

# Uma análise de desempenho dos métodos SCAN e BESAG&NEWELL na detecção de conglomerados espaciais

MARCELO AZEVEDO COSTA

RENATO MARTINS ASSUNÇÃO

LESTE - Laboratório de Estatística Espacial, Departamento de Estatística, ICEX UFMG,  
Caixa Postal 702, Belo Horizonte, MG, CEP 31270-901, Brasil  
{azevedo, assuncao}@est.ufmg.br

**Abstract.** Este artigo propõe uma análise comparativa dos métodos SCAN e BESAG&NEWELL definidos como testes genéricos de conglomerados espaciais. O objetivo do trabalho é explorar as peculiaridades de cada método avaliando o seu desempenho e o seu poder de identificação. O método BESAG&NEWELL necessita do ajuste de parâmetros por parte do usuário, sendo sensível à escolha dos mesmos. O resultado final é apresentado em um formato visual. O método SCAN incorpora os parâmetros cruciais a uma função de verossimilhança, permitindo obter uma estatística para o teste bem como um nível de significância via Monte Carlo. O presente trabalho procura descrever de forma alternativa o método de BESAG&NEWELL incorporando seus parâmetros a uma função de custo que deve ser minimizada sob a região de estudo. Dessa forma, é possível avaliar o desempenho do método de BESAG&NEWELL em relação ao método SCAN.

## 1 Introdução

Os estudos de detecção de conglomerados são muitos importantes para a identificação de regiões geográficas que apresentam risco elevado em relação à ocorrência de um determinado evento, seja uma doença ou uma epidemia, em suas diversas formas. Em especial, é de interesse tentar verificar a hipótese de que a epidemia apresenta uma distribuição espacial aleatória ou se possui um conglomerado de alta incidência. Sob esse aspecto, é preciso que seja definido o conceito de distribuição aleatória, neste caso, a hipótese nula ( $H_0$ ).

As áreas que apresentam um risco significativamente elevado, denominadas conglomerados, podem ser caracterizadas por áreas de caráter puramente temporal, quando a distribuição espacial dos casos é ignorada; puramente espacial, quando, por exemplo, especifica-se um determinado intervalo de tempo observando-se os casos sem que seja necessário conhecer o período exato no tempo de ocorrência do mesmo; ou espaço-temporal quando deve ser analisado não somente a existência de aglomerações no espaço mas também no tempo. De modo geral, o objetivo é encontrar um grupo de eventos limitado em relação ao tamanho e concentração tal que seja improvável a sua ocorrência por mero acaso. No presente trabalho, será considerado o caso de análise de conglomerados puramente espaciais.

Em relação aos métodos de detecção de conglomerados, os mesmos podem ser caracterizados em duas grupos gerais: O primeiro refere-se ao grupo de métodos nos quais os dados são coletados para testar a hipótese de um possível excesso ao redor de uma fonte suspeita, sendo a fonte identificada antes de se observar os dados. Este grupo de métodos é denominado *testes focados de conglomerados*.

O segundo grupo, chamado de *testes genéricos de conglomerados*, procuram identificar as áreas geográficas com um risco significativamente elevado, sem o conhecimento, a princípio, de quais e quantas áreas apresentam risco elevado. Neste trabalho, serão analisados dois métodos referentes ao segundo grupo: o método de varredura de SCAN [2] e o método proposto por Besag e Newell [1], sendo este aqui identificado como o método BESAG&NEWELL.

Independente da classificação quanto ao grupo, o modelo de ausência de conglomerados é o mesmo e representa a situação na qual o padrão espacial dos eventos é totalmente casual. Neste caso, pode-se supor que a taxa de incidência por pessoa é constante em todos os locais, sugerindo que o número esperado de casos em uma determinada área é proporcional ao número de pessoas em risco morando neste local, ou seja a população da região. Estas suposições permitem descrever o modelo nulo de distribuição aleatória.

Seja uma região geográfica, definida como região de estudo, por exemplo, a região geográfica referente a um determinado país. Esta região também pode estar associada, por exemplo, à bacia hidrográfica de um rio. Em qualquer caso, identifica-se uma região de estudo como uma área limitada ao qual deseja-se analisar a ocorrência de um determinado evento. Suponha que a região seja dividida em  $n$  sub-áreas, sendo associada a cada sub-área o número observado de casos,  $y_i$  e o número esperado de casos,  $E_i$ ,  $i = 1, \dots, n$ . Especificando  $N_i$ , como o número total de pessoas-ano em risco na área  $i$  e  $\lambda$  a taxa (anual) de ocorrência de casos. O modelo nulo de aleatoriedade ou ausência de conglomerados pode ser escrito como:

$$H_o : y_i \sim \text{Poisson}(E_i = \lambda N_i), \quad (1)$$

independentes,  $i = 1, \dots, n$

Neste trabalho, no caso de testes genéricos, será considerada a não existência de padronização por idade e sexo,  $E_i \propto N_i$ . A estimativa da taxa ( $\lambda$ ) de ocorrência dos casos pode então ser calculada através da Equação 2.

$$\hat{\lambda} = \frac{\sum_i y_i}{\sum_i N_i} \quad (2)$$

## 2 O método Besag & Newell

O método proposto por Besag e Newell [1] é um método visual, semelhante ao método GAM (*Geographical Analysis Machine*) [5], que procura identificar conglomerados verossímeis de formato circular. A área de risco é identificada por um emaranhado de círculos significativos, sobrepostos. Sendo que cada círculo contém em seu interior um número mínimo de  $k$  casos. Basicamente, o método procura identificar a ocorrência de conglomerados de casos sobre uma extensa região geográfica, subdividindo-a em sub-áreas identificadas pelas coordenadas do respectivo centróide.

Em termos gerais, o método BESAG&NEWELL fixa o número  $k$  de casos que devem ser buscados e, supondo círculos centrados nos centróides de cada sub-área, calcula o raio necessário de forma que cada círculo seja capaz de agrupar, pelo menos,  $k$  casos. Tal procedimento é realizado através de um algoritmo que aumenta sucessivamente o raio do círculo de forma a abranger o centróide mais próximo, incorporando o respectivo número de casos e população. Esta operação é realizada até que seja totalizado, no interior do círculo, um número de casos igual ou superior a  $k$ . A seguir, calcula-se uma estatística sobre a região obtida.

Sejam definidos o número de casos em toda a região de estudo,  $C = \sum_i y_i$ , e a população total em risco,  $M = \sum_i N_i$ . A variável aleatória  $L$  conta o número de outras áreas necessárias para acumular os  $k$  primeiros casos mais próximos do centróide  $i$

$$L = \min\{j : D_j \geq k\} \quad (3)$$

onde  $j$  representa o número de áreas próximas da área  $i$  necessárias para agrupar os  $k$  casos. Se  $l$  é o valor observado de  $L$ , o nível de significância do teste é definido por  $P(L \leq l)$ , calculado sob a hipótese nula ( $H_o$ ). Esta estatística também pode ser definida por  $P(L \leq l) = 1 - P(L > l)$ . Observe que  $L > l$  se e somente se existirem menos de  $k$  casos nas  $l$  primeiras áreas. Ou seja, a probabilidade de que o número de áreas necessárias para agrupar os  $k$  casos seja maior do que  $l$  é igual a probabilidade de que as  $l$  primeiras áreas possuam menos de  $k$  casos. Sob a

hipótese nula (Equação 1), pode-se definir o nível de significância do teste ( $p\_valor$ ) como:

$$P(L \leq l) = 1 - \sum_{j=1}^{k-1} P(N = j)$$

$$P(L \leq l) = 1 - \sum_{j=1}^{k-1} \frac{(mC/M)^j}{j!} e^{-mC/M} \quad (4)$$

onde  $m$  é o número de pessoas em risco nessas  $l$  áreas.

O método de BESAG&NEWELL desenha apenas os círculos significativos ( $p\_valor \leq 0.002$ ). O nível de significância é definido pelo usuário bem como o número  $k$  de casos. A sensibilidade do método aos parâmetros e a representação visual dos resultados dificulta a sua utilização em um estudo comparativo.

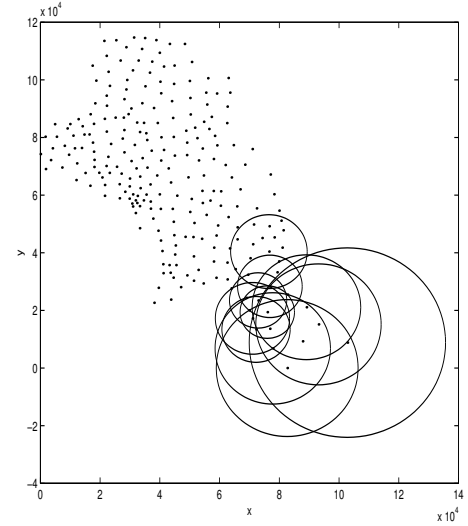


Figure 1: Exemplo do teste de BESAG&NEWELL para um cenário constituído por 245 centróides identificados por pontos no plano xy. O conglomerado candidato é identificado pela sobreposição de círculos. Os parâmetros utilizados neste exemplo foram:  $k = 17$  e  $p\_valor \leq 0.002$ .

A Figura 1 ilustra o comportamento do método para um cenário obtido através de uma base de dados de domínio público [4]. Os pontos representam municípios da região nordeste dos EUA. Estão distribuídos 600 casos, sendo que existe somente 1 conglomerado verossímil. O banco de dados utilizado é mencionado na seção 5.

## 3 O método de varredura SCAN

O método de Varredura (SCAN) proposto por Kulldorff e Nagarwalla [2] permite identificar conglomerados, envol-

vendo um número mínimo de parâmetros não cruciais para serem escolhidos pelo usuário.

Seja  $Z$  o conjunto das áreas  $z$  candidatas a formarem um conglomerado. Estes candidatos são os círculos de raio  $r$  arbitrário centrados em cada um dos  $n$  centróides definidos na região de estudo. Basicamente, os raios dos círculos são definidos de forma que, o aumento do raio implica na inclusão de um novo centróide. Utilizando esta abordagem, partindo de um centróide, tem-se um conjunto de possíveis conglomerados com os raios variando desde a situação onde somente o centróide em questão esteja inserido na região circular, até um círculo que contenha em seu interior, todos os centróides da região. Este número de áreas candidatas pode ser reduzido se for definido um limite para o raio, de modo que nenhum candidato a conglomerado  $z$  contenha mais do que uma certa porcentagem, como por exemplo 20% da população total da área.

O teste SCAN é fundamentado no método de máxima verossimilhança. O parâmetro é definido por  $(z, p, r)$ , onde  $z$  representa o círculo em  $Z$ ,  $p$  é a probabilidade de que um indivíduo qualquer dentro de  $z$  seja um caso e  $r$  é a probabilidade de que um indivíduo fora de  $z$  seja um caso. A região de interesse é definida pelo conglomerado onde  $p > r$ . Uma vez que a hipótese nula é na forma  $H_0 : p = r$  (cada indivíduo é igualmente provável de se tornar um caso), a hipótese alternativa pode ser descrita como:  $H_1 : z \in Z, p > r$ .

Definindo  $n_z$  como o número de indivíduos (população em risco) na região circular  $z$ ,  $c_z$  o número observado de casos nesta mesma região,  $\hat{p} = c_z/n_z$  e  $\hat{r} = (C - c_z)/(M - n_z)$ . A função de verossimilhança referente ao modelo de Bernoulli é dada por:

$$L(z, p, r) = p^{c_z} (1-p)^{(n_z - c_z)} r^{(C - c_z)} (1-r)^{(M - n_z - C + c_z)} \quad (5)$$

É importante observar que o valor do parâmetro que maximiza esta função não é necessariamente aquele correspondente ao círculo com maior taxa  $\hat{p}$ , nem aquele com o maior número de casos. Para encontrar o conglomerado mais verossímil fixa-se  $z \in Z$  e calcula-se  $p(z)$  e  $r(z)$  que maximiza a Equação 5 e, a seguir trabalha-se com  $L(z, p, r)$  para obter a solução que maximiza em  $z$ . Um possível candidato a conglomerado é definido por:

$$L(z, p(z), r(z)) = \sup_{z \in Z, p > r} p^{c_z} (1-p)^{(n_z - c_z)} r^{(C - c_z)} (1-r)^{(M - n_z - C + c_z)} \quad (6)$$

De maneira sucinta, é realizada uma varredura sobre todos os candidatos a conglomerados definido em  $Z$ , o conglomerado com máxima verossimilhança é a região  $\hat{z}$ , para

a qual  $L(z, p(z), r(z))$  é maximizada:

$$L(\hat{z}, p(\hat{z}), r(\hat{z})) \geq L(z, p(z), r(z))$$

para todo  $z \in Z$ .

Ao conglomerado verossímil é atribuída uma estatística do teste da razão de verossimilhança:

$$\kappa = \frac{L(\hat{z}, p(\hat{z}), r(\hat{z}))}{L_o} \quad (7)$$

onde o denominador  $L_o$  é obtido como:

$$L_o = \frac{C^C (M - C)^{M - C}}{M^M}$$

A distribuição de  $\kappa$  depende da distribuição da população e é muito difícil obtê-la analiticamente. Mas, a sua distribuição exata condicionada ao número total de casos observados pode ser obtida utilizando um procedimento de simulação Monte Carlo, através do seguinte algoritmo:

1 -  $S$  conjuntos independentes de dados possuindo o mesmo número de casos  $C$  que o conjunto original, obtidos como realizações de um distribuição multinomial e proporcional a população de cada área, são gerados. Para cada conjunto, calcula-se a estatística do teste da razão de verossimilhança  $(\kappa_1, \dots, \kappa_S)$ .

2 - A partir da ordenação dos valores de  $\kappa$  para os  $S$  conjuntos simulados, compara-se o valor de  $\kappa$ , associado ao conjunto de dados original. Se este estiver entre os maiores  $100(1 - \alpha)\%$  valores, rejeitar  $H_0$  ao nível de significância  $\alpha$ .

3 - Uma vez rejeitada  $H_0$ , o conglomerado  $\hat{z}$  associado ao valor de máxima verossimilhança do modelo não nulo é o conglomerado mais verossímil.

A principal vantagem do método SCAN, além do número mínimo de parâmetros não cruciais, consiste na capacidade do mesmo de reduzir o erro tipo I através da simulação Monte Carlo. O resultado final identifica uma região circular ao qual é associado o valor obtido da função de verossimilhança e um  $p\_valor$  referente à simulação Monte Carlo.

Um exemplo do uso do teste SCAN para a detecção de conglomerado pode ser visto na Figura 2. O método, além de determinar a posição geográfica do conglomerado, calcula a estatística  $\kappa$  e obtém sua distribuição e nível de significância via simulação Monte Carlo sob hipótese nula de aleatoriedade.

#### 4 Metodologia Proposta

Observando os métodos SCAN e BESAG&NEWELL descritos anteriormente, verifica-se que, aparentemente, apresentam abordagens bem distintas. O primeiro procura identificar uma única região circular candidata a conglomerado verossímil, calculando uma estatística  $\kappa$  e atribuindo um  $p\_valor$ , obtido sob  $H_0$ , a essa estatística. O segundo

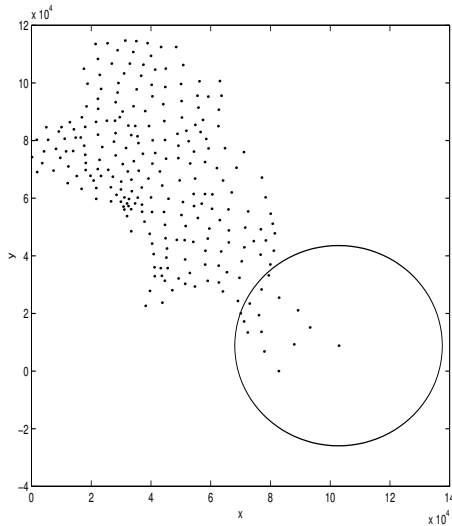


Figure 2: Exemplo do teste SCAN (varredura) para o mesmo cenário apresentado na Figura 1. O conglomerado verossímil, neste caso, é identificado pela circunferência. O teste obteve uma estatística  $\kappa = 74.085$  e um  $p\_valor = 0.001$  (Simulação Monte Carlo sob  $H_o$ )

método calcula um nível de significância ( $p\_valor$ ) para todos os conglomerados circulares que contenham pelo menos  $k$  casos, desenhando-os caso o valor da estatística esteja abaixo de um valor também fornecido pelo usuário. Neste último caso, o conglomerado candidato pode apresentar uma geometria não circular obtida através da sobreposição de círculos.

A metodologia proposta procura padronizar os métodos apresentados, descrevendo uma variação para o método de BESAG&NEWELL. A sua descrição é apresentada a seguir.

Inicialmente, deve-se tratar o problema da geometria. Como o método de Besag realiza o cálculo sob regiões circulares, pode-se definir, neste caso, que um círculo candidato a conglomerado verossímil é representado pela área  $z$  contida em  $Z$  que apresenta o menor valor possível para o nível de significância descrito pela Equação 4. O  $p\_valor$  de interesse é definido por:

$$p_k = \min_{z_k \in Z_k} \{p\_valor(k)\} \quad (8)$$

Onde  $Z_k$  representa o conjunto de todas as regiões circulares  $z_k$  contendo os  $k$  primeiros casos, contando a partir de cada centróide. Ou seja, para  $n$  centróides existem  $n$  regiões circulares, cada uma delas centralizadas em um respectivo centróide. Cada região contém um mínimo de  $k$  casos.

Apesar da padronização em relação à geometria da região de busca, a Equação 8 está condicionada ao valor de

$k$ . Este parâmetro, definido a princípio pelo usuário, pode ser incorporado ao novo método desde que o teste possa ser realizado para diferentes valores de  $k$ . Por exemplo:  $2 \leq k \leq 30$ . Define-se, então, uma nova estatística  $T$  em função da minimização de  $p_k$ .

$$T = \min_k p_k \quad (9)$$

A Equação 9 informa que o candidato a conglomerado verossímil é a região  $z$  que, para possíveis valores de  $k$ , apresenta o menor valor da estatística definida pela Equação 4. Tal característica é apresentada a seguir através da substituição do valor de  $p_k$  na Equação 9.

$$T = \min_k \left\{ \min_{z_k \in Z_k} \{p\_valor(k)\} \right\} \quad (10)$$

Utilizando esta nova abordagem, é possível comparar o poder dos testes SCAN e BESAG&NEWELL (proposto), uma vez que ambos irão apresentar como resultado uma região de geometria circular e um  $p\_valor$  (nível de significância, sob  $H_o$ ). Também é possível calcular, no teste de Besag, a distribuição exata da estatística  $T$  condicionada ao número total de casos observados, sob  $H_o$ . A abordagem é idêntica à descrita para o método SCAN: utiliza-se um procedimento de simulação Monte Carlo. Tal abordagem permite o controle do erro do tipo I.

## 5 O Banco de Dados

Para realizar os testes de comparação dos métodos SCAN e BESAG&NEWELL será utilizada uma base de dados de domínio público disponível em formato eletrônico: <http://www.commed.uchc.edu/biostat/datasets/>. Esta base de dados é constituída por um conjunto de coordenadas espaciais completando um total de 245 centróides. Cada centróide representa um município da região nordeste dos Estados Unidos da América que abrange os seguintes estados: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Pennsylvania, Delaware, Maryland e o Distrito de Columbia. A população em risco atribuída a cada centróide é definida pelo número de mulheres de acordo com o censo realizado em 1990. Esta base de dados foi previamente utilizado para avaliar a existência de conglomerados de mortalidade por câncer de mama [3].

Os métodos serão avaliados quanto à sua capacidade de detecção de uma única região circular ou conglomerado (*one hot spot cluster*). A base de dados apresenta cenários onde 600 casos encontram-se distribuídos sob a forma de conglomerados com 1, 2, 4, 8 e 6 municípios. Cada cenário representa um determinada distribuição dos casos (600) entre os centróides, totalizando um total de 10.000 cenários para cada dimensão de conglomerado. Os centros dos conglomerados são conhecidos e são classificados de acordo

com a sua posição: o conglomerado rural está centrado no município de Grand Isle na região norte de Vermont, o conglomerado urbano está centrado em Manhattan, New York e o conglomerado misto (intercessão da região rural e urbana), em Pittsburg, Allegheny. O tamanho e posição do conglomerado que se deseja identificar pode ser definido pelo seu centro e pelo seu tamanho, observando-se que o conglomerado é formado pelos vizinhos geográficos mais próximos do centro. Por exemplo, um conglomerado de 8 municípios centrado em New York é identificado agrupando-se os 7 municípios mais próximos (distância euclidiana). Utilizando esta abordagem é possível comparar o conglomerado identificado pelo método e o conglomerado real. Uma descrição detalhada do banco de dados, bem como diferentes opções para os cenários é descrita na literatura [4] e no endereço eletrônico indicado.

## 6 Conclusões

O presente trabalho é dividido em duas etapas: a primeira etapa encontra-se em fase de conclusão e é caracterizada pela metodologia proposta, ou seja, a adaptação do método de BESAG&NEWELL para a análise de conglomerados circulares e a escolha automática dos parâmetros críticos, definidos anteriormente pelo usuário. A segunda etapa, a ser realizada, consiste no teste exaustivo dos métodos citados, procurando identificar as suas características de simulação e as respectivas capacidades de identificação.

Espera-se que, ao final do trabalho, seja possível caracterizar o método de melhor desempenho, ou determinar as situações nas quais os métodos apresentam desempenhos distintos: ora superior ora inferior, quantificando a capacidade de identificação dos mesmos.

## Agradecimentos

Os autores agradecem ao CNPq pelo apoio financeiro.

## References

- [1] Julian Besag and James Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A*, (154):143–155, 1991.
- [2] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997.
- [3] Martin Kulldorff, Eric J. Feuer, Barry A. Miller, and Laurence S. Freedman. Breast cancer clusters in the northeast united states: A geographic analysis. *American Journal of Epidemiology*, 146(2):161–170, 1997.
- [4] Martin Kulldorff, Toshiro Tango, and Peter J. Park. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42:665–684, 2003.
- [5] S. Openshaw, A. W. Craft, M Charlton, and J. M. Birch. Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet*, i:272–273, 1988.