0,99$ ao utilizar-se um teste binomial padrão. Sob $H_0$ e condicionando-se ao número total de casos $C$, o número de casos tem distribuição Binomial com média $m_0 = Cn/N$ e variância $v_0 = Cn(N-n)/N^2$. Utilizando a aproximação pela normal padrão para a distribuição binomial, o valor crítico do número de casos $k$ necessário para rejeitar a hipótese nula em um teste unilateral é dado por um $k$ tal que $(k-m_0)/\sqrt{v_0} = 1,645$. Sob a hipótese alternativa, com um risco relativo $r$, o número de casos nas regiões pertencentes ao cluster tem distribuição Binomial com média $m_A = Cnr/(N-n+nr)$ e variância $v_A = Cnr(N-n)/(N-n+nr)^2$. Aproximando-se pela normal novamente, seleciona-se o risco relativo tal que $(k-m_A)/\sqrt{v_A} = 2,326$.

O método acima nos fornece então o valor $r$ do risco relativo sob o qual a população do cluster estará exposta. A tabela 4.1 fornece a população presente no cluster e o risco relativo. Como na simulação as regiões com zeros estruturais não podem receber casos, contabiliza-las no momento de calcular o risco relativo levaria a uma diluição do risco, uma vez que indivíduos nestas regiões não podem se tornar casos. Desta forma, dos 203.000 indivíduos por ano, apenas 188.000 serão levados em consideração nos cálculos. Utilizaremos $z$ para representar o conjunto de regiões que formam o cluster e $E$ para representar o conjunto de regiões que possuem zero estrutural.

A equação

$$n = \sum\sum_{(i,t) \in z \ \& \ (i,t) \notin E} n_{(i,t)} \tag{4.1}$$

nos fornece o total de pessoas que estão em todas as regiões e em todos os períodos de tempo que pertencem ao cluster e que não pertencem ao conjunto de regiões que possuem zero estrutural. Esse valor representa o número de pessoas inseridas dentro do cilindro definido pelo cluster $z$, desconsiderando as pessoas presentes nas regiões que possuem zero estrutural.

A população total será dada por

$$N = \sum\sum_{(i,t) \notin E} n_{(i,t)}, \tag{4.2}$$

que fornece então a população em todas as regiões e em todos os períodos que não pertencem ao conjunto de zeros estruturais. Com $n$, $N$ e $C$ que é igual a 470 é possível calcular $r$ com o método descrito no primeiro parágrafo da presente seção.

Tabela 4.1: População e risco relativos dos clusters

| | Cenário A | Cenário B | Cenário C | Cenário D |
|---|---|---|---|---|
| Risco relativo | 2,733781 | 2,661405 | 4,007512 | 2,733781 |
| População no cluster | 42000 | 45000 | 18000 | 42000 |

A tabela 4.1 confirma o esperado, quanto menor a população maior o riso relativo e quanto maior a população menor será o risco relativo.

## 4.3 Resultados

Os resultados exibidos a seguir buscam sintetizar o desempenho dos algoritmos Scan-ET e Scan ZIPET na presença de zeros estruturais. O scan zip espaço-tempo funciona tanto na circunstância em que é conhecida a posição dos zeros

estruturais, quanto quando, dentre o grande número de regiões com zeros estruturais, não se sabe a que grupo cada um deles pertence.

Tabela 4.2: Comparação entre os dois métodos em termos de Poder

| Método | Cenário A | Cenário B | Cenário C | Cenário D |
|---|---|---|---|---|
| Scan ZIPET | 0,594 | 0,609 | 0,58 | 0,477 |
| Scan Poisson-ET | 0,399 | 0,456 | 0,55 | 0,40 |

Uma rápida análise dos resultados indica uma deterioração no desempenho do algoritmo Scan Poisson-ET na presença de zeros estruturais, enquanto o algoritmo proposto apresenta um desempenho sistematicamente e significativamente melhor.

A tabela 4.2 exibe valores próximos para o Scan ZIPET nos cenários A,B e C. Os clusters destes três cenários foram construídos com formato circular, diferindo entre si em relação ao número de zeros estruturais, distribuição espacial destes e tamanho do cluster. Este resultado parece revelar que o poder da estatística Scan ZIPET é pouco influenciado por estes fatores. Este resultado já era esperado, pois de acordo com a construção da equação 3.23, as regiões com número de casos iguais a zero têm sua contribuição no denominador reduzida, evitando, assim, uma penalização no cálculo da razão de verossimilhança por incluir uma região com zero casos.

Ainda observando a tabela 4.2, só que agora avaliando o desempenho da estatística Scan Poisson-ET, encontramos uma melhora no poder à medida que o número de zeros estruturais decresce. Isto reforça o argumento do impacto negativo que o zero estrutural tem sobre o cálculo da razão de verossimilhança, inflando o denominador, quando a área investigada engloba uma ou mais regiões com zero estrutural.

A inserção de um cluster com formato não circular gera um decrescimento no poder para ambas as estatísticas, mantendo-se o poder da estatística Scan ZIPET superior ao da estatística Scan Poisson-ET.

A tabela 4.3 vem nos revelar a média da proporção da intersecção populacional entre o cluster detectado e o cluster real em relação ao cluster real. Obser-

Tabela 4.3: Comparação entre os dois métodos em termos de Sensibilidade

| Método | Cenário A | Cenário B | Cenário C | Cenário D |
|---|---|---|---|---|
| Scan ZIPET | 0,63 | 0,644 | 0,686 | 0,502 |
| Scan Poisson-ET | 0,40 | 0,475 | 0,637 | 0,36 |

Tabela 4.4: Comparação entre os dois métodos em termos de PPV

| Método | Cenário A | Cenário B | Cenário C | Cenário D |
|---|---|---|---|---|
| Scan ZIPET | 0,60 | 0,604 | 0,577 | 0,437 |
| Scan Poisson-ET | 0,46 | 0,521 | 0,537 | 0,42 |

vamos que o Scan ZIPET apresenta não só uma melhor capacidade de rejeitar a hipótese nula, como também o cluster detectado por este método é mais similar ao cluster real do que o outro método. Novamente, percebemos uma melhoria no Scan Poisson-ET quando comparamos a sua sensibilidade entre os cenários A e B.

A sensibilidade da estatística Scan ZIPET manteve sua média superior a 0,60 nos cenários de clusters circulares, chegando a atingir 0,686 no cenário C. É interessante destacar o baixo valor para a sensibilidade da estatística Scan Poisson-ET para o cenário D, revelando, assim, que a ocorrência conjunta de um cluster de formato não circular com a presença de zeros estruturais traz um grande prejuízo para a detecção das regiões pertencentes ao cluster.

Ao compararmos sensibilidade e ppv dos dois métodos, chegamos a conclusão de que os clusters detectados pelo Scan ZIPET são maiores do que os detectados pelo Scan Poisson-ET, uma vez que a sensibilidade supera o vpp na primeira estatística e o inverso na segunda, com exceção no cenário C para a estatística Scan Poisson-ET.

O vpp do Scan ZIPET é superior ao do Scan-ET em cenários com a presença de zeros estruturais.

# 5 Aplicação em Dados Reais

Para melhor visualização dos potenciais da estatística Scan ZIPET e sua comparação com a estatística Scan-ET, selecionamos uma base de dados para verificarmos o desempenho de ambas as estatísticas.

Para realização desta comparação buscou-se doenças que possuem baixa incidência. Dentre as diversas doenças que possuem esta característica, optou-se pela Tuberculose. Os dados foram obtidos através do site de saúde pública para o estado norte americano da Geórgia (http://oasis.state.ga.us/), selecionando casos de morte por tuberculose para os anos de 1998 até 2002.

## 5.1 Aplicação para Mortes por Tuberculose

O estado da Geórgia é localizado na região sudeste dos Estados Unidos, possui 159 condados e todos estes serão utilizados na presente análise. Sua população passou de 6.478.216 habitantes em 1990 para 8.186.453 habitantes em 2000.

A tuberculose é uma doença infecto contagiosa transmitida pelo *Mycobacterium tuberculosis*, também conhecido como bacilo de Koch. Apesar de afetar principalmente os pulmões, pode ocorrer em outros órgãos do corpo, como ossos, rins e meninges.

A transmissão da doença ocorre de pessoa para pessoa através de gotículas de saliva eliminadas ao se falar, espirrar ou tossir. Desta forma, ambientes com pouca ventilação e com grande concentração de pessoas favorecem a transmissão da bactéria.

Um dos principais sintomas é a de tosse seca por mais de duas semanas, que pode evoluir para uma tosse com secreção e, em casos mais avançados, uma tosse com pus e sangue. Outros sintomas são: febre, sudorese, cansaço excessivo, falta de apetite e emagrecimento acentuado.

O tratamento ocorre ao longo de seis meses através da utilização de três drogas, pirazinamida, isoniazida e rifamicina. Não pode haver abandono do tratamento, que deve ser seguido à risca.

### 5.1.1 Descrição dos Dados

Os casos de morte por tuberculose para o estado da Geórgia estão disponíveis para os anos de 1994 até 2008. Em consequência da grande quantidade de iterações necessárias ao se utilizar todo o período de tempo, optou-se por selecionar um subconjunto deste. Dentre os subconjuntos de tamanho 5, escolheu-se o período de 1998 até 2002. Buscando-se uma taxa de mortalidade intermediária, acima de 2 e abaixo de 5 casos por milhão, chegou-se ao intervalo de tempo mencionado. A taxa de mortalidade média dentro deste período é de 3,82 casos por milhão. As taxas de mortalidade ano a ano podem ser visualizadas na tabela 5.1.

Tabela 5.1: Taxa de mortalidade por 1.000.000 de habitantes de 1994 até 2008.

| ***1994*** | ***1995*** | ***1996*** | ***1997*** | ***1998*** |
|---|---|---|---|---|
| 6,287 | 5,868 | 5,466 | 5,335 | 4,578 |

| ***1999*** | ***2000*** | ***2001*** | ***2002*** | ***2003*** |
|---|---|---|---|---|
| 4,350 | 3,909 | 3,450 | 2,920 | 1,727 |

| ***2004*** | ***2005*** | ***2006*** | ***2007*** | ***2008*** |
|---|---|---|---|---|
| 3,511 | 1,653 | 2,136 | 1,676 | 1,445 |

Para uma comparação inicial entre as mortalidades dos condados no período selecionado utilizou-se a média das taxas de mortalidade, cuja fórmula é dada por:

$$\frac{\sum_{1998 \leq t \leq 2002} \frac{c_{it}}{n_{it}}}{5} \tag{5.1}$$

Aplicando a fórmula 5.1 para todos os condados e selecionando-se os 5 maiores valores, encontramos os condados de Terrell, Calhoun, Marion, Crisp, Atkinson. Todos estes condados possuem uma população mais baixa, o que justifica os altos valores para as médias das taxas de mortalidade. Estas informações podem ser observadas na tabela 5.2.

Tabela 5.2: Informações sobre os condados que apresentaram as maiores médias de taxas de mortalidade

| Condado | Número Total de Casos | Valor Esperado de Casos | População Média | Média da Taxa de Mortalidade |
|---|---|---|---|---|
| Terrell | 2 | 0,209 | 10929 | 36,516 |
| Calhoun | 1 | 0,120 | 6254 | 31,274 |
| Marion | 1 | 0,136 | 7100 | 28,944 |
| Crisp | 3 | 0,420 | 21963 | 27,414 |
| Atkinson | 1 | 0,145 | 7583 | 26,285 |

Uma seleção diferente consiste em observar os condados que possuem maior número de casos de morte por tuberculose. Seleção que pode ser observada na tabela 5.3.

Tabela 5.3: Informações sobre os cinco condados com maior número de mortes por tuberculose.

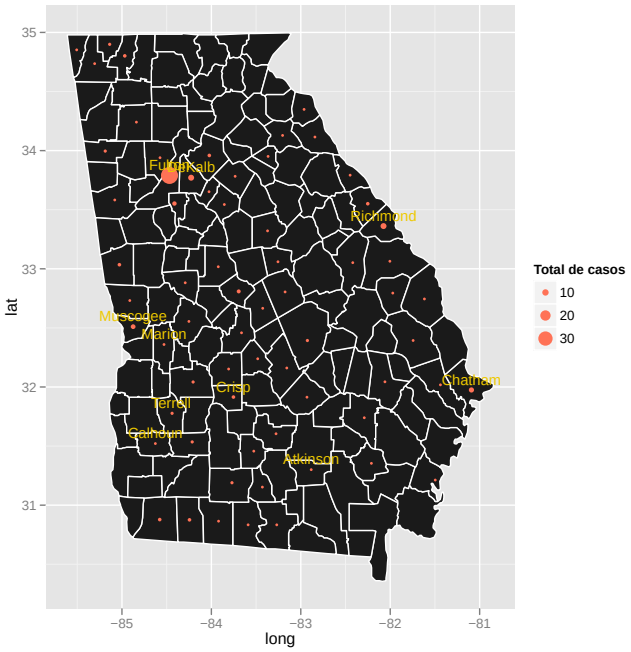
| Condado | Número Total de Casos | Valor Esperado de Casos | População Média | Média da Taxa de Mortalidade |
|---|---|---|---|---|
| Fulton | 37 | 15.542 | 812966 | 9.118 |
| Richmond | 9 | 3.806 | 199099 | 9.042 |
| DeKalb | 9 | 12.687 | 663617 | 2.719 |
| Chatham | 7 | 4.438 | 232146 | 6.027 |
| Muscogee | 6 | 3.556 | 185999 | 6.456 |

Se por um lado a tabela 5.2 apresenta baixos valores populacionais, a tabela 5.3 apresenta regiões com valor populacional elevado. Percebemos que com exceção do condado de DeKalb, todos os outros condados, de ambas as tabelas, exibiram número de casos superior ao valor esperado.

A figura 5.1 possibilita a visão do número de casos por condado, apresentando o nome dos condados que já foram mencionados até o momento. O tamanho do círculo presente em cada condado reflete o número de casos. Quanto

maior o círculo maior a quantidade de casos no condado. Condados que não apresentam círculos possuem número de casos igual a zero no período de 1998 a 2002. Percebemos que mesmo somando o número de casos nos 5 anos, ainda existem vários condados com zero casos e muitos com poucos casos.

Figura 5.1: Distribuição do número de casos por condado.




Percebe-se que os condados que apresentaram maiores médias da taxa de mortalidade se encontram próximos, localizando-se no canto inferior esquerdo do mapa. Os condados de Fulton e Dekalb possuem fronteira em comum e juntos possuem 36 casos, quando pouco mais de 28 casos eram esperados. Os condados de Richmond e Chatham localizam-se no limite territorial leste do estado. Em uma análise mais minuciosa seria interessante incluir os condados vizinhos pertencentes ao estado da Carolina do Sul. O mesmo vale para a inclusão de condados vizinhos dos outros estados que compartilham fronteira com o estado da Geórgia.

Apresentaremos na próxima seção os resultados utilizando o Scan ZIPET e o Scan-ET. Para a estatística Scan ZIPET foram utilizadas as duas definições de zero estrutural comentadas na seção 3.6. Para facilitar a notação, definiremos o Scan ZIPET como Scan ZIPET-ZER quando for utilizada a suposição de zero estru-

tural rígido e Scan ZIPET-ZEF quando for utilizada a definição de zero estrutural flexível.

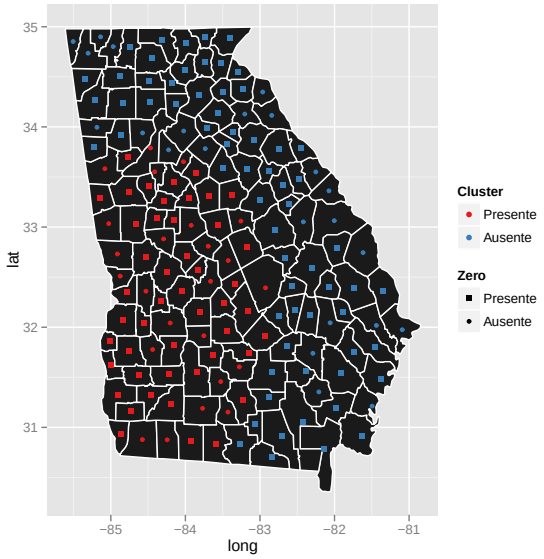
## 5.2 Resultados

Nesta seção serão apresentados os resultados para cada um dos métodos de forma paralela, de tal forma a facilitar a comparação.

O zero estrutural pode ser visto, nesta aplicação, como consequência de políticas de saúde pública aplicadas por alguns condados combinado a outros fatores presentes no condado que levem a um número de casos baixíssimo, ou quase impossível.

As figuras 5.2, 5.3 e 5.4 exibem o cluster encontrado segundo cada um dos métodos. O tom avermelhado foi utilizado para simbolizar a região do cluster. Para diferenciar entre a apresentação de valores e os zeros, foram utilizados dois símbolos diferentes.

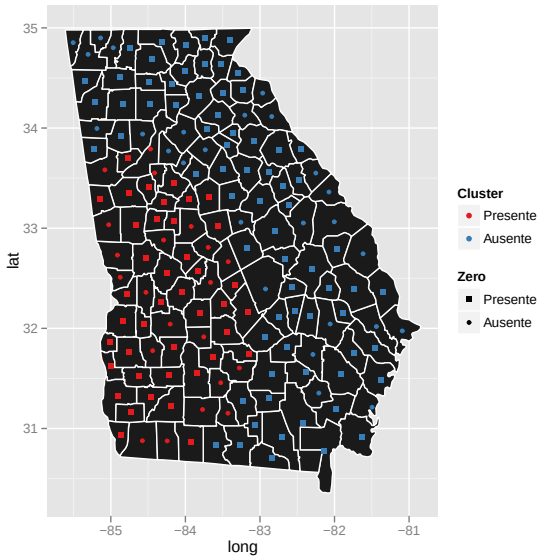
Figura 5.2: Regiões Detectadas através da estatística Scan-ET.




Observando a figura 5.2 e a figura 5.3 percebemos que os dois métodos, Scan-ET e Scan ZIPET-ZER, apresentaram clusters contendo quase que o mesmo conjunto de regiões, com a diferença que o cluster proveniente da estatística Scan

ZIPET-ZER não apresenta um conjunto de regiões presente no limite direito do cluster da estatística Scan-ET.

Figura 5.3: Regiões Detectadas através da estatística Scan ZIPET com zero estrutural rígido.




A figura 5.4 diferencia-se mais das duas outras, incluindo a parte central e sul do estado da Geórgia. As primeiras duas representações apresentam conjuntos de regiões que se concentram na faixa oeste do estado, incluindo parte do sul e parte do centro.

Dos 159 condados presentes no estado da Geórgia, 49 apresentaram 0 casos no período de 1994 até 2008, fortalecendo, assim, o ponto de que determinados fatores presentes em alguns condados sejam capazes de impossibilitar a ocorrência de casos, em particular, mortes por tuberculose. A distribuição espacial dos condados que apresentaram zero casos ao longo de 15 anos pode ser encontrada na figura 5.5.

A população total dos 159 condados nos 5 anos analisados é de 41.061.941. O crescimento populacional no estado de 1998 até 2002 foi de 8,86%. Forsyth foi o condado que apresentou o maior crescimento no período, 37,903%, enquanto que o condado de Stewart apresentou um encolhimento populacional de 5,067% na sua população. A população para o estado da Geórgia de 1998 até 2002 é apresentada na tabela 5.4

Figura 5.4: Regiões Detectadas através da estatística Scan ZIPET com zero estrutural flexível.

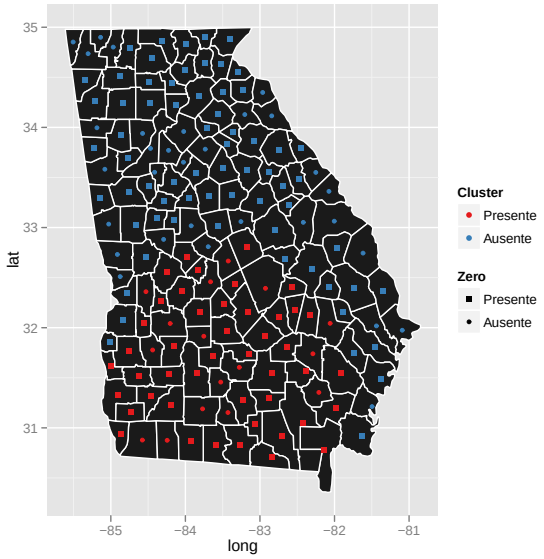



Tabela 5.4: População para o estado da Geórgia de 1998 até 2002.

| *1998* | *1999* | *2000* | *2001* | *2002* |
|---|---|---|---|---|
| 7.863.536 | 8.045.965 | 8.186.453 | 8.405.677 | 8.560.310 |

A variabilidade no crescimento populacional entre os condados reforça a importância de se levar em consideração a população para cada um dos anos cujos casos estão sendo analisados. A análise espaço-temporal utilizando casos de 1998 a 2002, mas fixando a população para o ano de 2000 poderia trazer conclusões distorcidas no sentido de não diferenciar entre um aumento efetivo do risco e um aumento do número de casos como consequência de um aumento populacional.

Os clusters apresentados por cada um dos métodos englobaram 3 anos, limite máximo de altura permitido para os algoritmos. O Scan-ET e o Scan ZIPET-ZER detectaram o cluster de 1998 a 2000, enquanto o Scan ZIPET-ZEF detectou para os anos de 1999 a 2001.

A população em cada um dos clusters em todo o período, assim como o número de condados que compuseram a base do cilindro, podem ser verificadas na tabela 5.5.

Percebe-se que, em ambas as estatísticas Scan ZIPET, a quantidade de condados presentes no cluster foi menor do que a quantidade presente no cluster de-

Figura 5.5: Distribuição espacial da ocorrência de zero casos de 1994 a 2008.

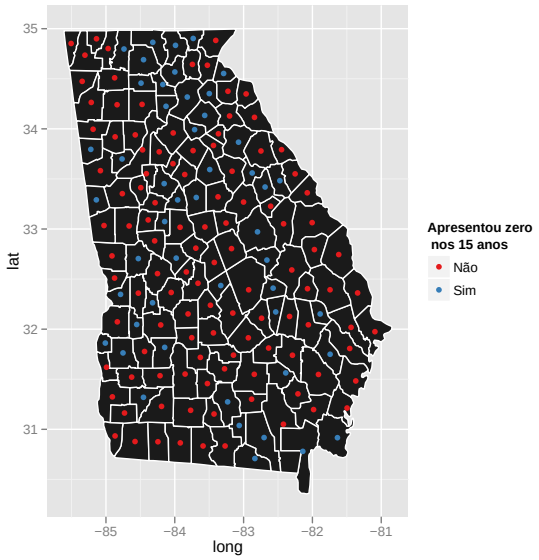



Tabela 5.5: Número de condados presentes no cluster, período de tempo no qual o cluster foi detectado e População presente dentro do cluster para cada um dos três métodos trabalhados.

| Variáveis | Scan-ET | Scan ZIPET ZER | Scan ZIPET ZEF |
|---|---|---|---|
| Número de condados | 71 | 62 | 61 |
| Período de tempo | 1998 - 2000 | 1998 - 2000 | 1999 - 2001 |
| População Dentro | 9.664.430 | 8.792.582 | 3.694.797 |

tectado através do Scan-ET. Em consequência, observa-se nos dois primeiros uma população inferior a deste último. A população dentro do cluster detectado pelo Scan-ZIPET ZEF é bem inferior às outras duas. A razão é que as regiões presentes nele diferiram em parte das encontradas através dos outros dois métodos.

No período estudado foram observados 157 casos de morte por tuberculose. A tabela 5.6 nos mostra a comparação para cada um dos métodos de como estes 157 casos deveriam ser distribuídos sob $H_0$, dentro e fora do cluster detectado, e como, de fato, ocorreu. Percebe-se valores próximos em relação ao número de casos observados dentro do cluster para as estatísticas Scan-ET e Scan ZIPET ZER. Comparando-se os valores observados com os esperados para dentro do cluster, destaca-se a estatística Scan ZIPET ZEF que apresentou um valor observado 76,97% acima do que seria esperado para as regiões presentes dentro do clus-

ter detectado, seguido da estatística Scan ZIPET ZER com 69,55% e por último a estatística Scan-ET com 65,08%.

Tabela 5.6: Número de casos observados e esperados para dentro e fora do cluster para os três métodos analisados.

| Variáveis | Scan-ET | Scan ZIPET ZER | Scan ZIPET ZEF |
|---|---|---|---|
| Número de Casos Observados Dentro do Cluster | 61 | 57 | 25 |
| Número de Casos Esperados Dentro do Cluster | 36,95 | 33,62 | 14,13 |
| Número de Casos Observados Fora do Cluster | 96 | 100 | 132 |
| Número de Casos Esperados Fora do Cluster | 120,05 | 123,38 | 142,87 |

A tabela 5.7 apresenta o log da razão de verossimilhança mais alto e em seguida a significância. Percebemos, ao se fixar uma significância de 0,05, que o único cluster significante seria o proveniente da estatística Scan ZIPET ZER.

O Scan ZIPET ZER apresentou um conjunto de regiões semelhante ao apresentado pelo Scan-ET, porém, o conjunto do primeiro se mostrou significante. Como já foi dito anteriormente, a inserção de uma região com um zero estrutural acarreta a elevação do número de casos esperados. Como a estatística Scan-ET não lida com esta particularidade o cluster tem sua verossimilhança reduzida.

Tabela 5.7: Log da razão de verossimilhança e significância para cada um dos três métodos analisados.

| Variáveis | Scan-ET | Scan ZIPET ZER | Scan ZIPET ZEF |
|---|---|---|---|
| Log da Razão de Verossimilhança | 9,116 | 10,429 | 17,632 |
| Significância | 0,055 | 0,041 | 0,111 |

Em relação a diferença observada entre os dois métodos Scan ZIPET, vale uma observação: a inserção do conceito de zero estrutural não traz como consequência intrínseca o aparecimento de clusters antes ocultos. Pode ocorrer de um cluster que havia sido considerado significante através da estatística Scan-ET não ser significativo após a aplicação da estatística Scan ZIPET. Como exemplo,

podemos afirmar que, fora de um cluster detectado, havia a presença de regiões com zero estrutural. O que estava ocorrendo, então, é que se atribuía para fora do cluster mais casos esperados do que de fato poderia ocorrer, uma vez que as regiões com zero estrutural não podem apresentar casos.

Com a inserção do zero estrutural nos cálculos, as populações que estavam gerando casos esperados, quando na verdade não tem probabilidade alguma de ocorrência, terão a contribuição da sua população reduzida de acordo com a probabilidade de o zero apresentado ser um zero estrutural. Isto implica em uma redução no número de casos esperados para fora do cluster e consequentemente em um aumento do número esperado para dentro do cluster, reduzindo assim a verossimilhança.

A redução da contribuição da população de uma região cujo zero foi candidato a zero estrutural ocorrerá tanto se a região estiver presente dentro do cluster analisado quanto fora. No entanto a presença ou não da região dentro do cluster leva a diferentes estimações, como pode ser observado na equação (3.49). A diferença está na intensidade da redução da população. Ao ser delimitado um conjunto de regiões que apresentam uma probabilidade mais elevada de apresentar ocorrências e dentro desta área observa-se uma região com zero casos, que é candidata a zero estrutural, esta região terá sua população reduzida com maior intensidade do que se esta região estivesse fora das regiões delimitadas.

O fato de uma região está inserida dentro de uma zona com probabilidade superior de ocorrência implica em uma probabilidade menor de apresentar zero casos, quando ainda assim uma região sob estas circunstâncias apresenta zero casos, aumenta-se a probabilidade de que esta apresente na verdade um zero estrutural (vide equação 3.49). O contrário também é válido, um conjunto de regiões que apresente probabilidade menor de ocorrência possui uma probabilidade maior de observar uma região com zero casos, desta forma, a ocorrência de zeros amostrais é mais natural e consequentemente probabilidade de uma região conter um zero estrutural é menor. Estas duas situações explicam a razão pela qual duas re-

giões com valor populacional idêntico teriam suas populações reduzidas de forma diferente se uma se encontrasse dentro e a outra fora do cluster.

Com estas considerações percebemos que a inserção do zero estrutural traz uma nova luz para a questão. A contabilização do número de casos esperados para a razão de verossimilhança deverá levar em conta apenas regiões que de fato têm condições de apresentar casos, ou seja, que não apresentem zeros estruturais. Porém, como a definição de zero estrutural não é absoluta, como discutido na seção 3.6, encontramos algumas discordâncias, de acordo com as definições de zero estrutural utilizadas. Isto pode ser visto ao se comparar os resultados das estatísticas Scan ZIPET ZER e Scan ZIPET ZEF.

# 6 Considerações Finais

A detecção de novas doenças, de novos fatores de risco ou de novos locais para fatores antigos consiste em informação valiosa para o processo de construção de políticas públicas na área de saúde. Desta forma, é importante definir se o padrão espacial observado constitui um cluster ou se pode ser atribuído de forma razoável ao acaso.

A inserção do zero estrutural levantou questões importantes para a aplicação. Estende-se a questão da modelagem estatística para o questionamento se os zeros apresentados são zeros amostrais ou zeros estruturais. A distinção entre as duas possibilidades, no caso espaço-temporal e na aplicação para casos de morte por tuberculose, revela a diferença que pode ocorrer de acordo com o que se assume sobre o comportamento dos zeros estruturais.

As reflexões apresentadas nos capítulos precedentes buscaram exemplificar o que ocorre quando todos os zeros são considerados zeros amostrais. Esta abordagem distorce a detecção de cluster quando se reconhece a possibilidade da presença de zeros estruturais. Clusters podem ser revelados ou encobertos de forma indevida.

As simulações foram construídas segundo a suposição de zero estrutural rígida. Por esta razão só foi aplicada a estatística Scan ZIPET ZER, que apresentou resultados superiores à estatística Scan-ET, para as simulações numéricas realizadas.

Em relação aos dados reais a estatística Scan ZIPET apresentou um conjunto de regiões significantes, para o caso de uma definição de zero estrutural mais rígida, e um conjuntos de regiões não significantes, para a definição de zero

estrutural mais flexível. Essa diferença nos resultados revela a necessidade de um método que possa indicar com melhor precisão quais regiões serão as candidatas para a estimação através do EM.

Comparando-se um mesmo conjunto de regiões entre as diferentes estatísticas nota-se que pode haver uma grande variabilidade na razão de verossimilhança de acordo com a aplicação que está sendo analisada, podendo ocorrer que a região mais provável de ser um cluster segundo um algoritmo tenha sua razão de verossimilhança completamente rebaixada segundo outro método.

Percebemos então que a inserção do zero estrutural e consequente redução da população traz impacto para o cálculo dos valores esperados. Se a redução ocorrer mais forte fora do cluster do que dentro, gera-se um redução no destaque do cluster. No caso contrário o cluster tem sua razão de verossimilhança aumentada e pode passar de não significante para significante.

## 6.1 Trabalhos Futuros

A estatística Scan ZIPET se ateve a uma base circular, mas sua extensão para outros formatos de base pode trazer contribuições interessantes ao incorporar zeros estruturais com um formato mais flexível de base. Uma possibilidade seria testar a utilização de uma base elíptica.

Seria interessante a criação de um método para a definição de quais regiões, e em que períodos de tempo, seriam candidatas a zero estrutural. Este método buscaria o ponto ótimo entre a flexibilização completa e a rigidez total.

Uma aplicação que traria comparações importantes, entre os métods Scan-ET e o Scan ZIPET, seria no contexto em que é possível estabelecer quais regiões e em quais períodos o zero apresentado é na verdade um zero estrutural.

Simular cenários sob diferentes suposições de zero estrutural e comparar o desempenho de cada uma das estatísticas Scan ZIPET ZER, Scan ZIPET ZEF e outras que utilizem definições intermediárias também é uma possibilidade.

Pode-se verificar o desempenho da estatística Scan ZIPET segundo definições alternativas do valor esperado, quando por exemplo, são fixados os totais de casos para os períodos de tempo e distribuídos proporcionalmente às populações no período, ao invés de somar todos os casos para todos os períodos e distribuir proporcionalmente as populações em todos os períodos.

# Referências Bibliográficas


Agarwal, D. K., Gelfand, A. E., e Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9(4):341–355.

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

Cançado, A. L. F., da-Silva, C. Q., e Silva, M. F. (2012). A spatial scan statistic for zero inflated poisson process. *Submetido*.

Choynowski, M. (1959). Maps based on probabilities. *Journal of the American Statistical Association*, 54(286):385–388.

Clayton, D. e Kaldor, J. (1987). Empirical bayes estimates of Age-Standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.

Dempster, A., Laird, N., e Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Gómez-Rubio, V., Ferrándiz, J., e López, A. (2003). Detecting clusters of disease with r. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria. Hornik, K.; Leisch, F. & Zeileis.

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

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

Johnson, R. A. e Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall.

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

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

Knox, E. G. e Bartlett, M. S. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 13(1):25–30.

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

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A*, 164(Part 1):61–72.

Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A., e Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in los alamos, new mexico. *American Journal of Public Health*, 88(9):1377.

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

Kulldorff, M. e Nagarwalla, N. (1995). Spatial disease cluster: Detection and inference. *Statistics in Medicine*, 14:799–810.

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

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220.

Marshall, R. J. (1991a). Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(2):283–294.

Marshall, R. J. (1991b). A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(3):421–441.

Naus, J. (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association*, 69(347):810–815.

Naus, J. I. (1965a). Clustering of random points in two dimensions. *Biometrika*, 52(1/2):263–267.

Naus, J. I. (1965b). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60(310):532–538.

Oliveira, F. L., Duczmal, L. H., Cançado, A. L., e Tavares, R. (2011). Nonparametric intensity bounds for the delineation of spatial clusters. *International Journal of Health Geographics*, 10(1):1.

Openshaw, S., Charlton, M., Craft, A., e Birch, J. (1988). Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet*, 331(1):272–273.

Sonesson, C. e Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 166(1):5–21.

Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L., e Clark, L. C. (1989). Monitoring for clusters of disease; application to leukemia incidence in upstate new york. Technical Report 840, School of Operations Research and industrial Engineering College of Engineering Cornell University, Ithaca, NY.

Whittemore, A. S., Friend, N., Brown, B. W., e Holly, E. A. (1987). A test to detect clusters of disease. *Biometrika*, 74(3):631–635.