APLICABILIDADE DE MÉTODOS DE DETECÇÃO DE AGLOMERADOS ESPACIAIS PARA A IDENTIFICAÇÃO DE ÁREAS DE ALTO RISCO DE DOENÇAS ENDÊMICAS

C. T. Codeço[1] e F. F. Nobre[2]

RESUMO -- Este trabalho procura avaliar a aplicabilidade de métodos estatísticos de detecção de aglomerados espaciais na identificação de áreas de alto risco para doenças endêmicas. Aglomerados de diferentes formas e intensidades, simulados a partir de dados de mortalidade infantil por diarréia no Estado do Rio de Janeiro (1980), são utilizados para o cálculo dos potenciais estatísticos dos dois métodos escolhidos: a estatística μ_S e o "Procedimento de Permutação para a Avaliação de Aglomerados" (CEPP). Os resultados mostram que enquanto a estatística μ_S é mais afetada pelo tamanho dos municípios que formam o aglomerado, a capacidade de detecção do CEPP sofre maior influência da densidade populacional.

Palavras-Chave: Epidemiologia, Aglomerados Espaciais, Simulação.

INTRODUÇÃO

Aglomerado espacial é definido como uma alta taxa de incidência de casos ocorrendo em uma região delimitada, taxa essa superior ao que seria esperado por simples aleatoriedade. Segundo Marshall (1991), existem dois mecanismos gerais de formação de aglomerados que são relevantes em epidemiologia: aquele baseado em um aumento do risco na região devido a algum fator ambiental (por exemplo, radiação) que torna as pessoas que nela vivem independentemente sujeitas a riscos maiores; e a interação espacial, onde o aglomerado se deve ao fato da presença de uma pessoa doente influenciar o adoecimento de outra (os eventos da doença são dependentes entre si).

A busca de explicações para a presença de aglomerados passa pelo levantamento das possíveis variáveis pertinentes à doença quanto à sua variação espacial. A escala do aglomerado é fundamental na definição das variáveis pertinentes. Em uma escala internacional, fatores sociais, étnicos, econômicos, ecológicos, genéticos e muitos outros podem estar envolvidos com o padrão de ocorrência da doença e dificilmente algum deles poderá ser descartado. A um nível mais local, diferenças ecológicas, genéticas e outras podem ser melhor conhecidas e mapeadas e isso pode ajudar na focalização dos fatores geradores destes aglomerados espaciais (Rothman, 1987).

[1] Aluna de Mestrado do Programa de Engenharia Biomédica - COPPE/UFRJ
[2] Professor Adjunto do Programa de Engenharia Biomédica - COPPE/UFRJ
Caixa Postal 8510 - Rio de Janeiro, RJ - BRASIL

// Trabalho recebido em 12/set/95 e aceito em 17/set/96 //

Muitos métodos estatísticos foram desenvolvidos para a detecção de aglomerados espaciais. Esta metodologia de natureza exploratória apresenta o intuito de fornecer subsídios para o conhecimento dos processos etiológicos ainda desconhecidos de certas doenças, como alguns tipos de câncer (mal de Hodgkin, leucemia) , defeitos de nascimento e doenças crônicas (Symons et alli, 1983; Abel e Becker, 1987; Bender, 1987; Schulte et alli, 1987; Raubertas, 1989; Turnbull et alli, 1990; Glaser, 1990). O princípio empregado parte da idéia de que, se detectamos a existência de um aglomerado espacial de uma doença numa dada região, podemos pensar que existe algum fator ligado àquela região que está determinando esta alta incidência. A partir daí, podem ser levantadas hipóteses de possíveis fatores etiológicos.

Outro objetivo associado à detecção de aglomerados espaciais é a identificação de regiões de alto risco para uma doença, o que pode auxiliar no processo de tomada de decisão e planejamento de medidas para o controle epidemiológico. Doenças como a malária têm recebido grande atenção devido aos seus elevados índices de incidência em grande parte do país, justificando o estudo de metodologias que agilizem a análise do seus padrões de espalhamento e forneçam informações referentes às áreas com prioridade de ação (Vasconcellos, 1993).

No entanto, a aplicação desta metodologia a dados de doenças endêmicas precisa ser avaliada com cuidado. Como a maioria dos métodos foi desenvolvida para doenças raras, algumas premissas estatísticas podem não ser adequadas e levar a resultados espúrios. Neste trabalho, procuramos avaliar o desempenho de dois métodos, o "Procedimento de Avaliação de Aglomerados por Permutação" (no original, "cluster evaluation permutation procedure", CEPP) (Iwano, 1989) e a estatística μ_S (Raubertas, 1988), quando aplicados a dados de mortalidade infantil por diarréia em 1980, no Estado do Rio de Janeiro.

São muitas as variáveis que podem influenciar a performance de detecção de aglomerados espaciais. Wartenberg e Greenberg (1993) citam, como exemplos, a heterogeneidade populacional, o risco diferencial para pessoas de diferentes idades, etnias e sexo, efeitos migratórios e o tamanho, forma e posição do(s) aglomerado(s) no mapa em estudo. Análise preliminar da estatística μ_S e do CEPP, empregando dados simulados, mostra que são mais eficientes na detecção de aglomerados de doenças frequentes do que raras. Além disso, o tamanho dos municípios da área de estudo mostraram influenciar a capacidade de detecção dos métodos. Aglomerados situados em áreas representadas por municípios pequenos foram melhor diagnosticados do que aqueles situados em regiões de municípios grandes (Codeço, 1995).

No presente estudo, avaliamos a aplicabilidade desta metodologia mediante a geração de simulações baseadas nos dados reais de mortalidade infantil por diarréia. De forma geral, dados reais contêm uma série de características que dificilmente podem encontrar correspondência em dados completamente simulados, como por exemplo, os limites irregulares do mapa, a possível proximidade dos aglomerados em relação às bordas e a heterogeneidade da taxa da doença. Devido a isto, procuramos preservar estas características dos dados originais e realizar simulações apenas para gerar aglomerados de forma e posição conhecidas, permitindo assim verificar a capacidade de detecção de cada método.

MÉTODOS DE DETECÇÃO DE AGLOMERADOS

Métodos de detecção de aglomerados são, classicamente, testes de hipótese onde a Hipótese Nula, H_0, é de que a ocorrência do evento de interesse em uma dada região é aleatória. Para populações homogêneas, a H_0 usual é de que a taxa da doença é constante em todos os municípios, isto é, o número de casos varia apenas de acordo com a população de risco de cada região (Waller e Lawson, 1995). Cada método, então, procura formular estatísticas de teste que resumam a similaridade dos valores observados em regiões geográficas adjacentes, de forma a rejeitar o modelo nulo na presença de aglomerados espaciais.

Estatística μ_s

Este método, proposto por Raubertas (1988), é aplicável quando taxas da doença estão disponíveis por área (por exemplo, taxa de mortalidade por município) e detecta um aglomerado quando existem municípios próximos cujas taxas estão acima do valor esperado. A comparação entre a taxa de mortalidade observada da doença (π_i) no município i e o valor esperado (p_i) (calculado a partir do modelo nulo, onde a taxa da doença é constante) gera um índice d_i associado a cada município:

$$d_i = \frac{\pi_i - p_i}{\sqrt{p_i}} \tag{1}$$

Valores altos positivos de d_i indicam taxa observada acima do esperado no município i, valores negativos indicam taxas abaixo do esperado. Porém, segundo a definição de aglomerado, não basta que municípios apresentem taxas elevadas, é preciso que haja uma relação de vizinhança entre eles. Desta forma, é gerado um índice a_i que associa a cada município i a soma dos valores d_i de todos os seus vizinhos:

$$a_i = \sum_{j \in N_i} w_{ij} \left(d_j - \frac{\sum_{j=1}^{n} d_j}{n} \right) \tag{2}$$

onde: i, j são municípios;
n é o número total de municípios na área de estudo;
N_i é o conjunto dos municípios vizinhos de i, e
w_{ij} é o peso dado ao grau de vizinhança entre i e j.

Geralmente $w_{ij}=1$, se i e j são vizinhos e $w_{ij}=0$, se i e j não o são. Neste trabalho, estipulamos raios de abrangência e aqueles municípios cujas sedes municipais se encontravam a uma distância menor do que este raio foram considerados vizinhos. Critérios de vizinhança mais complexos também são possíveis e têm sido sugeridos por alguns autores (Vasconcellos, 1993; Cliff e Ord, 1981). Por fim, o índice a_i é normalizado através da divisão pelo seu desvio padrão (obtido por simulação) para ter valor esperado zero e desvio padrão 1. Valores altos de a_i sugerem a presença de aglomerados. A estatística de teste final é:

$$\mu_s = \frac{1}{n}\sum_{i=1}^{n}(a_i)^2 \qquad (3)$$

A distribuição de μ_S no modelo nulo de taxa de mortalidade constante foi obtida por simulação via Monte Carlo, com 1000 simulações (Raubertas, 1988). O valor crítico para o teste de hipótese corresponde ao 995° valor obtido da ordenação dos μ_S simulados segundo H_0.

A identificação dos municípios envolvidos na formação do aglomerado é feita através dos índices a_i. Uma vez o teste tenha indicado a presença de aglomerados (μ_S observado > μ_S crítico), o método define como os municípios que mais contribuiram para o resultado positivo aqueles com valores de a_i maiores do que 2 (isto é, que excedem pelo menos dois desvios padrão em relação ao valor esperado).

CEPP - Procedimento de Permutação para a Avaliação de Aglomerados

O CEPP (Iwano, 1989; Turnbull et alli, 1990) apresenta uma abordagem diferente da estatística μ_S e parte do pressuposto de que, se cada um dos n municípios da área de estudo for agrupado aos seus vizinhos até criar k "células" que possuam populações de risco iguais, na ausência de aglomerados esperaremos que estas "células" apresentem aproximadamente o mesmo número de casos. Assim, Iwano propõe que, após a formação destas k células, sejam analisadas aquelas que apresentarem maiores números de casos e verificar se estes valores excedem ou não o esperado na ausência de aglomerados. Se realmente exceder, isto indicará que a região compreendida por esta célula possui um risco significativamente alto para a doença em estudo e é a área que deve ser considerada como de vigilância prioritária.

O procedimento de montagem das células é o seguinte: Escolhe-se qual vai ser o tamanho de cada célula (dado pelo valor P, que significa o número de pessoas em cada célula); em seguida, para cada município *i* (i = 1,...,n, onde n é o número total de municípios), verifica-se qual a população de risco p_i. Se $p_i < P$, o município mais próximo de i, chamado de j, é analisado. Se $p_i + pj = P$, então os dois municípios formarão a célula r (r = 1,...,k). Se a soma for maior do que P, então apenas uma fração de j é incorporada na célula r. Se este valor for menor do que P, o próximo vizinho de i é analisado e assim por diante. Uma descrição detalhada deste procedimento pode ser encontrada em Turnbull et alli (1990).

Com as k células montadas, verifica-se o número de casos, C_r, que foi alocado em cada uma. Escolhemos analisar, neste trabalho, a significância das três células com maiores valores de C_r. A célula com maior C_r é chamada M_1, a segunda é M_2 e a terceira é M_3. A presença de aglomerados é testada comparando-se estes valores com a distribuição empírica, obtida por simulação com 1000 replicações sob a hipótese de aleatoriedade (Turnbull et alli, 1990).

A escolha do parâmetro P é essencial para o CEPP e deve considerar o tamanho médio dos municípios. Ele não pode ser tão pequeno que não junte os municípios, nem tão grande que todas as n células sejam formadas pelos mesmos municípios. Turnbull et alli (1990) dizem que o valor de P depende da doença e do padrão de exposição, mas que estes parâmetros são geralmente difíceis de determinar na prática.

PROCEDIMENTO DE SIMULAÇÃO

Os dois métodos foram aplicados a aglomerados simulados sobre o mapa do Rio de Janeiro. O procedimento para avaliação dos métodos consiste na simulação de varios padrões, todos contendo regiões definidas como aglomerados. Foram criados quatro tipos de aglomerados espaciais (mostrados na figura 1), definidos em funcão de características regionais da região. Os padrões para aglomerados considerados foram:

1. GRio: engloba 10 municípios do Grande Rio, caracterizados por uma alta densidade populacional e por um formato aproximadamente circular, com uma parte de seu perímetro correspondendo à borda do mapa (Oceano Atlântico).

2. Interior: são 9 municípios de densidade populacional relativamente baixa formando uma área aproximadamente circular no interior do estado.

3. Dutra: é um aglomerado aproximadamente elíptico que se estende por 11 municípios atravessados pela via Dutra, importante eixo de fluxo de população.

4. Macaé: aglomerado elíptico localizado em área de municípios grandes. Embora possua a mesma ordem de magnitude dos outros aglomerados, ele é composto por apenas 4 municípios: Macaé, Campos, Casimiro de Abreu e Conceição de Macabu.

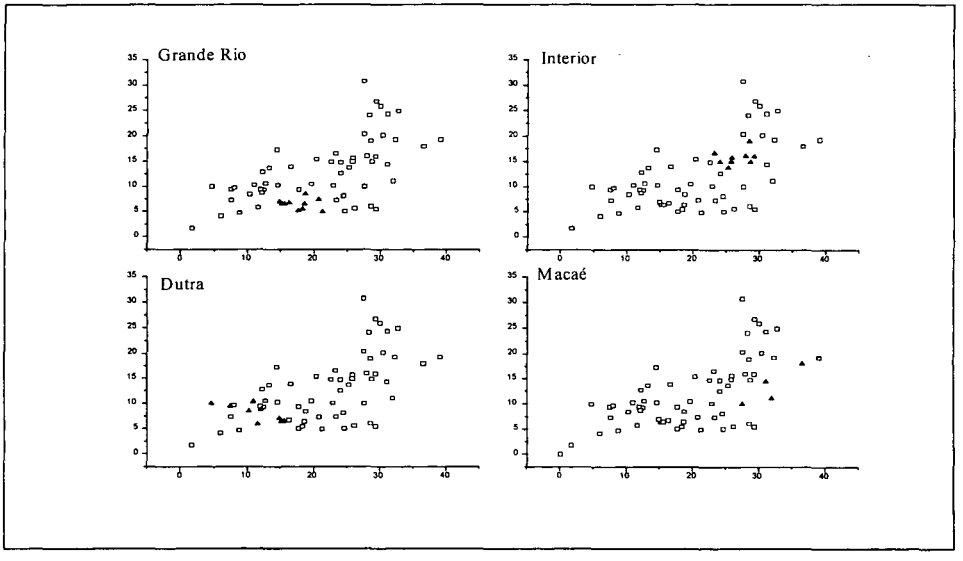

Figura 1. Localização dos aglomerados simulados no Estado do Rio de Janeiro

O método de simulação consiste em uma randomização estratificada dos dados. Inicialmente, as taxas de mortalidade são divididas em dois grupos, de forma a ter um conjunto com as taxas a serem aleatoriamente alocadas nas regiões pertences ao aglomerado. Este conjunto deve conter um número maior de valores do que o número de regiões presentes no aglomerado. O procedimento de simulação consistiu em alocar os dados deste grupo aos municípios do aglomerado, e as taxas não alocadas passavam a fazer parte do outro grupo para a alocação aleatória às demais regiões. Esta estratégia foi adotada, visto estarmos interessados na avaliação dos métodos quanto a sua aplicação a dados de doenças frequentes, ao contrarío de outros estudos que focalizam doenças raras. Para avaliarmos cada padrão, utilizamos a distribuição empírica construida através de 1000 simulações para a situação de ocorrência aleatória, isto é, sem aglomerados.

Os dados utilizados se referem à mortalidade infantil por diarréia em 64 municípios do Estado do Rio de Janeiro (1980). A população alvo considerada é a de nascidos vivos do censo de 1980. A figura 2 mostra a distribuição espacial das taxas de mortalidade da doença no estado. Optou-se por trabalhar com os dados de 1980, uma vez que nesta época, o estado apresentava municípios com areas mais variadas do que a divisão geográfica atual . A distribuição de frequência das taxas e a acumulada são apresentadas na figura 3.

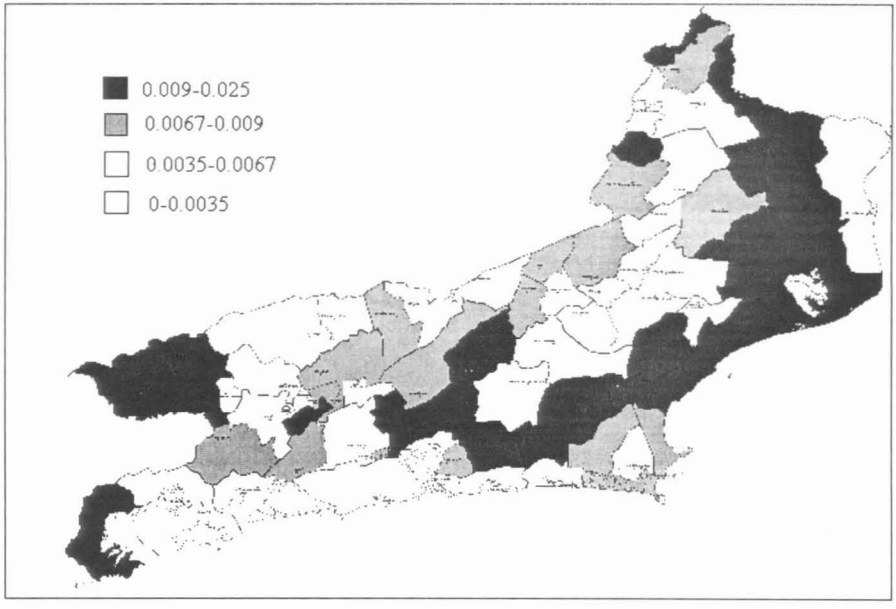

Figura 2. Taxa de mortalidade infantil por diarréia nos 64 municípios do Estado do Rio de Janeiro, em 1980. As cores representam os quartis da distribuição das taxas.

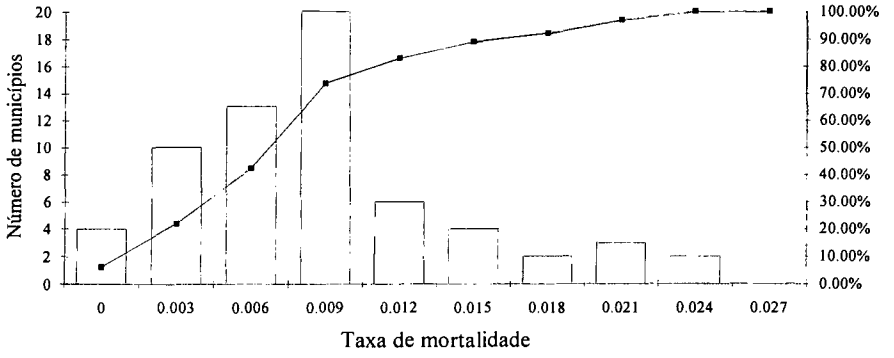

Figura 3. Histograma de frequências e polígono de frequências acumuladas das taxas de mortalidade infantil por diarréia nos 64 municípios do Estado do Rio de Janeiro (1980).

A randomização das taxas de mortalidade, cujo histograma é apresentado na figura 3, foi realizada de forma a gerar aglomerados com três intensidades:

- Aglomerados fortes: as taxas de mortalidade nos municípios do aglomerado estão no quartil superior da distribuiçáó de taxas, correpondendo a 17 valores para alocação aleatória nos municípios pertences ao aglomerado.

- Aglomerados moderados: taxas de mortalidade tiradas do último tercil da distribuição, correspondendo a 22 observações.

- Aglomerados fracos: taxas de mortalidade obtidas a partir dos valores maiores do que a mediana da distribuição, que resulta em 37 taxas.

Os pontos de corte foram definidos após análise da distribuição das taxas de mortalidade e considerando o número máximo de municípios existentes nos aglomerados (neste caso, 11 municípios para o padrão espacial definido como Dutra). Assim, foram gerados doze tipos de aglomerados, pela combinação das quatro posições com as tres intensidades. Como as populações de risco em cada município foram mantidas em seus lugares e apenas as taxas foram randomizadas, os óbitos associados a cada município em cada situação simulada foi obtido pela multiplicação da população pela taxa.

Utilizamos como raios de vizinhança para o cálculo da estatística μ_S 20, 30, 50 e 70 km. O CEPP foi aplicado com valores de P iguais a 1, 2, 4 e 6 vezes a média populacional, isto é, 4700, 9400, 18800 e 28200 pessoas em cada célula, analisando-se as três células com maior número de casos (M_1, M_2, M_3).

RESULTADOS E DISCUSSÃO

Estatística μ_S

Os parâmetros utilizados para comparar os resultados da estatística μ_S quanto a sua capacidade de identificar os municípios envolvidos na formação do aglomerado foram:

$$\text{taxa de acertos} = \frac{\text{número de municipios corretamente identificados}}{\text{número total de municipios presentes no aglomerado}}$$

$$\text{taxa de erro} = \frac{\text{número de municipios indicados incorretamente}}{\text{número total de municipios identificados}}$$

A taxa de acertos indica o quanto do aglomerado foi identificado e a taxa de erro indica quanto da resposta do método referente a indicação dos municípios envolvidos está errada.

A estatística μ_S rejeitou a hipótese nula de uniformidade em todos os aglomerados simulados, indicando a presença de aglomerados para todos os raios de vizinhança. Variações ocorreram, porém, na capacidade de identificação dos municípios pertencentes ao aglomerado. A Tabela 1 mostra a taxa de acertos e de erro da estatística μ_S.

A taxa de erro foi relativamente baixa para todos os aglomerados com exceção de Macaé, com valores muito altos. Isto pode ser explicado pelo fato deste aglomerado ser formado por poucos municipios e muitos outros com taxas altas de mortalidade estarem presentes no resto do estado, embora aleatoriamente distribuidos. Mesmo assim, a taxa de acerto foi alta para aglomerados fortes nesta região.

O Grande Rio foi a região com resultados mais estáveis, com taxas de acerto bem altas e de erros baixas, só aumentando para o raio 70, que excede em muito a magnitude do aglomerado. O Interior apresentou taxas de acertos mais baixas, em torno de 70% e erros um pouco mais altos do que o Grande Rio. Para a Dutra, os resultados foram altos para aglomerado forte mas decaíram para aglomerados mais fracos.

Os resultados parecem indicar que a heterogeneidade da taxa de mortalidade por diarréia dificulta a identificação dos municípios pertencentes ao aglomerado, uma vez que ambos os métodos apresentaram potencialidade máxima para a detecção de aglomerados simulados quando a taxa encontrada nas áreas fora do aglomerado eram uniformemente baixas (Codeço, 1995). Isto fica mais claro quando o aglomerado está numa região pouco representada por centróides (isto é, tem municípios grandes) onde o resto do estado, com sua heterogeneidade, leva à inclusão de muitos outros municípios erroneamente no resultado. Porém, apesar desta diversidade e do número de municípios falso positivos, o método apresentou um bom desempenho quanto à taxa de acertos, que foi alta na maioria das situações.

Tabela 1. Taxa de acerto e taxa de erro da estatística μ_S para a identificação dos municípios pretencentes aos aglomerados simulados.

	Raio 20		Raio 30		Raio 50		Raio 70	
Aglomerado	**acerto**	**erro**	**acerto**	**erro**	**acerto**	**erro**	**acerto**	**erro**
GRio								
Forte	0,8	0	1	0	1	0,07	1	0,13
Médio	0,8	0	1	0,02	1	0,05	1	0,18
Fraco	0,5	0,04	0,7	0,04	0,8	0,11	0,9	0,13
Interior								
Forte	0,78	0,16	0,55	0,13	0,67	0,11	0,78	0,04
Médio	0,74	0,09	0,77	0,07	0,89	0,04	1	0,11
Fraco	0,77	0,02	0,77	0,07	0,67	0,04	0,78	0,09
Dutra								
Forte	0,73	0,07	0,64	0,02	0,91	0,09	0,83	0,17
Médio	0,64	0,04	0,81	0,07	0,63	0,13	0,45	0,02
Fraco	0,45	0,02	0,54	0,04	0,64	0,09	0,58	0,17
Macaé								
Forte	0,75	0,7	0,75	0,62	0,75	0,72	0,75	0,77
Médio	0,5	0,75	0,25	0,78	0,75	0,57	0,75	0,81
Fraco	0,25	0,89	0,25	0,67	0,25	0,83	0,5	0,71

Vasconcellos (1993), ao utilizar a estatística μ_S para detectar aglomerados de casos de malária no estado de Tocantins, mostrou que a utilização da hipótese nula descrita por Raubertas (1988, 1989) de uniformidade das taxas é mais adaptada a doenças raras. Doenças endêmicas apresentam geralmente uma variabilidade nas suas taxas, mesmo na ausência de aglomerados, que não são compatíveis com o modelo nulo de taxas de mortalidade uniformes. Esta heterogeneidade faz com que a estatística μ_S acuse a presença de aglomerados mesmo quando estes não existem (isto é, aumenta o número de falsos positivos). Devido a isto, Vasconcellos propõe uma hipótese nula alternativa que admite a existência de variações significativas nas taxas de mortalidade, mas que estes desvios se encontram randomicamente distribuidos. A rejeição desta hipótese nula significa que os municípios que apresentam desvios semelhantes estão agrupados, formando aglomerados. A Tabela 2 mostra os resultados da aplicação da estatística μ_S aos dados simulados, utilizando como Hipótese Nula o modelo de Vasconcellos.

Pela Tabela 2, podemos ver que, para a maioria dos aglomerados simulados, o método não detectou sua presença, ao utilizar a Hipótese Nula alternativa. Apenas o aglomerado localizado no Grande Rio foi detectado independentemente da sua intensidade e do raio de vizinhança empregado. Para estes, tanto a taxa de erro como a de acerto tenderam a aumentar com o aumento do raio.

Tabela 2. Taxa de acerto e taxa de erro da estatística μ_S para a identificação dos municípios pertencentes aos aglomerados simulados, utilizando a H_0 proposta por Vasconcellos. O símbolo * indica aglomerados não detectados (aceitou H_0).

	Raio 20		Raio 30		Raio 50		Raio 70	
Aglomerado	**acerto**	**erro**	**acerto**	**erro**	**acerto**	**erro**	**acerto**	**erro**
GRio								
Forte	0,7	0	0,7	0	0,8	0,11	0,9	0,31
Médio	0,7	0	0,9	0	1	0,23	1	0,23
Fraco	*	*	*	*	0,8	0,27	0,6	0,4
Interior								
Forte	*	*	*	*	*	*	*	*
Médio	*	*	*	*	*	*	*	*
Fraco	*	*	*	*	*	*	*	*
Dutra								
Forte	*	*	*	*	0,45	0,37	*	*
Médio	*	*	*	*	*	*	*	*
Fraco	*	*	*	*	*	*	*	*
Macaé								
Forte	*	*	*	*	*	*	*	*
Médio	*	*	*	*	*	*	*	*
Fraco	*	*	*	*	*	*	*	*

Os resultados encontrados mostram que, enquanto que a Hipótese Nula proposta por Raubertas é sensível à heterogeneidade das taxas (Vasconcellos, 1993), o modelo nulo de Vasconcellos é por demais conservativo e não reconheceu a maioria dos aglomerados simulados, com exceção daqueles situados no Grande Rio, o que pode ser explicado pelo fato do método ser mais sensível a aglomerados situados em áreas de maior densidade populacional (Raubertas, 1988).

CEPP

Para o CEPP, usamos como indicador de sua performance a taxa de acertos, isto é, quantas vezes o método identificou como significativas as células M_1, M_2 e M_3 que realmente estavam dentro do aglomerado, utilizando 4 valores de P (tamanhos de célula). A Tabela 3 mostra a taxa de acertos e a proporção de resultados positivos, isto é, nos 12 testes realizados (3 células e 4 valores de P) qual a proporção de resultados considerados significativos (aglomerados detectados).

Com exceção do resultado para aglomerado fraco no interior, todos os outros testes indicaram a presença de aglomerados. É interessante notar que o CEPP só foi capaz de indicar com eficiência as localidades do aglomerado Grande Rio e Dutra, que cobrem áreas mais populosas. A taxa de acertos foi muito baixa para aglomerado no interior, reforçando os resultados obtidos por Iwano (1989) de que o método é mais sensível a aglomerados em áreas mais populosas.

Tabela 3. Taxa de acertos na identificação do aglomerado do CEPP aplicado aos aglomerados simulados. Os valores de P se referem ao tamanho das células. O símbolo * indica aglomerados não detectados (aceitou H_0).

Aglomerado	P = 1	P = 2	P = 4	P = 6
GRio				
Forte	1	1	1	1
Médio	1	1	1	1
Fraco	0,75	0,74	0,72	0,73
Interior				
Forte	0,05	0,02	0,06	0,20
Médio	0,21	0,05	0,04	0,04
Fraco	*	*	0,00	0,00
Dutra				
Forte	1	1	0,60	0,52
Médio	1	1	0,65	0,65
Fraco	1	1	0,70	0,68
Macaé				
Forte	0,60	0,61	0,35	0,21
Médio	0,33	0,35	0,30	0,32
Fraco	0,30	0,32	0,26	0,20

CONCLUSÃO

Apesar de muitos métodos terem sido descritos para a detecção de aglomerados espaciais em dados epidemiológicos, poucos são aqueles que além de acusar a presença deste padrão são também capazes de identificar as localidades envolvidas. A estatística μ_S e o CEPP possuem estas duas características e mostraram-se sensíveis para os aglomerados simulados neste estudo.

Interessante notar que ambos os métodos, embora tenham algoritmos bem diferentes, são sensíveis aproximadamente às mesmas condições. A potencialidade dos métodos aumenta quando os aglomerados estão localizados em áreas mais populosas, ou quando eles se estendem por um número grande de municípios. Para trabalhar com áreas de municípios grandes, o ideal seria utilizar dados em escalas menores, como distritos.

A escolha do raio de abrangência (estatística μ_S) ou do valor de P (CEPP) influenciou mais na identificação dos municípios do que na detecção do aglomerado. Em geral, os dois métodos deram resultados positivos para todos os aglomerados propostos, independentemente do valor escolhido para o critério de vizinhança. Porém, o aumento do raio de abrangência para além da dimensão do aglomerado levou a um aumento da taxa de erros na estatística de Raubertas. Quanto ao CEPP, encontramos melhores taxas de acerto para valores pequenos de P.

Por fim, a partir das simulações realizadas, podemos dizer que ambos os métodos são capazes de detectar os aglomerados propostos mas a estatística μ_S apresentou resultados melhores quanto à

identificação das localidades envolvidas. A estatística μ_S é mais influenciada pelo número de municípios que formam o aglomerado e cuidado deve ser tomado quando houver suspeita de aglomerado situado em área de municípios grandes, pois a taxa de erros tende a aumentar. Por outro lado, a densidade populacional foi mais importante na determinação do potencial do CEPP, que apresenta taxas de acerto muito baixas na localização de aglomerados em áreas menos populosas.

AGRADECIMENTOS

À Marília Carvalho, pesquisadora da ENSP / FIOCRUZ, pela gentil doação dos dados de mortalidade infantil por diarréia. Ao CNPq e IDRC, pelo apoio financeiro.

REFERÊNCIAS

ABEL, U. and BECKER, N. (1987). "Geographical Clusters and Common Patterns in Cancer Mortality of the Federal Republic of Germany". *Archive of Environmental Health.* v. 42(1), p. 51-57.

BENDER, A. P. (1987). "On Disease Clustering" (letter). *American Journal of Public Health.* v. 77, p. 742.

CLIFF, A. D. and ORD, J. K. (1981). *Spatial Processes. Models and Applications.* Pion Limited.

CODEÇO, C. T. (1995). *Comparação de Métodos de Detecção de Aglomerados Espaciais em Epidemiologia.* Tese de Mestrado, Programa de Engenharia Biomédica, COPPE / UFRJ, Rio de Janeiro.

GLASER, S. L. (1990). "Spatial Clustering of Hodgkin's Disease in the San Francisco Bay Area". *American Journal of Epidemiology.* v. 132(1), p. S167-177.

IWANO, E. J. (1989). *A comparation of cluster detection procedures.* Tese de Mestrado, Cornell University, USA.

MARSHALL, R. J. (1991). "A review of methods for the statistical analysis of spatial patterns of disease". *Journal of the Royal Statistical Society A*, v. 154, p. 421-441.

RAUBERTAS, R. F., BROWN, P., CATHALA, F. E and BROWN, I. (1989). "The Question of Clustering of Creutzfeldt-Jakob Disease". *American Journal of Epidemiology.* v. 129(1), p. 146-154.

RAUBERTAS, R. F. (1988). "Spatial and Temporal Analysis of Disease Ocurrence for Detection of Clustering". *Biometrics.* v. 44, p. 1121-1129.

ROTHMAN, K. J. (1987). "Clustering of disease". *American Journal of Public Health.* v. 77, p.13-15.

SCHULTE, P. A. R. L., EHRENBERG, E. and SINGARAL, M. (1987). "Investigation of occupational cancer clusters: Theory and Practice". *American Journal of Public Health.* v. 77, p.52-56.

SYMONS, M. J., GRIMSON, R. C. and YUAN, Y. C. (1983). "Clustering of rare events". *Biometrics*. v. 39, p.193-205.

TURNBULL, B. W., IWANO, E. J., BURNETT, W. S., HOWE, H. L. and CLARK, L. C. (1990). "Monitoring for Clusters of Disease: Application to Leukemia Incidence in Upstate New York". *American Journal of Epidemiology*. v. 132 (Supl.1), p. S136-143.

VASCONCELLOS, A. A. T. (1993). *Análise de "clusters" espaciais para dados epidemiológicos.* Tese de Mestrado, Programa de Engenharia Biomédica, COPPE / UFRJ, Rio de Janeiro.

WARTENBERG, D. and GREENBERG, M. (1993). "Solving the Cluster Puzzle Clues to Follow and Pitfalls to Avoid". *Statistics in Medicine*. v. 12, p.1763-1770.

WALLER, L. A. and LAWSON, A. B. (1995). "The Power of Focused Tests to Detect Disease Clustering". *Statistics in Medicine*. v. 14, p.2291-2308.

EVALUATION OF STATISTICAL METHODS OF SPATIAL CLUSTER DETECTIONON THE IDENTIFICATION OF HIGH RISK AREAS FOR ENDEMIC DISEASES

C. T. Codeço[1] and F. F. Nobre[2]

ABSTRACT -- The goal of this work is to study the ability of statistical methods of spatial cluster detection to identify high risk areas for endemic diseases. Clusters of different shapes and intensities, simulated from diarrhea infant mortality in Rio de Janeiro (1980), were used to compute the statistical power of two chosen methods: the μ_S statistics and the "Cluster Evaluation Permutation"(CEPP). The results show that μ_S statistic is influenced by the municipality sizes and the CEPP is affected by their population density.

Key - words: Epidemiology, Spatial Cluster, Simulation.

[1] Aluna de Mestrado do Programa de Engenharia Biomédica - COPPE/UFRJ
[2] Professor Adjunto do Programa de Engenharia Biomédica - COPPE/UFRJ
Caixa Postal 8510 - Rio de Janeiro, RJ - BRASIL