UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Extensões da Estatística Scan Espacial utilizando Técnicas de Otimização Multiobjetivo

Ricardo Tavares

Tese de doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Doutor em Estatística.

Orientador: Prof. Luiz Henrique Duczmal

Belo Horizonte/MG, 07 de Julho de 2009.

Dedicatória

Dedico este trabalho à minha mãe, Maria das Graças Tavares, ao meu pai, Francisco Canindé Tavares, ao meu irmão, Ronaldo Tavares e à minha esposa, Helen Cristina Alkmim Tavares. Todos foram testemunhas oculares do meu esforço durante minha vida estudantil, na qual passei, com a ajuda de Deus e deles, por muitos obstáculos. E que jamais serão esquecidos por mim. Muito Obrigado Mamãe, Papai e Ronaldo, por todo o apoio que vocês me deram ao longo desses anos. Muito obrigado meu amor por você estar sempre do meu lado de forma carinhosa e compreensiva durante todo o Doutorado e por saber me entender como ninguém. Eu te amo e vou te amar por toda a minha vida ...

Amo vocês! Muito !!!

Epígrafe

“A educação faz um povo fácil de ser liderado, mas difícil de ser dirigido; fácil de ser governado, mas impossível de ser escravizado. ”

Henry Peter

“Não se mede o valor de um homem pelas suas roupas ou pelos bens que possui. O verdadeiro valor de um homem é o seu caráter, suas idéias e a nobreza dos seus ideais.”

Charles Chaplin

“Preocupe-se mais com a sua consciência do que com sua reputação. Porque sua consciência é o que você é e a sua reputação é o que os outros pensam de você. E o que os outros pensam, é problema deles.”

Bob Marley

Agradecimentos

Ao Grande Deus, por ter me protegido dos males e por me trazer saúde e força de vontade para adquirir novos conhecimentos, buscando a honestidade e a qualidade profissional.

À minha querida e amável esposa, por todo o amor, carinho e compreensão que tem me proporcionado durante esta nova fase da minha vida. Eu te amo meu amor e não me esquecerei do que te prometi em 07/07/07.

A todos da minha família que de uma forma ou de outra me apoiaram durante esta etapa, especialmente aos meus pais pelos conselhos e educação que me passaram, fazendo o possível e o impossível para que eu chegasse aonde cheguei, honrando nossas origens e preservando a nossa humildade, dignidade e honestidade.

Ao Sr. Raimundo (sogro) e à Sra. Arlete (sogra) que sempre me deram carinho e abrigo principalmente quando eu estava em situação financeira díficil.

Ao meu amigo e orientador Prof. Luiz Duczmal por toda e imensa dedicação para com este trabalho. Ele jamais reclamou dos meus outros compromissos e sempre buscou me orientar da melhor forma possível ao longo de toda a pesquisa. Caro Professor Luiz, o Sr. foi um verdadeiro pai para mim, pois quando não aceitei a bolsa do Doutorado tive a sua confiança e apoio humano incondicional. Muito obrigado de coração!!!

Às Professoras Sueli, Mercedes e Arminda por todos os ensinamentos e atenção para comigo.

Aos Professores Frederico, Gregório, Sabino, Roberto Quinino, Michel e Antônio Eduardo, por todas as ajudas, consideração e confiança que tiveram por mim.

Aos Professores Adrian e Enrico por terem dado um voto de confiança para com o meu esforço.

Aos demais Professores que formam este Departamento de destaque no meio estatístico e acadêmico.

Aos Professores Ricardo Takahashi e Eduardo Carrano pelas valiosas contribuições para a melhoria deste trabalho.

Aos Professores Hélio dos Santos Migon e Cibele Queiroz que, mesmo distantes, aceitaram participar deste momento tão ímpar na minha vida.

Ao Professor Valdério Reisen que contribuiu muito durante o meu mestrado e foi um dos responsáveis pelo meu ingresso neste doutorado.

Aos meus amigos alunos e ex-alunos do Departamento de Estatística da UFMG, em especial àqueles que ingressaram nas turmas deste doutorado em 2005/2 (Joab, Lupércio e José Rodrigues) e em 2006/1 (Fábio Demarqui, Maristela e Magda). As ajudas do Fábio Demarqui foram fundamentais. Fica aqui a minha gratidão a vocês por tudo.

Aos Amigos Anderson, André, Max, Cristiano, Elias, Erik, Carlito, Ivan, Alexandre Loureiros, Mário e Fábio Colombiano pelas valiosas ajudas ao longo dessa jornada.

Aos amigos que compõem o corpo técnico do Departamento de Estatística, principalmente a Rogéria, o José Carlos, a Cristina, a Rose e a Marcinha. Eles são muito bacanas.

Ao Departamento de Matemática da Universidade Federal de Ouro Preto (UFOP) e aos meus colegas docentes que sempre me apoiaram durante esta etapa de esforços, em especial, àqueles da área de Estatística Profa. Maria Cláudia, Prof. Álvaro, Profa. Thaís e Prof. Flávio (que me incentivou a trabalhar nesta área e que me apoiou bastante).

A todos que compõem o Departamento de Estatística da Universidade Federal do Rio Grande do Norte (UFRN) e que foram fundamentais para eu chegar até aqui, em especial aos Professores Formiga, Franciné, Damião e Medeiros.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelos investimentos depositados nos projetos de pesquisa envolvendo Estatística e Probabilidade.

Enfim, a todos àqueles que de uma forma ou de outra tentaram me ajudar a chegar ao término desse trabalho.

Resumo

Este trabalho apresenta três novas extensões da estatística scan espacial de Kulldorff para a detecção e inferência de clusters espaciais em estudos cuja localização está associada a áreas delimitadas por polígonos. Considere um mapa com m regiões em que se conhece os casos observados de um certo evento de interesse (por exemplo, infecção ou óbito por alguma doença) e a população de cada área. Podemos perguntar: os casos estão distribuídos de forma aleatória nessas áreas? Ou ainda, se existe uma área do mapa que possui uma quantidade discrepante de casos em relação às demais? O nosso interesse é testar as seguintes hipóteses: H_0: Não existe cluster no mapa vs H_1: Existe um cluster no mapa.

Na primeira parte, nós propomos uma nova ferramenta para testar hipóteses sobre a adequação de fatores ambientalmente definidos para a formação de clusters localizados de doenças, através da avaliação comparativa da significância dos clusters mais prováveis detectados sob mapas cujas estruturas de vizinhança foram modificadas de acordo com estes fatores. Um algoritmo genético multiobjetivo para a estatística scan é utilizado para encontrar clusters em um mapa dividido em um número finito de regiões, cuja adjacência é definida por uma estrutura de grafo. Este detector de clusters maximiza dois objetivos, a estatística espacial scan e a regularidade do formato do cluster. Ao invés de especificar localizações para o possível cluster a priori, como acontece para algoritmos baseados nos algoritmos focados, alteramos a adjacência básica induzida por limites geográficos comuns entre as regiões. Na nossa abordagem, a conectividade entre as regiões é reforçada ou enfraquecida, conforme certas características ambientais de interesse associadas com o mapa. Construímos vários cenários plausíveis, cada um modificando a estrutura de adjacência em áreas específicas do mapa, e executamos o algoritmo genético multiobjetivo para selecionar as melhores soluções

de clusters para cada um desses cenários. As significâncias estatísticas dos clusters mais prováveis são estimadas através de simulações Monte Carlo. Os clusters com os menores valores p estimados, junto com os seus respectivos mapas de características ambientais alterados são apresentados para a análise comparativa. Conseqüentemente a probabilidade de detecção do cluster é aumentada ou diminuída, de acordo com as mudanças feitas na estrutura de adjacência do grafo, relacionada à seleção das características ambientais. A eventual identificação de características ambientais específicas às quais induzem os clusters mais significativos permitem ao pesquisador aceitar ou rejeitar diferentes hipóteses a respeito da relevância dos fatores geográficos. Estudos de simulação numérica e uma aplicação para clusters de malária no Brasil são apresentados.

Na segunda parte, desenvolvemos uma nova metodologia para analisar clusterização em mapas de regiões. Situações em que um cluster da doença não tem um formato regular são razoavelmente comuns. Além disso, mapas com clusters múltiplos, quando não existe um cluster primário claramente dominante, também ocorrem frequentemente. Nós desenvolvemos um método para analisar mais cuidadosamente os vários níveis de clusterização que surgem naturalmente em um mapa de doenças. A estatística espacial scan é a medida usual de força de um cluster. Uma outra medida importante é a sua regularidade geométrica. Um algoritmo genético multiobjetivo foi desenvolvido para identificar clusters de formato irregular. Uma busca é executada tentando maximizar dois objetivos competitivos: a estatística scan e a regularidade do formato (que usa o conceito de compacidade). A solução apresentada é um conjunto Pareto, consistindo de todos os clusters encontrados os quais não são simultaneamente piores em ambos objetivos. Uma avaliação da significância é conduzida em paralelo para todos os clusters no conjunto Pareto através de simulações Monte Carlo, e então o cluster mais provável é encontrado. Ao invés de usar um algoritmo genético, nossa nova metodologia incorpora a simplicidade da estatística scan circular, podendo detectar e avaliar clusters de formato irregular. Nós definimos a ocupação circular (CO) de um candidato a cluster como sua população dividida pela população do menor círculo que a contém. O conceito de CO, computacionalmente mais rápido, substitui aqui a definição de compacidade como a medida

de regularidade do formato. A estatística scan é avaliada para cada uma das *m* regiões do mapa tomadas individualmente. As regiões são ordenadas de forma decrescente conforme os valores da estatística scan. Um procedimento Monte Carlo é usado para a avaliação da significância. A presença de "joelhos" nos conjuntos Pareto indica transições repentinas na estrutura dos clusters, correspondendo aos rearranjos devido à coalescência de clusters fracamente ligados (geralmente desconectados). Cada conjunto Pareto contendo os clusters mais prováveis dentro de um determinado nível de informação geográfica, podem ser aglutinados para fornecer uma descrição global mais completa. O método scan circular multiobjetivo é um procedimento eficiente que permite a visualização da estrutura de clusterização de um mapa. A comparação dos conjuntos Pareto para os casos observados, com aqueles calculados sob a hipótese nula fornece valiosas pistas para a distribuição espacial da doença. O procedimento proposto pode ser fundamental para monitorar clusters incipientes e em diversas escalas geográficas simultaneamente, o que o torna uma ferramenta promissora em vigilância sindrômica, especialmente para doenças contagiosas, em que existem interações espaciais de curto e longo alcance.

Na terceira parte, exploramos o novo conceito de "estatística espacial scan desagregada". Esta parte da tese ainda está em desenvolvimento, e assim apresentaremos apenas um trabalho introdutório com alguns exemplos. Apresentamos uma variante multiobjetivo da estatística espacial scan de Kulldorff, definindo a busca para o cluster mais provável como um problema multiobjetivo. Duas funções foram consideradas para maximização do conjunto multiobjetivo: o número de casos e o risco relativo dentro da zona candidata a cluster. Mostramos através de exemplos que esta nova abordagem apresenta algumas características atrativas: ela é capaz de distinguir "famílias" distintas de clusters de significância geográfica dentro do conjunto das soluções potenciais, agrupadas pelas suas posições relativas no espaço de *casos* versus *risco relativo*. Assim, a estrutura de clusterização é facilmente disponível para o pesquisador, e inferências podem ser desenvolvidas através desta nova ferramenta.

Palavras-chave: Estatística Espacial Scan, Otimização Multiobjetivo, Algoritmo Genético,

Compacidade, Conjunto Pareto, Scan Reforçada, Conjuntos Seletivos, Ocupação Circular, Scan Seletiva, Curvas de Níveis, Scan Desagregrada.

Abstract

This work presents three new extensions of Kulldorff's Spatial Scan Statistic for the detection and inference of spatial clusters. Consider a map divided into m regions with known populations at risk and number of cases of some disease. We would like to know if the cases are randomly distributed over the m regions or not; if the cases are not randomly distributed, is it possible to locate a specific area within the map with an abnormal concentration of cases? We are interested in testing the alternative hypothesis (there is a cluster in the map) against the null hypothesis (there are no clusters in the map).

In the first part, we propose a novel tool for testing hypotheses concerning the adequacy of environmentally defined factors for local clustering of diseases, through the comparative evaluation of the significance of the most likely clusters detected under maps whose neighborhood structures were modified according to those factors. A multi-objective genetic algorithm scan statistic is employed for finding spatial clusters in a map divided in a finite number of regions, whose adjacency is defined by a graph structure. This cluster finder maximizes two objectives, the spatial scan statistic and the regularity of cluster shape. Instead of specifying locations for the possible clusters a priori, as is currently done for cluster finders based on focused algorithms, we alter the usual adjacency induced by the common geographical boundary between regions. In our approach, the connectivity between regions is reinforced or weakened, according to certain environmental features of interest associated with the map. We build various plausible scenarios, each time modifying the adjacency structure on specific geographic areas in the map, and run the multi-objective genetic algorithm for selecting the best cluster solutions for each one of the selected scenarios. The statistical significances of the most likely clusters are estimated through Monte Carlo simulations. The clusters with the

lowest estimated p-values, along with their corresponding maps of enhanced environmental features, are displayed for comparative analysis. Therefore the probability of cluster detection is increased or decreased, according to changes made in the adjacency graph structure, related to the selection of environmental features. The eventual identification of the specific environmental conditions which induce the most significant clusters enables the practitioner to accept or reject different hypotheses concerning the relevance of geographical factors. Numerical simulation studies and an application for malaria clusters in Brazil are presented.

In the second part, we develop a new methodology for analyzing clustering in maps of regions. 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. A search is executed aiming to maximize two objectives, namely the scan statistic and the regularity of shape (using the compactness concept). The solution presented is a Pareto-set, consisting of all the clusters found which are not simultaneously worse in both objectives. A significance evaluation is conducted in parallel for all clusters in the Pareto-set through Monte Carlo simulation, then finding the most likely cluster. Instead of using a genetic algorithm, our novel method incorporates 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, computationally faster and relying on familiar concepts, substitutes here the compactness definition as the 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. 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. As each Pareto-set contains the most likely clusters within a certain level of geographical information, they could be joined to provide an overall complete description. The multi-objective circular scan is a fast method that 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 spatial-temporal clusters at several geographic scales simultaneously is a promising tool in syndromic surveillance, especially for contagious diseases when there is a mix of short and long range spatial interactions.

In the third part, we explore the novel concept of "disaggregated spatial scan statistic". This part the thesis is still under development, so we will present only introductory work and a few examples. We present a multi-objective variant of Kulldorff's Spatial Scan Statistic, defining the search for the most likely cluster as a multiobjective problem. Two functions were considered for maximization in the multi-objective setting: the rate and the number of cases within the cluster. We show through examples that this novel approach presents some attractive features: it is capable of distinguishing "families" of clusters of geographical significance within the set of potential solutions, grouped by their relative position in the rates versus cases space. Thus the clustering structure is readily available to the practitioner, and more detailed inference could be derived through this new tool.

Key words: Spatial Scan Statistic, multi-objective optimization, genetic algorithm, compactness, Pareto-set, reinforced scan, levels of clustering, circular occupation, seletive scan, levels curves, disaggregated scan.

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Capítulo 1

Revisão Bibliográfica

O tratamento da informação geográfica vem influenciando de maneira crescente várias áreas do conhecimento. É importante questionar se os dados referentes aos fenômenos pesquisados estão distribuídos espacialmente de forma aleatória ou se estão aglomerados em apenas alguma(s) área(s) específica(s). E isto pode ser fundamental, por exemplo, para os orgãos competentes que podem estar interessados em conhecer quais são os municípios onde estão as maiores incidências de uma patologia (malária, crimes, dengue, etc). Assim, poderão direcionar recursos de modo eficiente para prevenção e tratamento dos problemas de saúde de cada região.

Neste trabalho adotaremos o uso do termo inglês *cluster* para representar "conglomerado", pois o seu uso já está bastante difundido. Um cluster espacial é uma parte de um mapa em que a ocorrência de casos de um fenômeno de interesse é discrepante do restante do mapa, isto é, alta demais ou baixa demais. Esse fenômeno é, muitas vezes, a infecção por alguma doença ou a ocorrência de algum crime.

Algoritmos para a detecção e avaliação da significância estatística de clusters espaciais são importantes ferramentas geográficas em Epidemiologia, vigilância sindrômica, monitoramento de crimes e ciências ambientais. A elucidação da etiologia das doenças, a disponibilidade de alarmes confiáveis para detectar surtos intencionais ou não de uma certa doença, o estudo de padrões espaciais de atividades criminais e monitoramento geográfico de mudanças ambientais são tópicos recentes de intensa pesquisa. Métodos para encontrar clusters espa-

ciais foram revisados em Elliott et al. (1995), Waller and Jacquez (2000), Kulldorff (1999), Lawson et al. (2000), Moore and Carpenter (1999), Balakrishnan and Koutras (2002), Glaz et al. (2001), Lawson (2001), Buckeridge et al. (2005) e Duczmal et al. (2009). Em muitas situações necessitamos reconhecer clusters espaciais em uma classe geométrica mais geral. Várias propostas para encontrar clusters espaciais de formato arbitrário são revisadas e descritas a seguir.

Atualmente, o método mais popular para encontrar clusters espaciais é a estatística espacial scan de Kulldorff (1997, 1999), que é uma descendente da estatística espacial scan de Naus (1965). A significância do cluster mais provável é estimada através de simulações de Monte Carlo (Dwass, 1957). Este método pode ser usado para dados pontuais com localizações geográficas exatas ou para dados agregrados, onde uma região de estudo é particionada em células. A Estatística Scan Circular (Kulldorff and Nagarwalla, 1995) varre todos os possíveis conjuntos de regiões conectadas cujos centros estejam dentro de um círculo com raio variando conforme o percentual de população dentro deste círculo. Um estudo da avaliação do poder da Estatística Scan foi realizado por Lima (2004) e Kulldorff et al. (2003). Muitas propostas têm sido sugeridas para encontrar clusters espaciais de formato arbitrário, como por exemplo Duczmal et al. (2008), mas a maioria delas sendo uma extensão da estatística scan de Kulldorff. Duczmal et al. (2009) apresenta um levantamento bibliográfico recente da estatística Scan, de suas extensões e aplicações.

Ao procurar clusters sem limitar a liberdade de seu formato geométrico, tem-se uma redução no poder de detecção. Isto acontece porque a coleção de todas as zonas conectadas, independentemente do formato, é muito numerosa. O valor máximo da função objetivo está provavelmente associada com clusters em formatos de “árvores”, que simplesmente ligam as células com maior razão de verosimilhança do mapa, sem contribuir com a descoberta de soluções geograficamente significativas que delineiam corretamente o verdadeiro cluster. Ou seja, há muito ruído, o que dificulta a distinção das legítimas soluções. Este problema ocorre na maioria dos detectores de cluster com formato irregular. Em seguida revisaremos alguns trabalhos que buscam levar em consideração formatos irregulares.

A cota do nível superior (ULS) da estatística scan (Patil and Taillie, 2004) controla a liberdade do formato, pois explora uma coleção pequena de grafos conectados para formarem as zonas candidatas *z*, avaliadas conforme sua taxa (número de casos divididos pela população em risco) na área de estudo com n regiões. A árvore ULS é construída pelas zonas selecionadas com maiores taxas e formadas por apenas uma região individual, que são máximos locais para a taxa. As regiões vizinhas da área de estudo são sucessivamente conectadas às regiões individuais representadas pelas folhas, formando zonas maiores com taxas mais baixas que são identificadas como os nós internos mais baixos da árvore ULS. Eventualmente, aquelas zonas agregadas formando conjuntos maiores possuem taxas menores e são representadas como nós internos próximos à raiz. A própria raiz representa a área completa do estudo. A coleção de zonas representadas pelos nós da árvore ULS constitui o espaço de parâmetros ULS reduzido e sua cardinalidade será no máximo *n*. A árvore ULS precisa ser calculada novamente para cada nova réplica de Monte Carlo. Este método é rápido, mas possivelmente pode ignorar muitos clusters interessantes, devido à pequena cardinalidade da árvore ULS. Esta questão é abordada em Patil et al. (2006), em que uma extensão do conjunto ULS original é construída. Modarres and Patil (2007) discutiram uma extensão da estatística scan ULS para dados bivariados para os modelos Binomial e Poisson e estudaram a sensibilidade dos pontos quentes para o grau de associação entre as variáveis.

Duczmal and Assuncao (2004) propuseram um algoritmo Simulated Annealing (SA) para a detecção de agrupamentos espaciais de formato geométrico arbitrário em um mapa de ocorrências e população georreferenciadas. A coleção de zonas de formato irregular conectadas consiste de todas aquelas zonas para as quais os subgrafos correspondentes são conectados. Como esta coleção é muito grande, torna-se impraticável calcular a razão de verossimilhança da estatística scan para todas elas. O algoritmo SA tenta visitar apenas as zonas mais promissoras. Duas zonas são vizinhas quando elas diferem por uma única região. Para cada região individual da área de estudo, a estatística scan circular é usada para definir um cluster inicial z_0. O algoritmo escolhe algum vizinho z_1 sobre todos os vizinhos de z_0. No próximo passo, outro vizinho z_2 é escolhido sobre os vizinhos de z_1, e assim por diante, até um limiar pré-

predefinido do número de regiões ser atingido. Dessa forma, em cada passo uma nova zona é construída, adicionando-se ou excluindo-se uma célula da zona do passo anterior. Ao invés de sempre se comportar como um algoritmo guloso (o algoritmo escolhe sempre o vizinho de maior razão de verossimilhança (RV) em cada passo), então ele avalia se existiu pouca ou nenhuma melhoria da RV durante os últimos passos. Neste caso, o algoritmo SA opta por escolher um vizinho de forma aleatória. Isto é feito para tentar evitar que o algoritmo fique preso em máximos locais da RV. A busca é reiniciada muitas vezes, cada vez usando células individuais do mapa como zonas iniciais. Assim, o efeito desta estratégia é manter o programa explorando livremente as zonas mais promissoras do espaço de configuração e abandonando as direções que não são interessantes. A melhor solução encontrada pelo programa, a que maximiza a RV, é o cluster mais provável.

Tango and Takahashi (2005) apresentaram uma estatística espacial scan com formato flexível (FS) que faz uma busca exaustiva de todos os possíveis clusters conectados em primeira ordem de vizinhança dentro de um conjunto cercado pelos k vizinhos mais próximos de uma dada região. Para cada região i, a estatística FS considera os k círculos concêntricos mais todos os conjuntos de regiões conectadas, cujos centróides estão localizados dentro do k-ésimo maior círculo concêntrico. O método é repetido para cada região do mapa, permitindo que todos os clusters conectados sejam enumerados até um limite de tamanho k. O conjunto de clusters potenciais é armazenado e avaliado sobre a hipótese nula sem a necessidade de reconstruí-lo a cada tempo. Takahashi et al. (2007) estenderam a estatística FS para detectar clusters com formato irregular levando em consideração o espaço e o tempo.

Kulldorff et al. (2006) apresentaram uma versão elíptica da estatística espacial scan, generalizando o formato circular da janela da varredura. Eles usaram uma janela de varredura elíptica de locação, formato (excentricidade), ângulo e tamanho variáveis. A scan elíptica teve um maior poder para detectar clusters mais alongados que a estatística scan circular.

Conley and MacGill (2005) propuseram um algoritmo genético para explorar um espaço de configuração de múltiplas aglomerações de elípses para conjuntos de dados pontuais. O método empregava uma estratégia para limpar a melhor configuração encontrada e em se-

guida simplificar geometricamente o cluster. Sahajpal et al. (2004) também usaram um algoritmo genético para encontrar clusters baseados nas interseções de círculos de diferentes tamanhos e centros em conjuntos de dados pontuais.

Duczmal et al. (2007) desenvolveram um algoritmo genético scan para a detecção e inferência de clusters espaciais de formato irregular. Assumindo um mapa dividido em regiões com uma dada população em risco e casos de uma determinada doença, as operações com grafos são minizadas por se basearem numa rápida geração de filhos e avaliarem a estatística espacial scan de Kulldorff. A função de penalidade de Duczmal et al. (2006), baseada no conceito de compacidade geométrica, é empregada para evitar irregularidade excessiva do formato geométrico do cluster. Este algoritmo estocástico possui variância das soluções menor que o SA e é mais flexível que o scan elíptico. O poder deste método é igual ao do SA para clusters moderadamente irregulares e é superior para clusters muito irregulares.

Duczmal et al. (2008) propuseram uma abordagem para delineação geográfica de clusters espaciais de doenças com formatos irregulares tratando como um problema de otimização multiobjetivo. Um critério quantitativo para escolher a melhor solução para cluster foi apresentado por maximizar simultaneamente dois objetivos competitivos: a regularidade da forma e a estatística espacial scan (LLR).

Através de extensivos testes numéricos, Abrams et al. (2006) mostraram que, sob a hipótese nula de não existir cluster no mapa, a distribuição empírica dos valores da estatística espacial scan de Kulldorff para clusters circulares é aproximada pela distribuição Gumbel. Os autores mostraram que usando esta abordagem semi-paramétrica apenas 100 réplicas Monte Carlo já são suficientes para fornecer a mesma adequação na estimação da significância encontrada por utilizar 10000 réplicas da distribuição empírica usual.

Kulldorff et al. (2003) apresentaram uma coleção grande de conjuntos de dados *benchmark* gerados sob diferentes modelos de clusters e de hipóteses nulas, para serem usados para avaliação do poder. Estes conjuntos de dados são usados para comparar o poder da estatística espacial scan.

Este trabalho tem como objetivo propor algumas extensões da estatística espacial scan

de Kulldorff utilizando técnicas de otimização multiobjetivo para levar em consideração o formato arbitrário das zonas candidatas a clusters e também os vários níveis de clusterização presentes em um mapa. A pesquisa apresenta três estatísticas para detecção de clusters espaciais: estatística scan de adjacência modificada, estatística scan multiseletiva e estatística scan desagregada. Os métodos de cada extensão proposta foram implementados em Dev C++ (Laplace et al., 2007), e teve o auxílio do R (R Development Core Team, 2007) para confeccionar os mapas e gráficos.

A estrutura desta tese envolve outros seis capítulos. O capítulo 2 descreve a Estatística Scan proposta por Kulldorff and Nagarwalla (1995). O algoritmo genético e as técnicas multiobjetivo são introduzidas no capítulo 3. As extensões da Estatística Scan propostas neste trabalho são apresentadas nos capítulos 4, 5 e 6, enquanto as considerações finais são descritas no capítulo 7.

Capítulo 2

Estatística Scan Espacial

A estatística de teste para o método de detecção de conglomerados espaciais proposto por Kulldorff é uma descendente da estatística scan espacial de Naus (1965).

2.1 Estatística Scan

Considere um mapa dividido em m regiões, com população total N e casos totais C de algum fenômeno observável. A análise é feita condicionalmente no número total de casos e assim C é tratado como uma constante conhecida. Defina zona como qualquer conjunto de regiões conectadas e denote-a por z. Seja Z o conjunto das áreas z candidatas a formarem um cluster. Estes candidatos z são os círculos de raio r arbitrário centrados em cada um dos m centróides das regiões do mapa. O teste proposto por Kulldorff (1997) baseia-se no método de máxima verossimilhança. O parâmetro neste caso é $(z; p; q)$ em que z denota o círculo em Z, p é a probabilidade de que um indivíduo qualquer dentro de z seja um caso enquanto que essa probabilidade para um indivíduo fora de z é q. Tais probabilidades são independentes para todos os indivíduos. Supondo que não existe cluster no mapa (hipótese nula), o número de casos em cada região segue uma distribuição Poisson, com média proporcional à razão entre C e N, multiplicada pela população das respectivas zonas. Defina $L(z)$ como a verossimilhança sob a hipótese alternativa de que existe um cluster na zona z ($H_A : p > q$), e L_0 a verossimilhança sob a hipótese nula ($H_0 : p = q$). Sejam $\mu(z) = \frac{n(z)}{N}C$ o número esperado de casos dentro da zona z, sob a hipótese nula, $n(z)$ a população dentro de z, $c(z)$ o número

de casos dentro de z. Para o modelo em que os dados possuem uma distribuição de Poisson, Kulldorff (1997) mostrou que a função de verossimilhança é dada por

$$L(z,p,q) = \frac{e^{-pn(z)-q(N-n(z))}}{C!} p^{c(z)} q^{C-c(z)} \prod_{j=1}^{m} n(j) \tag{2.1}$$

A razão de verossimilhança, λ, pode ser escrita como

$$\lambda = \frac{Sup_{H_A}\{L(z)\}}{Sup_{H_0}\{L(z)\}} = \frac{Sup_{z\in Z, p>q}\{L(z,p,q)\}}{Sup_{p=q}\{L(z,p,q)\}} = \frac{L(\hat{Z})}{L_0} \tag{2.2}$$

Por definição, $L_0 = \frac{e^{-C}}{C!}\left(\frac{C}{N}\right)^C \prod_{j=1}^{m} n(j)$.

A estatística de teste λ do teste da razão de verossimilhança é expressa por

$$\lambda = \begin{cases} Sup_{z\in Z} \dfrac{\left(\frac{c(z)}{n(z)}\right)^{c(z)} \left(\frac{C-c(z)}{N-n(z)}\right)^{C-c(z)}}{\left(\frac{C}{N}\right)^C}, \text{ se } \dfrac{c(z)}{n(z)} > \dfrac{C-c(z)}{N-n(z)}; \\ 1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{, caso contrário.} \end{cases}$$

O cluster mais verossímil é a zona $\hat{z}$ para a qual $L(z)$ é maximizada $L(\hat{z}) \geq L(z)\ \forall z \in Z$. Kulldorff and Nagarwalla (1995) obtiveram a distribuição exata de $(\lambda \mid C)$ via um procedimento de simulação Monte Carlo, pois além da distribuição de λ depender da distribuição da população o que a torna muito difícil de ser obtida analiticamente, a aproximação assintótica usual por uma distribuição Qui-quadrado da transformação $-2log\lambda$ não é válida pois as condições de regularidade não são satisfeitas.

Uma forma mais simplificada para a razão de verossimilhança acima é denotar por $I(z) = \frac{c(z)}{\mu(z)}$ o risco relativo dentro de z, $O(z) = \frac{C-c(z)}{C-\mu(z)}$ o risco relativo fora de z.

$$LR(z) = \frac{L(z)}{L_0} = \begin{cases} I(z)^{c(z)} O(z)^{C-c(z)}, \text{ se } I(z) > 1; \\ 1 \qquad\qquad\qquad\qquad \text{, caso contrário.} \end{cases}$$

O cluster mais plausível é a zona $\hat{z}$ para a qual $LR(z)$ é maximizada $LR(\hat{z}) \geq LR(z)\ \forall z \in Z$. Como a função logaritmo é uma função estritamente crescente e $LR(z)$ cresce muito

rapidamente, maximizar $log\{LR(z)\}$ é equivalente a maximizar $LR(z)$. Usualmente emprega-se a expressão $LLR(z) = log\{LR(z)\}$.

Uma janela circular sobre a área em estudo define uma zona z constituída pelas regiões cujos centróides caem dentro da janela. A zona Z definida é o conjunto de todas as zonas obtidas por janelas centradas em cada centróide e de raios variando entre zero e algum número maior que zero, conforme ilustra a Figura 2.1. Este geralmente é dado em termos percentuais (25% por exemplo) da população total do mapa.

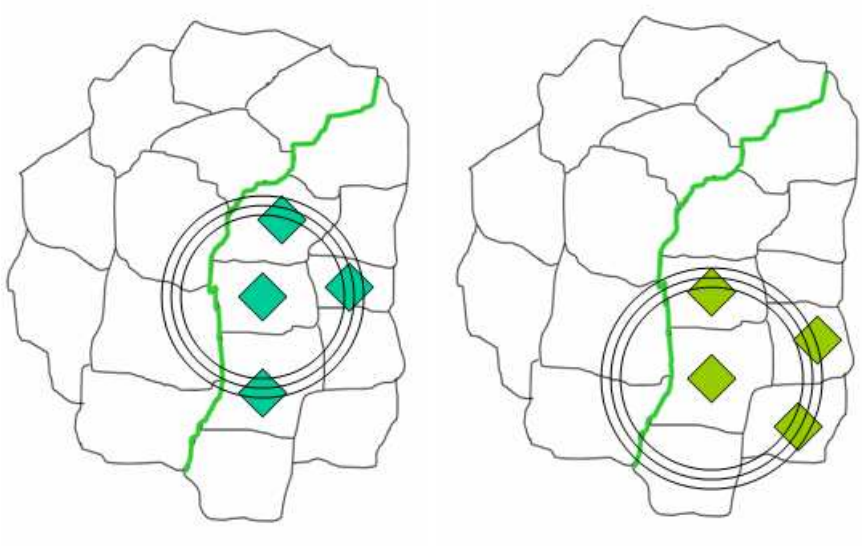


Figura 2.1: Ilustração da varredura circular em dois centróides do mapa.

Alternativamente poderíamos detectar um cluster levando em conta simplesmente a incidência de casos em cada zona. Ou seja, o número de casos observados dividido pela população, ou ainda, o risco relativo que é o número observado de casos dividido pelo número esperado de casos. A significância desta incidência não é considerada nessas duas idéias, pois um aumento no risco relativo é tão mais significativo quanto maior seja a população em risco. Em outras palavras, embora uma zona possa apresentar um risco relativo alto, se sua população é pequena, ela se torna pouco significante.

A proposta de Kulldorff (1997) contorna o problema de significância aqui descrito. Ele sugere uma varredura em todas as m^2 zonas circulares possíveis (com algumas possíveis re-

petições), procurando a de maior $LR(z)$. Necessita-se comparar este valor com o $max(LR(z))$ para os mapas de casos distribuídos aleatoriamente sob a hipótese nula. Esse procedimento é repetido milhares de vezes para cada conjunto de casos distribuídos aleatoriamente.

Utiliza-se simulações de Monte Carlo para testar a significância do teste. Distribui-se aleatoriamente o número total de casos do mapa entre as regiões, supondo a hipótese nula verdadeira. Compara-se o resultado observado de $LR(z)$ com os dados obtidos das simulações de Monte Carlo e caso haja divergência entre eles, acredita-se na existência de conglomerados espaciais ou clusters.

2.2 Algoritmo

O algoritmo para realizar o método de detecção de conglomerados proposto por Kulldorff (1997) é o seguinte:

1. Escolher uma região no mapa em estudo;

2. Calcular as distâncias até as outras regiões, ordenando-as em ordem crescente, e guardando-as em um vetor;

3. Criar um círculo centrado na região escolhida no passo 1 e continuamente aumentar o seu raio de acordo com as distâncias encontradas no passo 2. Para cada região que entrar no círculo atualizar o número de casos $c(z)$ e a população $n(z)$ dentro do círculo z. Calcular $LR(z)$ para cada par $(c(z), n(z))$. Registrar o círculo com maior $LR(z)$ até o momento;

4. Repetir os passos 1, 2 e 3 para cada região do mapa;

5. Utilizar simulações de Monte Carlo para avaliar a significância do teste:

   (a) Gerar um conjunto de casos independentes, em que C casos são distribuídos ao acaso entre as m regiões de acordo com a hipótese nula, isto é, cada região tem um número esperado de casos $\mu(z)$, e a distribuição de casos segue uma Multinomial;

(b) Calcule $T = \max_z\{LR(z)\}$ de acordo com os passos 1-4;

(c) Repita os passos (a) e (b) para um número grande B de simulações;

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

6. Se a hipótese nula for rejeitada, então a zona $\hat{z}$ associada com a maximização de $LR(z)$ é o cluster mais plausível e deve ser armazenada para que se faça o mapa destacando o cluster encontrado.

2.3 Características

O teste proposto por Kulldorff (1997) tem alguns pontos a serem destacados. As suas principais vantagens são: (a) levar em consideração a densidade populacional não constante no mapa; (b) procurar clusters sem especificação prévia de sua localização e tamanho; (c) se a hipótese nula é rejeitada, o teste fornece a localização do cluster mais verossímil que levou à rejeição; (d) é uniformemente mais poderoso[1]; (e) fornecer uma estimativa para o valor p. Além disso, o método pode ser modificado para levar em conta qualquer número de variáveis de risco conhecidas, tais como idade e sexo.

O método do Scan circular (Kulldorff and Nagarwalla, 1995) apresenta as seguintes desvantagens: (a) o método fixa a forma geométrica dos candidatos a clusters como círculos ou quadrados. Isto tende a criar clusters compactos englobando muitas vezes regiões que, de fato, não fazem parte do conglomerado (superestimação), ou, pode deixar de englobar regiões que de fato estariam no cluster, mas que não foram consideradas (subestimação) devido

[1]Um teste uniformemente mais poderoso é um teste de hipótese que tem o maior poder (probabilidade do teste rejeitar corretamente a hipótese nula) entre todos os possíveis testes de um dado tamanho. Mais detalhes em Casella and Berger (2002).

ao fato do centróide estar fora do círculo avaliado; (b) o método tem poder baixo diante de clusters com formatos irregulares; (c) o método tem um poder baixo contra alternativas com um grande número de pequenos clusters localizados em posições bastante diferentes.

A estatística Scan espacial definida em Kulldorff (1997) não exige nenhuma restrição no formato dos clusters. No entanto, a delineação geográfica de clusters com formato irregular apresenta algumas dificuldades. A liberdade em avaliar clusters de qualquer formato geométrico diminui o poder de detecção (Duczmal et al., 2006). Isto acontece porque a coleção de todas as zonas conectadas, independentemente do seu formato, é muito grande. O valor máximo da função objetivo provavelmente está associada ao "verdadeiro" formato dos clusters, que apenas liga as regiões com maiores razões de verossimilhança do mapa sem contribuir com a descoberta de soluções geograficamente significativas que delineiam corretamente o cluster "verdadeiro". Esse "ruído" existente contribui para que as legítimas soluções não sejam distinguidas das demais. O problema ocorre em muitos algoritmos de detecção de clusters com formato irregular e pode ser abrandado em parte por limitar o número máximo de regiões que devem compor o cluster. A solução que tem sido adotada em diversos trabalhos científicos da área consiste em aplicar uma função de penalidade que leve em conta também o formato geométrico dos clusters. Duczmal et al. (2006) propuseram que o valor da estatística scan fosse reduzida de acordo com a irregularidade do formato do cluster através de sua compacidade geométrica, que generalizou a penalidade adotada para o caso especial das elipses, Kulldorff et al. (2006). Com a necessidade de variar a quantidade de penalização conforme o formato, diversas soluções de clusters são encontradas, desde os clusters mais circulares até aqueles com formas muito irregulares. Desta maneira, as técnicas de otimização multiobjetivo entraram na pesquisa por algoritmos mais interessantes que levassem em conta múltiplos objetivos durante a busca por clusters espaciais.

O capítulo 3 apresenta as técnicas de otimização multiobjetivo relacionadas a este trabalho.

Capítulo 3

Técnicas de Otimização Multiobjetivo

Muitos problemas do mundo real apresentam vários objetivos que podem ser conflitantes entre si. Por exemplo, suponhamos que desejamos comprar um automóvel com base em dois objetivos: maximizar o desempenho e minimizar o preço. Neste contexto, a *otimização* é a tarefa de encontrar uma ou mais soluções que atendam aos dois objetivos mencionados. Se o exemplo em questão envolvesse apenas uma *função objetivo*, desempenho ou preço, esse problema seria classificado como de *otimização monoobjetivo*, e sua solução é denominada *solução ótima*. Como o nosso interesse é atender simultaneamente aos dois objetivos, o problema é qualificado como sendo de *otimização multiobjetivo*. Neste caso, geralmente não há uma única solução ótima, mas um conjunto de alternativas com *compromissos* (*trade-offs*) diferentes, chamado de *soluções ótimas de Pareto*, ou *soluções não-dominadas*. Apesar da existência de múltiplas soluções ótimas de Pareto, na prática, geralmente apenas uma destas soluções será escolhida.

A Figura 3.1 ilustra o exemplo do automóvel, em que encontramos sete modelos disponíveis no mercado. A Figura 3.2 destaca as três soluções mais interessantes no sentido de maximizar o desempenho e minimizar o seu preço simultaneamente. Pode-se notar que nenhum dos retângulos hachurados (Figura 3.2(a)-(c)) contém soluções, indicando que não existem soluções que são melhores simultaneamente em relação a preço e desempenho, para cada um dos três pontos escuros (as 3 soluções ótimas do conjunto de 7 pontos, Figura 3.2(d)). Em outras palavras, apesar de não ser possível encontrar algum automóvel que seja melhor

do que todos os outros em relação ao preço e desempenho, ainda assim podemos escolher as três soluções ótimas da figura, que se caracterizam por não serem piores do que nenhuma outra simultaneamente nos dois objetivos (preço e desempenho). Assim, as três soluções ótimas podem ser pensadas como as três únicas soluções que não são imediatamente eliminadas de nossa consideração; todas as demais 4 soluções são eliminadas por serem piores do que alguma outra nos dois objetivos simultaneamente.

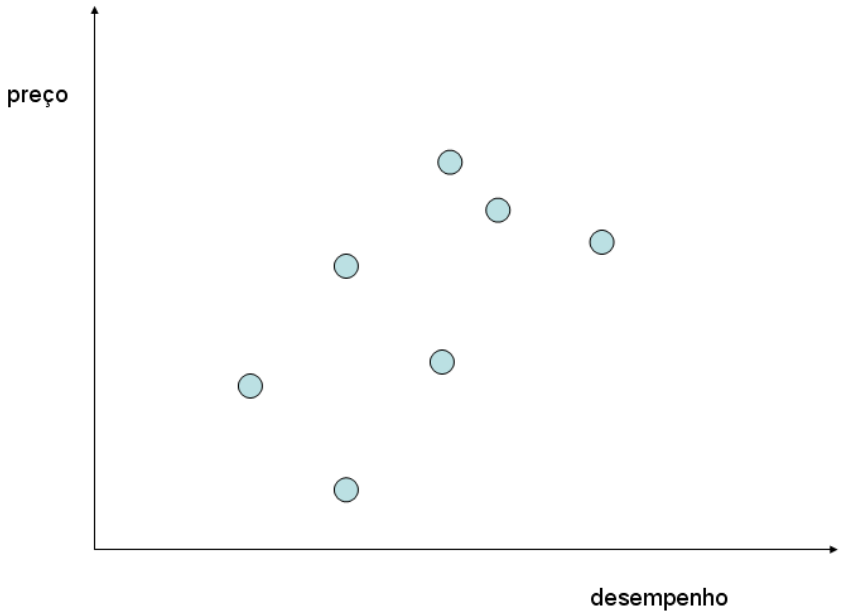


Figura 3.1: Desempenho x Preço para sete modelos de automóveis disponíveis no mercado.

Ao formular esse problema como a otimização de uma função monoobjetivo, a decisão da solução ótima é pré-estabelecida. Um problema grave ocorre, pois a escolha dessa função monoobjetivo é artificial, ao penalizar arbitrariamente um dos objetivos. Por exemplo, poderíamos escrever a função de avaliação de preço como f_1 e a função de avaliação de desempenho como f_2, e nossa função monoobjetivo g seria uma combinação linear de f_1 e f_2, $g = af_1 + bf_2$. A solução encontrada seria a solução que estivesse na curva de nível mais extrema da função g (Figura 3.3). O problema desse enfoque é que escolhemos os pesos a e b antes de conhecer os pontos soluções, e portanto o processo de escolha da solução ótima é determinado antes do processo de otimização. Pesos diferentes dão origem a funções g distin-

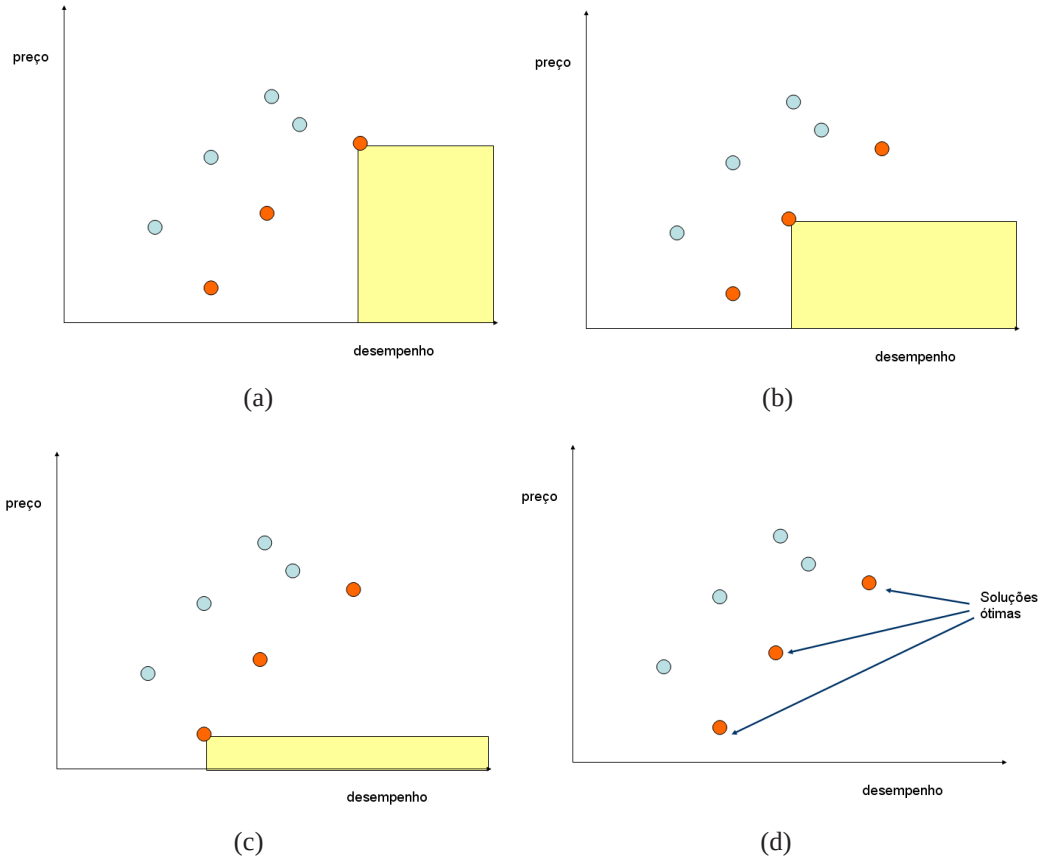
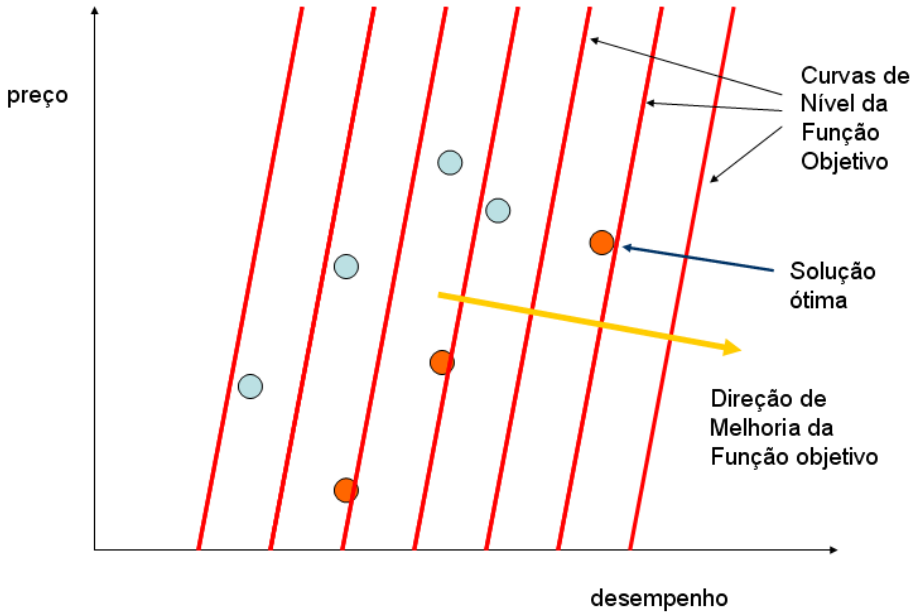


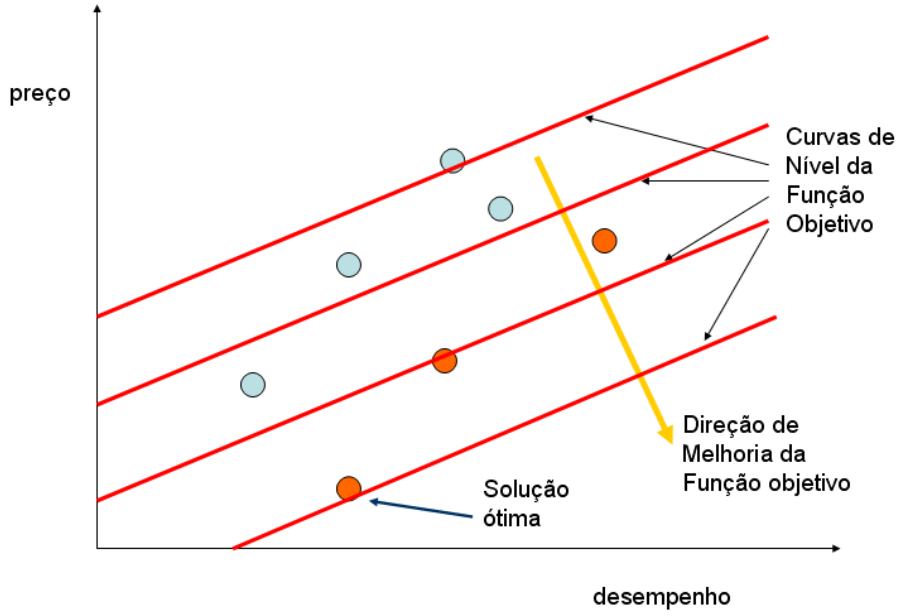
Figura 3.2: Escolhas ótimas que maximizam o desempenho e minimizam o preço simultaneamente.

tas (Figuras 3.3(a) e 3.3(b)), e portanto diferentes soluções. O enfoque multiobjetivo fornece uma estratégia mais robusta, pois permite decidir qual é o subconjunto das soluções não-dominadas, eliminando assim todas as soluções dominadas que são piores simultaneamente nos dois objetivos. Assim nossa atenção pode se concentrar apenas no conjunto de soluções não-dominadas, e um critério de desempate pode ser elaborado com um conhecimento mais completo de todas as alternativas.

No enfoque multiobjetivo, essa decisão pode ser feita após se encontrar todas as soluções candidatas a ótimas (conjunto Pareto-ótimo), que será definido na seção 3.2.



(a)



(b)

Figura 3.3: Exemplos de direções de melhoria da função objetivo se o problema fosse mono-objetivo.

Portanto, em otimização multiobjetivo há pelo menos duas tarefas importantes: uma tarefa de otimização para encontrar as soluções ótimas de Pareto (que envolve um procedimento computacional) e uma tarefa de tomada de decisão para escolher uma única solução, dentre as soluções ótimas de Pareto. A última geralmente necessita de informações extras dadas por um especialista do problema que está sendo tratado.

As seções 3.1, 3.2 e 3.3 são apresentados os algoritmos genéticos, alguns conceitos em otimização multiobjetivo e os algoritmos genéticos multiobjetivo, respectivamente. Tais algoritmos serão úteis para se compreender as abordagens desenvolvidas por Duczmal et al. (2007) e Duczmal et al. (2008). O algoritmo de Duczmal et al. (2008) foi aplicado em uma das propostas desta tese.

3.1 Algoritmos Genéticos

Os algoritmos genéticos foram desenvolvidos por Holland (1975) com o objetivo de abstrair e explicar os processos adaptativos da evolução biológica. As idéias de Holland foram desenvolvidas e divulgadas por seu aluno Goldberg em Goldberg (1989), que desenvolveu programas de computador baseados nos mecanismos evolutivos.

Uma aplicação comum dos Algoritmos Genéticos é o problema de busca. Dado um conjunto de *indivíduos*, deseja-se encontrar aquele ou aqueles que melhor atendam a certas condições especificadas. Tais algoritmos transformam uma *população* de indivíduos, cada um com um valor associado de adaptabilidade, chamado de *aptidão*, numa nova geração de indivíduos usando os princípios Darwianos de *seleção natural* dos mais aptos. Cada indivíduo na população representa uma possível *solução* para um dado problema. O que o Algoritmo Genético faz é procurar aquela que seja muito boa ou a melhor para o problema analisado, construindo gerações sucessivas de populações de indivíduos cada vez mais aptos à otimização da *função objetivo* de interesse.

Os operadores genéticos que constituem a base de um algoritmo genético são:

1. Um operador de *cruzamento*, que gera novos indivíduos a partir da combinação da

informação contida em dois ou mais indivíduos;

2. Um operador de *mutação*, que utiliza a informação contida em um indivíduo para, estocasticamente, gerar outro indivíduo;

3. Um operador de *seleção*, que decide se um indivíduo vai estar presente na próxima geração, baseado em sua aptidão.

O operador de seleção garante aos melhores indivíduos da população corrente a preferência para o processo de reprodução, permitindo que estes indivíduos possam passar as suas características às próximas gerações. O operador de cruzamento é responsável pela propagação das características dos indivíduos mais aptos da população, o que dará origem a novos indivíduos. O operador de mutação garante a introdução e manutenção da diversidade genética na população. Enquanto os operadores de cruzamento e de mutação aumentam a variação da população estudada, o operador de seleção direciona a busca.

Com base em uma população inicial, os algoritmos genéticos vão formando uma sequência de gerações. A cada iteração os operadores genéticos são aplicados à população corrente, e uma nova população é obtida.

Vamos agora mostrar mais detalhadamente como o operador de cruzamento produz novos candidatos a clusters a partir de uma população de candidatos.

Como foi dito anteriormente, o objetivo do cruzamento é gerar novos indivíduos, denominados filhos, a partir da combinação das características de outros elementos, tipicamente dois, denominados pais. Como os filhos reúnem características de ambos os pais, é natural imaginar que ele se encontra em algum ponto do "caminho" que os une. Alguns estarão eventualmente mais próximos de um dos pais do que de outro, mas espera-se que cada filho carregue consigo pelo menos uma pequena quantidade de características de cada um dos pais. Em problemas de variáveis contínuas é comum, por exemplo, a geração de filhos que estão no segmento de reta (o caminho mais curto, considerando a distância Euclideana) que liga os dois pais. Num contexto de variáveis discretas, porém, o conceito de caminho entre soluções não está, na maioria das vezes, definido implicitamente ou intuitivamente, pela ausência da

noção de vizinhança. Muitas vezes é necessário que se defina uma métrica adequada à natureza do problema, para que se possa trabalhar com o conceito de vizinhança. A partir daí é que será possível definir um caminho partindo de um pai, saltando de um indivíduo para um de seus vizinhos, e assim sucessivamente, até que se alcance o outro pai. O objetivo do nosso operador de cruzamento é, então, obter uma sequência de indivíduos que se encontram no caminho entre dois subgrafos pais.

A seguir tem-se os aspectos estruturais para representar um mapa através de um grafo, conforme descritos em Carrano (2007) e em Cancado (2009).

Definição 3.1 (Grafo) *Um grafo G é um par* $G = (V, A)$, *onde* $V = \{v_1, v_2, ..., v_n\}$ *é o conjunto de seus vértices e A é o conjunto de todas as arestas* $a_{i,j}$, *onde* v_i *e* v_j *são adjacentes, com* v_i, $v_j \in V$.

A cada vértice v_k, $k = 1, ..., n$, associa-se um dos n centróides e, portanto, cada vértice está associado a uma região. Se duas regiões i e j têm uma fronteira em comum, então os vértices v_i e v_j correspondentes são adjacentes e, portanto, ligados por uma aresta $a_{i,j}$. A representação do mapa através de um grafo apresenta algumas vantagens sobre outros tipos de estruturas. Conceitos de caminhos e conexidade estão bem definidos para estruturas de grafos. Além disso são conhecidos vários algoritmos de manipulação e busca eficientes sobre essas estruturas.

- Dados dois subgrafos A e B, tais que $A \cap B \neq \varnothing$, chamados pais, sejam $C = A \cap B$ e D o maior subgrafo conexo cujos vértices estão em C, ou seja, D é o maior subconjunto conexo dos vértices que formam o conjunto C;

- Atribuiremos um nível para cada vértice do pai A. Cada um dos n_d vértices de D (que também são vértices de A) recebe o nível zero;

- Escolhemos aleatoriamente um vértice v_1 adjacente a qualquer vértice de $A_0 = D$, com $v_1 \in A - A_0$, e a ele associamos o nível um. Depois, escolhemos aleatoriamente um vértice v_2 adjacente a qualquer vértice de $A_1 = D \cup \{v_1\}$, com $v_2 \in A - A_1$, e a ele

associamos o nível 2. No i-ésimo passo, escolhemos aleatoriamente um vértice v_i adjacente a qualquer vértice de $A_{i-1} = D \cup \{v_1, v_2, ..., v_{i-1}\}$, com $v_i \in A - A_{i-1}$;

- Repetimos esse passo até que todos os n_a vértices de $A - D$ tenham sido escolhidos e tenham recebido seus respectivos níveis (veja o exemplo de atribuição de níveis na Figura 3.4, no meio).

Percebe-se no procedimento acima que a escolha dos níveis não é única.

A Figura 3.4, de Cancado (2009), ilustra uma geração de filhos a partir de dois pais {a, b, c, d, e} e {c, f, g, h, i} dentro do mapa (acima, à esquerda). Os pais têm a região c em comum. A numeração dos níveis exemplificada (no meio, acima) gera os filhos {b, c, d, e, g}, {b, c, d, f, g} e {b, c, f, g, h} (apontados com setas pontilhadas). Os filhos {a, b, c, d, e} e {c, f, g, h, i} (apontados com setas sólidas) são idênticos a seus pais.

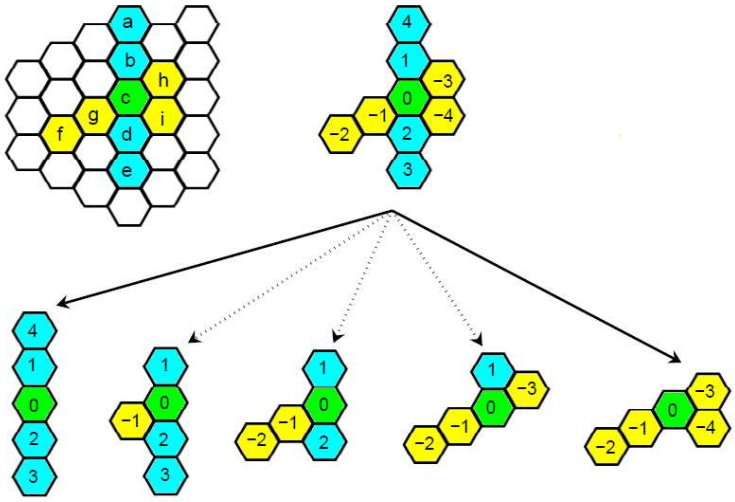


Figura 3.4: Geração de filhotes em um algoritmo genético (Cancado (2009)).

Os n_a vértices do pai A mais o nó virtual r (formado pela fusão dos vértices no conjunto D), juntamente com os segmentos orientados (v_j, v_k), onde v_k foi escolhido como adjacente a v_j no k-ésimo passo $(j < k)$, mais os segmentos orientados (r, v_k), onde v_k é adjacente ao conjunto D, formam a árvore T_A (veja Figura 3.5) que tem a seguinte propriedade:

O processo descrito para determinação dos níveis dos vértices do pai A é feito também para os n_b vértices de $B - D$, porém usando níveis negativos ao invés de positivos.

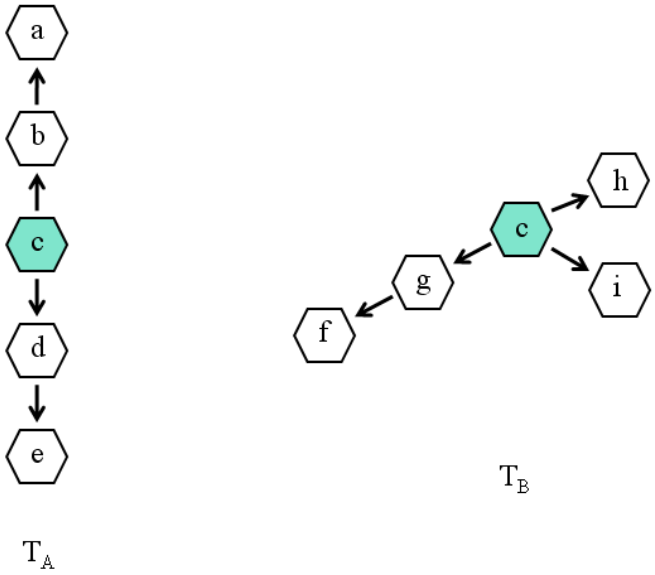


Figura 3.5: Árvores T_A e T_B.

A partir daí contruímos os filhos de A e B. Os níveis dos vértices do pai A são $\{0,\ 1,\ 2,\ 3, ...,\ n_a\}$ e do pai B $\{0,\ -1,\ -2,\ -3, ...,\ -n_b\}$. Suponha, sem perda de generalidade, que $n_a \geq n_b$. Então cada filho de A e B é formado pelos vértices associados aos níveis de cada uma das seguintes sequências, formadas a partir do pai A e em cada passo, retirando o vértice de nível mais afastado de zero do pai A (ou seja, o mais positivo) e adicionando o vértice de nível mais próximo de zero do pai B (ou seja, o menos negativo):

$$
\begin{gathered}
\{n_a - 1,\ ...,\ 1,\ 0,\ -1\} \\
\{n_a - 2,\ ...,\ 1,\ 0,\ -1,\ -2\} \\
\vdots \\
\{n_a - n_b,\ ...,\ 1,\ 0,\ -1,\ -2,\ ...,\ -n_b\} \\
\{n_a - n_b - 1,\ ...,\ 1,\ 0,\ -1,\ -2, ...,\ -n_b\} \\
\vdots \\
\{2,\ 1,\ 0,\ -1,\ -2,\ ...,\ -n_b\} \\
\{1,\ 0,\ -1,\ -2,\ ...,\ -n_b\}
\end{gathered}
\tag{3.1}
$$

Se alguma sequência tem dois níveis correspondentes ao mesmo vértice (um positivo e outro negativo para vértices em $C - D$), então basta levar em conta apenas um dos níveis. A cada vértice retirado ou adicionado saltamos de um grafo para seu vizinho. Como os filhos sempre são obtidos retirando e adicionando um vértice, o conjunto de filhos obtido no final constitui um caminho formado por passos de tamanho dois[1]. O próximo resultado representa uma grande vantagem desse processo de cruzamento.

A idéia por trás dessa operação é que os filhos formam uma transição suave entre os pais A e B. Note que o primeiro filho se parece bastante com o pai A e que o último se parece bastante com o pai B.

A cada geração, o algoritmo genético faz várias tentativas de cruzamento, uma vez que o cruzamento só é possível caso haja interseção não-vazia entre os pais. Essas tentativas cessam caso ele atinja o número máximo ct_{max} de cruzamentos tentados ou cbs_{max} de cruzamentos bem sucedidos.

O operador mutação também é responsável por gerar novos filhos, porém de forma diferente. Esse operador substitui aleatoriamente um indivíduo por um de seus vizinhos, e esse processo adiciona ou remove indivíduos que são os vértices de um subgrafo. A Figura 3.6

[1]Obviamente poderíamos trabalhar com passos de tamanho 1, ou mesmo outros tamanhos. Essa escolha levou em conta que (1) a geração de todos os filhos pode nos conduzir a um número demasiadamente grande de soluções, consumindo muito tempo e (2) a avaliação incremental permite que avaliemos mais de dois filhos, sem aumento significativo do tempo. A escolha de passos de tamanho dois parece ser um bom compromisso entre tempo e número de soluções avaliadas.

mostra um exemplo de Mutação num subgrafo hipotético. Nesse sentido, torna-se necessário testar a conexidade nesse subgrafo modificado. Este modo de gerar filhos é computacionalmente mais caro e por isso é realizado em apenas uma pequena fração do mapa.

Finalmente, o operador seleção baseia-se na comparação da aptidão de indivíduos. O indivíduo com maior aptidão terá maior probabilidade de ser colocado na próxima geração. Portanto, a seleção é responsável pela construção da próxima geração, em que os filhotes e os pais são ordenados segundo o critério de otimização. Neste trabalho tem-se que 10% da nova geração é formado pelos melhores pais e 90% pelos melhores filhos.

O cluster será a região com maior LLR e sua significância é obtida de forma similar aos métodos já descritos no capítulo 2.

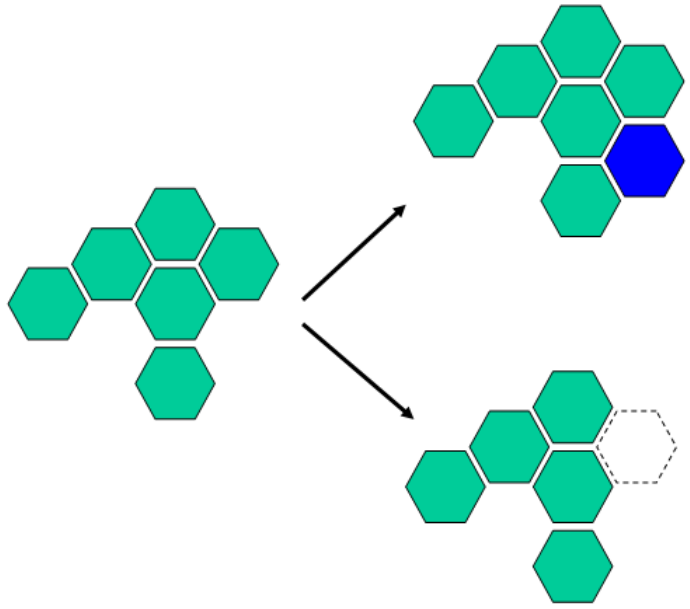


Figura 3.6: A geração dos filhotes por Mutação.

Para que se consiga chegar a um desempenho satisfatório, o algoritmo genético deve ser ajustado através de alguns parâmetros que determinam seu funcionamento. Esses parâmetros são:

- N: tamanho da população

- p_m: probabilidade de mutação
- cbs_{max}: número máximo de cruzamentos bem sucedidos
- ct_{max}: número máximo de cruzamentos tentados
- g_{max}: número de gerações, utilizado como critério de parada do algoritmo

Especificamente para o algoritmo descrito nessa tese o tamanho da população N está implicitamente definido como o número de regiões do mapa. O procedimento de geração da população inicial gera um indivíduo partindo de cada uma das N regiões. A probabilidade de mutação p_m utilizada ne maioria dos algoritmos genéticos está próxima de 0,05. No caso do nosso algoritmo, a mutação pode não ser possível, já que se a escolha for por retirar uma região pode não haver nenhuma região passível de ser retirada sem tornar o indivíduo desconexo. Por esse motivo, escolhemos uma probabilidade maior que 0,05. A taxa de mutação foi definida como 0,1. Um valor acima de 0,1 faria com que a busca se tornasse demasiadamente aleatória. Para o número máximo de cruzamentos bem sucedidos cbs_{max} optamos por utilizar o número de cruzamentos em um algoritmo padrão, que é de $\lceil N/2 \rceil$ cruzamentos. Com esse número de cruzamentos, um algoritmo normal (com um cruzamento que gerasse dois indivíduos por cruzamento) obteria N novos indivíduos filhos. No nosso algoritmo, caso esse número de cruzamentos seja atingido, em geral teremos mais que N indivíduos filhos. Para isso, verificamos para o nosso problema que com um número de tentativas de cruzamento $ct_{max} = 2N$ raramente não atingimos $\lceil N/2 \rceil$ cruzamentos bem sucedidos. Por fim, como critério de parada utilizamos o número máximo de gerações g_{max} que foi fixado em 40. A configuração de parâmetros utilizada nesta proposta foi baseada em estudos realizados por Cancado (2009).

Um algoritmo genético genérico pode ser visualizado da seguinte forma:

Algoritmo Genético Genérico

Inicialize a população de indivíduos (geração i = 1)

Avalie indivíduos na população (função objetivo ou aptidão)

Repita (evolução)

Selecione indivíduos para reprodução

Aplique operadores de cruzamento e/ou mutação

Avalie indivíduos gerados na população

Selecione indivíduos para sobreviver (geração i = i + 1)

Até critério de parada (objetivo final ou máximo de gerações)

Fim

3.2 Otimização Multiobjetivo

Nos algoritmos multiobjetivo têm-se a necessidade de otimizar simultaneamente duas ou mais funções objetivo $f_1, f_2, ..., f_n$. Assim, o problema de otimização multiobjetivo pode ser escrito na forma

$$\max_x \quad f(x) = (f_1(x), f_2(x), ..., f_n(x))$$

Na maioria das vezes os objetivos $f_1, f_2, ..., f_n$ são conflitantes, no sentido de que dificilmente uma mesma escolha de parâmetros x otimiza todos os objetivos simultaneamente. Por essa razão a busca pela melhor solução em um problema com mais de um objetivo está intimamente ligada ao conceito de dominância, dado a seguir.

Definição 3.2 (Dominância) *Seja $f(x) = (f_1(x), ..., f_n(x))$ uma função definida em um espaço X. Um ponto $x_1 \in X$ domina outro ponto $x_2 \in X$ (denota-se $x_1 \succ x_2$) se $f_i(x_1) \geq f_i(x_2)$, $i = 1, ..., n$ e se existe pelo menos um índice $k \in \{1, ..., n\}$ tal que $f_k(x_1) > f_k(x_2)$.*

Em outras palavras, um ponto x_1 domina o ponto x_2 se a avaliação de x_1 for melhor que a avaliação de x_2 em um objetivo e não for pior em nenhum outro objetivo. Caso o problema seja de minimização, a definição para $x_1 \prec x_2$ vale trocando os sinais $\geq$ e $>$ por $\leq$ e $<$, respectivamente.

Com o conceito de dominância podemos agora definir o objeto essencial na resolução de problemas de otimização multiobjetivo, a *solução Pareto-ótima*.

Definição 3.3 (Solução Pareto-ótima) *Diz-se que uma solução $x^* \in X$ é Pareto-ótima se não existe $x \in X$ tal que x domina x^*.*

Uma solução Pareto-ótima pode ainda ser chamada de solução não-dominada ou solução eficiente. O *conjunto Pareto-ótimo* é formado por todas as soluções Pareto-ótimas. A Figura 3.7 apresenta um exemplo com pontos dominados ($+$) e os pontos que formam o conjunto de Pareto ($\oplus$).

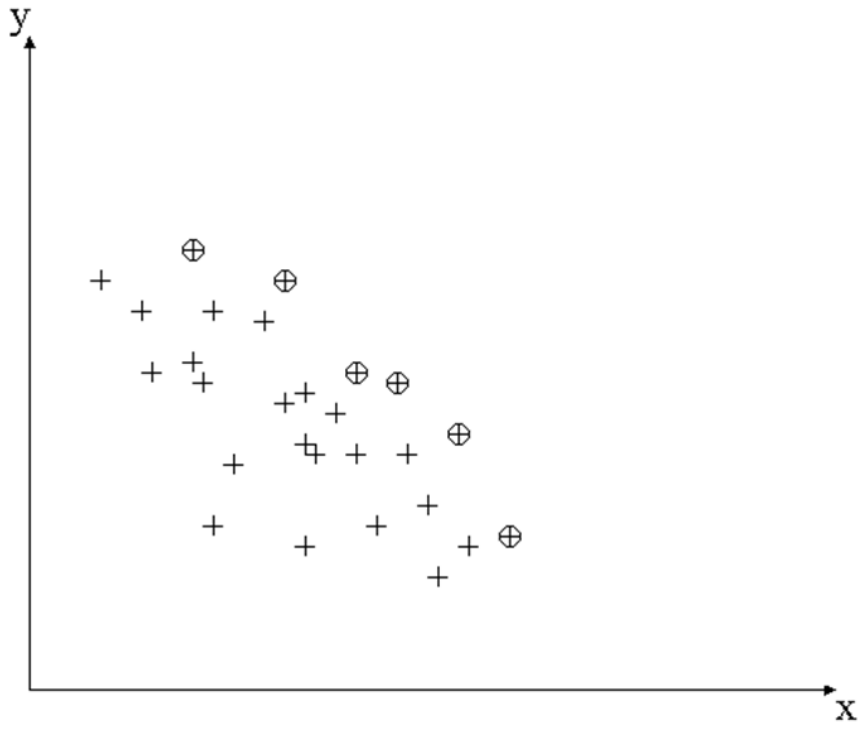


Figura 3.7: Ilustrando o conjunto de Pareto

O algoritmo genético apresentado na seção 3.1 será modificado na seção 3.3 para lidar simultaneamente com as duas grandezas: a compacidade K e a estatística espacial scan.

3.3 Algoritmos Genéticos Multiobjetivo

Os algoritmos evolutivos têm sido utilizados com sucesso para determinar o conjunto Pareto-ótimo nos problemas de otimização multiobjetivo. Chankong and Haimes (1983) e Schaffer (1984) apresentaram implementações práticas para tratamento de problemas multiobjetivo. Fonseca and Fleming (1993) foram pioneiros ao proporem o Algoritmo Genético Multiobjetivo (MOGA) que ordena as soluções não-dominadas. Nessa proposta, não se garantia a diversidade e havia um alto custo computacional. Com base no NSGA *(Non-dominated Sorting Genetic Algorithm)* de Srinivas and Deb (1995), Deb (2001); Deb et al. (2002) propuseram o NSGA-II em que o operador de seleção favorece elitismo e o algoritmo preserva diversidade por compartilhar a aptidão baseada em ordenamento das soluções. Muitos outros trabalhos surgiram neste intervalo, como Coello (1996), Fonseca and Fleming (1995), Nepomuceno et al. (2003), Takahashi et al. (2003), Takahashi et al. (2004) e Carrano et al. (2006) que desenvolveram diferentes algoritmos genéticos multiobjetivo para finalidades diversas.

Coello (1996) classifica os algoritmos evolutivos para otimização multiobjetivo em três categorias: técnicas que utilizam funções de agregação, técnicas não baseadas na teoria de Pareto e técnicas baseadas na teoria de Pareto. As extensões da Estatística Scan propostas neste trabalho estão relacionadas apenas as técnicas multiobjetivo baseadas na teoria de Pareto. Duas são as finalidades quando se deseja determinar o conjunto Pareto de problemas multiobjetivo via métodos evolucionários: guiar a busca na direção do conjunto ótimo de Pareto e manter a diversidade da população na fronteira de Pareto.

Conforme descreve Carrano (2007), o conjunto de Pareto pode auxiliar numa tomada de decisão planejada em diversos problemas cujos objetivos sejam conflitantes, pois de acordo com a disposição das soluções no conjunto Pareto é possível avaliar o compromisso dentre os objetivos envolvidos. O analista pode avaliar o efeito de substituir uma solução por outra, sabendo o que irá perder em um objetivo e ganhar em outro.

Duczmal et al. (2008) desenvolveram um algoritmo multiobjetivo baseado no algoritmo genético para detecção e inferência de clusters espaciais. Dois objetivos competetivos são

envolvidos na busca dos clusters: uma medida de regularidade do formato e o valor da estatística espacial scan. Eles propuseram critérios quantitativos para escolher a melhor solução do conjunto dos clusters possíveis. Desta maneira, a escolha que geralmente se dava de forma arbitrária e subjetiva passou a ser através de metodologia teoricamente sistemática para encontrar tal solução.

As Figuras 3.8, 3.9 e 3.10 ilustram a evolução da população no algoritmo genético multiobjetivo. Os gráficos se referem a *LLR* $\times$ *compacidade* ao longo das gerações 1, 30 e 500, respectivamente.

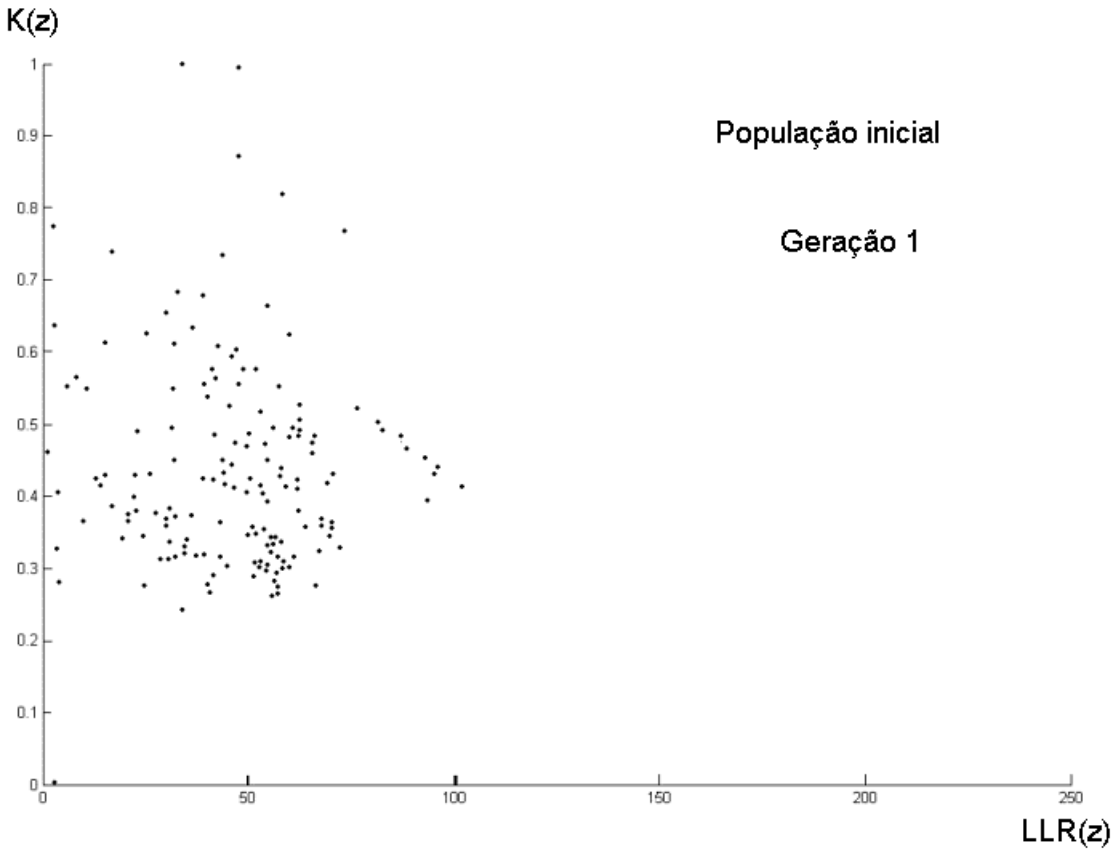


Figura 3.8: Evolução da população no algoritmo genético multiobjetivo, geração 1.

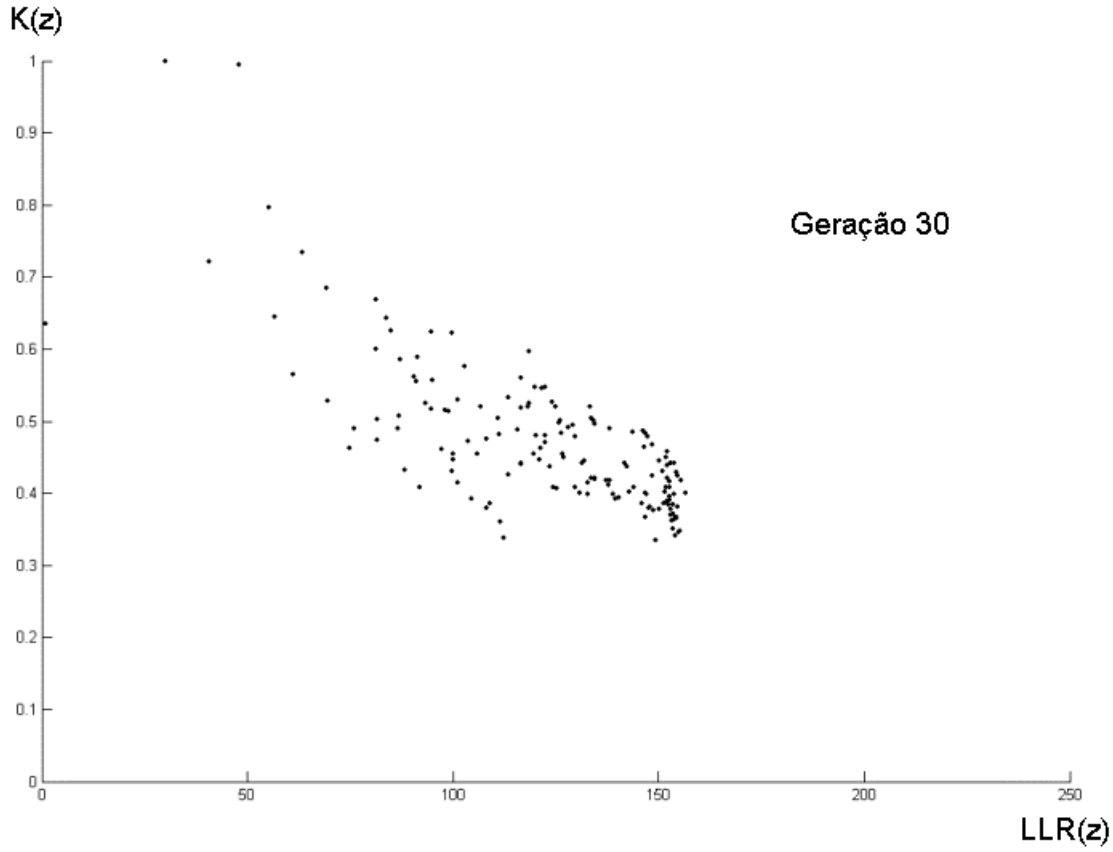


Figura 3.9: Evolução da população no algoritmo genético multiobjetivo, geração 30.

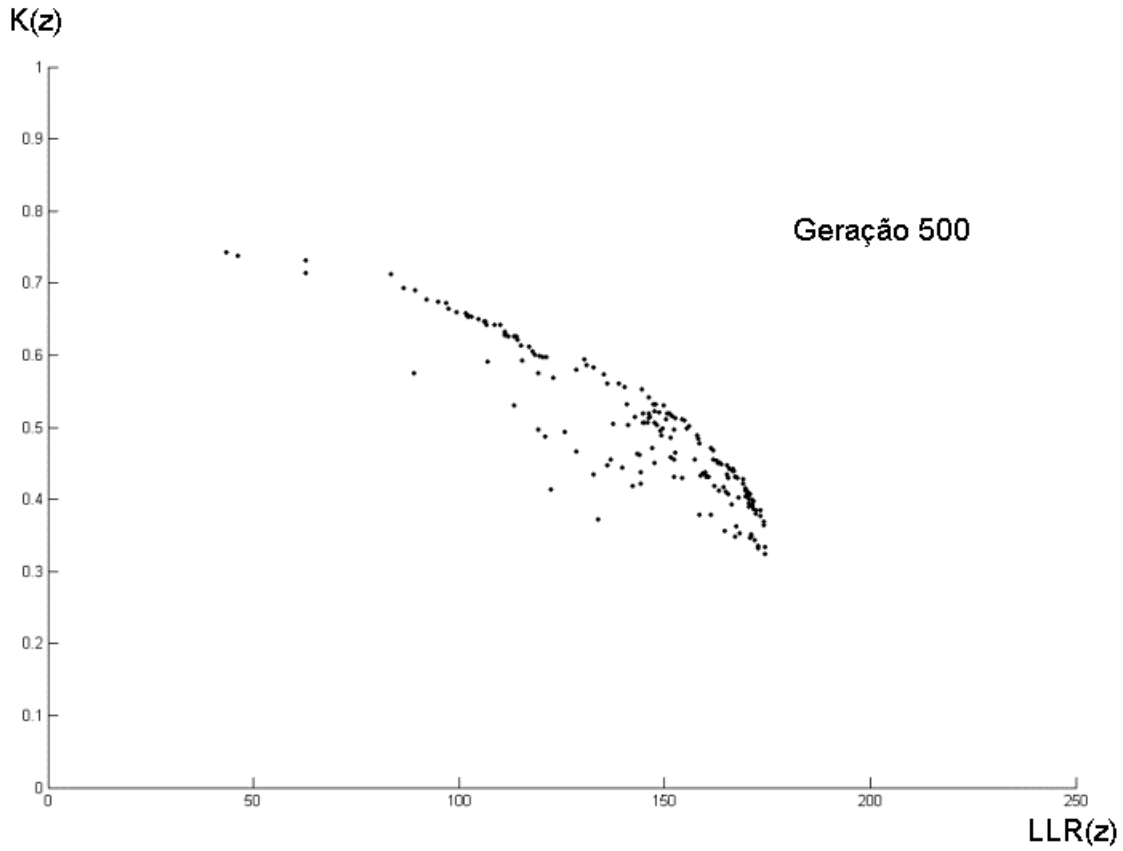


Figura 3.10: Evolução da população no algoritmo genético multiobjetivo, geração 500.

Com base nessas figuras, observa-se o movimento dos pontos em direção a valores maiores de *LLR* e *K*. Nota-se também que a convergência é muito rápida para pontos com alta compacidade. A população pode conter múltiplas cópias de alguns indivíduos, principalmente nas gerações finais. O conjunto de Pareto da última geração é considerado a solução dada pelo algoritmo genético. As últimas populações vão se tornando cada vez mais próximas de seus respectivos conjuntos de Pareto, e essa proximidade pode ser utilizada como um critério de convergência conforme destaca Cancado (2009).

Uma das extensões propostas neste trabalho utiliza operadores específicos propostos por Duczmal et al. (2008) para explorar a estrutura do problema de busca de clusters espaciais mais verossímeis em um mapa composto por *m* regiões. O algoritmo genético multiobjetivo utilizado nesta tese está descrito na seção 4.2 do capítulo 4.

Capítulo 4

Estatística Scan de Adjacência Modificada: um método semi-focado

4.1 Introdução

Atualmente, ferramentas desenvolvidas para a detecção de clusters espaciais de doença são utilizados por epidemiologistas para estudos etiológicos (Lawson et al., 2000) e vigilância sindrômica (Duczmal and Buckeridge, 2006; Kulldorff et al., 2005, 2006, 2007). O conjunto de clusters potenciais é construído, no caso mais simples, como a coleção de zonas definidas por janelas de formato regular, tais como círculos ou quadrados (Kulldorff and Nagarwalla, 1995). Em muitos cenários, porém, estamos interessados na detecção de clusters espaciais que não estão restritos à forma regular. Recentemente, outros métodos foram propostos para detectar clusters espaciais de formato irregular, incluindo Patil and Taillie (2004), Duczmal and Assuncao (2004), Duczmal and Buckeridge (2006), Duczmal et al. (2006, 2007, 2008, 2009), Iyengar (2004), Sahajpal et al. (2004), Conley and MacGill (2005), Tango and Takahashi (2005), Neill and Moore (2006), Assuncao et al. (2006) e Kulldorff et al. (2003, 2006, 2007). Uma estatística é então empregada para avaliar cada um dos clusters potenciais. A Estatística Scan Espacial (Kulldorff, 1997) é muito popular e é usada pelos softwares SatScan (Kulldorff, 1999) e ClusterSeer (TerraSeer, 2004). O cluster mais provável é definido como o cluster do conjunto potencial que maximiza a estatística. Um método Monte Carlo é usado para calcular o valor da probabilidade de significância (p) do cluster mais provável.

Ao invés de trabalhar com um mapa livre de característica ambientais, faz sentido adicionar tanta informação relevante quanto possível. Besag and Newell (1991) classificaram os testes de conglomerados em Gerais e Focados. Nos testes focados, 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 se observar os dados. Os testes gerais procuram identificar as áreas geográficas com um risco significativamente elevado sem especificar previamente quais e quantas áreas seriam estas. Neste capítulo, propomos uma alternativa semi-focada que, ao invés de considerar um mapa livre de características ambientais, tenta adicionar o máximo de informação relevante relacionada com a distribuição espacial de casos da doença avaliada. Esta abordagem deve ser comparada aos testes gerais, que não especificam qualquer característica especial para as diferentes regiões do mapa. Cada uma destas classes de algoritmos tem suas próprias especificidades e vantagens.

A extensão proposta neste capítulo explora uma abordagem alternativa baseada no conhecimento prévio de características ambientais em que se suspeita serem significantes para explicar a distribuição espacial dos casos da doença estudada. Um exemplo de tal característica sugere a conexão mais forte das regiões encontradas no mesmo trajeto do vento predominante quando o estudo se refere a poluentes transportados por via aérea. Um outro exemplo é a proximidade a um rio, lago ou praia, relativo à transmissão de uma doença pela água: o rio pode conectar as regiões que não são imediatamente vizinhas. Em uma situação inteiramente diferente, o mesmo rio pode obstruir a vizinhança entre regiões que possuem um limite comum, se considerarmos uma doença transmitida por animais contaminados e que não podem cruzar o rio. Neste caso, a conectividade da vizinhança é enfraquecida pela presença do mesmo rio.

A metodologia proposta neste capítulo pretende testar hipóteses relacionadas às causas ambientais geográficas através da avaliação comparativa da significância dos clusters mais prováveis detectados sob os correspondentes mapas modificados conforme aquelas causas. A questão aqui é decidir se o cluster mais provável detectado em um mapa com a estrutura de adjacência é menos significante que o cluster mais provável detectado no mapa alterado

com adjacência alterada. Caso a resposta seja positiva, tem-se indicações de que o efeito ambiental introduzido pela alteração na estrutura de adjacência teve um importante papel na detecção do cluster. Sendo a resposta negativa, então não existe razões para acreditar que a presença do cluster possa ser atribuída à causa ambiental mencionada.

Como mostram os exemplos apresentados, para cada cenário introduzimos modificações nos diferentes lugares do mapa: naturalmente, esperamos enfraquecer a detecção dos clusters se modificarmos algum vizinho que não esteja completamente relacionado com a causa ambiental. Desenvolvemos uma nova metodologia, baseada em um algoritmo genético de otimização multiobjetivo, que foi desenvolvido por Duczmal et al. (2008) para selecionar a melhor solução de cluster, entre as muitas possíveis soluções encontradas. Esse algoritmo busca maximizar dois objetivos competitivos: nominalmente a razão de verossimilhança da estatística scan (Kulldorff, 1997) e a regularidade do formato do cluster (Duczmal et al., 2006).

Neste caso, executamos sequencialmente o algoritmo genético multiobjetivo para vários conjuntos de parâmetros ou cenários, e em seguida obtivemos um grande número de soluções de clusters, que são comparadas e julgadas em termos de suas significâncias. Os conjuntos de parâmetros controlam a intensidade da adjacência, que é baseada na similaridade/dissimilaridade ambiental entre regiões. O objetivo é testar quais condições ambientais específicas induzem clusters mais significativos, permitindo assim ao algoritmo aceitar ou rejeitar diferentes hipóteses sobre a relevância de fatores ambientais geográficos.

O algoritmo genético multiobjetivo é descrito na seção 4.2. A seção 4.4 introduz a nova metodologia de mapas reforçados com estruturas ambientalmente definidas. O método é numericamente avaliado na seção 4.5 e uma aplicação é discutida na seção 4.6.

4.2 O Algoritmo Genético Multiobjetivo

O algoritmo genético proposto por Duczmal et al. (2007) foi utilizado para a detecção e inferência de clusters espaciais usando a estatística scan em um mapa dividido em regiões. O algoritmo tenta maximizar uma função objetivo, modificando uma população inicial de

indivíduos para um número de gerações. A variação da população é aumentada através dos operadores cruzamento e mutação. O operador seleção escolhe os indivíduos que permanecerão na geração seguinte, mantendo o tamanho da população fixado durante o processo. O operador cruzamento cria novos indivíduos filhos, ou as zonas, misturando as características de dois pais aleatoriamente escolhidos a cada vez, conforme visto no capítulo 3. Desse modo, vários filhos são gerados, sendo estes as zonas intermediárias entre as duas zonas extremas A e B. O operador seleção ordena as zonas conforme o valor da função objetivo (estatística scan). Sendo assim, com o avanço do algoritmo através das gerações, espera-se encontrar indivíduos com valores da função objetivo cada vez mais elevados.

A significância estatística do cluster mais provável de casos observados é calculado através de uma simulação Monte Carlo (Dwass, 1957), como descrito no capítulo 2. Sob a hipótese nula, os casos simulados são distribuídos aleatoriamente sobre o mapa e a estatística scan é calculada para o cluster mais provável. A quantidade de casos simulados é igual ao total de casos observados no mapa. Este procedimento é repetido milhares de vezes e os valores obtidos são comparados com aquele do cluster mais provável de casos observados, produzindo uma estimativa de valor p. O algoritmo tem uma convergência rápida e um bom poder de detecção relativamente quando comparado com outros algoritmos da literatura.

A medida quantitativa utilizada para avaliar a regularidade do formato geométrico do cluster foi a compacidade geométrica, proposta por Duczmal et al. (2006). Dado um objeto geométrico z no plano, defina $A(z)$ como a área de z e $H(z)$ como o perímetro do fecho convexo de z. A compacidade de z é dada por $K(z) = \frac{4\pi A(z)}{H(z)^2}$. Isto é equivalente a dividir $A(z)$ pela área do círculo com perímetro $H(z)$. A área $A(z)$ é quase sempre fornecida pelo banco de dados usuais, e o perímetro $H(z)$ é calculado através de uma rotina implementada por Duczmal et al. (2006). A compacidade de uma zona depende de sua forma, mas não de seu tamanho. O objeto que apresenta a maior compacidade é o círculo que tem uma compacidade igual a um, enquanto um quadrado, por exemplo, tem compacidade $\frac{\pi}{4}$. A compacidade pode ser usada para filtrar a presença de clusters formados por árvores muito grandes e com alto valor de LLR e baixo valor de compacidade, que não tenha qualquer significado geográfico

no mapa.

O algoritmo genético é modificado para tratar simultaneamente as duas quantidades: a compacidade (K) e estatística espacial scan de Kulldorff (LLR), constituindo o algoritmo genético multiobjetivo proposto por Duczmal et al. (2008). Os pares (LLR_i, K_i), que indicam a compacidade e a estatística scan calculadas para cada indivíduo i (ou zona), são plotados no plano cartesiano. A cada zona *z* associamos o ponto $(LLR(z), K(z))$ no conjunto $[0,\infty) \times (0,1]$ contido no plano cartesiano. O operador seleção é definido em termos dos dois objetivos maximizando a compacidade e a estatística scan. Este operador baseia-se no conceito de dominância: um ponto é considerado dominado se ele for pior que outro ponto em pelo menos um objetivo, enquanto não for melhor que este ponto em qualquer outro objetivo (Chankong and Haimes, 1983). O conjunto Pareto é o conjunto que não contém qualquer solução dominada (Takahashi et al., 2003).

Os operadores de cruzamento e de mutação são idênticos àqueles usados no algoritmo genético do capítulo 3. A população genética inicial de M indivíduos também é construída de forma idêntica. O operador de seleção é modificado como se segue. No início de cada geração começamos com a lista da geração atual, que consiste do conjunto dos pais selecionados na geração anterior (ou na população inicial, para a primeira geração). O operador cruzamento escolhe pais aleatoriamente na lista da geração atual, produzindo um grande número de novos filhos. Aqueles indivíduos novos são adicionados à lista da geração atual. A lista da geração seguinte, inicialmente vazia, armazena os indivíduos que sobreviverão para a próxima geração. Isto é feito calculando o conjunto Pareto da lista da geração atual, que é então transferida para a próxima geração inicialmente vazia. O mesmo conjunto Pareto é também removido da lista da geração corrente. Então, um novo conjunto Pareto dos indivíduos restantes é calculado novamente, sendo o procedimento repetido até que a lista da nova geração cresça até conter pelo menos M indivíduos. Os indivíduos eventualmente excessivos adicionados na última etapa são removidos aleatoriamente para formar uma nova lista da geração seguinte com exatamente M indivíduos. A lista da geração atual é finalmente substituída pela lista da geração seguinte. O operador cruzamento constrói novamente novos filhos

e o procedimento descrito neste parágrafo é repetido para um número de gerações sucessivas. Observamos então um deslocamento coletivo dos pontos para valores geralmente mais elevados de K e de LLR. Sendo assim, o conjunto Pareto da última geração é considerado a solução dada pelo algoritmo.

O algoritmo genético multiobjetivo calcula o conjunto de soluções eficientes (conjunto Pareto) da coleção de todas as soluções encontradas e, através de simulações Monte Carlo, a significância dessas soluções é avaliada conforme a estratégia apresentada na seção 4.3.

4.3 Avaliação da Significância dos Clusters

Para calcular a significância estatística dos pontos do conjunto de Pareto do mapa de casos observados devemos compará-los com os conjuntos de Pareto obtidos para cada um dos mapas de casos simulados sob a hipótese nula, obtidos através de uma simulação de Monte Carlo. O algoritmo genético multiobjetivo é executado várias vezes para mapas contendo casos distribuídos aleatoriamente conforme a distribuição Multinomial sob a hipótese nula, em que a probabilidade de ocorrência de casos em cada região é proporcional à população naquela região.

O processo de obtenção do conjunto de Pareto é repetido para cada uma dessas alocações aleatórias de casos. Esses conjuntos de Pareto são agrupados, formando uma coleção de milhares de pontos distribuídos no espaço $LLR \times K$, que é a faixa $(0,\infty) \times (0,1]$ (Figura 4.1). Nesse caso, ao invés de encontrar o ponto crítico, acima do qual consideramos que um cluster é significativo, devemos encontrar uma curva crítica. Esta curva crítica divide o plano em duas regiões de maneira que um ponto do plano será considerado um cluster significativo se estiver acima dessa curva. Recaímos então na questão de como encontrar essa curva crítica.

A fim de viabilizar a extensão paramétrica para este contexto, vamos dividir o espaço $(0,\infty) \times (0,1]$ em m faixas horizontais paralelas, sendo a j-ésima faixa o espaço $(0,\infty) \times (s_j, s_{j+1}]$, com $s_j < s_{j+1}$. Agora, para cada uma dessas faixas podemos utilizar a abordagem empírica ou paramétrica, simplesmente utilizando os pontos obtidos na simulação de Monte

Carlo que caem dentro de cada faixa (Figura 4.3). Os valores s_j e s_{j+1} devem ser escolhidos próximos o bastante de forma que a distribuição não mude muito para valores diferentes de compacidade no intervalo (s_j, s_{j+1}) e também que a faixa contenha um número suficiente de pontos que nos permita fazer inferência.

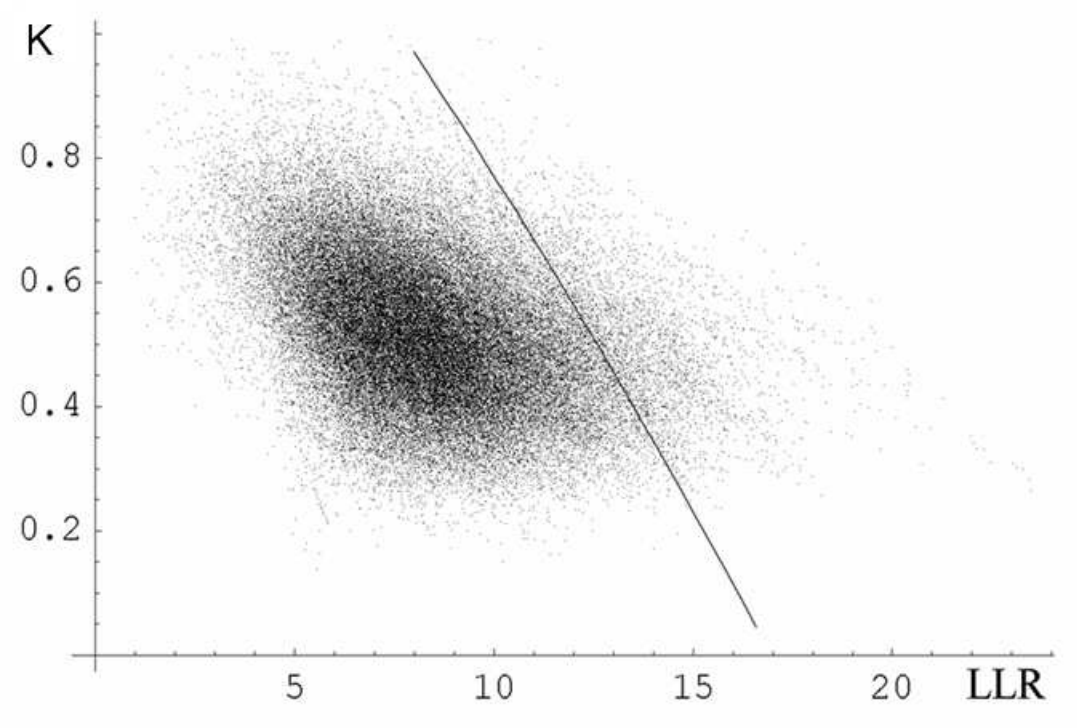


Figura 4.1: Nuvem de Pareto de casos simulados.

Abrams et al. (2006) mostraram através de testes exaustivos que, sob a hipótese nula, a estatística scan parece ter comportamento similar ao de uma distribuição Gumbel. A explicação é que esta distribuição é apropriada para estudar valores extremos, sendo o caso da estatística de Kulldorff que busca encontrar a zona cuja verossimilhança seja máxima. Johnson et al. (1995); Coles (2001) mostraram que para uma dada sequência de variáveis aleatórias $\{X_1, X_2, \cdots, X_n\}$ independentes e identicamente distribuídas, a distribuição assintótica do máximo $Y = max\{X_1, X_2, \cdots, X_n\}$, se existe, é uma distruição de valores extremos. Desta forma, podemos fazer um número menor de execuções para que se possa estimar os

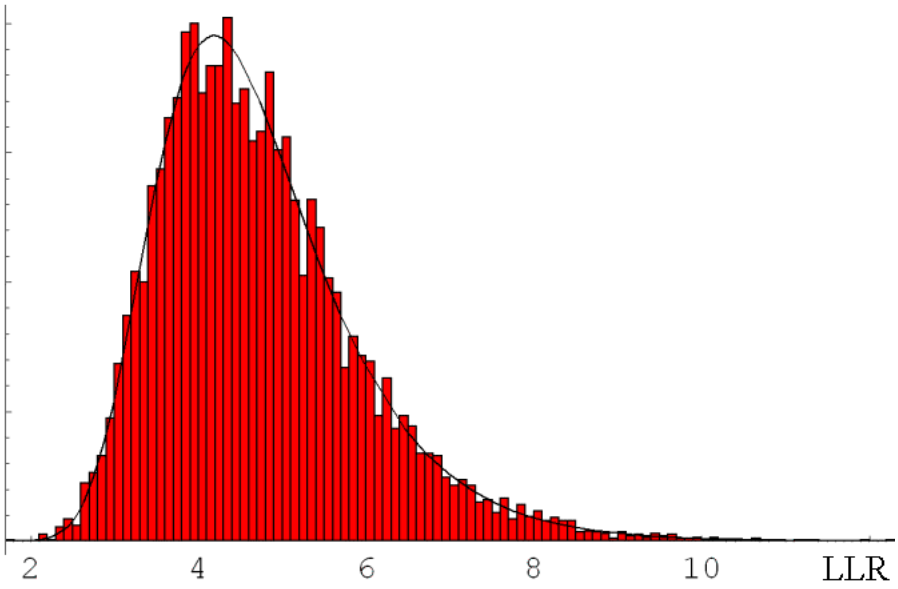


Figura 4.2: Distribuição Empírica da Estatística Scan e o respectivo ajuste pela distribuição Gumbel.

parâmetros da distribuição e assim calcular o valor p do teste. Isto evita a necessidade de simular um número grande de réplicas Monte Carlo para que se possa fazer inferência. E com base nesta idéia, em seguida calculamos a aproximação da distribuição Gumbel que melhor ajusta aos valores de LLR dos pontos presentes em cada faixa construída. A Figura 4.2 mostra um exemplo da distribuição empírica da Estatística Scan construída através de 30000 simulações de Monte Carlo do algoritmo genético sob hipótese nula para um mapa de regiões e a distribuição de Gumbel correspondente.

A vantagem do cálculo paramétrico do valor p das soluções é que podemos fazer inferência com um número razoável de simulações. Ainda que a nuvem de pontos obtidos sob a hipótese nula não avance o suficiente para envolver o conjunto de Pareto obtido para os casos observados, podemos estimar o valor para cada ponto do conjunto, utilizando uma distribuição ajustada.

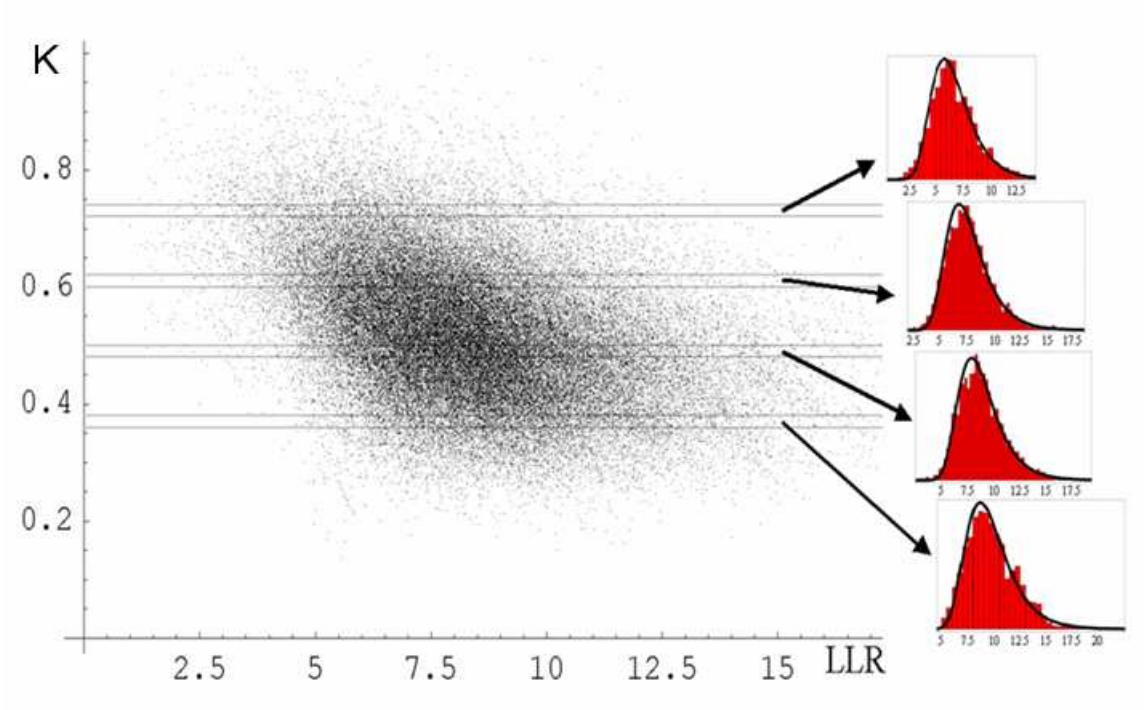


Figura 4.3: Utilização da Distribuição Gumbel no cálculo das isolinhas.

A função de densidade de probabilidade da distribuição de Gumbel para máximos é dada por:

$$f_G(x;\mu,\theta) = \frac{1}{\theta} e^{-e^{\left(\frac{\mu-x}{\theta}\right)}} e^{\left(\frac{\mu-x}{\theta}\right)}, \quad x \in \Re$$

e a função de probabilidade acumulada é dada por:

$$F_G(x;\mu,\theta) = e^{-e^{\left(\frac{\mu-x}{\theta}\right)}}, \quad x \in \Re$$

em que $\mu \in \Re$ é o parâmetro de locação e $\theta > 0$ é o parâmetro de escala.

No caso da abordagem por faixas, o espaço $(0,\infty) \times (0,1]$ é particionado pelas faixas $(0,\infty) \times (s_j, s_{j+1}]$, com $s_j < s_{j+1}$. Para cada uma dessas faixas utilizamos os pontos que caem em seu interior para podermos estimar os parâmetros necessários para se chegar à distribuição que se ajusta aos dados daquela faixa. Seja f_j a função de densidade de probabilidade da distribuição para a faixa $(0,\infty) \times (s_j, s_{j+1}]$ e seja $y_{obs} = (l,k)$ um ponto do conjunto de Pareto encontrado pelo algoritmo para os casos observados tal que $k \in (s_j, s_{j+1}]$. A função distribuição F_j da faixa j que contém o ponto k é então utilizada para calcular seu valor p, através da integral:

$$\text{Valor P} = \int_{l}^{\infty} f_j(t) d_t$$

Essa integral pode ser estimada como a proporção de pontos da faixa que caem à direita de l, em relação ao número total de pontos dessa faixa.

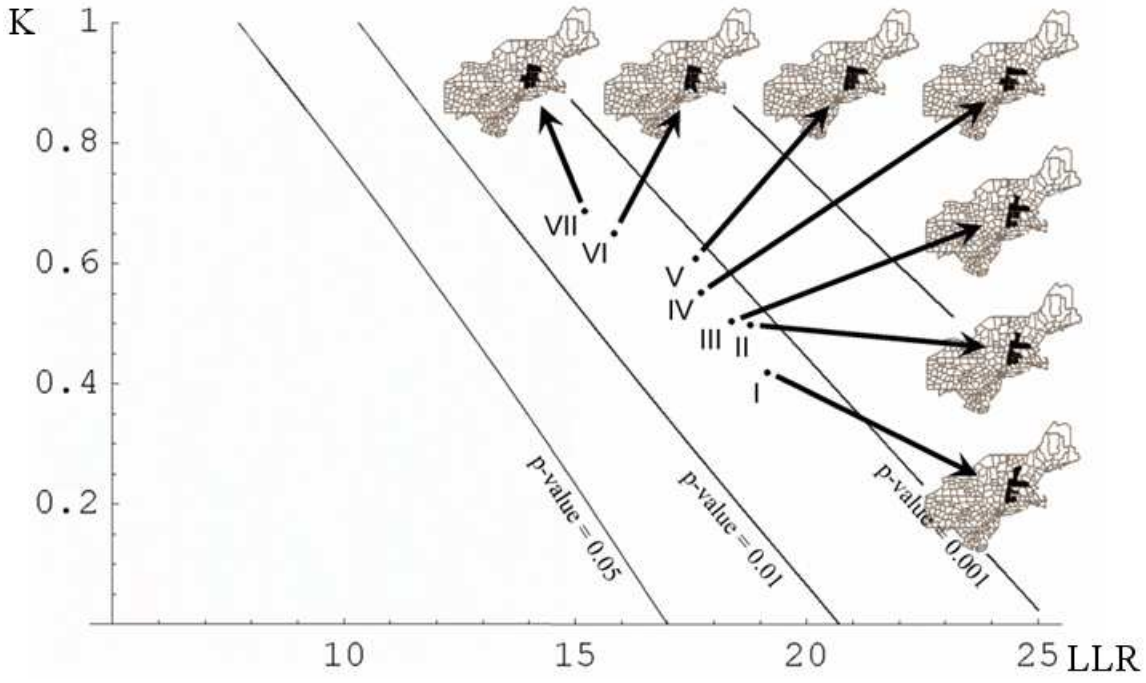


Figura 4.4: Avaliação da significância de sete candidatos a cluster.

Esse mesmo processo pode ser repetido para qualquer ponto (x,y) do espaço $(0,\infty) \times$

$(0,1]$, permitindo criar por interpolação as curvas de nível com um mesmo valor *p*, que serão chamadas aqui de isolinhas de valor *p*. Isso permite estender a idéia de ponto crítico do processo de Monte Carlo de Dwass (1957), originalmente concebida para uma única função objetivo, para otimização multi-objetivo. Uma outra maneira de se estimar a integral anterior se baseia na estimação semi-paramétrica da distribuição de pontos de cada faixa, em que os parâmetros de locação e escala são estimados através dos pontos e usados para definir a distribuição de Gumbel (Figura 4.3). Na Figura 4.1 é mostrada a isolinha de valor 0.05. Um exemplo de cluster artificial em que foram encontrados sete diferentes clusters para o conjunto de Pareto é mostrado na Figura 4.4. Nesta figura são mostradas as isolinhas de valor 0.05, 0.01, 0.001 e 0.0001, com os sete pontos de Pareto e seus clusters associados. Observe que o cluster VII é o cluster mais regular (e de menor LLR) e o cluster I é o cluster mais irregular (e de maior LLR). Pelo critério acima, o cluster mais significativo é o cluster V; esse critério é usado para desempatar os clusters no conjunto de Pareto e sugerir qual é o cluster solução.

4.4 Estatística Scan de Adjacência Modificada

Nesta extensão discutimos como um conhecimento a priori de informações ambientais pode ser incorporado na estrutura de vizinhança de uma mapa e ser útil para o estudo da distribuição espacial de casos de alguma doença específica. A estrutura de vizinhança mais simples assume que duas regiões em um mapa são conectadas apenas quando elas têm uma parte de suas fronteiras geográficas em comum. Nesta seção, estendemos este conceito para que o grau de conectividade entre duas regiões seja aumentado ou diminuído. As regiões vizinhas que têm características ambientais em comum são definidas como fortemente conectadas.

A Figura 4.5 mostra como as características ambientais tais como a presença de um rio ou uma montanha podem alterar a adjacência entre duas regiões de um mapa, indicado aqui por uma estrutura de grafo. A Figura 4.5A apresenta a adjacência usual, em que fronteiras geográficas comuns definem o grafo. Neste caso, apenas os vizinhos de primeira ordem são permitidos. Na Figura 4.5B, o rio induz adjacência modificada, no sentido de que regiões próximas ao rio devem ter vizinhos de segunda ordem considerados em sua estrutura de adjacência. Isto significa que duas regiões geograficamente desconectadas próximas ao rio devem ser consideradas vizinhas no grafo reforçado. Na Figura 4.5C, por exemplo, as quatro regiões de alta incidência (hachuradas) não constituem uma zona no mapa da Figura 4.5A (inicialmente desconectadas), mas constituem uma zona conectada no mapa da 4.5B. A Figura 4.5D apresenta uma situação em que os vizinhos devem ser enfraquecidos, pois as ligações das regiões situadas nos lados opostos da montanha são removidas da estrutura do grafo.

Adicionalmente, permitimos amplitudes maiores de vizinhos para as regiões mais populosas, refletindo o fato de que o fluxo de trabalho (Duczmal and Buckeridge, 2006), que é o movimento de indivíduos entre sua residência e o local de trabalho, está relacionado ao tamanho da população. Isto faz uma grande cidade estender sua influência não apenas às regiões mais próximas, mas também entre as regiões vizinhas de segunda ordem ou até mesmo de terceira ordem, ou seja, sua influência ultrapassa os limites da adjacência geográfica usual. A Figura 4.6A mostra a estrutura de vizinhança usual. Na Figura 4.6B a adjacência é modi-

ficada para incluir os vizinhos de segunda ordem. Na Figura 4.6C, os vizinhos de primeira ordem da grande cidade formam um clique, ou seja, todos os nós são conectados. O clique da Figura 4.6D é formado por todos os vizinhos de primeira e segunda ordem. Por exemplo, as regiões I, II e III constituem uma zona conectada no mapa da Figura 4.6D, mas não em outros mapas.

Uma consequência de se anexar uma variável de grau de conectividade entre as regiões é aumentar o sinal para a detecção de clusters potenciais em uma área que compartilha das mesmas características ambientais. Ao mesmo tempo, esta estratégia reduz o sinal para possíveis clusters espúrios que possuem fatores dissimilares. Assim, esperamos um aumento no poder de detecção para clusters que são apropriadamente relacionados às causas ambientais. Uma coleção de cenários diferentes deve ser analisada, em que podemos variar o grau de conectividade espacial entre as regiões, e alterar a seleção das características ambientais que são inicialmente estudadas. Esta estratégia permite testar, simultaneamente, um grande número de diferentes hipóteses sobre as localizações mais plausíveis do cluster. Esperamos que a maioria dos cenários prove que eles são inadequados indicando clusters que não são estatisticamente significativos, e alguns poucos (ou mesmo nenhum) cenários produzam clusters altamente significativos.

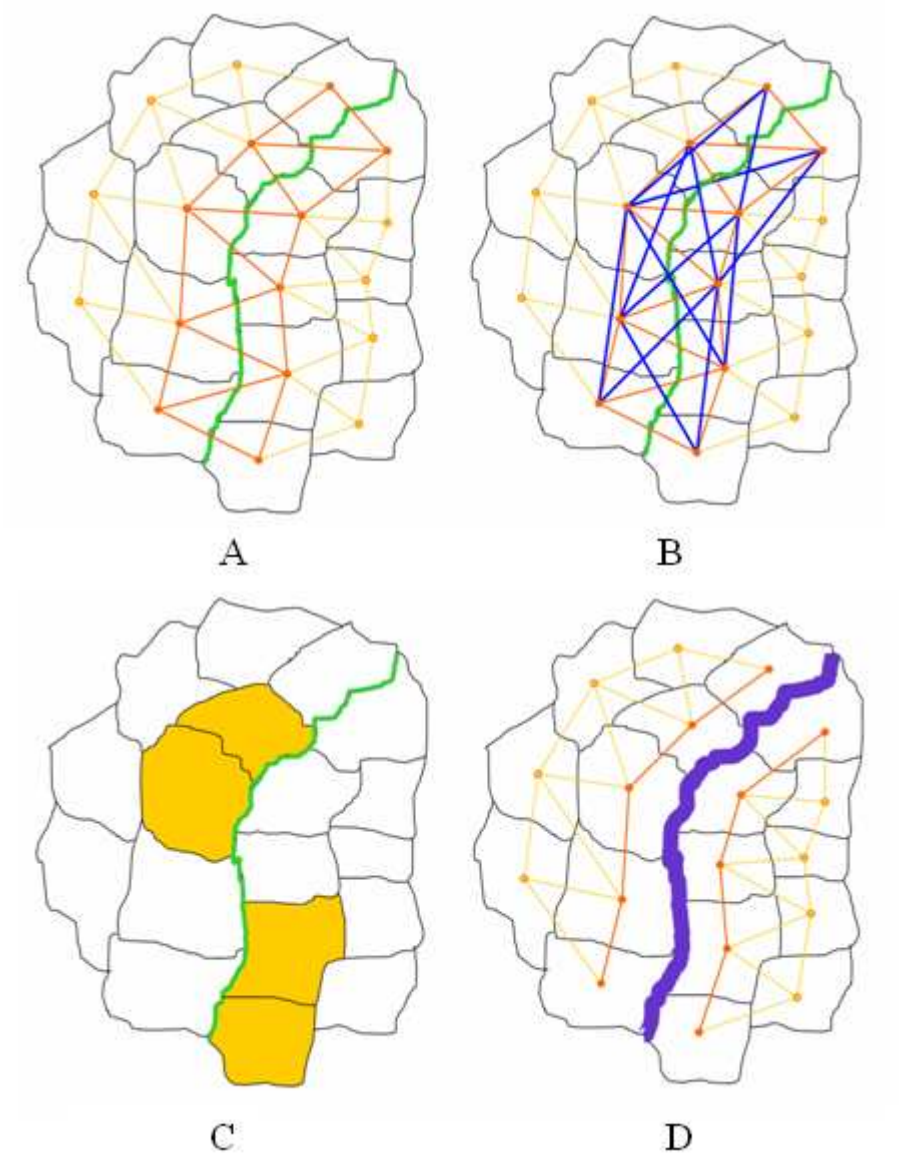


Figura 4.5: Mudanças na adjacência com base em características ambientais diversas.

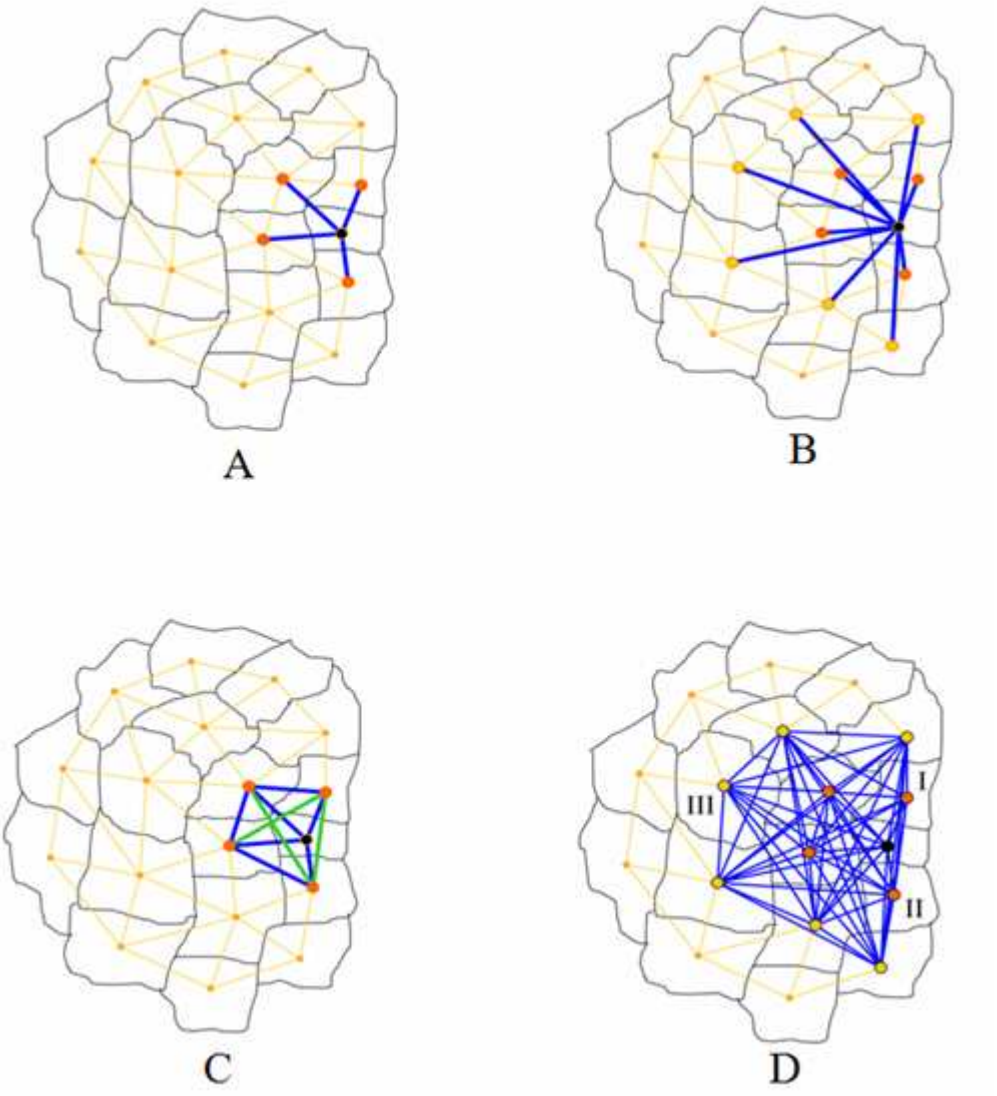


Figura 4.6: Reforço para considerar o efeito de uma área altamente populosa entre os vizinhos.

O algoritmo genético multiobjetivo é executado sequencialmente para vários conjuntos de parâmetros ou cenários e por consequência obtemos um número grande de soluções de clusters, que são comparadas e julgadas em termos de suas significâncias. Um conjunto de parâmetros controla a força de vizinhança entre as regiões, que são baseados nas similaridades ambientais dessas regiões. O objetivo é testar quais condições ambientais específicas induzem clusters mais significativos, e com isso permitir ao profissional da área aceitar ou rejeitar diferentes hipóteses a respeito da relevância dos fatores geográficos.

O artigo de Smith et al. (2002) ilustra uma situação que se adequa ao nosso objetivo do parágrafo anterior. A Figura 4.7 indica o número de dias desde a primeira ocorrência de raiva no mapa do estado de Connecticut (EUA) em 1991. As regiões mais escuras representam regiões em que a primeira ocorrência de casos de raiva ocorreu há mais tempo, enquanto que as mais claras são regiões em que a primeira ocorrência de raiva foi mais recente. A Figura 4.7 sugere que os rios que cortam essa região atuam como barreiras para a propagação da raiva, criando um lapso de tempo maior para a disseminação da raiva entre regiões situadas em margens opostas a um rio.

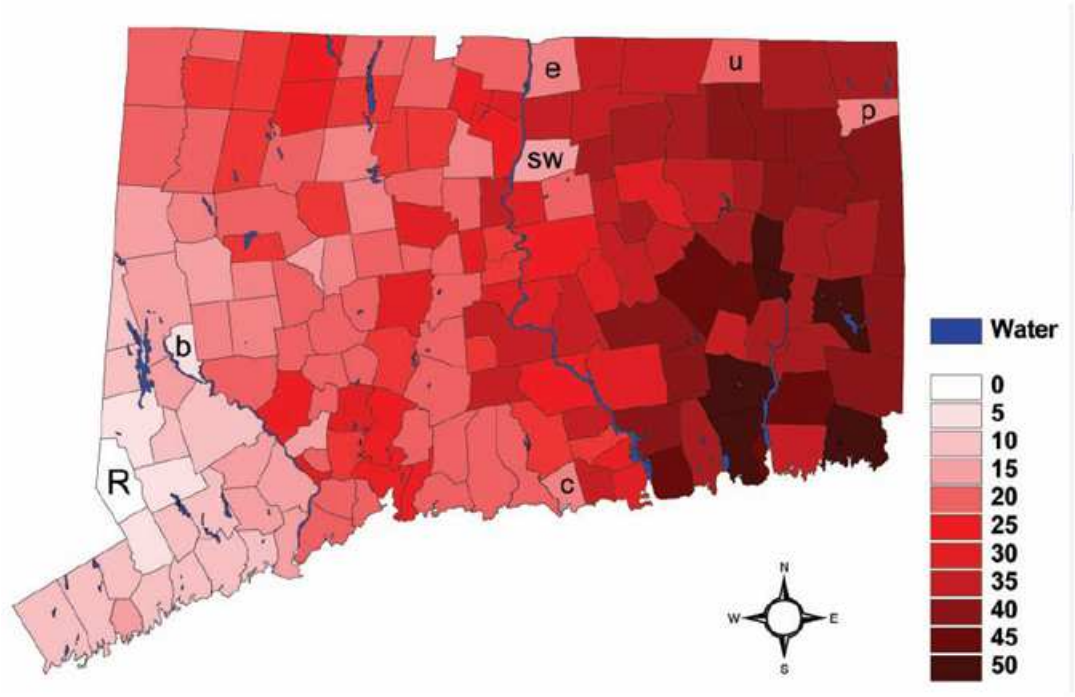
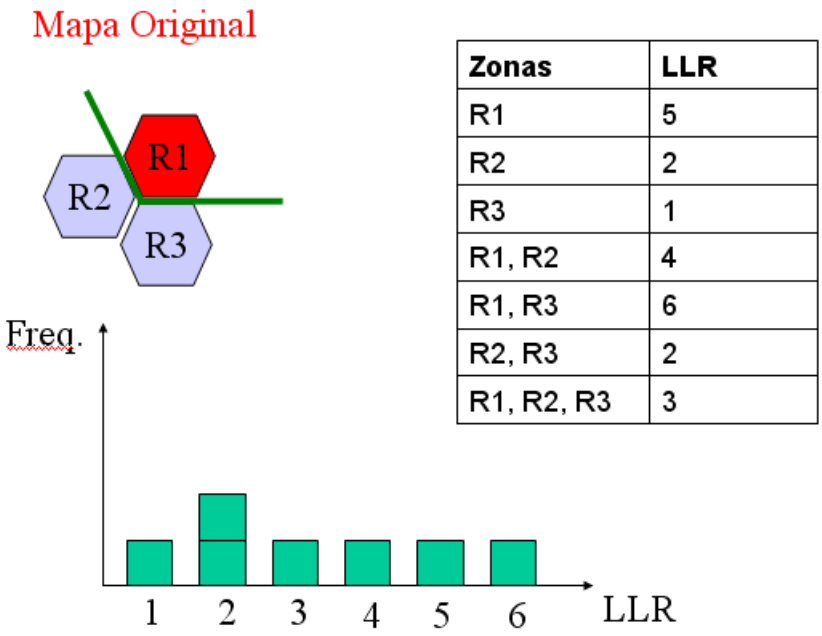
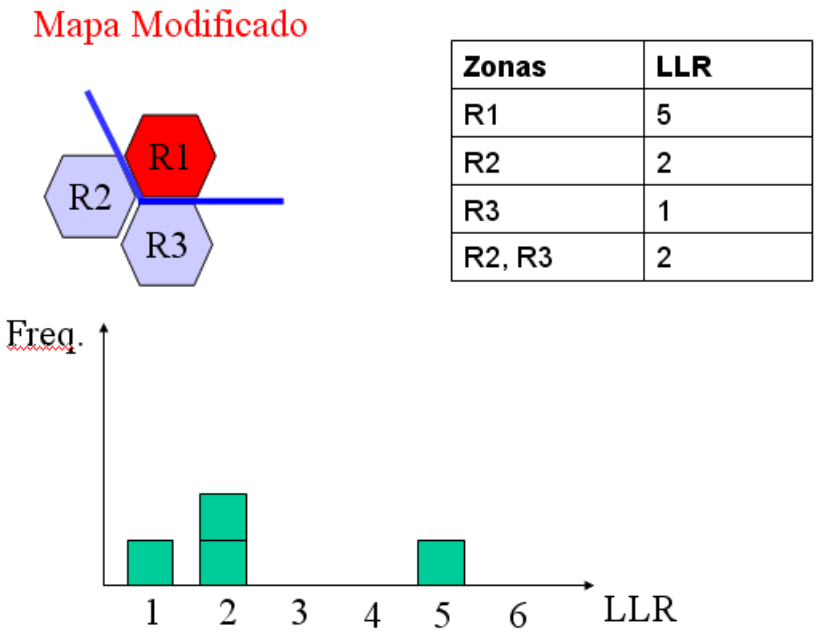


Figura 4.7: Dias desde a primeira ocorrência de raiva no estado de Connecticut, EUA, 1991.



Zonas	LLR
R1	5
R2	2
R3	1
R1, R2	4
R1, R3	6
R2, R3	2
R1, R2, R3	3

(a)



Zonas	LLR
R1	5
R2	2
R3	1
R2, R3	2

(b)

Figura 4.8: Avaliando a LLR das zonas após a estrutura de adjacência ter sido alterada.

A Figura 4.8 apresenta um exemplo em que se tem 3 regiões vizinhas a um rio num mapa hipotético. Analisamos duas situações, com e sem modificação da estrutura de adjacência. Queremos testar a hipótese de que o rio atua como uma barreira para a propagação da raiva. No primeiro caso (Figura 4.8(a)), as zonas são avaliadas conforme o mapa original e percebemos que o cluster mais provável é aquele formado pelas regiões 1 e 3 cuja LLR foi igual a 6. Suponha agora que o rio seja um fator que se correlacione com o grau de incidência da patologia analisada (Figura 4.8(b)). Observa-se que ao modificarmos a adjacência das regiões levando em conta o fato delas terem fronteira comum com o rio, o conjunto das possíveis zonas é alterado. Agora, o cluster mais provável é aquele constituído apenas pela região 1 que tem uma LLR igual a 5. No entanto, várias zonas foram eliminadas no segundo mapa (Figura 4.8(b)), por conterem regiões situadas em margens opostas do rio. Mesmo que a LLR tenha diminuído um pouco de 6 para 5, ao reforçar o mapa percebemos que houve um destaque para a LLR igual a 5. Isso não acontece no mapa original (Figura 4.8(a)) em que tivemos candidatos com LLR igual a 4, 5 e 6. Portanto, temos razões para acreditar que a significância do cluster pode ser alterada pelo enfraquecimento da adjacência. Em outras palavras, o cluster mais provável do segundo mapa é nitidamente distinto dos demais candidatos a cluster. Isso não acontece no primeiro mapa (Figura 4.8(a)).

Isso mostra que a técnica descrita neste capítulo pode ser usada para testar a hipótese de que o rio atua como uma barreira para a propagação de doenças como a raiva.

As seções 4.5 e 4.6 apresentam os resultados obtidos para o desempenho do método em situações artificias e também numa aplicação real considerando cenários distintos para o reforço.

4.5 Avaliação Numérica

Com base no mapa da Amazônia Brasileira (excluíndo o estado do Tocantins), composto por 56 microregiões e abrangendo 12.297.604 habitantes, criamos um cluster artificial em que os casos se referem a incidência de uma certa patologia. Avaliamos o poder de detecção

do algoritmo em diferentes cenários (a)-(g) descritos na Tabela 4.1. O cenário básico original é formado por um cluster artificial com cinco regiões mostrado na Figura 4.10. Os cenários de (b) a (g), apresentados nas Figuras 4.11 a 4.16, são obtidos reforçando o grafo nas regiões marcadas de cinza nos respectivos mapas, os quais têm a estrutura de vizinhança alterada por adicionar os vizinhos de segunda ordem.

Tabela 4.1: *Descrição da vizinhança modificada para os cenários simulados.*

Cenário	Descrição
(a)	Grafo básico (sem reforço, apenas a vizinhança de primeira ordem).
(b)	Todas as regiões do cluster básico (desconexo) são modificadas nas regiões: 19, 22, 23, 32, 45.
(c)	Reforço completamente fora do cluster nas regiões: 1, 2, 3, 4, 5, 6, 7.
(d)	Reforço completamente fora do cluster nas regiões: 33, 34, 41, 46, 56.
(e)	Reforço parcialmente fora do cluster nas regiões: 32, 33, 34, 45, 56.
(f)	Reforço parcialmente fora do cluster nas regiões: 31, 32, 44, 45, 48.
(g)	Todas as regiões do cluster básico e mais três regiões conexas são modificadas nas regiões: 19, 20, 22, 23, 26, 32, 44, 45.

A Figura 4.9 apresenta a numeração utilizada na construções dos clusters artificiais conforme a Tabela 4.1.

Para cada cenário (a)-(g), executamos o algoritmo genético multiobjetivo para 10.000 réplicas Monte Carlo da distribuição de 596 casos sob a hipótese nula e calculamos as isolinhas dos valores críticos para um nível de significância de 0,05, mostrados abaixo de cada grafo nas Figuras 4.10 a 4.16. A hipótese nula é diferente em cada cenário, pois a estrutura de adjacência é diferente devido o reforço utilizado em cada situação. Da mesma forma, para todos os cenários aplicamos o algoritmo sob 10.000 réplicas Monte Carlo da distribuição de 596 casos sob a hipótese alternativa de que existe um cluster constituído por cinco regiões do mapa. A proporção de pontos à direita da isolinha é uma estimativa do poder médio do algoritmo para esta hipótese alternativa. O poder de detecção foi calculado pela proporção de pontos que superaram a isolinha do valor crítico 0,05 para cada faixa de compacidade (K). Os resultados são apresentados na Tabela 4.2. Cada coluna corresponde a um cenário e contém o poder e a quantidade de pontos (entre parênteses) que ocorreram em cada faixa de compacidade. A compacidade do cluster verdadeiro é aproximadamente igual a 0,11. Esta compacidade é muito baixa devido à não conectividade do cluster, e a linha da Tabela 4.2

referente a esta faixa de compacidade é destacada em negrito.

Tabela 4.2: *Comparações do poder para os cenários dos grafos reforçados (a)-(g) por faixa de compacidade K.*

Faixa de K	(a)	(b)	(c)	(d)	(e)	(f)	(g)
0,00-0,05	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)	- (0)
0,05-0,10	- (0)	0,85 (137)	0,67 (6)	0,67 (3)	0,50 (16)	1.00 (5)	0,82 (147)
0,10-0,15	**0,81 (98)**	**0,91 (11343)**	**0,87 (4305)**	**0,81 (542)**	**0,82 (1554)**	**0,83 (1840)**	**0,90 (12221)**
0,15-0,20	0,86 (1745)	0,92 (39222)	0,90 (15218)	0,87 (6514)	0,91 (14184)	0,90 (17010)	0,91 (41377)
0,20-0,25	0,91 (7275)	0,92 (56534)	0,91 (19680)	0,91 (14919)	0,93 (24544)	0,92 (28656)	0,92 (60537)
0,25-0,30	0,92 (12481)	0,89 (46785)	0,92 (19985)	0,91 (16225)	0,92 (23890)	0,93 (29167)	0,90 (51238)
0,30-0,35	0,92 (11544)	0,80 (30224)	0,92 (18217)	0,91 (18855)	0,91 (26367)	0,92 (32581)	0,75 (25380)
0,35-0,40	0,92 (13921)	0,70 (20013)	0,89 (17140)	0,91 (23633)	0,90 (28043)	0,91 (35125)	0,59 (15973)
0,40-0,45	0,92 (15357)	0,66 (14238)	0,87 (16013)	0,91 (22093)	0,89 (30023)	0,83 (22627)	0,59 (10356)
0,45-0,50	0,92 (14703)	0,63 (8843)	0,84 (10117)	0,92 (19381)	0,88 (26827)	0,67 (8277)	0,47 (2168)
0,50-0,55	0,94 (11438)	0,23 (1790)	0,84 (6614)	0,89 (11141)	0,76 (7157)	0,35 (3232)	0,05 (588)
0,55-0,60	0,94 (10863)	0,07 (510)	0,81 (9955)	0,82 (6229)	0,39 (2720)	0,29 (2651)	0,02 (180)
0,60-0,65	0,91 (8715)	0,01 (287)	0,43 (2145)	0,72 (5459)	0,28 (1887)	0,05 (1058)	0,04 (464)
0,65-0,70	0,85 (4782)	0,01 (293)	0,48 (1830)	0,49 (1124)	0,34 (1138)	0,12 (893)	0,04 (399)
0,70-0,75	0,77 (2379)	0,03 (213)	0,82 (1449)	0,66 (883)	0,01 (144)	0,06 (249)	0,07 (177)
0,75-0,80	0,80 (5374)	0,02 (411)	0,64 (1903)	0,27 (955)	0,16 (596)	0,87 (725)	0,88 (418)
0,80-0,85	0,55 (2997)	0,06 (397)	0,24 (1135)	0,15 (851)	0,15 (775)	0,02 (139)	- (0)
0,85-0,90	0,02 (350)	0,01 (225)	0,01 (456)	0,01 (253)	0,04 (54)	0,02 (330)	0,75 (253)
0,90-0,95	0,03 (251)	0,02 (154)	0,02 (350)	0,02 (238)	0,02 (45)	0,05 (331)	0,02 (167)
0,95-1.00	0,10 (6352)	0,04 (2871)	0,03 (5302)	0,04 (4739)	0,04 (4244)	0,04 (3856)	0,04 (2970)
Médio	0,8605	0,8285	0,8345	0,8549	0,8571	0,8457	0,8329

Da Tabela 4.2 podemos ver que para os cenários (b) e (g), nos quais todas as regiões do cluster verdadeiro foram modificadas, o poder aumenta nas faixas próximas àquela que contém a compacidade do cluster, do cenário básico original. Percebe-se ainda que para esses cenários, o poder decresce mais rapidamente com o aumento das faixas de compacidade. Por outro lado, para os cenários (c) e (d), em que o reforço aconteceu completamente fora do cluster verdadeiro, a concentração de pontos para as faixas de menor compacidade é muito inferior e o decréscimo do poder é mais gradual, similarmente à situação do cenário básico. Por fim, os cenários (e) e (f), em que o reforço aconteceu de forma parcial, a concentração de pontos foi apenas moderada e o poder decresceu.

As Figuras 4.10 a 4.16 apresentam o mapa do cluster artificial, o grafo associado ao reforço atribuído ao mapa, e os pontos do conjunto Pareto com as isolinhas de valor p 0,05 para cada um dos sete cenários avaliados. Juntamente com esse conjunto de pontos são mostradas as isolinhas de valor p 0,05 contra os pontos obtidos por simulação Monte Carlo sob a hipótese alternativa: pontos pretos indicam pontos do conjunto Pareto além da isolinha,

e os pontos em cinza claro indicam o restante dos pontos do conjunto Pareto.

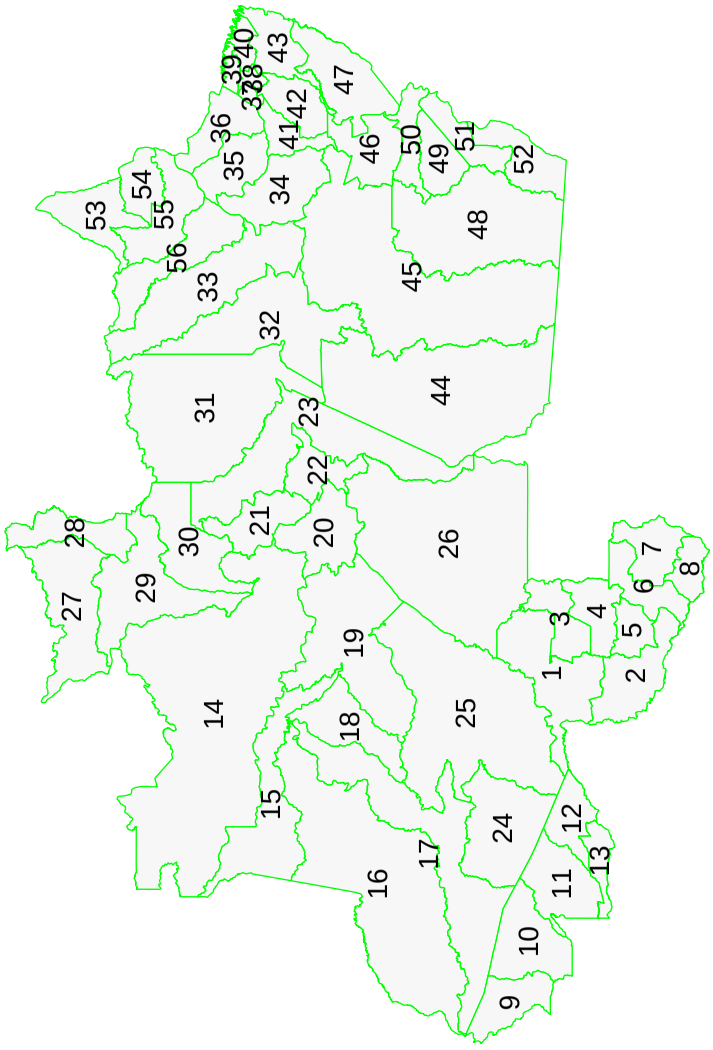


Figura 4.9: Codificação atribuída às microregiões do Norte, IBGE, 2000.

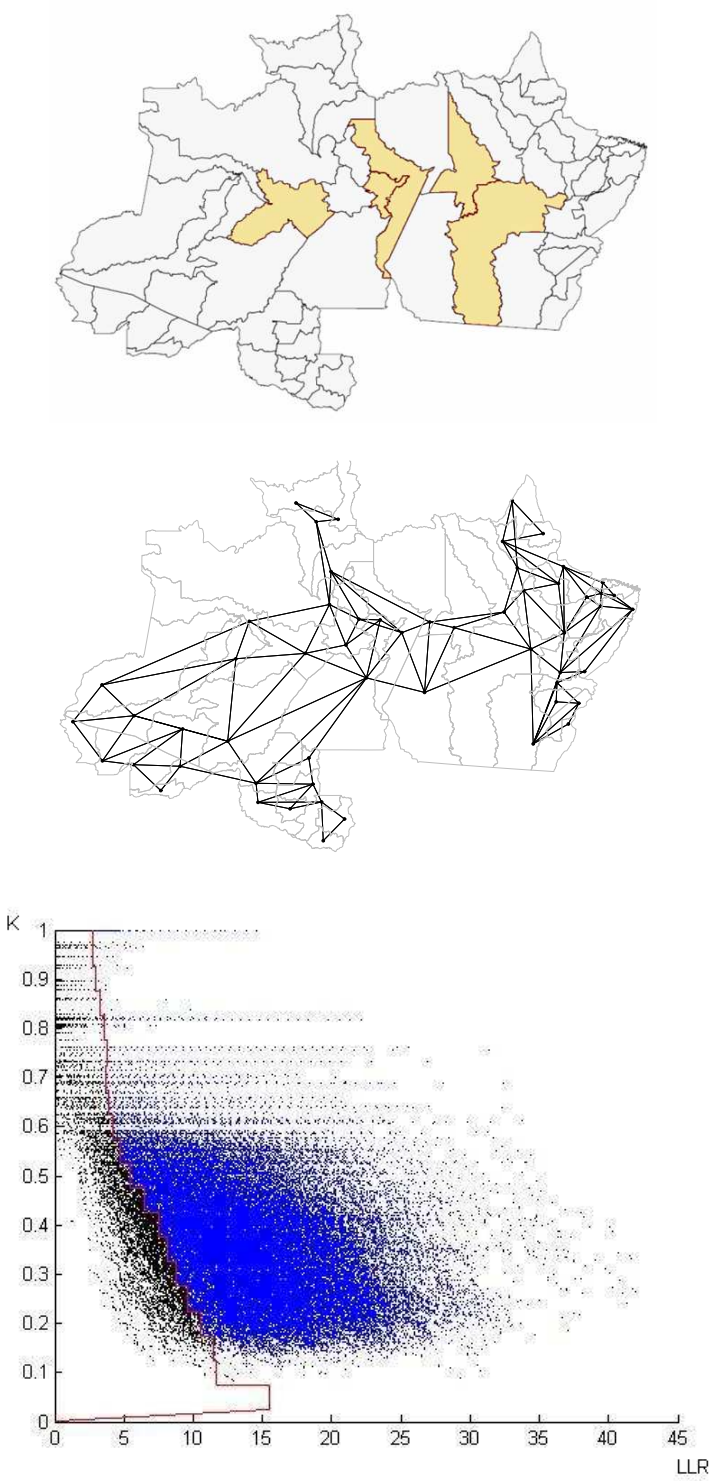


Figura 4.10: (a) Grafo básico (sem reforço, apenas a vizinhança de primeira ordem).

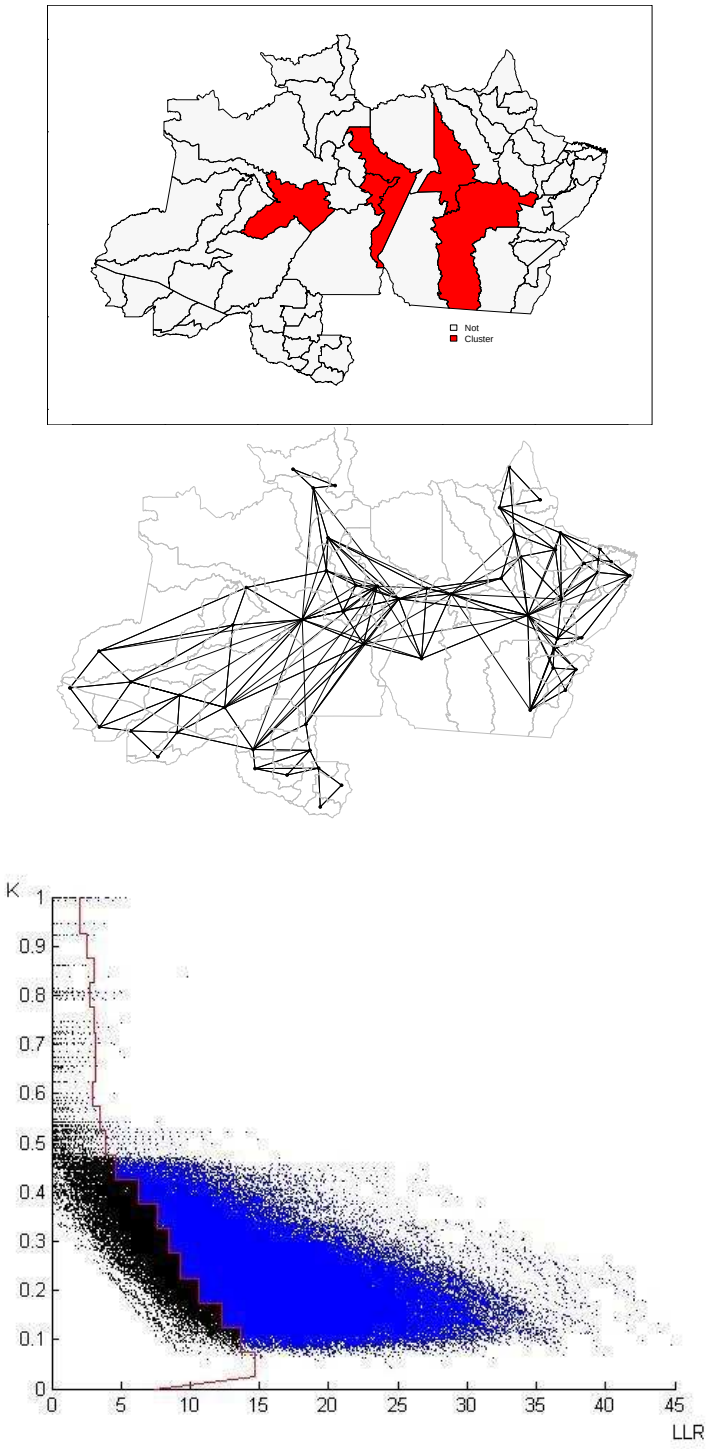


Figura 4.11: (b) Todas as regiões do cluster básico (desconexo) são moficadas nas regiões: 19, 22, 23, 32, 45.

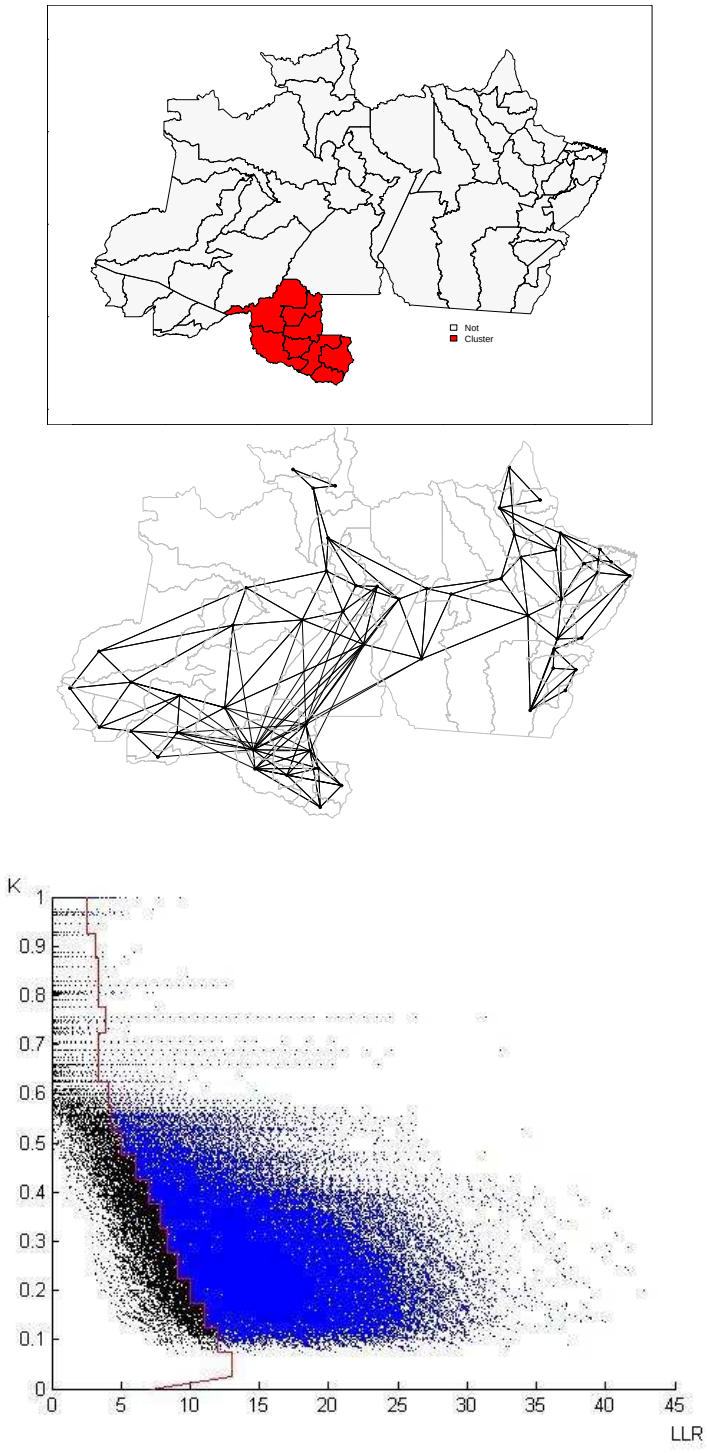


Figura 4.12: (c) Reforço completamente fora do cluster nas regiões: 1, 2, 3, 4, 5, 6, 7.

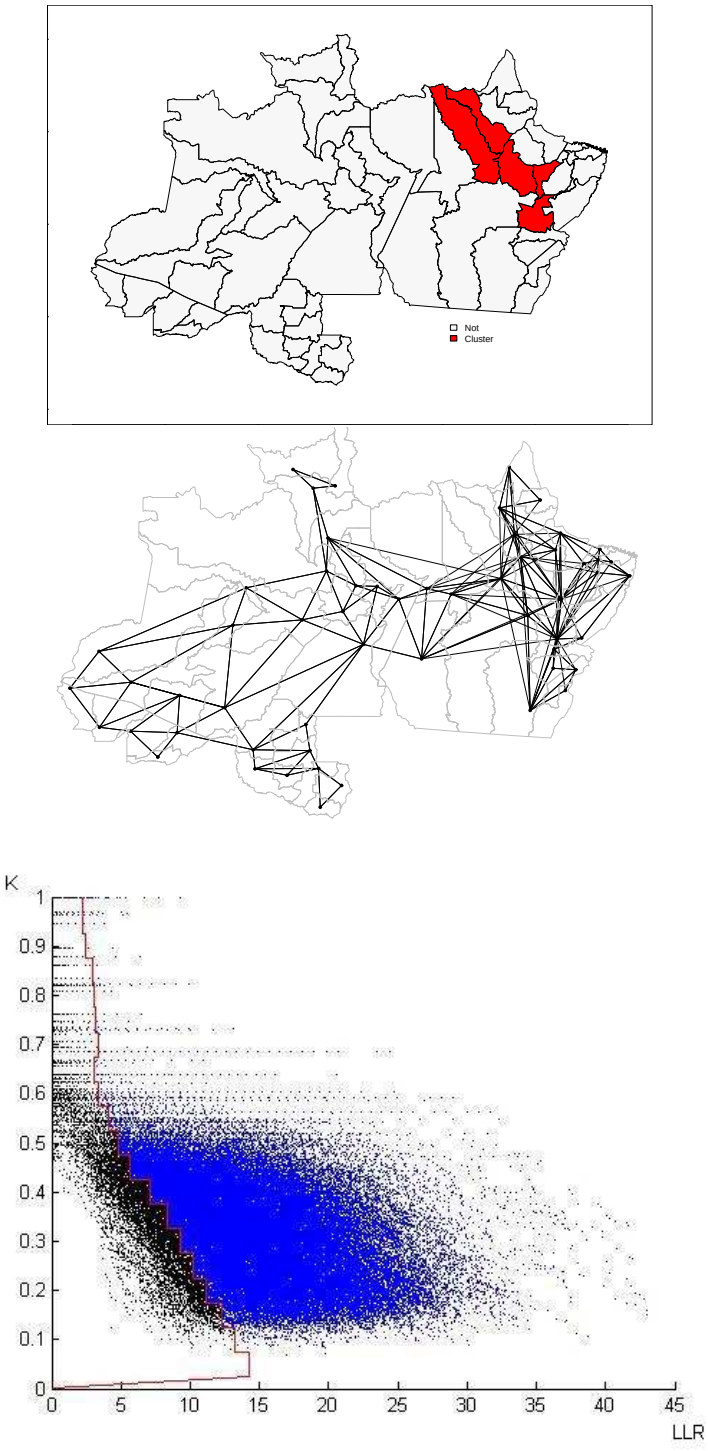


Figura 4.13: (d) Reforço completamente fora do cluster nas regiões: 33, 34, 41, 46, 56.

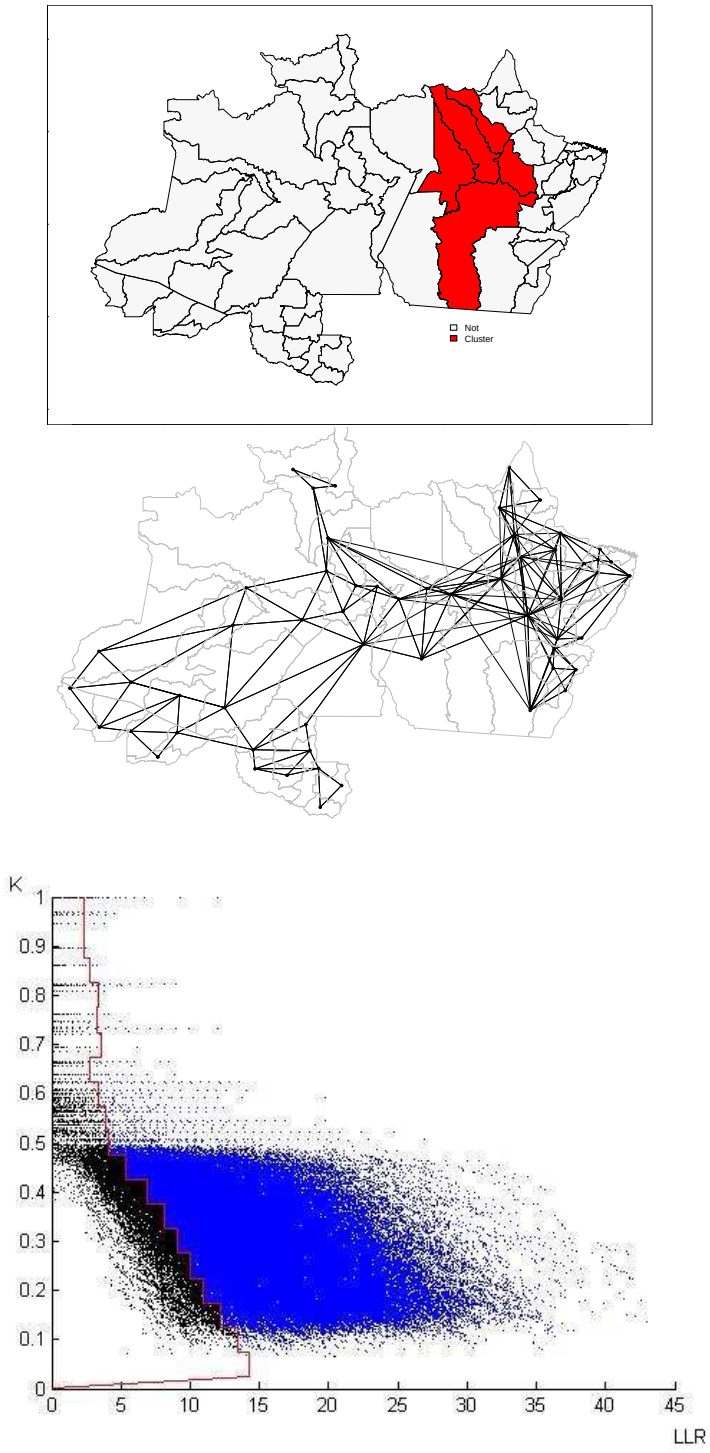


Figura 4.14: (e) Reforço parcialmente fora do cluster nas regiões: 32, 33, 34, 45, 56.

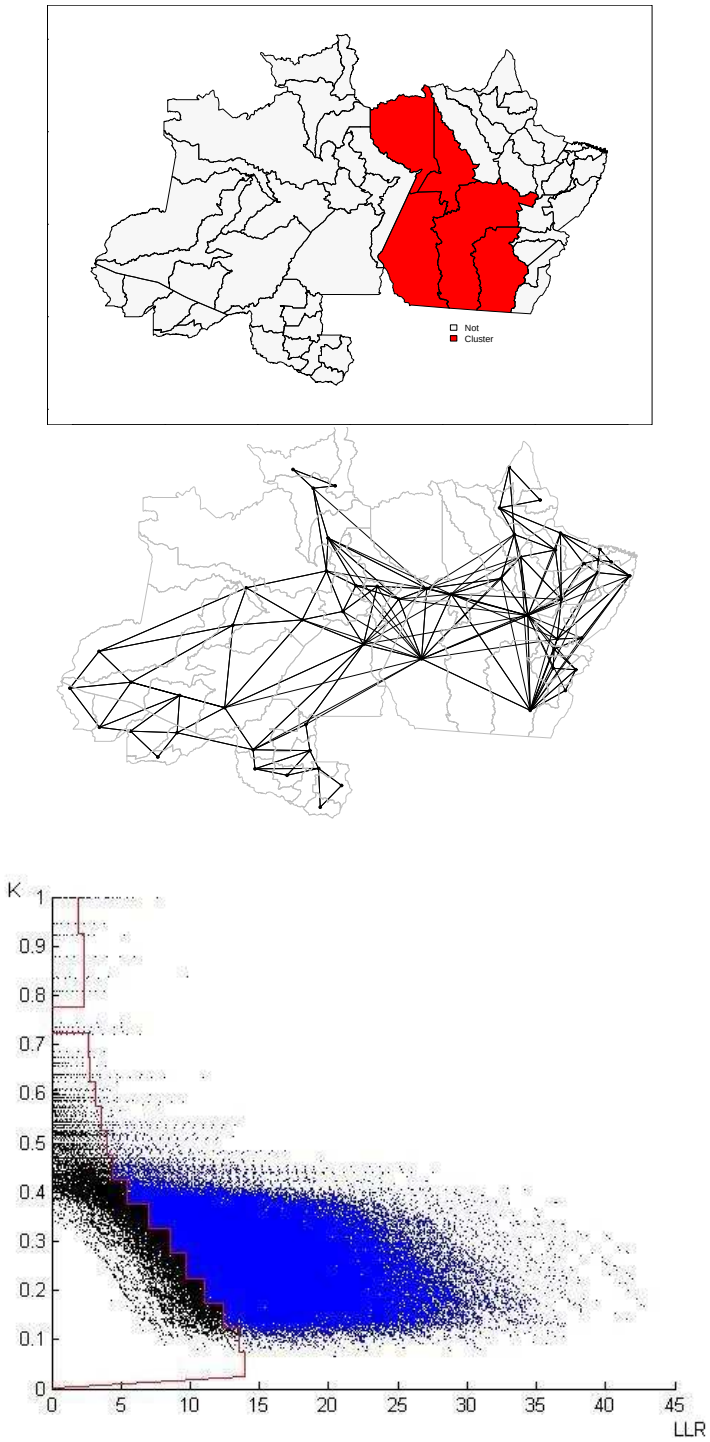


Figura 4.15: (f) Reforço parcialmente fora do cluster nas regiões: 31, 32, 44, 45, 48.

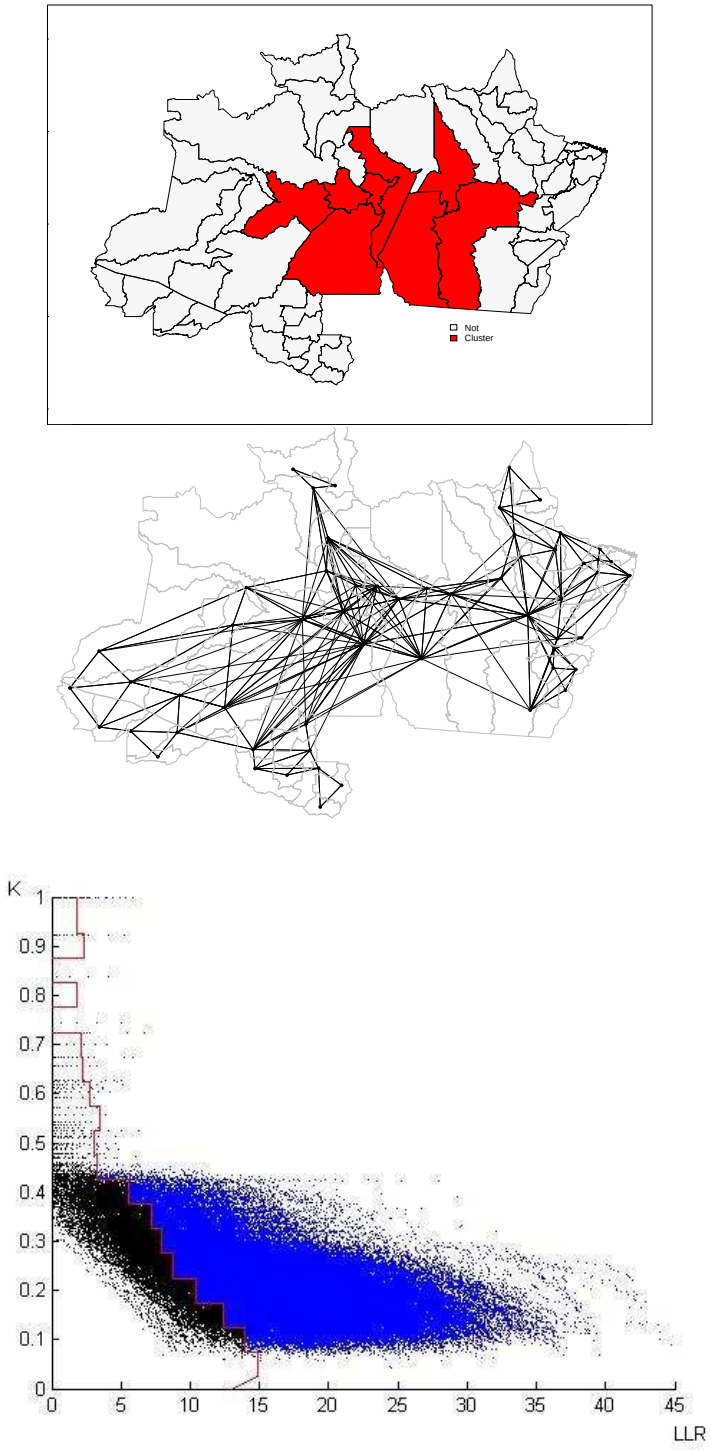


Figura 4.16: (g) Todas as regiões do cluster básico, mais três regiões conexas, são modificadas nas regiões: 19, 20, 22, 23, 26, 32, 44, 45.

4.6 Aplicação: Óbitos de Malária na Amazônia Brasileira

Analisamos o mapa de mortalidade causada por malária na Amazônia Brasileira (Figura 4.17). Um total de 596 óbitos foram registrados durante o período de 1998 a 2002, para uma população de 12.297.604 habitantes (www.datasus.gov.br). O nosso mapa da região norte desconsidera o estado do Tocantins pois este possui muitos municípios pequenos. O mapa considerado contém 310 municípios. Nesta proposta estamos considerando os clusters relacionados com condições ambientais tais como precipitação de chuva. Na seção 4.6.1 aplicamos o teste de Mantel, um método clássico para avaliar correlação entre dados geográficos, biológicos e ambientais, a fim de comparar os resultados obtidos com nossa nova metodologia. A seção 4.6.2 descreve os resultados do método dos grafos reforçados.

4.6.1 Teste de Mantel

O teste de Mantel (Mantel, 1967) é uma abordagem clássica muito usada em análise espacial em Ecologia para explicar a distribuição de espécies em termos de variáveis ambientais. Trata-se de uma correlação entre matrizes de distâncias e estas matrizes podem ser definidas de várias formas, e assim o teste assume uma variedade de casos especiais. O teste parcial de Mantel em três matrizes de distância se adequa aos interesses deste trabalho. O teste parcial de Mantel é uma estatística de teste para avaliar a correlação linear entre a matriz de dados biológicos A e a matriz de dados ambientais B quando uma matriz de distância C é fixada. No nosso estudo desejamos estudar a relação entre as matrizes dos dados ambientais (precipação de chuva) e biológicos (taxa de malária), fixando o efeito das distâncias geográficas (Legendre and Fortin, 1989). A questão que o teste parcial de Mantel responde é a seguinte: amostras que são biologicamente similares também são ambientalmente similares quando se leva em conta a estrutura espacial do mapa?

Para cada mês, as entradas das matrizes A, B e C simétricas e de ordem $m \times m$ são definidas como:

$$\begin{cases} A: & A_{ij} = 0 \text{ se as regiões i e j são ambas chuvosas ou ambas secas em cada mês e} \\ & A_{ij} = 1 \text{ caso contrário.} \\ B: & B_{ij} = \mid r_i - r_j \mid \text{ em que } r_i \text{ e } r_j \text{ são as taxas de malária para as regiões i e j,} \\ & \text{respectivamente.} \\ C: & C_{ij} = dist(i,j) \text{ é a distância geográfica entre os centróides das regiões i e j.} \end{cases}$$

O coeficiente parcial de Mantel é definido como,

$$r_M(AB,C) = \frac{r_M(AB) - r_M(AC)r_M(BC)}{\sqrt{1-r_M(AC)^2}\sqrt{1-r_M(BC)^2}}$$

em que a estatística simples de Mantel $r_M(XY) = \frac{1}{n-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m}\frac{X_{ij}-\bar{X}}{S_X}\frac{Y_{ij}-\bar{Y}}{S_Y}$ é a correlação de Pearson entre os $n = m(m-1)/2$ elementos correspondentes das partes superiores dessas matrizes triangulares. $\bar{X}$, $\bar{Y}$ e S_X, S_Y são respectivamente a média e o desvio padrão sob as n entradas acima da diagonal das matrizes X e Y.

Como os elementos de uma matriz de distância não são independentes, a significância da estatística do coeficiente parcial de Mantel é estimada através de um teste de permutação, em que se permutam as linhas e as colunas da matriz A, e novamente se calcula o coeficiente do teste parcial de Mantel. Este procedimento é repetido muitas vezes e a distribuição dos valores para a estatística permutada é comparada com o valor da estatística para os dados originais.

O pacote **Ecodist** do software R (Goslee and Urban, 2007) foi usado para calcular os coeficientes parciais de Mantel e seus respectivos valores p (1000 réplicas Monte Carlo) para cada um dos doze cenários mensais de chuva na Amazônia Brasileira, e os resultados são apresentados na Tabela 4.3. Apenas para os meses de abril e novembro (em negrito) encontramos correlações significativas e o coeficiente parcial de Mantel foi elevado apenas para o cenário de chuva referente a novembro.

Tabela 4.3: *Correlação entre precipitação de chuva e taxa de malária.*

Cenário de chuva	coeficiente parcial de Mantel	valor p
Janeiro	-0,040	0,695
Fevereiro	0,044	0,591
Março	-0,007	0,775
Abril	**-0,047**	**0,021**
Maio	-0,020	0,753
Junho	0,056	0,605
Julho	0,089	0,158
Agosto	-0,007	0,577
Setembro	-0,009	0,406
Outubro	0,013	0,881
Novembro	**0,363**	**0,001**
Dezembro	-0,061	0,339

4.6.2 Teste dos Grafos Reforçados

Os resultados do teste de Mantel foram comparados com os resultados oriundos do nosso método dos grafos reforçados. Construímos estudos de casos em que as modificações ambientais foram introduzidas apenas por alterar a estrutura de vizinhança do mapa. Deste modo, adicionamos ou subtraímos informações de vizinhança no mapa a fim de testar o efeito do fator ambiental na detecção do cluster dos dados biológicos referentes a óbitos causados pela malária, cuja taxa de incidência é apresentada na Figura 4.17. A adjacência geográfica usual é mostrada na Figura 4.18. Com base nos registros meteorológicos mensais dos dias chuvosos para o ano 2000 (www.inpe.br), Figura 4.19, construímos doze cenários, selecionando os municípios mais úmidos para cada mês (Figura 4.20). Em cada cenário, reforçamos o grafo correspondente. A escala na Figura 4.19 mede o número de dias chuvosos durante o respectivo mês. Para cada cenário, as regiões que tinham mais que um certo número de dias chuvosos foram selecionadas e marcadas de cor cinza no mapa (Figura 4.20). A vizinhança de primeira ordem dessas regiões marcadas de cinza é estendida à vizinhança de segunda ordem. O limiar para a seleção não é o mesmo para todos os meses, pois quando tínhamos um mês mais seco, foi necessário diminuir o limiar (adotou-se 9 dias chuvosos) e captar a di-

nâmica das precipitações ao longo das áreas no mapa, evitando assim mapas vazios. O outro limiar utilizado foi 18 dias em meses que aparentemente o volume de precipitação era grande. Os grafos reforçados por esta estratégia são mostrados na Figura 4.21. Usando os doze cenários, aplicamos o algoritmo multiobjetivo para detectar cluster de diferentes formatos, desde aqueles mais circulares até os mais irregulares possíveis, conforme seja a disposição dessas soluções dentro do conjunto Pareto. Esses clusters são plotados juntamente com suas curvas de valores p na Figura 4.22. Os números situados abaixo de cada curva referem-se aos valores p, dados por 10^{-3}, 10^{-6}, $\cdots$, 10^{-45}. Obviamente, os clusters (representados por pontos) mais significativos estão situados mais à direita destas curvas.

Os dados originais observados e os gerados de forma aleatória sob a hipótese nula através das réplicas Monte Carlo foram comparados usando as mesmas características de reforço nos respectivos grafos. Isto pode ser observado nas Figuras 4.22 e 4.23, em que as isolinhas de valor p construídas sob a hipótese nula são diferentes para cada cenário. A Figura 4.23 compara o grafo do cenário básico, correspondente ao mapa original sem reforço, com dois outros cenários, julho e novembro. Os clusters E e F (Figura 4.24) deixaram de pertencer ao conjunto Pareto no cenário de novembro. Eles foram dominados por outros clusters (não explicitamente referidos na Figura 4.24). A Figura 4.24 apresenta alguns clusters pertencentes aos conjuntos de Pareto da Figura 4.23. A solução A é o cluster primário e aparece no conjunto solução de quase todos os cenários. Os outros dois clusters B e C são secundários e foram incluídos para comparação. O cluster C é menos significativo e se compara com os demais clusters B-I. Os clusters A, B, C e D são facilmente distinguidos ao se olhar o mapa da Figura 4.17. Os clusters E, F e G apareceram nos cenários básico e de julho. Os clusters H e I apareceram no cenário de novembro. Observe que os valores p para os clusters A, E, F e G crescem para o cenário de julho, devido ao reforço ter sido colocado nas áreas sem clusters significativos. Isso aumenta a probabilidade de ocorrência de clusters espúrios.

Os clusters legítimos tais como E, F e G devem competir com mais ruídos no mapa, e consequentemente, a sua significância diminui. No entanto, quando o reforço do grafo ocorre próximo ao cluster significativo A no cenário de novembro, o efeito oposto acontece. Um

conjunto completo de novos clusters com LLR muito alto (valor p muito baixo) aparece, entre eles os clusters H e I, mostrados na Figura 4.24. Observe que esses clusters estão desconectados devido ao reforço do grafo. Uma análise similar poderia ser conduzida para os demais cenários (sem contar julho), comparando-os com o cenário de novembro, assim mostrando que o cenário de novembro é essencialmente diferente dos meses restantes.

Este resultado é coerente com o teste de Mantel acima, que diz que a matriz de chuva de novembro é a única matriz correlacionada significativamente com os dados da taxa de malária e tem o coeficiente essencialmente não nulo. Além deste resultado, o teste dos grafos reforçados apresenta o cluster mais provável relativo ao cenário de novembro.

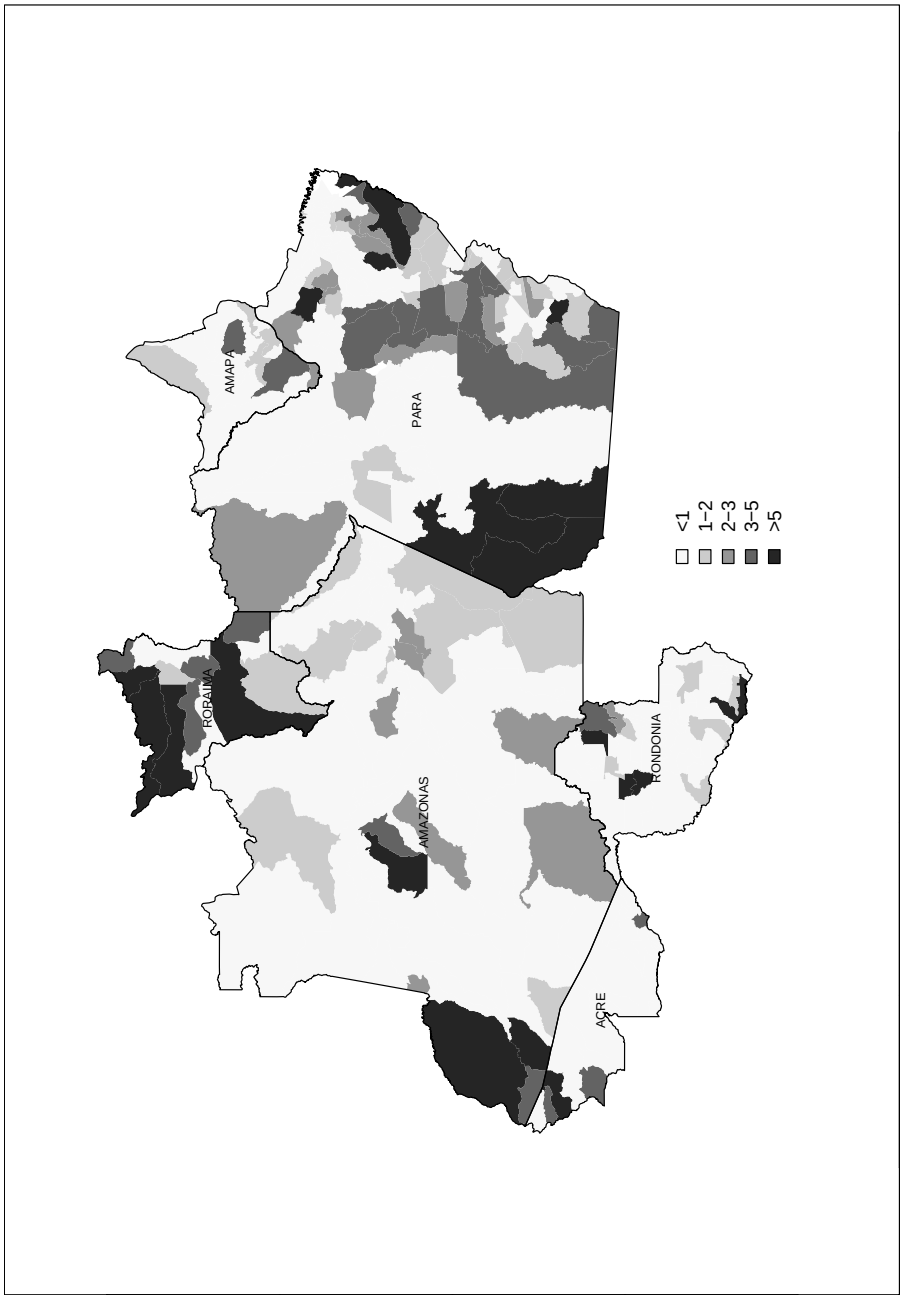


Figura 4.17: Taxa de mortalidade (por 100 mil habitantes) causadas por Malária na Região Norte entre 1998-2002.

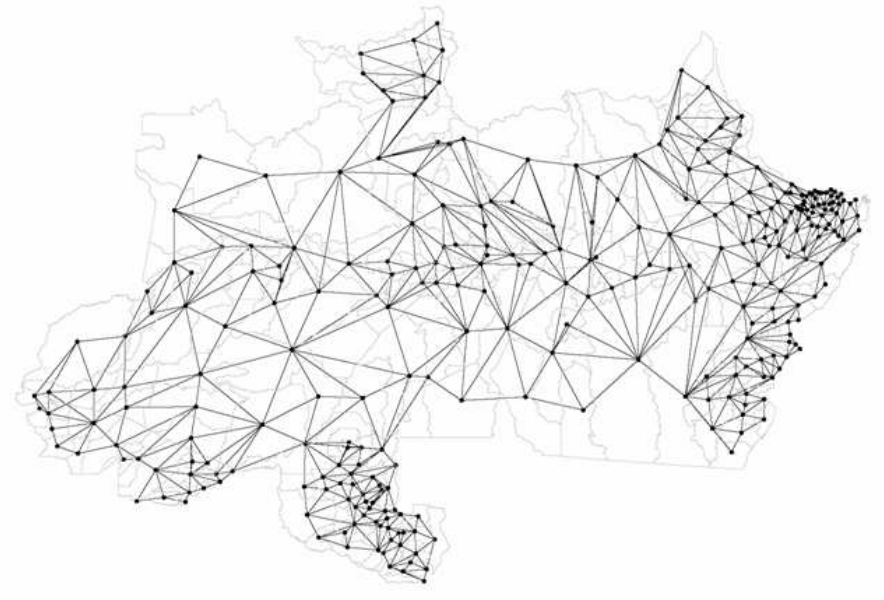


Figura 4.18: Estrutura de vizinhança do mapa da Região Norte.

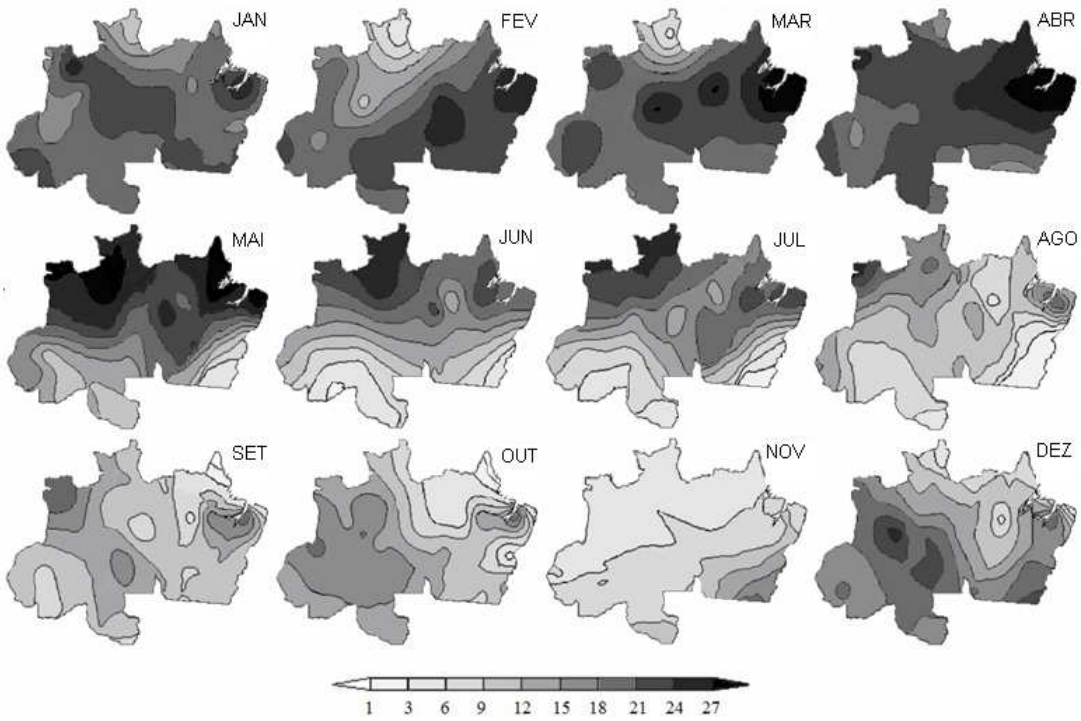


Figura 4.19: Número de dias chuvosos por mês na Amazônia Brasileira, 2000.

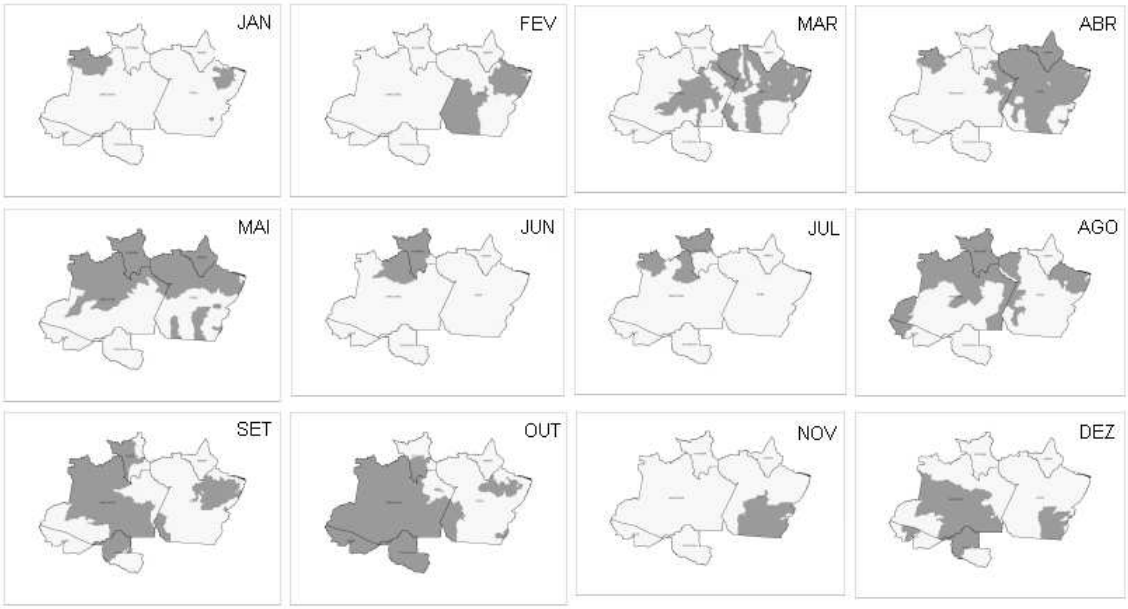


Figura 4.20: Municípios mais úmidos por mês na Amazônia Brasileira, 2000.

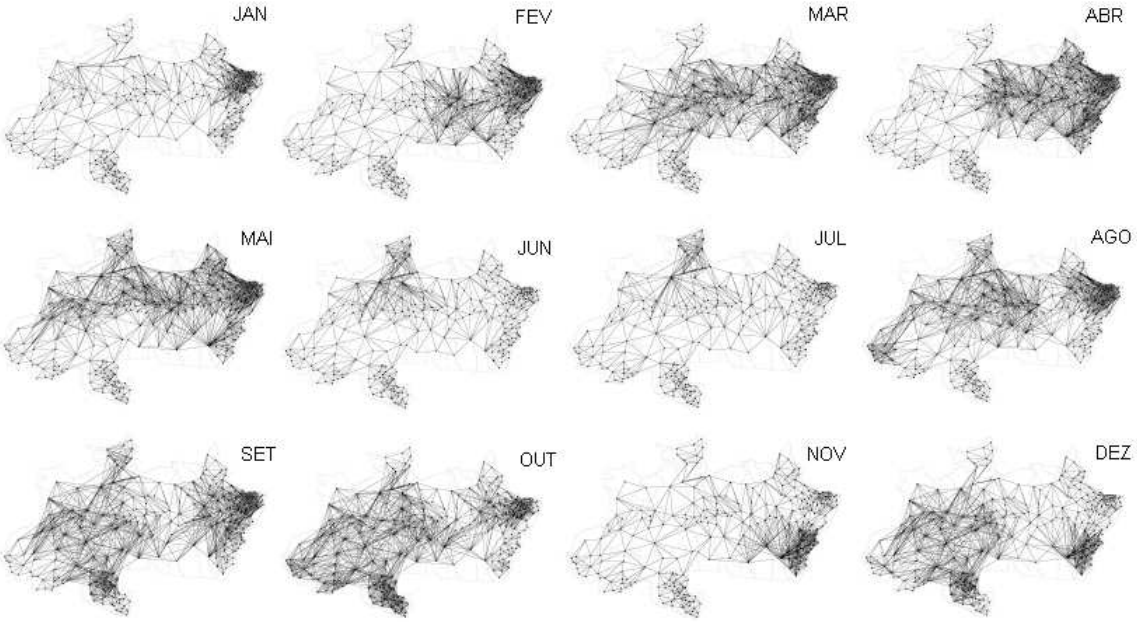


Figura 4.21: Grafos reforçados nos municípios mais úmidos por mês.

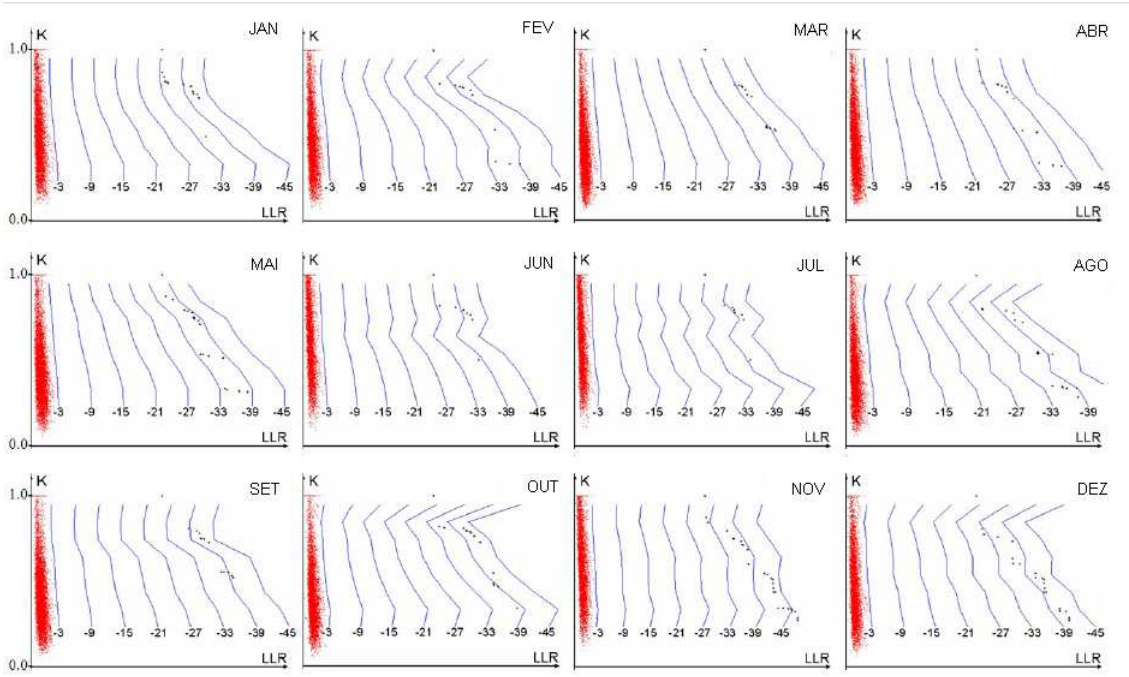


Figura 4.22: Significância das soluções de Pareto por mês. As isolinhas de valor p referem-se a $10^{-3}, 10^{-9}, \ldots, 10^{-45}$.

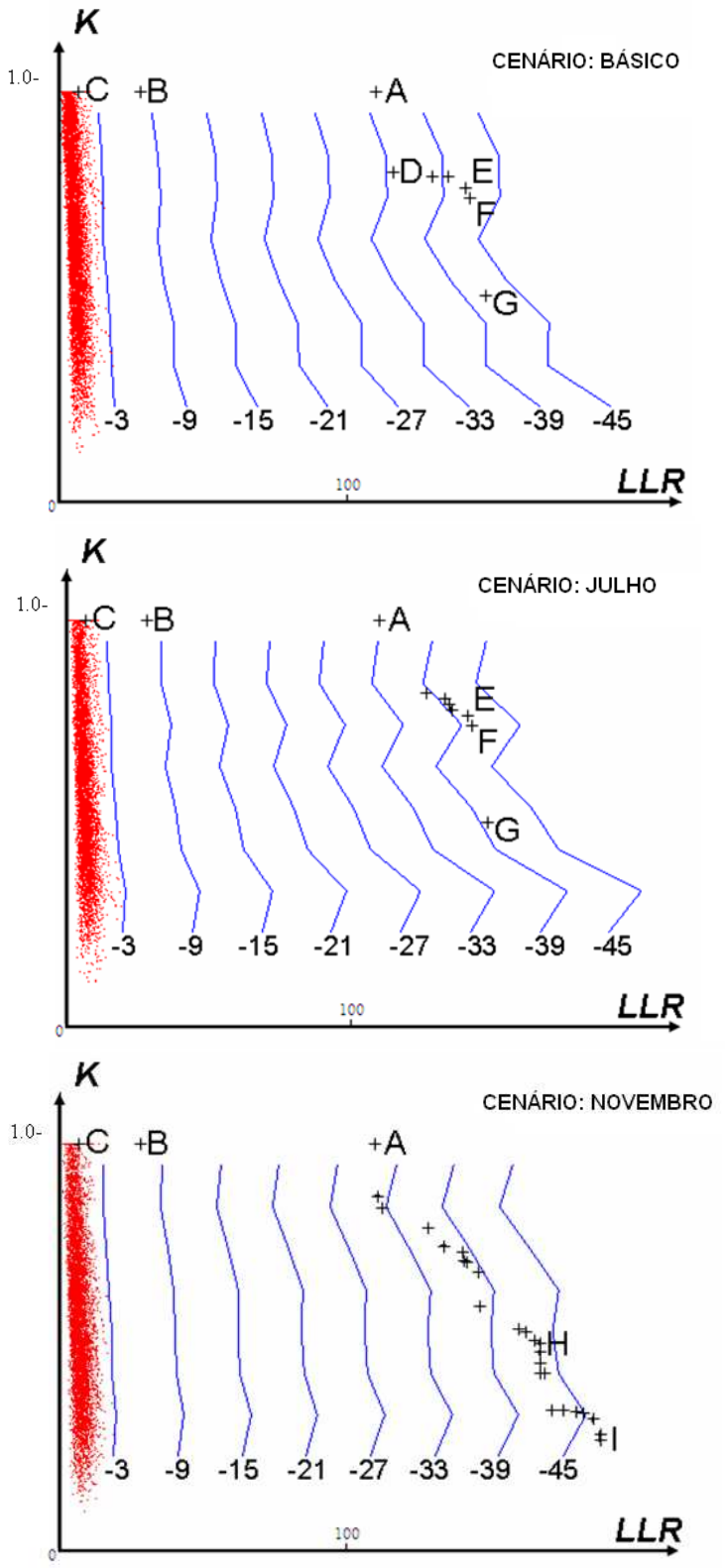


Figura 4.23: Visualizando alguns clusters diante das isolinhas de valor p referentes a $10^{-3}, 10^{-9}, ..., 10^{-45}$.

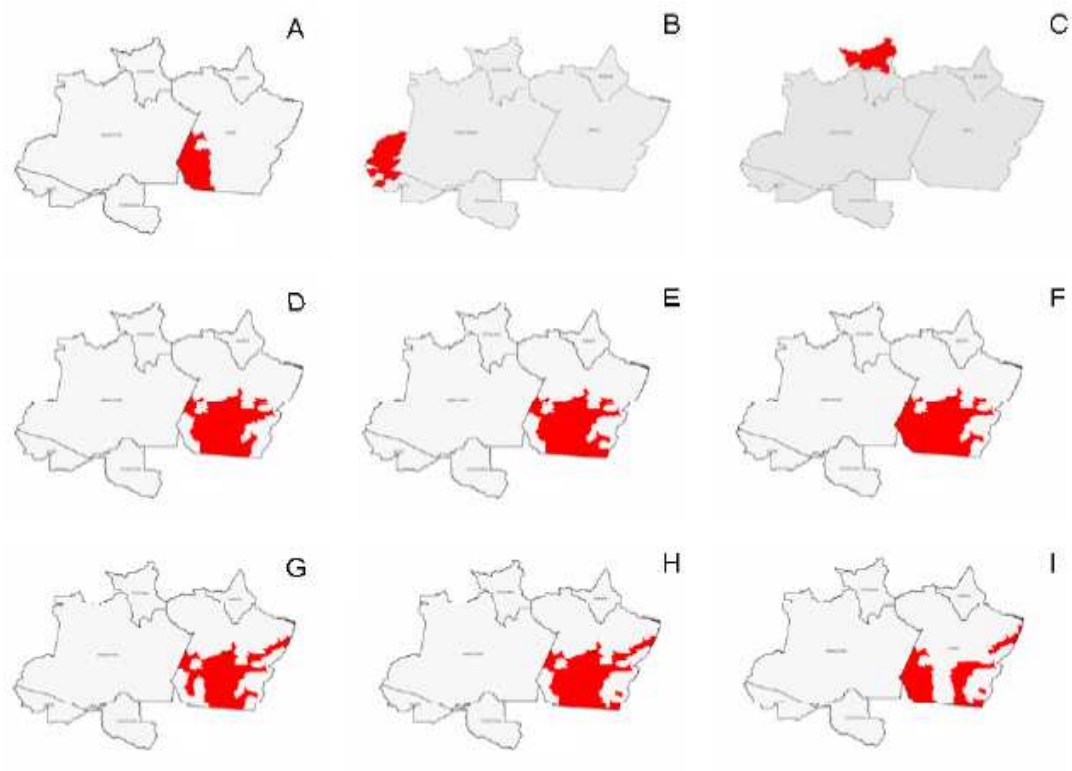


Figura 4.24: Visualizando alguns clusters no mapa.

Capítulo 5

Estatística Scan Multiseletiva

Comumente têm-se presenciado situações em que clusters espaciais de uma determinada patologia não apresentam um formato regular, ou ainda, situações em que não se tem claramente um cluster primário definido. O método proposto neste capítulo analisa os diversos níveis de clusterização que aparecem naturalmente em mapas subdivididos em regiões.

A Estatística Scan Espacial é a medida mais utilizada para quantificar a intensidade de um cluster. Duczmal et al. (2008) trabalharam com técnicas multiobjetivo usando o algoritmo genético para localizar clusters espaciais levando em conta, além da intensidade do cluster, a sua regularidade geométrica. Em Tango (2007) foi proposta uma estatística de teste da razão de verossimilhança modificada em que se avalia o risco para cada região individual do mapa. Esta scan modificada inclui uma variável indicadora baseada no valor p para a zona composta da região individual *i*. Dado um valor p=α_1 pré-especificado, e se p_i é o valor p da zona composta pela região individual *i*, então a razão de verossimilhança da scan modificada para um cluster incluindo i é considerada igual a zero quando $p_i > \alpha_1$.

A estatística scan multiseletiva é uma extensão da estatística scan diferente da proposta por Tango (2007), no sentido de que criamos conceitos novos para detectar e avaliar a significância dos clusters através de técnicas de otimização multiobjetivo. A nova estatística foi apresentada em Moura et al. (2007). O método foi proposto para analisar mais precisamente os vários níveis de clusterização que surgem naturalmente em um mapa dividido em *m* regiões. Neste trabalho, serão apresentados novos resultados a partir do desenvolvimento

da metodologia necessária para avaliar o poder de teste da estatística scan multiseletiva e estudados o desempenho do método para clusters resultantes de processo de difusão.

Ao invés de usar um algoritmo genético, este método incorpora a simplicidade e a velocidade do scan circular, podendo detectar e avaliar clusters de formato irregular. A ocupação circular (OC) de um candidato a cluster é definida como sua população dividida pela população dentro do menor círculo que o contém. O conceito de OC substitui aqui a compacidade como a medida de regularidade do formato e é computacionalmente mais rápido. Uma modificação multiobjetivo do algoritmo scan circular é aplicada usando os objetivos como sendo a OC e a LLR. A comparação dos conjuntos Pareto para os casos observados com aqueles calculados sob a hipótese nula de não existir cluster no mapa, fornece pistas para a distribuição espacial do número de casos da doença estudada. O potencial para monitorar clusters espaço-temporais incipientes em várias escalas geográficas simultaneamente é uma ferramenta promissora em vigilância sindrômica, especialmente para doenças contagiosas quando existe uma mistura de interações espaciais de curta e longa amplitudes. A presença de "joelhos" nos conjuntos Pareto indicam transições nas estruturas dos clusters, correspondendo a rearrajos devido à coalescência da malha (geralmente desconectada).

O método multiseletivo se diferencia do algoritmo genético multiobjetivo a partir dos conjuntos seletivos, pois o procedimento se inicia num contexto em que as regiões de baixa LLR são apagadas. Em outras palavras, apenas as regiões de alta LLR serão consideradas para o mapa que será utilizado em cada nível de clusterização. E são estes conjuntos seletivos os responsáveis por admitir zonas desconexas, o que o algoritmo genético não incorpora na sua busca. Quando o nível de desconexidade é alto, as regiões que compõem tal conjunto seletivo tendem a ter um LLR muito alto, e assim surge a necessidade do outro objetivo desenvolvido para penalizar zonas muito desconexas ou muito irregulares (ocupação circular). A penalização de clusters de acordo com a irregularidade de seus formatos pode ser vista esquematicamente na Figura 5.1, cuja penalização é da forma $OC(z)LLR(z)$. Clusters circulares não são penalizados (sua verossimilhança é multiplicada por 1) enquanto que clusters altamente irregulares ou desconexos são bastante penalizados (sua verossimilhança é multi-

plicada por um número próximo de zero). A ocupação circular de uma zona é obtida ao se dividir a sua população pela população do menor círculo que a contém, amenizando assim o efeito da irregularidade e da desconexidade gerada pelo respectivo conjunto seletivo.

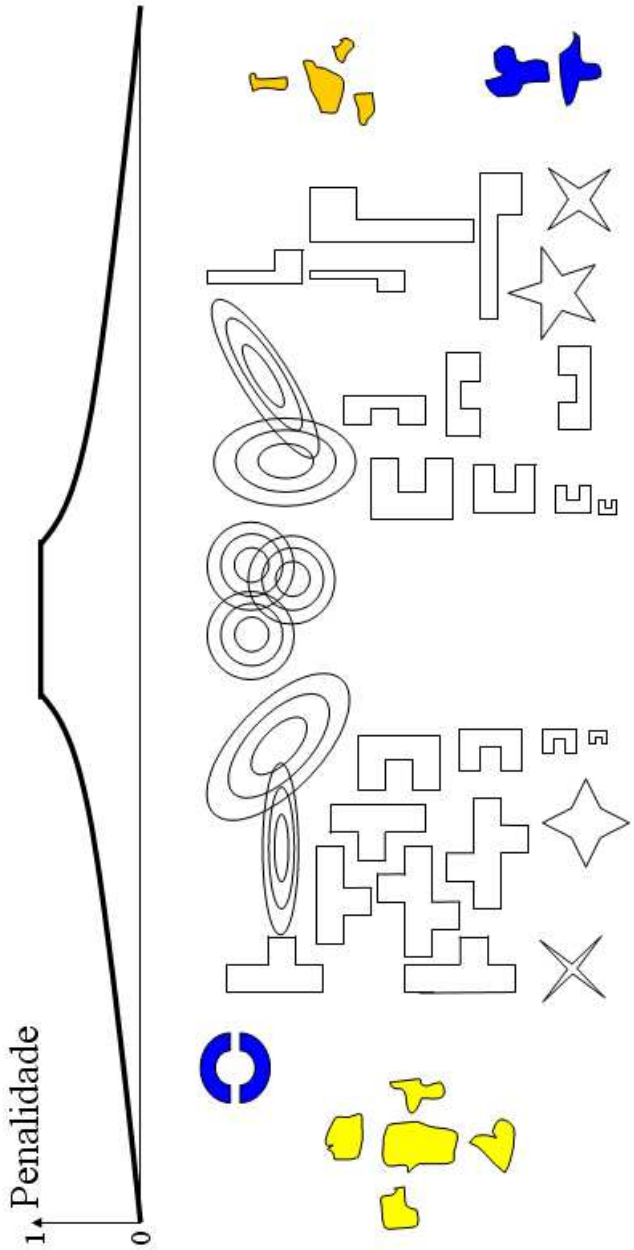


Figura 5.1: Formas geométricas possíveis para as zonas ao utilizar a scan multiseletiva.

Com a necessidade de analisar a LLR para alguns conjuntos seletivos diferentes, o método também considera, a exemplo do genético multiobjetivo, dois objetivos: a LLR (estatística scan) e a ocupação circular (regularidade do formato).

Um dos pontos mais relevantes deste método é que ele, através dos conjuntos seletivos, apaga as regiões com incidência baixa e admite clusters desconexos que têm uma importância significativa em vários contextos práticos. Em muitas situações é possível imaginar um cluster "espalhado" e portanto desconexo devido ao fato do sinal da incidência não ter uma ligação direta com os vizinhos mais próximos e sim com algumas áreas mais distantes. Como por exemplo, ao se estudar a vulnerabilidade social em uma metrópole, em que áreas de baixo IDH (Índice de Desenvolvimento Humano) são em geral distantes ou desconexas. Uma outra situação, se refere à difusão de doenças transmitidas por aves contaminadas que migram de um município à outro município próximo porém não vizinho ao município de partida.

5.1 Conjuntos Seletivos

Os conjuntos seletivos são obtidos a partir das regiões ordenadas segundo suas verossimilhanças. Suponha um mapa com m regiões $\{r_1, r_2, \cdots, r_m\}$. Defina $L_i = LLR(r_i)$ como sendo o valor do logaritmo da razão de verossimilhança (da estatística scan espacial) da zona contendo apenas a região r_i. Ordene as m regiões do mapa de modo que $L_1 > L_2 > \cdots > L_m$. Defina o subconjunto $R_j = \{r_1, r_2, \cdots, r_j\}$ como sendo o conjunto seletivo de tamanho j ($j = 1, 2, \cdots, m$). Perceba que $R_1 \subset R_2 \subset \cdots \subset R_m$. Os conjuntos R_1, $\cdots$, R_m não são necessariamente conexos.

A Figura 5.2 apresenta uma ilustração sobre o conceito de conjuntos seletivos. A Figura 5.2(a) representa o mapa de Minas Gerais em que as 0,4%(ou seja, 0,004*853=3) regiões com maiores LLR individuais (referentes aos óbitos por homicídios, 1998 a 2002) são destacadas em cinza escuro. Da mesma forma, a Figura 5.2(b) apresenta 0,8% (7) regiões; a Figura 5.2(c), 1,6%(14); a Figura 5.2(d), 3,2%(27) e assim por diante até a Figura 5.2(i) que detém as 100% (853) regiões do mapa.

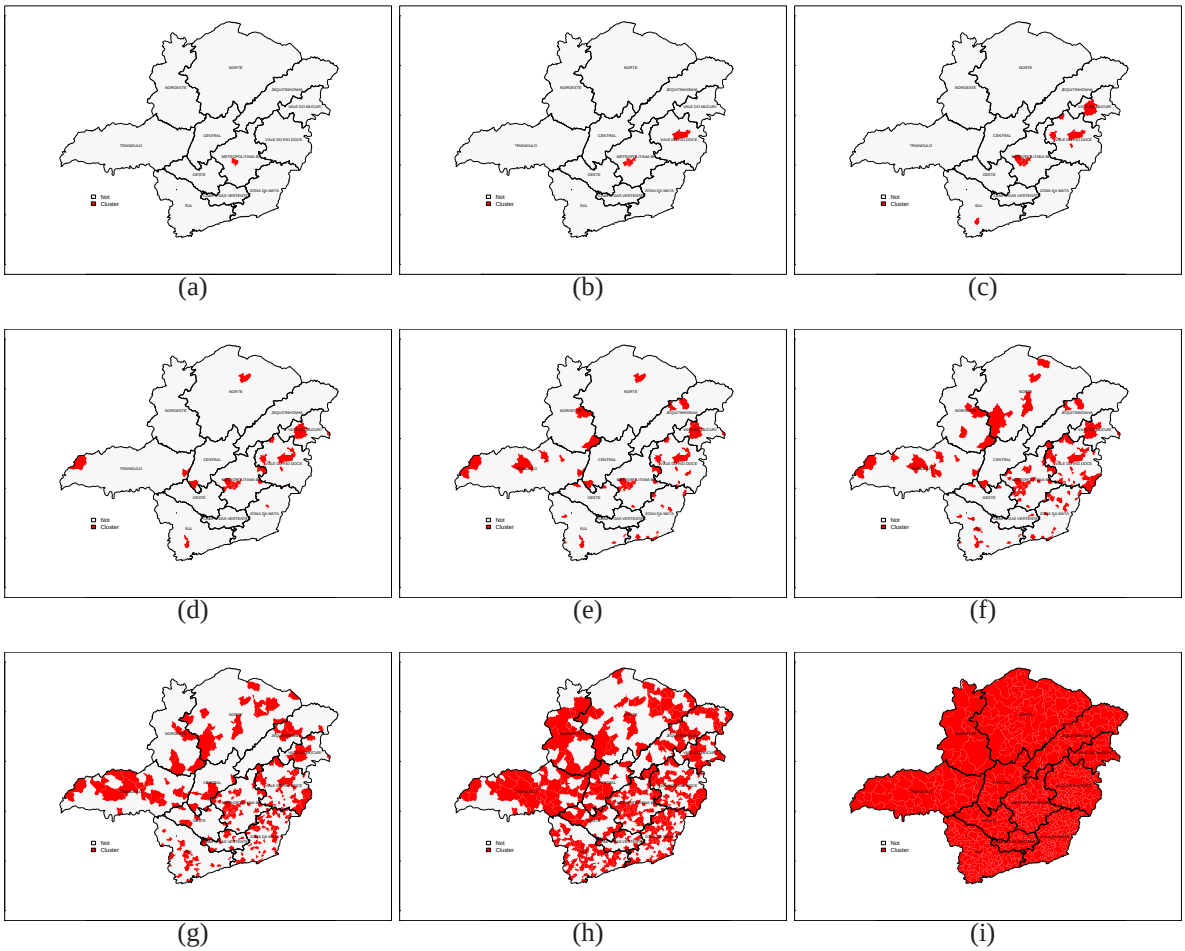


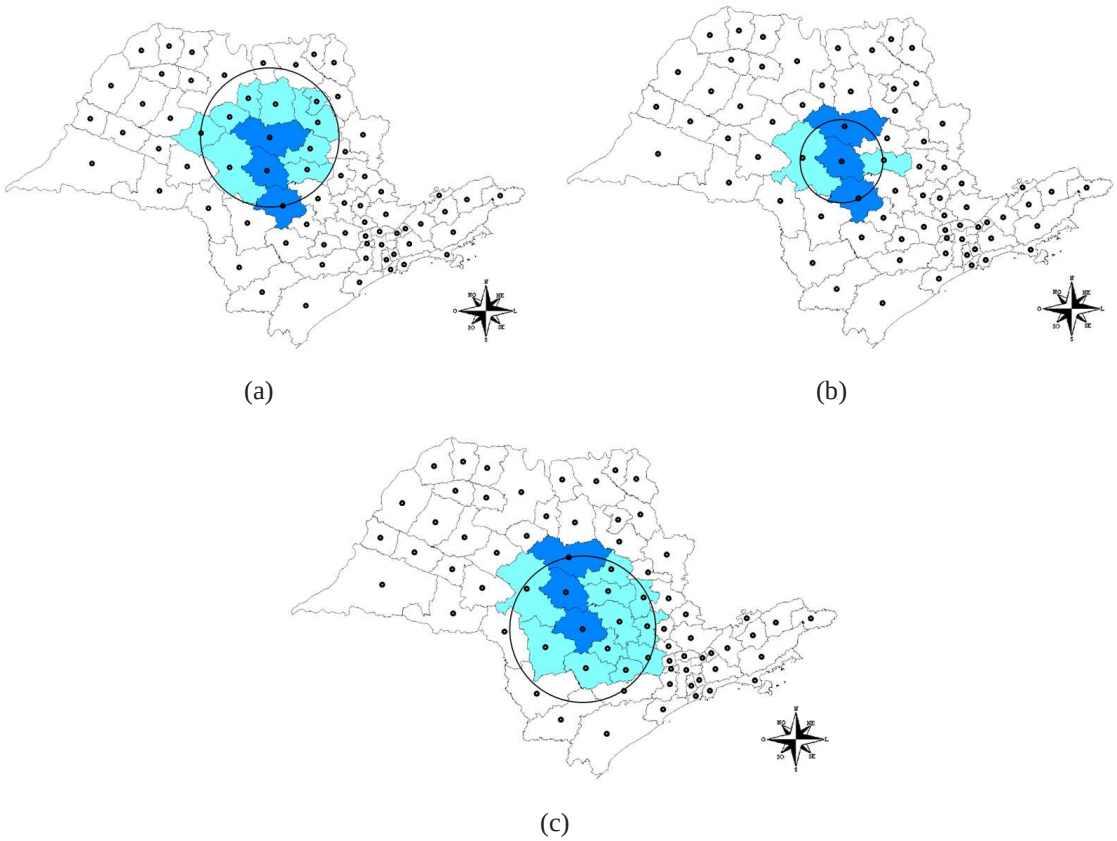
Figura 5.2: Ilustração dos conjuntos seletivos

5.2 Ocupação Circular

A ocupação circular de uma zona z é definida como a razão de sua população pela população do menor círculo que a contém. Neste trabalho, esta medida substitui a compacidade e estende o conceito de zona para englobar áreas desconexas.

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 de todas as regiões do mapa original cujos centróides estão dentro de C. A ocupação circular da zona z, $OC(z)$, é dada pelo quociente $P(z)/P(C)$, ou seja, visualizando um dos mapas da Figura 5.3, seria a razão da população da zona (cinza escuro) pela população do círculo (cinza escuro

+ cinza claro). O problema é que esta medida pode não ser única. A Figura 5.3 apresenta três situações possíveis para o valor da ocupação circular dentro de uma mesma zona z. O problema foi resolvido calculando a ocupação circular usando o máximo de todos os quocientes possíveis, $OC(z) = max_{r_j \in z}\{P(z)/P(C_{r_j})\}$ em que C_{r_j} é o menor círculo centrado na região r_j que contém a zona z. A ocupação circular varia entre 0 e 1, em que valores próximos de 1 representam clusters com formatos mais regulares e grau de conexidade maior, enquanto que valores próximos de 0 representam clusters mais irregulares e grau de conexidade menor.



(a) (b)

(c)

Figura 5.3: Círculos centrados em cada uma das três regiões (superior, central e inferior) em cinza escuro.

Desta forma, para cada zona z desejamos maximizar dois objetivos: a $LLR(z)$ e a $OC(z)$.

Num conjunto C formado por n zonas $z_1, z_2, \cdots, z_n$, considere os pares ordenados

$$(LLR(z_i), OC(z_i)).$$

O uso de técnicas multiobjetivo é necessário para avaliar possíveis clusters em níveis de regularidade e conexidade diferentes. A Figura 5.4 ilustra o conjunto Pareto-ótimo, representado por $\oplus$, que apresenta as diversas soluções com diferentes níveis de clusterização.

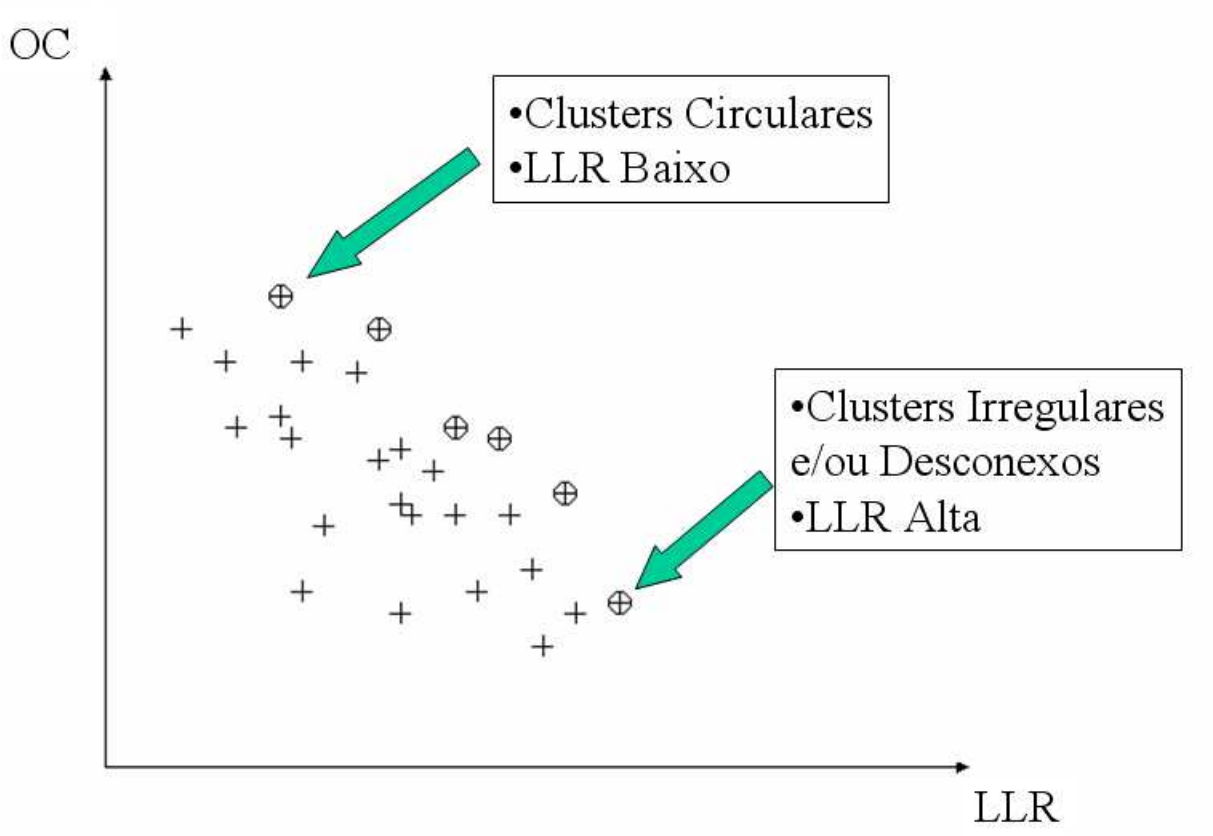


Figura 5.4: Conjunto Pareto-ótimo: os diferentes níveis de clusterização.

O valor de *LLR* é calculado para cada uma das m regiões do mapa e ordenados decrescentemente. Seja $R(j)$ o conjunto contendo as j primeiras regiões. A extensão multiobjetivo do algoritmo scan circular é aplicada sucessivamente para cada conjunto $R(j)$. Em cada círculo, a zona candidata a ser um cluster é composta pelas regiões pertencentes a $R(j)$ e que estão

dentro do círculo. Na prática, escolhe-se apenas alguns poucos valores de j tais como $\lceil m \rceil$, $\lceil m/2 \rceil$, $\lceil m/4 \rceil$, $\cdots$, $\lceil 1 \rceil$. Para cada valor de j, constrói-se um conjunto Pareto-ótimo $P(j)$. Com base em todos esses conjuntos de Pareto, obtém-se o conjunto de Pareto-ótimo Global $P(0)$. Por fim, um procedimento de Monte Carlo é implementado para avaliar a significância dos clusters desse conjunto $P(0)$.

5.3 Algoritmo Scan Multiseletivo

Formalmente, o algoritmo do método proposto nesta extensão consiste dos seguintes passos:

1. Calcule e ordene decrescentemente as verossimilhanças de cada uma das regiões do mapa;

2. Calcule a matriz de distâncias entre as regiões considerando a ordenação realizada no passo anterior;

3. Para cada uma das regiões ordenadas, construa o vetor de população acumulada das j-ésimas regiões vizinhas;

4. Para cada valor de a, $0 < a \leq 1$, considere as $100a\%$ regiões com maiores verossimilhanças, explicitamente as regiões $r_1, r_2, \cdots, r_A$, com A sendo o maior inteiro menor ou igual a $a.m$:

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

   4.2. Construa o vetor de população corrente, calculada por $PC = \sum_{i=1}^{A} Pop[r_i]$;

   4.3. Utilizando apenas as regiões $r_1, r_2, \cdots, r_A$ construa todas as possíveis zonas circulares, com a restrição de que a população de cada zona não exceda o mínimo entre a população corrente e 25% da população total do mapa;

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 o conjunto Pareto-ótimo de C_A e denote-o por $P(C_A)$;

5. Calcule o conjunto Pareto-ótimo dos conjuntos Pareto de todos os valores de a, ou seja, $\bigcup_a P(C_A)$, e denote-o por P (conjunto Pareto-ótimo Global);
6. Use Monte Carlo para avaliar a significância de cada ponto do conjunto Pareto-ótimo (P), que serão os clusters.

Quando o conjunto seletivo corresponde a 100% das regiões $(OC = 1)$, o algoritmo acima equivale ao algoritmo scan circular usual.

O método foi avaliado por meio de simulações e aplicado a uma situação real, e os resultados são apresentados na seção 5.4 e 5.5, respectivamente.

Os dados utilizados neste capítulo são de fontes de natureza pública. Os dados de população foram obtidos junto ao portal do Ministério da Saúde (http://www.datasus.gov.br), enquanto que os dados referentes aos óbitos causados por homicídios e por bronquite foram do Sistema de Informações sobre Mortalidade (SIM) do Ministério da Saúde. Esses dados se referem ao período de 1998 a 2002.

5.4 Avaliação Numérica

Nosso método foi avaliado para cinco diferentes clusters artificiais, regulares e irregulares, de pequena e grande população, conexos e desconexos, utilizando o mapa de 853 municípios de Minas Gerais, com dados reais de população.

O risco relativo (maior que 1) dentro de cada cluster foi estabelecido de modo que clusters de população pequena tivessem risco relativamente grande e clusters de maior população tivessem risco relativo pequeno. Isso foi feito com o propósito de rejeitar a hipótese nula com a mesma probabilidade θ qualquer que seja o cluster, para um algoritmo que soubesse

a localização exata do cluster, seguindo o procedimento de Kulldorff et al. (2003) e descrito em Lima (2004).

5.4.1 Estimação de Risco Relativo para os Clusters

Para calcular o risco relativo de um cluster, Kulldorff et al. (2003) consideram o seguinte procedimento. Seja n_z a população em risco do cluster, e N a população total do mapa. Dado o número total de casos C, o número c_z de casos observados no cluster sob a hipótese nula, de não existir cluster espacial no mapa, tem distribuição Binomial com parâmetros (C, τ_z) com $\tau_z = \frac{n_z}{N}$. A média e a variância desta distribuição são dadas, respectivamente, por:

$$m_0 = \frac{n_z C}{N} \quad e \quad v_0 = \frac{n_z C(N-n_z)}{N^2}$$

Usando a aproximação normal para a distribuição binomial, o número crítico de casos k para que o teste unilateral rejeite a hipótese nula com o nível de significância $0 < \alpha < 1$ é tal que:

$$\Phi\left(\frac{k-m_0}{\sqrt{v_0}}\right) = \alpha \quad \Longrightarrow \quad \frac{(k-m_0)}{\sqrt{v_0}} = \Phi^{-1}(\alpha)$$

em que $\Phi(.)$ é a função de distribuição acumulada da Normal padrão. Se $\alpha = 0,05$ e $\theta = 1-\alpha$ temos que $\Phi^{-1}(\theta) = 1,645$, daí o valor crítico k é tal que $\frac{(k-m_0)}{\sqrt{v_0}} = 1,645$. Sob a hipótese alternativa, com o risco relativo ρ_z para a região do cluster, o número de casos nesta região tem distribuição Binomial com média $m_a = \frac{n_z C \rho_z}{(N-n_z+n_z\rho_z)}$ e variância $v_a = \frac{n_z C \rho_z (N-n_z)}{(N-n_z+n_z\rho_z)^2}$. Observe, neste caso, que $\tau_z = \frac{n_z \rho_z}{(N-n_z+n_z\rho_z)}$. Usando novamente a aproximação Normal, selecionamos o risco relativo ρ_z tal que $\frac{(k-m_a)}{\sqrt{v_a}} = \Phi^{-1}(\theta)$. Desta forma o risco relativo é escolhido de modo que o poder atingido por um teste "ideal" (isto é, um método idealizado em que se informa exatamente a localização do cluster para que este seja detectado) para cluster espacial tem um limite superior igual a θ. Observe que este valor de θ não corresponde ao poder do teste do método a ser testado, mas sim para o poder do método "ideal". O método testado tem poder muito menor que θ em geral (Kulldorff et al., 2003). Neste trabalho foi escolhido o valor de θ igual a $0,999$.

5.4.2 Estimação do Poder

Seja C_b o conjunto de todos os candidatos a cluster cuja ocupação circular seja maior ou igual a b, para valores $0 < b \leq 1$. Assim, o scan circular usual encontra clusters no conjunto C_1, e o scan multiseletivo encontra clusters no conjunto C_0. Para qualquer outro valor de b, estamos restringindo nossa busca de modo a recusar clusters cuja ocupação circular seja menor que b. Em outras palavras, estaremos limitando o grau de desconexidade ou irregularidade que estamos dispostos a tolerar para os clusters encontrados.

Para fixar as idéias, considere os valores $b = 1.0, 0.9, 0.8, ..., 0.1, 0.0$. Observe que $C_{1.0} \subset C_{0.9} \subset ... \subset C_{0.1} \subset C_{0.0}$, isto é, o conjunto de busca aumenta com a diminuição de b.

A avaliação do poder de detecção do algoritmo foi feita de modo a se estimar a proporção de vezes em que um cluster é detectado, separadamente para cada C_b. Com isso estaremos verificando como o algoritmo se comporta, frente a diferentes graus de restrição aplicados à liberdade de forma permitida aos clusters.

Por exemplo, se não quisermos clusters conexos, na maioria das vezes podemos utilizar $b = 1.0$ ou mesmo $b = 0.8$. Se quisermos permitir clusters desconexos mas cujas regiões não estejam muito afastadas entre si, provavelmente o valor $b = 0.5$ deve ser satisfatório. Na prática o usuário do sistema deve efetuar uma análise exploratória e verificar o que acontece com a conexidade dos clusters soluções encontrados pelo algoritmo quando variamos o valor de b. É de se esperar que o valor de b não seja muito diferente do valor da ocupação circular do cluster verdadeiro, quando este existir.

As Figuras 5.6, 5.7, 5.8, 5.9 e 5.14, mostram o comportamento do scan multiseletivo quando variamos os valores de b para $b = 1.0, 0.9, 0.8, ..., 0.1, 0.0$.

5.4.3 Clusters Conexos

Para avaliar o desempenho do método, foi construído quatro clusters artificiais de formas e de populações diferentes, porém todos conexos. Os clusters simulados foram: *Circular* (cluster circular), *Fino* (cluster em formato de tira), *Ppeq* (cluster com população pequena)

e *Pgra* (cluster com população grande). Algumas características desses quatro clusters estão na Tabela 5.1.

Tabela 5.1: *Características dos quatro clusters artificiais conexos.*

Características	*Circular*	*Fino*	*Ppeq*	*Pgra*
Número de municípios	22	5	11	7
Número de habitantes	383.488	89.010	57.521	3.725.982
Risco Relativo	1,506	2,152	2,496	1,171
Ocupação Circular	0,880	0,434	0,099	0,953

A Figura 5.5 apresenta os diferentes formatos desses quatro clusters no mapa. A Tabela 5.2 apresenta a comparação do poder de detecção entre a Estatística Scan (OC=1) e a Estatística Scan Multiseletiva nas diversas situações de formato do cluster (conforme a faixa de OC). Para tornar o método menos dependente da escolha exata da faixa, a ocupação circular é apresentada através de faixas acumulativas que vão de um certo valor de OC até o nível máximo de regularidade para a Estatística Scan Circular (OC=1). Para os clusters artificiais de formato mais regular (*Circular* e *Pgra*), percebe-se que existe pelo menos uma faixa de valores da ocupação circular em que a Estatística Scan Multiseletiva tem poder muito próximo da Estatística Scan Circular, porém, com a vantagem da não existência de subestimação e superestimação do cluster. Observa-se também que em situações circulares, a Estatística Scan tem poder de detecção maior que a proposta deste capítulo. Para os clusters com geometria mais irregular (*Fino* e *Ppeq*), a Estatística Scan Multiseletiva tem poder de teste superior à Scan Circular em pelo menos uma faixa cuja ocupação circular seja menor que 1,0. Isso evidencia que é possível melhorar o poder de teste para detectar clusters irregulares usando a Estatística Scan Multiseletiva.

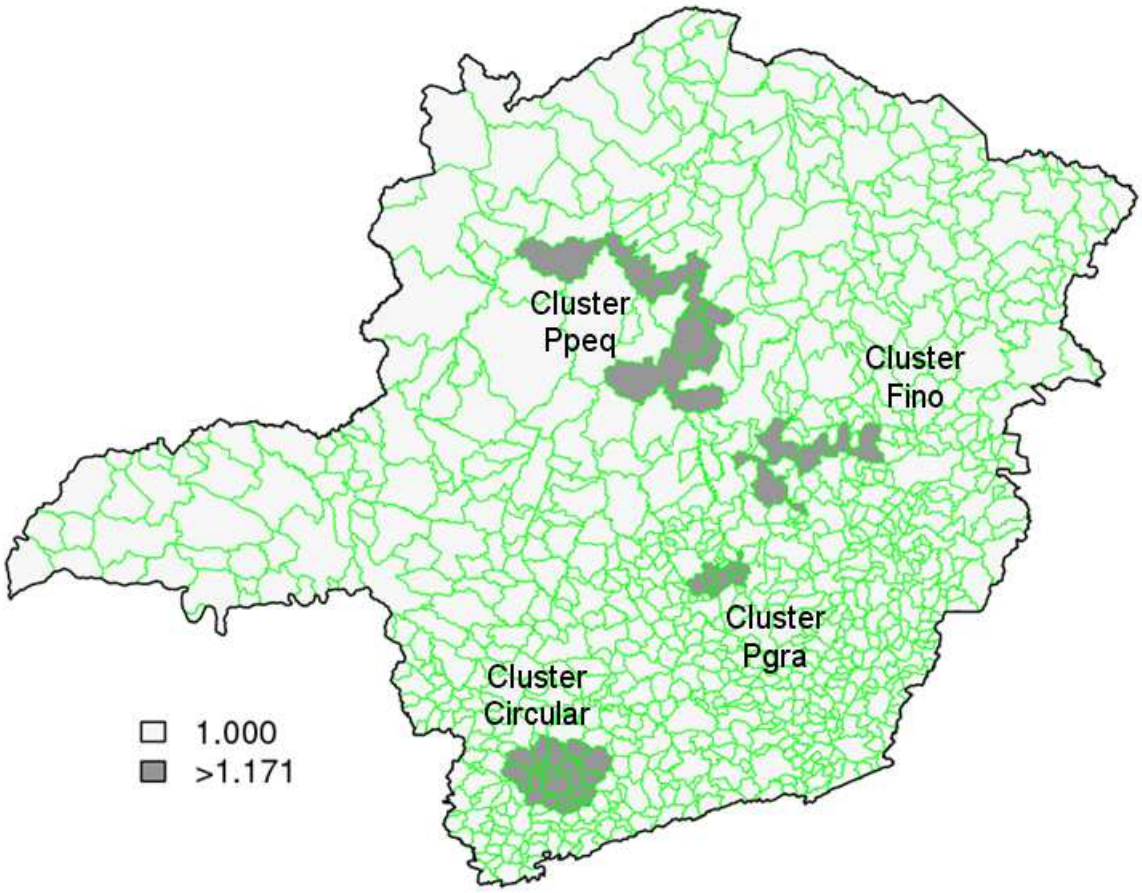


Figura 5.5: Os quatro clusters artificiais conexos cujo risco relativo é superior a 1,171.

Tabela 5.2: *Poder da Estatística Multiseletiva para os quatro clusters artificiais avaliados.*

Ocupação Circular	*Circular*	*Fino*	*Ppeq*	*Pgra*
1	0,7870	0,4380	0,0481	0,4632
0,90 - 1,00	0,7807	0,4300	0,0482	0,4631
0,80 - 1,00	0,7799	0,4224	0,0524	0,4628
0,70 - 1,00	0,7910	0,4284	0,0645	0,4624
0,60 - 1,00	0,7868	0,4771	0,0899	0,4623
0,50 - 1,00	0,7497	0,4812	0,1076	0,4611
0,40 - 1,00	0,7178	0,4702	0,1219	0,4591
0,30 - 1,00	0,6613	0,4444	0,1368	0,4553
0,20 - 1,00	0,5744	0,3935	0,1697	0,4490
0,10 - 1,00	0,4221	0,3150	0,2095	0,4252
0,00 - 1,00	0,1609	0,1985	0,2258	0,1726

Observe que para os clusters *Circular* (Figura 5.6) e *Pgra* (Figura 5.9) o poder se mantém quase inalterado para valores altos de b e decai para valores menores de b. Isso é bastante esperado, pois nesses casos os clusters são quase circulares e o algoritmo scan circular usual deve detectá-los bem, enquanto que a ampliação do conjunto de busca C_b para valores baixos de b apenas introduz ruído (clusters altamente irregulares cujo LLR é elevado), a partir de certo valor.

No entanto, para os clusters *Fino* (Figura 5.7) e *Ppeq* (Figura 5.8) isso não acontece, pois são clusters altamente irregulares. Isso faz com que a ampliação do conjunto de busca C_b forneça mais sinal (clusters candidatos com formato compatível com o cluster verdadeiro) em comparação com o ruído introduzido ao se aumentar o conjunto de busca. Portanto, observamos que para faixas de ocupação circular próximas da verdadeira ocupação circular do cluster artificial, o scan multiseletivo tem seu poder aumentado em comparação com o scan circular usual. Assim, o máximo do poder de detecção é atingido para o cluster *Fino*, enquanto que o poder cresce monotonicamente para o cluster *Ppeq*. Note que o valor de b para o poder de detecção máximo não fica muito diferente do valor da ocupação circular do cluster verdadeiro (Tabela 5.2).

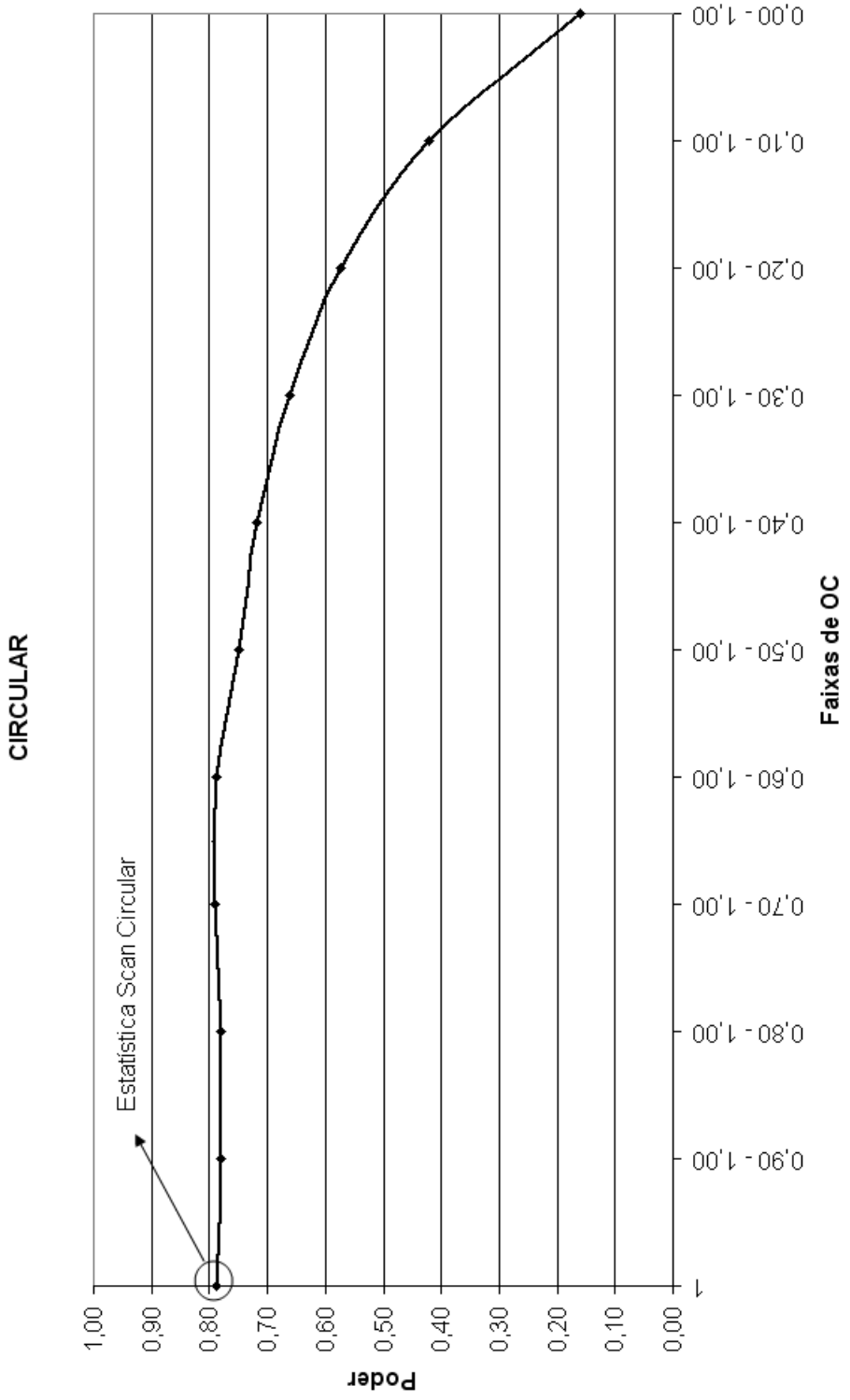


Figura 5.6: Poder para o cluster *Circular*

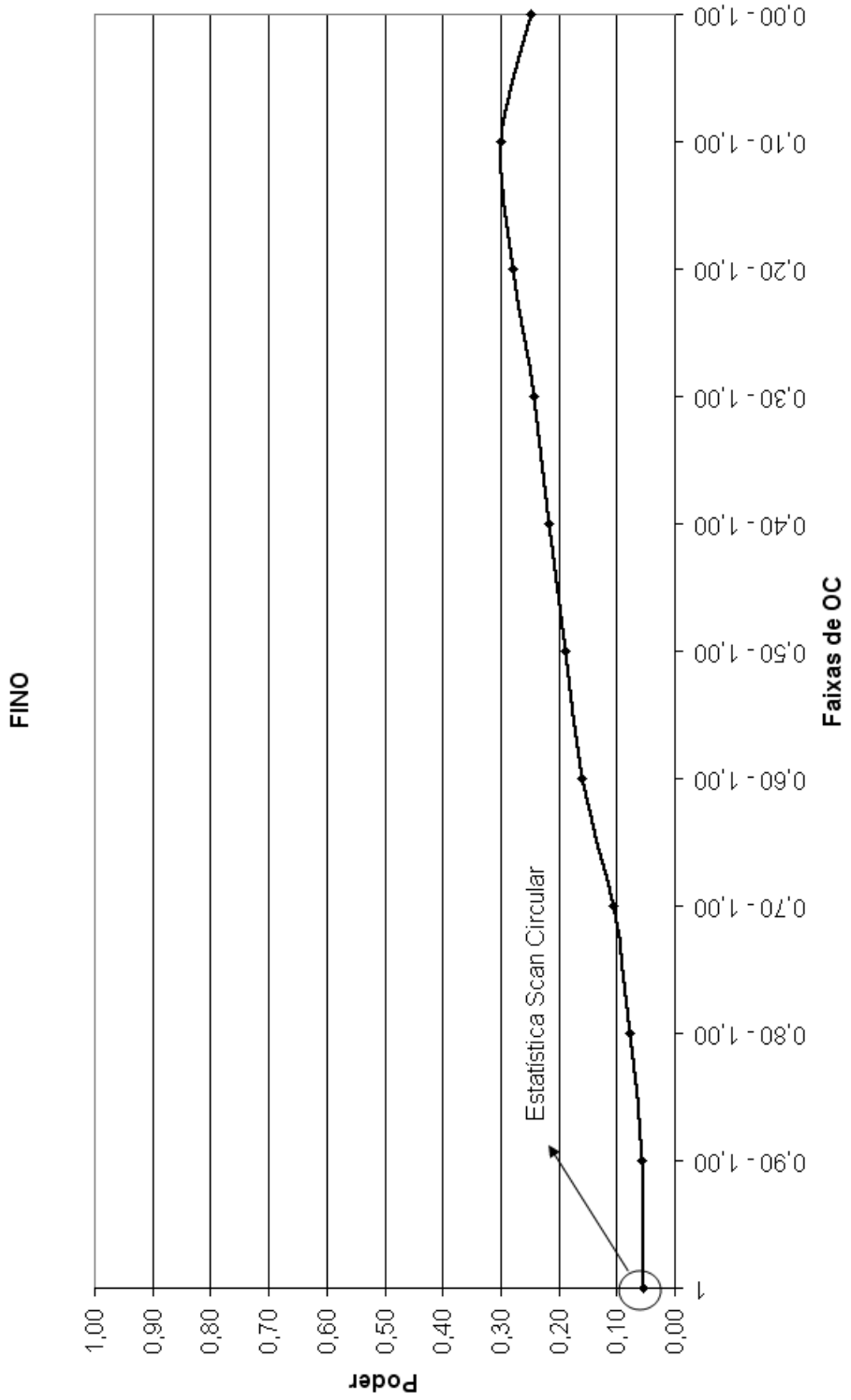


Figura 5.7: Poder para o cluster *Fino*

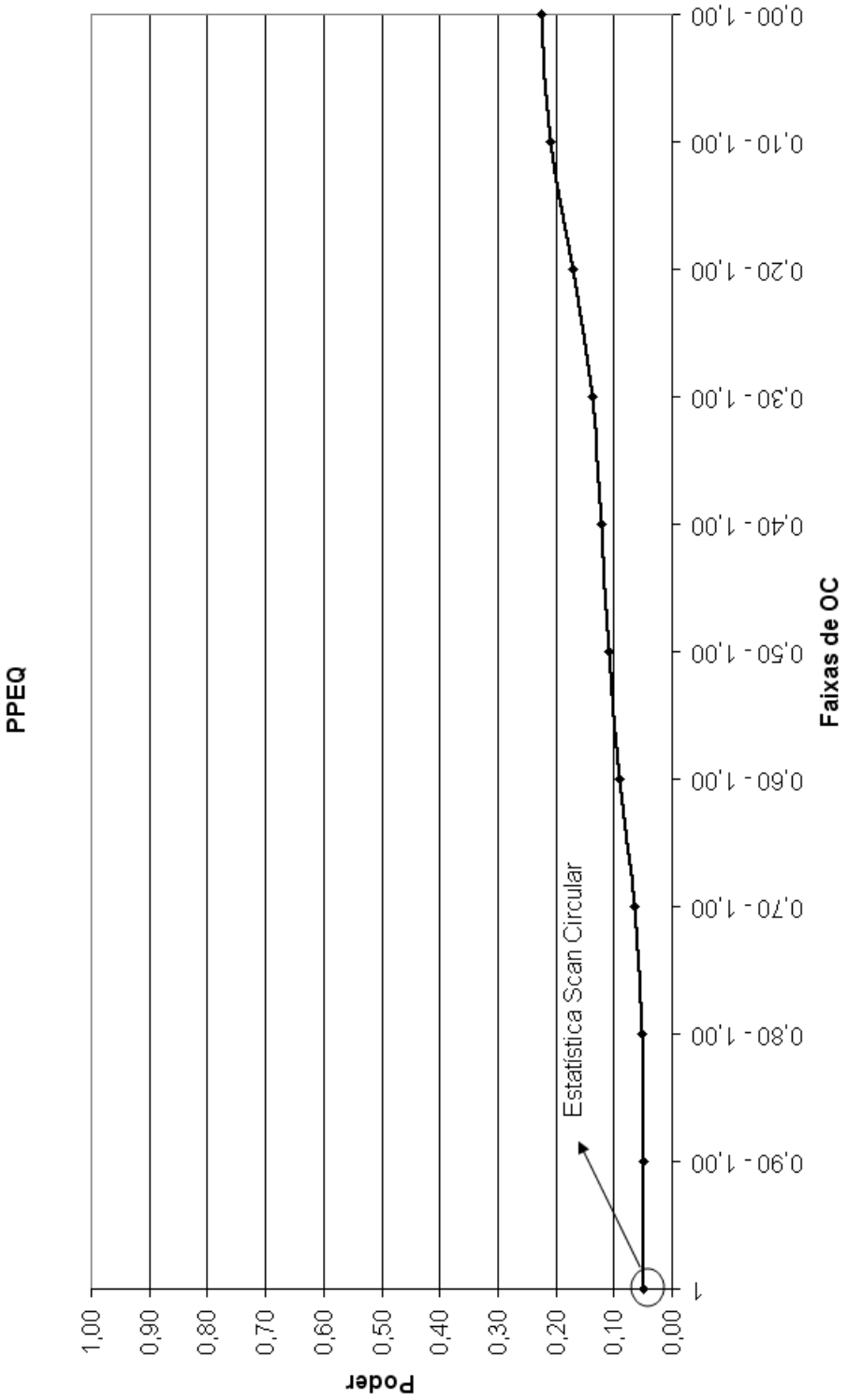


Figura 5.8: Poder para o cluster *Ppeq*

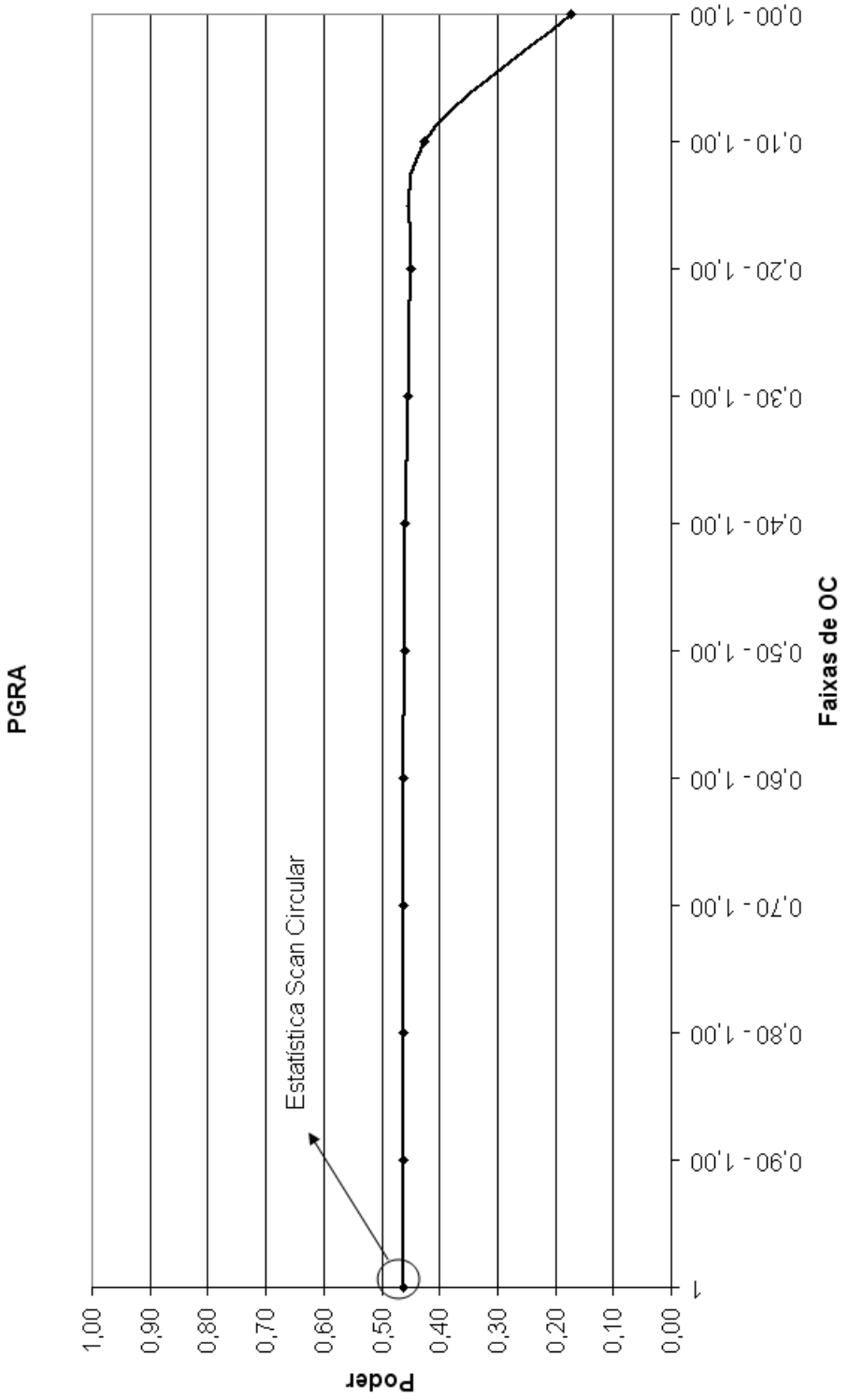


Figura 5.9: Poder para o cluster *Pgra*

5.4.4 Cluster Desconexo gerado por Difusão

A avaliação do método para clusters desconexos também é fundamental. Assim, construímos um cluster artificial desconexo baseado num processo de difusão de doenças transmitidas por aves contaminadas que migram de um município à outro próximo porém não vizinho ao município de partida. O cluster foi denominado *Patos* (cluster desconexo que representa a propagação da gripe aviária em patos selvagens). As características desse cluster simulado segue na Tabela 5.3.

Tabela 5.3: *Característica do cluster desconexo (Patos) baseado no processo de difusão.*

Características	*Patos*
Número de municípios	9
Número de habitantes	143.340
Risco Relativo	1,873
Ocupação Circular	0,372

As Figuras 5.10, 5.11 e 5.12 ilustram passos consecutivos do processo de difusão do vírus da gripe aviária transmitido por patos selvagens de um município a outros municípios, não necessariamente vizinhos geograficamente.

Em nosso modelo, a transmissão começa a partir de um município inicial (região central em cinza escuro), que é a única a reportar casos de gripe aviária no primeiro passo. No passo seguinte aves (patos selvagens) migram para três municípios próximos (mas não vizinhos), escolhidos aleatoriamente pelo modelo de difusão, e novos casos são reportados nos quatro municípios. O município inicial reporta um número de casos ainda maior do que no primeiro passo. A movimentação das aves é indicada pelas setas. No terceiro passo, cinco novos municípios próximos são atingidos aleatorimente por aves contaminadas que migraram dos municípios já contaminados no segundo passo, como indicado pelas setas. Nesse terceiro passo, o município inicial já reporta uma redução no número de casos, enquanto que os três municípios que foram atingidos no segundo passo reportam um número crescente de casos. Além do procedimento de escolha aleatória de municípios próximos que são atingidos

a cada passo, o modelo utiliza uma função de crescimento vegetativo de casos dentro de cada município. No nosso exemplo, esse número de casos máximo é atingido após um passo após a contaminação, e passa a decrescer a partir do segundo passo que se segue à contaminação.

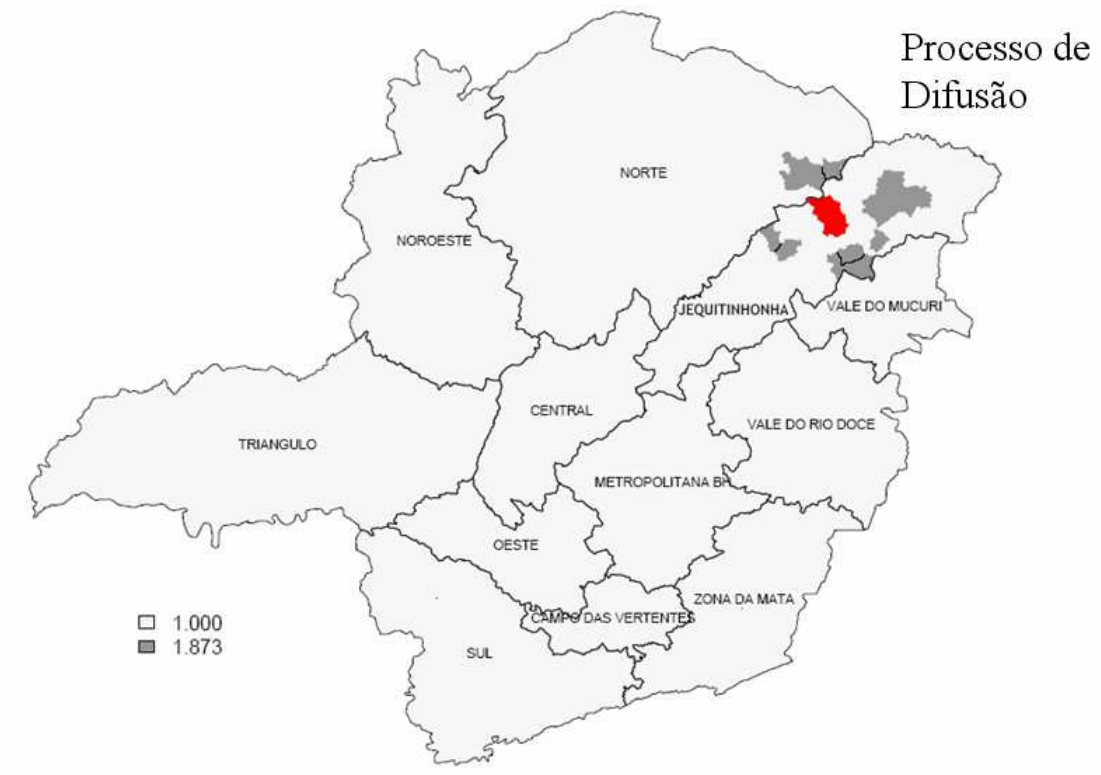


Figura 5.10: O primeiro estágio do processo de difusão que gera o cluster desconexo (*Patos*) com risco relativo igual a 1,873.

Para o cluster desconexo dos *Patos* (Figura 5.14) a ampliação do conjunto de busca C_b também fornece mais sinal (clusters candidatos com formato compatível com o cluster verdadeiro) em comparação com o ruído introduzido ao se aumentar o conjunto de busca. Portanto, observamos que pelo menos até certo ponto (para valores decrescentes de b) o scan multiseletivo tem seu poder aumentado em comparação com o scan circular usual. Nota-se que o valor de b para o poder de detecção máximo não fica muito diferente do valor da ocupação circular do cluster verdadeiro (Tabela 5.4).

A Tabela 5.4 apresenta o poder de detecção entre a Estatística Scan (OC=1) e a Estatística

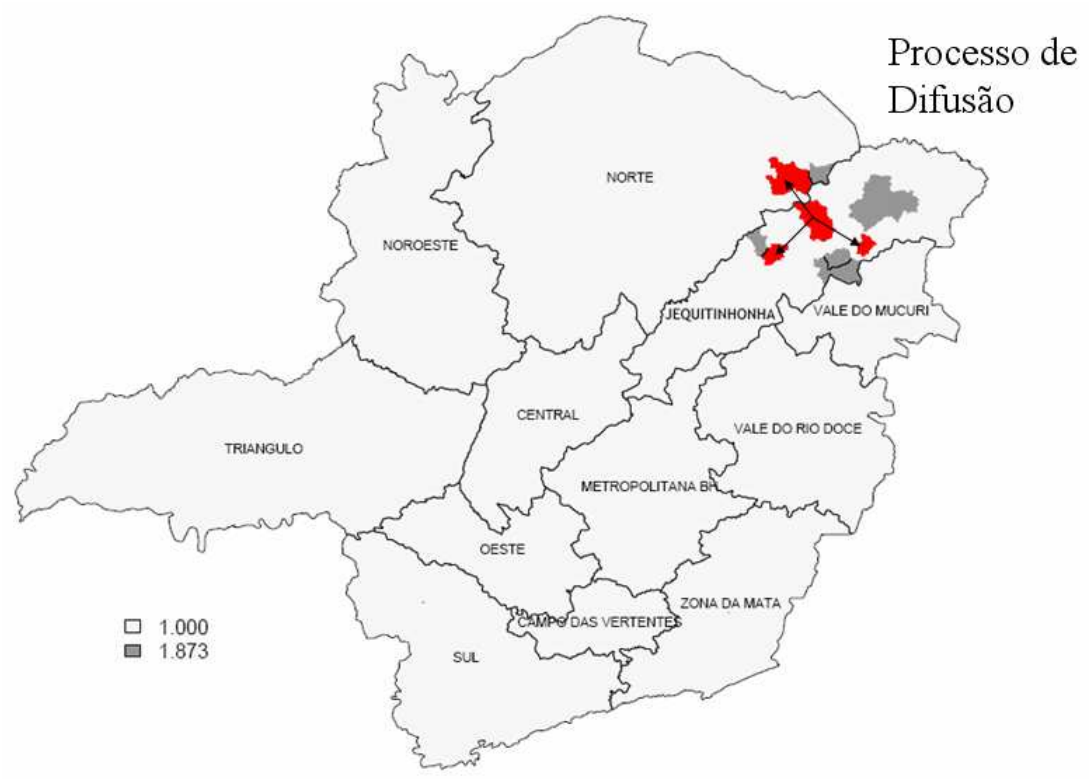


Figura 5.11: O segundo estágio do processo de difusão que gera o cluster desconexo (*Patos*) com risco relativo igual a 1,873.

Scan Multiseletiva na situação cujo o cluster é desconexo. Para o cluster dos patos que tem uma geometria mais irregular e desconexa, a Estatística Scan Multiseletiva teve poder de teste superior à Scan Circular em pelo menos uma faixa cuja ocupação circular seja menor que 1,0. O poder máximo (igual a 0,4812) é atingido na faixa de ocupação circular que admite clusters irregulares até um nível de 0,5 (faixa de OC [0,50 - 1,00]). Portanto, a estatística também apresentou melhorias no poder do teste para o cluster desconexo.

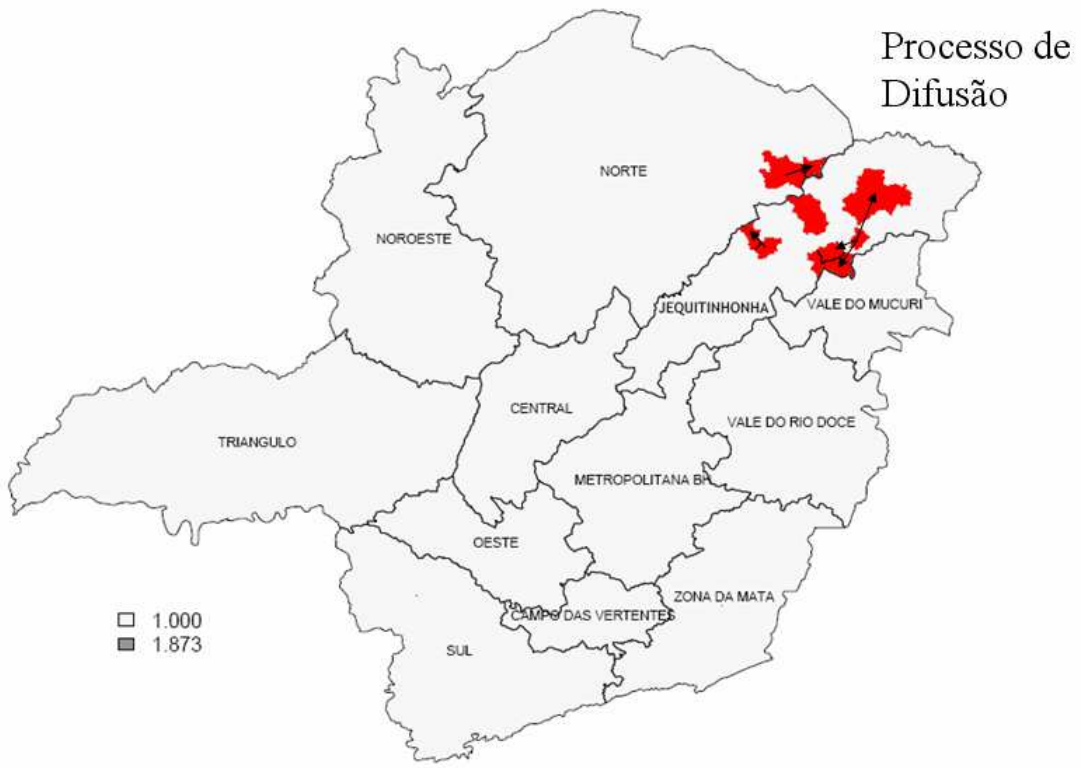


Figura 5.12: O terceiro estágio do processo de difusão que gera o cluster desconexo (*Patos*) com risco relativo igual a 1,873.

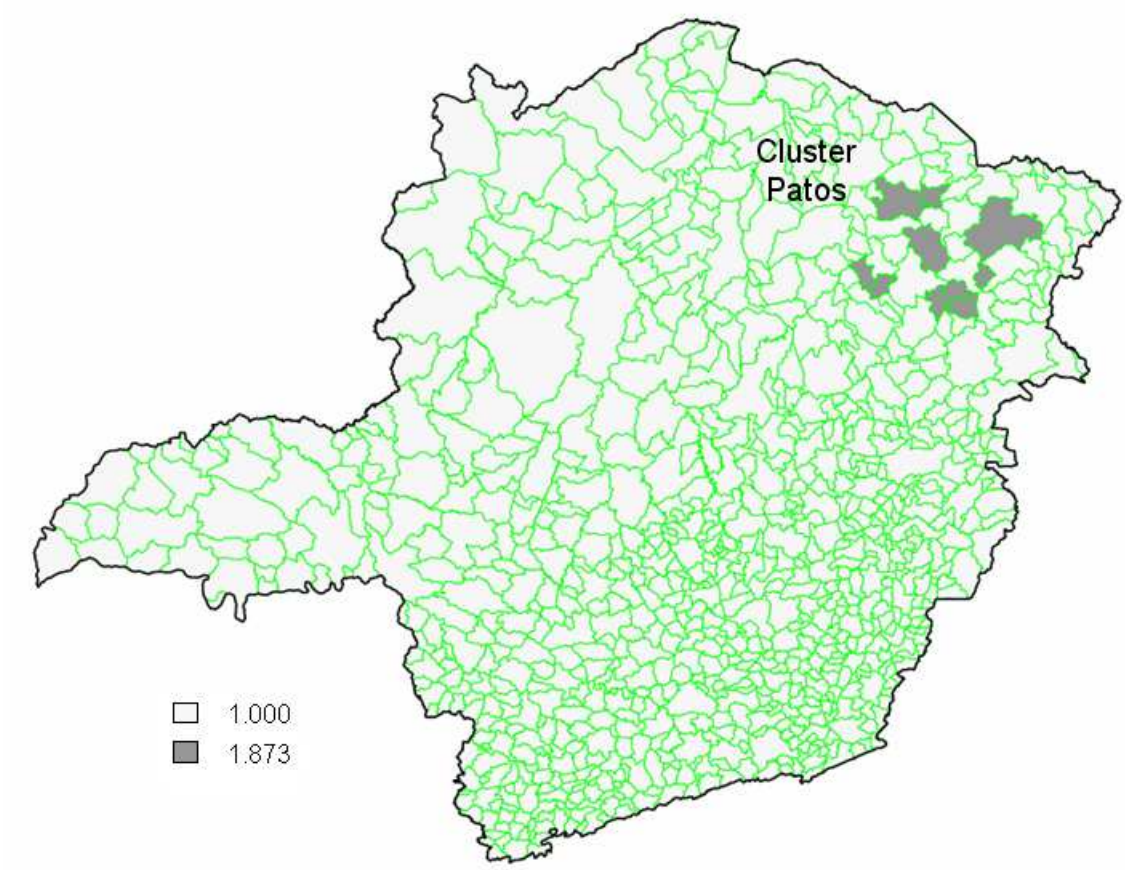


Figura 5.13: O cluster artificial desconexo (*Patos*) gerado pela difusão com risco relativo igual a 1,873.

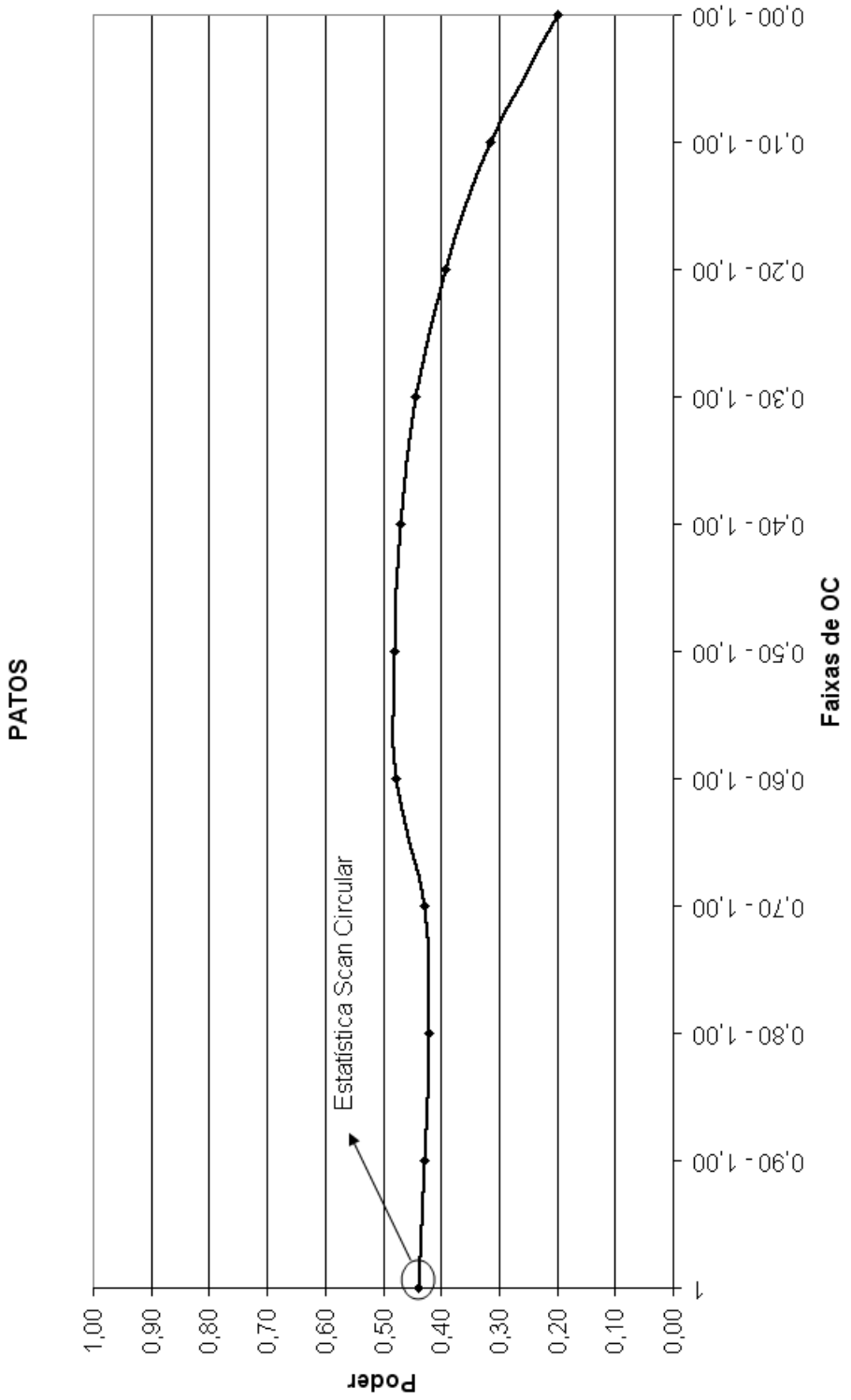


Figura 5.14: Poder do cluster artificial desconexo (*Patos*)

Tabela 5.4: *Poder da Estatística Multiseletiva para o cluster artificial desconexo (Patos).*

Ocupação Circular	Patos
1	0,4380
0,90 - 1,00	0,4300
0,80 - 1,00	0,4224
0,70 - 1,00	0,4284
0,60 - 1,00	0,4771
0,50 - 1,00	0,4812
0,40 - 1,00	0,4702
0,30 - 1,00	0,4444
0,20 - 1,00	0,3935
0,10 - 1,00	0,3150
0,00 - 1,00	0,1985

A Figura 5.15 mostra a união de pontos de 1000 conjuntos de Pareto obtidos por simulação de Monte Carlo sob hipótese nula, juntamente com a curva crítica de valor p igual a 0,05.

As Figuras seguintes 5.16 e 5.17 mostram a união de pontos de 1000 conjuntos de Pareto obtidos por simulação de Monte Carlo sob hipótese alternativa: a Figura 5.16 foi obtida utilizando-se o modelo alternativo do cluster *Circular* (Figura 5.5) e a Figura 5.17 foi obtida utilizando-se o modelo alternativo do cluster de gripe aviária da Figura 5.13.

Pode-se observar que no primeiro caso existe uma concentração maior de pontos à direita da isolinha crítica próximo à ordenada correspondente ao valor de ocupação circular 0,8, coerente com o valor de ocupação circular 0,88 do cluster *Circular* da Figura 5.5. No entanto, no gráfico da Figura 5.17 a concentração de pontos à direita da isolinha crítica aumenta em torno dos valores de OC entre 0,4 e 0,6, coerente com o valor de ocupação circular 0,4 do cluster circular da Figura 5.5.

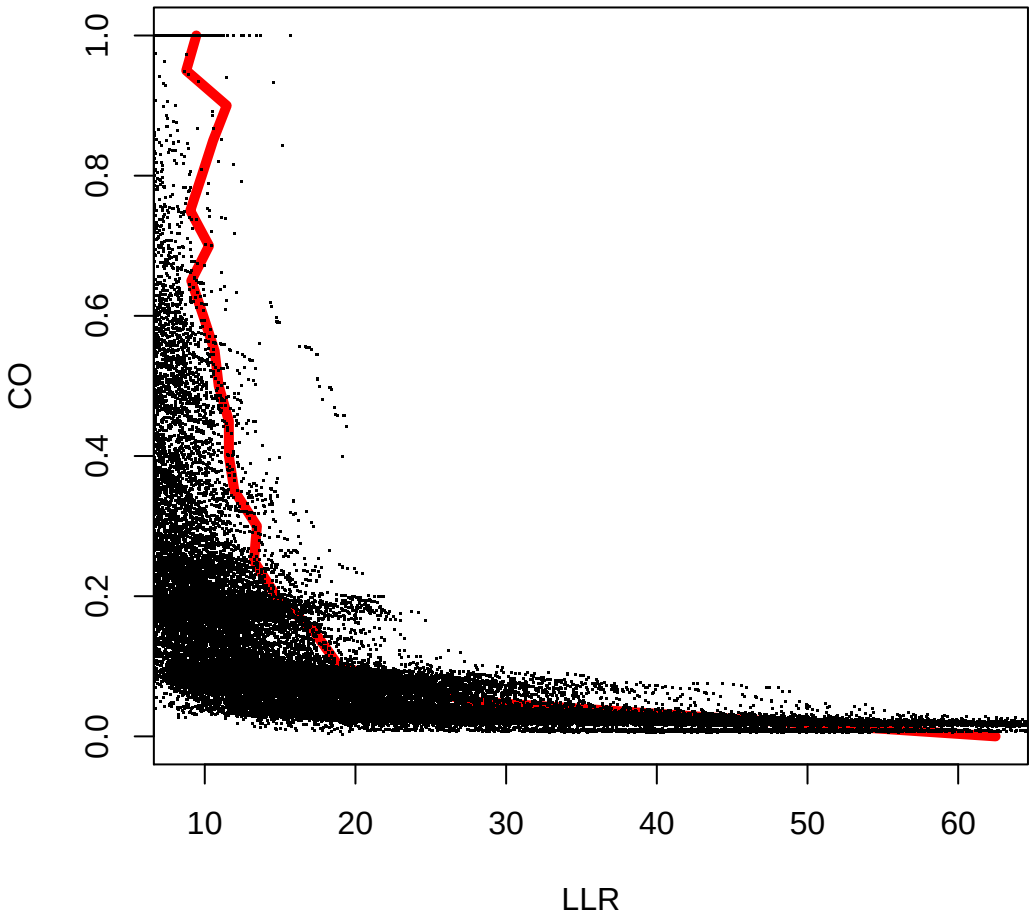


Figura 5.15: Isolinha de valor p = 0,05 para a hipóteses nula.

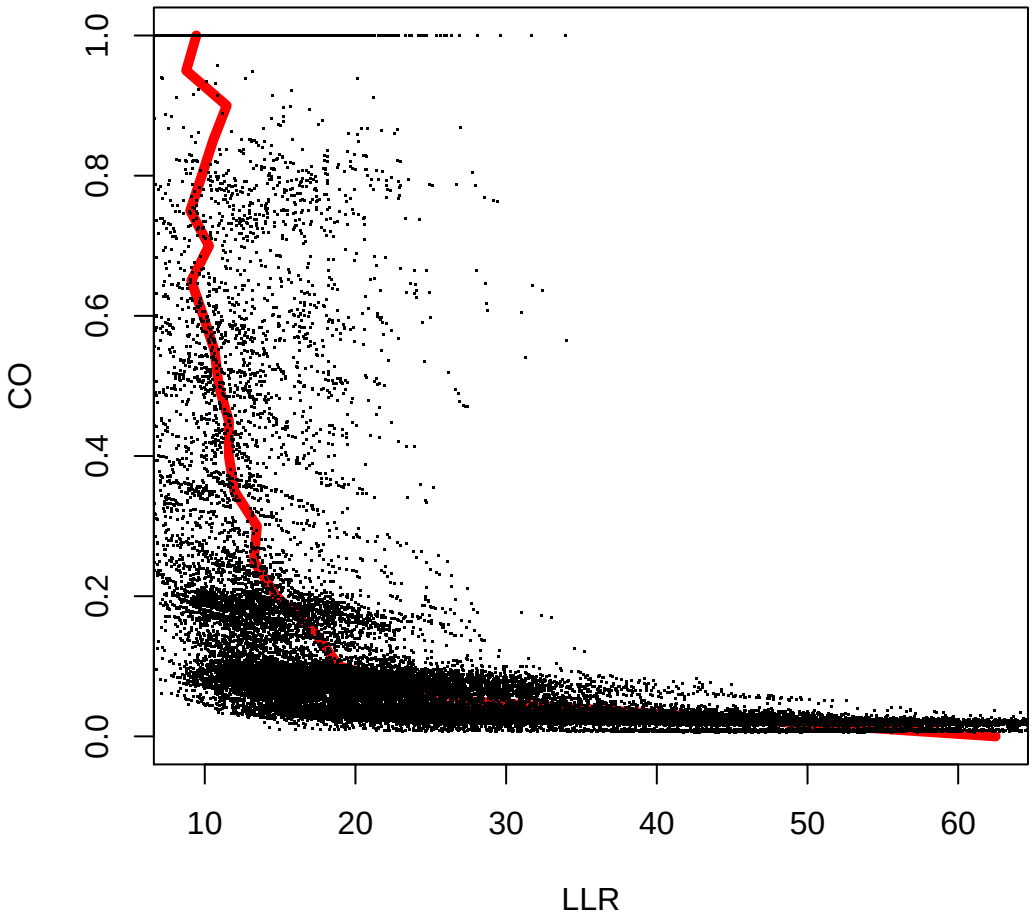


Figura 5.16: Poder de teste para o cluster artificial *Circular*.

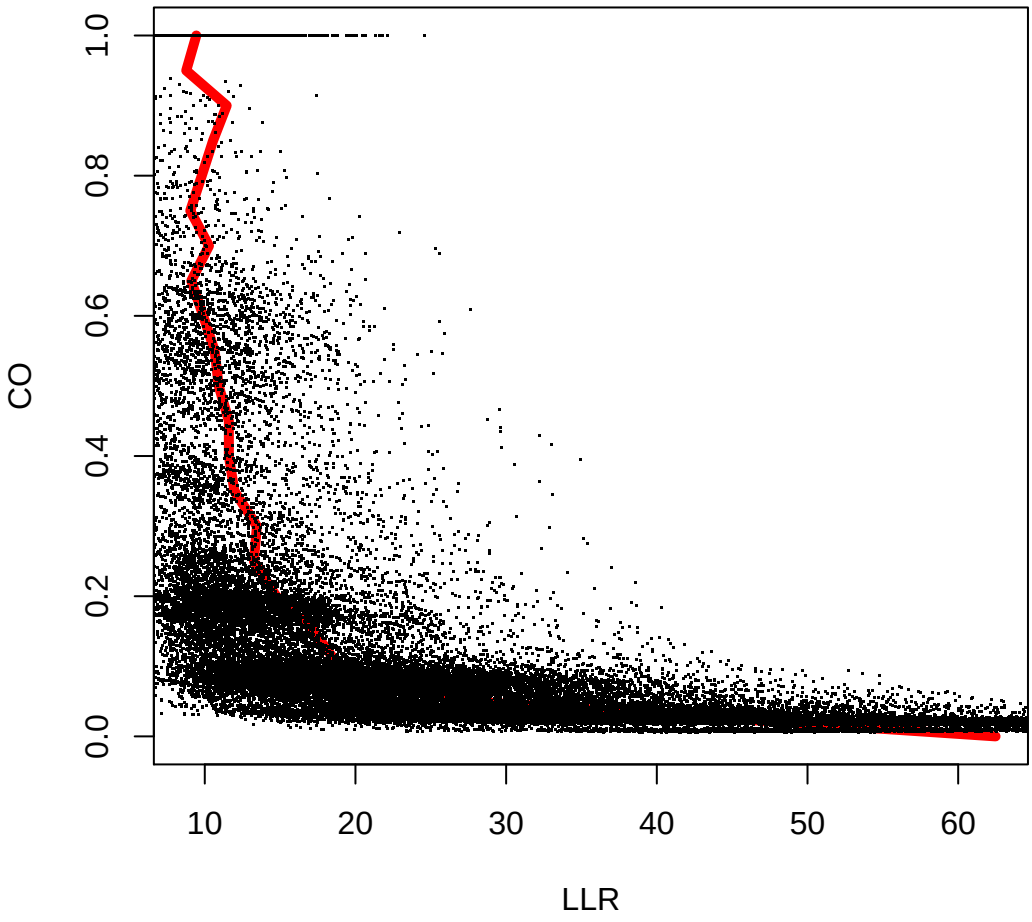


Figura 5.17: Poder de teste para o cluster de gripe aviária dos patos selvagens.

5.5 Aplicação: Homicídios nos municípios de Minas Gerais

Nesta seção vamos detectar clusters de homicídios no mapa de 853 municípios do estado de Minas Gerais, registrados no período de 1998-2002 (Moura, 2006). O mapa de taxa de homicídios aparece na Figura 5.18. Na Figura 5.19 mostra-se o conjunto de Pareto global formado pela união dos Paretos para cada um dos conjuntos seletivos com a igual a 0.002, 0.004, 0.008, 0.016, 0.032, 0.032, 0.064, 0.125, 0.250, 0.500 e 1.000, indicados com diferentes símbolos na parte inferior esquerda da Figura 5.19. Alguns clusters do conjunto de Pareto Global são indicados pelas setas, com seus respectivos mapas e apresentados na Figura 5.21. O conjunto de Pareto Global é visto com maior detalhe na Figura 5.20.

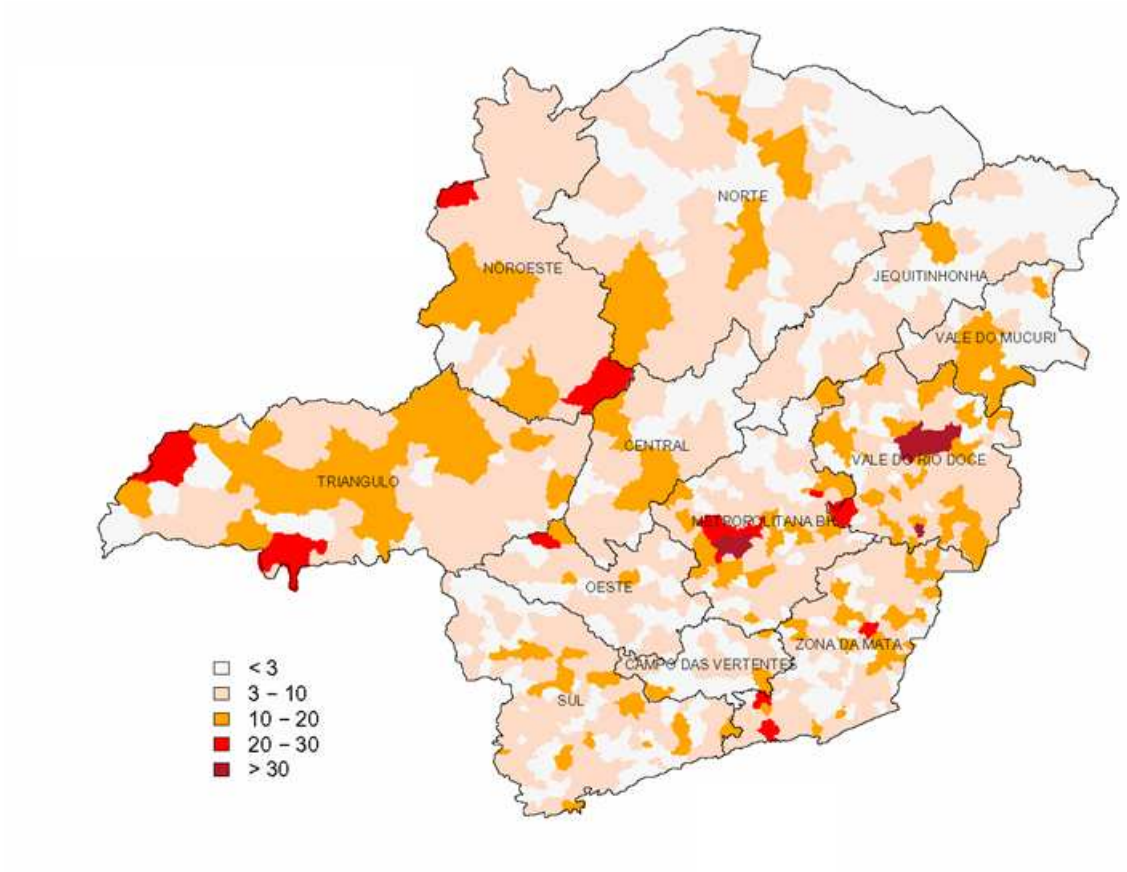


Figura 5.18: Taxa de homicídios (por 100 mil hab.) em Minas Gerais, 1998-2002.

A Figura 5.19 mostra o conjunto dos Paretos para cada um dos conjuntos seletivos consi-

derados. Cada ponto desse conjunto 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 a de um circulo. Um fato interessante a ser notado é que saltos bruscos no gráfico dos pontos do conjunto de Pareto geralmente indicam mudanças significativas nas estruturas dos clusters.

A Figura 5.21 apresenta os conjuntos Pareto-ótimo para cada um dos conjuntos seletivos abordados (valores de *a* desde 0,002 a 1,000). Para algumas soluções foi destacado o mapa do cluster referente a estas soluções. Com base nestes conjuntos, observa-se a existência de alguns níveis de clusterização que acompanham de forma satisfatória as desconexidades existentes no mapa de risco (Figura 5.18). Um fato importante a ser destacado é que saltos bruscos no conjunto Pareto-ótimo Global geralmente indicam mudanças significativas nas estruturas dos clusters. Tais soluções variam desde clusters com alto LLR e baixa OC, até clusters de baixo LLR e com alta OC. O primeiro mapa (inferior direito) corresponde à solução do Pareto-ótimo com maior LLR, mas, ao mesmo, tempo esse cluster possui um formato muito irregular, sendo totalmente desconexo. À medida que se diminui o valor do LLR observa-se que os clusters passam a ter um formato mais regular e a quantidade de regiões conexas tende a diminuir. Por outro lado, na medida em que a OC vai aumentando o cluster tende a se tornar mais conexo e seu LLR vai diminuindo. O último mapa (superior esquerdo) mostra o mapa de maior OC e menor LLR, e representa o cluster encontrado pelo método scan circular de Kulldorff.

Nesse exemplo, nota-se que o formato de um cluster passa de não conexo para conexo exatamente em um desses saltos. Tal salto é melhor visualizado na Figura 5.20 que apresenta o Pareto Global referente a todos os conjuntos seletivos avaliados. Os mapas mostrados na figura 5.21 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. Na Figura 5.21, 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) ob-

servamos 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 bastante 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, porém essas regiões têm baixa população e só aparecem se o risco for realmente muito alto. 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 está fortemente correlacionada à conexidade. O mapa do lado esquerdo e superior se refere à solução candidata que possui maior ocupação circular (ocupação circular igual a um) e menor razão de verossimilhança. Esse cluster seria o cluster encontrado pelo método scan circular.

Portanto, esses níveis de clusterização encontrados no Pareto-ótimo (Figura 5.21) acompanham de forma satisfatória as desconexidades existentes no mapa de risco (Figura 5.18).

Observe que os dois primeiros clusters nos mapas inclusos na parte superior esquerda da Figura 5.21 são os únicos clusters conexos representados. Os demais clusters são desconexos. O caso extremo é o mapa incluso do cluster na parte inferior direita, que foi obtido com o conjunto seletivo correspondente a $a = 0,064$, que é o cluster (desconexo) com o maior valor de LLR.

A curva de nível de menor valor p que intercepta o conjunto de Pareto é mostrada na Figura 5.20. Pode-se ver que nesse exemplo o cluster de maior significância dentre os clusters do Pareto Global é o cluster do segundo mapa incluso na parte superior da Figura 5.21 que é justamente o cluster conexo de maior LLR.

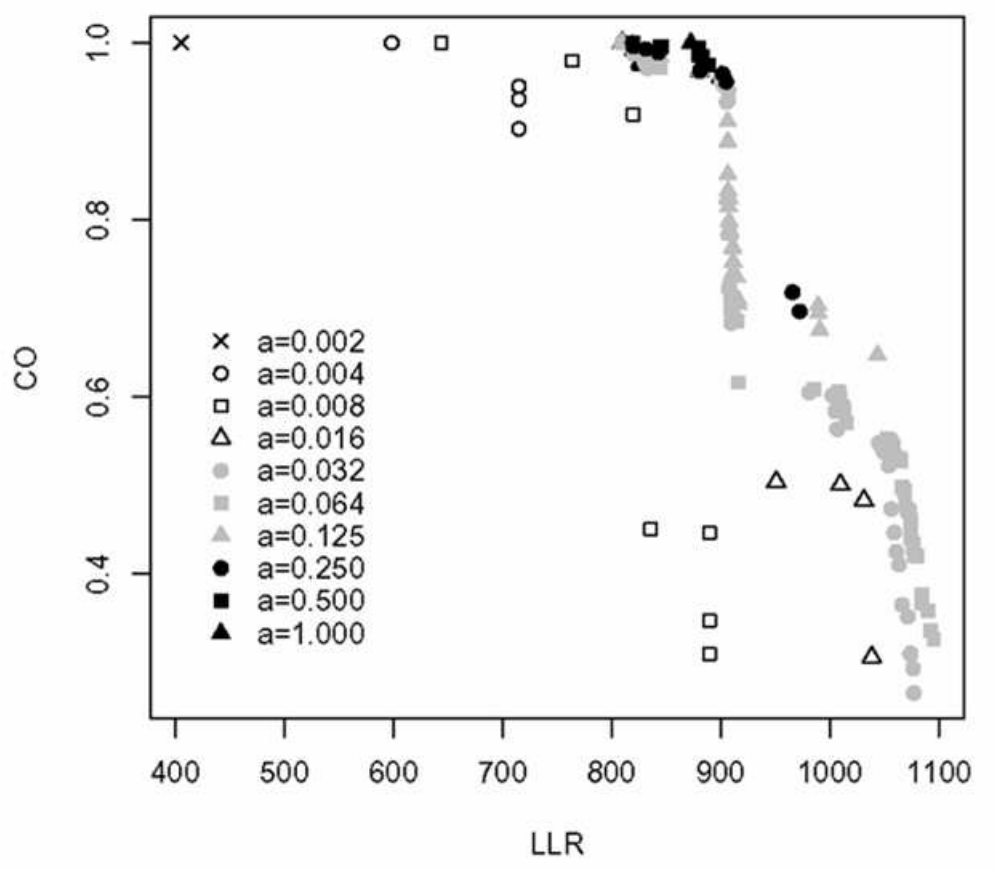


Figura 5.19: Conjuntos de Pareto para os diversos conjuntos seletivos.

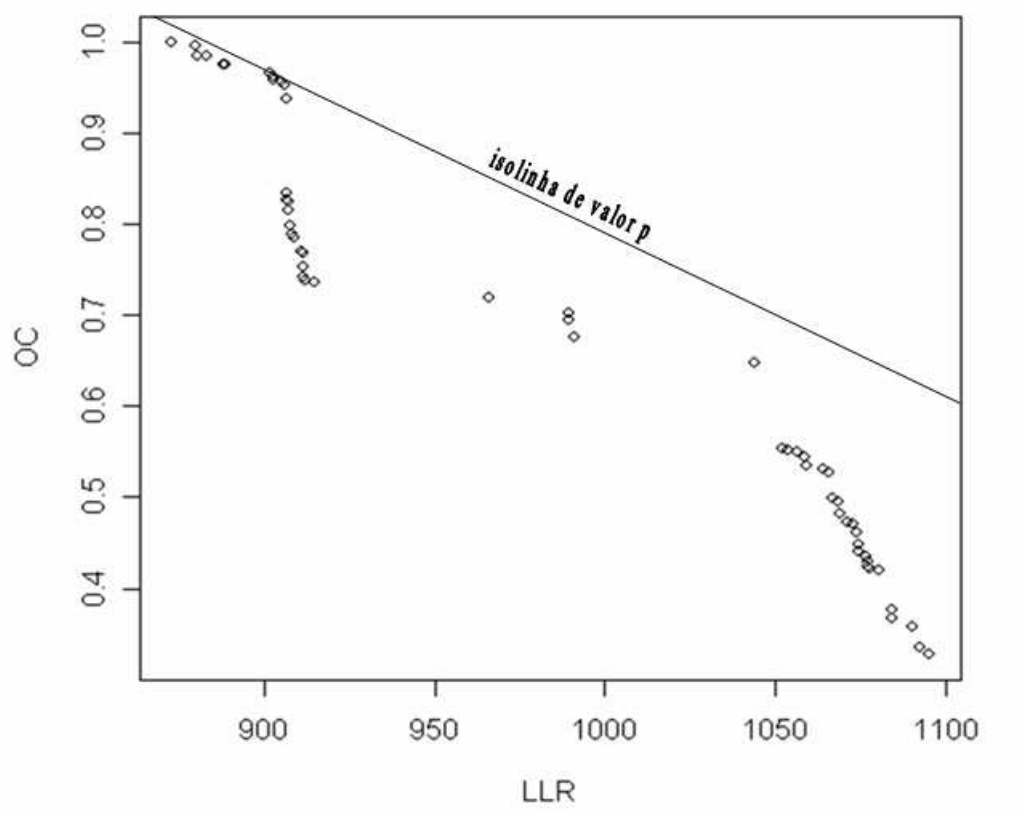


Figura 5.20: Conjunto de Pareto Global com a visualização da isolinha de valor p.

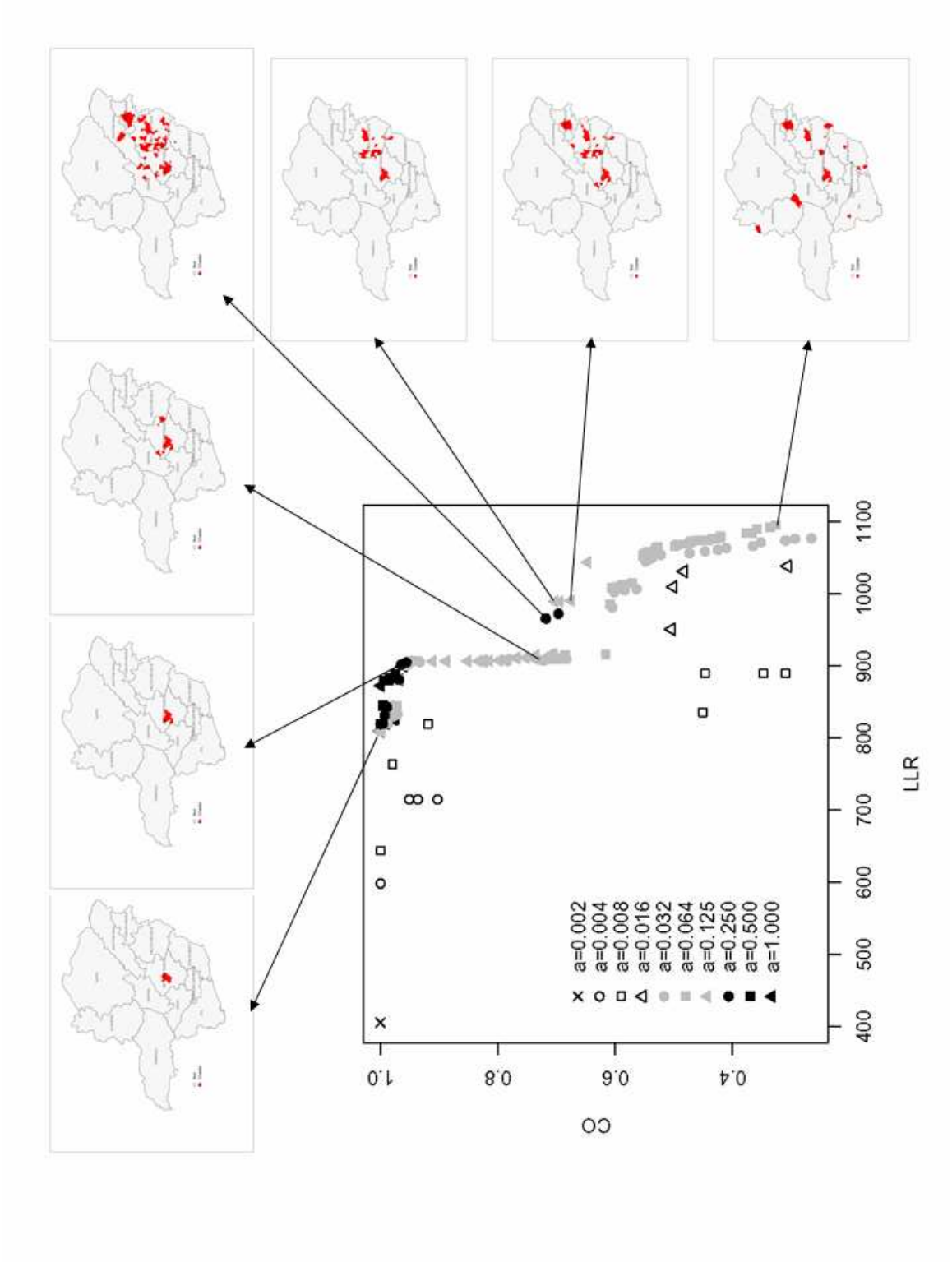


Figura 5.21: Conjunto de Pareto com a visualização de algumas soluções de clusters para o mapa.

Capítulo 6

Estatística Scan Desagregada

6.1 Introdução e Descrição do Método

Dentre os trabalhos pioneiros que propuseram técnicas para detectar clusters espaciais podemos citar Choynowski (1959), Openshaw et al. (1987), Cuzick and Edwards (1990), Turnbull et al. (1990) e Besag and Newell (1991). Alguns usaram o número de casos e outros sugeriram a taxa bruta como medida quantitativa para detectar os clusters no mapa. Kulldorff and Nagarwalla (1995) e Kulldorff (1997) propuseram uma Estatística que levasse em conta o número de casos e o risco relativo, dentro e fora da zona simultaneamente, que é a Estatística Scan Circular que define uma verossimilhança para avaliar os possíveis clusters no mapa.

É bastante comum um mapa possuir mais de um cluster com incidência destacada de casos ou de determinadas características estudadas. Mapas de riscos com vários clusters são melhor estudados através de soluções múltiplas com a finalidade de obtermos clusters de vários níveis de magnitude (primários, secundários, terciários, etc). A extensão de nosso trabalho se propõe a analisar a estatística scan circular como um problema de otimização multiobjetivo, isto é, considerando dois aspectos (o número de casos observados e o risco relativo), que avalia os vários níveis de clusterização existentes naturalmente em um mapa de doenças composto de m regiões.

A extensão proposta aqui, que é uma versão preliminar de nosso trabalho em andamento,

utiliza uma versão particular do algoritmo multiobjetivo sem usar o algoritmo genético. A estatística scan proposta neste trabalho é desagregada em duas partes: os casos observados e o risco relativo dentro de cada zona. Esses dois objetivos são maximizados numa avaliação realizada em cada uma das m regiões do mapa tomadas individualmente. A idéia consiste em considerarmos apenas as informações referentes à parte interna da zona e estudar com mais detalhe a zona primária e a sua localização no mapa e no conjunto de Pareto frente as demais soluções de menor significância.

Sejam c_z o número de casos na zona z e μ_z o número de casos esperado dentro da zona z. O conjunto de Pareto será baseado no par $(c_z, c_z/\mu_z)$.

O algoritmo do scan circular será utilizado, com apenas uma diferença: para cada cluster circular os valores do número de casos e do risco relativo de cada zona são plotados num gráfico $c_z \times c_z/\mu_z$. Em seguida, avaliamos a significância de cada solução de Pareto.

A análise das soluções é feita diferenciando-as por camadas de Pareto e tendo como parâmetro as curvas de níveis geradas pela função de verossimilhança da Estatística Scan. Estas estratégias são descritas nas seções 6.2 e 6.3.

6.2 Curvas de nível para o LLR(z)

A razão de verossimilhança já definida na seção 2.1 é expressa por

$$LR(z) = \left(\frac{c(z)}{\mu(z)}\right)^{c(z)} \left(\frac{C-c(z)}{C-\mu(z)}\right)^{C-c(z)}, \quad c(z) > \mu(z)$$

Sabemos que o risco relativo para uma zona z é definido como $RR(z) = \frac{c(z)}{\mu(z)}$, e que o número de casos esperados para esta zona é obtido por $\mu(z) = C\frac{n(z)}{N}$.

Desta forma, podemos olhar para a razão de verossimilhança da seguinte maneira,

$$LR(z) = (RR(z))^{c(z)} \left(\frac{C-c(z)}{C-\frac{c(z)}{RR(z)}}\right)^{C-c(z)}.$$

Assim, trata-se de uma função $f : \Re^2 \to \Re$ tal que $f(x,y) = y^x \left(\frac{C-x}{C-\frac{x}{y}}\right)^{C-x}$, em que para uma zona z, os casos são representados por x e o risco relativo por y. Se aplicarmos o logaritmo nesta função, temos que:

$$g(x,y) = xlog(y) + (C-x)log\left(\frac{C-x}{C-\frac{x}{y}}\right)$$

Portanto, as curvas de nível que são utilizadas neste trabalho são dadas pelos conjuntos $\{(x,y) : g(x,y) = B\}$ em que $g : \Re^2 \to \Re$ e B é uma constante arbitrária.

6.3 Camadas do Pareto-ótimo

Considere $\mathbb{P}$ o conjunto de pontos $(c_z, c_z/\mu_z)$ para todas as zonas circulares definidas pelo algoritmo scan circular para um mapa de m regiões com casos observados. Seja $\mathbb{P}_1$ o conjunto de Pareto obtido de $\mathbb{P}$, e seja $\mathbb{P}_2$ o conjunto de Pareto de $\mathbb{P} - \mathbb{P}_1$, $\mathbb{P}_3$ o conjunto de Pareto de $\mathbb{P} - \mathbb{P}_1 - \mathbb{P}_2$, etc. O conjunto $\mathbb{P}_k$ é chamado de *k-ésima camada de Pareto de* $\mathbb{P}$.

6.4 Resultados Preliminares

Nesta seção apresentamos os resultados de uma aplicação para dados de mortalidade causada por bronquite para as microregiões de Minas Gerais durante os anos de 1998 a 2002. Percebe-se claramente uma incidência maior de bronquite no sul do estado.

Na Figura 6.4 plotamos os pontos $(c_z, c_z/\mu_z)$ para todas as zonas do scan circular juntamente com as curvas de nível de log da verossimilhança. A Figura 6.2 mostra apenas os pontos da primeira camada de Pareto com as curvas de nível do log da verossimilhança. Os pontos para a segunda camada de Pareto são apresentados na Figura 6.3. Pode-se observar que vários clusters com valores de LLR próximos entre si ficam situados em locais distantes no gráfico. Além disso, joelhos na estrutura do conjunto do Pareto podem indicar a presença de clusters com características notáveis dentro do mapa. Tais clusters se destacariam como tendo mais casos em relação aos vizinhos à esquerda no Pareto e como tendo maior risco

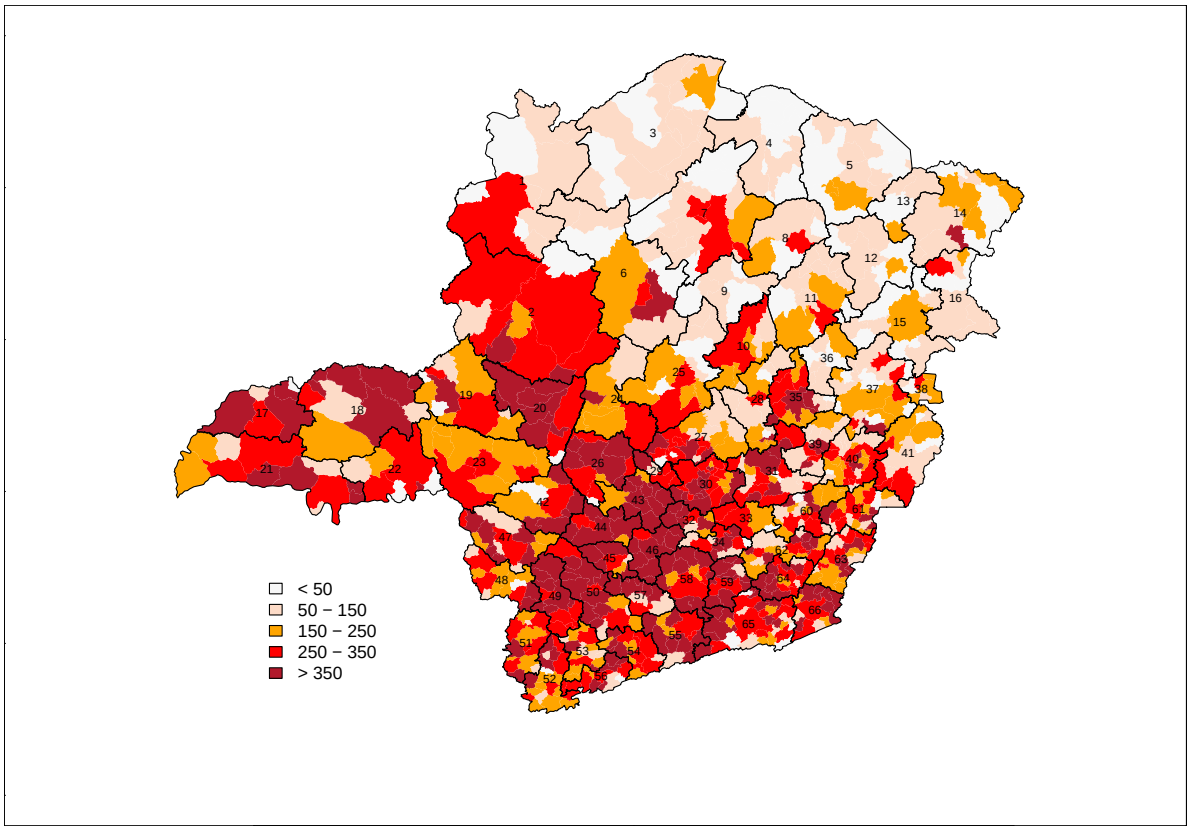


Figura 6.1: Taxa de mortalidade causada por bronquite (por 100 mil habitantes) para as microregiões de Minas Gerais, 1998-2002.

relativo em relação aos seus vizinhos à direita no Pareto. Alguns desses clusters estão indicados nas Figuras 6.5 e 6.6. A Figura 6.6(c) representa o cluster com maior LLR dentre todos aqueles que estão no conjunto Pareto-ótimo.

Neste exemplo, utilizamos as curvas de nível para valores de B iguais a 25, 50, 75, 100, 120, 140, 160, 180, 200 e 220. Estas constantes são escolhidas de forma que abranja todo o conjunto de Pareto. As Figuras 6.2 e 6.3 apresentam a primeira e segunda camada do Pareto-ótimo.

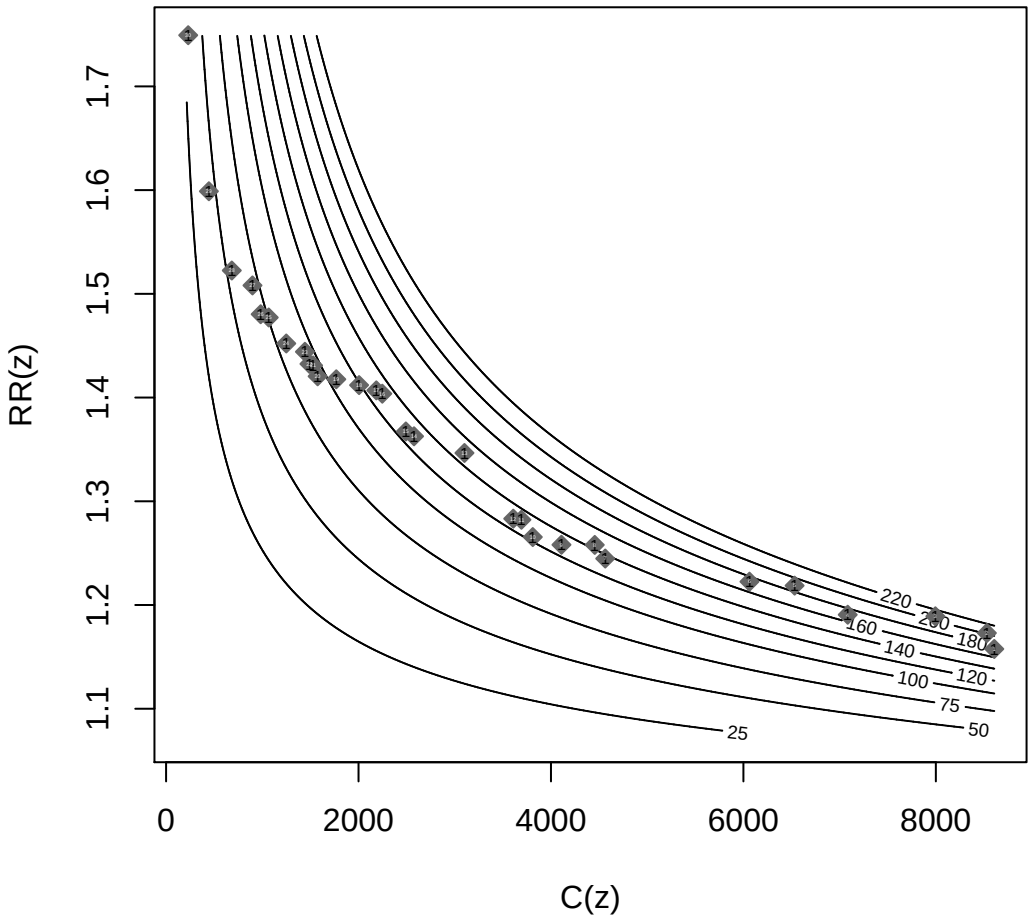


Figura 6.2: Avaliando a significância dos clusters considerando a primeira camada do conjunto Pareto-ótimo através das curvas de níveis do LLR da Estatística Scan de Kulldorff.

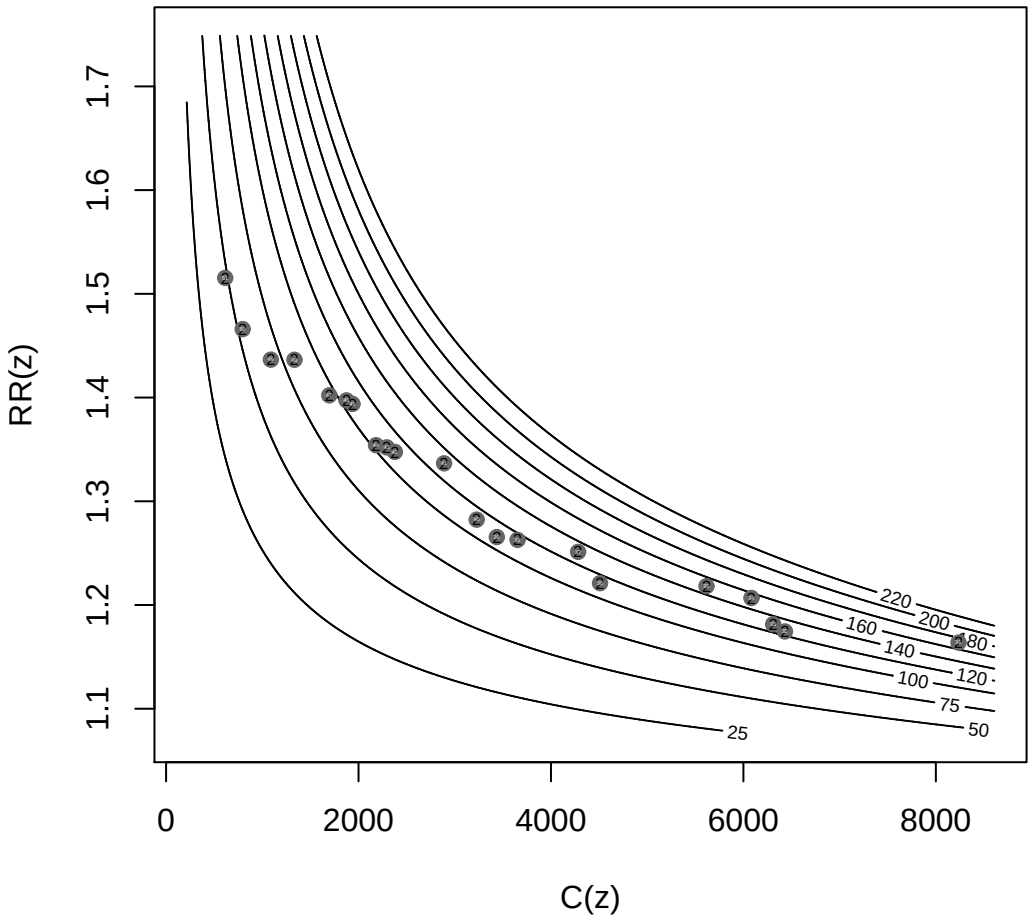


Figura 6.3: Avaliando a significância dos clusters considerando a segunda camada do conjunto Pareto-ótimo através das curvas de níveis do LLR da Estatística Scan de Kulldorff.

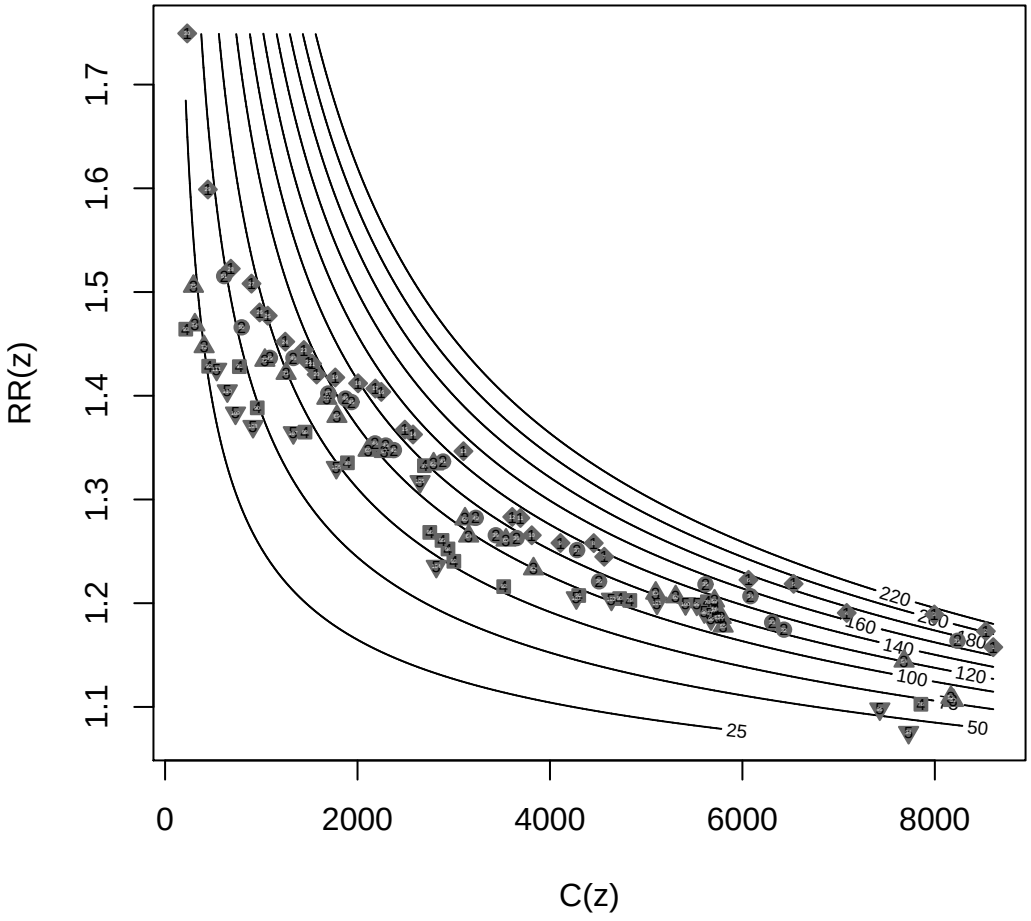


Figura 6.4: Avaliando a significância dos clusters diante de todas as camadas do conjunto Pareto-ótimo através das curvas de níveis do LLR da Estatística Scan de Kulldorff.

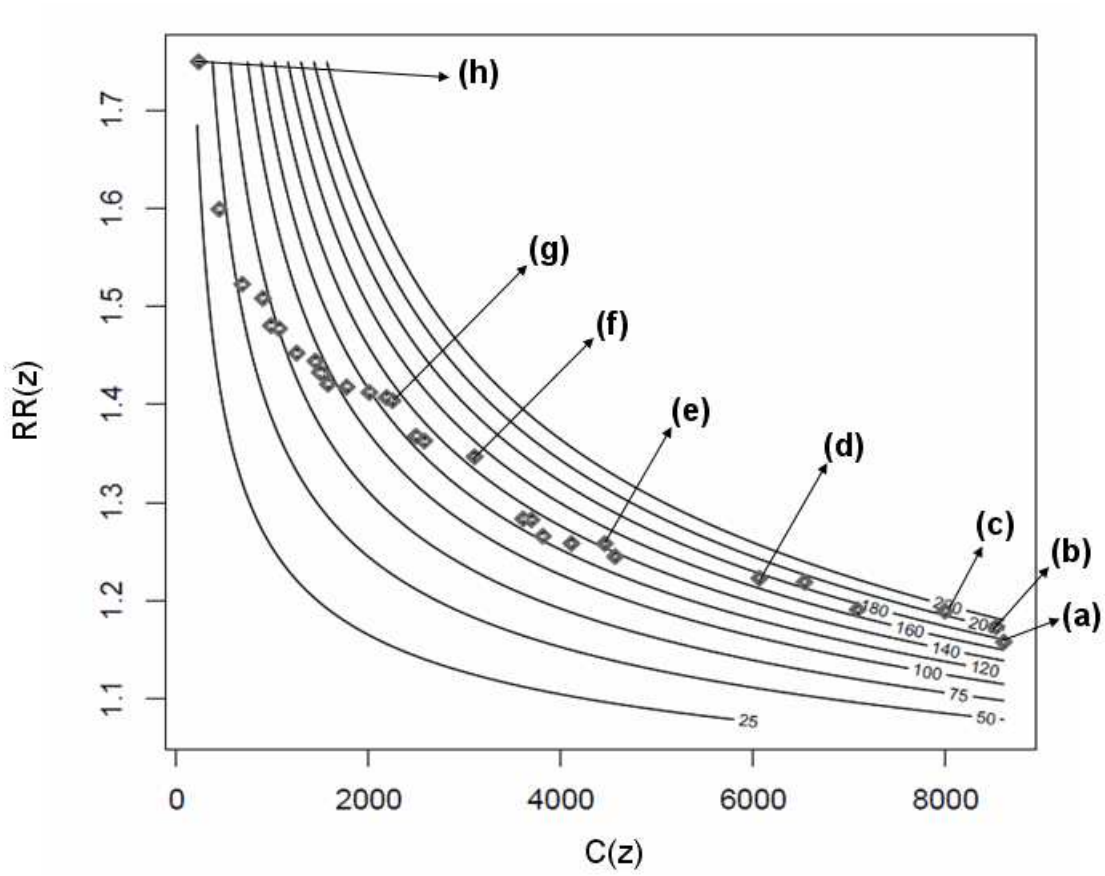
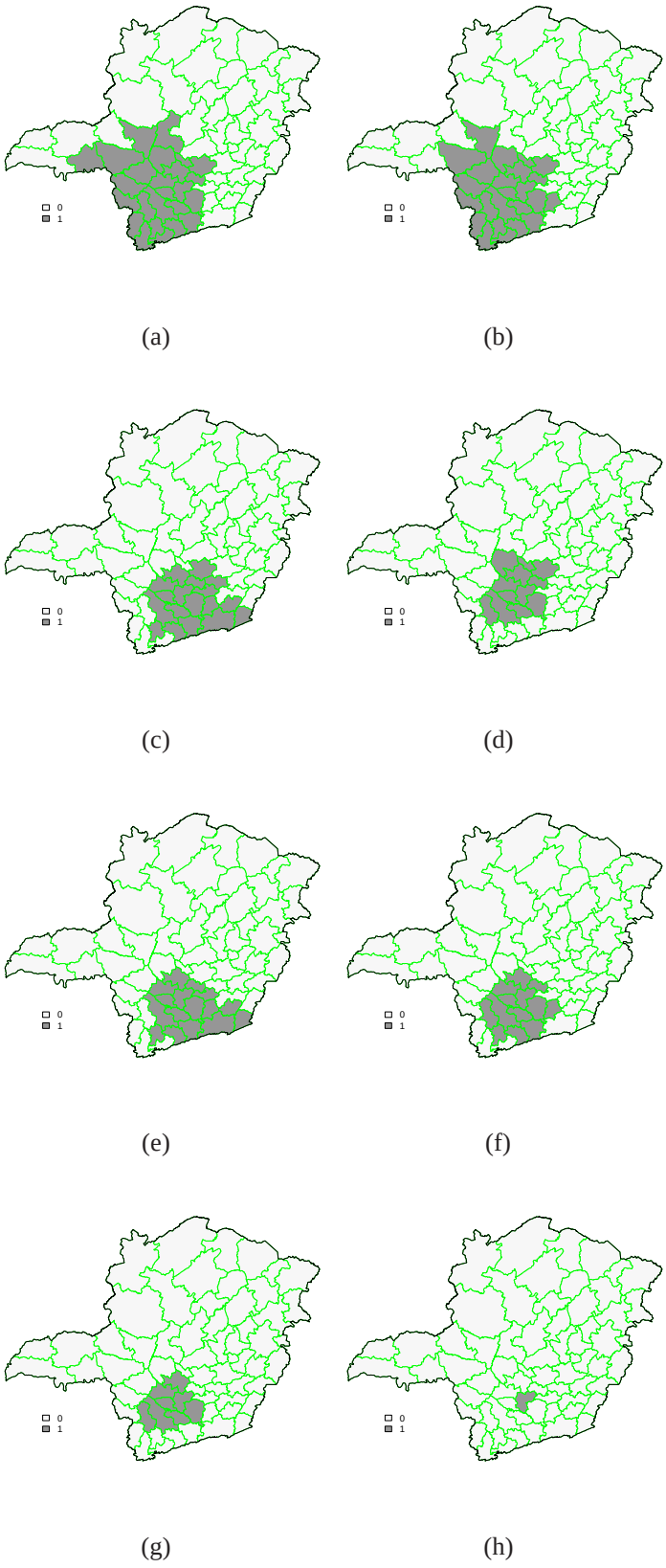


Figura 6.5: Visualizando alguns clusters no conjunto Pareto-ótimo.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figura 6.6: Alguns clusters encontrados pela estatística scan desagregada para a primeira camada ou Pareto-ótimo.

Capítulo 7

Considerações Finais

7.1 Conclusões

A estratégia desenvolvida nesta tese de reforçar o grafo de adjacência, construindo muitos cenários diferentes sugeridos por fatores ambientais, é uma alternativa aos métodos focados já existentes. Ao invés de forçar a ocorrência de um cluster pré-especificado, como no método focado, nós criamos as circunstâncias em que sua probabilidade de ocorrência é maior em alguns lugares selecionados no mapa. Desta forma, pode-se verificar o aumento ou diminuição da significância do cluster dependendo do cenário de precipitação existente naquela região. Acreditamos que esta nova ferramente seja útil para testar hipóteses a respeito da influência de fatores ambientais na formação de cluster de determinadas doenças. No nosso estudo, utilizamos dados ambientais relacionados a chuva para construir cenários distintos para o reforço do grafo. Com as simulações, observamos que usar o algoritmo multiobjetivo fornece meios de encontrarmos mais soluções que não se limitem à solução primária, e ainda que estas não sejam apenas soluções regulares. Mostrou-se que a detecção do cluster foi influenciada pelo reforço aplicado a cada cenário específico, sugerindo que o fator ambiental selecionado (chuva) possa ser significativamente relacionado à ocorrência de clusters espaciais de malária na Amazônia brasileira. Os testes numéricos mostram que o poder da deteção é aumentado quando o reforço no grafo coincide com o cluster real, e permanece o mesmo quando o reforço ocorre em outras posições do mapa. A comparação com o teste de

Mantel mostrou que o nosso método de reforçar o grafo é capaz de encontrar a localização dos clusters espaciais mais significativos quando este se correlaciona de maneira mais forte com o fator ambiental que foi utilizado para alterar a estrutura de vizinhança.

A Estatística Scan Multiseletiva mostrou-se muito eficiente na detecção de clusters de formato regular e irregular, controlando ou evitando as superestimações e subestimações. Esta proposta se baseia em conceitos simples e generaliza a definição de zona que agora abrange situações regulares e irregulares, conexas e desconexas. Observou-se também que saltos significativos no conjunto de Pareto resultam em mudanças significativas no formato do cluster, inclusive modificando o grau de conexidade da solução.

A Estatística Scan Desagregada poderá explicar mais detalhes sobre os níveis de clusterização existentes em um conjunto de Pareto, bem como estudar as relações entre as camadas do Pareto e seus significados no mapa. Nesta proposta contribuímos com a utilização das curvas de nível da LLR para auxiliar na escolha de uma das soluções ótimas. Utilizou-se também as camadas do conjunto de soluções eficientes para analisarmos melhor a clusterização.

As técnicas multiobjetivo parecem ser promissoras para detecção e inferência de clusters espaciais, principalmente devido à possibilidade de visualizarmos a estrutura multi-cluster das diversas soluções. Além disso, o cálculo da significância bi-objetivo de cada solução se mostrou uma ferramenta eficaz para a escolha do melhor cluster.

7.2 Trabalhos Futuros

Um trabalho que pode ser estendido é comparar estas três extensões com a estatística Scan Circular de Kulldorff num mesmo conjunto de dados, como por exemplo, no conjunto de dados da Nova Inglaterra que já foi bastante explorado por diversos métodos. Tal comparação seria realizada usando as medidas de sensibilidade e valor de predição positiva, e não apenas o poder de detecção, bem como o erro tipo II.

O estudo de caso descrito para a Estatística Scan Modificada neste trabalho foi limitado à apenas uma característica ambiental (chuva), e uma extensão pode ser implementada para

uma situação multivariada levando em conta diversos fatores ambientais, por exemplo.

Em um trabalho futuro estudaremos os grafos de semelhança entre zonas candidatas a cluster pela Estatística Desagregada utilizando-se a Estatística Scan Multiseletiva ao invés da Scan Circular.

7.3 Produção Bibliográfica durante o Doutorado

Artigos em Periódicos

- DUCZMAL, L.; TAVARES, R.; PATIL, Ganapati; CANCADO, André L. F.. Testing spatial cluster occurrence in maps equipped with environmentally defined structures. Environmental and Ecological Statistics, 2009. (submetido após pequenas ressalvas).

Conferências

- DUCZMAL, L. H. ; TAVARES, R. ; CANÇADO, A. L. F. ; PATIL, G. P. . Finding Spatial Clusters in Maps Equipped with Environmentally Defined Structures with Disease Policy Case Studies. In: Joint Statistical Meeting, 2009, Washington. Proceedings of the 2009 Joint Statistical Meeting, 2009.

- DUCZMAL, L. H. ; CANÇADO, A. L. F. ; TAKAHASHI, R. H. C. ; FERREIRA NETO, S. J. ; MOURA, F. R. ; DUARTE, A. R. ; TAVARES, R. . Multi-Objective Spatial Scans for Disease Cluster Detection. In: International Workshop in Applied Probability, 2008, Compiegne. Proceedings of the International Workshop in Applied Probability 2008, 2008.

- MOURA, Flávio ; DUCZMAL, L. ; TAVARES, Ricardo ; TAKAHASHI, R. H. C.. Exploring multi-cluster structures with the multi-objective circular scan. In: Syndromic Surveillance Conference, 2007. Advances in Disease Surveillance. v. 2. p. 48-48.

- DUCZMAL, L. ; PATIL, G. P. ; CANÇADO, André Luiz Fernandes ; TAVARES, Ricardo . Detection of spatial clusters in maps equipped with environmentally defined

structures. In: 7th Annual International Conference on Digital Government Research, 2006, San Diego, CA. ACM Proceedings - Conference on Digital Government Research, 2006.

Capítulo de Livro

- DUCZMAL, L. ; DUARTE, A. Ribeiro ; TAVARES, Ricardo . Extensions of the scan statistic for the detection and inference of spatial clusters. In: Joseph Glaz; Vladimir Pozdnyakov; Sylvan Wallenstein. (Org.). Scan Statistics - Methods and Applications. 1 ed. Hamilton: Springer, 2009, v. 1, p. 157-182.

Artigos em Desenvolvimento

- MOURA, Flávio; DUCZMAL, L.; TAVARES, Ricardo; TAKAHASHI, R. H. C.. Exploring multi-cluster structures with the multi-objective circular scan, 2009.

- DUCZMAL, L.; TAVARES, Ricardo. Disaggregated spatial scan statistic: a tool for distinguishing distinct “families” of clusters, 2009.

Referências Bibliográficas

Abrams, A., Kulldorff, M., and Kleinman, K. (2006). Empirical/assymptotic p-values for monte carlo-based hypothesis testing: an application to cluster detection using the scan statistic. *Advances in Disease Surveillance*, 1:1–1.

Assuncao, R. M., Costa, M. A., Tavares, A., and Neto, S. J. F. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25:723–742.

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*. New York, John Wiley & Sons.

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

Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., and Moore, A. W. (2005). Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, 38:99–113.

Cancado, A. L. F. (2009). *Detecção de clusters espaciais através de otimização multiobjetivo*. Doutorado em engenharia elétrica, Universidade Federal de Minas Gerais, Belo Horizonte / MG.

Carrano, E. G. (2007). *Algoritmos Evolucionários Eficientes para Otimização de Redes*. Doutorado em engenharia elétrica, Universidade Federal de Minas Gerais, Belo Horizonte / MG.

Carrano, E. G., Soares, L. A. E., Takahashi, R. H. C., Saldanha, R. R., and Neto, O. M. (2006). Electric distribution network multiobjective design using a problem-specific genetic algorithm. *IEEE Transactions on Power Delivery*, 21(2):995–1005.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. California, Duxbury.

Chankong, V. and Haimes, Y. Y. (1983). *Multiobjective Decision Making: Theory and Methodology*. North-Holland.

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

Coello, C. A. C. (1996). *An Empirical Study of Evolutionary Techniques for Multiobjective Optimization in Engineering Design*. Phd., department of computer science, Tulane University, New Orleans, Louisiana.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. London, Springer-Verlag.

Conley, J., G. M. and MacGill, J. (2005). A genetic approach to detecting clusters in point data sets. *Geographical Analysis*, 37:286–314.

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society*, 52 (B):73–104.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*. New York, John Wiley & Sons.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multi-objective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6 (2):182–197.

Duczmal, L. H. and Assuncao, R. M. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysiss*, 45:269–286.

Duczmal, L. H. and Buckeridge, D. L. (2006). A workflow spatial scan statistic. *Statistics in Medicine*, 25:743–754.

Duczmal, L. H., Cancado, A. L. F., and Takahashi, R. H. C. (2008). Delineation of irregularly shaped disease clusters through multi-objective optimization. *Journal of Computational & Graphical Statistics*, 17:243–262.

Duczmal, L. H., Cancado, A. L. F., Takahashi, R. H. C., and Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, 52:43–52.

Duczmal, L. H., Duarte, A. R., and Tavares, R. (2009). *Extensions of the scan statistic for the detection and inference of spatial clusters*, chapter 1, pages 1–24. Springer, Hamilton, 1 edition.

Duczmal, L. H., Kulldorff, M., and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters. *Journal of Computational & Graphical Statistics*, 15:428–442.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28:181–187.

Elliott, P., Martuzzi, M., and Shaddick, G. (1995). Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research*, 4:137–159.

Fonseca, C. and Fleming, P. (1993). Genetic algorithms for multi-objective optimization: Formulation, discussion and generalization. *In Proceedings of the Fifth International Conference on Genetic Algorithms, San Mateo, California (pp. 416-423). Morgan Kauffman Publishers*, pages 416–423.

Fonseca, C. and Fleming, P. (1995). An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation*, 3(1):1–16.

Glaz, J., Naus, J., and Wallestein, S. (2001). *Scan Statistics*. In Springer Series in Statistics, Springer, New York.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.

Goslee, S. C. and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22(7):1–19.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Harbor: University of Michigan Press, Michigan.

Iyengar, V. S. (2004). *Space-time Clusters with flexible shapes*. IBM Research Report RC23398 (W0408-068).

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*. 2nd edn. Wiley Series in Probability and Statistics, vol. 2., John Wiley & Sons.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations, and applications. *In Scan Statistics and Applications*, pages 303–322.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. M., and Mostashari, F. (2005). A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3):216–224.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, 25:3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K., and Platt, R. (2007). Multivariate scan statistics for disease surveillance, 26:1824-1833.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14:799–810.

Kulldorff, M., Tango, T., and Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42:665–684.

Laplace, C., Berg, M., and Lai, H. (2007). *Bloodshed Dev-C++*. GNU - General Public License.

Lawson, A. (2001). *Statistical methods in spatial epidemiology*, pages 197–206. Wiley.

Lawson, A., Biggeri, A., Bohning, D., Lesare, E., Viel, J. F., and Bertollini, R. (2000). *Disease Mapping and Risk Assessment for Public Health*. Wiley, London.

Legendre, P. and Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, 80:107–138.

Lima, M. (2004). Avaliação do poder do teste da estatística scan para múltiplos clusters. Mestrado em estatística, Universidade Federal de Minas Gerais, Belo Horizonte / MG.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220.

Modarres, R. and Patil, G. P. (2007). Hotspot detection with bivariate data. volume 137(11). Syndromic Surveillance Conference, Celebration of the Centennial of the Birth of Samarendra Nath Roy (1906-1964).

Moore, D. A. and Carpenter, T. E. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic Reviews*, 21:143–161.

Moura, F. d. R. (2006). Detecção de clusters espaciais via algoritmo scan circular seletivo. Mestrado em estatística, Universidade Federal de Minas Gerais, Belo Horizonte / MG.

Moura, F. d. R., Duczmal, L. H., Tavares, R., and Takahashi, R. H. C. (2007). Exploring multi-cluster structures with the multi-objective circular scan. volume 2, pages 48–48, Indianapólis. Syndromic Surveillance Conference, Advances in Disease Surveillance.

Naus, J. I. (1965). Clustering of random points in two dimensions. *Biometrika*, 52:263–267.

Neill, D. and Moore, A. (2006). Methods for detecting spatial and spatio-temporal clusters. *Handbook of Biosurveillance*, pages 243–254.

Nepomuceno, E. G., Takahashi, R. H. C., Amaral, G. F. V., and Aguirre, L. A. (2003). Nonlinear identification using prior knowledge of fixed points: a multiobjective approach. *International Journal of Bifurcation and Chaos*, 13(5):1229–1246.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark i geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1:335–358.

Patil, G. P., Modarres, R., Myers, W. L., and Patankar, P. (2006). Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics. *Environmental and Ecological Statistics*, 13:365–377.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11:183–197.

R Development Core Team (2007). R: A language and environment for statistical computing. `http://www.r-project.org`.

Sahajpal, R., Ramaraju, G. V., and Bhatt, V. (2004). Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. *International Conference on Intelligent Sensing and Information Processing*.

Schaffer, J. D. (1984). *Multiple Objective Optimization with Vector Evaluated Genetic Algorithms*. Ph.d., Vanderbilt University.

Smith, D. L., Lucey, B., Waller, L. A., Childs, J. E., and Real, L. A. (2002). Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *PNAS*, 99(6):3668–3672.

Srinivas, N. and Deb, K. (1995). Multiobjective function optimization using nondominated sorting genetic algorithms. *Evol. Comput.*, 2 (3):221–248.

Takahashi, K., Kulldorff, M., Tango, T., and Yie, K. (2007). A flexible space-time scan statistic for disease outbreak detection and monitoring. *Advances in Disease Surveillance*, 2:70.

Takahashi, R. H. C., Palhares, R. M., Dutra, D. A., and Gonçalves, L. P. S. (2004). Estimation of pareto sets in the mixed h2/h-infinity control problems. *International Journal of Systems Science*, 35(1):55–67.

Takahashi, R. H. C., Vasconcelos, J. A., Ramirez, J. A., and Krahenbul, L. (2003). A multiobjective methodology for evaluating genetic operators. *IEEE Transactions on Magnetics*, 39(3):1321–1324.

Tango, T. (2007). A spatial scan statistic scanning only the regions with elevated risk. volume 4, page 117, Indianapólis. Syndromic Surveillance Conference, Advances in Disease Surveillance.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11.

TerraSeer (2004). `http://www.terraseer.com`.

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*, 132:S136–S143.

Waller, L. A. and Jacquez, G. M. (2000). Disease models implicit in statistical tests of disease clustering. *Epidemiology*, 6:584–590.