Departamento de Estatística
Universidade de Brasília

Juliano Cézar Sant'Anna
Matrícula: 180002872

Estatística de varredura espacial Touchard baseada em expectância

Brasília
11 de Março de 2020

Juliano Cézar Sant'Anna

Estatística de varredura espacial Touchard baseada em expectância

Projeto apresentado para obtenção do título de Mestre em Estatística na Universidade de Brasília

Departamento de Estatística - Universidade de Brasília

Orientador: André LF Cançado

Brasília
11 de Março de 2020

Dedico este trabalho a Deus, aos meus pais Gisele e Hudson, à Roseney, à minha parceira Karima e aos meus irmãos Davi e Rafael (in memorian).

Agradecimentos

Ao meu orientador o Doutor Professor André Luiz Fernandes Cançado, pela orientação prestada, pela sua disponibilidade e apoio que sempre demonstrou.

A minha família pelo auxílio, incentivo e carinho a mim dedicados.

Aos amigos pela ajuda nas horas de estudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

As estatísticas Scan são amplamente utilizadas tanto na detecção de clusters espaciais quanto no teste de sua significância. Usualmente, ao lidar com contagens, utiliza-se essa técnica associada à distribuição de Poisson. No entanto, ela não representa bem dados nos quais ocorrem superdispersão, subdispersão ou excesso de zeros. A distribuição Touchard proposta por Matsushita et al. (2018) é uma generalização da Poisson que funciona como solução única para esses três problemas. Este trabalho tem como objetivo adaptar a estatística Scan baseada em expectância, proposta por Neill (2006) para dados modelados pela distribuição Touchard. A técnica foi comparada com o Scan Poisson baseado em expectância para dados simulados e para um estudo de casos de suicídios nos municípios do Acre em 2017. Foram alcançados resultados satisfatórios acerca do método apresentado.

Palavras-chave: Estatística. Estatística Scan. Cluster espacial. Distribuição Touchard.

Abstract

Scan statistics are widely used as much in detecting spatial clusters as in testing their significance. Usually, when dealing with counts, this technique is used associated with the Poisson distribution. However, it does not represent well data in which overdispersion, underdispersion or excess of zeros occur. The Touchard distribution proposed by Matsushita et al. (2018) is a Poisson generalization that works as an unique solution to these three problems. This work aims to adapt the Scan statistic based on expectation, proposed by Neill (2006) for data modeled by the Touchard distribution. The technique was compared with the Scan Poisson based on expectation for simulated data and for a study of suicide cases in the municipalities of Acre in 2017. Satisfactory results on the presented method were achieved.

Keywords: Statistics. Scan Statistics. Spatial cluster. Touchard distribution.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O campo de Análise Espacial de Dados possui diversos métodos, os quais têm como objetivo investigar os dados levando em consideração sua localização e seus aspectos geográficos. Dentre esses estão as estatísticas de varredura espacial (estatísticas Scan). Essas estatísticas são utilizadas para se detectar concentrações anômalas de pontos através de uma varredura ao longo do mapa em estudo.

Naus (1965b) propôs uma estatística de varredura, que foi bastante estudada nas décadas seguintes, ficando mais conhecida pela abordagem de Kulldorff (1997), que a estendeu a configurações multidimensionais e janelas de corte de diferentes tamanhos utilizando as distribuições de Poisson e Bernoulli.

A estatística Scan de Kulldorff (1997) está relacionada ao problema de agrupamentos de pontos no espaço, isto é, ela identifica *clusters* espaciais de dados. Essa técnica teve diversas extensões ao longo dos anos como: para dados ordinais (Jung et al., 2007), utilizando a distribuição exponencial para dados de sobrevivência (Huang et al., 2007), ajustado para múltiplos clusters (Zhang et al., 2010) e uma abordagem não paramétrica baseada em postos multivariados para dados não Gaussianos (Cucala, Genin, Occelli, and Soula, 2019).

Dentre as aplicações dessa técnica e suas extensões podemos citar estudar concentrações de doenças infecciosas transportadas pelo ar (Chen et al., 2013); (Souris et al., 2014), doenças sexualmente transmissíveis (Jennings et al., 2005); (González et al., 2015), diabetes (Green et al., 2003), ecologia (Seidel and Boyce, 2015), criminologia (Zeoli et al., 2014), entre outras. Além da função exploratória da técnica de Kulldorff (1997), a de localizar *clusters* espaciais, ela é capaz de testar a significância desses conglomerados através de testes de razão de verossimilhança, ou seja, verificar se eles ocorrem ao acaso ou não.

De acordo com Neill (2006) essa técnica é baseada na população, o que significa que ela usa como referência para seus cálculos a população de cada região. Para ele há também o Scan com base em expectância que usa como referência o número esperado de casos estimado através de dados históricos.

Esse trabalho tem como objetivo adaptar a estatística Scan baseada em expectância, proposta por Neill (2006) para dados modelados pela distribuição Touchard (Matsushita et al., 2018). Apesar de cada método ter suas vantagens, nesse caso, a abordagem baseada em expectância foi usada pois, assumindo que há uma quantidade suficiente de dados históricos e que podemos estimar com precisão a contagem esperada em cada localização espacial então, de acordo com Neill (2006), estatísticas baseadas em expectância terão maior poder de detecção do que as estatísticas baseadas na população.

As técnicas citadas são muitas vezes utilizadas com a distribuição de Poisson, já que esta é adequada para a modelagem de contagens, como no caso em que os dados se referem

a casos de uma doença. Porém, a distribuição de Poisson não representa bem dados nos quais ocorrem superdispersão, subdispersão ou excesso de zeros.

Pensando nisso a distribuição de Touchard foi proposta por Matsushita et al. (2018) como uma solução única para esses três problemas. Dito isso, esse trabalho propõe uma generalização da estatística Scan com distribuição de Touchard.

2 Metodologia

2.1 Revisão Bibliográfica

Antes de apresentar a técnica desenvolvida, é necessário trazer algumas definições importantes e outras ferramentas que serviram de base para a estatística Scan.

2.1.1 Análise de Cluster e Análise de Cluster Espacial

Para entender análise de cluster espacial, é preciso saber o que são clusters.

Na estatística esse termo se refere a um grupo de objetos homogêneos entre si e heterogêneos em relação a outros agrupamentos. A análise de agrupamentos engloba diversas técnicas que objetivam classificar em grupos, através de algoritmos, observações similares de forma a facilitar a interpretação dos dados.

A análise de clusters espaciais é um pouco diferente já que busca encontrar áreas com características comuns, seja para um simples estudo exploratório que precede uma pesquisa maior ou para se chegar a conclusões acerca de padrões escondidos nos dados, através da relação entre a ocorrência de eventos e sua localização geográfica.

A presença de agrupamentos costuma ser ilustrada por meio de mapas, nos quais as observações são dados pontuais, isto é, os eventos são expressos como pontos identificados no mapa através de pares de coordenadas (x, y). Por exemplo: posicionamento de espécies de fauna ou flora, de crimes hediondos, crianças com leucemia, entre outros. A ferramenta descrita nesse estudo, utiliza dados pontuais.

2.1.2 Técnicas de Análise Espacial de Dados

Como mencionado anteriormente, o Scan Circular de Kulldorff (1997) teve como base diversas técnicas, algumas das quais são revisadas nas seções seguintes.

2.1.2.1 Mapas Baseados em probabilidade de Choynowski (1959)

Antes da técnica de Choynowski (1959), mapas estatísticos eram utilizados apenas como uma ferramenta descritiva, trazendo frequências absolutas ou porcentagens para representar a distribuição geográfica do objeto de estudo. Levando isso em conta, ele propôs um mapa a partir do qual pudessem ser feitas inferências acerca da distribuição espacial de algum fenômeno. Partindo da suposição de que os dados se distribuem segundo uma distribuição Uniforme, ele calculou as probabilidades de cada um dos desvios observados da média.

Uma vantagem desse mapa é que ele evita que se chegue a conclusões baseadas em variações amostrais. Por outro lado as probabilidades não têm fácil interpretação.

2.1.2.2 Distribuição do Tamanho do Máximo de *clusters* em uma linha - Naus (1965b)

Utilizando uma distribuição Uniforme$(0, 1)$, são extraídos N pontos de forma independente. Define-se como evento de interesse a existência de um subintervalo de $(0, 1)$ de tamanho t que contém, pelo menos, n dos N pontos amostrados. Logo Naus encontra a probabilidade desse evento para $n > N/2$.

Apesar de se tratar de pontos na reta, isto é, apenas uma dimensão, essa técnica pode ser relacionada à análise de *clusters* espacial, se ao invés de se utilizar as coordenadas (latitude, longitude), utilizar a soma delas, reduzindo o problema para uma dimensão.

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

Nesse artigo, Naus trata de agrupamentos de pontos em duas dimensões. São escolhidos, ao acaso, N pontos de um quadrado unitário representados por suas coordenadas. O evento de interesse é similar ao do tópico anterior, porém dessa vez no R^2. Ou seja, busca-se um subretângulo de um quadrado de unidade 1, com lados u e v, que contém pelo menos n dos N pontos escolhidos. O autor encontra limites superiores e inferiores para a probabilidade desse evento, que convergem conforme u e v se aproximam de 0. O artigo refuta trabalhos anteriores, mostrando que o formato da janela de corte importa para maximizar a probabilidade de encontrar um *cluster* grande.

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

A técnica desenvolvida por Openshaw et al. (1988) (Geographical Analysis Machine) foi capaz de identificar a presença de *clusters*, em um conjunto de dados sobre casos de leucemia infantil no Norte da Inglaterra sobre o qual métodos prévios não perceberam.

O método GAM é uma ferramenta exploratória, que constrói uma grade de pontos no mapa, e para cada ponto são desenhados diversos círculos sobrepostos, de tamanhos diferentes, regularmente espaçados, que abrangem toda a área de interesse. Em seguida, são contados os pontos contidos em cada círculo, para serem comparados ao número esperado caso fossem distribuídos de acordo com uma distribuição Poisson. Isso é feito utilizando simulações de Monte Carlo. Os círculos significativos são desenhados no mapa. O problema desse método é que múltiplos testes são realizados, o que aumenta a possibilidade de erro tipo 1. Para contornar isso pode ser usado o ajuste de Bonferroni, o que em contrapartida pode tornar o teste muito conservador em razão de sua fórmula que depende do número de testes realizados. Outro problema é que esse método é computacionalmente intensivo.

2.1.2.5 A detecção de *clusters* de Doenças Raras - Besag and Newell (1991)

Com seu trabalho, Besag and Newell (1991) queriam encontrar *clusters* de doenças raras. Sua região de estudo é dividida em pequenas zonas administrativas e em cada uma o número de observações está associado a um centróide. Para cada uma dessas i zonas são realizados testes de significância, nos quais a i-ésima zona é o centro e outras zonas são agregadas, por distância, até atingir um número de casos pré-definido. A estatística do teste é o número mínimo de zonas necessárias para acumular pelo menos c casos, seguindo a ordenação por distância. O nível de significância do teste é calculado de forma aproximada por uma probabilidade dada pela distribuição Poisson. Após esse cálculo, uma boa medida de diagnóstico, de acordo com os autores, é fazer um mapa com todos os círculos que foram significativos a um nível de 5%.

Eles desenvolveram uma técnica com menor carga computacional que o método GAM e mais fundamentação matemática. Como desvantagem, um tamanho mínimo para o *cluster* precisa ser definido a priori.

Depois de relembrar algumas técnicas da área, vamos tratar da estatística Scan de Kulldorff (1997) e uma modificação proposta por Neill (2006), as bases para esse trabalho.

2.1.2.6 Kulldorff (1997) e Neill (2006)

A estatística Scan de Kulldorff (1997) traz alguns aspectos em comum com os trabalhos anteriores, pois é capaz de detectar clusters em um processo pontual multidimensional. O algoritmo emprega janelas circulares com centros e raios variados ao longo da região de estudo.

Apesar das similaridades, os trabalhos anteriores falham em relação a resolver de forma conjunta os problemas de múltiplos testes de hipótese com janelas de busca passando por diversas regiões de vários tamanhos ou formatos e de considerar a proximidade espacial como um fator na análise. Já Kulldorff (1997) resolve esses problemas considerando um teste de razão de verossimilhança para o qual a hipótese nula é de que não há presença de clusters, e a alternativa é de que existe um cluster. Além disso, o teste é realizado apenas para o cluster mais evidente, e não para cada cluster candidato, o que resolve o problema de múltiplos testes.

O Scan Circular pode ser descrito da seguinte forma:

Seja um mapa repartido em t regiões, cada uma com sua respectiva contagem de eventos de interesse c_i, associada a um centróide, e população exposta ao risco $n_i, i = 1, .., t$. O número total de casos de interesse e a população total do mapa são dados, respectivamente, por $C = \sum_{i=1}^{t} c_i$ e $N = \sum_{i=1}^{t} n_i$.

Considere agora um subconjunto de regiões adjacentes denominado zona z. Então podemos definir:

- $c_z = \sum_{i \in z} c_i$ número de eventos de interesse dentro de z.
- $c_{\bar{z}} = \sum_{i \notin z} c_i$ número de eventos de interesse fora de z.
- $n_z = \sum_{i \in z} n_i$ tamanho da população dentro de z.
- $n_{\bar{z}} = \sum_{i \notin z} n_i$ tamanho da população fora de z.

Cada zona z é um candidato a cluster. O próximo passo é, através do teste de razão de verossimilhança, avaliar a significância estatística da zona mais verossímil. As hipóteses do teste são:

$$\begin{cases} H_0 : \text{A probabilidade de um indivíduo vir a ser um caso é a mesma em qualquer zona do mapa.} \\ H_1 : \text{Existe uma zona } z \text{ tal que a probabilidade de um indivíduo vir a ser um caso é maior em } z \text{ do que fora de } z. \end{cases}$$

A rejeição da hipótese inicial implica a existência de um cluster da característica em estudo.

Kulldorff (1997) considera a distribuição de Poisson para modelar o número de casos em cada região. A seguir iremos apresentar a estatística Scan obtida para esse modelo.

2.1.2.7 Modelo Poisson

Considere que $c_i \sim Poisson(\alpha_i n_i)$. Portanto,

$$P(c_i = k) = \frac{e^{-\alpha_i n_i}(\alpha_i n_i)^k}{k!}. \tag{2.1}$$

Pela definição de c_z temos que $c_z \sim Poisson(\alpha_z n_z)$, em que:

- α_z: é a probabilidade de um indivíduo, que pertence a uma região na zona z, ser um caso.
- $\alpha_{\bar{z}}$: é a probabilidade de um indivíduo, que pertence a uma região fora da zona z, ser um caso.

E as hipóteses, nesse caso, são:

$$\begin{cases} H_0 : \alpha_z = \alpha_{\bar{z}} = \alpha_0, \text{para toda zona z.} \\ H_1 : \text{Existe uma zona } z \text{ tal que } \alpha_z > \alpha_{\bar{z}}. \end{cases} \tag{2.2}$$

Para obtermos a estatística teste iremos computar as verossimilhanças sob H_0 e sob H_1 e encontrar os EMV (Estimadores de Máxima Verossimilhança) para ambos os casos. Sob H_0 temos:

$$L_0(\mathbf{c};\alpha_0) = \prod_{i=1}^{t} \frac{e^{-\alpha_0 n_i}(\alpha_0 n_i)^{c_i}}{c_i!} = e^{-\alpha_0 N}\alpha_0^C \prod_{i=1}^{t} \frac{(n_i)^{c_i}}{c_i!} \tag{2.3}$$

Aplicando o logaritmo temos:

$$\log L_0(\mathbf{c};\alpha_0) = -\alpha_0 N + C\log\alpha_0 + \sum_{i=1}^{t} \log \frac{(n_i)^{c_i}}{c_i!} \tag{2.4}$$

Derivando em relação a α_0 e igualando a zero:

$$\begin{aligned} \frac{\partial \log L_0(\mathbf{c};\alpha_0)}{\partial \alpha_0} &= -N + \frac{C}{\alpha_0} = 0 \\ N &= \frac{C}{\alpha_0} \\ \hat{\alpha}_0 &= \frac{C}{N} \end{aligned} \tag{2.5}$$

Agora sob H_1:

$$\begin{aligned} L_1(\mathbf{c};\alpha_{\bar{z}},\alpha_z) &= \prod_{i\in z} \frac{e^{-\alpha_z n_i}(\alpha_z n_i)^{c_i}}{c_i!} \prod_{i\notin z} \frac{e^{-\alpha_{\bar{z}} n_i}(\alpha_{\bar{z}} n_i)^{c_i}}{c_i!} \\ &= e^{-\alpha_z n_z}\alpha_z^{c_z} e^{-\alpha_{\bar{z}} n_{\bar{z}}}\alpha_{\bar{z}}^{c_{\bar{z}}} \prod_{i\in z} \frac{(n_i)^{c_i}}{c_i!} \prod_{i\notin z} \frac{(n_i)^{c_i}}{c_i!} \\ &= e^{-\alpha_z n_z}\alpha_z^{c_z} e^{-\alpha_{\bar{z}} n_{\bar{z}}}\alpha_{\bar{z}}^{c_{\bar{z}}} \prod_{i=1}^{t} \frac{(n_i)^{c_i}}{c_i!} \end{aligned} \tag{2.6}$$

Aplicando o logaritmo obtemos:

$$\log L_1(\mathbf{c};\alpha_{\bar{z}},\alpha_z) = -\alpha_z n_z + c_z \log\alpha_z - \alpha_{\bar{z}} n_{\bar{z}} + c_{\bar{z}} \log\alpha_{\bar{z}} + \sum_{i=1}^{t} log \frac{(n_i)^{c_i}}{c_i!} \tag{2.7}$$

Derivando em relação a $\alpha_{\bar{z}}$ e α_z, e igualando ambos a zero:

$$\begin{aligned} \frac{\partial \log L_1(\mathbf{c};\alpha_{\bar{z}},\alpha_z)}{\partial \alpha_z} &= -n_z + \frac{c_z}{\alpha_z} = 0 \\ n_z &= \frac{c_z}{\alpha_z} \\ \hat{\alpha}_z &= \frac{c_z}{n_z} \end{aligned} \tag{2.8}$$

$$\begin{aligned} \frac{\partial \log L_1(\mathbf{c};\alpha_{\bar{z}},\alpha_z)}{\partial \alpha_{\bar{z}}} &= -n_{\bar{z}} + \frac{c_{\bar{z}}}{\alpha_{\bar{z}}} = 0 \\ n_{\bar{z}} &= \frac{c_{\bar{z}}}{\alpha_{\bar{z}}} \\ \hat{\alpha}_{\bar{z}} &= \frac{c_{\bar{z}}}{n_{\bar{z}}} \end{aligned} \tag{2.9}$$

Então define-se a razão de verossimilhança λ dessa forma:

$$\lambda = \frac{\sup\limits_{\alpha_z > \alpha_{\bar{z}}} L_1}{\sup\limits_{\alpha_z = \alpha_{\bar{z}}} L_0} = \frac{e^{-\hat{\alpha}_z n_z} \hat{\alpha}_z^{c_z} e^{-\hat{\alpha}_{\bar{z}} n_{\bar{z}}} \hat{\alpha}_{\bar{z}}^{c_{\bar{z}}} \prod\limits_{i=1}^{t} \frac{(n_i)^{c_i}}{c_i!}}{e^{-\hat{\alpha}_0 N} \hat{\alpha}_0^{C} \prod\limits_{i=1}^{t} \frac{(n_i)^{c_i}}{c_i!}} \\ = e^{-(\hat{\alpha}_z n_z + \hat{\alpha}_{\bar{z}} n_{\bar{z}} - \hat{\alpha}_0 N)} \hat{\alpha}_z^{c_z} \hat{\alpha}_{\bar{z}}^{c_{\bar{z}}} \hat{\alpha}_0^{-C} \tag{2.10}$$

Substituindo-se as estimativas dos parâmetros temos:

$$\lambda_z = e^{-(\frac{c_z}{n_z} n_z + \frac{c_{\bar{z}}}{n_{\bar{z}}} n_{\bar{z}} - \frac{C}{N} N)} \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{c_{\bar{z}}}{n_{\bar{z}}}\right)^{c_{\bar{z}}} \left(\frac{C}{N}\right)^{-C} \tag{2.11a}$$

Como $C = c_z + c_{\bar{z}}$, então:

$$\lambda_z = e^0 \left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{c_{\bar{z}}}{n_{\bar{z}}}\right)^{c_{\bar{z}}} \left(\frac{C}{N}\right)^{-c_z + c_{\bar{z}}} = \left(\frac{c_z}{n_z}\frac{N}{C}\right)^{c_z} \left(\frac{c_{\bar{z}}}{n_{\bar{z}}}\frac{N}{C}\right)^{c_{\bar{z}}} \tag{2.11b}$$

Logo para cada zona z:

$$\lambda_z = \begin{cases} \left(\frac{c_z}{n_z}\frac{N}{C}\right)^{c_z} \left(\frac{c_{\bar{z}}}{n_{\bar{z}}}\frac{N}{C}\right)^{c_{\bar{z}}}, \text{se } \frac{c_z}{n_z} > \frac{c_{\bar{z}}}{n_{\bar{z}}}. \\ 1, \text{caso contrário.} \end{cases} \tag{2.11c}$$

A condição $\frac{c_z}{n_z} > \frac{c_{\bar{z}}}{n_{\bar{z}}}$ é explicada pelo fato de estarmos interessados apenas em clusters de alta incidência, isto é, aglomerados nos quais a proporção de casos seja maior que no resto da área em estudo.

Para encontrar o cluster mais verossímil é necessário calcular a razão de verossimilhança para cada zona candidata, de forma a encontrar a de maior valor, ou seja, a estatística do teste é $T = \sup\limits_{z} \lambda_z$. É importante notar que o mesmo valor que maximiza λ_z maximiza $\log \lambda_z$, por isso é preferível calcular $T^* = \log T$, uma vez que esta é mais adequada para cálculos computacionais.

Após encontrar a zona z mais verossímil, a próxima tarefa é verificar sua significância. Para isso é necessário saber a distribuição de T^* sob H_0. Como não conseguimos encontrar essa distribuição de forma analítica, utiliza-se simulação de Monte Carlo, de modo a obter uma distribuição empírica e por fim calcular o p-valor.

2.1.3 Construindo as zonas candidatas

Como foi explicado, o algoritmo calcula a razão de verossimilhança para cada zona z. Porém, antes disso, devem-se definir as regiões que irão compor cada uma dessas zonas.

No algoritmo Scan Circular, o conjunto de zonas Z é determinado utilizando janelas circulares de centros e raios diferentes, que por meio de uma varredura do mapa formam as zonas. Isso significa que se a janela engloba as coordenadas dos centróides de determinadas regiões, logo essas regiões compõem uma zona.

Essa composição é feita através de medidas de distância entre os centróides de cada região.

Seja $d_{i,j}$ a distância euclidiana entre os centróides da região i e j com coordenadas (x_i, y_i) e (x_j, y_j), respectivamente. Então temos a seguinte fórmula para essa distância no R^2 e seus axiomas:

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2.12a}$$

$$d_{i,j} \geq 0 \tag{2.12b}$$

$$d_{i,j} = 0 \text{ se } i = j \tag{2.12c}$$

$$d_{i,j} = d_{j,i} \tag{2.12d}$$

Seja n_{max} o tamanho máximo da população para cada zona z, então para construir o conjunto Z de zonas-candidatas, adota-se o seguinte procedimento: considere, inicialmente, a zona z_1 formada pela região 1; em seguida, a zona z_2 formada pela região 1 e a região mais próxima da região 1; depois, a zona z_3 formada pela região 1 mais as duas regiões mais próximas da região 1, e assim sucessivamente. As zonas são formadas adicionando-se as regiões mais próximas à região 1, enquanto valer a desigualdade $n_z < n_{max}$, isto é, enquanto o tamanho da população dentro da zona for menor que n_{max} esse processo continua. Em seguida, o procedimento é reiniciado começando da zona formada apenas pela região 2, e depois a partir de cada uma das demais regiões.

Note que a inclusão das regiões nas zonas por ordem de distância em relação à região inicial equivale a adotar uma sequência de janelas circulares com centro na região inicial com raios variando de zero (em que apenas uma região é incluída) até o limite em que a população dentro da zona respeita $n_z < n_{max}$.

Para poder vizualizar o funcionamento desse algoritmo, a figura 1 traz uma ilustração simplificada do mesmo. Isto é, foi representada a formação de duas zonas com regiões iniciais diferentes.

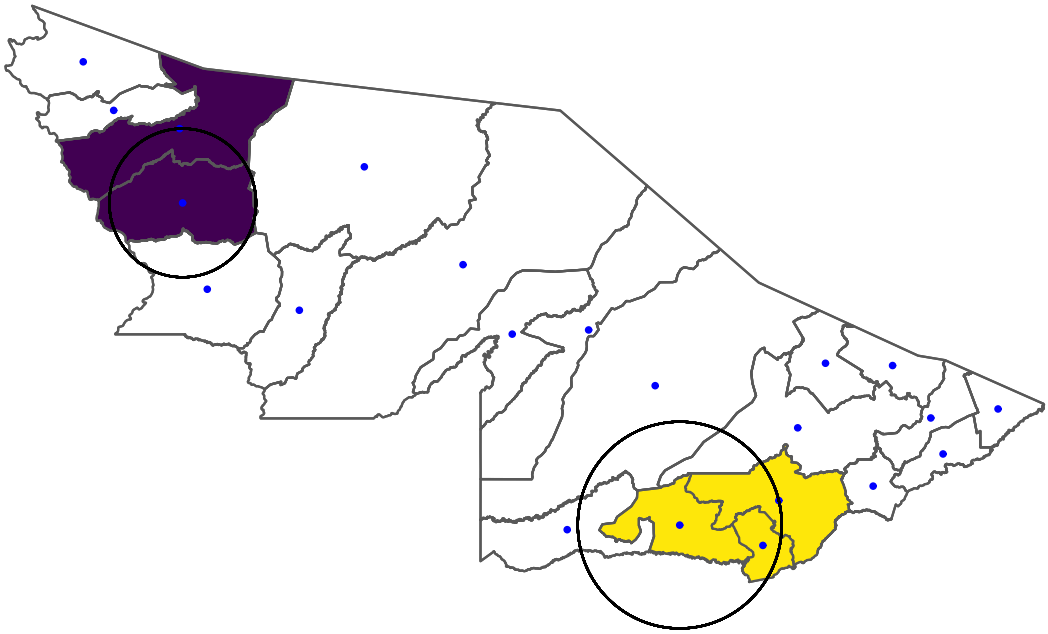

Figura 1 – Exemplo do algoritmo scan circular para o mapa do estado do Acre, zona roxa formada por Porto Walter e Cruzeiro do Sul e zona amarela formada por Brasiléia, Epitaciolândia e Xapuri.

Os pontos azuis representam os centróides de cada município, enquanto os círculos pretos são as janelas de corte circulares que formam cada zona. Podemos ver que os municípios presentes em cada zona são apenas aqueles cujos centróides estão contidos na janela circular.

2.1.4 Simulação de Monte Carlo

Para todas as zonas contidas no conjunto Z de zonas candidatas a cluster calcula-se o valor do logaritmo da razão de verossimilhança. A zona de maior valor da razão de verossimilhança é a zona mais verossímil.

Em seguida deve-se testar a significância do cluster através da distribuição de T^* sob H_0. Porém, como dito anteriormente, isso não é possível analiticamente. Portanto utiliza-se o método de Monte Carlo, que envolve a criação de réplicas dos dados por meio de amostragem aleatória para obtenção de uma distribuição empírica possibilitando o cálculo do p-valor.

O procedimento detalhado é o seguinte:

- O número total de casos C, é redistribuído aleatoriamente sob H_0 ao longo das regiões a partir de uma distribuição Multinomial $(C, \frac{n_i}{N})$, isto é, considerando probabilidades proporcionais às populações das regiões.

- Em seguida, obtém-se T^* para esse novo conjunto.

- Esse processo é repetido M vezes, gerando uma distribuição empírica da Estatística sob a hipótese inicial.

- Obtém-se o Percentil 95% da distribuição empírica e compara-se o valor de T^* obtido para os dados reais com o percentil, isto é, se $T^* > P_{95}$ rejeitamos, com 5% de significância, a hipótese inicial de que a probabilidade de um indivíduo vir a ser um caso é a mesma em qualquer zona do mapa e o cluster mais verossímil é considerado um cluster.

2.1.5 Scan baseado em população e baseado em expectância

Para Neill (2006), antes de verificar a existência de aglomerados espaciais, é importante considerar a informação conhecida, isto é, a referência (variável referência) n_i correspondente a cada região. No caso do Scan Circular de Kulldorff (1997), n_i representa a população na i-ésima região. Porém, dependendo da situação, podem estar disponíveis informações adicionais, como acerca do número de ocorrências de eventos em vários momentos no tempo, o que resultaria em uma análise ainda mais precisa. Nesse caso, essas contagens históricas podem ser utilizadas para se estimar n_i, que seria definida como a contagem esperada de eventos de interesse para cada região.

Levando isso em consideração, para Neill (2006), é possível seguir duas abordagens para o uso da estatística Scan: a primeira seria a baseada em população na qual, sob H_0, o número esperado de casos em cada região é proporcional à referência, a população; enquanto a segunda seria baseada em expectância na qual, sob H_0, o número esperado de casos em cada região é igual à referência, que pode ser, por exemplo, a média histórica de casos.

No caso populacional, as contagens c_i por região são distribuídas seguindo uma distribuição Poisson com média $\alpha_i n_i$, na qual α_i é a probabilidade definida na subseção 2.1.2.7. No caso da expectância, c_i segue a mesma distribuição com algumas diferenças. Uma é a referência, como já foi dito, a outra é a respeito de α_i, que nesse caso, é interpretado como o risco relativo, ou seja, a razão entre a contagem c_i e a contagem esperada (média histórica obtida por meio de séries temporais).

As hipóteses para a abordagem com base em expectância são:

- $H_0 : \left\{ \quad \alpha_z = \alpha_{\bar{z}} = 1, \text{para toda zona } z. \right.$

- $H_1 : \left\{ \begin{array}{l} \text{existe uma zona } z \text{ tal que } \alpha_z > 1. \\ \alpha_{\bar{z}} = 1. \end{array} \right.$

Caso H_0 não seja rejeitada concluímos que as contagens observadas superiores às contagens esperadas foram obra do acaso. Se rejeitarmos H_0 então existe um cluster significativo, isto é, risco relativo significativamente maior que 1.

Se pensarmos em termos de ocorrência de uma doença como câncer podemos dizer que o risco de uma pessoa da zona z desenvolver câncer é significativamente maior do que alguém fora dela.

Como o modelo baseado em população já foi mostrado, vamos demonstrar o teste para o caso de expectância utilizando a distribuição de Poisson.

2.1.5.1 Modelo Poisson baseado em expectância

Seja c_i o número de contagens por região distribuído como $c_i \sim Poisson(\alpha_i n_i)$. Então, considerando as hipóteses já descritas, como não há parâmetros a se estimar na hipótese nula, a verossimilhança sob H_0 é:

$$L_0(\mathbf{c}; n_i) = \prod_{i=1}^{t} \frac{e^{-n_i}(n_i)^{c_i}}{c_i!} \tag{2.13}$$

Agora sob H_1 é necessário obter $\hat{\alpha}_z$:

$$\begin{aligned} L_1(\mathbf{c}; n_i, \alpha_z) &= \prod_{i \in z} \frac{e^{-\alpha_z n_i}(\alpha_z n_i)^{c_i}}{c_i!} \prod_{i \notin z} \frac{e^{-n_i}(n_i)^{c_i}}{c_i!} \\ &= e^{-\alpha_z n_z} \alpha_z^{c_z} e^{-n_{\bar{z}}} \prod_{i \in z} \frac{(n_i)^{c_i}}{c_i!} \prod_{i \notin z} \frac{(n_i)^{c_i}}{c_i!} \\ &= e^{-\alpha_z n_z} \alpha_z^{c_z} e^{-n_{\bar{z}}} \prod_{i=1}^{t} \frac{(n_i)^{c_i}}{c_i!} \end{aligned} \tag{2.14}$$

Aplicando o logaritmo:

$$\log L_1(\mathbf{c}; n_i, \alpha_z) = -\alpha_z n_z + c_z \log \alpha_z - n_{\bar{z}} + \sum_{i=1}^{t} log \frac{(n_i)^{c_i}}{c_i!} \tag{2.15}$$

Derivando em relação a α_z, e igualando a zero:

$$\begin{aligned} \frac{\partial \log L_1(\mathbf{c}; n_i, \alpha_z)}{\partial \alpha_z} &= -n_z + \frac{c_z}{\alpha_z} = 0 \\ n_z &= \frac{c_z}{\alpha_z} \\ \hat{\alpha}_z &= \frac{c_z}{n_z} \end{aligned} \tag{2.16}$$

Então define-se a razão de verossimilhança λ dessa forma:

$$\begin{aligned}\lambda = \sup_{\alpha_z>1}\frac{L_1}{L_0} &= \frac{\sup_{\alpha_z>1}\alpha_z^{c_z}e^{-\alpha_z n_z}e^{-n_{\bar{z}}}\prod_{i=1}^{t}\frac{(n_i)^{c_i}}{c_i!}}{\prod_{i=1}^{t}\frac{e^{-n_i}(n_i)^{c_i}}{c_i!}} \\ &= e^{(-\hat{\alpha}_z n_z - n_{\bar{z}} + N)}\hat{\alpha}_z^{c_z}, \\ &= e^{(-\hat{\alpha}_z n_z + n_z)}\hat{\alpha}_z^{c_z}.\end{aligned} \tag{2.17}$$

Agora substituindo as estimativas dos parâmetros temos:

$$\lambda = \begin{cases} e^{(n_z-c_z)}\left(\frac{c_z}{n_z}\right)^{c_z}, \text{se } c_z > n_z. \\ 1, \text{caso contrário.}\end{cases} \tag{2.18}$$

Como é mais comum o uso da log-razão de verossimilhança:

$$\log\lambda = \begin{cases} (n_z - c_z) + c_z(\log c_z - \log n_z), \text{ se } c_z > n_z. \\ 0, \text{ caso contrário.}\end{cases} \tag{2.19}$$

Portanto a estatística do teste, do mesmo modo que na outra abordagem, é $T^* = \log\sup_z \lambda$.

2.1.5.2 Diferenças entre as abordagens

Para Neill (2006), caso haja acesso às contagens esperadas, o Scan baseado em expectância deve ser usado, pois será mais poderoso com respeito a detecção de clusters. Em contrapartida, na ausência desses valores esperados, o mais indicado seria o Scan baseado em população.

Além da disponibilidade dos dados, outro ponto importante é que essas técnicas se comportam de forma diferente em determinadas situações. Por exemplo, se as contagens observadas por meio da janela de corte são bem maiores que o esperado, então o teste baseado em expectância irá rejeitar H_0, concluindo que as contagens são significativas. Por outro lado, o teste com base em população só irá chegar à mesma conclusão se houver variação espacial na quantidade de aumento, ou seja, se esse for constante comparando dentro e fora de qualquer zona z/janela de corte o teste não notaria.

Logo, como este estudo assume que podemos estimar de forma precisa as contagens esperadas, adotou-se o paradigma do Scan baseado em expectância.

2.1.6 Scan Touchard baseado em expectância

2.1.6.1 Motivação

A distribuição de Poisson costuma ser usada pelos métodos Scan em razão de sua forma simples e também por, em muitos casos, ser adequada para representar contagens de

eventos independentes ao longo do espaço. O que a torna perfeitamente capaz de modelar observações que cumpram suas suposições. Todavia, essa distribuição não produz um bom modelo na ausência de alguma das suposições do modelo, como:

- superdispersão : ocorre quando há uma variabilidade nos dados maior do que o esperado, a variância do modelo é maior que sua média.

- subdispersão : acontece no caso em que a variabilidade é menor do que o esperado, a variância é menor que a média.

- excesso de zeros : ocorrem, por exemplo, em eventos raros; geralmente quando o modelo é uma mistura; podem ser tanto zeros amostrais, quanto estruturais.

  - zeros amostrais: no caso em que o número de contagens é zero. Por exemplo: se estivermos modelando a quantidade de uma certa espécie de planta por região, esse zero significaria que a j-ésima região não tinha nenhuma planta dessa espécie, apesar do fato de que a mesma pode ser encontrada nesse tipo de vegetação.

  - zeros estruturais: o número de contagens é zero, mas, aproveitando o exemplo anterior, isso ocorre porque naquela região não há condições necessárias para a existência daquela espécie, e portanto um valor diferente de zero é impossível.

O problema é que a distribuição Poisson possui um único parâmetro, λ, que é tanto sua média quanto variância, tornando-a inadequada para modelos com as características acima. Diversas distribuições foram propostas para lidar com isso, como:

- Binomial Negativa: pode lidar com superdispersão, dado o fato que ela possui um parâmetro a mais, podendo ser usado para ajustar a variância de forma independente da média.

- Poisson Generalizada: É outra solução que tem um parâmetro a mais em relação à Poisson, trazendo a possibilidade de modelagem de superdispersão.

- Poisson Dupla: É a combinação exponencial de Poisson, com o intuito de modelar os dois problemas de dispersão.

- Conway-Maxwell-Poisson: Possui dois parâmetros, modela os dois problemas. Quando o parâmetro $v = 1$ equivale à Poisson.

- Poisson Inflada de Zeros (ZIP) : É indicada para dados de contagem com abundância de zeros. Consiste em uma mistura: assume 0, ou segue uma Poisson.

- Binomial Inflada de Zeros (ZIB) : É indicada para dados de contagem com abundância de zeros. Consiste em uma mistura: assume 0, ou segue uma Binomial.

Extensões da estatística Scan com essas misturas de distribuições como Scan-ZIP e Scan-ZIP+EM de Cançado et al. (2014) e a Scan Bayesiana ZIB (BZIB) de Cançado et al. (2017) não resolvem todos os três problemas. Existem algumas dessas misturas de distribuições que o fazem como a ZINB (Binomial Negativa Inflada de Zeros), porém é possível obter o mesmo resultado com a distribuição Touchard, só que de forma mais simples, por não ser uma mistura, e com importantes propriedades como fazer parte da família exponencial e da de distribuições de séries de poder. Por esse motivo esse trabalho tem como foco o desenvolvimento de uma estatística Scan Touchard. Antes de mais detalhes vamos apresentar a distribuição e algumas de suas propriedades.

2.1.6.2 Distribuição Touchard

Consiste em uma generalização da distribuição de Poisson com um parâmetro extra proposta por Matsushita et al. (2018). Foi proposta como uma possibilidade para modelagem de dados de contagem não-poisson, tendo recebido seu nome em razão de sua associação com os polinômios de Touchard.

Seja X uma variável aleatória inteira não-negativa, então sua distribuição de probabilidade é:

$$P(X = x) = \frac{\lambda^x (x+1)^\delta}{x!\tau(\lambda, \delta)}, \tag{2.20}$$

em que $x \in \mathbb{N}$, $\lambda > 0$ e $\delta \in \mathbb{R}$; λ e δ são os parâmetros da distribuição e

$$\tau(\lambda, \delta) = \sum_{j \in \mathbb{N}} \frac{\lambda^j (j+1)^\delta}{j!} \tag{2.21}$$

é a função que normaliza a expressão anterior.

Com relação a esperança e variância, temos:

$$\mu = \mathrm{E}(X) = \mathrm{E}\left[\left(\frac{X+2}{X+1}\right)^\delta\right]\lambda, \tag{2.22}$$

$$\sigma^2 = Var(X) = \mathrm{E}\left[(X+1)\left(\frac{X+2}{X+1}\right)^\delta\right]\lambda - \mu^2. \tag{2.23}$$

Uma maneira de se medir a associação da média com a variância é através da razão:

$$r = \frac{\sigma^2}{\mu} = \frac{\mathrm{E}\left[(X+1)\left(\frac{X+2}{X+1}\right)^\delta\right]}{\mathrm{E}\left[\left(\frac{X+2}{X+1}\right)^\delta\right]} - \mu. \tag{2.24}$$

Se $\delta = 0$, $r = 1$ e a distribuição de $X \sim$ Poisson (λ). No caso de $\delta > 0$, temos $r < 1$ apontando subdispersão e para $\delta < 0$, $r > 1$ implicando superdispersão.

2.2 Scan Touchard

Pelos motivos citados em outras seções, o método com base em população não será abordado neste trabalho. Por simplicidade, a estatística Scan Touchard será desenvolvida a partir do método baseado em expectância.

Ao se buscar um cluster espacial, queremos encontrar uma zona z onde o processo de geração de casos tem uma intensidade basal (risco relativo) maior em z do que fora de z. Agora, suponhamos que, além do aumento na intensidade, a contagem de casos sofra alguma perturbação. Por exemplo, é comum que para algumas doenças haja problemas de subnotificação, o que diminui o número de casos reportados, mesmo que o número real de casos seja grande. A distribuição Touchard seria capaz de modelar essa situação adotando $\delta < 0$, por exemplo. Isso motiva o teste apresentado a seguir, no qual a distribuição Touchard foi reparametrizada de forma que o parâmetro λ equivale a αn_i com α representando a intensidade do processo subjacente gerador de casos e n_i é a contagem esperada de eventos de interesse por região.

$$H_0 : c_i \sim Touchard(n_i, 0) \tag{2.25}$$

$$H_1 : \begin{cases} c_i \sim Touchard(\alpha n_i, \delta), \ se\ i \in z\ (\alpha > 1, \delta \neq 0) \\ c_i \sim Touchard(n_i, 0), \ se\ i \notin z \end{cases} \tag{2.26}$$

Isto é, estamos interessados na zona z cuja intensidade do processo (α) é maior que 1. Então sob H_0 a verossimilhança é

$$L_0(c, n) = \prod_{i=1}^{n} \frac{n_i^{c_i} e^{-n_i}}{c_i!} \tag{2.27}$$

e sob H_1 é

$$L_1(c, n, \alpha) = \prod_{i \in z} \frac{(\alpha n_i)^{c_i}(c_i + 1)^{\delta}}{ci!\tau(\alpha n_i, \delta)} \prod_{i \notin z} \frac{n_i^{c_i} e^{-n_i}}{c_i!}. \tag{2.28}$$

Em seguida tomando logaritmo de L_1 tem-se

$$\log(L_1) = c_z \log(\alpha) + \sum_{i=1}^{n}(c_i \log n_i) + \delta \sum_{i \in z} \log(c_i + 1) - \sum_{i=1}^{n} \log(c_i!) - \sum_{i \in z} \log(\tau(\alpha n_i, \delta)) - n_{\bar{z}} \tag{2.29}$$

Para achar estimativas de α e δ é necessário usar um método numérico. A razão de verossimilhança será dada por

$$\lambda_z = \sup_{\alpha > 1} \frac{L_1}{L_0} = \sup_{\alpha > 1} \frac{\alpha^{c_z} \prod_{i \in z}(c_i + 1)^{\delta}}{e^{-n_z} \prod_{i \in z} \tau(\alpha n_i, \delta)} = \frac{\hat{\alpha}^{c_z} \prod_{i \in z}(c_i + 1)^{\hat{\delta}}}{e^{-n_z} \prod_{i \in z} \tau(\hat{\alpha} n_i, \hat{\delta})} \tag{2.30}$$

Já a log-razão de verossimilhança é

$$\begin{cases} c_z \log(\hat{\alpha}) + n_z + \delta \sum_{i \in z} \log(c_i + 1) - \sum_{i \in z} \log(\tau(\hat{\alpha} n_i, \hat{\delta})), \text{ se } \hat{\alpha} > 1 \\ 0, \text{ caso contrário} \end{cases} \tag{2.31}$$

Logo, assim como no caso da Poisson, podemos computar a log-razão de verossimilhança para cada zona candidata e testar a significância da zona mais verossímil. Para tanto, adota-se procedimento semelhante ao anterior, gerando réplicas aleatórias dos dados sob H_0 e obtendo, para cada uma delas o valor da estatística de teste. Em seguida compara-se o valor da estatística de teste (T^*) para os dados observados com sua distribuição empírica sob H_0.

3 Resultados

O método Scan Touchard foi desenvolvido usando a linguagem R. Os códigos podem ser disponibilizados mediante solicitação pelo email `juliano6@hotmail.com`.

Esse método teve seu desempenho avaliado e comparado com o Scan Poisson baseado em expectância (Neill (2006)). Para isso, foram utilizados conjuntos de dados gerados artificialmente por meio das distribuições Poisson e Touchard, assim como dados reais.

3.1 Simulações

Para comparar os métodos Scan Poisson e Touchard foi criado um mapa hipotético composto de 203 regiões vizinhas com formatos hexagonais. Como foi utilizada a abordagem baseada em expectância, n_{max} foi definido de forma diferente. Isto é, o tamanho máximo de n_z, n_{max}, foi 15% do total de casos esperados N.

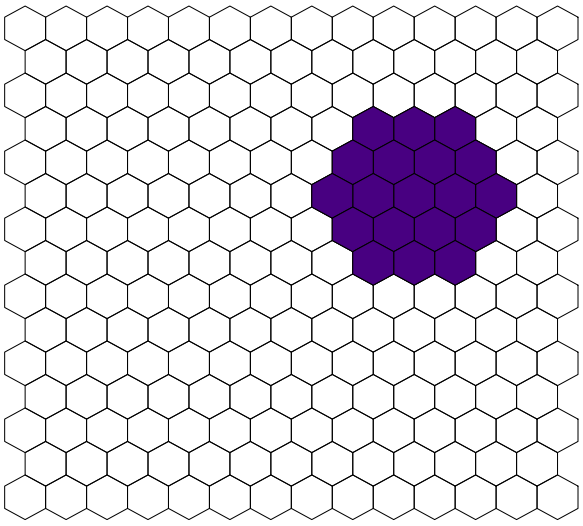

Figura 2 – Mapa com a localização do cluster verdadeiro

As 19 regiões destacadas na figura 2 foram escolhidas para formar o cluster verdadeiro.

Para cada região i, n_i, número de casos esperados, foi definido de forma aleatória usando uma distribuição uniforme. Enquanto c_i, número de casos observados, foi gerado a partir de uma $Touchard(\alpha n_i, \delta)$, se $i \in z$ e $Poisson(n_i)$, (equivalente à $Touchard(n_i, 0)$) se $i \notin z$ de forma que as regiões que compõem o cluster tivessem intensidade $\alpha > 1$, ou seja, para que o cluster seja de alta incidência.

Para essa análise foram realizadas 1000 simulações, isto é, foram gerados 1000 distribuições aleatórias de casos observados c_i e para cada um desses repetiu-se processo de

detecção e verificação de significância do cluster. Os valores de n_i foram os mesmos para cada uma das simulações. Além disso, em cada simulação, toma-se apenas o cluster mais significativo.

Foram analisados vários cenários com o parâmetro α variando ao longo do vetor $(1.4; 1.6; 1.8; 2)$ e δ assumindo os valores $(-4; -2; 0; 2)$. Para cada um desses cenários os resultados foram comparados através de gráficos e métricas de desempenho.

Os gráficos a seguir mostram as regiões detectadas como cluster mais significativo, ao longo das 1000 simulações, pelos dois métodos conforme variam os parâmetros α e δ. Quanto mais escuro o hexágono que contém o centróide de determinada região, mais vezes essa região fez parte do cluster significativo durante as simulações.

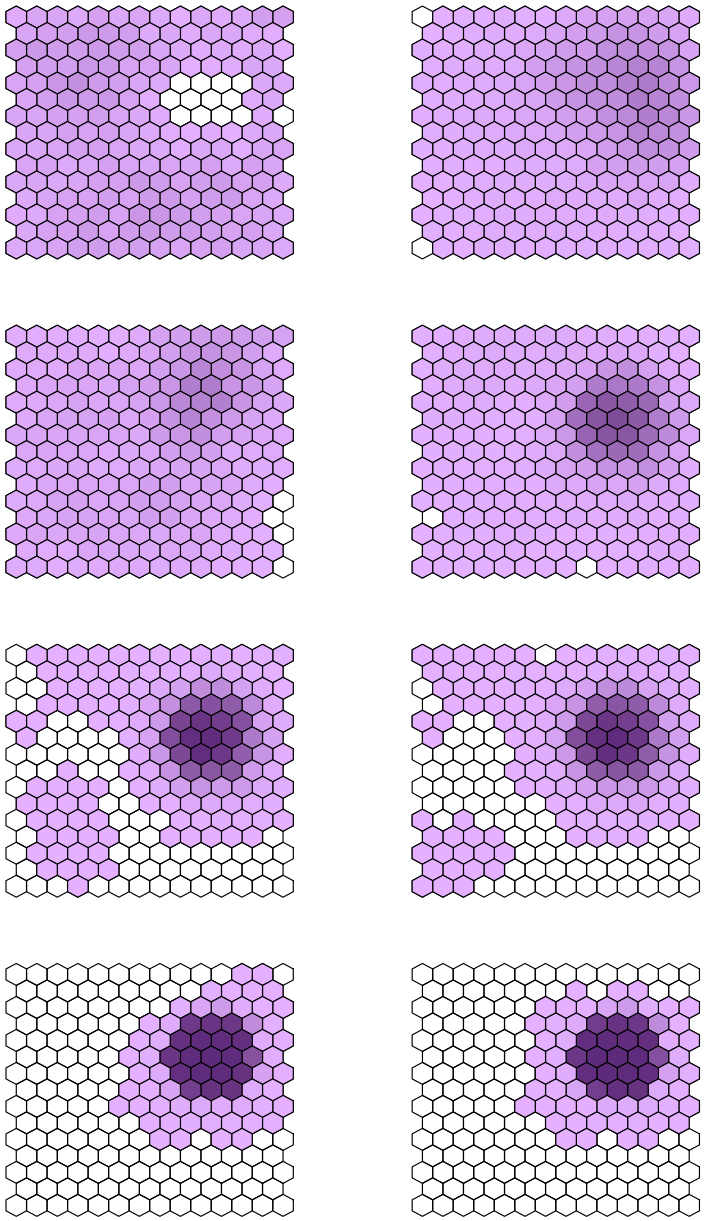

Figura 3 – Clusters detectados pelo Scan Poisson (esquerda) e clusters detectados pelo Scan Touchard (direita) para $\alpha = 1.4$ e $\delta = (-4; -2; 0; 2)$, respectivamente.

No caso da figura 3 para $\delta = -4$ nenhum dos métodos foi eficaz na detecção do cluster verdadeiro. No entanto, percebe-se que enquanto o Scan Touchard possui um leve sombreamento próximo à zona correta, o Poisson detectou praticamente o oposto do que deveria. Conforme δ vai aumentando ambos os métodos passam a enxergar melhor as regiões do cluster verdadeiro, no entanto, ainda que em menor quantidade, continuam obtendo vários falsos positivos. Importante notar que para o caso em que o scan Touchard equivale ao Scan Poisson ($\delta = 0$) ambos se enganam, pois em algumas das simulações, encontram um cluster que não contêm nenhuma das regiões pertencentes ao cluster verdadeiro.

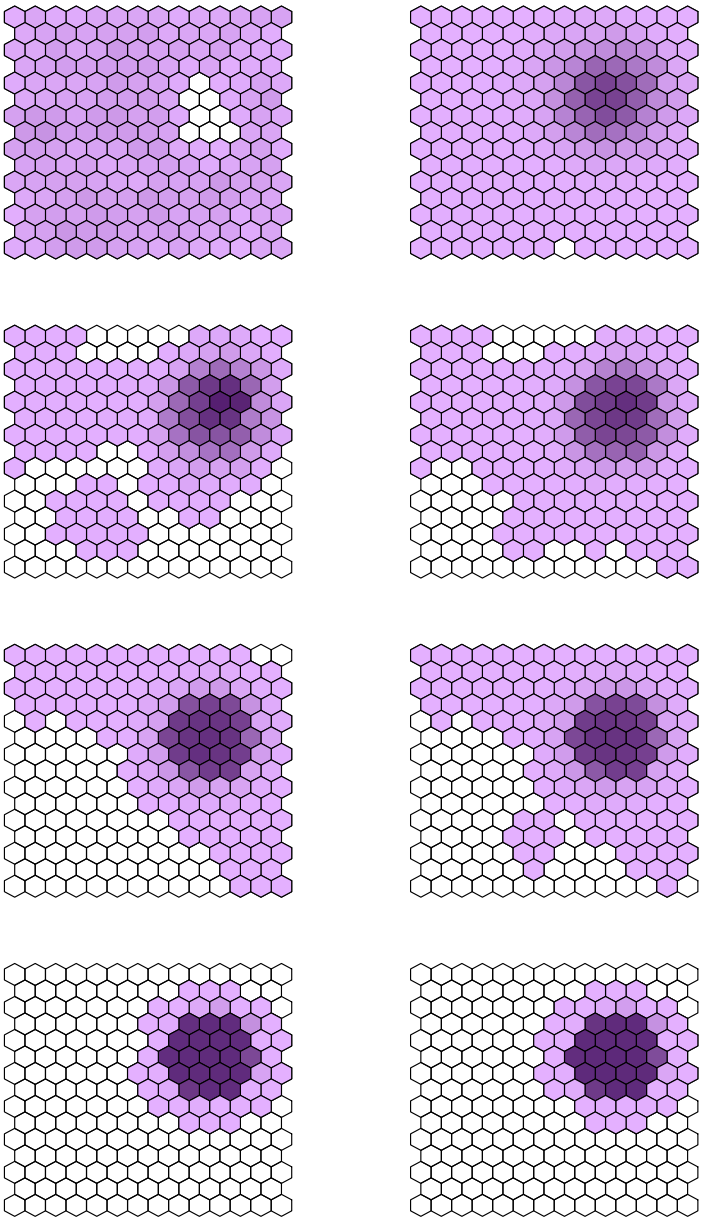

Figura 4 – Clusters detectados pelo Scan Poisson (esquerda) e clusters detectados pelo Scan Touchard (direita) para $\alpha = 1.6$ e $\delta = (-4; -2; 0; 2)$, respectivamente.

Na figura 4 nota-se que para $\delta = -4$ o Scan Poisson novamente não consegue distinguir o sinal do ruído, detectando muito mais regiões fora do que dentro do agrupamento verdadeiro. Olhando para $\delta = 0$, percebemos que com o aumento da intensidade do processo, isto é $\alpha = 1.6$, o Poisson teve um resultado ligeiramente melhor, quando comparado com o Touchard.

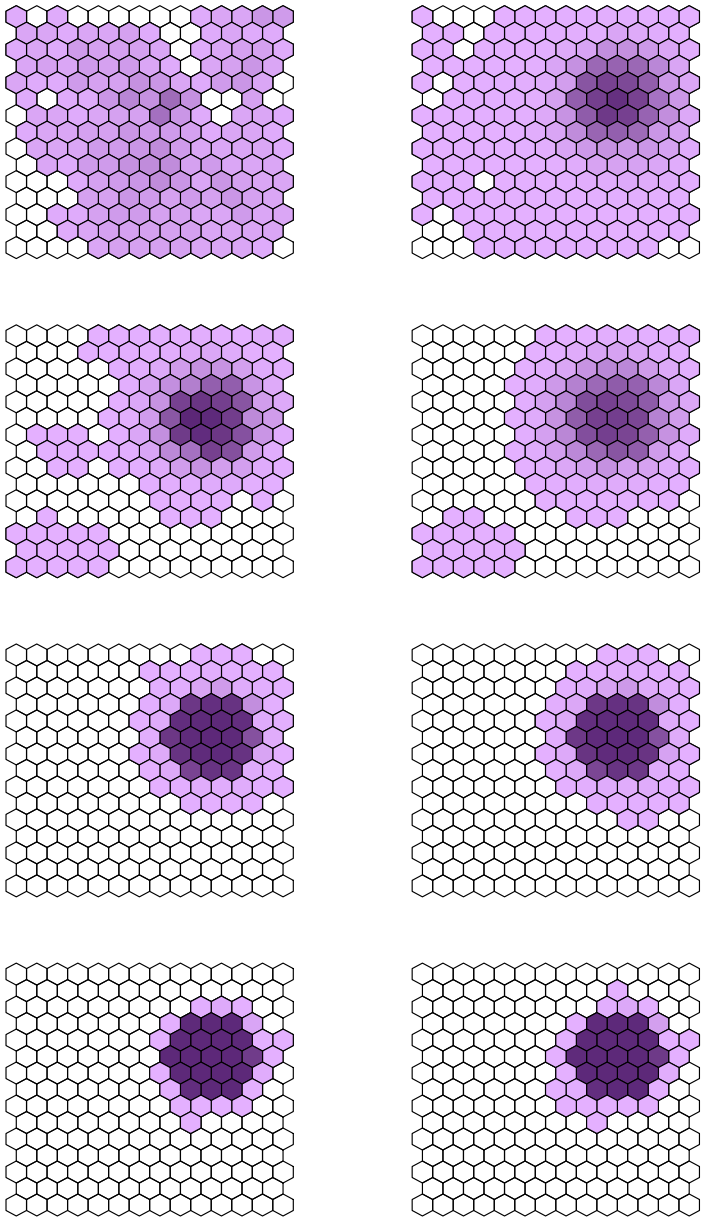

Figura 5 – Clusters detectados pelo Scan Poisson (esquerda) e clusters detectados pelo Scan Touchard (direita) para $\alpha = 1.8$ e $\delta = (-4; -2; 0; 2)$, respectivamente.

Apesar do desempenho longe do ideal, a figura 5 mostra mais uma vez a superioridade do Scan Touchard para valores negativos. Os outros casos apresentam equivalência entre os métodos.

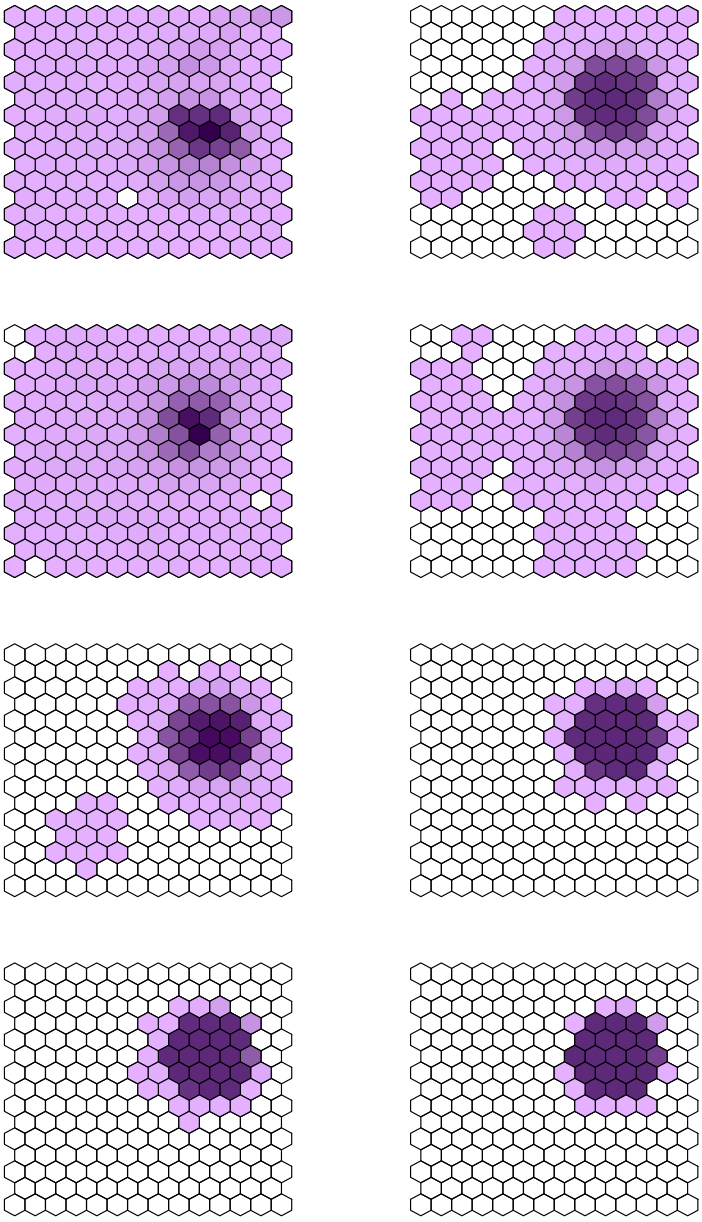

Figura 6 – Clusters detectados pelo Scan Poisson (esquerda) e clusters detectados pelo Scan Touchard (direita) para $\alpha = 2$ e $\delta = (-4; -2; 0; 2)$, respectivamente.

A figura 6 traz um comportamento um pouco diferente, pois para $\alpha = 2$ o Scan Touchard se sobressai para todos os valores de δ testados. Isso pode indicar que para processos subjacentes muito fortes o Scan Poisson acaba detectando mais ruído.

Com relação às medidas de desempenho temos:

$$\text{Poder do teste} = \frac{\sum_{i=1}^{1000} I(T^* > P_{95})}{1000} \tag{3.1}$$

Para calcular o poder do teste utilizam-se as estatísticas T^* que foram significativas, isto é, que cumpriram a condição $T^* > P_{95}$. Essa medida representa a proporção de testes para

os quais H_0 foi rejeitada nas 1000 simulações.

$$\text{Sensibilidade} = \frac{\sum\limits_{i \in (\hat{z}_j \cap v)} n_i}{\sum\limits_{i \in (v)} n_i} \tag{3.2}$$

$$\text{VPP} = \frac{\sum\limits_{i \in (\hat{z}_j \cap v)} n_i}{\sum\limits_{i \in (\hat{z}_j)} n_i} \tag{3.3}$$

- cluster detectado na j-ésima simulação: $\hat{z}_j$

- cluster verdadeiro: v

Ambas as métricas acima visam comparar o cluster encontrado ao verdadeiro por meio do número de casos esperados. A Sensibilidade representa a proporção do cluster verdadeiro que foi detectada pelo método. Já o VPP (Valor Preditivo Positivo) representa a proporção do cluster detectado que faz parte do verdadeiro. Essas duas fórmulas são calculadas para cada uma das simulações, logo para obter informações resumidas acerca de todas as simulações calcula-se por fim a Sensibilidade e o VPP médios.

Além dessas medidas, como observado anteriormente no Scan Touchard, para achar estimativas de α e δ é preciso utilizar métodos de otimização tornando esse método computacionalmente mais oneroso do que o Poisson, uma vez que não é possível computar as estimativas analiticamente. Já para o Scan Poisson estima-se o parâmetro α através da razão $\frac{c_z}{n_z}$. Logo, também foi calculada a média dessas estimativas ($\bar{\alpha}$ e $\bar{\delta}$) ao longo de todas as simulações de forma a verificar se os valores estavam próximos dos parâmetros verdadeiros.

Os resultados podem ser observados na tabela 1.

Tabela 1 – Métricas de Desempenho

α	δ	VPP Médio		Sensibilidade Média		Poder		$\bar{\alpha}$		$\bar{\delta}$
		Touchard	Poisson	Touchard	Poisson	Touchard	Poisson	Touchard	Poisson	
1.4	**-4**	0.306	0.029	0.216	0.006	0.328	0.197	1.971	2.201	-3.301
	-2	0.582	0.256	0.472	0.141	0.264	0.201	1.957	1.872	-2.785
	0	0.816	0.818	0.764	0.770	0.451	0.471	1.508	1.608	0.691
	2	0.935	0.933	0.929	0.925	0.995	0.992	1.397	1.751	2.283
1.6	**-4**	0.600	0.127	0.477	0.022	0.503	0.138	2.038	2.050	-4.120
	-2	0.791	0.767	0.696	0.534	0.289	0.179	1.964	1.713	-2.319
	0	0.857	0.872	0.858	0.863	0.821	0.834	1.636	1.701	0.422
	2	0.971	0.969	0.976	0.981	1.000	1.000	1.543	1.925	2.595
1.8	**-4**	0.688	0.322	0.601	0.067	0.545	0.067	2.153	2.619	-4.225
	-2	0.796	0.822	0.750	0.662	0.410	0.287	2.153	1.738	-2.377
	0	0.936	0.942	0.933	0.935	0.987	0.986	1.741	1.831	0.613
	2	0.981	0.987	0.986	0.986	1.000	1.000	1.741	2.164	2.576
2.0	**-4**	0.867	0.600	0.805	0.252	0.943	0.519	2.167	1.827	-4.212
	-2	0.878	0.893	0.816	0.780	0.951	0.943	2.003	2.359	-1.402
	0	0.978	0.979	0.975	0.974	1.000	1.000	1.939	2.071	0.617
	2	0.993	0.991	0.995	0.996	1.000	1.000	1.911	2.292	2.874

Percebe-se pela tabela 1 que conforme aumentamos os valores dos parâmetros, todas as métricas de performance também aumentam. Isto é, os métodos Scan passam a detectar melhor as regiões pertencentes ao cluster verdadeiro. Em especial, no caso do α, esse comportamento era esperado uma vez que quanto maior ele fica, maior é a proporção de casos observados em relação aos esperados tornando o cluster mais evidente.

Comparando agora os métodos scan, percebe-se que o Scan Touchard se mostra superior ao Poisson na maioria dos cenários com $\delta < 0$ considerando todas as métricas. A diferença é especialmente grande para $\delta = -4$, o que realça a dificuldade da distribuição Poisson ao lidar com amostras com superdispersão. Já para $\delta = 0$ os dois métodos foram parecidos com leve vantagem para o Scan Poisson, o que faz sentido já que a Touchard nesse caso equivale a uma Poisson. Nas amostras em que poderia haver subdispersão os resultados foram similares independente dos parâmetros.

Uma última vantagem do Scan Poisson foi o tempo de execução que foi bem menor que o Scan Touchard, pois o segundo precisa encontrar os parâmetros do modelo por meio de otimização.

Em resumo, o parâmetro α está relacionado a intensidade do processo. Assim, quanto maior o valor de α, mais intenso é o processo adjacente que gera os casos, tornando o cluster, portanto, mais evidente.

Com relação ao parâmetro δ, valores negativos mitigam a força do processo. Assim, mesmo que o α seja elevado, o número observado de casos pode não ser significativamente maior que o esperado sob a hipótese de os dados seguirem distribuição Poisson. Já a abordagem via modelo Touchard é capaz de indentificar essa intensidade elevada nesses casos com mais facilidade.

Por outro lado, quando $\delta > 0$ o processo fica ainda mais intenso, e aparentemente

ambos os modelos são capazes de identificar corretamente o cluster, embora o modelo Poisson sistematicamente superestime o risco relativo.

O modelo Touchard, portanto mostra-se mais adequado em situações em que o sinal do cluster possa estar disfarçado por algum motivo como, por exemplo, subnotificação de casos ou presença de zeros estruturais. Nos outros casos, aparentemente os dois modelos são equivalentes.

Quanto às médias das estimativas ($\bar{\alpha}$ e $\bar{\delta}$), os valores de $\bar{\alpha}$ ficaram bem mais próximos do parâmetro verdadeiro do que os de $\bar{\delta}$.

3.2 Aplicação em Dados Reais

3.2.1 Descrição do conjunto de dados

Com intuito de mostrar a aplicabilidade da técnica proposta na detecção de clusters de alta incidência, foram utilizados dados de óbitos por lesões autoprovocadas voluntariamente (suicídio), por ano do Óbito (1996 a 2017), registrado pelo município de residência no Estado do Acre. As observações foram obtidas pela plataforma TABNET do DATASUS e são oriundas do Sistema de Informações sobre Mortalidade (SIM). Elas podem ser acessadas pelo endereço eletrônico `http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sim/cnv/obt10AC.def`. Já os mapas apresentados nas figuras 7 e 8 foram gerados no software R, a partir do shapefile do estado do Acre disponibilizada pelo IBGE no endereço eletrônico `ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais`. Isso foi feito através de vários pacotes do R, desde o rgdal para importação do shapefile até o ggplot2 para produzir a ilustração do mapa.

Parte da motivação por trás do uso desses dados é a possível presença de sub-registro de suicídios, uma vez que, de acordo com a Organização Mundial da Saúde (World Health Organization, 2014), erros de classificação e subnotificação são mais prováveis de acontecer nos dados associados à mortalidade por suicídio, em comparação com outras causas de morte. Isso ocorre, por exemplo, pela possibilidade de perda de seguros e direitos por parte de familiares e pelo fato de suicídio ser muitas vezes motivo de vergonha para as famílias.

A forma de óbito em estudo, suicídio, foi definida de acordo com a lista **CID-10** [1]. Mais especificamente pelos códigos da **CID–10: X60 até X84** que estão descritos na tabela 2.

[1] **CID-10** se refere à Classificação Internacional de Doenças e é publicada pela Organização Mundial de Saúde (OMS)

Tabela 2 – Lesões autoprovocadas intencionalmente

Categoria	Descrição
X60	Auto-intoxicação por e exposição, intencional, a analgésicos, antipiréticos e anti-reumáticos, não-opiáceos.
X61	Auto-intoxicação por e exposição, intencional, a drogas anticonvulsivantes [antiepilépticos] sedativos, hipnóticos, antiparkinsonianos e psicotrópicos não classificados em outra parte.
X62	Auto-intoxicação por e exposição, intencional, a narcóticos e psicodislépticos [alucinógenos] não classificados em outra parte.
X63	Auto-intoxicação por e exposição, intencional, a outras substâncias farmacológicas de ação sobre o sistema nervoso autônomo.
X64	Auto-intoxicação por e exposição, intencional, a outras drogas, medicamentos e substâncias biológicas e às não especificadas.
X65	Auto-intoxicação voluntária por álcool.
X66	Auto-intoxicação intencional por solventes orgânicos, hidrocarbonetos halogenados e seus vapores.
X67	Auto-intoxicação intencional por outros gases e vapores.
X68	Auto-intoxicação por e exposição, intencional, a pesticidas.
X69	Auto-intoxicação por e exposição, intencional, a outros produtos químicos e substâncias nocivas não especificadas.
X70	Lesão autoprovocada intencionalmente por enforcamento, estrangulamento e sufocação.
X71	Lesão autoprovocada intencionalmente por afogamento e submersão.
X72	Lesão autoprovocada intencionalmente por disparo de arma de fogo de mão.
X73	Lesão autoprovocada intencionalmente por disparo de espingarda, carabina, ou arma de fogo de maior calibre.
X74	Lesão autoprovocada intencionalmente por disparo de outra arma de fogo e de arma de fogo não especificada.
X75	Lesão autoprovocada intencionalmente por dispositivos explosivos.
X76	Lesão autoprovocada intencionalmente pela fumaça, pelo fogo e por chamas.
X77	Lesão autoprovocada intencionalmente por vapor de água, gases ou objetos quentes.
X78	Lesão autoprovocada intencionalmente por objeto cortante ou penetrante.
X79	Lesão autoprovocada intencionalmente por objeto contundente.
X80	Lesão autoprovocada intencionalmente por precipitação de um lugar elevado.
X81	Lesão autoprovocada intencionalmente por precipitação ou permanência diante de um objeto em movimento.
X82	Lesão autoprovocada intencionalmente por impacto de um veículo a motor.
X83	Lesão autoprovocada intencionalmente por outros meios especificados.
X84	Lesão autoprovocada intencionalmente por meios não especificados.

3.2.2 Analise Exploratória

A unidade federativa do Acre possui 22 municípios, ao longo destes registrou 696 casos de suicídio entre 1996 e 2017. Nesse estudo, os casos de óbitos por suicídio esperados por município (n_i) foram inferidos utilizando modelos preditivos de séries temporais para os casos observados de 1996 até 2016. Isto é, para cada município foram ajustados modelos ARIMA[2], que em seguida são utilizados para prever os casos esperados para 2017.

O processo de modelagem foi feito utilizando-se a função auto.arima() (do pacote forecast do software R) que usa uma variação do algoritmo desenvolvido por Hyndman and Khandakar (2008) que combina testes de raiz unitária, minimização de AICc[3] e EMV [4] para obter um modelo ARIMA para cada série. Os modelos ARIMA foram escolhidos para simplificar o estudo, uma vez que a análise desse conjunto de dados tem fins meramente ilustrativos. Para uma análise mais detalhada seriam explorados outros modelos como métodos de alisamento exponencial, métodos não paramétricos e modelos dinâmicos para contagens baixas.

A partir dos modelos encontrados pelo auto.arima() a função forecast faz a previsão dos valores esperados (n_i) para cada município em 2017. Já os casos observados (c_i) são os próprios valores observados para o ano de 2017. As coordenadas dos centróides associados às observações c_i e n_i foram obtidos pelo software QGIS.

A tabela 4 traz as contagens para cada cidade do Acre.

[2] Modelo auto-regressivo integrado de médias móveis
[3] Critério de Informação de Akaike Corrigido
[4] Estimação de Máxima Verossimilhança

Tabela 4 – Casos observados e esperados por município ordenados pela População

	Nome do município	População	n_i	c_i
1	Rio Branco	407319	17.5249	31
2	Cruzeiro do Sul	88358	3.7902	6
3	Sena Madureira	43310	0.7143	4
4	Tarauacá	41976	5.0000	4
5	Feijó	35092	4.1051	4
6	Senador Guiomard	24808	0.6667	0
7	Brasiléia	24274	0.5714	0
8	Plácido de Castro	18862	0.7619	0
9	Xapuri	18685	0.4762	4
10	Marechal Thaumaturgo	18539	0.0001*	0
11	Mâncio Lima	18528	1.1477	1
12	Rodrigues Alves	18502	0.0001*	0
13	Porto Acre	18176	4.0000	1
14	Epitaciolândia	18120	0.4762	2
15	Acrelândia	16304	0.0001*	0
16	Porto Walter	11017	0.4975	0
17	Bujari	10079	0.0175	2
18	Capixaba	9947	0.2857	0
19	Manoel Urbano	8570	4.0000	5
20	Jordão	8140	1.0025	0
21	Assis Brasil	6393	0.1429	0
22	Santa Rosa do Purus	5952	0.0001*	0

* As previsões que resultaram em valor menor que $\epsilon = 0.0001$ foram igualadas a ϵ para evitar problemas numéricos.

Pela tabela 4 é possível notar que Rio Branco, capital do Acre, possui um número bem maior de casos observados e esperados do que todas as outras cidades, o que se justifica, pelo menos em parte, pelo tamanho da sua população (407319) que é mais de quatro vezes maior que a de Cruzeiro do Sul (88358), segunda cidade mais populosa. Algumas regiões como Sena Madureira, Xapuri, Epitaciolândia e Bujari tiveram mais que o dobro de mortes por suicídio do que o esperado o que é interessante, pois observando a figura 7, percebe-se a proximidade espacial entre elas e também de Rio Branco, um comportamento que pode ser indicativo da presença de um cluster. Além disso, nota-se que a metade dos municípios não registrou ocorrência de óbitos por suicídio, o que pode acontecer tanto por ser um evento raro, quanto por problemas de subnotificação.

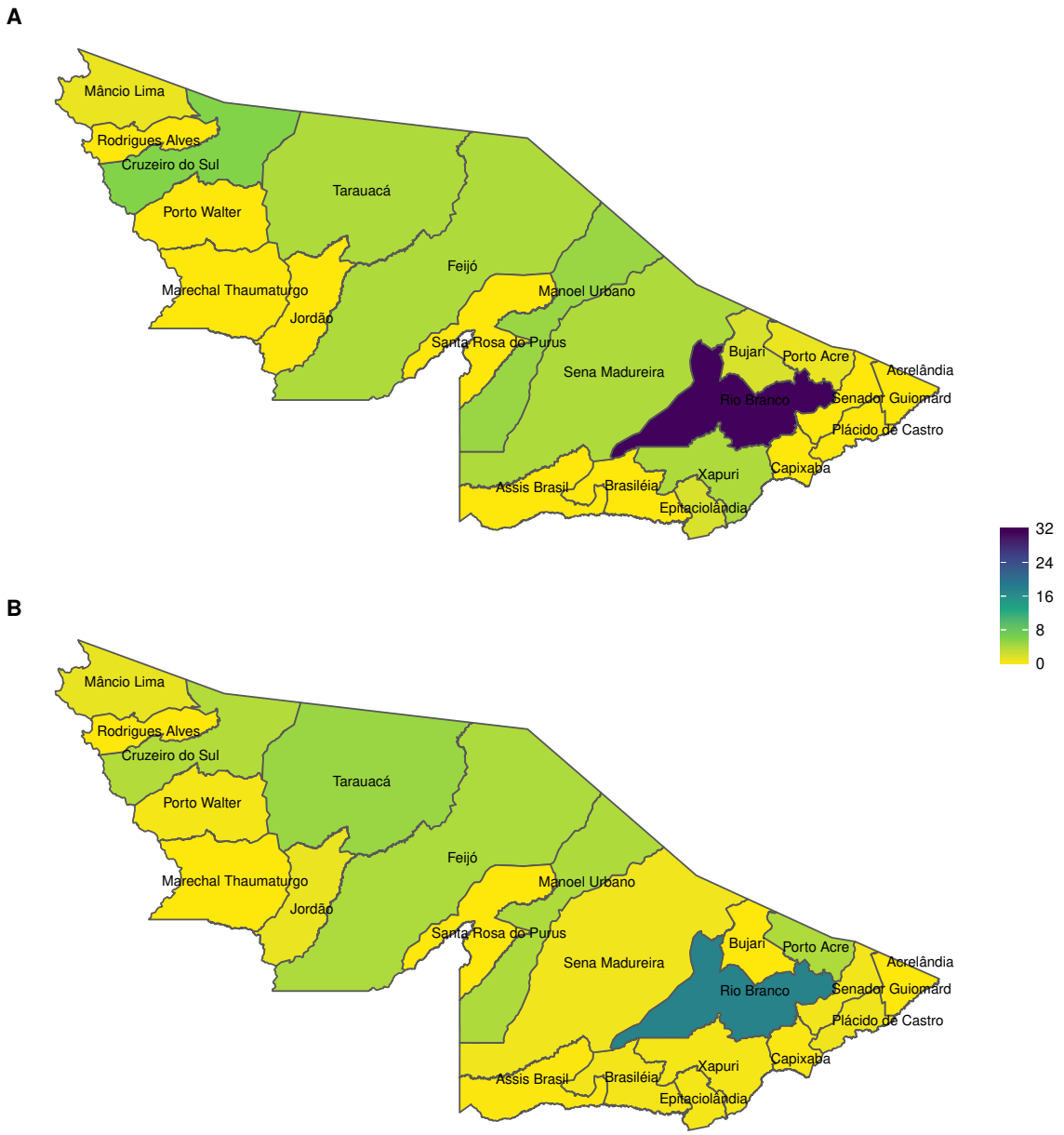

Figura 7 – A: Distribuição de óbitos observados (A) e distribuição de óbitos esperados (B) por lesões autoprovocadas intencionalmente ao longo dos municípios do Acre.

A figura 7 também demonstra que as cidades com maior número de casos observados são aquelas com maior extensão territorial, com exceção de Rio Branco que é um pouco menor.

3.2.3 Scan Touchard e Scan Poisson

Ambos os métodos Scan, quando aplicados encontraram clusters significativos ao nível de 5% como pode ser visto pela figura 8 e tabela 5.

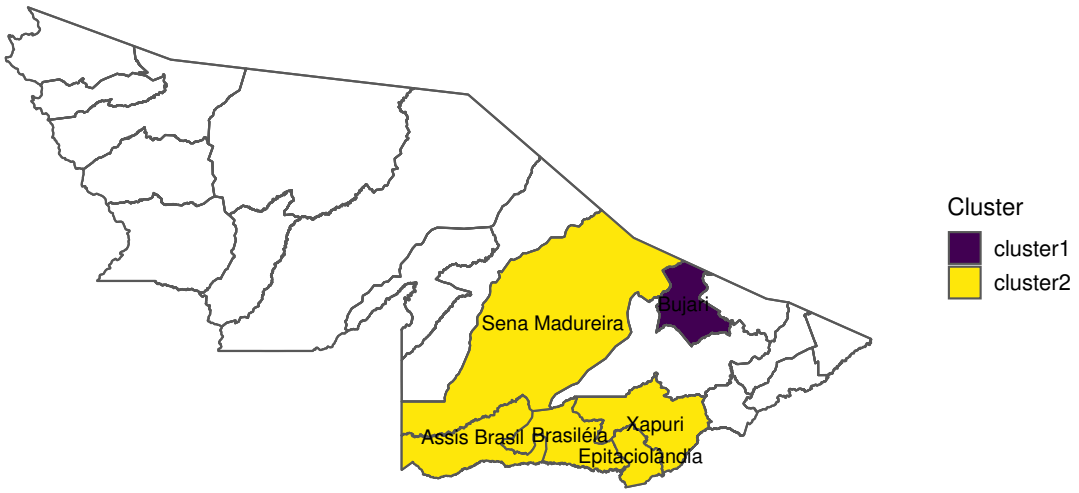

Figura 8 – Clusters de suicídios detectados pelo Scan Poisson em roxo e amarelo; e cluster detectado pelo Scan Touchard em amarelo.

Tabela 5 – Resultados Scan Touchard e Scan Poisson

№	de regiões	c_z	n_z *	$\hat{\alpha}_z$		$\hat{\delta}_z$	log λ_z		P-valor	
				T	P		T	P	T	P
1	1	2	0.02	-**	114.01	-**	-**	7.49	-**	0.01
2	5	10	2.38	8.42	4.20	-2.20	7.20	6.73	0.02	0.03

** - representa ausência de informação
* n_{max} para os dados reais foi de 25% de N.

O Scan Touchard detectou apenas um cluster significativo (p-valor = 0.02) que é igual ao cluster 2 detectado pelo Poisson, ilustrado na figura 8. O Scan Poisson percebeu dois clusters significativos: o primeiro (p-valor = 0.01) contém apenas Bujari; e o segundo (p-valor = 0.03) é formado pelas cidades de Brasiléia, Epitaciolândia, Xapuri, Assis Brasil e Sena Madureira. É interessante notar que essas regiões estão no entorno de Rio Branco, e apesar desse município ter apresentado quase o dobro de casos observados em relação aos casos esperados, não foi selecionado como parte do cluster.

c_z e n_z representam, respectivamente, os 10 suicídios ocorridos em Brasiléia, Epitaciolândia, Xapuri, Assis Brasil e Sena Madureira em 2017 e os 2.38 suicídios que esperava-se ocorrer nessa zona nesse ano. Logo, pode-se dizer que houve mais de quatro vezes mais óbitos por essa causa do que o previsto.

Agora comparando o valor de α_z, que pode ser entendido como a intensidade do processo, o método baseado na distribuição Touchard parece indicar que essa intensidade

pode ser ainda maior do que o percebido pelo Scan Poisson, pois pode haver algum ruído mascarando o verdadeiro aumento na intensidade basal do processo. Isso pode estar ocorrendo devido a uma possível subnotificação dos casos.

A divergência entre o Scan Touchard e o Scan Poisson pode ser justificada em parte também pelo valor da estimativa de δ (-2.20) ser um número negativo. Já que, como verificado na análise de dados simulados, nesse tipo de situação o método Poisson costuma errar consideravelmente a detecção do cluster verdadeiro, seja percebendo falsos negativos ou falsos positivos como na detecção do município Bujari como cluster.

4 Conclusão

Neste estudo foi proposto um método de identificação de agrupamentos espaciais embasado na distribuição Touchard que modela separadamente a média e a variância do modelo com o objetivo de melhor lidar com dados nos quais há a presença de subdispersão, superdispersão e excesso de zeros.

As técnicas aqui comparadas foram baseadas em expectância uma vez que quando há acesso às séries históricas, como foi o caso dos conjuntos utilizados, eles produzem melhores resultados.

A análise de dados simulados mostrou que, apesar de em alguns casos, como para $\delta = 0$, o Scan Poisson ter tido um desempenho ligeiramente superior, o Scan Touchard demonstrou que, na maioria dos cenários, seu desempenho é superior ou equivalente apoiando a idéia de que para as situações citadas ao longo do estudo o Scan Touchard baseado em expectância deve ser utilizado.

A análise dos óbitos por lesões autoprovocadas voluntariamente nos municípios do Acre no ano de 2017, por apresentar problemas de subnotificação e excesso de zeros, o modelo Touchard mostrou-se mais adequado, assim como nas simulações para $\delta < 0$.

Apesar de ambas as técnicas terem encontrado o cluster formado por Brasiléia, Epitaciolândia, Xapuri, Assis Brasil e Sena Madureira, o método Poisson selecionou um possível falso positivo, a cidade de Bujari, isto é, cometeu um erro do tipo 1. Além disso, o Scan Touchard distinguiu-se por identificar melhor a intensidade do processo subjacente α_z responsável pela geração de casos.

Em resumo, os resultados encontrados demonstram a validade do Scan Touchard baseado em expectância para dados de contagem não Poisson.

5 Considerações para trabalhos futuros

Para cada simulação selecionou-se apenas o cluster mais significativo. Para um próximo pode-se escolher todos clusters significativos para cada simulação.

Pode-se também comparar outras técnicas Scan com aquela desenvolvida nesse trabalho, como a Scan-ZIP e Scan-ZIP+EM de Cançado et al. (2014) e outras.

Nesse estudo não trabalhou-se com clusters de baixa incidência, o que pode ser um próximo passo.

Por fim, como os códigos foram implementados no R, é possível no futuro criar um função para o Scan Touchard, talvez dentro do pacote **touchard** já presente no R.

Referências

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

André LF Cançado, Cibele Q da Silva, and Michel F da Silva. A spatial scan statistic for zero-inflated poisson process. *Environmental and ecological statistics*, 21(4):627–650, 2014.

André LF Cançado, Lucas B Fernandes, and Cibele Q da Silva. A bayesian spatial scan statistic for zero-inflated count data. *Spatial Statistics*, 20:57–75, 2017.

JH Chen, C Weng, and HG Chnag. Using space-time scan statistic to detect pertussis and shigellosis outbreaks. In *CSTE Annual Conference*, 2013.

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

Lionel Cucala, Michaël Genin, Florent Occelli, and Julien Soula. A multivariate nonparametric scan statistic for spatial data. *Spatial statistics*, 29:1–14, 2019.

Raquel González, Orvalho J Augusto, Khátia Munguambe, Charlotte Pierrat, Elpidia N Pedro, Charfudin Sacoor, Elisa De Lazzari, John J Aponte, Eusébio Macete, Pedro L Alonso, et al. Hiv incidence and spatial clustering in a rural area of southern mozambique. *PloS one*, 10(7), 2015.

Chris Green, Robert D Hoppa, T Kue Young, and JF Blanchard. Geographic analysis of diabetes prevalence in an urban area. *Social science & medicine*, 57(3):551–560, 2003.

Lan Huang, Martin Kulldorff, and David Gregorio. A spatial scan statistic for survival data. *Biometrics*, 63(1):109–118, 2007.

Rob Hyndman and Yeasmin Khandakar. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, Articles*, 27(3):1–22, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i03. URL `https://www.jstatsoft.org/v027/i03`.

Jacky M Jennings, Frank C Curriero, David Celentano, and Jonathan M Ellen. Geographic identification of high gonorrhea transmission areas in baltimore, maryland. *American journal of epidemiology*, 161(1):73–80, 2005.

Inkyung Jung, Martin Kulldorff, and Ann C Klassen. A spatial scan statistic for ordinal data. *Statistics in medicine*, 26(7):1594–1607, 2007.

Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997. doi: 10.1080/03610929708831995.

Raul Matsushita, Donald Pianto, Bernardo B. De Andrade, Andre Cançado, and Sergio Da Silva. The touchard distribution. *Communications in Statistics - Theory and Methods*, 0 (0):1–11, 2018. doi: 10.1080/03610926.2018.1444177. URL `https://doi.org/10.1080/03610926.2018.1444177`.

JL Naus. Clustering of random points in two dimensions. *Biometrika*, 52(1-2):263–266, 1965a.

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

Daniel B Neill. Detection of spatial and spatio-temporal clusters. In *Tech Rep CMU-CS-06-142, PhD thesis*. Carnegie Mellon University, 2006.

Stan Openshaw, Martin Charlton, Alan William Craft, and JM Birch. Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet*, 331(8580): 272–273, 1988.

Dana P Seidel and Mark S Boyce. Patch-use dynamics by a large herbivore. *Movement ecology*, 3(1):7, 2015.

Marc Souris, Dubravka Selenic, Supaluk Khaklang, Suwannapa Ninphanomchai, Guy Minet, Jean-Paul Gonzalez, and Pattamaporn Kittayapong. Poultry farm vulnerability and risk of avian influenza re-emergence in thailand. *International journal of environmental research and public health*, 11(1):934–951, 2014.

World Health Organization. *Preventing suicide: A global imperative*. World Health Organization, 2014.

April M Zeoli, Jesenia M Pizarro, Sue C Grady, and Christopher Melde. Homicide as infectious disease: Using public health methods to investigate the diffusion of homicide. *Justice quarterly*, 31(3):609–632, 2014.

Zhenkui Zhang, Renato Assunção, and Martin Kulldorff. Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 2010, 2010.