DETECÇÃO DE CLUSTERS ESPACIAIS VIA ALGORITMO SCAN MULTI-OBJETIVO

FLÁVIO DOS REIS MOURA

FLAVIOSMOURAS@YAHOO.COM.BR

DISSERTAÇÃO SUBMETIDA AO DEPARTAMENTO DE ESTATÍSTICA DA UNIVERSIDADE FEDERAL DE MINAS GERAIS EM AGOSTO DE 2006 EM CUMPRIMENTO ÀS EXIGÊNCIAS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ESTATÍSTICA.

ÁREA DE CONCENTRAÇÃO: **ESTATÍSTICA ESPACIAL**

ORIENTADOR: **PROF. LUIZ HENRIQUE DUCZMAL**

AGOSTO / 2006

BELO HORIZONTE / MG

Dedicatória

À minha família

Agradecimentos

Em primeiro lugar a Deus por tudo.

A meus pais, Oscar Moura e Maria Coelho dos Reis Moura pelo amor e dedicação irrestrita para com seus filhos.

A meus irmãos Jose Antônio, Cristiano, Cristina e Lucilene, pelo carinho, apoio e amizade e por todas as boas coisas que convivemos.

Ao meu orientador Prof. Luiz Henrrique Duczmal, PhD, pelo conhecimento incentivo, dedicação, dicas e especialmente pela orientação paciente e irrestrita.

A minha namorada Eva Rodrigues Vieira pelo amor e compreensão.

Ao discente Ricardo Tavares, do curso de Estatística da UFMG, sem cujo o auxilio este trabalho não estaria completo.

Aos meus amigos pelo apoio, compreensão incentivo, em especial a Ricardo Tavares, Fabrício Avelar, Spencer Barbosa, Anderson Duarte, Luiz Cláudio, Marcelo Lindolfo, Ana Paula, Washington, Rubens Crispim, Wemerson Ribeiro, Jurandir, Petrúcia, Alexandre Celestino, Keila, Erik, Carlito, Fábio Marques, Vander Luiz, Geraldo Manaus, Helem Alkimim , Lucineia do Amaral e Viviane.

As professora Cristina Marques e Cibele Queiroz, pelo apoio e confiança em minha capacidade.

Aos professores Gregório Saraiva, Frederico Cruz, Henrrico Colosmo pelo apoio e pela oportunidade.

Ao professor Ricardo Takahashi, pelo apoio e conselhos sempre muito sensatos.

Aos demais professores e funcionários do Departamento de Estatística da Universidade Federal de Minas Gerais.

Finalmente novamente a Deus, Agradeço pelas lições recebidas, pelas oportunidades, pelas dádivas e pelos amigos e amigas que ele coloca a nossa vida.

RESUMO

Situações em que clusters espaciais de doenças não têm um formato regular são muito comuns. Além disso, mapas com múltiplos clusters, que não têm um cluster primário claramente dominante, ocorrem freqüentemente. Nós desenvolvemos um método para analisar mais detalhadamente os diversos níveis de clusterização que aparecem naturalmente em mapas de doenças divididos em *m* regiões.

A estatística scan espacial é uma medida usual da intensidade de um cluster. Outra medida importante é a regularidade geométrica. O algoritmo genético multi-objetivo foi desenvolvido anteriormente para identificar o formato geométrico dos clusters. Este método realiza uma busca para maximizar dois objetivos, a estatistica scan e a regularidade da forma (o conceito de compacidade). A solução encontrada é um conjunto de Pareto, consistindo de todos os clusters encontrados que não são piores que nenhum outro cluster em ambos objetivos simultaneamente. A avaliação da significância é feita paralelamente para todos os cluster através de simulações de Monte Carlo. Este procedimento determina a melhor solução.

Ao invés de usarmos o algoritmo genético, nós desenvolvemos um novo método que incorpora a simplicidade do método scan circular, sendo capaz de detectar e avaliar clusters de formato irregulares. Nós definimos a ocupação circular (OC) de uma zona candidata a cluster como a sua população dividida pela população dentro do menor círculo que a contém. O conceito de OC é computacionalmente rápido, utiliza um conceito mais intuitivo, e substitui aqui o conceito de compacidade como outra medida de regularidade de forma. A estatística scan é calculada para cada uma das *m* regiões do mapa examinado-as individualmente. As regiões são ordenadas decrescentemente de acordo com o valor da estatística scan. Seja R(k) o conjunto contendo as k primeiras regiões. A modificação multi-objetivo do algoritmo scan circular é aplicada

sucessivamente para cada conjunto R(k). Em cada círculo, a zona candidata a ser um cluster consiste das regiões pertencentes a R(k) e que estão no círculo. Na prática nós escolhemos somente alguns poucos valores de k tais como m, m/2, m/4,. . .1. Para cada valor de k nós construímos um conjunto de Pareto P(k). Reunimos todos os conjuntos de Pareto em um gráfico e calculamos o cojunto de Pareto Globlal P(0). Um procedimento de Monte Carlo é usado para avaliar a significância dos clusters.

A presença de "joelhos" no conjunto de Pareto indica transições repentinas na estrutura dos clusters, correspondendo a rearranjos devido à coalescência de clusters fracamente ligados (geralmente desconectados). Cada conjunto de Pareto contém os cluster mais prováveis dentro de um certo nível de informação geográfica. Eles são relacionados, refletindo a distribuição dos casos, estrutura de população e vizinhança do mapa. Computacionalmente, o método é somente algumas vezes mais demorado que o scan circular usual.

O scan circular multi-objetivo permite enxergar a estrutura de clusters de um mapa. A comparação do conjunto de Pareto de casos observados com aquele calculados sobre a hipótese nula fornece indicações valiosas sobre a ocorrência de clusters espaciais de doenças. O potencial para monitoramento de clusters incipientes e em diversas escalas geográficas simultaneamente o torna uma ferramenta promissora em vigilância sindrômica, especialmente para doenças contagiosas em que existem interações de curto e longo alcance.

ABSTRACT

Situations where a disease cluster does not have a regular shape are fairly common. Moreover, maps with multiple clustering, when there is not a clearly dominating primary cluster, also occur frequently. We would like to develop a method to analyze more thoroughly the several levels of clustering that arise naturally in a disease map divided into *m* regions.

The spatial scan statistic is the usual measure of strength of a cluster. Another important measure is its geometric regularity. A genetic multi-objective algorithm was developed elsewhere to identify irregularly shaped clusters. That method conducts a search aiming to maximize two objectives, namely the scan statistic and the regularity of shape (the compactness concept). The solution presented is a Pareto-set, consisting of all the clusters found which are not worse in both objectives simultaneously. The significance evaluation is conducted in parallel for all the clusters in the Pareto-set through a Monte Carlo simulation. This procedure determines the best cluster solution.

Instead of using a genetic algorithm, we designed a novel method that incorporated the simplicity of the circular scan, being able to detect and evaluate irregularly shaped clusters. We define the circular occupation (CO) of a cluster candidate roughly as its population divided by the population inside the smallest circle containing it. The CO concept, being computationally faster, and relying on familiar concepts, is easier to grasp and substitutes here the compactness concept as another measure of regularity of shape. The scan statistic is evaluated for each of the *m* regions of the map taken individually. The regions are ranked accordingly in decreasing order. Let *R(k)* be the set containing the first *k* regions. A multi-objective modification of the circular scan algorithm [8] is successively applied for each set *R(k)*. For each circle, the candidate cluster consists of the regions belonging to *R(k)* within it, and the quotient in the CO calculation takes into account all the regions of the original map inside the circle. In practice we choose only some few *k* values such as

m,m/2,m/4,…,1. For each value of *k* we build the Pareto-set *P(k)*. We display all the Pareto-sets in a graph and after joining all of them we compute the global Pareto-set *P(0)*. A Monte Carlo procedure is used for significance evaluation.

The presence of "knees" in the Pareto-sets indicates sudden transitions in the clusters structure, corresponding to rearrangements due to the coalescence of loosely knitted (usually disconnected) clusters. Each Pareto-set contains the most likely clusters within a certain level of geographical information. They are related, reflecting the distribution of cases, populations and neighborhood structure of the map. Computationally, the method is only a few times slower than the usual circular scan.

The multi-objective circular scan allows peering into the clustering structure of a map. The comparison of Pareto-sets for observed cases with those computed under null-hypothesis provides valuable hints for the spatial occurrence of diseases. The potential for monitoring incipient clusters at several geographic scales simultaneously makes this a promising tool in syndromic surveillance, especially for contagious diseases when there is a mix of short and long range spatial interactions.

SUMÁRIO

LISTA DE FIGURAS

INTRODUÇÃO

É de interesse dos pesquisadores da área de saúde identificar áreas de riscos distintos em meio a regiões maiores com risco aproximadamente constante. Neste sentido, a estatística fornece ferramentas para o estudo de aglomerados de eventos, ou casos, espacialmente distribuídos (conglomerados ou clusters espaciais), que têm recebido bastante atenção na literatura. Faz-se necessário saber que, uma vez detectado o conglomerado de casos, pode dar-se início a estudos mais sofisticados na tentativa de se encontrar o meio causador destes casos. Lawson et al. (1999) apresentam várias situações, nas quais a detecção do conglomerado de doenças foi um passo importante no estabelecimento de sua etiologia até então desconhecida.

Hoje, a literatura apresenta vários métodos de detecção de conglomerados. Contudo, vale ressaltar que, alguns destes métodos se mostraram estatisticamente inapropriados. Outros, ainda que estatisticamente apropriados, carregam o problema de ajuste de múltiplos testes ou vício de pré-seleção (TANGO, 1999). Os métodos mais comuns e usuais partem do pressuposto de que existe um mapa dividido em regiões e que, para cada uma dessas regiões, é conhecida a população em risco e o número de casos observados.

Estes métodos, ou testes, utilizando janelas móveis que se superpõem à área em estudo, fazem a contagem do número de casos das regiões cujos centróides caem dentro da cada janela. Cada um dos possíveis conjuntos de regiões definidos pelas janelas é chamado de zona. Estes tipos de testes alteram sistematicamente o tamanho das janelas e avaliam a significância estatística do número de casos que caem dentro dela. Este procedimento permite que seja determinada a melhor zona que vai conter as regiões de riscos mais elevados dentro do mapa, de forma estatisticamente significativa. Para efeitos ilustrativos, quando o interesse é avaliar espacialmente o comportamento de determinada doença em um estado, cada município do mapa pode ser entendido como uma

região e cada observação de uma doença em um indivíduo desta população pode ser entendido como um caso.

CAPÍTULO 1. Revisão Bibliográfica

Nos últimos anos o estudo para detecção de conglomerados espaciais vem ganhando espaço na literatura e, assim, vários métodos foram propostos. Um conglomerado é uma área de risco significativamente distinto (elevado ou baixo), mas não explicado pelas covariáveis conhecidas. Para esta definição, em inglês, se usa o termo *cluster*.

Estes conglomerados são ditos puramente espaciais quando a ocorrência dos casos é mais alta em algumas áreas do que em outras. Quando a incidência dos casos é mais alta durante um determinado intervalo de tempo, esses conglomerados são puramente temporais. Quando a análise é feita levando-se em conta tanto o espaço quanto o tempo, ou seja, a ocorrência dos casos é temporariamente maior em algum local do que em outros locais, esses conglomerados são ditos espaciais - temporais.

Kulldorff(1995) destaca mais de 100 diferentes métodos de detecção de conglomerados, que estão classificados de acordo com as características e hipóteses feitas sobre o cluster. Besag e Newell (1991) classificaram os testes em gerais e focados. Nos testes focados de conglomerados, os dados são coletados para testar a hipótese de um possível excesso de casos ao redor de uma fonte suspeita e esta fonte deve ser identificada antes de observar os dados. Os testes gerais de conglomerados procuram identificar as áreas geográficas com um risco significativamente elevado sem especificar previamente quais e quantas áreas seriam estas.

Para os dois tipos de testes o modelo de nulidade é sempre o mesmo supondo que não há conglomerado na região, isto é, o risco é constante na área em estudo implicando que o número esperado de casos em uma região é proporcional ao número de pessoas em risco morando neste local. Segundo Wartenberg (1990), existem basicamente duas classes de cluster, *Hot-spot* e *Clinal*.

No *Hot-spot,* o risco é elevado e uniforme nas regiões que formam o conglomerado e fora destas regiões não há elevação do risco. No *Clinal,* o risco é elevado no centro do conglomerado e decresce à medida que vai se afastando do centro, e para regiões muito afastadas do centro o risco adicional é desprezível.

Considerando o centro do conglomerado como foco de risco à saúde, e havendo suposição de que há uma área em torno deste foco com uma taxa elevada e uniforme de doença então o melhor modelo adotado seria o *Hot-spot*. Se por outro, houver a suspeita de que a taxa de doença é elevada apenas em uma pequena região em torno do foco, seguida de um declínio ao longo do resto do mapa então, deve ser adotado um modelo da classe *Clinal*. Estas duas características são usadas para formular modelos estatísticos que investigam os conglomerados de eventos através dos testes de hipóteses.

Quando o objetivo da investigação for detectar pequenos clusters localizados, o modelo Hot-spot é o mais indicado. Este assume que a população está dividida em dois grupos formados por expostos e não-expostos, também conhecido como estudo caso-controle.

O primeiro método criado para detecção de conglomerados espaciais foi baseado em "quadrats", proposto por Choynowsky (1959). O interesse estava em estudar a distribuição espacial de casos de tumores no cérebro de uma certa região da Polônia, divida em municípios. Baseando-se em taxas brutas por área, este método apesar de simples, não levava em conta a variabilidade das taxas. Assim, áreas com uma pequena população tinham taxas com grande variabilidade. Este teste consistia em avaliar as áreas individualmente e determinar se o número de casos era significativamente alto considerando um nível de significância α . Com Choynowsky testando cada quadrante separadamente surge o problema de múltiplos testes. Outro problema deste método era a incapacidade de detectar clusters que não seguissem as delimitações geográficas dos municípios da região em estudo.

O GAM – Geographical Analysis Machine, desenvolvido por Openshaw et al. (1987) se baseia na idéia de Choynowsky. O método faz uso de múltiplos círculos de raio *R* sobrepostos, permitindo que os conglomerados possam ter formas diferentes daquelas impostas pelas delimitações geográficas dos municípios da região em estudo. Turbull et al. (1990), desenvolveram o CEPP – Cluster Evaluation Permutation Procedure, que também usa zonas circulares sobrepostas. Os círculos são construídos de forma que tenham o mesmo tamanho populacional *P*. Besag e Newell (1991), desenvolveram o teste TBN, em que o número de casos *K* é definido como o tamanho do conglomerado a ser procurado. Este método consiste em, fixado o tamanho *K*, centrar o círculo em um ponto na região, ir aumentando o seu raio e agregando os centróides vizinhos até que o círculo tenha agregado o menor número de centróides necessários para que o número de casos dentro do círculo tenha no mínimo *K* casos. Tango (1995) desenvolveu o teste C_λ ,onde o tamanho do conglomerado é determinado por λ que, neste caso, é o parâmetro de escala de alguma função $g(\lambda)$ que mede a proximidade entre as áreas pertencentes ao conglomerado.

Contudo, estes métodos apresentam um grande problema na definição à priori do parâmetro que caracteriza o tamanho do cluster: No GAM, o raio *R* do conglomerado, no CEPP, o raio *P* populacional, no TBN, o raio *K* de casos, e no C_λ , o parâmetro λ. Desta forma, os testes são repetidos usando valores diferentes para os parâmetros uma vez que as características do conglomerado em questão são desconhecidas. Conseqüentemente, além do vício de pré-seleção, os vários testes simultâneos resulta no problema de ajustes de múltiplos testes. Isto é quando se faz um teste para comparar dois ou mais parâmetros conjuntamente com um nível alfa de significância e a hipótese de igualdade entre os parâmetros é rejeitada, deve se fazer os testes para comparar os parâmetros dois a dois, e dessa forma considerar para cada teste o mesmo nível de significância, alfa, o nível de confiança considerando-se a independência dos testes, passa a ser $(1\text{-alfa})^n$, e conseqüentemente o nível de significância passa a ser maior que alfa.

Para resolver este problema e permitir uma avaliação global dos resultados, Kulldorff e Nagarwalla (1995) propuseram, a partir das idéias do GAM e CEPP, uma estatística para detectar áreas com elevada taxa de incidência. Baseado na razão de verossimilhança e utilizando uma estatística de varredura multidimensional, *scan statistic* em inglês, este método possui três propriedades básicas: geometria da área sendo varrida; a distribuição de probabilidade que gera os casos sob a hipótese de completa aleatoriedade espacial; tamanho e forma da janela de varredura.

Com a estatística de varredura espacial, a janela, de forma e tamanho variáveis, é movida sobre a área geográfica em estudo. Cada forma, tamanho e localização define uma área candidata a ser um conglomerado. Para cada área candidata, a verossimilhança é calculada baseada na observação e número esperado de casos dentro e fora dessa área.

A área com a máxima verossimilhança define o conglomerado mais provável. A significância estatística deste conglomerado é determinada gerando um grande número de conjunto de dados aleatórios sobre a hipótese nula, e então calculando a máxima verossimilhança para cada um destes conjuntos aleatórios de dados exatamente da mesma forma como a calculada para os dados reais.

Quando se aplicar a estatística de varredura espacial, uma escolha natural para a forma da janela é a circular (scan circular), esta é a forma mais compacta que pode ser obtida. Isto tem sido usado na prática (KULLDORFF, 1997). Entretanto, ao usar janelas circulares, o método apresenta algumas deficiências. É possível identificar um conglomerado maior do que o real, se o conglomerado real tiver um formato muito diferente de um círculo. Nesse caso, será encontrado conglomerado formado por áreas compactas englobando muitas regiões que de fato não fazem parte do conglomerado real. Aqui, temos um problema de superestimação do conglomerado como mostra a Figura 1. A Figura 2 mostra um outro problema, o de subestimação do conglomerado, ou seja, é possível encontrar um conglomerado pequeno que inclui poucas regiões do conglomerado real.

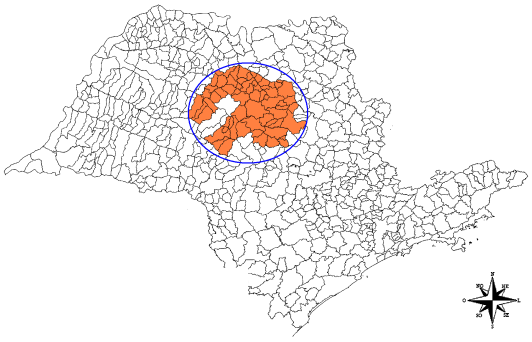

***Figura 01**: Superestimação de conglomerado*

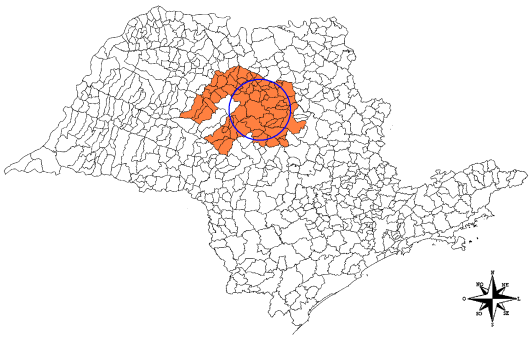

***Figura 02**: Subestimação de conglomerado*

O teste passa a ter um baixo poder se existem pequenos clusters isolados na região em estudo. Para solucionar o primeiro problema, o uso da razão de verossimilhança ponderada, no qual os pesos funcionam como um fator de correção para o teste, foi proposto por Ronald e Murray (2001). Duczmal & Assunção (2003) acrescentaram no mapa uma estrutura de grafo e varreram o mapa utilizando o Simulated Annealing (SA). Neste caso as áreas encontradas como conglomerados são zonas conectadas com formato irregular. Este método detecta conglomerados de formato geométrico arbitrário e com um poder não muito menor do que o método de Kulldorff para janelas circulares.

Outras formas também são possíveis de serem usadas, como por exemplo, elipses, quadrados ou triângulos. Estes métodos podem ter um alto poder se a forma do conglomerado verdadeiro for não circular, o que freqüentemente pode ocorrer. Kulldorff (2003) propôs o scan elíptico que usa elipses ao invés de círculos no formato das janelas de varredura. No caso de subestimação de conglomerados não há estudo conhecido.

CAPÍTULO 2. Método scan circular

O método scan circular tem sido bastante estudado e testado em diversas situações. A principal deficiência mostrada pelo método tem sido sua incapacidade de detectar corretamente um cluster de forma irregular, ou seja, cluster com forma muito diferente da circular. Na verdade, o método scan circular detecta o cluster, mas o cluster encontrado é uma superestimação ou uma subestimação do cluster real. Em certas situações, como será mostrado em um exemplo o cluster real possui 8 regiões e, o scan circular indica um cluster com até 34 regiões.

Muitos métodos têm sido desenvolvidos para detecção de cluster de forma arbitrária, mas, no geral, o scan circular tem mostrado ser mais eficiente que todos estes. Alguns métodos são bastante eficientes na detecção de cluster de forma irregular, mas têm deficiências na detecção de cluster com forma próxima a circular. Outros métodos carregam o problema da superestimação ser ainda maior que a do scan circular. Outro problema da maioria dos novos métodos desenvolvidos é a demora no tempo de execução do algoritmo, o que muitas vezes dificulta a sua popularização. Nas seções seguintes serão descritos os dois métodos de varredura circular de Kulldorff, que utilizam os modelos Binomial e Poisson.

2.1. Modelo Binomial

Seja S uma área dividida em m regiões. Para cada uma dessas regiões é necessário conhecer as coordenadas geográficas, o número de indivíduos e número de casos. Tome z como uma única zona criada pelo método, Z como o conjunto de todas as possíveis zonas circulares criadas por este método, p como a probabilidade de um indivíduo pertencente a zona z vir a ser um

caso e q a probabilidade de indivíduo fora da zona *z* vir a ser um caso. Sob a hipótese nula de que não há nenhum conglomerado na região, temos que $p = q$, neste caso a hipótese alternativa supõe a existência de uma única zona z, tal que $p > q$, assim temos:

$$\begin{cases} H_0 : p = q \\ H_1 : p > q, \quad z \in Z \end{cases}$$

Seja N o número total de indivíduos sob risco na área *S*, C o número total de casos na área *S*, n_z o número de indivíduos sob risco na zona z e c_z o valor observado da variável aleatória C_z, que representa o número de casos na zona z. Admita ainda que o modelo Bernoulli seja apropriado para o número de casos. Com isso a função de verossimilhança para a zona z será dada por:

$$L(z,p,q) = p^{c_z}(1-p)^{n_z - c_z} q^{C-c_z}(1-q)^{(N-n_z)-(C-c_z)}, \qquad (01)$$

O valor de p que maximiza a verossimilhança não é necessariamente aquele que corresponde a maior taxa $\hat{p} = \frac{c_z}{n_z}$, nem aquele com o maior número de casos c_z. Para identificar a zona mais provável de ser o conglomerado, dentre todas as possíveis, o teste proposto por Kulldorff e Nagarwalla usa a razão de verossimilhança,

$$\lambda(z) = \frac{\sup_{z \in Z, p>q} L(z,p,q)}{\sup_{p=q} L(z,p,q)}, \text{ com } \{p,q \in (0,1)\} \qquad (02)$$

Sob H_0 os EMV's de p e q são dados por $\hat{p} = \hat{q} = C/N$. Então o denominador da equação 02 é reduzido a

$$\sup_{p \in (0,1)} p^C (1-p)^{N-C} = \frac{C^C (N-C)^{N-C}}{N^N} = L_0. \qquad (03)$$

Observe que L_0 é constante e depende somente do número total de casos e não da sua distribuição espacial. Sob a hipótese alternativa os valores de p e q que maximizam a verossimilhança, para

uma zona fixa z sobre o espaço $0 < q < p < 1$, são dados por $\hat{p}(z) = c_z / n_z$ e $\hat{q}(z) = (C - c_z)/(N - n_z)$, se $\frac{c_z}{n_z} > \frac{C - c_z}{N - n_z}$ Daí segue que,

$$L(z) = \begin{cases} \left(\frac{c_z}{n_z}\right)^{c_z}\left(\frac{n_z - c_z}{n_z}\right)^{n_z - c_z}\left(\frac{C - c_z}{N - n_z}\right)^{C - c_z}\left(\frac{(N - n_z) - (C - c_z)}{N - n_z}\right)^{(N - n_z) - (C - c_z)} & se \quad \frac{c_z}{n_z} > \frac{C - c_z}{N - n_z} \\ \frac{C^C (N - C)^{N - C}}{N^N} & se, \quad \frac{c_z}{n_z} \leq \frac{C - c_z}{N - n_z} \end{cases} \qquad (04)$$

Desta forma a equação 02 pode ser escrita como:

$$\lambda(z) = \begin{cases} L(z)/L_0, & se \quad (c_z/n_z) > (C - c_z)/(N - n_z) \\ 1 \quad , & se \quad (c_z/n_z) \leq (C - c_z)/(N - n_z) \end{cases} \qquad (05)$$

Para detectar a zona como sendo o conglomerado mais provável é escolhido a zona $\hat{z}$ para a qual a $L(z, p(z), q(z))$ é maximizada. A distribuição de $\lambda(z)$, sob H_0, é obtida via simulação de Monte Carlo.

2.2. Modelo Poisson

Se considerarmos que os dados seguem uma distribuição de Poisson, e definindo $\mu(z)$ como sendo o número esperado de casos na zona z, sob a hipótese nula é possível mostrar que:

$$\lambda(z) = \begin{cases} \left(\frac{c_z}{\mu(z)}\right)^{c_z}\left(\frac{C - c_z}{C - \mu(z)}\right)^{C - c_z}, & se \quad c_z > \mu(z) \\ 1 \quad , & caso \quad contrário. \end{cases} \qquad (06)$$

onde $\mu(z) = n_z(C/N)$ e $\mu(\bar{z}) = (N - n_z)(C/N)$ (KULLDORFF, 1997).

Note que a equação (06) é equivalente a (04).

Neste modelo podem ser incorporados fatores de riscos tais como sexo, idade, raça, nível escolar etc.

A escolha do modelo Bernoulli ou Poisson vai depender dos dados em estudos. Para um número pequeno de casos, até 10% da população sob risco, os modelos se aproximam um do outro. Por outro lado, se temos um estudo caso-controle é preferível usar o modelo Bernoulli e caso exista algum fator de risco é preferível usar o modelo de Poisson.

Para detectar a zona como sendo o conglomerado mais provável é escolhido a zona $\hat{z}$ para o qual a $L(z, p(z), q(z))$ é maximizada. A distribuição de $\lambda(z)$, sob H_0, é obtida via simulação de Monte Carlo.

2.3. Algoritmo do Método Circular

Nesta seção, é mostrado o algoritmo circular para a detecção de cluster. Esse funciona com segue:

1. Escolher um ponto na região em estudo.

2. Calcular as distâncias até os outros pontos, ordenando-as em ordem crescente, e guardando-as em um vetor.

3. Para cada ponto da região repetir os passos 1 e 2.

4. Escolher novamente um ponto da região.

5. Criar um círculo centrado no ponto escolhido no passo 4 e continuamente aumente o seu raio de acordo com as distâncias encontradas no passo 2. para cada ponto que entrar no círculo atualize o número de casos c_z e a população n_z dentro do círculo Z.

6. Repetir os passos 4 e 5 para cada ponto. Calcule λ para cada par (c_z, n_z) usando um dos modelos Bernoulli ou Poisson. Registre o círculo com maio λ.

7. Utilizar simulações Monte Carlo para avaliar a significância do teste.

7.1 Gerar B conjuntos de dados independentes, em cada réplica aleatória possui o mesmo número de casos C que o conjunto de dados original. Estes C casos são distribuídos ao acaso entre as m áreas de acordo com a hipótese nula.

7.2 Em cada um dos B conjuntos de dados gerados, calcular a estatística do teste da razão de verossimilhança obtendo$\lambda_1, \lambda_2, \ldots \lambda_B$.

7.3 Ordenar os valores de λ dos B conjuntos simulados e observados no conjunto de dados original. Denote o posto da estatística λ associado ao conjunto de dados original por R. Se R estiver entre os $100\alpha\%$ maiores postos, rejeite a hipótese nula ao nível de significância de α. O valor-p associado a este teste é 1- R/(B + 1).

7.4 Se a hipótese nula for rejeitada, então a zona $\hat{Z}$ associada com a máxima verossimilhança do modelo alternativo é o cluster mais provável.

CAPÍTULO 3. Método Scan Circular Multi-Objetivo

Neste capítulo apresentaremos um novo método para detecção de clusters espaciais. O novo método é uma generalização do método scan circular, com o objetivo de encontrar clusters de qualquer formato, e não somente os de formato circular. Este método fornece como solução um conjunto de clusters, em que cada um dos clusters é um elemento de um conjunto especial, chamado conjunto de Pareto.

Estudos anteriores já mostraram que o método scan circular apresenta um grande poder para detecção de cluster único, ou seja, quando existe um único cluster bem definido no mapa em estudo. No entanto esse poder diminui na presença de clusters múltiplos. O método scan circular, ao buscar clusters em formato de círculo, pode ter seu poder de detecção reduzido quando no mapa existe um cluster de formato irregular. Como já visto anteriormente, este método faz uma varredura completa no mapa através de círculos, ou seja, as possíveis regiões candidatas a serem clusters no mapa aparecerão sempre com um formato próximo de um círculo. Isso pode gerar super-estimações ou sub-estimações do cluster real, principalmente quando o cluster tem formato irregular. O método proposto neste trabalho é capaz de detectar clusters de formato irregular, mesmo fazendo o uso do círculo para fazer uma varredura completa do mapa. O método scan circular multi-objetivo baseia-se em dois novos conceitos: o *conjunto seletivo* e a *ocupação circular*. Esses conceitos serão definidos formalmente nas próximas seções.

Informalmente, para construirmos os conjuntos seletivos procedemos da seguinte maneira. Dado um mapa com *m* regiões, consideremos inicialmente cada uma dessas regiões como sendo uma zona e calculamos, usando a estatística scan, o valor de logaritmo da razão de verossimilhança (LLR) em cada uma dessas zonas. Em seguida colocamos em ordem decrescente todos os *m* valores dos LLR obtidos. O conjunto correspondente, por exemplo, às 5% regiões de maiores LLR no mapa formarão nosso primeiro *conjunto seletivo*, ou *mapa seletivo*. Essas regiões são ditas regiões *ativas*

e as demais 95% regiões são ditas regiões *inativas*. Por outro lado, se escolhermos as 10% regiões de maiores LLR, teremos um conjunto seletivo com as 10% regiões de maiores LLR. Aumentamos as porcentagens para 15%, 20%, 25%, ... , até atingir 100%, que é o conjunto seletivo com todas as regiões do mapa. Observe que o conjunto seletivo com as 100% regiões de maiores LLR é exatamente o mapa original. Percebemos também que o primeiro conjunto seletivo está contido no segundo conjunto seletivo, e assim por diante, até chegar ao mapa original, que contém todos os subconjuntos seletivos anteriores. Usando somente as regiões ativas, aplicamos o método scan circular para construirmos as possíveis zonas candidatas a clusters. Para cada uma dessas zonas calculamos a *ocupação circular*, que é uma medida da presença populacional e geométrica da zona assim construída no mapa seletivo, em relação ao mapa original (mapa com todas regiões). Podemos também dizer que um conjunto seletivo com r regiões filtra a informação dessas r regiões com maiores LLR no mapa. Quando os conjuntos seletivos possuem poucas regiões, essas regiões frequentemente aparecem distribuídas de maneira desconexa. A medida que olhamos para conjuntos seletivos com maior número de regiões, esses conjuntos seletivos vão geralmente se tornando mais conexos. É interessante observar que os clusters irregulares frequentemente aparecem em conjuntos seletivos com poucas regiões. Os clusters com formatos mais próximos aos circulares aparecem em conjuntos seletivos com maior número de regiões.

As figuras 03, 04, 05, 06, 07, 08, 09, 10, 11, 12 mostram conjuntos seletivos formados com as 0.4%, 0.8%, 1.6%, 3,2%, 6,4%, 12,5%, 25%, 50%, 100% regiões de maiores verossimilhanças no mapa de taxa de homicídios nos municípios de Minas Gerais acumulados durante os anos de 1998 a 2002.

***Figura 03**: Conjunto Seletivo Com 0.40% das Regiões Ativas*

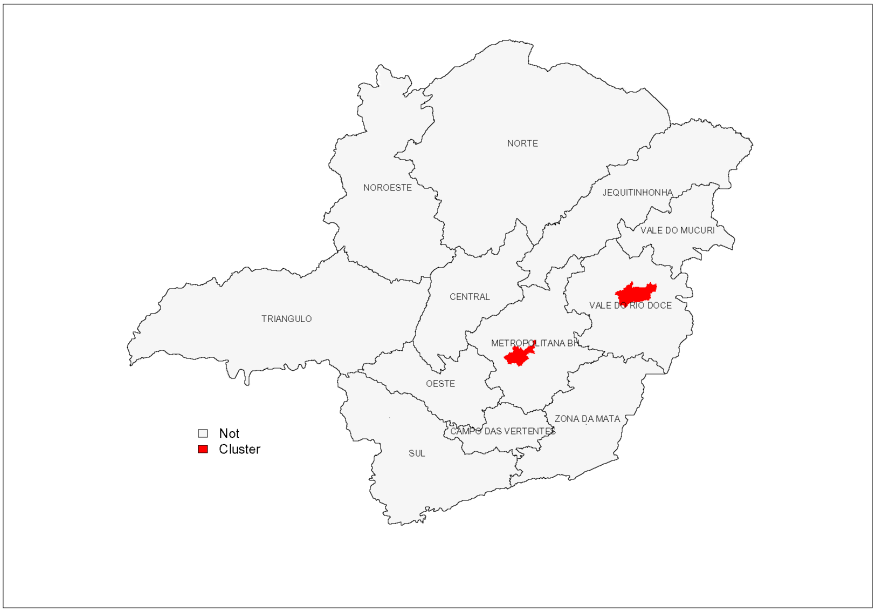

***Figura 04**: Conjunto Seletivo Com 0.80% das Regiões Ativas*

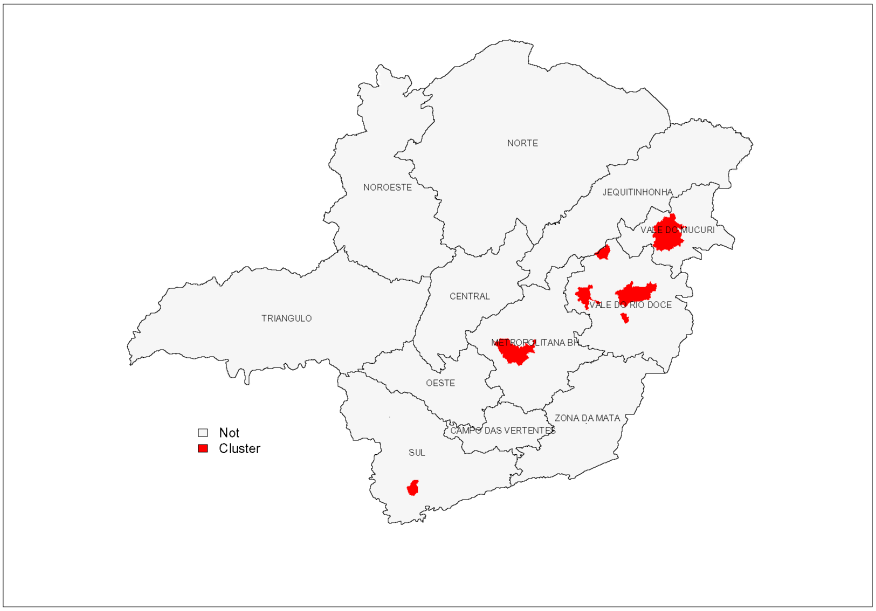

***Figura 05**: Conjunto Seletivo Com 1.60% das Regiões Ativas*

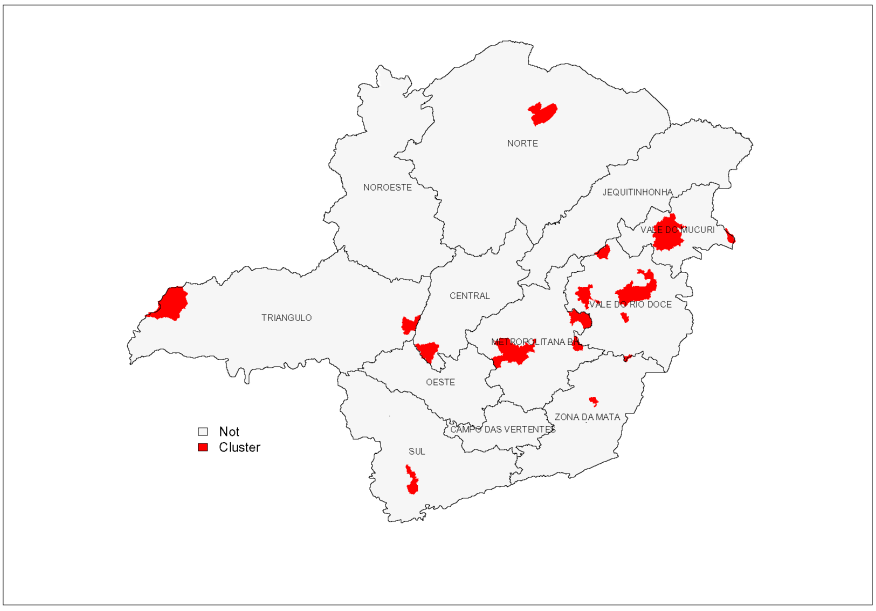

***Figura 06**: Conjunto Seletivo Com 3.20% das Regiões Ativas*

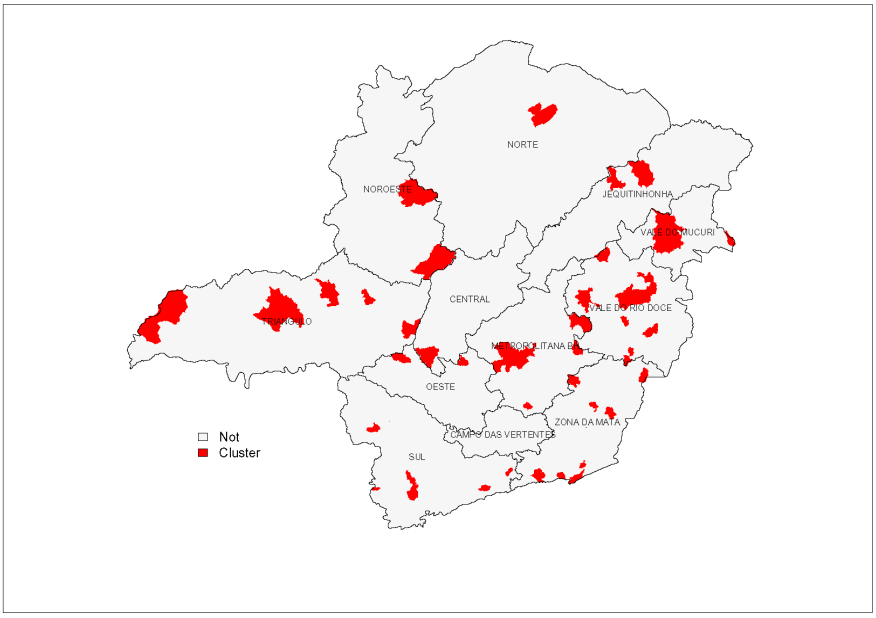

***Figura 07**: Conjunto Seletivo Com 6.40% das Regiões Ativas*

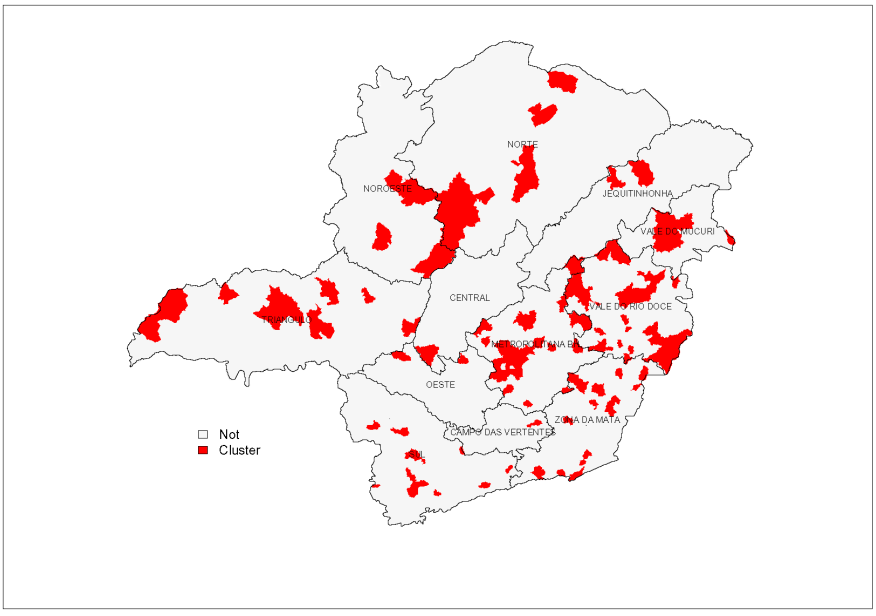

***Figura 08**: Conjunto Seletivo Com 12.50% das Regiões Ativas*

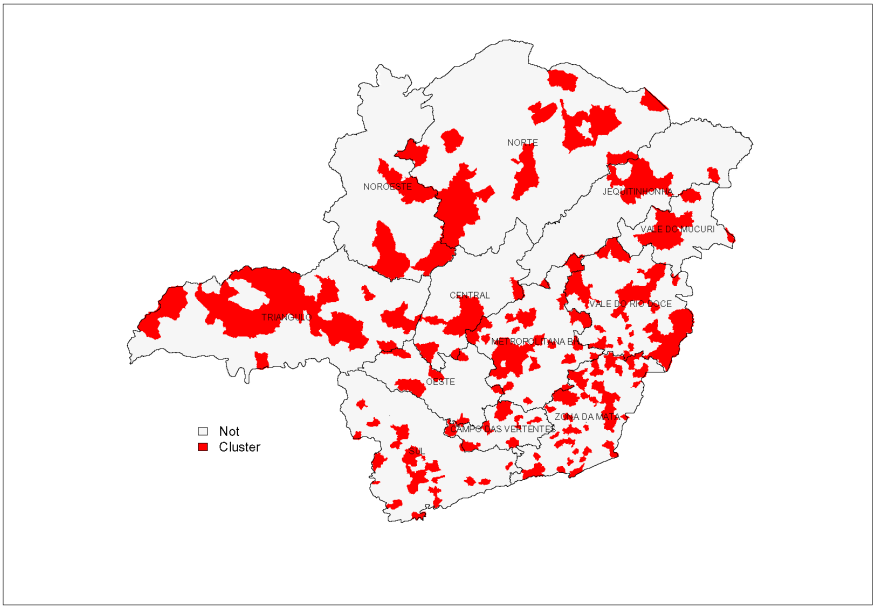

***Figura 09**: Conjunto Seletivo Com 25% das Regiões Ativas*

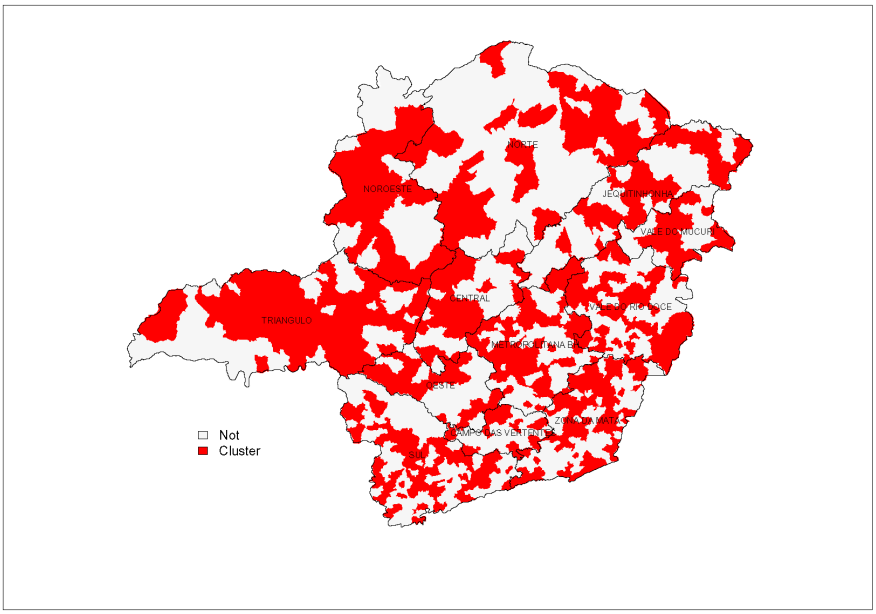

***Figura 10**: Conjunto Seletivo Com 50% das Regiões Ativas*

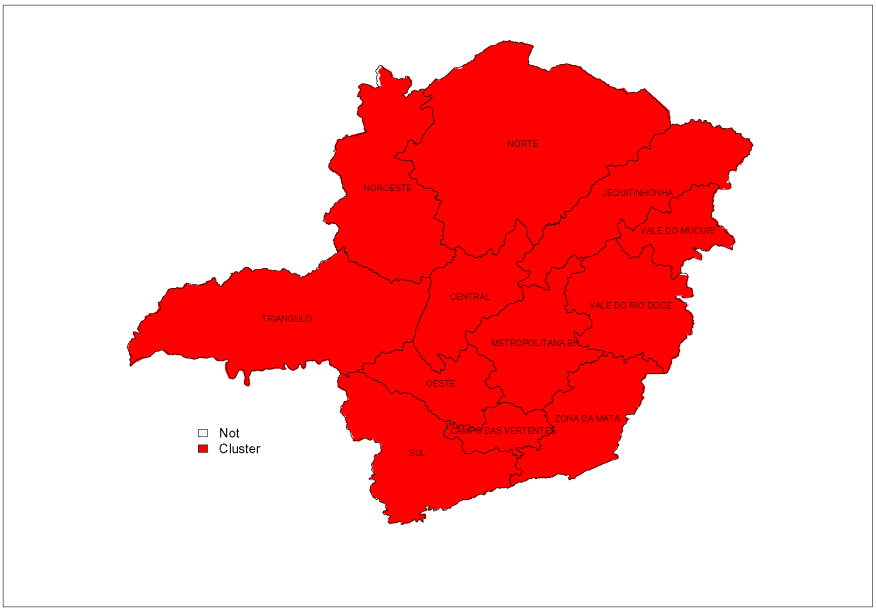

***Figura 11**: Conjunto Seletivo Com 100% das Regiões Ativas.*

Outro conceito importante que aparece com o algoritmo scan circular multi-objetivo é o conceito de ocupação circular. A ocupação circular é uma medida da presença populacional e geométrica de um cluster em relação ao mapa original. Dado um conjunto seletivo S e um círculo C, seja Z a zona formada pelas regiões de S cujos centróides estão dentro de C. Seja P(Z) a população de Z e seja P(C) a população somada de todas as regiões do mapa original cujos centróides estão dentro de C. Gostaríamos de definir OC(z), a ocupação circular de Z pelo quociente P(Z)/P(C). No entanto existe um problema com essa definição. Da maneira como está definida, a ocupação circular de uma zona pode não ser única. As figuras 12, 13 e 14 ilustram a definição da ocupação circular para uma mesma zona Z, em que o centro do círculo é escolhido de modo diferente, de acordo com o centróide de cada uma das três diferentes regiões componentes. Contornamos esse problema redefinindo a ocupação circular OC(z) como o máximo dos quocientes da população da zona Z pela população de cada círculo. Em cada um dos círculos dividimos a população das regiões pintadas de preto, (zona que queremos calcular a presença geométrica no mapa) pela população total do círculo (população das regiões pintadas de preto somada com a população das regiões pintadas de cinza), regiões cujos centróides caem dentro do círculo. O círculo da figura 13 possui a menor população, e como o numerador é constante para os três círculos, pois é exatamente a população da zona pintada de preto, temos que o quociente do circulo da figura 14 será o maior entre os três quocientes calculados. Esse quociente é o que chamamos de ocupação circular da zona formada pela regiões pintadas de preto. A ocupação circular é um número real entre 0 e 1. Naturalmente, clusters com ocupação circular próxima de 1 são mais regulares, e aqueles com ocupação circular próxima de zero são irregulares ou mesmo desconexos.

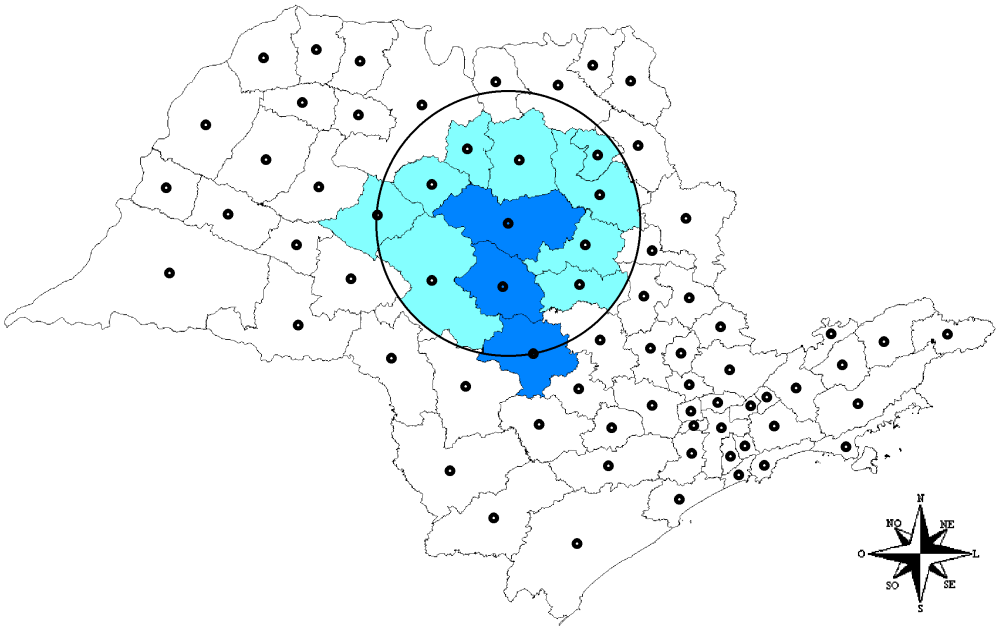

***Figura 12**: Circulo com Centro na Região Ativa Superior*

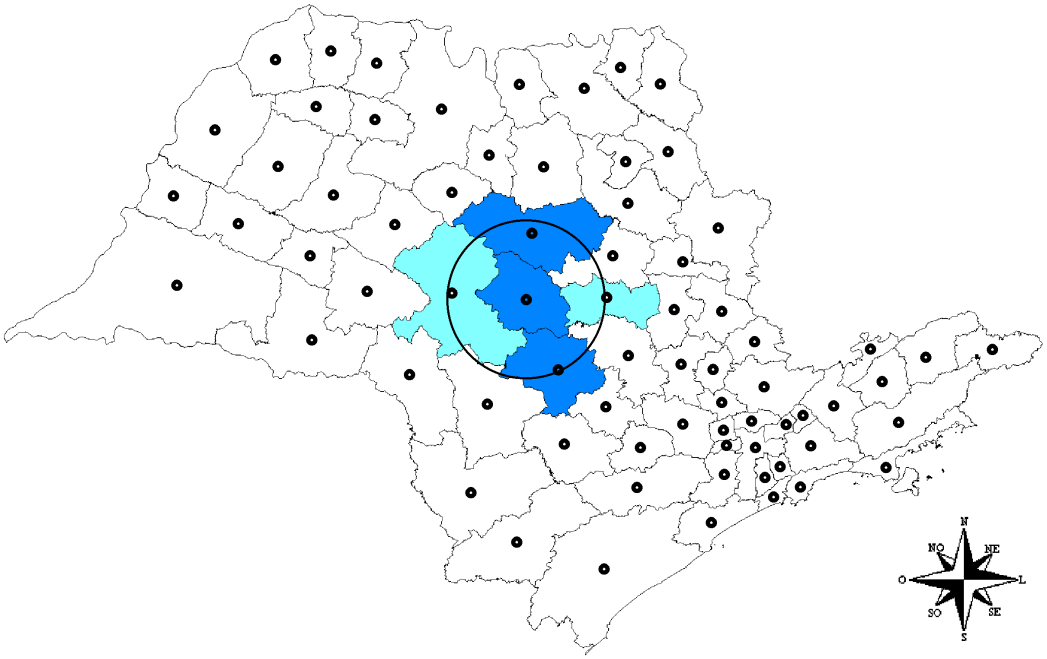

***Figura 13**: Círculo com Centro na Região Ativa Central*

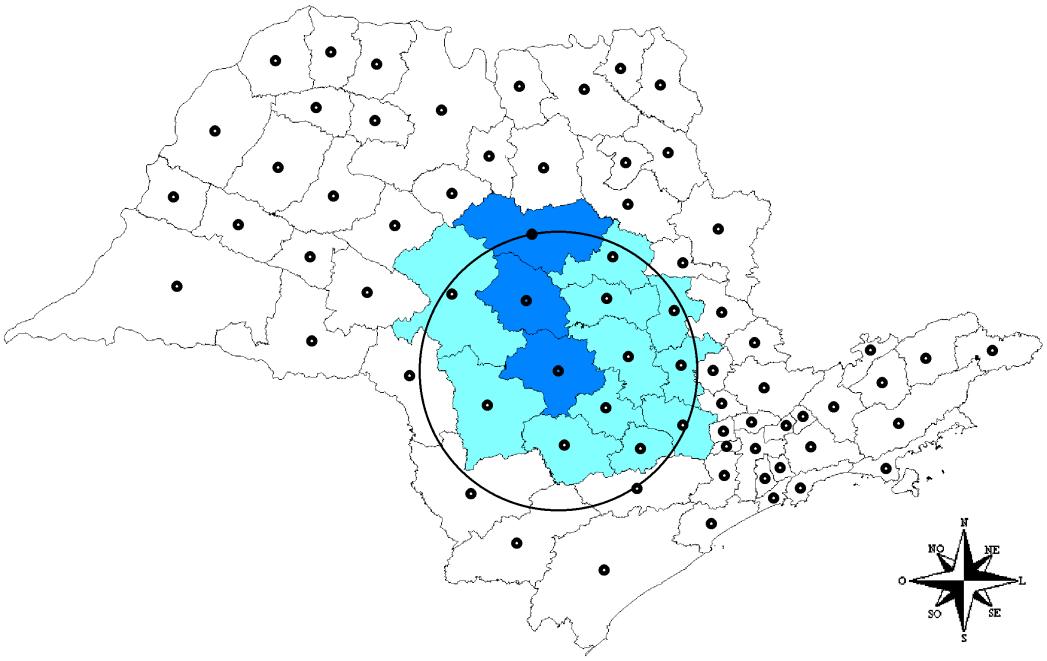

***Figura 14**: Círculo com Centro na Região Ativa Inferior*

O propósito do algoritmo scan circular multi-objetivo é encontrar zonas Z no conjunto de todas as zonas possíveis, numa tentativa de maximizar dois objetivos: a razão de verossimilhança e a ocupação circular. Para isso usaremos um conceito desenvolvido por Pareto, chamado de *conjunto de Pareto*. O conjunto de Pareto pode ser visto do seguinte modo. Seja M contido em RxR um conjunto finito de pares ordenados (x,y). Um par ordenado (X_1, Y_1) é pior que um par ordenado (X_2, Y_2) na direção x se $X_1 < X_2$. Analogamente, um par ordenado (X_1, Y_1) é pior que um par ordenado (X_2, Y_2) na direção y se $Y_1 < Y_2$.. Um par ordenado é dito ser pertencente ao conjunto de Pareto de M se esse par ordenado não for pior que nenhum outro par ordenado de M simultaneamente nas duas direções x e y. Assim, seja (X_1, Y_1) pertencente ao conjunto de Pareto de M. Se existir um outro ponto (X_2, Y_2) tal que $X_1 < X_2$ então $Y_1 >= Y_2$. Se existir um outro ponto (X_3, Y_3) tal que $Y_1 < Y_3$ então $X_1 >= X_3$. Na figura 15 , o conjunto de Pareto, indicado pelos círculos abertos, consiste nos pontos que não são simultaneamente piores que nenhum outro ponto em ambos objetivos x e y.

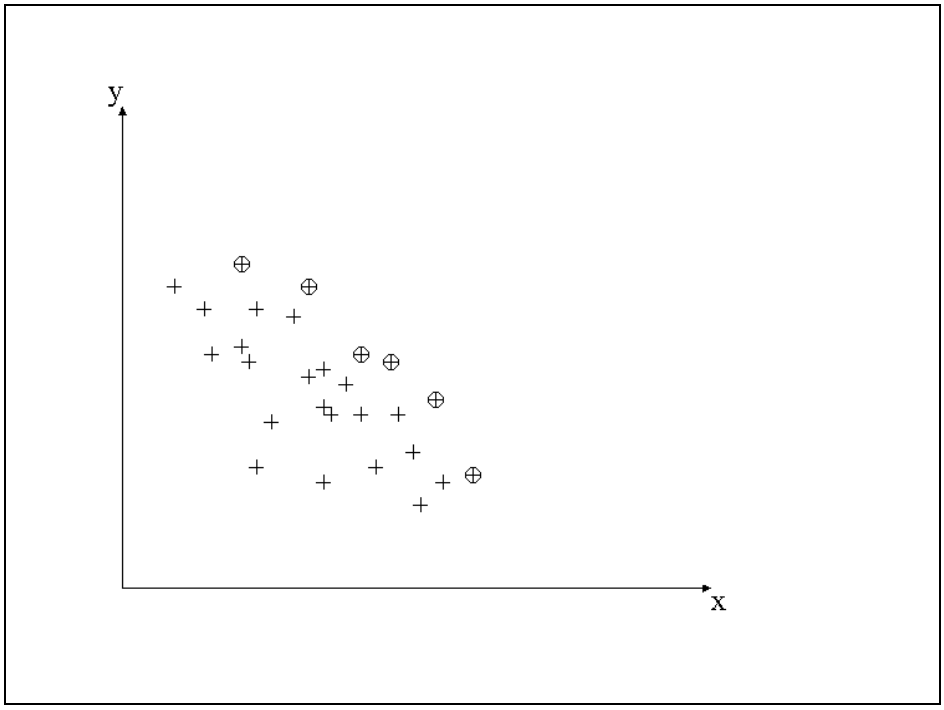

***Figura 15**: Conjunto de Pareto.*

O método scan circular multi-objetivo analisa separadamente cada um dos conjuntos seletivos do mapa. Consideremos inicialmente um conjunto seletivo como, por exemplo, o conjunto seletivo contendo as 5% regiões do mapa de maiores LLR. Calculamos a distância entre todas essas regiões, e as guardamos em uma matriz de distâncias, ordenadas em ordem crescente de distância. Centrado em cada uma das regiões ativas (regiões do mapa seletivo) construímos através de círculos todas as possíveis zonas, do mesmo modo que é feito no método scan circular. Repetimos esse procedimento para cada uma das regiões ativas. Para cada uma dessas zonas z assim construídas, calculamos LLR(z) e OC(z). Obtemos com esse procedimento um conjunto finito de pontos, ou pares ordenados (LLR(z), OC(z)), correspondentes a todas as zonas z obtidas. A otimização multi-objetivo consiste em maximizar dois objetivos: o primeiro objetivo é LLR(z), e o segundo objetivo é OC(z). Nosso próximo passo será extrair os melhores pontos desse conjunto; em outras palavras, queremos encontrar as zonas que são os clusters mais prováveis. Para isso obtemos o conjunto de Pareto desse conjunto de pontos.

Finalizado essa primeira parte, repetimos tudo que foi feito no primeiro conjunto seletivo para os demais conjuntos seletivos. Para cada novo conjunto seletivo analisado obtemos um novo conjunto de Pareto. Se tivermos s conjuntos seletivos, obtemos s conjuntos de Pareto. Fazendo a união de todos esses conjuntos, obtemos um novo conjunto formado pelos s conjuntos de Pareto. A partir desse conjunto extraímos um novo conjunto de Pareto, denominado Pareto dos Paretos, ou Pareto Global. Cada ponto do Pareto Global representa uma zona candidata a ser um cluster no mapa.

Para avaliar a significância de cada cluster usamos um procedimento Monte Carlo. Dado um mapa com *m* regiões, seja *n* o número de casos no mapa. Sob a hipótese nula de que não existe cluster no mapa, distribuímos aleatoriamente através da distribuição multinomial os *n* casos sobre o mapa. O número esperado de casos em cada região é proporcional à população de cada região. Aplicamos o algoritmo scan circular multi-objetivo acima descrito para analisar este novo mapa de

casos simulados, e como resposta, obtemos um conjunto de Pareto Global. Repetimos esse procedimento 9999 vezes, obtendo assim 9999 conjuntos de Pareto Globais. Representado num gráfico LLR x OC os pontos correspondentes aos clusters encontrados, teremos uma nuvem de pontos que será chamada de nuvem de Pareto de casos simulados. Observe que essa nuvem está contida na faixa semi-infinita $(0,\infty)\times(0,1]$. Veja a figura 16 que mostra uma nuvem de Pareto obtida pelo scan multi-objetivo quando executado 9999 vezes.

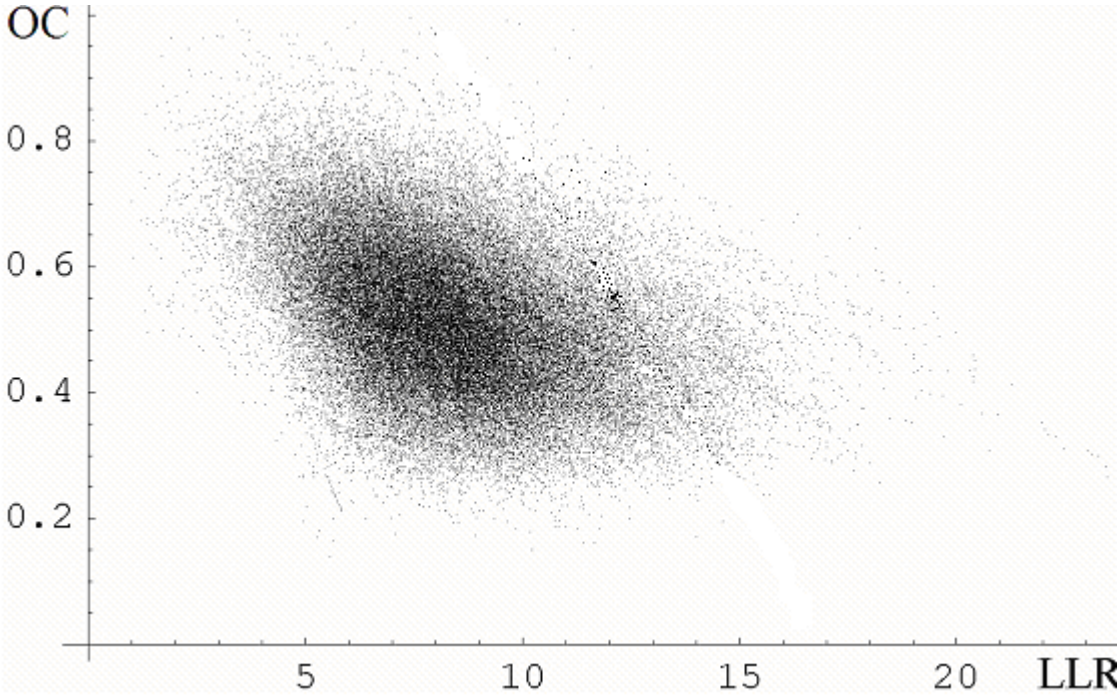

***Figura 16**: Nuvem de Pareto de casos simulados sobre a hipótese nula.*

Nossa preocupação agora será atribuir um valor-p para cada cluster do Pareto Global do mapa de casos observados. Para isso, vamos dividir a faixa $(0,\infty)\times(0,1]$ em faixas paralelas horizontais, por exemplo faixas da forma $(0,\infty)\times(s_i,s_{i+1})$, onde $s_i = i/10$, $i = 0,\ldots,9$. Dentro de cada uma dessas faixas estudamos o comportamento de cada ponto com relação ao LLR. Se um ponto $P_0 = (OC_0, LLR_0)$ do Pareto Global do mapa de casos observados cai numa faixa $(0,\infty)\times(s_i,s_{i+1})$, verificamos qual a porcentagem dos pontos da nuvem pertencentes a esta faixa

que possui LLR maior ou igual que LLR_0. Esse valor é a estimativa do valor-p do cluster correspondente a P_0. Repetimos esse procedimento para todos os pontos do Pareto Global do mapa de casos observados.

Nós estamos interessados em valores-p muito pequenos, como ocorre frequentemente em mapas com dados reais. Nesse caso, como seria exigido um grande número de simulações para obter alguns poucos pontos simulados que ficassem à direita dos pontos observados, esse problema é contornado fazendo uso de uma distribuição paramétrica que aproxime a distribuição empírica da nuvem de pontos. Aqui usaremos a distribuição Gumbel, que é a distribuição de valores extremos. Veremos que essa distribuição paramétrica se ajusta bem à distribuição empírica da nuvem de pontos em cada faixa. Usamos os valores de LLR dos pontos de cada faixa para calcular os parâmetros da distribuição Gumbel correspondente a essa faixa e a partir daí usamos esta distribuição paramétrica para calcular os valores-p muito pequenos dos pontos correspondentes aos clusters de casos observados. A figura 17 mostra a aproximação da distribuição Gumbel em algumas faixas de OC.

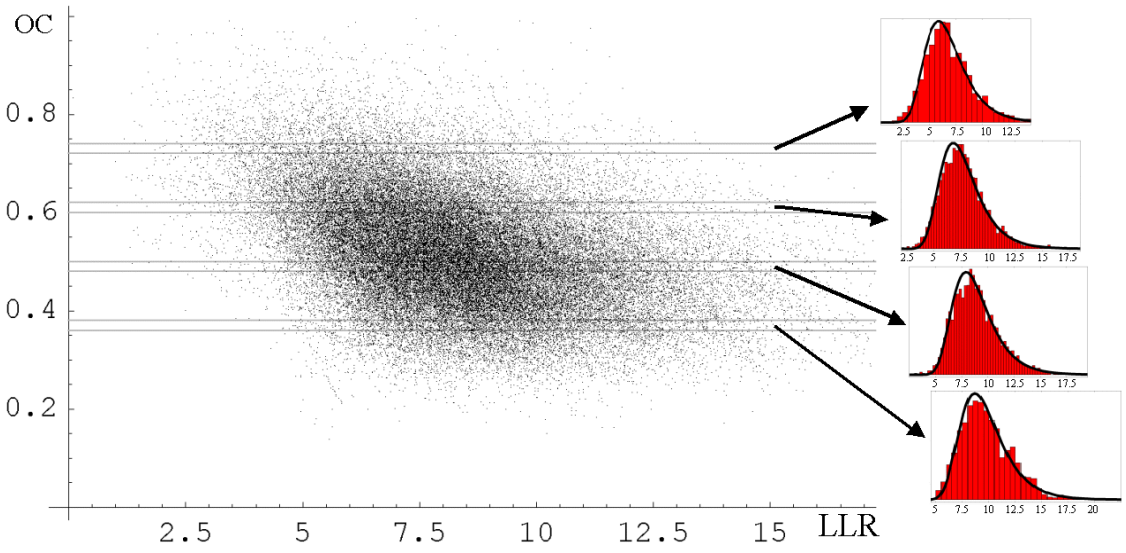

***Figura 17**: Aproximação da Distribuição Gumbel.*

Como veremos adiante, existe uma maneira mais precisa de se calcular os valores-p dos pontos correspondentes aos clusters de casos observados. A partir da família de distribuições de Gumbel para cada uma das faixas, construiremos uma superfície bidimensional no domínio da faixa $(0,\infty) \times (0,1]$. A partir dessa superfície construímos as curvas de nível de valor-p, ou isolinhas de valor-p. Os pontos do conjunto de Pareto dos dados observados são comparados com as isolinhas construídas a partir dos dados simulados sob hipótese nula. Um cluster é dito mais significativo que um outro se ele estiver mais à direita em relação às isolinhas. As figuras 18A e 18B mostram as isolinhas de valor-p construídas com dados de homicídios em Minas Gerais de 1998 a 2000. Para construir esse gráfico, foram utilizados 999 Paretos globais com o procedimento Monte Carlo.

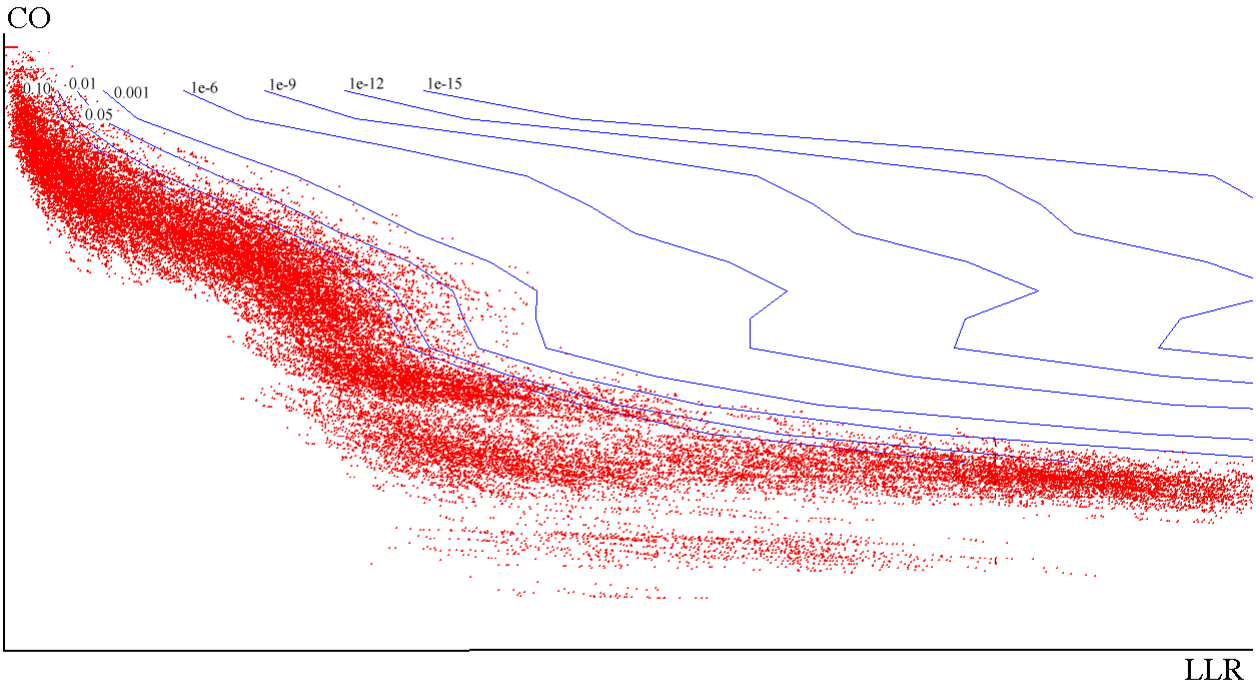

***Figura 18a**: Isolinhas de Valor-P para o mapa de homicídios de Minas Gerais. Os números indicam o p-valor das isolinhas.*

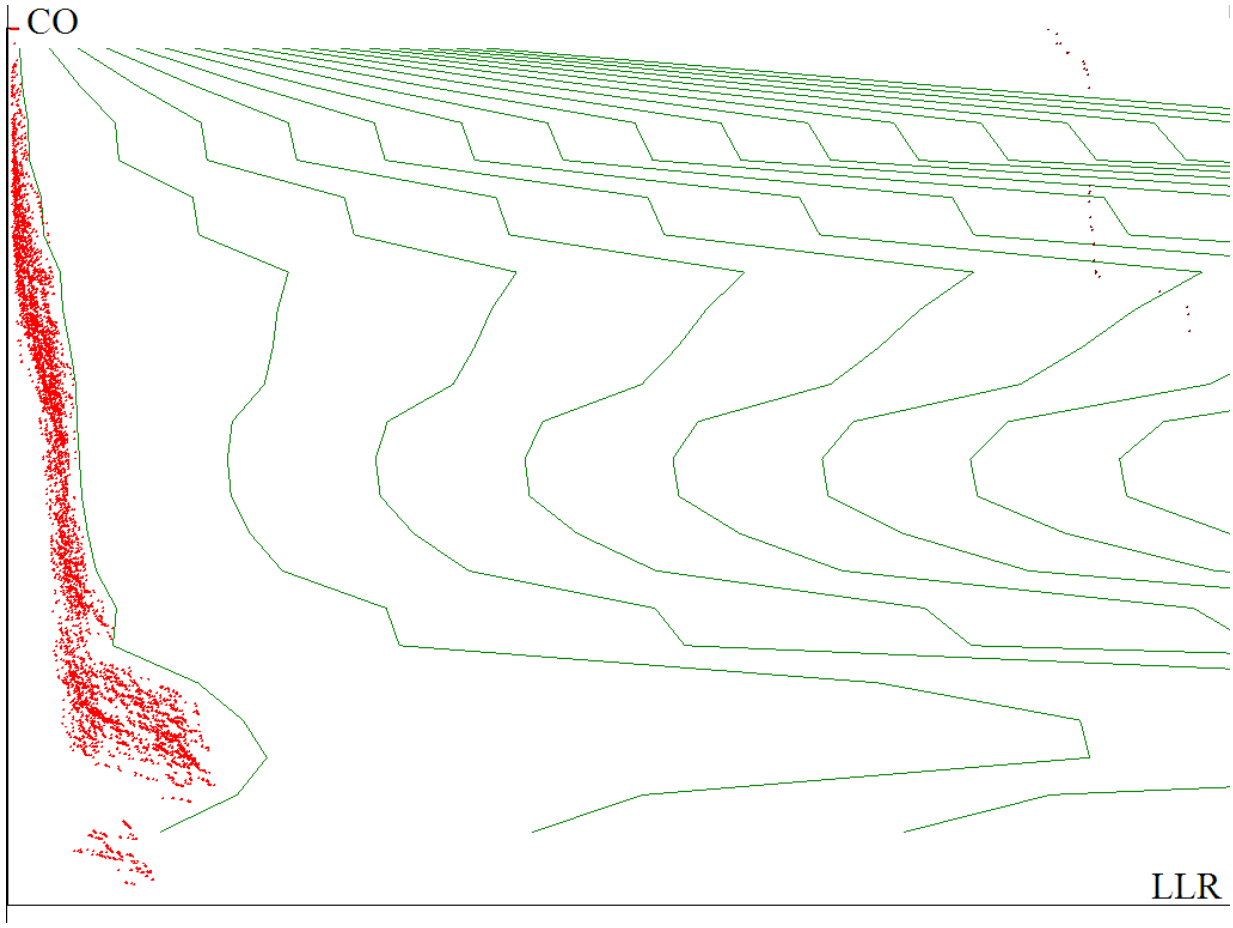

***Figura 18b**: Isolinhas de Valor-P do mapa de homicídios de Minas Gerais, incluindo os pontos de casos observados, acima à direita.*

Neste capítulo vamos formalizar rigorosamente essas idéias. Na seção 3.1 definiremos os subconjuntos seletivos das regiões do mapa original. Na seção 3.2 definiremos ocupação circular. Na seção 3.3 definiremos o conjunto de Pareto. Na seção 3.4 falaremos da aproximação pela distribuição Gumbel. Na seção 3.5 falaremos da significância do cluster e na seção 3.6 apresentaremos o algoritmo scan circular seletivo.

3.1. Subconjuntos Seletivos

Dado um mapa com m regiões $r_1, r_2, \ldots, r_m$, defina L_i=LLR($\{r_i\}$) como sendo o logaritmo da razão de verossimilhança da zona contendo apenas a região r_i. Reorganize as m regiões do mapa de modo que as regiões $r_1, r_2, \ldots, r_m$ sejam ordenadas de tal forma que $L_1 \geq L_2 \geq L_3 \geq \ldots \geq L_m$. Defina o subconjunto $R_k = \{r_1, \ldots, r_k\}$, para $k = 1, \ldots, m$. Observe que $R_1=\{r_1\}$, $R_2=\{r_1, r_2\}$, ... e $R_m = \{r_1, r_2, \ldots, r_m\}$ e que $R_1 \subset \ldots \subset R_m$. Note também que R_m é o conjunto de todas as regiões do mapa original. Diremos que R_k é o conjunto seletivo de tamanho k, e contém as k regiões de maior verossimilhança no mapa.

Num certo sentido o conjunto R_k retém as k regiões mais importantes do mapa, filtrando a informação das k regiões em que o número de casos é mais significativo relativamente à sua população. Ao longo desse trabalho, ao fixarmos o conjunto R_k diremos que as k regiões pertencentes a R_k são regiões ativas e as demais são regiões inativas.

3.2. Ocupação Circular

Nosso objetivo nessa seção é definir uma medida quantitativa da irregularidade de formato geométrico de uma zona ou cluster, que chamaremos de ocupação circular.

Seja S uma zona formada pelas regiões $s_1, s_2, s_3, \ldots s_r$, de um subconjunto seletivo qualquer de um mapa original. Para cada s_i, $0 < i < r+1$, construa o círculo C_i como sendo o menor círculo com centro no centróide da região s_i que contenha todos os centróides de S. Seja D_i a zona circular formada por todas as regiões do mapa original cujos centróides pertencem ao circulo C_i. Seja P_i a

população da zona D_i e P a população da zona S. Defina OC(S) = max { $P/P_1, \ldots, P/P_r$ }. Chamamos OC(S) a ocupação circular da zona S. Em outras palavras OC(S) mede a presença das regiões de S no menor círculo que contém S, ponderada pelas populações de suas regiões componentes.

3.3. Conjunto de Pareto

Num conjunto C de n zonas $z_1, z_2, \ldots z_n$ considere os pares ordenados (L_i , OC_i), indicando a verossimilhança e a ocupação circular calculada para cada zona z_i. Esses pares são plotados em um plano cartesiano. O *operador de seleção* é agora definido em termos de dois objetivos: maximizar a ocupação circular e maximizar a verossimilhança. Este operador está baseado no conceito de *dominância*: um ponto é dito ser dominado por outro ponto se ele é pior que o outro ponto em pelo menos um objetivo, ao mesmo tempo em que ele não é melhor que aquele outro ponto em nenhum outro objetivo. O *conjunto de Pareto P* do conjunto *C* é o subconjunto dos pontos de *C* que não são dominados por nenhum ponto do complementar de *P*. Na Figura 15, o conjunto de Pareto indicado pelos círculos abertos consiste nos pontos que não são simultaneamente piores que nenhum outro ponto em ambos objetivos X e Y. Observe que o conjunto de Pareto é invariante por mudanças de escala e transformações monotônicas das funções objetivo.

3.4. A aproximação pela distribuição Gumbel

Através de testes numéricos extensivos, Abrams et al. (2005) mostraram que sob a hipótese nula, a distribuição empírica dos valores da estatística scan de Kulldorff para o algoritmo scan circular é aproximada pela distribuição Gumbel,

$$f(x) = \theta^{-1} \exp\{-\exp[(x-\mu)/\theta] - (x-\mu)/\theta\},$$

com parâmetros μ *(forma)* e θ *(escala)*.

Nesta seção estendemos os resultados encontrados por Abrams para a distribuição empírica da Estatística Scan, do Algoritmo Circular Multi-Objetivo, sob a hipótese nula. Testes numéricos sugerem que esta também seja razoavelmente bem aproximada pela distribuição Gumbel. Nós não tentaremos dar aqui uma prova rigorosa deste resultado. O raciocínio segue o mesmo argumento usado para a scan circular: a estatística scan do algoritmo circular multi-objetivo é também uma distribuição de valor extremo. Nós estamos interessados em calcular apenas valores-p . Mais tarde apresentamos os resultados das simulações mostrando a adequação da aproximação pela distribuição Gumbel.

3.5. Calculando a Significância do Cluster

Nesta seção nós mostramos como calcular a significância estatística dos pontos na solução do conjunto de Pareto global para o mapa de casos observados, quando comparados com centenas de conjuntos soluções de Pareto calculado para cada um dos mapas de casos simulados sobre a hipótese nula.

Uma simulação Monte Carlo consiste em executar o algoritmo scan circular multi-objetivo várias vezes para mapas de casos distribuídos aleatoriamente de acordo com a distribuição de Poisson sobre a hipótese nula, em que a média dos casos localizados em cada região é proporcional à população dessa região. Este processo encontra o conjunto de Pareto para cada alocação aleatória dos casos. Estes conjuntos de Paretos são juntados, obtendo uma coleção com milhares de pontos distribuídos no espaço *LLR x OC*, ou seja na faixa $(0,\infty) \times (0,1]$. Seja $D(l,c)$ a verdadeira distribuição bivariada de uma coleção arbitrariamente grande de pontos em $(0,\infty) \times (0,1]$.

Nosso objetivo agora é obter uma boa aproximação para $D(l,c)$. Extensivos testes numéricos sugerem que a distribuição marginal de $D(l,c)$ na variável l é aproximada pela

distribuição Gumbel, como segue. A faixa $(0,\infty) \times (0,1]$ é particionada em um número de faixas paralelas $(0,\infty) \times (s_j, s_{j+1})$, $s_j < s_{j+1}$. A média e a variância dos valores de LLR dos pontos contidos na faixa são usados para calcular os parâmetros para distribuição Gumbel ($\hat{\mu} = \overline{LLR}$ e $\hat{\theta} = \left(\text{var}(LLR)\right)^{\frac{1}{2}}$) a uma faixa particular. Seja G_j a distribuição Gumbel para a faixa $(0,\infty) \times (s_j, s_{j+1})$. Os valores $s_j < s_{j+1}$ são escolhidos de modo que a distribuição marginal de *D* não mude muito no intervalo , e que também existam pontos suficientes dentro da faixa para avaliar apropriadamente os parâmetros de G_j. Seja $P_0 = (l_0, c_0)$ um ponto pertencente ao conjunto de Pareto dos casos observados. A distribuição Gumbel G_j para a faixa $(0,\infty) \times (s_j, s_{j+1})$ contendo o ponto P_0 é usada para calcular o valor-p estimado para P_0 como $\int_{l_0}^{\infty} G_j(t)dt$.

3.6. Algoritmo do Scan circular multi-objetivo

O algoritmo consiste nos seguintes passos.

1- Usando a notação da seção 3.1, considere as regiões $r_1, r_2, \ldots, r_m$ do mapa ordenadas de modo que $L_1 \geq L_2 \geq L_3 \geq \ldots \geq L_m$, onde $L_k = LLR(\{r_k\})$ e $\{r_k\}$ é a zona que contém somente a região r_k.

2- Calcule a matriz de distâncias entre os centróides das m regiões do mapa original. Assim dist[i,k] = distância do centróide da região r_i ao centróide da região r_k, e indicedist[i,j] = índice da j-ésima região mais próxima de r_i. Dessa forma indicedist[i,1]=i, indicedist[i,2] = índice da região mais próxima de r_i, e indicedist[i,3] = índice da segunda região mais próxima de r_i, e assim por diante até atingir a região mais afastada de r_i no mapa.

3- Construa a matriz de população acumulada (PA), em que o elemento da posição i, j é dado por

$$PA[i, j] = \sum_{k=1}^{j} P[indicedist[i,k]]$$

que é a população dos j-ésimos vizinhos mais próximos da região r_i, (incluindo a própria região r_i).

4- Para cada valor de a, 0 < a < = 1, considere as 100a% regiões com maiores verossimilhanças, explicitamente as regiões $r_1, r_2, \ldots, r_A$, onde A é o menor inteiro maior ou igual a a.m.

4.1- Construa a submatriz da matriz de distância definida no passo 2, que contenha apenas os dados das regiões $r_1, r_2, \ldots, r_A$.

4.2 Construa PC = População Corrente = $\sum_{k=1}^{A} Pop[r_i]$.

4.3- Utilizando apenas as regiões $r_1, r_2, \ldots, r_A$ construa todas a zonas circulares possíveis, com a seguinte restrição : a população de cada zona não pode exceder $min\{$ PC, 0.25**poptotal*$\}$.

4.4- Para cada zona Z do passo 4.3 calcule LLR(Z) e OC(Z), obtendo o conjunto C_A dos pares ordenados (LLR(Z), OC(Z)).

4.5- Calcule P(C_A), o conjunto de pareto do conjunto C_A.

5.1- Construa $\bigcup P(C_A)$, união dos conjuntos P(C_A).

5.2 . Calcule P como sendo o pareto do conjunto $\bigcup P(C_A)$

6- Use simulação Monte Carlo para encontrar a significância de cada cluster.

CAPÍTULO 4: Aplicação

4.1. Detecção de Clusters de Homicídios em Minas Gerais.

Aplicamos o scan circular multi-objetivo para dados de homicídios no estado de Minas Gerais, acumulando os casos de 1998 a 2002. A figura 19 mostra os conjunto de Pareto para o mapa dos casos observados, para os valores de *a* indicados na figura.

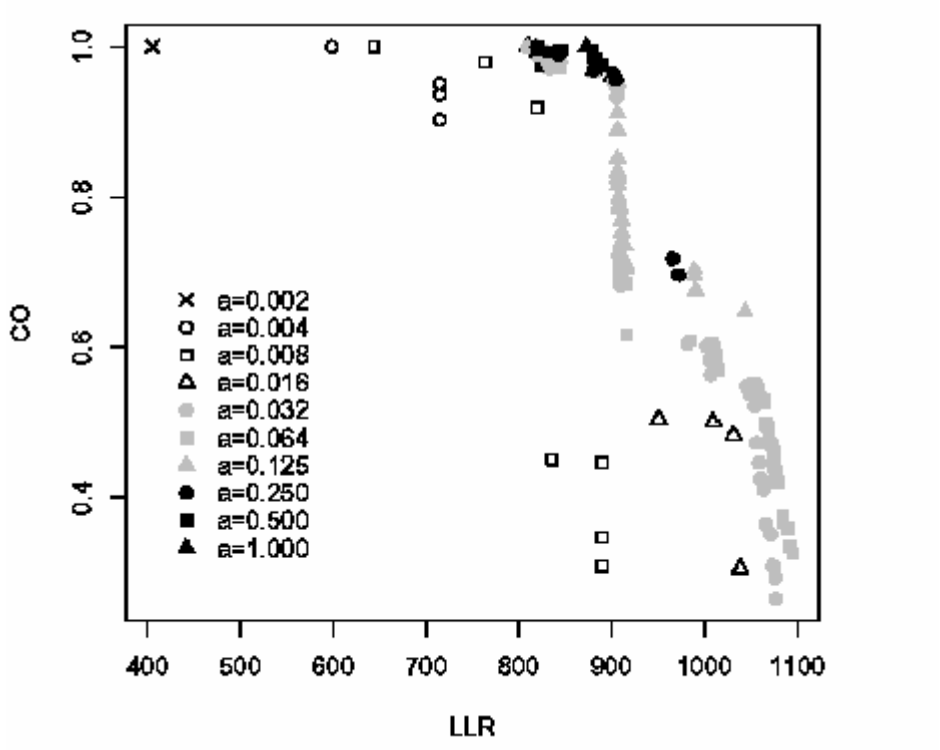

***Figura 19**: Todos os conjuntos de Pareto dos clusters de homicídios de Minas Gerais.*

A figura 20 mostra o conjunto dos Paretos dos Paretos ou Paretos Globais extraídos a partir do figura 19. Observe que o Pareto Global é formado por pontos de vários conjuntos seletivos diferentes, como pode ser visto comparando-se com a figura 19. O fato de calcularmos o conjunto

de Pareto por faixas é de extrema importância, pois caso contrário perderíamos informações provenientes de vários conjuntos seletivos.

Cada ponto do conjunto de Pareto global é uma zona candidata a um possível cluster, cujo formato podemos notar observando o valor da ocupação circular. À medida que esta cresce, o cluster tende a ficar conexo e com formato geométrico mais próximo o de um círculo (veja figura 22.)

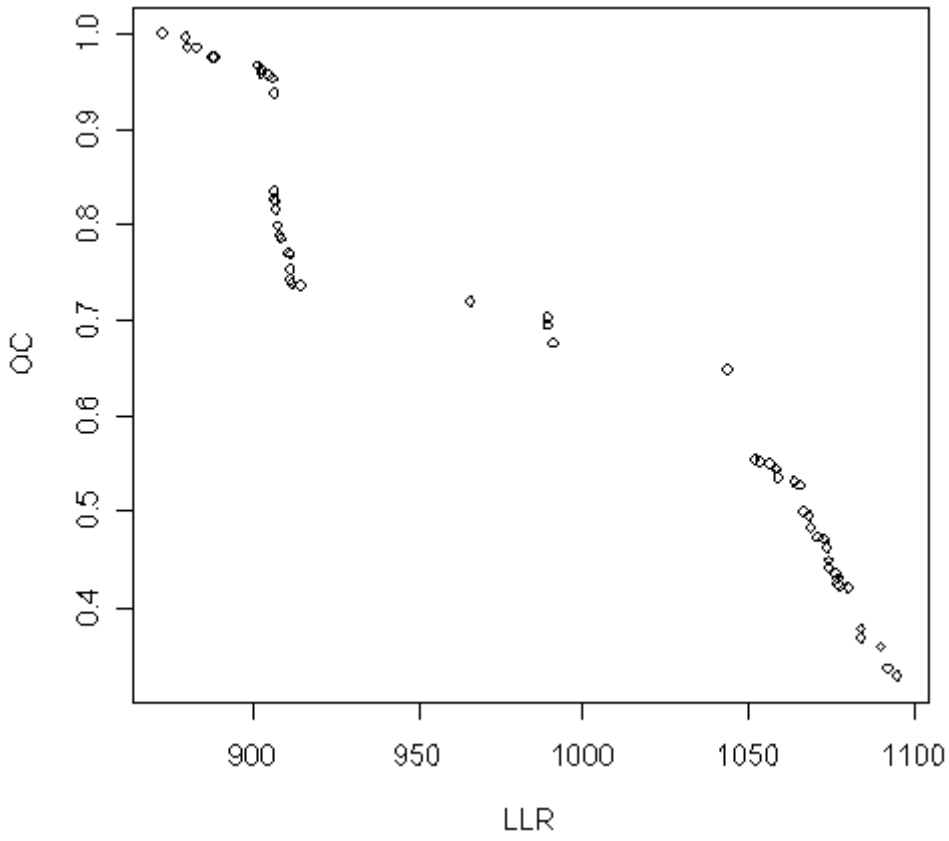

***Figura 20**: Conjunto de Pareto Global dos clusters de homicídios de Minas Gerais.*

Outro fato importante a ser notado é que saltos bruscos no gráfico dos pontos do conjunto de Pareto Global geralmente indicam mudanças significativas nas estruturas dos clusters. No nosso exemplo nota-se que o formato de um cluster passa de um cluster não conexo para conexo exatamente em um desses saltos. Os mapas mostrados na figura 22 apresentam diversos formatos de

clusters variando de cluster com alta verossimilhança e baixa ocupação circular, até clusters de baixa verossimilhança mas de alta ocupação circular.

O primeiro mapa corresponde ao ponto de Pareto com maior verossimilhança, mas ao mesmo tempo esse cluster possui um formato muito irregular, sendo totalmente desconexo. À medida que diminuímos a razão de verossimilhança LLR(Z) observamos que os clusters passam a ter um formato mais regular e a quantidade de regiões desconexas tende a diminuir. É importante notar que mapas de cluster com alta verossimilhança e baixa ocupação circular são bastantes parecidos com o mapa de taxa de risco. Existem algumas regiões que aparecem no mapa de risco e não aparecem nos mapas de cluster de alta verossimilhança e baixa ocupação circular, mas essas regiões têm pequenas populações e só aparecem se o risco for realmente muito alto. Caso contrário estas regiões são descartadas pela estatística scan.

Outro fator importante mostrado no mapas é que saltos de tamanhos maiores em relação às soluções no conjunto de Pareto produzem mudanças bruscas no formato do cluster, inclusive em termos de conexidade. À medida que a verossimilhança vai diminuindo e a ocupação circular vai aumentando o cluster tende a se tornar conexo. Em outras palavras podemos dizer que alta ocupação circular é quase um sinônimo de conexidade. O último mapa mostra o mapa de maior ocupação circular e menor razão de verossimilhança LLR (ocupação circular igual a um). Esse cluster seria o cluster encontrado pelo método scan circular. A figura 22 ilustra bem o que foi dito acima. A figura 21 mostra o mapa de risco relativo para homicídios em Minas Gerais nos anos de 1998 a 2002.

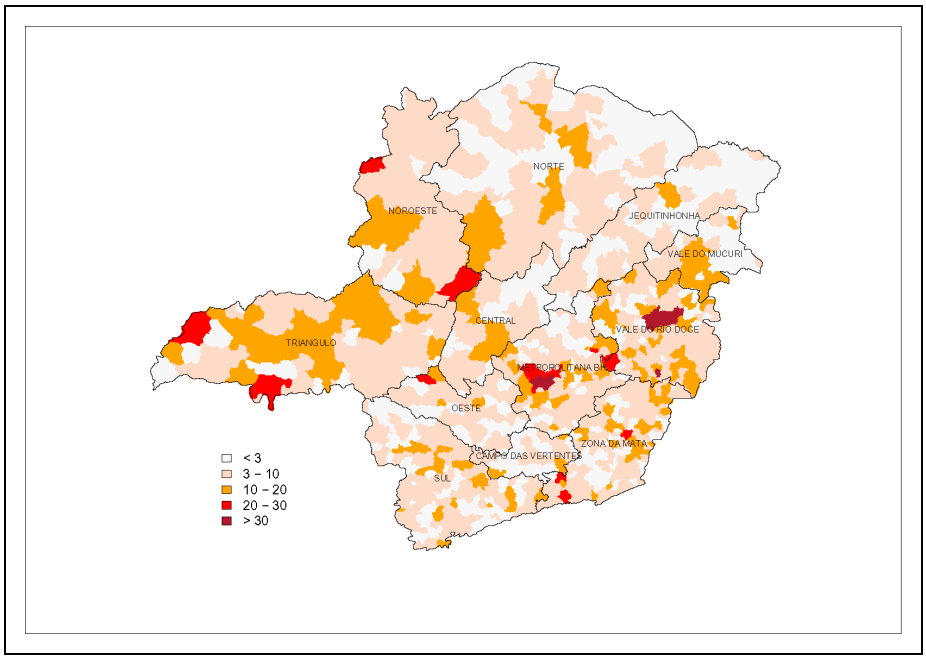

***Figura 21**: Taxa de mortes por homicídios em 100 mil habitantes em Minas Gerais.*

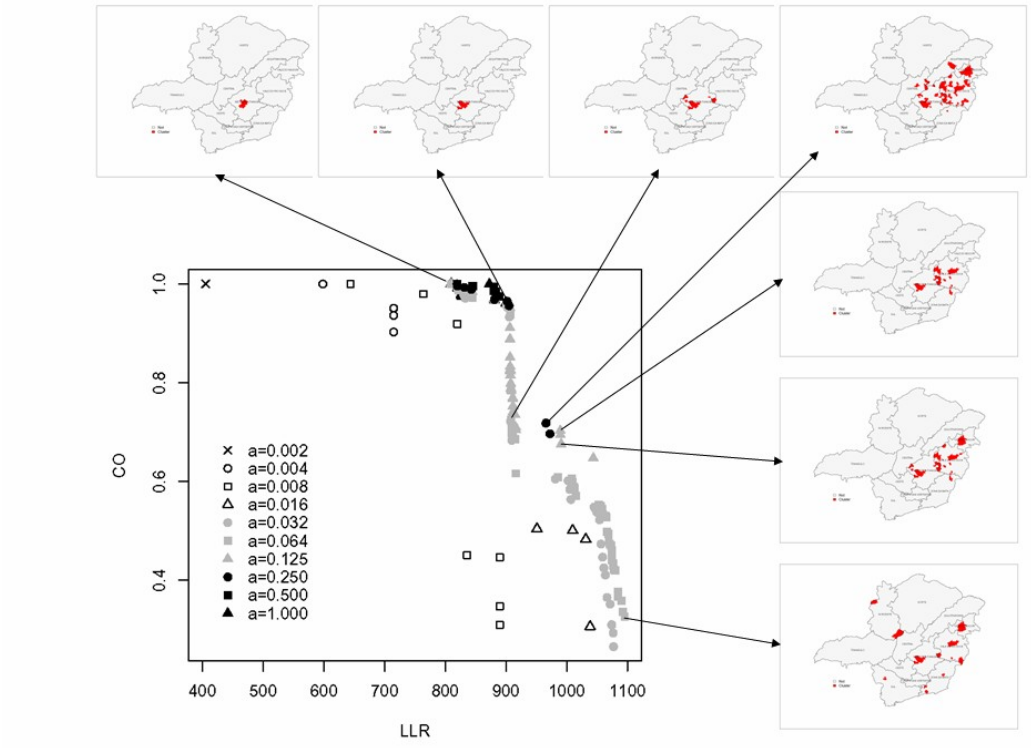

***Figura 22**: Cluster Representados por Pontos de Paretos.*

CAPÍTULO 5: Considerações finais

5.1. Conclusões

O algoritmo scan circular multi-objetivo mostrou-se muito eficiente na detecção de clusters tanto de formato regular quanto irregular. Além disso ele foi eficiente no controle das superestimações e a subestimações. Porém, o resultado mais relevante que aparece com o método scan circular multi-objetivo é o novo modo como olhamos para um cluster.

Agora o cluster encontrado não é mais uma zona conexa única e sim um conjunto de zonas que são pontos de um conjunto de Pareto. Essas zonas, possíveis candidatas a serem um cluster no mapa, variam de zonas com formatos totalmente irregulares e até mesmas desconexas, mas com alta razão de verossimilhança (LLR), até zonas circulares que geralmente apresentam valores de LLR menor. O que temos a partir de agora é uma estrutura multi-cluster. Esses resultados são devidos principalmente ao fato de usarmos conjuntos seletivos para escolher pontos para formar o conjunto de Pareto e também ao fato do algoritmo construir um conjunto de Pareto para cada conjunto seletivo e depois com esses pontos obtidos obter-se um novo conjunto de Pareto, chamado Pareto dos Paretos. Isso possibilita que tenhamos informações ricas sobre a maneira com que esses clusters são formados no mapa, ou seja, temos um ganho muito grande de informação de muitos possíveis tipos de cluster existentes.

Note que o objetivo agora não é mais encontrar somente um único cluster e sim um conjunto de clusters, ou seja, temos uma solução multi-cluster. Conhecemos o nível de significância de cada um desses clusters através do uso da simulação Monte Carlo. Em outras palavras conhecemos a ordem de importância de cada cluster.

Outro aspecto importante a ser notado no método scan circular multi-objetivo é a simplicidade dos conceitos usados. A maneira que fazemos a varredura do mapa, através de círculos, da mesma maneira como é feita no método scan circular. Assim não necessitamos de técnicas heurísticas estocásticas para seleção dos melhores conjuntos de regiões candidatas a serem um cluster.

O uso de subconjuntos seletivos permite a detecção de clusters de quaisquer formatos, pois os clusters de formatos irregulares aparecem naturalmente quando analisamos conjuntos seletivos com poucas regiões. Os clusters de formatos mais regulares, ou seja, com formatos mais próximos de um círculo, aparecem nos conjuntos seletivos com maior nmero de regiões, em geral em conjuntos seletivos com mais de 50% das regiões. Para medir a presença geométrica do cluster no mapa original não fazemos uso do conceito de compacidade apresentada no algoritmo simulated anneling e também usado no algoritmo genético. Trocamos esse conceito pelo conceito de ocupação circular, que é um conceito bem mais simples e que produz resultados igualmente satisfatórios.

A simplicidade do algoritmo scan circular multi-objetivo se traduz em rapidez na sua execução. O método scan circular multi-objetivo nos revela uma importante relação entre a posição dos pontos no conjunto de Pareto e a relação estrutural entre os clusters. Saltos significativos no conjunto de Pareto resultam em mudanças significativas no formato do cluster. Por exemplo, o cluster pode tornar-se conexo em um desses saltos.

5.2. Trabalhos Futuros

- Usar o método scan circular multi-objetivo para verificar a existência de cluster de mortes por malária na Amazônia Brasileira.

- Estudar o poder do método scan circular multi-objetivo.

- Comparar resultados encontrados pelo método scan circular multi-objetivo com resultados encontrados por outros métodos de detecção de clusters.

- Aplicação do scan circular multi-objetivo no monitoramento de clusters incipientes, especialmente para doenças contágiosas em que existem interações de curto e longo alcance.

Referências Bibliográficas

[1] ABRAMS A, KULLDORFF M, KLEINMAN K, 2005. Empirical/Assymptotic P-values for Monte Carlo Based Hypothesis Testing: an Application to Cluster Detection Using the Scan Statistic. *2005 Syndromic Surveillance Conference.*

[2]ASSUNÇAO R, COSTA M, TAVARES A, FERREIRA S, 2006. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine, 25; 1-21.*

[3] BESAG, J., NEWELL, J., 1991. The detection of clusters in rare diseases. *J. Roy. Statist. Soc. Ser., A*, 154, 143-155.

[4] DUCZMAL, L., ASSUNÇÃO, R., 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis, 45, 269-286.*

[5] DUCZMAL, L., CANÇADO, A.L.F., TAKAHASHI, R.H.C., BESSEGATO, L. 2006. A genetic algorithm approach to the detection and inference of irregularly shaped spatial disease clusters. *(submitted)*

[6] DUCZMAL, L., CANÇADO, A.L.F., TAKAHASHI, R.H.C., 2006. Delineation of Irregularly Shaped Disease Clusters through Multi-Objective optimization. Pre-print *(submitted)*

[7] DUCZMAL, L., KULLDORFF, M., HUANG, L., 2006, Evaluation of Scan Statistics for Irregularly Shaped Spatial Clusters. *J. Comput. Graph. Stat.* 2;15, 1-15;

[8] KULLDORFF, M., NAGARWALLA, N., 1995. Spatial disease clusters: detection and inference. *Stat. Med.*, 14, 779-810.

[9] KULLDORFF, M., 1997. A spatial scan statistic. *Communications in Statististics: Theory and Methods,* 26, pgs. 1481-1496.

[10] KULLDORFF M, TANGO T, PARK PJ., 2003. Power comparisons for disease clustering sets, *Comp. Stat. & Data Anal.*, 42, 665-684.

[11] KULLDORFF, M., HUANG L., PICKLE, L., DUCZMAL, L., 2006. Na elliptic spatial scan statistic. *Stat. Med*. (in press).

[12] LIMA, M., 2004. Dissertação de Mestrado: Avaliação do Poder do Teste da Estatística Scan para Múltiplos Clusters. *Departamento de Estatística da UFMG.*