Comparação dos Métodos Scan Circular e Flexível na Detecção de Aglomerados Espaciais de Dengue

José C. S. Melo[1], Ana C. O. Melo[1], Ronei M. Moraes[1],

[1] Laboratório de Estatística Aplicada ao Processamento de Imagens e Geoprocessamento (LEAPIG),
Departamento de Estatística, Universidade Federal da Paraíba
João Pessoa, Paraíba, Brasil
zka07@hotmail.com, anaclaudiaemelo@gmail.com, ronei@de.ufpb.br

Abstract. A detecção de aglomerados espaciais é útil para identificar localidades com valores diferenciados significativos ou não do ponto de vista estatístico em uma região geográfica de interesse. Do ponto de vista epidemiológico, essa detecção auxilia a promover políticas públicas diferenciadas para o combate à uma doença. Neste artigo objetivou-se comparar o desempenho das Estatísticas Scan Circular e Scan Flexível para detecção de aglomerados espaciais usando dados reais de dengue na Paraíba.

Keywords: Métodos de Aglomeração Espacial, Scan Circular, Estatística Scan Flexível, Epidemiologia do Dengue.

1 Introdução

A Epidemiologia é a ciência que estuda uma doença segundo os seus padrões, causas e efeitos. Ela visa prover a base de conhecimento para a promoção e cuidados em saúde de acordo com as especificidades de cada localidade e de sua população específica [1]. Ela pode ainda auxiliar a tomada de decisão em saúde por gestores de modo a prover diferentes políticas de acordo com os dados epidemiológicos de cada localidade [2], elegendo diferentes níveis de prioridade para cada localidade de acordo com a região geográfica na qual ela está inserida [3]. Problemas como esse, remetem ao uso dos métodos de aglomeração espacial, que se utilizam de informações georreferenciadas para identificar localidades com valores diferenciados significativos ou não do ponto de vista estatístico.

Vários métodos para detecção de aglomerados espaciais estão disponíveis na literatura científica. Alguns são baseados em matrizes de proximidade, como o Índice de Moran e a Estatística de Getis & Ord [4]. Outros são baseados em grafos de vizinhança, como a Estatística Scan Circular [5] e a Estatística Scan Flexível [6]. Em situações práticas, os métodos são baseados em metodologias diferentes e, portanto, produzem resultados diferentes. Além disso, não há uma informação de referência para se avaliar quais aglomerados são verdadeiros ou não. Assim, são usadas formas indiretas de avaliação, baseadas por exemplo nos mapas de risco [7].

O mosquito Aedes aegypti, que é o vetor transmissor do dengue, foi detectado nas principais cidades do Brasil na década de 1970, depois de ter sido erradicado na década de 1950 [8]. O combate à doença é mais efetivo se for possível detectar a presença do vetor. Do ponto de vista epidemiológico, isso se concretiza a partir da detecção de aglomerados espaciais da doença na região geográfica em estudo e promovendo políticas públicas direcionadas àquelas sub-regiões.

Portanto, dado que essa detecção é fundamental para direcionar políticas que possam ser efetivas no combate à doença, torna-se necessário usar a melhor metodologia disponível para tal. Tango e Takahashi afirmam que a Estatística Scan Flexível, proposta por eles, classifica os conglomerados não circulares de maneira mais eficiente que a Estatística Scan Circular proposta por Kulldorff [5]. Desse modo, esse artigo visa comparar o desempenho dos métodos Scan Circular e Scan Flexível na detecção de aglomerados espaciais de dengue na Paraíba, usando para isso dados notificados no ano de 2013

2 Métodos

2.1 A Estatística Scan Circular

Considerando a situação em que, a região geográfica em estudo é dividida em m sub-regiões ou geo-objetos (por exemplo, municípios, distritos, bairros, etc). O número de casos na sub-região i é denotado pela variável aleatória N_i com valor observado n_i $(i = 1, \dots, m)$ e $\mathrm{n} = \mathrm{n}_1 + \cdots + \mathrm{n}_\mathrm{m}$. A hipótese H_0 afirma que não há aglomerados espaciais na sub-região i, os N_i são variáveis independentes de Poisson tal que

$$\mathrm{H}_0 : \mathrm{E}(\mathrm{N}_\mathrm{i}) = \xi_\mathrm{i}, \mathrm{N}_\mathrm{i} \sim \mathrm{Poisson}(\xi_\mathrm{i}), \mathrm{i} = 1, \dots, \mathrm{m}, \quad (1)$$

onde $Poisson(\xi)$ denota uma distribuição de Poisson com média ξ, e o ξ_i é o número esperado de casos da sub-região i sob a hipótese nula. Assim calculamos ξ_i como

$$\xi_\mathrm{i} = \mathrm{n}\,\frac{\mathrm{w_i}}{\sum_{\mathrm{k}=1}^{\mathrm{m}} \mathrm{w_k}},\ \mathrm{i} = 1, \dots, \mathrm{m}, \quad (2)$$

onde w_i denota o tamanho da população na sub-região i. Usaremos as coordenadas do centro do município para especificar a posição geográfica de cada sub-região i [9].

Kulldorff [5] propõe, para a situação descrita acima, a estatística scan circular que gera uma janela Z em cada centroide das sub-regiões, que é o ponto centro geométrico de uma sub-região. Neste caso, uma janela consiste no círculo criado a partir do centroide. Para qualquer destes centroides, o raio do círculo varia continuamente desde zero até um percentual da população em risco a ser coberta, estabelecido pelo usuário. Logo se uma janela contém o centroide de uma sub-região, o raio do círculo crescerá até englobar, nesta janela, o percentual da população estabelecido. Seja $\mathrm{Z_{ik}}(\mathrm{k} = 1, \dots, \mathrm{K_i})$ denotando a janela composta pelos

k-1 vizinhos à sub-região *i*, então todas as janelas a serem verificadas pela estatística scan circular estão incluídas no conjunto

$$Z = Z_1 = \{Z_{ik} | 1 \leq i \leq m, 1 \leq k \leq K_i\}.$$

Com a utilização da notação da janela $Z \in Z$, a hipótese nula (1) é expressa como

$$H_0 : E\big(N(Z)\big) = \xi(Z), \text{ para todo } Z \in Z, \quad (3)$$

onde $N()$ e $\xi()$ denotam, respectivamente, a variável aleatória para o número de casos e o número esperado de casos sob H_0 dentro da janela especificada. A hipótese alternativa H_1, afirma que existe pelo menos uma janela $Z \in Z$, para o qual o risco é mais elevado no interior da janela, quando comparado com o exterior da mesma, que é,

$$H_1 : E\big(N(Z)\big) > \xi(Z), \text{ para algum } Z \in Z, \quad (4)$$

É possível calcular a probabilidade de observar o número de casos observados dentro e fora da janela, respectivamente para cada janela de Z. Kulldorff [5] propõe que sob H_0, a estatística da razão de verossimilhança é calculada por

$$\lambda_K = \max_{Z \in Z} \lambda_K(Z) = \max_{Z \in Z} \left(\frac{n(Z)}{\xi(Z)}\right)^{n(Z)} \left(\frac{n-n(Z)}{n-\xi(Z)}\right)^{n-n(Z)} I\left(\frac{n(Z)}{\xi(Z)} > \frac{n-n(Z)}{n-\xi(Z)}\right), \quad (5)$$

onde *n()* denota o número de casos observados dentro da janela especificada e *I()* é a função indicadora.

2.2 A Estatística Scan Flexível

A proposta de Tango e Takahashi [6] é de criar uma janela de forma flexível em cada centroide da sub-região, ligando as sub-regiões vizinhas. O processo, portanto ocorre da seguinte forma, para qualquer sub-região *i*, criamos o conjunto de janelas de forma flexível com comprimento *k*, o que consiste em *k* sub-regiões conectadas incluindo *i* e vamos mover *k* de 1 até o comprimento máximo pré-estabelecido K de vizinhos mais próximos. Para evitar a detecção de um conjunto de forma improvável, as sub-regiões ligadas são restritas aos subconjuntos do conjunto de sub-regiões *i* e os K vizinhos mais próximos à região *i*. Ao final, como na estatística scan circular, várias janelas diferentes de formas arbitraria e sobrepostas umas às outras, são criadas. Seja $Z_{ik(j)}$, $j = 1, \ldots, j_{ik}$ denotando a janela de ordem *j*, a qual é um conjunto de *k* sub-regiões conectadas a partir da sub-região *i*, onde j_{ik} é a janela *j* satisfazendo $Z_{iK(j)} \subseteq Z_{iK}$ para $k = 1, \ldots, K_i = K$. Então, todas as janelas a serem verificados são incluídas no conjunto

$$\mathcal{Z} = \mathcal{Z}_2 = \{\mathbb{Z}_{ik(j)} | 1 \leq i \leq m, 1 \leq k \leq K_i, 1 \leq j \leq j_{ik}\}.$$

De forma mais clara, para qualquer sub-região *i*, a estatística scan circular considera *K* círculos concêntricos que denotamos por Z_1, enquanto que a estatística scan flexível considera *K* círculos concêntricos mais todos os conjuntos de sub-regiões ligados (incluindo a única região *i*) cujos centroides estão localizados dentro do K-ésimo maior círculo concêntrico que denotamos por .

Portanto, o tamanho de $\mathcal{Z}_2$ é muito maior do que o de, que é no máximo mK. Outro ponto a ser destacado é que, o comprimento máximo de K deve ser inferior a 30, pois a carga computacional, devido ao grande número de possíveis combinações de janelas tornar-se-ia muito pesada. O valor de K, padrão no *software* FleXScan [6] é definido como 15.

A janela Z^* que contem a razão de máxima verossimilhança é definida como a *MLC,* ou seja, o aglomerado mais provável. No entanto, não é interessante que Z^*continue aumentando o seu raio, quando já englobou em seu círculo os aglomerados espaciais de maior risco, apenas para atingir o percentual da população pré-estabelecido pelo usuário, pois desta forma, englobará na mesma janela também aglomerados espaciais de menor risco [6, 10]. Tango [11] propôs que no processo de varredura na janela baseada em $\lambda_K(Z)$, exista uma possibilidade de que existam duas janelas disjuntas $\mathbb{Z}_1$ e Z_2 e várias regiões $\{i_1\}, \dots, \{i_r\}$ tal que

$$\lambda_k\left(\{\mathbb{Z}_1, \mathbb{Z}_2, \{i_1\}, \dots, \{i_r\}\}\right) > \max\{\lambda_k(\mathbb{Z}_1), \lambda_k(\mathbb{Z}_2)\}, \quad (6)$$

onde

$$\frac{n(\mathbb{Z}_1)}{\xi(\mathbb{Z}_1)} > 1,\ \frac{n(\mathbb{Z}_2)}{\xi(\mathbb{Z}_2)} > 1 \quad e \quad \frac{n_i}{\xi_i} \leq 1\ (i = 1, \dots, r).$$

Para evitar fenômenos indesejáveis, Tango [5] propôs a seguinte razão de verossimilhança restrita considerando, para cada sub-região, um risco individual:

$$\lambda_T(Z) = \left(\frac{n(Z)}{\xi(Z)}\right)^{n(Z)} \left(\frac{n-n(Z)}{n-\xi(Z)}\right)^{n-n(Z)} I\left(\frac{n(Z)}{\xi(Z)} > \frac{n-n(Z)}{n-\xi(Z)}\right) \prod_{i \in Z} I(p_i < \alpha_1), \quad (7)$$

onde p_i é o p-*value* uni caudal do teste para $H_0 : E(N_i) = \xi_i$ e é dado pelo p-*value* médio.

$$p_i = P_r\{N_i \geq n_i + 1 | N_i \sim \text{Poisson}(\xi_i)\} + \frac{1}{2} P_r\{N_i = n_i | N_i \sim \text{Poisson}(\xi_i)\}, \quad (8)$$

onde a função indicadora $I(p_i < \alpha_1)$ funciona como critério de seleção para os valores que estão na fronteira e α_1 é o nível de significância pré-definido para a região individual. A opção de usar o *p-value* médio é para ajustar o *p-valor* para pequenos ξ_i e contar os resultados. Por conseguinte, tal como no caso da estatística scan flexível original, o valor de p do teste scan flexível com base na razão de verossimilhança (7) é obtido através do teste da hipótese de Monte Carlo.

Para analisar os mapas das estatísticas espaciais, toma-se como referência o mapa de Risco Relativo (RR). Para a obtenção do mapa coroplético (como é denominado na Geografia qualquer mapa colorido) que seja referente ao risco do dengue no estado, se faz necessário o cálculo do RR. Este, por sua vez, permitirá a comparação da informação de diferentes áreas, padronizando os dados e, com isso, retirando o efeito das diferentes populações. Este indicador representa a intensidade da ocorrência de um fenômeno com relação a todas as regiões de estudo [3]. A equação do RR de uma área é denotada por:

$$RR_i = \frac{x_i/n_i}{\sum x_i / \sum n_i}, (i = 1, \dots, m) \quad (9)$$

onde x_i é o número de ocorrência do fenômeno em uma região e n_i é a população dessa região.

3 Resultados

Observando o mapa de RR (Fig. 1), nota-se que nas regiões leste e centro-sul possuem risco relativo baixo (inferior a 0,5 vezes o risco do Estado da Paraíba) em comparação com as demais áreas do mapa. Dos 223 municípios da Paraíba, 52 possuem risco elevado (superior ou igual a 1,5 vezes o risco do Estado da Paraíba) e cerca de 20 municípios não apresentaram risco. Quando se tem risco igual ou próximo a 1 significa que o risco do município é o mesmo que o do Estado. Por conseguinte, risco igual a 0 indica que não foi observado risco no município levando em conta o risco do Estado. Observa-se no mapa de RR que o maior número de municípios que possuem risco alto encontra-se ao oeste, havendo uma pequena concentração na parte central do Estado.

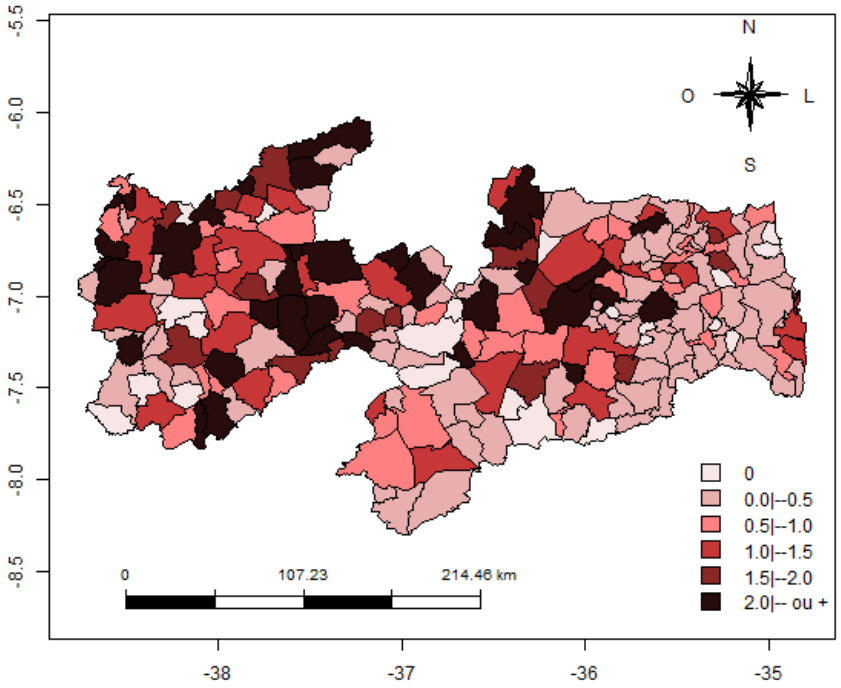

Fig. 1. Mapa do Risco Relativo para o dengue no ano de 2013 na Paraíba.

Analisando o método Scan Flexível (Fig. 2), observa-se que para cada aglomerado significativo que é detectado atribui-se uma cor, essa cor diferencia o aglomerado de forma que possam ser identificados os municípios que se relacionam. Com respeito à epidemiologia do dengue, nota-se a disparidade entre a área litorânea do estado (ao leste) e a área que representa o sertão paraibano (ao oeste), no que diz respeito ao número de municípios detectados. No litoral foram detectados poucos municípios, enquanto a maioria foi detectada no sertão.

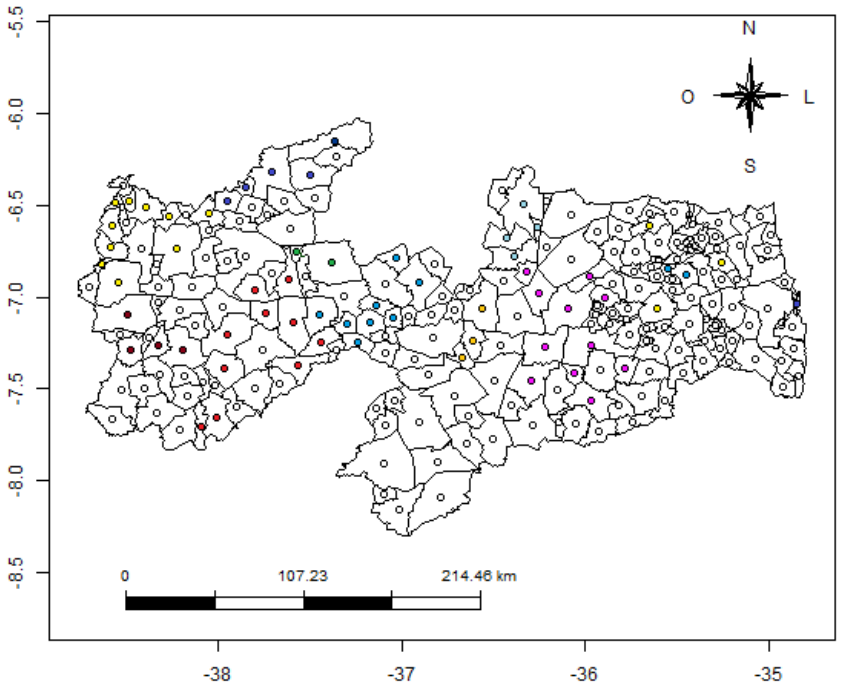

Fig. 2. Resultado da Estatística Scan Flexível para o dengue no ano de 2013 na Paraíba.

A Estatística Scan Circular (Fig. 3) apresentou resultado similar a Flexível, no qual a parte oeste do estado apresentou o maior número de municípios detectados. Todavia, o Scan detectou nove municípios a mais do que o método flexível. Também vale ressaltar que a parte centro-sul do estado, em ambos os métodos, não foram detectados municípios, o que é compatível com o mapa de RR (Fig. 1).

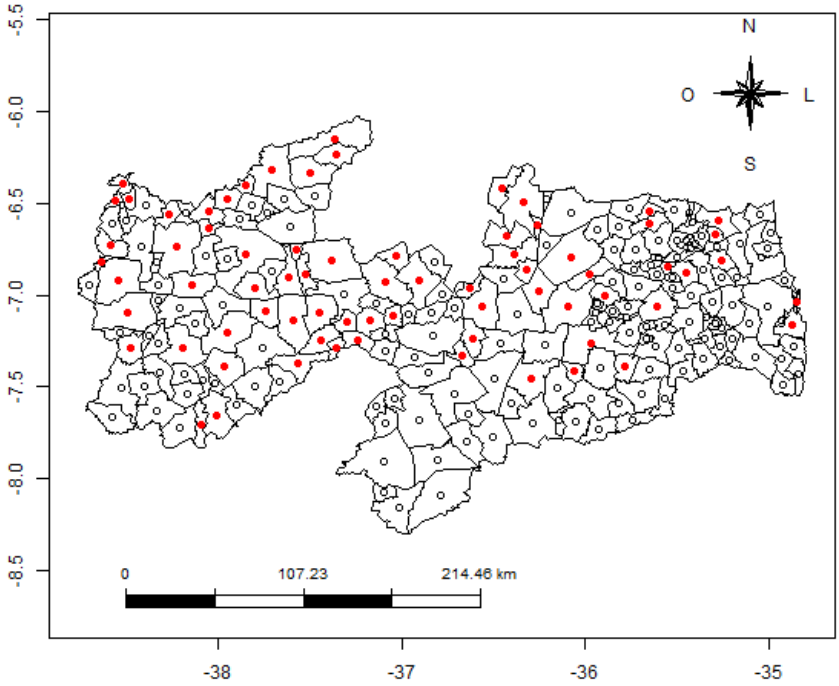

Fig. 3. Resultado da Estatística Scan Circular para o dengue no ano de 2013 na Paraíba.

Levando-se em conta o mapa RR como referência, a Estatística Scan Flexível detectou 63 municípios, sendo 50 deles valores de risco alto, deixando de detectar 2 municípios nessa situação. Ele também incluiu 3 municípios de risco baixo e 10 risco aproximadamente igual ao risco do Estado nesses aglomerados. A Estatística Scan Circular detectou 72 municípios, dos quais 52 deles apresentam risco alto. Observa-se que a Estatística Scan Circular conseguiu detectar todas as áreas de alto risco, quando comparado ao mapa de RR, porém também incluiu 3 municípios de risco baixo (onde apenas um deles foi também detectado pela Estatística Scan Flexível) e 17 risco aproximadamente igual ao risco do Estado (sendo 5 deles também detectados pela Estatística Scan Flexível) nesses aglomerados.

4 Conclusão

Levando em consideração o mapa RR como referência, ambos os métodos superestimaram os número de municípios com valores de risco alto. A Estatística Scan Flexível, entretanto, não detectou dois desses municípios de valores de risco alto,

enquanto a Estatística Scan Circular o fez para todos eles. A Estatística Scan Flexível adicionou aos seus aglomerados 13 municípios com valores de risco baixo ou risco aproximadamente igual ao risco do Estado. Em contrapartida, a Estatística Scan Circular incluiu 20 municípios cujos valores de risco não são altos nos seus aglomerados. Portanto, para a epidemiologia do dengue analisada neste trabalho, ambas as formas de Estatística Scan superestimaram os aglomerados espaciais detectados, mas a Estatística Scan Flexível o fez para um número menor de casos. Essas diferenças certamente se devem ao formato geométrico da janela utilizada por cada método.

Como trabalhos futuros, espera-se estender essa comparação para a epidemiologia de outras doenças de modo a aprofundar o conhecimento das vantagens e desvantagens de cada método em várias situações distintas.

Referências

1. Bailey, L.; Vardulaki, K.; Langham, J.; Chandramohan, D. Introduction to Epidemiology, 1st ed. London: Open University Press (2007).

2. Sanderson C.; Gruen, R. Analytical Models for Decision Making. London: Open University Press (2006).

3. Rothman, K.; Lash, T.; Greenland, S. Modern Epidemiology. Wolters Kluwer (2012).

4. Anselin, L. Spatial data analysis with GIS: an introduction to application in the social sciences. National Center for Geographic Information end Anlisis. University of California - Santa Barbara. August (1992).

5. Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods; 26:1481--1496 (1997)

6. Tango T, Takahashi K. A Flexibly Shaped Spatial Scan Statistic for Detecting Clusters. International Journal of Health Geographics; 4:11. DOI: 10.1186/1476-072X-4-11 (2005)

7. Moraes, R. M.; Nogueira, J. A. & Sousa, A. C. A. A New Architecture for a Spatio-Temporal Decision Support System for Epidemiological Purposes, in Proceedings of the 11th International FLINS Conference on Decision Making and Soft Computing (FLINS2014), Agosto, , Brazil, pp. 17--23 (2014)

8. Braga, I. A.; Valle, D. Aedes Aegypti: Histórico do Controle no Brasil. Epidemiologia e Serviços de Saúde, 16(2), 113--118 (2007)

9. Tango, T., & Takahashi, K.. A Flexible Spatial Scan Statistic with a Restricted Likelihood Ratio for Detecting Disease Clusters. Statistics in medicine, 31(30), 4207--4218 (2012)
10. Tango T. A Test for Spatial Disease Clustering Adjusted for Multiple Testing. Statistics in Medicine; 19:191--204. DOI: 10.1002/(SICI)1097-0258(20000130)(2000)
11. Tango T. A Spatial Scan Statistic With a Restricted Likelihood Ratio. Japanese Journal of Biometrics; 29:75--95 (2008)