UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
INSTITUTO DE CIÊNCIAS EXATAS - ICEx
DEPARTAMENTO DE ESTATÍSTICA

Tese de Doutorado:

# INFERÊNCIA MULTIVARIADA DE *CLUSTERS* ESPACIAIS

Fábio Rocha da Silva
Orientador: Prof. Dr. Luiz Henrique Duczmal
Coorientador: Prof. Dr. Alexandre Celestino Leite Almeida

**FÁBIO ROCHA DA SILVA**

# INFERÊNCIA MULTIVARIADA DE *CLUSTERS* ESPACIAIS

Tese apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial a obtenção do título de Doutor em Estatística.

Orientador: Prof. Dr. Luiz H. Duczmal
Coorientador: Prof. Dr. Alexandre Celestino Leite Almeida

Universidade Federal de Minas Gerais
Belo Horizonte
Abril de 2015

Dedico este trabalho à memoria de minha mãe Maria Rocha da Silva.

# Agradecimentos

Agradeço primeiramente a Deus, pela proteção constante e por todas as maravilhosas oportunidades concedidas fundamentais para esta conquista.

Quero deixar expresso o meu agradecimento a alguém que já partiu, à minha amada mãe Maria Rocha da Silva, é a ela que devo esta minha força de vontade de lutar por um futuro melhor, de me realizar pessoalmente e profissionalmente. Tive sempre o seu ombro amigo para contar as minhas dúvidas, os meus receios e as minhas mágoas. Agradeço-lhe, com todo o meu coração, a sua disponibilidade, o seu carinho e a sua palavra de incentivo nos momentos de desmotivação, em que ela sempre dizia : "Você vai conseguir". Se de fato consegui, foi graças à sua crença inabalável em mim, e quando alguém acredita muito em nós, nós próprios começamos a acreditar também.

Sem o suporte emocional e a compreensão inesgotável da minha namorada, Dayanna Aguiar, não me teria sido possível conciliar o trabalho, o doutorado e a vida familiar/conjugal. Enquanto me mantive trabalhando, investigando e escrevendo esta tese de doutorado, nunca ouvi uma única palavra de descontentamento da sua parte e isso, por si só, já é um apoio colossal. Ela esteve sempre do meu lado, mesmo quando a angústia se apoderava de mim e o cansaço falava mais alto, nesses momentos ela mostrou-se sempre compreensiva e deu-me força para continuar a minha caminhada.

Dedico também esta tese a pessoas muito importantes para mim, meu pai Manoel de Fátima, meu irmão Heraldo Rocha e meu sobrinho Miguel que são tudo para mim. Sem o amor deles este objetivo não teria sido alcançado.

Aos irmãos que Deus colocou em minha vida e escolhi para conviver: Cristiano de Carvalho e José Luiz Padilha. Ambos compartilharam comigo todos os momentos felizes, mas também estiveram juntos nos momentos tristes da minha caminhada, e nunca me deixaram desistir de nada, nem mesmo de probabilidade avançada.

Ao meu orientador, Luiz Henrique Duczmal, pelos ensinamentos,dedicação e paciência durante a realização deste trabalho. Ao meu co-orientador, Alexandre Celestino, pelos valiosos conhecimentos em Estatística Espacial, pela motivação, paciência, comprometimento, e confiança na orientação deste projeto.

Aos demais educadores do Departamento de Estatística da UFMG, e ao professor André Cançado (UNB), por aceitar o convite para participar da banca examinadora (Qualificação), pelas contribuições na tese.

Aos professores, funcionários e colegas do Departamento de Estatística da UFMG, em especial a "Ana", Maria Cristina, Rogéria, "Rosi", "Marcinha"e Maísa pela solicitude de sempre, solidariedade perante os percalços da vida e principalmente pela amizade proporcionada em todos estes anos e que ficará para sempre.

E, finalmente, peço a Deus que abençoe aqueles que sempre estiveram do meu lado me dando apoio, que o homem lá de cima nunca lhes deixem faltar o que vocês não me

deixaram faltar, a “FÉ”.

# Resumo

Da SILVA, F. R. **Inferência Multivariada de *Clusters* Espaciais**.
Tese (Doutorado) - Instituto de Ciências Exatas, Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

A vigilância sindrômica fornece informações que auxiliam a identificar problemas de saúde pública e responder apropriadamente quando ocorrem. Estas informações em saúde pública são cruciais para pautar ações de controle e prevenção de uma variedade de condições de saúde, como doenças infecciosas, doenças crônicas e comportamentos diversos relacionados à saúde.

Na vigilância sindrômica, geralmente, há vários conjuntos de dados que podem ser usados para detectar *clusters* espaciais. Por exemplo, para se detectar um surto de gripe podem ser usados os registros médicos e o número de remédios para gripe vendidos em farmácias ou para se detectar um surto Dengue pode-se usar os casos efetivamente notificados e o número de ovos capturados em armadilhas para a fêmea do mosquito transmissor da dengue.

Neste trabalho é proposto um método de detecção de *clusters* espaciais que se baseia na estatística *scan* espacial incorporando simultaneamente informações de dois conjuntos de dados através de uma ferramenta multiobjetivo, de modo que um surto será detectado se ocorrer em apenas um ou em vários conjuntos de dados.

O conceito de significância de um *cluster* espacial será estendido de maneira natural através do uso da função de aproveitamento, sendo empregado como critério de decisão para a escolha da melhor solução. A introdução do conceito de conjunto de Pareto nesse problema, seguido da escolha da solução mais significativa, permitirá que a escolha da melhor solução seja feita de maneira rigorosa.
**Palavras-chaves:** Estatística *Scan* Espacial, Otimização Multiobjetivo, Conjunto de Pareto.

# Abstract

Syndromic surveillance provides information that helps to identify public health issues and respond appropriately when they occur. This information is crucial to prevent and control a variety of health conditions such as infectious diseases, chronic diseases and various health-related behaviors.

In syndromic surveillance, often, there are several sets of data that can be used to detect spatial clusters. For example, to detect an outbreak of influenza can be used medical records and the number of cold medicines sold in pharmacies or to detect an outbreak dengue can effectively use the reported cases and the number of eggs caught in traps the female transmits dengue mosquito.

In this paper we propose a method of detecting spatial clusters using statistical and spatial scan while incorporating multiple sets of information data via a multiobjective tool, so that a burst is detected if it occurs in one or multiple sets of data.

The concept of significance of a spatial cluster will be extended in a natural way by using the recovery function and is used as a decision criterion for choosing the best solution. The introduction of the concept set of Pareto this problem, followed by the choice of the most meaningful solution, allow the choice of the best solution is performed in a rigorous manner.

**Keywords:** Scan Statistic, Multiobjective Optimization, Pareto set.

# Sumário

**Página**

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

A demanda por sistemas capazes de detectar mudanças nos padrões espaciais de ocorrência de eventos tem crescido em diversas áreas do conhecimento tais como o monitoramento de acidentes de trânsito, o monitoramento de crimes em grandes cidades e o monitoramento de doenças (vigilância sindrômica). Em particular, existe uma crescente demanda por sistemas capazes de identificar, em tempo real, *clusters* espaciais.

Um *cluster* espacial é um agregado de regiões próximas no espaço de incidência atípica, cuja probabilidade de ocorrência por mero acaso é pequena. Um sistema para detecção deste tipo de *cluster* é chamado de sistema de vigilância espacial.

Tradicionalmente sistemas de vigilância sindrômica contam com profissionais de saúde que relatam as condições notificáveis, geralmente com a confirmação biológica de casos. Embora estes registros sejam uma parte fundamental em níveis nacionais e internacionais de saúde, estes sistemas têm problemas bem conhecidos, tais como atrasos na notificação e dificuldade em identificar atividade incomum da doença. Por esta razão os sistemas de vigilância sindrômica têm procurado usar fontes de dados alternativas, que permitam que profissionais de saúde detectem um surto da doença com a maior brevidade possível.

A expansão de métodos de vigilância não tradicionais têm ocorrido ao longo das duas últimas décadas, sendo incorporados aos sistemas de saúde pública de rotina em muitos países. Estes métodos oferecem, em tempo real, dados de uma variedade de fontes de uma forma automatizada e que podem permitir a identificação precoce de ameaças emergentes à saúde pública, fornecendo melhores estimativas da incidência de surtos sazonais.

Desta forma, para ampliar a capacidade do setor de saúde no controle das doenças faz-se necessário desenvolver novos instrumentos para a prática da vigilância incorporando informações contidas em diversos conjuntos de dados. Por exemplo, na tentativa de detecção de *clusters* espaciais com dados de dengue, podem ser utilizados os registros de casos de dengue e o número de ovos capturados em armadilhas para a fêmea do mosquito; para se detectar um surto de gripe podem ser usados os dados de visitas ambulatoriais e vendas de remédios para gripe em farmácias; ou ainda, se estamos interessados em investigar a existência de *clusters* espaciais para dados de leucemia, pode-se utilizar dados de casos de leucemia linfocítica aguda, da leucemia mielóide aguda, de leucemia crônica ou de todas as leucemias combinadas.

A estatística *scan* espacial desenvolvida por Kulldorff (1997) é uma medida usual da intensidade de um *cluster* espacial. Estatísticas *scan* espaciais foram adaptadas para detectar *clusters* espaciais usando vários conjuntos de dados (Burkom, 2003; Kulldorff *et al.*, 2007; Jonsson *et al.*, 2010) . Neste contexto, apenas procedimentos *ad-hoc* foram

propostos para resolver o problema de detecção de *clusters* mais prováveis e fornecer a significância estatística destes *clusters*.

Um algoritmo multiobjetivo (Duczmal *et al.*, 2008) foi desenvolvido anteriormente para identificar o formato geométrico dos *clusters* espaciais. Este método realiza uma busca para maximizar dois objetivos, a estatística *scan* e a regularidade da forma para um *cluster* candidato. A solução encontrada é um conjunto de Pareto, que consiste em todos os *clusters* espaciais encontrados que não são piores que nenhum outro *cluster* espacial em ambos objetivos simultaneamente. A avaliação da significância é feita paralelamente para todos os *clusters* espaciais através de simulações de Monte Carlo e este procedimento determina a melhor solução (Fonseca *et al.*, 2005).

Usaremos as ideias do algoritmo multiobjetivo para o desenvolvimento de um método que vai ser utilizado para a detecção e inferência de *clusters* espaciais que envolvam duas fontes de dados. Desenvolveremos neste trabalho um novo método que incorpora a simplicidade do método *scan* circular (Kulldorff e Nagarwalla, 1995), sendo capaz de detectar e avaliar *clusters* espaciais usando duas fontes de dados.

Objetiva-se produzir um instrumento estatístico eficiente para antecipação e, consequentemente, a ampliação da capacidade preventiva do setor de saúde para que este possa otimizar suas atividades e recursos, visando a prevenção de doenças, a promoção da saúde e minimização dos danos à população exposta a estes riscos.

## 1.1 Motivação e Justificativa

Métodos estatísticos aplicados à análise de dados, obtidos periodicamente pelos sistemas de vigilância (em saúde pública, criminologia ou ambiental), são importantes para detectar *clusters* de eventos que podem indicar uma rápida mudança no padrão dos dados observados.

Na Estatística Espacial, o problema de detecção de *clusters* espaciais é abordado através de testes de hipóteses: sob a hipótese nula a taxa de ocorrência de eventos é constante para todas as regiões e sob a hipótese alternativa existe uma elevada taxa de ocorrência de eventos em alguma sub-região conexa do mapa. A localização e o tamanho do *cluster* espacial são desconhecidos *a priori*.

Muitos testes baseados em algoritmos computacionalmente intensivos têm sido propostos na literatura para detectar *clusters* espaciais de eventos. A estatística *scan* espacial (Kulldorff, 1997) é atualmente o método mais usado em vários campos do conhecimento para a detecção de *clusters* com um formato circular.

Os departamentos de saúde têm acesso a várias fontes de dados sobre diversas doenças, tais como visitas hospitalares, vendas de medicamentos e outros dados úteis para detectar automaticamente as epidemias emergentes da doença. A rápida detecção de *clusters* espaciais de doença traz benefícios importantes para a saúde pública, como a intervenção rápida que pode evitar uma pandemia e contribuir para redução da mortalidade.

Existem dificuldades com relação ao monitoramento de algumas doenças: o diagnóstico definitivo, especialmente de doenças raras, requerem muitos testes e vários dias para estabelecer um diagnóstico preciso. Para superar o atraso inerente da saúde pública tradicional, com base em diagnósticos finais, os Centros de Controle de Doenças têm discutido a implementação de um sistema de monitoramento sobre as admissões e visitas ambulatoriais. Sintomas como febre, diarreia ou vômitos também podem ser utilizados para

auxiliar na detecção precoce de surtos de doenças.

Se, por exemplo, estamos interessados em investigar a existência de *clusters* espaciais de câncer uterino, pode ser difícil decidir se observaremos somente os casos de câncer de útero, ou somente os casos de câncer de colo de útero, ou uma combinação dos casos destes.

As mudanças climáticas podem provocar impactos na saúde humana: condições extremas de temperatura e precipitação podem facilitar a disseminação de doenças transmitidas por vetores, como a malária, a dengue e a febre amarela. Uma aplicação interessante seria descobrir a existência ou não de *clusters* espaciais de dengue, malária e/ou febre amarela na região Norte do Brasil.

A utilização de várias fontes de dados tem-se mostrado essencial para fornecer informações confiáveis em relação ao aparecimento de potenciais ameaças à saúde, em oposição a apenas uma fonte de informação. Algumas razões para se usar várias fontes de dados são:

1. Em grande parte das aplicações reais nenhuma fonte de dados é capaz de captar todos os indivíduos com uma certa doença.

2. Algumas doenças manifestam-se normalmente com um sintoma único, ao passo que outras doenças podem causar uma ampla variedade de diferentes sintomas em indivíduos diferentes.

3. Algumas doenças afetam principalmente pessoas de uma determinada faixa etária ou têm sintomas distintos dependendo da faixa etária da pessoa.

4. Algumas doenças afetam de forma distinta homens e mulheres.

A partir de uma revisão bibliográfica, notamos que existe uma carência de trabalhos que apresentam técnicas para procedimentos de detecção de *clusters* espaciais utilizando dois conjuntos de dados.

O desafio é, então, desenvolver sistemas que usem múltiplas fontes de dados que possam identificar focos de doenças nos seus estágios iniciais, permitindo uma resposta da saúde pública oportuna e eficaz.

Com esta pesquisa pretendemos apresentar um método para detecção e inferência de *clusters* espaciais, na presença de duas fontes de dados. Propomos um critério quantitativo para escolher a melhor solução, encontrando o conjunto Pareto-ótimo no espaço de soluções, seguido de um critério de decisão que consiste em maximizar a significância sobre este conjunto.

## 1.2 Revisão da literatura

Os principais métodos para detecção de *clusters* espaciais foram revisados por Elliott *et al.* (1995), Lawson e Kulldorff (1999), Buckeridge *et al.* (2005), Duczmal *et al.* (2009), Balakrishnan e Koutras (2011) e mais recentemente por Lawson (2013).

Algoritmos para a detecção de *clusters* espaciais são ferramentas úteis em estudos epidemiológicos, conforme Lawson *et al.* (1999). Para a detecção antecipada de surtos de doenças infecciosas, resultados podem ser encontrados nos trabalhos de Duczmal e Buckeridge (2006); Kulldorff *et al.* (2005, 2006, 2007) e Neill (2009). A estatística *scan* espacial

de Kulldorff (1997), utilizada para a detecção de *clusters* espaciais, está implementada no software *SaTScan* (Kulldorff, 1999). A precisão das estimativas de p-valor fornecidas pelo *Scan* Circular é discutida por Abrams *et al.* (2010). Almeida *et al.* (2011) propõem uma correção no cálculo do p-valor da estatística *scan* acrescentando a informação do tamanho dos *clusters* candidatos.

Profissionais de vigilância em saúde pública relatam condições notificáveis, geralmente com a confirmação biológica. Porém estes sistema tem problemas, incluindo atrasos na notificação, e dificuldade em identificar uma atividade incomum (Ortiz *et al.*, 2009)

A pandemia mundial de gripe *H1N1* de 2009 foi uma das motivações para adotar e avaliar muitos destes métodos. Diferentes registros têm sido utilizados como fontes de dados para a vigilância em saúde pública (Sugawara *et al.*, 2012; Liu *et al.*, 2013; Tian *et al.*, 2013; Pivette *et al.*, 2014; Todd *et al.*, 2014), incluindo dados sobre o absentismo no trabalho ou na escola (Kara *et al.*, 2012), as chamadas para serviços de assistência à saúde (Rolland *et al.*, 2006; Rodman *et al.*, 1998), as chamadas de emergência (Josseran *et al.*, 2010), ou vendas de remédios sem prescrição.

A vigilância de doenças infecciosas pode ser fortemente afetada pela demanda de cuidados por parte dos indivíduos (Dailey *et al.*, 2007). Como muitas pessoas irão se automedicar para doença leve, a vigilância das vendas sem prescrição médica tem sido sugerida como um complemento para a vigilância baseada em cuidados de saúde para estimar a magnitude e a dinâmica de uma doença (Mostashari *et al.*, 2003; Wagner *et al.*, 2004; Sočan *et al.*, 2012; Van Boeckel *et al.*, 2014; Todd *et al.*, 2014).

O desenvolvimento da Internet e a explosão das mídias sociais também tem proporcionado muitas novas oportunidades para a vigilância sindrômica (Morse, 2012; Chawla e Davis, 2013). O uso da Internet para questões de saúde deve-se em grande parte, à disponibilidade de recursos e aplicações de tecnologia da informação de saúde (Chakma *et al.*, 2009; Weitzman *et al.*, 2011; Khoury *et al.*, 2013).

Estes desenvolvimentos online, além de uma demanda por dados efetivos mais oportunos, amplamente disponíveis e a baixo custo, levou a novas formas de obtenção de dados epidemiológicos (Salathe *et al.*, 2012; Hay *et al.*, 2013).

Por exemplo, ao longo das últimas duas décadas, a tecnologia da Internet tem sido utilizada para identificar surtos de doenças, controlar a propagação de doenças infecciosas e monitorar práticas de auto-cuidado dos indivíduos (Chary *et al.*, 2013; Chunara *et al.*, 2013; Minniear *et al.*, 2013).

O uso dessas ferramentas modernas de comunicação para a vigilância da saúde pública tem provado ser menos onerosa e mais rápida do que os modos de vigilância tradicionais (Boicey, 2013). A Internet deu origem a várias fontes de "Big Data", tais como *Facebook* (Gittelman *et al.*, n.d.), *Twitter* (Li e Cardie, 2013; Lee *et al.*, 2013; Yoon *et al.*, 2013; Hingle *et al.*, 2013) e *Google* (Dugas *et al.*, 2013). Estes canais de comunicação on-line e locais de mercado fornecem uma riqueza de dados coletados passivamente que podem ser extraídos para fins de saúde pública, tais como as características sócio-demográficas, comportamentais, e as construções sociais e culturais.

Segundo Wu e Gruenwald (2010) múltiplos conjuntos de dados são fundamentais para fornecer informações úteis e confiáveis sobre o surgimento de ameaças potenciais à saúde, em comparação com os métodos que usam apenas um único conjunto de dados. A Estatística *scan* espacial de Kulldorff (1997) foi adaptada para analisar duas fontes de dados simultaneamente (Burkom, 2003; Jonsson *et al.*, 2010; Kulldorff *et al.*, 2007).

Em um trabalho de vigilância sindrômica em que se aplicou a estatística *scan* espaço-temporal a múltiplas fontes de dados, Burkom (2003) usou a abordagem de somar as contagens de todos os conjuntos de dados; somar, em cada região do mapa, os casos observados e os casos esperados para cada um dos conjuntos de dados. Os dados combinados são, então, usados para descobrir se há um *cluster* espacial no mapa em estudo. O inconveniente deste método é que se forem combinados os bancos de dados pode-se perder informação, ou então, um dos bancos de dados pode encobrir a informação contida em outro banco de dados. Isto é, se um *cluster* espacial estiver presente apenas em um dos conjuntos de dados, ele poderá ser ocultado pela variação aleatória presente nos outros conjuntos de dados.

Burkom (2003) também menciona outra abordagem: analisar cada um dos conjuntos de dados separadamente e, depois, usando algum ajuste para se combinar os resultados, tentar detectar *clusters* espaciais. Dois ajustes possíveis são: a soma das estatísticas *scan* obtidas para os conjuntos de dados e o máximo da estatística *scan* dentre todos os conjuntos de dados. A desvantagem desta abordagem é a possível falta de poder de detecção.

Em Kulldorff *et al.* (2006) foi apresentada uma extensão da estatística *scan* espaço-temporal que incorpora simultaneamente dois conjuntos de dados em uma única função de verossimilhança. Isto é feito através da soma das razões de log-verosimilhanças individuais para os conjuntos de dados para os quais a contagem de casos observados é maior do que a contagem de casos esperado. Porém, neste trabalho não existe uma maneira de se analisar a significância de um *cluster* espacial candidato.

## 1.3 Objetivos

### 1.3.1 Objetivo Geral

Propor uma metodologia de detecção e inferência de *clusters* espaciais, na presença de dois conjuntos de dados espaciais, tendo como critério quantitativo na escolha da melhor solução, o conjunto Pareto-ótimo no espaço de soluções.

### 1.3.2 Objetivos específicos

- Avaliar o poder de detecção do método na presença de *clusters* espaciais simulados;
- Comparar, por meio de simulação, o método proposto com métodos e modelos existentes na literatura;
- Aplicação do modelo desenvolvido em dados reais.

## 1.4 Estrutura do texto

Esta tese está organizada em capítulos. No capítulo 2 introduz-se a estatística de teste na qual é baseada a busca de *clusters* - a estatística *scan* espacial. Essa revisão abrange o método *Scan* Circular clássico.

No capítulo 3 serão apresentados métodos para a detecção de *clusters* espaciais para duas fontes de dados, baseados na estatística *scan* espacial. No capítulo 4 será apresentada a abordagem de otimização multiobjetivo para atender o objetivo principal, proposto neste trabalho.

No capítulo 5 apresentamos o método proposto nesta tese para o caso bi-objetivo.

No capítulo 6 descrevemos como avaliar o comportamento do algoritmo multiobjetivo em termos de poder, sensibilidade e valor preditivo positivo.

No capítulo 7 comparamos o algoritmo proposto com os outros procedimentos para detecção de *clusters*, usando duas fontes de dados. O método multiobjetivo, proposto nesta tese, é usado em aplicações a dados reais no capítulo 8. As considerações finais desta tese e as propostas de continuidade de trabalho estão no capítulo 9.

# Capítulo 2

# Estatística *Scan*

A verificação de padrões anormais da distribuição geográfica da incidência de algum fenômeno de interesse é de suma importância para que se possa planejar políticas de intervenção em saúde ou segurança pública. Estudos referentes a *clusters* espaciais que apresentam discrepância na ocorrência do fenômeno de interesse são encontrados em diversos trabalhos.

O método *scan* circular tem sido bastante estudado e testado em diversas situações. Para um bom entendimento do referencial teórico sobre o assunto de interesse dessa tese, a próxima seção se propõe a explicar o método da estatística *scan* espacial, a principal técnica utilizada na detecção de *clusters*.

## 2.1 Estatística *Scan* Espacial

Nesta seção será feita uma breve revisão da estatística *scan* clássica introduzida em Kulldorff (1997).

Consideremos um mapa dividido em $m$ regiões $R_1, \ldots, R_m$, com população total $N$ e número total de casos $C$, para algum fenômeno de interesse. Assuma que a população e o número de casos em cada uma das regiões sejam também conhecidos e denotados por $n_i$ e $c_i$ com $i \in \{1, \ldots, m\}$, respectivamente. Define-se como zona qualquer subconjunto conexo de regiões do mapa em estudo e seja $Z$ o conjunto de todas as zonas do mapa.

Com a suposição de que os casos se distribuem no mapa segundo o modelo de *Poisson*, o número de casos $C_i$ em cada região $R_i$ é uma variável aleatória de *Poisson*, cujo parâmetro $\mu_i$ é o número esperado de casos na região $R_i$, dado por $C \cdot (n_i/N)$, sob $H_0$.

O procedimento proposto por Kulldorff (1997) constitui-se na construção de um teste de hipótese, cuja hipótese nula é a de não existência de *cluster* no mapa em estudo, enquanto que a hipótese alternativa pressupõe a existência de pelo menos um *cluster* no mapa em estudo. Este teste é o clássico teste da razão de verossimilhanças.

Pode-se mostrar que a razão entre a função de verossimilhança sob a hipótese alternativa e a função de verossimilhança sob a hipótese nula, para a distribuição de casos em alguma zona $z$, é dada por:

$$LR(z) = \begin{cases} \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z} & \text{se } c_z > \mu_z \\ 1 & \text{caso contrário} \end{cases} \tag{2.1}$$

A zona $z$ mais verossímil é aquela que maximiza a função $LR(z)$ com respeito ao conjunto $Z$. Desta forma, a estatística de teste fica definida por $\max_{z \in Z} LR(z)$.

Em geral, a função $LR(z)$ assume valores muito grandes. Para amenizar esse problema, utiliza-se o logaritmo da razão de verossimilhança, $LLR(z)$. Dado que a função logaritmo é monotonamente crescente, a zona $z$ que maximiza $LR(z)$ também maximiza $LLR(z)$. A expressão $LLR(z)$ é dada por:

$$LLR(z) = \begin{cases} (c_z) \log\left(\frac{c_z}{\mu_z}\right) + (C - c_z) \log\left(\frac{C - c_z}{C - \mu_z}\right) & \text{se } c_z > \mu_z \\ \\ 0 & \text{caso contrário} \end{cases} \tag{2.2}$$

De posse de uma estatística que permita avaliar cada zona, resta encontrar aquela que apresenta avaliação máxima. Porém, a maior dificuldade da estimação de *clusters* reside exatamente na maximização da estatística $LLR(z)$ sobre o conjunto $Z$ de todas as zonas possíveis.

Isto porque, embora seja finito, o conjunto $Z$ é, em geral, tão grande que torna a maximização de $LLR(z)$ impraticável através de uma busca exaustiva. Se há $m$ regiões no mapa de estudo, existem $2^m - 1$ possíveis subconjuntos de regiões, dos quais deveríamos verificar quais são conexos, para construir o conjunto $Z$.

Desta forma, alguma técnica para redução do espaço de busca deve ser utilizada. A técnica mais utilizada para este fim é denominada *Scan* Espacial Circular (Kulldorff e Nagarwalla, 1995).

O método *scan* circular restringe o espaço de busca apenas às zonas que têm formato circular. Para isso o método utiliza janelas circulares que varrem o mapa em busca da zona $z^*$, sendo que $z^*$ é a zona que possui o maior valor para a $LLR(z)$. Para cada região do mapa definimos um centroide, que é um ponto arbitrário em seu interior. Assim, uma janela circular sobre o mapa em estudo define uma zona que é constituída pelas regiões cujos centroides se encontram dentro da janela.

Seja $d_{i,j}$ a distância entre os centroides $i$ e $j$, das regiões $R_i$ e $R_j$, respectivamente. O método *scan* circular escolhe as janelas da seguinte forma: selecione uma região $R_k$, $1 \leq k \leq m$. Ordenam-se as demais $m - 1$ regiões do mapa quanto à distância ao centroide $k$, em ordem crescente, obtendo a sequência de regiões $\{R_{k_1}, R_{k_2}, \ldots, R_{k_{m-1}}\}$, onde $d_{k,k_1} \leq d_{k,k_2} \leq \ldots \leq d_{k,k_{m-1}}$. As janelas são escolhidas como sendo círculos cujos centros coincidem com o centroide $k$ e com raios iguais a $d_{k,k_1}, d_{k,k_2}, \ldots, d_{k,k_s}$, onde $s$ é tal que $d_{k,k_s} \leq r_{\max} < d_{k,k_{s+1}}$, sendo $r_{\max}$ é o raio máximo permitido. Cada janela gera uma zona e o processo é repetido para $k = 1, \ldots, m$. Para cada janela avalia-se a zona correspondente através da estatística *scan*. O *cluster* mais verossímil é aquele que maximiza $LLR(z)$.

## 2.2 Significância Estatística

A princípio, a zona $z^*$ que maximiza a razão de verossimilhança é uma candidata a *cluster*. Somente após a verificação de sua significância estatística, a zona $z^*$ poderá ter seu status alterado para *cluster* detectado.

Se a distribuição de probabilidade da estatística $T = \max_z LLR(z)$ fosse conhecida sob a hipótese de que não há *clusters* no mapa, poder-se-ia calcular o valor de $T_{\text{crit}}$ acima do

qual poder-se-ia considerar, sob $H_0$, uma solução discrepante, simplesmente encontrando $T_{\text{crit}}$ tal que $P(T > T_{\text{crit}}) = \alpha$, onde $T$ é a estatística sob a hipótese nula e $\alpha$ é o nível de significância, sendo $\alpha = P\,(\text{Rejeitar } H_0 | H_0 \text{verdadeiro}) = P(T > T_{crit} | H_0)$.

Um valor de $T$ abaixo de $T_{\text{crit}}$ pode ocorrer por mero acaso $(1-\alpha) \times 100\%$ das vezes, mas um valor acima de $T_{\text{crit}}$ só acontece por acaso com probabilidade menor ou igual a $\alpha$ e, portanto, a solução pode ser considerada um *cluster*. A probabilidade de que o valor observado da estatística *scan* ocorra por mero acaso sob $H_0$ é chamada de p-valor, p-valor $= P(T \geq T_{obs} | H_0)$.

O problema é que, a princípio, não é conhecida a distribuição da estatística *scan*. Os métodos de detecção são utilizados para encontrar o *cluster* que maximiza a estatística *scan*. Antes de podermos afirmar que essa solução é um *cluster*, deve-se levar em conta que um *cluster* deve apresentar um número anormal de casos. Em outras palavras, não se pode afirmar que uma medida é discrepante das demais simplesmente por ela ser a maior dentre todas as avaliadas. Essa medida deve ser comparada com um universo de medidas. A partir desse universo é que será possível estabelecer uma medida crítica, acima da qual uma medida pode ser considerada anormal.

Desta forma, para considerarmos que a solução encontrada pelo método de detecção é um *cluster* devemos comparar sua avaliação com as avaliações de soluções encontradas para vários cenários aleatórios sob $H_0$. Só a partir dessa comparação é que será possível afirmar se a solução é ou não um *cluster*.

Para testar a significância da estatística de teste, serão utilizadas as simulações de Monte Carlo como apresentado por Dwass *et al.* (1957). Uma simulação Monte Carlo consiste em construir milhares de réplicas do mapa original em que o número total de casos $C$ está fixo e os casos em cada região são distribuídos aleatoriamente sob a hipótese nula de não existência de *cluster* no mapa.

Para cada réplica tem-se um valor da estatística *scan* e o conjunto delas, obtido pelas várias réplicas, gera uma distribuição empírica da estatística de teste. O p-valor da estatística de teste $LLR(z^*)$ para o mapa dos casos observados pode ser estimado determinando o posto ocupado pelo seu valor dentre os valores da distribuição empírica da estatística de teste sob a hipótese nula, ou seja, o p-valor do *cluster* mais provável do mapa original é estimado como sendo a razão entre o número de valores empíricos de $LLR(z)$ que ultrapassam o valor de $LLR(z^*)$ pelo número total de mapas gerados aleatoriamente na simulação.

## 2.3 Algoritmo da Estatística *Scan* Circular

A seguir será apresentado o algoritmo de detecção de *cluster* proposto por Kulldorff (1997). Considerando um mapa de ocorrências aleatórias dividido em m regiões, define-se o centroide de cada região como um ponto arbitrariamente escolhido nesta região do mapa.

1. Escolher o centroide de uma das regiões em estudo;

2. Representar o conjunto das distâncias entre dois centroides quaisquer em uma matriz simétrica denominada matriz de distâncias. Cada linha $i$ da matriz representa as distâncias entre o centroide $i$ e os demais centroides das regiões do mapa. Ou

seja, cada linha $i$ da matriz de distâncias representa um vetor contendo as distâncias entre o centroide da região $i$ e os centroides das demais regiões;

3. Em seguida, para cada um desses vetores, as distâncias são ordenadas em ordem crescente;

4. Centrada em cada uma das regiões constrói-se um círculo de raio variável, de tal forma que, o raio desse círculo aumente de acordo com as distâncias crescentes até que a população das regiões englobadas pelo círculo atinja um percentual máximo pré estabelecido da população total, ou um certo número pré estabelecidos de regiões. Para cada círculo o número de casos e a população são atualizados e calcula-se o logaritmo da razão de verossimilhança;

5. Calcular o valor da estatística de teste $T = \max_{z} LLR(z)$;

6. Utilizar a simulação de Monte Carlo para avaliar a significância do teste, como descrito anteriormente;

7. Se $H_0$ for rejeitada, a zona $z^*$, que maximiza $LLR(z)$, será o *cluster* mais verossímil ou provável.

O algoritmo *scan* circular gerará como resultado o *cluster* mais verossímil.

# Capítulo 3

# Algoritmos que levam em conta duas fontes de dados

A estatística *scan* é a técnica mais usada para a detecção automática de *clusters* espaciais de eventos, sendo comumente usada pelos órgãos responsáveis pela saúde pública através do popular software SaTScan (Kulldorff, 2011) para detecção de surtos de doenças.

A estatística *scan* espacial foi aplicada a uma ampla variedade de estudos epidemiológicos para detecção de *clusters* (ver por exemplo, (Viel *et al.*, 2000; Sankoh *et al.*, 2001; Perez *et al.*, 2002) ). Nas comparações de poder de testes para detecção de *clusters* de doenças, a estatística *scan* demonstrou ser a mais poderosa para a detecção de *cluster* (Kulldorff *et al.*, 2003; Song e Kulldorff, 2003).

Por esta razão muitas variantes da estatística *scan* foram sugeridas para encontrar *clusters* espaciais. Alguns métodos para encontrar *clusters* espaciais e suas aplicações foram revisados por Duczmal *et al.* (2009).

As agências governamentais responsáveis pela segurança pública devem responder rapidamente a ameaças potenciais, incluindo guerras, surtos de doenças, ondas de crime, desastres naturais e ataques terroristas. Respostas rápidas a tais eventos podem reduzir substancialmente os custos resultantes para a sociedade, enquanto as respostas tardias podem ter resultados catastróficos. Por esta razão os sistemas de vigilância monitoraram uma grande quantidade de dados para tentar detectar e identificar padrões emergentes.

Portanto, incorporar estes múltiplos conjuntos de dados na detecção de *clusters* espaciais, é fundamental para fornecer informações úteis e confiáveis sobre o surgimento de ameaças potenciais à saúde. A Estatística *scan* foi adaptada para analisar múltiplas fontes de dados.

Porém, como foi visto na revisão bibliográfica, poucos trabalhos foram propostos para se incorporar a informação de diversas fontes de dados na detecção de *clusters* espaciais. Neste capítulo iremos descrever as seguintes abordagens, existentes na literatura, que incorporam a informação de dois conjuntos de dados usando a estatística *scan*:

1. **Soma das razões de log verossimilhança na zona z** ($llr_1 + llr_2$): Uma abordagem para lidar com dois conjuntos de dados é calcular a razão de log-verossimilhança para cada zona em cada uma das fontes de dados separadamente e, depois, somar estas razões de verossimilhança;

2. **Máximo das razões de log verossimilhança na zona z** ($max(llr_1, llr_2)$)**:** Ou-

tra abordagem para lidar com dois conjuntos de dados é calcular a razão de log-verossimilhança para cada zona em cada uma das fontes de dados separadamente e, depois, dentro de cada zona, obter a razão de verossimilhança máxima dentre os dois conjuntos de dados;

3. *A norma da Soma (Naive)*: Burkom (2003), em um trabalho aplicado de vigilância sindrômica, para usar a estatística *scan*, somou, em cada região do mapa, a população em risco, os casos observados e os casos esperados, para cada um dos conjuntos de dados. Os dados combinados foram, então, usados para se avaliar se há um *cluster* espacial no mapa em estudo.

Descreveremos brevemente estes três métodos nas próximas Seções.

## 3.1 Soma das razões de log verossimilhança na zona z $(llr_1 + llr_2)$

Burkom (2003) sugeriu uma abordagem para lidar com dois conjuntos de dados: analisar cada um dos conjuntos de dados separadamente e, depois, usando algum ajuste para se combinar os resultados, tentar detectar *clusters*. Dois ajustes usados são a soma das estatísticas *scan* obtidas para os conjuntos de dados, que será descrito nesta seção, e o máximo da estatística *scan* dentre os dois conjuntos de dados que será descrito na Seção 3.2.

O método da soma das estatísticas *scan*, que aqui denotaremos por $(llr_1 + llr_2)$, pode ser descrito através dos seguintes passos:

1. Para cada zona, construída como descrito na Seção 2.3, serão atualizados:

   - $c_i(z)$: o número de casos do conjunto de dados $i = 1, 2$ para a zona $z$;
   - $pop_i(z)$ a população do conjunto de dados $i = 1, 2$ para a zona $z$.

2. Para cada zona $z$ construída no mapa de casos calcula-se o logaritmo da razão de verossimilhança para o conjunto de dados 1 $(llr_1(z))$ e para o conjunto de dados 2 $(llr_2(z))$ de acordo com a equação 2.2;

3. Para cada zona $z$ serão somadas as razões de log-verossimilhança obtidas para cada conjunto de dados, ou seja, obteremos, para cada zona $z$, a soma $llr_1(z) + llr_2(z)$;

4. Calcular o valor da estatística de teste $T = \max_z (llr_1(z) + llr_2(z))$.

   A zona $z^*$ que maximiza a estatística $T$ obtida acima é apenas uma candidata a *cluster* e temos que verificar a sua significância estatística, para descobrir se a zona $z^*$ é ou não *cluster* detectado. Para testar a significância da estatística de teste, serão utilizados as simulações de Monte Carlo: Serão construídos milhares de réplicas, em paralelo, para cada um dos conjunto de dados, do mapa original, da seguinte maneira:

   - O número total de casos $C_i$, para cada conjunto de dados $i = 1, 2$, estará fixo;

- os casos em cada região, em cada um dos conjuntos de dados, serão distribuídos aleatoriamente sob hipótese nula de não existência de *cluster* no mapa.
- Calcula-se para cada zona $z$ construída no mapa de casos calcula-se o logaritmo da razão de verossimilhança para o conjunto de dados 1 e para o conjunto de dados 2 e estas razões de log-verossimilhança serão somadas.

Para cada réplica tem-se a estatística $T^* = \max_z (llr_1(z) + llr_2(z))$. O p-valor da estatística de teste $T$ pode ser estimado determinando o posto ocupado pelo seu valor dentre os valores da distribuição empírica da estatística de teste sob a hipótese nula, ou seja, o p-valor do *cluster* mais provável do mapa original é estimado como sendo a razão entre o número de valores empíricos de $(llr_1(z) + llr_2(z))$ que ultrapassam o valor de $(llr_1(z^*) + llr_2(z^*))$ e o número total de mapas gerados aleatoriamente na simulação;

5. Utilizando as simulações de Monte Carlo será obtida a significância do teste;
6. Se $H_0$ for rejeitada, a zona $z^*$, que maximiza $(llr_1(z) + llr_2(z))$, será o *cluster* mais verossímil ou provável.

A desvantagem desta abordagem é a potencial falta de poder de detecção de *clusters*.

## 3.2 Máximo das razões de log verossimilhança na zona z $(max(llr_1, llr_2))$

O *Máximo das razões de log verossimilhança* na zona z $(max(llr_1, llr_2))$ ao invés de obter a soma das razões de log-verossimilhança dos conjuntos dados, como no método anterior, obtém o máximo entre estas verossimilhanças em cada zona e pode ser descrito através dos seguintes passos:

1. Para cada zona $z$ , as razões de log-verossimilhança são calculadas para cada conjunto de dados, obtendo $llr_1(z)$ e $llr_2(z)$;
2. Para cada zona $z$, obtém-se a razão de log-verossimilhança máxima entre os conjuntos de dados, ou seja, será obtido, para cada zona $z$, $U(z) = max(llr_1(z), llr_2(z))$;
3. O valor da estatística de teste $T = \max_z (U(z))$ é calculada.

   Para se calcular a significância da zona $z^*$, serão construídos milhares de réplicas, em paralelo, para cada um dos conjunto de dados, do mapa original, conforme foi descrito na Seção 3.1. O p-valor do *cluster* mais provável do mapa original é estimado como sendo a razão entre o número de valores empíricos de $(U(z))$ que ultrapassam o valor de $(U(z^*))$ e o número total de mapas gerados aleatoriamente na simulação;
4. Utilizando as simulações de Monte Carlo será obtida a significância do teste;
5. Assim como no método anterior a zona $z^*$, que maximiza $(U(z))$, será o *cluster* mais verossímil.

## 3.3 A norma da soma (*naive*)

A *norma da soma* (*naive*) é o mais simples dentre os métodos apresentados anteriormente listados. Burkom (2003) para aplicar a estatística *scan* a múltiplas fontes de dados, combinou os dados somando, em cada região do mapa, a população em risco, os casos observados e os casos esperados de todos os conjuntos de dados.

Os dados combinados foram, então, usados para se descobrir se há, ou não, um *cluster* espacial no mapa em estudo.

Logo, o método consiste basicamente em combinar os diversos bancos de dados em um só e proceder uma análise univariada dos dados combinados. O método pode ser assim descrito:

1. Para cada zona $z$ os casos e a população em risco dos conjuntos de dados $i = 1, 2$ serão somados:
   - $c(z) = c_1(z) + c_2(z)$.
   - $\text{pop}(z) = \text{pop}_1(z) + \text{pop}_2(z)$.
2. Para cada zona será obtida a razão log-verossimilhança usando $c_i$ e a $\text{Pop}(z)$;
3. O valor da estatística de teste, para os dados combinados $W = \max_z LLR(z)$, será calculada;
4. Serão construídos milhares de réplicas, em paralelo, para cada um dos conjunto de dados, do mapa original de acordo com 2.2. Em cada réplica os dados serão combinados como foi descrito no passo 2 e, então, será obtida a estatística $W = \max_z LLR(z)$;
5. Utilizando as simulações de Monte Carlo será obtida a significância do teste;
6. Se $H_0$ for rejeitada, a zona $z^*$, que maximiza a estatística de teste para os dados reais, será o *cluster* mais verossímil ou provável.

A principal desvantagem deste método é que, se um surto está presente apenas em um dos conjuntos de dados, ele pode não ser detectado, por seu efeito ser "escondido"pela variação aleatória presente no outro conjunto de dados.

Estas três técnicas podem ser estendidas facilmente para qualquer quantidade de conjunto de dados.

Desta maneira, as técnicas de otimização multiobjetivo proporcionam algoritmos mais interessantes que levam em conta múltiplos objetivos durante a busca por *clusters* espaciais. O capítulo 4 apresenta as técnicas de otimização multiobjetivo relacionadas a este trabalho.

# Capítulo 4

# Abordagem multiobjetivo

Neste capítulo serão introduzidos alguns conceitos sobre otimização multiobjetivo e será apresentada uma maneira de se obter a significância de *clusters* candidatos.

Como o problema tratado nesta tese de doutorado é um problema de maximização, nossas definições levarão em conta que o problema de otimização a ser resolvido é um problema de maximização, ao contrário do que é feito normalmente na literatura.

## 4.1 Otimização multiobjetivo

Um problema de otimização multiobjetivo é um problema com dois ou mais objetivos que precisam ser otimizados simultaneamente. É importante mencionar que os objetivos podem ser conflitantes entre si e que o problema pode estar sujeito a restrições. Isso faz com que o conceito de otimalidade utilizado em otimizações mono-objetivo não possa ser utilizado. Em otimização multiobjetivo, o conceito de otimalidade baseia-se na noção introduzida por Edgeworth (1881) e depois generalizada por Pareto (1964).

Um problema de otimização multiobjetivo é composto por um conjunto de funções-objetivo a serem otimizadas (maximizadas ou minimizadas) e um conjunto de restrições que devem ser satisfeitas para que a solução seja factível. Supondo a existência de $n$ funções-objetivo que formam o vetor $\boldsymbol{f}(x) = (f_1(x), f_2(x), \ldots, f_n(x))$, sujeitas possivelmente às restrições $g_i(x) \geq 0$, $i = 1, \cdots, r$, o problema pode ser formulado como:

$$\begin{aligned} \text{maximizar} \quad & \boldsymbol{f}(x) \\ \text{sujeito a} \quad & g_i(x) \geq 0,\ i = 1, \cdots, r \end{aligned} \tag{4.1}$$

Analisando um único objetivo, o conjunto imagem desta função possui elementos pertencentes à reta, portanto, podem ser classificados pela ordem existente na reta. Quando parte-se para uma abordagem multiobjetivo, o conjunto imagem da função objetivo possui elementos pertencentes ao $\mathbb{R}^n$, não possuindo, então, uma relação de ordem total. Para estabelecer uma relação de ordem neste tipo de conjunto, utiliza-se o conceito de dominância:

**Definição 4.1** (Dominância)**.** *Seja um problema de maximização e seja* $\boldsymbol{f}(x) = (f_1(x), \ldots f_n(x))$ *uma função definida em um espaço* $X$*. Um ponto* $x \in X$ *domina outro ponto* $y \in X$ *(denota-se* $x \succ y$*) se* $f_i(x) \geq f_i(y)$*,* $i = 1, \ldots n$ *e se existe pelo menos um índice* $k \in \{1, \ldots, n\}$ *tal que* $f_k(x) > f_k(y)$*.*

Em outras palavras, um ponto $x$ domina o ponto $y$ se a avaliação de $x$ for melhor que a avaliação de $y$ em pelo menos um objetivo e não for pior em nenhum outro objetivo, como pode ser visto na figura 4.1.

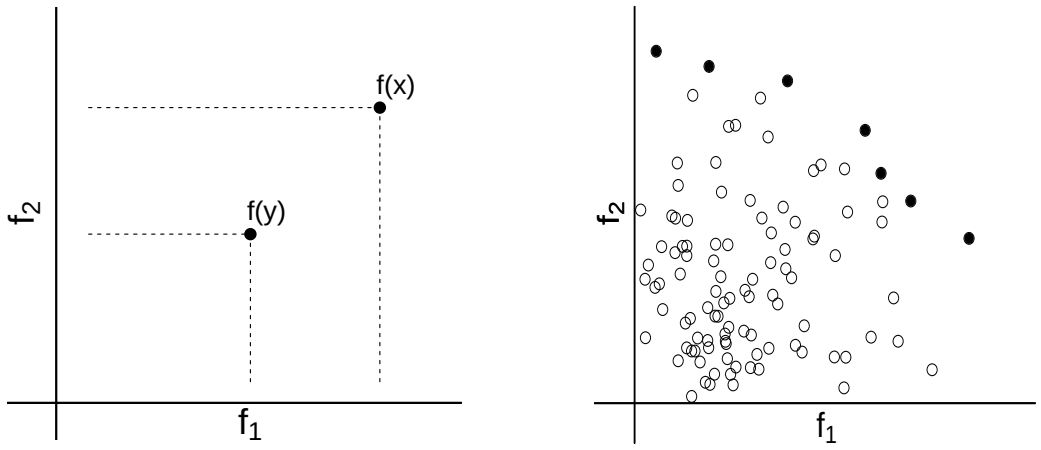



Figura 4.1: O ponto x domina o ponto y e o Conjunto Pareto-ótimo (•) e pontos dominados (○).

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

**Definição 4.2** (Solução Pareto-ótima)**.** *Diz-se que uma solução $x^* \in X$ é Pareto-ótima se não existe outro elemento em $X$ que domina $x^*$.*

Note que dizer que uma solução é Pareto-ótima não significa dizer que ela é melhor que todas as (ou que algumas das) outras soluções, mas que ela não é pior que nenhuma outra. Uma solução Pareto-ótima pode ainda ser chamada de solução não dominada ou solução eficiente. O conjunto Pareto-ótimo é formado então por todas as soluções Pareto-ótimas. A Figura 4.1 ilustra o conceito de pontos dominados e o de pontos que formam o conjunto de Pareto-ótimo.

## 4.2 Inferência multiobjetivo

Nesta tese de doutorado será utilizada a *superfície de aproveitamento* para o cálculo do p-valor para o caso multiobjetivo (Cançado *et al.*, 2010). A seguir será explicado como inferir o p-valor das soluções utilizando esta técnica.

### 4.2.1 Superfície de aproveitamento

As definições presentes nesta seção foram discutidas por Cançado *et al.* (2010), a partir dos trabalhos de da Fonseca *et al.* (2001) e Fonseca *et al.* (2005).

Considere um problema de maximização bi-objetivo, com objetivos $f_1$ e $f_2$. Seja $\varepsilon = \{x_j,\ j = 1, \ldots, Q\}$ o conjunto de todas as soluções obtidas em uma realização da estratégia de otimização, e sua imagem será $\mathcal{I} = \{Y_j = (f_1(x_j); f_2(x_j)),\ j = 1, \ldots, Q\}$, contida no espaço de objetivos contido no $\mathbb{R}^2$. Uma solução $x$ é chamada de não-dominada se $x_i$ não é dominada por qualquer outra solução em $\{x_j,\ j \neq i = 1, \ldots, Q\} \in \varepsilon$. Seja $\{x_j^*,\ j = 1, \ldots, q\} \subset \varepsilon$ o conjunto de soluções não-dominadas de $\varepsilon$, o subconjunto $\mathcal{Y} = \{Y_j^* = (f_1(x_j^*), f_2(x_j^*)),\ j = 1, \ldots, q\} \subset \mathcal{I}$ é definido como o resultado de uma única execução de um algoritmo bi-objetivo.

Pode-se associar a $\mathcal{Y}$ uma fronteira que divide o espaço de objetivos em duas regiões $R_1$ e $R_0$: $R_1$ é a região consistindo de pontos dominados por, ou igual a, pelo menos um ponto em $\mathcal{Y}$; e $R_0$ consistindo dos pontos que não são dominados por nenhum dos pontos em $\mathcal{Y}$, conforme pode ser visto na Figura 4.2.

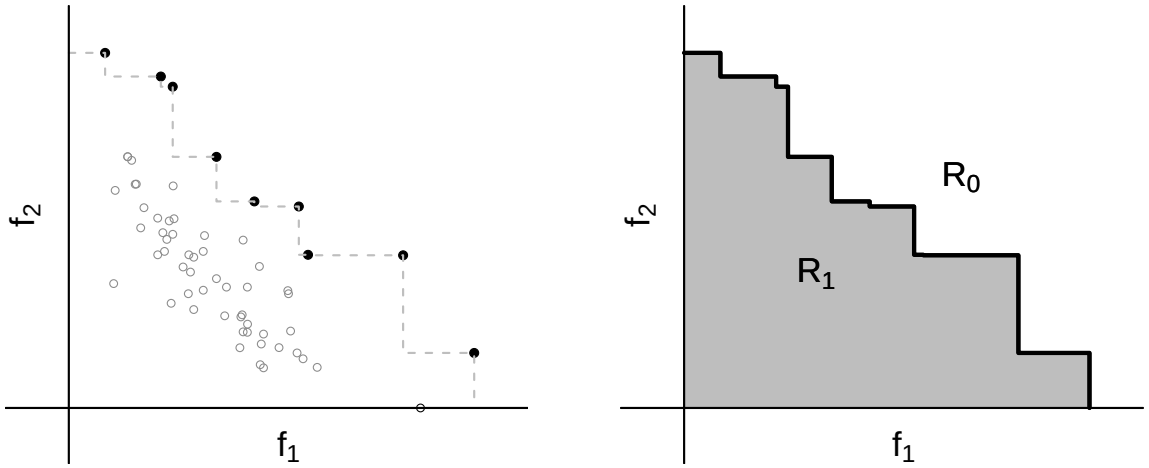



Figura 4.2: Fronteira entre $R_0$ e $R_1$.

Quando uma solução $x$ é dominada por, pelo menos, uma solução de $\mathcal{Y}$, tem-se que $x$ é atingida por $\mathcal{Y}$. Na Figura 4.2, qualquer solução localizada na região $R_1$ é atingida por $\mathcal{Y}$.

Considerando entradas de dados distintas, a cada realização do algoritmo serão produzidas diferentes saídas, obtendo-se assim múltiplas fronteiras, como pode ser visto na Figura 4.3(a).

Pontos situados no canto superior direito da Figura 4.3 (a) não são atingidos por nenhuma das fronteiras. Pontos localizados no canto inferior esquerdo são atingidos por todas as fronteiras. E pontos situados entre as diferentes fronteiras foram atingidos em algumas realizações, mas em outras não. Assim pode-se dividir o espaço em $n+1$ regiões de acordo com a frequência em que estas regiões são atingidas. As fronteiras dessas regiões são chamadas de superfícies de aproveitamento (Figura 4.3 (b)). Essas frequências são utilizadas para estimar a probabilidade de se atingir um ponto no espaço de objetivos, quando um grande número de realizações é executado. A função de aproveitamento avaliada em um ponto $Y$ no espaço de objetivos pode ser estimada pelos conjuntos das saídas $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ obtidas através de $n$ realizações independentes do algoritmo, como:

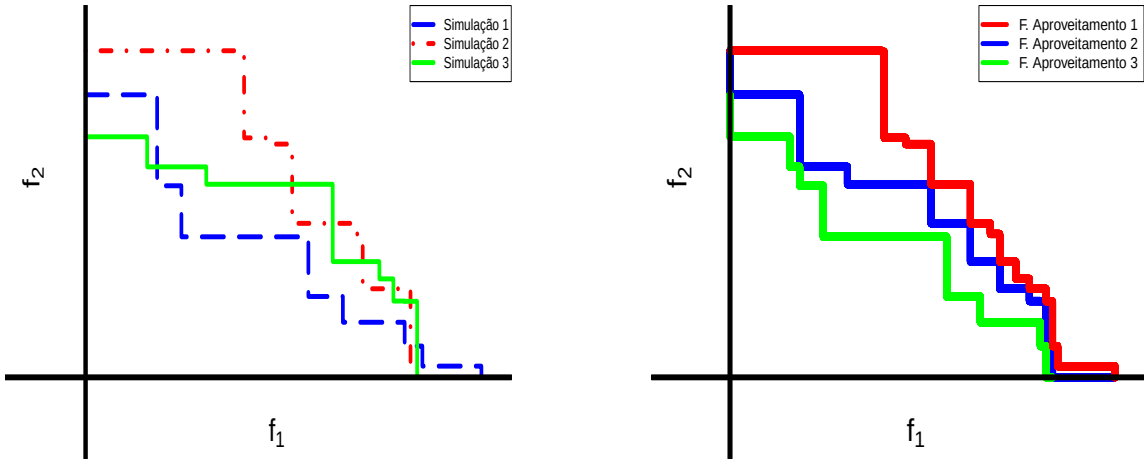



Figura 4.3: Múltiplas Fronteiras de Pareto e suas respectivas Superfícies de Aproveitamento.

$$A_n(Y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\mathcal{Y}_i \trianglerighteq Y\right),$$

em que $\mathbb{I}$ é a função indicadora (igual a 1 se $\mathcal{Y}_i \trianglerighteq Y$ e zero caso contrário) e o símbolo "$\trianglerighteq$" significa que pelo menos um elemento de $\mathcal{Y}_i$ domina ou é igual a $Y$.

No problema específico, em estudo neste trabalho, o interesse está em estimar o p-valor de *clusters* candidatos não-dominados, representados por pontos no espaço de objetivos. Formalmente, define-se $A(Y)$ como o $\lim_{n\to\infty} A_n(Y)$ quando ele existe. Agora, dado que $0 < p \leq 1$, a isolinha do valor p é definida como a imagem inversa $A^{-1}(p)$.

Sob certas condições de suavidade, $A^{-1}(p)$ é uma superfície unidimensional dividindo o espaço de objetivos em duas regiões $R_0$ e $R_1$, tais que se $Y \in R_0$ então $A(Y) > p$, e se $Y \in R_1$ então $A(Y) \leq p$. Na prática, dadas $n$ saídas $y_1, \ldots y_n$ pode-se construir aproximações das isolinhas de $p$ valor para cada $p = i/(n+1),\ i = 1, \ldots, n+1$ através das funções de aproveitamento estimadas $A_n(Y)$.

# Capítulo 5

# Inferência Multivariada de *Clusters* Espaciais

Para incorporar informações de duas fontes de dados em uma detecção de *clusters* espaciais mais coerente será proposto um método de detecção e inferência sobre *clusters* que utiliza ferramentas estatísticas de análise multiobjetivo em conjunto com a estatística *scan* espacial.

Descrevemos brevemente o método proposto nesta tese:

- Considere duas fontes de dados distintas ;
- A $j$-éssima função objetivo avaliará a força dos *clusters* candidatos usando a estatística *scan* usando apenas a informação do conjunto de dados $j$, $j = 1, 2$;
- As melhores soluções de *clusters* são encontrados pela maximização de duas funções objetivo simultaneamente, com base no conceito de dominância (descrito na seção 4.1).. Com isto será possível analisar e selecionar as melhores soluções de *cluster*, dadas pelo conjunto não dominado;
- Para avaliar a significância estatística das soluções, uma técnica estatística baseada no conceito de função de aproveitamento é utilizada (descrita na seção 4.2). Com isto é possível atribuir, de forma rigorosa, a significância estatística de cada *cluster* candidato.

Neste capítulo apresentaremos um novo método para detecção de *clusters* espaciais. Este método fornece como solução um conjunto de *clusters*, em que cada um dos *clusters* é um elemento de um conjunto especial, chamado *conjunto de Pareto*.

O propósito do algoritmo *scan circular* multiobjetivo é, então, encontrar zonas $Z$ no conjunto de todas as zonas possíveis, numa tentativa de maximizar dois objetivos: a razão de verossimilhança para o conjunto de dados 1 e razão de verossimilhança para o conjunto de dados 2.

Considere um mapa com $m$ regiões, em que cada conjunto de dados $i = 1, 2$, tenha população igual a $N_{ij}$ (a população do conjunto de dados $i,\ i = 1, 2$ na região $j,\ j = 1, \ldots, m$) sendo que, destes, $n_{ij}$ são casos (casos do conjunto de dados $i,\ i = 1, 2$ na região $j,\ j = 1, \ldots, m$).

A $i$-ésima função objetivo avalia a força do *cluster* candidato somente para o $i$-ésimo conjunto de dados através do cálculo do logaritmo da razão de verossimilhança, conforme

descrita na seção 2.3. Assim, o valor da função $LLR(z)$ em cada uma das zonas candidatas para cada um dos $i$ objetivos, para as todas as zonas construídas de acordo com Kulldorff (1997).

As melhores soluções de *cluster* são encontradas pela maximização das funções objetivos $(LLR_1(z), LLR_2(z))$ simultaneamente, baseado no conceito de dominância (Definição 4.1), ou seja, um ponto é dito ser dominado se sua avaliação é pior que outro ponto em pelo menos um objetivo, e não é melhor em nenhum dos demais objetivos. O conjunto de Pareto (Definição 4.2) é composto por todos os pontos que não são dominados por nenhuma outra solução. Para um melhor entendimento, apresenta-se um exemplo ilustrado, na Figura 5.1, para as funções objetivo.

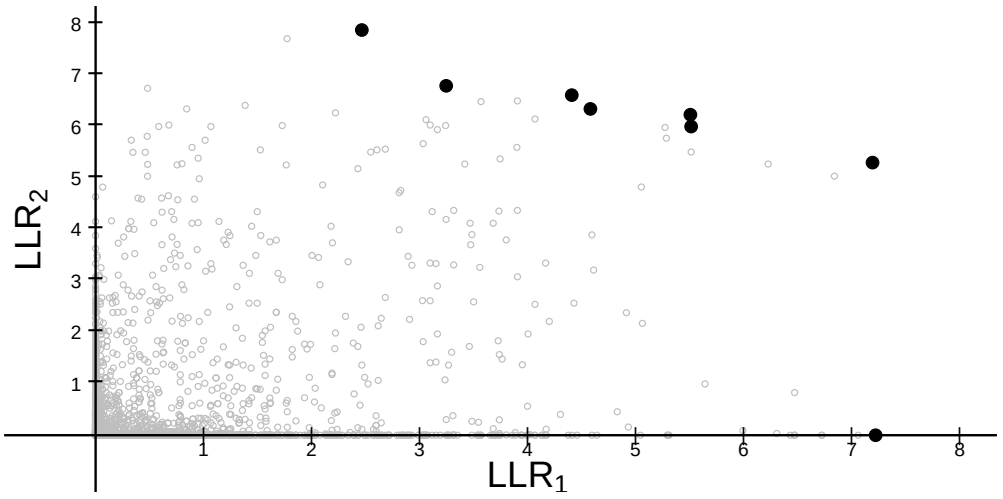
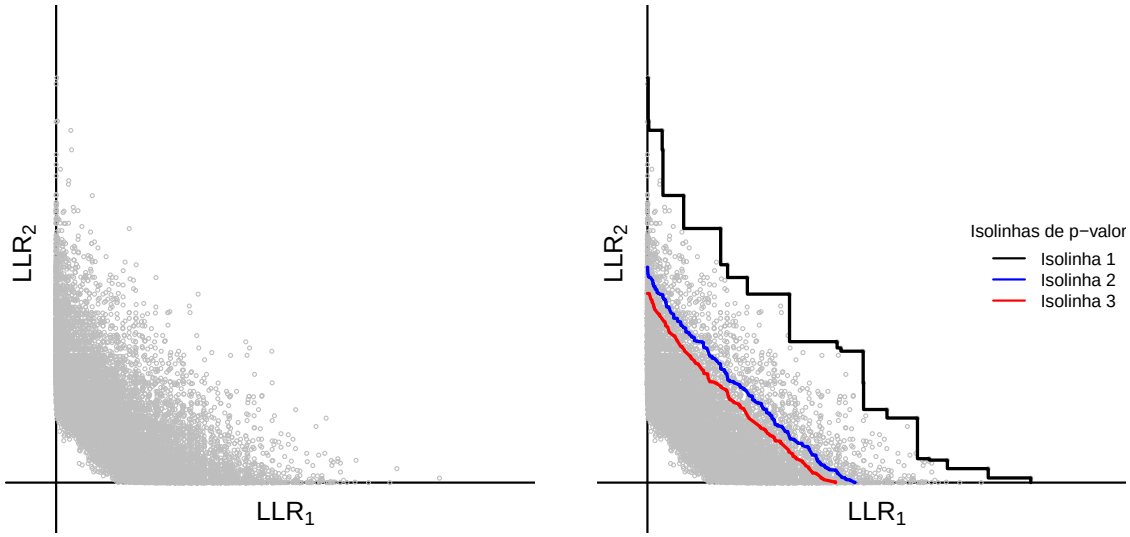



Figura 5.1: Pontos do conjunto de pareto (•).

Na Figura 5.1 observam-se os valores de $LLR_i,\ i = 1, 2$ obtida para cada *cluster* candidato. Os pontos pertencentes ao conjunto de Pareto são apenas *clusters* candidatos. Para se afirmar se algum destes *clusters* candidatos é, ou não, um *cluster* detectado, deve-se testar a sua significância estatística. Nossa preocupação, então será atribuir um p-valor para cada *cluster* do Pareto nos mapas de casos observados.

Para avaliar a significância de cada *cluster* pertencente ao Pareto obtido nos mapas de casos observados usamos o procedimento de Monte Carlo. Sob a hipótese nula de que não existe *cluster* no mapa, distribuímos, para cada conjunto de dados $i,\ i = 1, 2$, aleatoriamente através da distribuição multinomial os $n_{ij}$ casos sobre o mapa. O número esperado de casos em cada região, para cada conjunto de dados $i,\ i = 1, 2$ é proporcional à população de cada região. Aplicamos o algoritmo *scan* circular multiobjetivo descrito anteriormente para analisar este novo mapa de casos simulados, e como resposta, obtemos um conjunto de Pareto. Repetimos esse procedimento um número grande de vezes, obtendo assim vários conjuntos de Pareto.

Esses conjuntos de Pareto são agrupados, formando uma coleção de milhares de pontos distribuídos no espaço $LLR_1(z) \times LLR_2(z)$, conforme pode ser visto na figura 5.2.




(a) (b)

Figura 5.2: (a) Os mil conjuntos Pareto-ótimo obtidos por simulações de Monte Carlo sob a hipótese nula são representados pelos pontos; (b) são apresentadas as isolinhas de p-valor: pontos localizados acima da isolinha 1 tem p-valor menor que 0,000999001, um ponto localizado entre a isolinha 1 e a isolinha 2 tem p-valor entre 0,001 e 0,05 e um ponto localizado entre a isolinha 2 e a isolinha 3 tem p-valor entre 0,01 e 0,05.

Como este é um problema multiobjetivo, precisa-se de um procedimento para se calcular a significância estatística para estas soluções. Nesse caso, ao invés de encontrar o ponto crítico, acima do qual considera-se que um *cluster* é significativo, deve-se encontrar uma curva crítica. Essa 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.

Para se encontrar esta curva crítica será adotado o método proposto por Cançado *et al.* (2010), onde o conceito usual de significância é estendido de forma natural para o problema multiobjetivo através do conceito de isolinhas de p-valor. Essas isolinhas são calculadas através da função de aproveitamento, já descrita nesta tese. Para se obter as superfícies de aproveitamento será utilizado o algoritmo descrito em Knowles (2005).

A Figura 5.2(b) ilustra estas aproximações de algumas isolinhas de p-valor resultantes de $n = 1.000$ conjuntos Pareto-ótimos obtidos por simulações de Monte Carlo sob a hipótese nula.

A seguir será apresentado o algoritmo de detecção de *cluster* proposto nesta tese. Considerando um mapa de ocorrências aleatórias dividido em $m$ regiões, define-se o centroide de cada região como um ponto arbitrariamente escolhido nesta região do mapa.

1. De acordo com o algoritmo descrito na seção 2.3 encontre a matriz de distância entre as $m$ regiões em que cada linha $i$ da matriz representa as distâncias ordenadas em ordem crescente entre o centroide da região $i$ e os demais centroides das regiões do mapa;

2. Centrada em cada uma das regiões será construído um círculo de raio variável, de tal forma que, o raio desse círculo aumente de acordo com as distâncias crescentes até que a população das regiões englobadas pelo círculo atinja um percentual máximo pré estabelecido da população total, ou certo número de regiões pré estabelecidos anteriormente. Para cada círculo, que chamaremos aqui de zona $z$, o número de casos e a população para cada conjunto de dados $i$, $i = 1, 2$ são atualizados e calcula-se o logaritmo da razão de verossimilhança para cada conjunto de dados $LLR_1(z)$ e $LLR_2(z)$;

3. Calcule o conjunto de Pareto para $LLR_1(z)$ e $LLR_2(z)$;

4. Faça $B$ simulações de Monte Carlo e obtenha, para cada uma das $B$ simulações de Monte Carlo sob $H_0$, o conjunto de Pareto;

5. Utilize os $B$ conjuntos de Pareto para obter as isolinhas de p-valor através das superfícies de aproveitamento;

6. Avalie os pontos $LLR_1(z)$ e $LLR_2(z)$ pertencentes ao conjunto de Pareto para os dados observados, através das isolinhas de p-valor;

7. Rejeite a hipótese nula se existir pelo menos uma zona $z^*$, tal que o ponto $(LLR_1(z^*), LLR_2(z^*))$ estiver em uma região correspondente a um p-valor inferior ao nível de significância pré fixado.

Este trabalho comparará o método proposto para se detectar *clusters* na presença de duas fontes de dados distintas com três outros métodos já descritos anteriormente. Precisa-se, de alguma maneira, avaliar a eficiência do algoritmo proposto. Uma das medidas usadas é o poder do teste. Além desta medida, no próximo capítulo serão definidos os conceitos de sensibilidade e o valor de predição positiva

# Capítulo 6

# Medidas de eficiência de um método de detecção de *clusters* espaciais

Quando se opta por um método de detecção de *clusters* espera-se que ele seja bom o suficiente para encontrar um *cluster*, quando este existe. Para se avaliar a qualidade dos algoritmos apresentados nesta tese serão calculados o seu poder de detecção, a sua sensibilidade e o seu valor de predição positiva.

**Definição 6.1** (Poder do teste)**.** *O poder de um teste de hipóteses é a probabilidade de rejeição da hipótese nula quando a hipótese nula de fato é falsa.*

Em outras palavras, o poder do teste mede a habilidade de um teste detectar corretamente uma hipótese alternativa. O poder deste método é, então, a capacidade do algoritmo de encontrar um *cluster* quando ele realmente existe. Neste trabalho o poder será estimado através de execuções de Monte Carlo, executando o algoritmo várias vezes em cenários artificiais, feitos de forma que neles há presença de um *cluster* para cada um dos conjuntos de dados.

Para o algoritmo multiobjetivo serão obtidos, para cada simulação de Monte Carlo, o conjunto de pontos não dominados de acordo com o que foi descrito na seção 1.1. Estes conjuntos, obtidos nas simulações de Monte Carlo, serão comparados com a isolinha de p-valor 0.05, que será obtida por meio de simulações Monte Carlo sob a hipótese de que não há *cluster* no mapa. A proporção de conjuntos não dominados que tem pelo menos um ponto localizado acima da isolinha de p-valor de 0.05 é uma estimativa do poder do algoritmo para cada um dos cenários fixados de acordo com Cançado *et al.* (2010).

Para os algoritmos da seção 3 e para cada simulação Monte Carlo executada sob a hipótese de que há um *cluster* em cada banco de dados, o poder do algoritmo será a proporção do número de vezes que um dos *clusters* foi detectado no mapa em estudo sob o total de execuções de Monte Carlo.

Duas outras medidas são extensivamente usadas para avaliação da eficácia de uma algoritmo de detecção de *cluster*: a *sensibilidade* e o *valor de predição positivo* (*VPP*).

**Definição 6.2** (Sensibilidade)**.** *É a proporção de regiões do cluster real que pertencem ao cluster detectado.*

Consideremos a existência de $n$ casos no mapa, que são denotadas por $c_i,\ i = 1, \ldots, n$.

Então, a sensibilidade fica definida aqui, como:

$$\text{Sensibilidade} = \frac{\sum_{i=1}^{n} \mathbb{1}(c_i \in \textit{Cluster Detectado } \cap \textit{Cluster Real})}{\sum_{i=1}^{n} \mathbb{1}(c_i \in \textit{Cluster Real })}, \tag{6.1}$$

em que $\mathbb{1}(\cdot)$ é uma função indicadora de um evento.

**Definição 6.3** (Valor Positivo Preditivo ( *VPP*))**.** *É a proporção do número de regiões do cluster detectado que pertencem realmente ao cluster verdadeiro entre todos as regiões de um cluster detectado.*

O *VPP* expressa a probabilidade de que um *cluster* detectado venha a ser um *cluster* real. Neste trabalho, o *VPP* é assim definido:

$$\text{VPP} = \frac{\sum_{i=1}^{n} \mathbb{1}(c_i \in \textit{Cluster}\text{Detectado } \cap \textit{Cluster}\text{ Real })}{\sum_{i=1}^{n} \mathbb{1}(c_i \in \textit{Cluster}\text{ Detectado})} \tag{6.2}$$

Como trabalhamos com duas fontes de dados, temos que definir a sensibilidade e o *VPP* levando isto em conta. Assim, para duas fontes distintas de dados e, considerando que existem $n_1$ casos no primeiro conjunto de dados, denotados por $R_{1i},\ i = 1, \ldots, n_1$, e que existam $n_2$ casos conjunto de dados, que serão denotados por $R_{2j},\ j = 1, \ldots, n_2$. Então, a sensibilidade e o *VPP* ficam definidos por:

$$\begin{aligned} \text{Sensibilidade} &= 0.5 \cdot \frac{\sum_{i=1}^{n_1} \mathbb{1}(c_{1i} \in \textit{Cluster Detectado } \cap \textit{Cluster Verdadeiro } n^o 1)}{\sum_{i=1}^{n_1} \mathbb{1}(c_{1i} \in \textit{Cluster Verdadeiro } n^o \textit{ 1 })} + \\ &+ 0.5 \cdot \frac{\sum_{j=1}^{n_2} \mathbb{1}(c_{2i} \in \textit{Cluster Detectado} \cap \textit{Cluster Verdadeiro } n^o 2)}{\sum_{j=1}^{n_2} \mathbb{1}(c_{2i} \in \textit{Cluster Verdadeiro } n^o 2)}, \end{aligned} \tag{6.3}$$

e:

$$\begin{aligned} \textit{VPP} &= 0.5 \cdot \frac{\sum_{i=1}^{n_1} \mathbb{1}(c_{1i} \in \textit{Cluster Detectado } \cap \textit{Cluster verdadeiro } n^o 1)}{\sum_{i=1}^{n_1} \mathbb{1}(c_{1i} \in \textit{Cluster Detectado})} + \\ &+ 0.5 \cdot \frac{\sum_{j=1}^{n_2} \mathbb{1}(c_{2i} \in \textit{Cluster Detectado} \cap \textit{Cluster verdadeiro } n^o 2)}{\sum_{j=1}^{n_2} \mathbb{1}(c_{2i} \in \textit{Cluster Detectado})}. \end{aligned} \tag{6.4}$$

# Capítulo 7

# Resultados das Simulações

Nesta seção serão comparados o desempenho do método multiobjetivo com outros três métodos descritos na capítulo 3: $\mathbf{llr_1} + \mathbf{llr_2}$ consiste calcular a razão log verossimilhança para cada um dos conjuntos de dados e, dentro de cada zona, somar estas razões de log-verossimilhanças; $\mathbf{max}(\mathbf{llr_1}, \mathbf{llr_2})$ consiste em calcular a razão de log-verossimilhança definida por Kulldorff (1997) para todos os conjuntos de dados e, após isto, para cada zona no mapa, é obtida a razão de log-verossimilhança máxima entre todos as fontes de dados em cada zona no mapa, e o método denotado por *naive* é um método que consiste em combinar os conjuntos de dados, através da soma da população em risco e os casos em cada um dos conjuntos de dados.

## 7.1 *Clusters* com formato circular

Para a estimação do poder, da sensibilidade e do *VPP* do algoritmo multiobjetivo e dos três outros métodos, utilizou-se quatro situações diferentes de acordo com a Figura 7.1, para três diferentes cenários, que serão descritos nas próximas seções.

Para cada um destes, em cada uma das duas fontes de dados, foi gerado um *cluster* artificial em um mapa formado por duzentas e três regiões, com um total dois mil e trinta casos.

Estes *clusters* serão denotados por *clusters* reais e os *clusters* detectados são os *clusters* encontrados pelos algoritmos aqui considerados.

1. *clusters* sobrepostos: *clusters* artificiais, que foram construídos por meio dos bancos de dados, coincidem;

2. *Clusters* com grande sobreposição: Os *clusters* artificiais foram construídos com uma grande interseção entre eles;

3. *Clusters* com pouca sobreposição: Os *clusters* artificias apresentam uma pequena interseção entre eles;

4. *Clusters* sem sobreposição: Os *clusters* foram construídos de forma a não ter interseção entre eles.

Em cada simulação, para cada cenário, os dois mil e trinta casos foram distribuídos, em cada fonte de dados, fixado um risco relativo esperado igual para todas as regiões

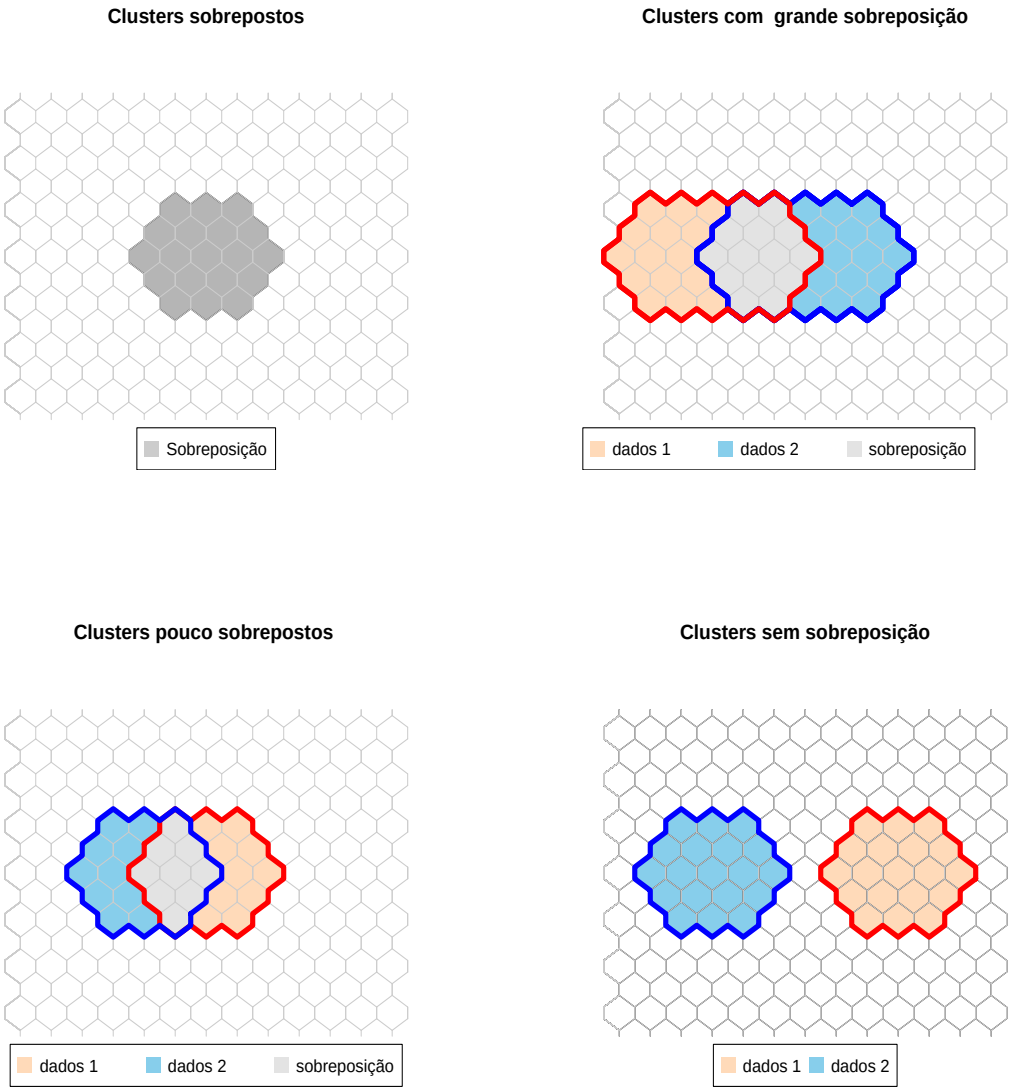



Figura 7.1: Da esquerda para a direita de cima para baixo: Dois *clusters* sobrepostos; *Clusters* com uma grande região de interseção; *Clusters* com uma interseção moderada; *Clusters* totalmente separados no mapa.

que não fazem parte do *cluster*, e um risco relativo maior do que um para as regiões que formam um *cluster* de acordo com Kulldorff *et al.* (2003).

Foram feitas, para cada cenário, vinte mil simulações sob a hipótese alternativa e, para cada uma destas simulações, foram produzidos conjuntos de pontos não dominados de acordo com o algoritmo multiobjetivo descrito por Cançado *et al.* (2010). Estes conjuntos foram comparados com a isolinha de p-valor 0.05, obtida por meio de vinte mil simulações Monte Carlo sob a hipótese nula. A proporção de conjuntos não dominados que têm pelo menos um ponto localizado acima da isolinha de p-valor de 0.05 é uma estimativa do poder do algoritmo para cada um dos cenários fixados acima. As medidas de sensibilidade e *VPP* foram obtidos para o *cluster* com menor p-valor dentre os pontos não dominados em cada simulação.

### 7.1.1 Cenário 1

Para este cenário, o valor de risco relativo ($r$), para cada um dos bancos de dados, é escolhido de forma que se tenha uma probabilidade de 0,99 de que em cada distribuição aleatória se forme um *cluster* exatamente nas regiões com risco $r$, ou seja, para cada um dos *clusters* $P(T_v > T_{crit}) = 0,99$, em que $T_v$ é o valor da estatística $T$ para o *cluster* verdadeiro (formado pelas regiões com risco $r$) (Kulldorff *et al.*, 2003).

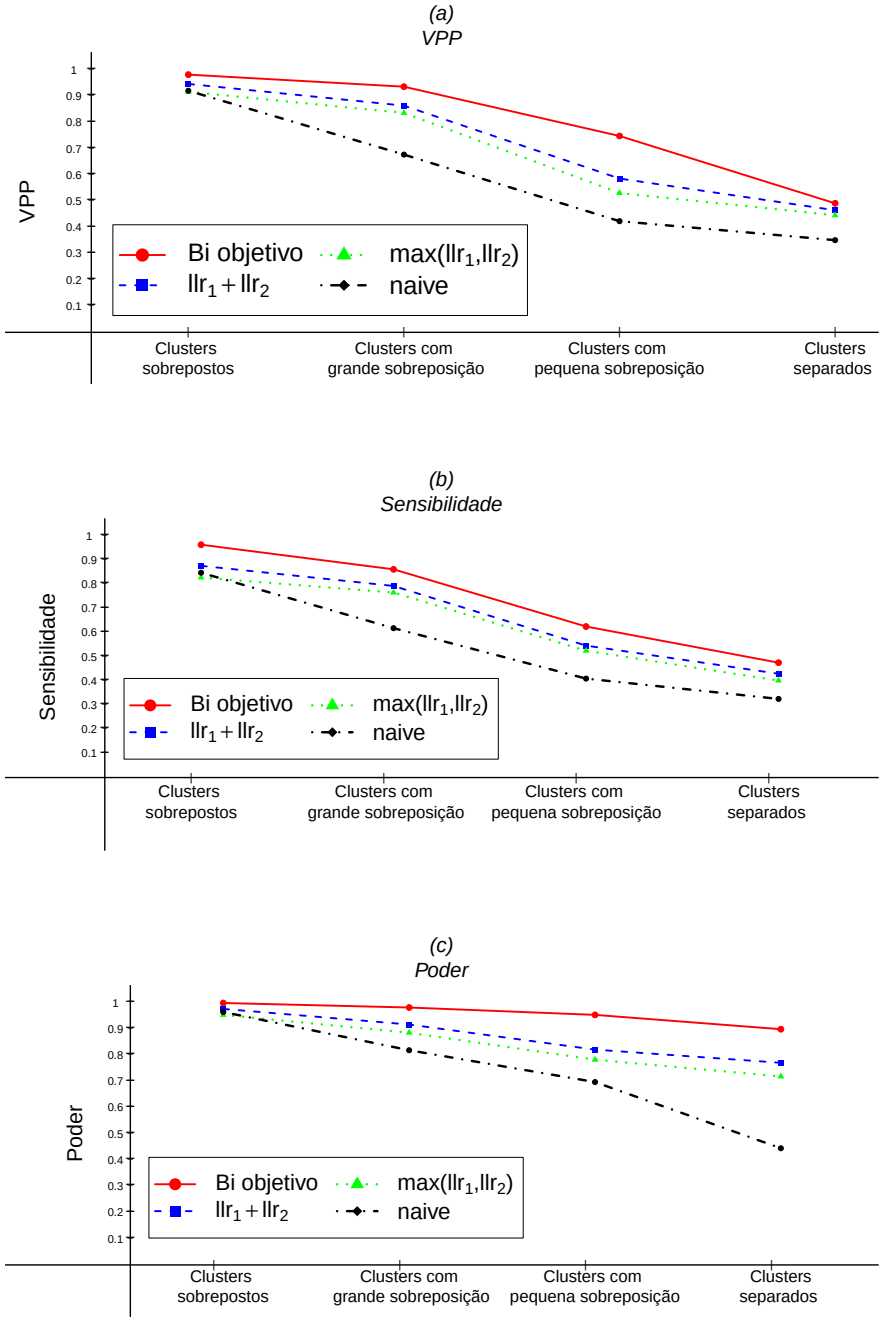



Figura 7.2: (a) Comparação do *VPP* entre os métodos para as combinações de *clusters*; (b) Comparação da sensibilidade entre os métodos para as combinações de *clusters* e (c) Comparação do poder do teste entre os métodos para as combinações de *clusters*;.

### 7.1.2 Cenário 2

Assim como na seção 7.1.1, a probabilidade de haver um *cluster* em cada uma das fontes de dados é a mesma, porém, a população sob risco de cada uma das 203 regiões

do conjunto de dados um é igual a dez mil e a população sob risco de cada uma das 203 regiões do conjunto de dados dois é igual a mil. Um exemplo de situação onde isto pode ocorrer é o estudo de uma doença em uma população de homens e em uma população de mulheres.

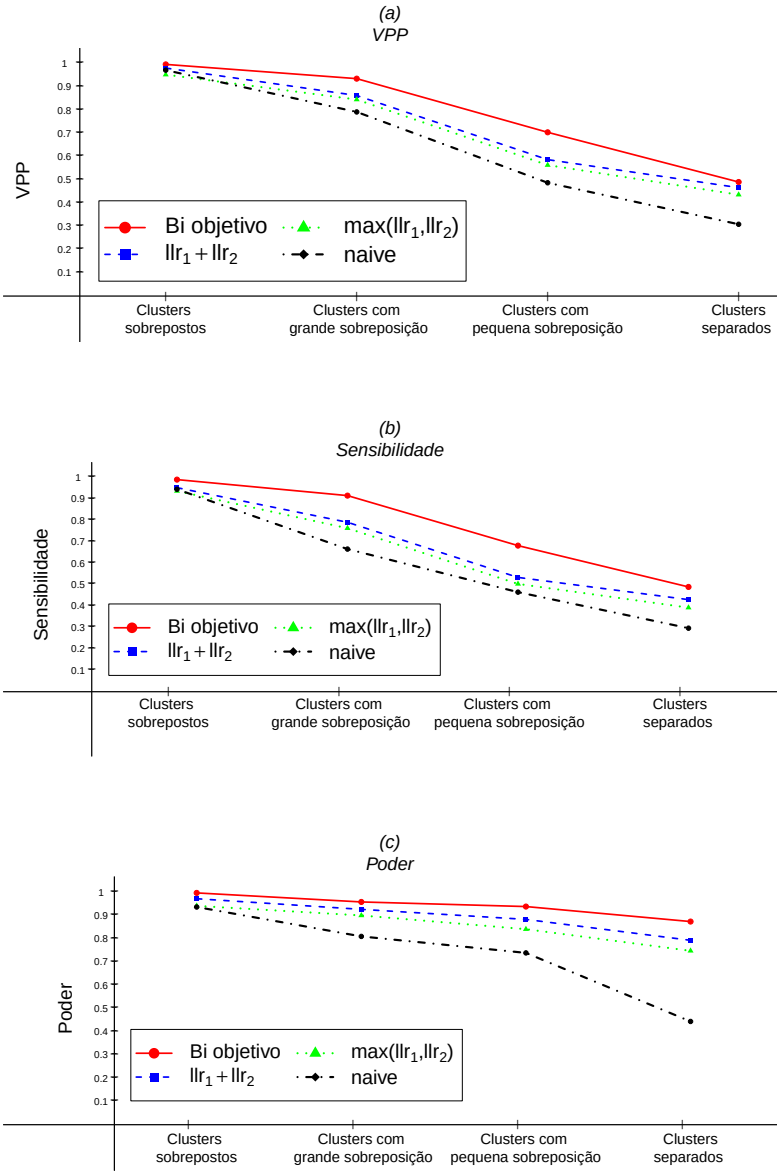



Figura 7.3: (a) Comparação do *VPP* entre os métodos para as combinações de *clusters*; (b) Comparação da sensibilidade entre os métodos para as combinações de *clusters* e (c) Comparação do poder do teste entre os métodos para as combinações de *clusters*;.

### 7.1.3 Cenário 3

Nesta seção a população sob risco é a mesma nos dois conjuntos de dados: para cada uma das 203 regiões do conjunto de dados é igual a 1.000 indivíduos. Porém, neste conjunto de simulações a probabilidade de haver um *cluster* em cada um dos bancos de dados é diferente: na primeira fonte de dados risco relativo foi escolhido de forma que se tenha uma probabilidade de 0,99 de que em cada distribuição aleatória se forme um *cluster* exatamente nas regiões com risco $r_1$ e, para o conjunto de dados 2, risco relativo foi escolhido para que se tenha uma probabilidade de 0,9 de que em cada distribuição aleatória se forme um *cluster* exatamente nas regiões com risco $r_2$.

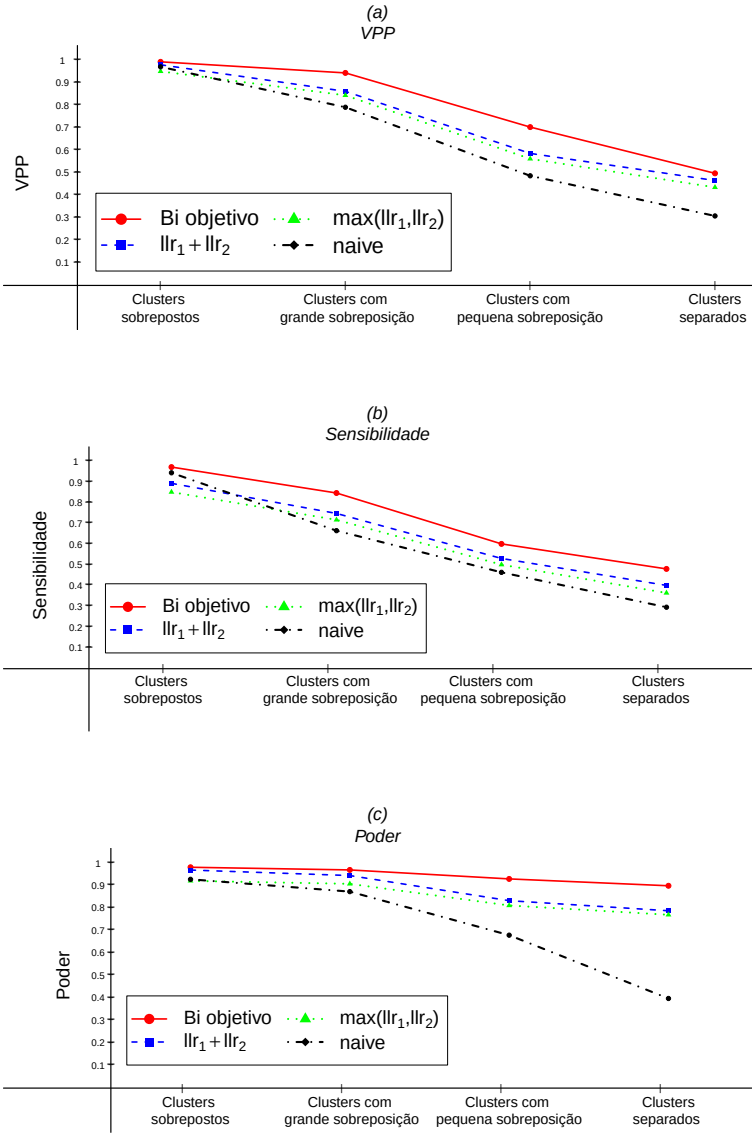



Figura 7.4: (a) Comparação do *VPP* entre os métodos para as combinações de *clusters*; (b) Comparação da sensibilidade entre os métodos para as combinações de *clusters* e (c) Comparação do poder do teste entre os métodos para as combinações de *clusters*;.

Para os métodos do capítulo 3 o poder, a sensibilidade e o *VPP* foram obtidas, em cada simulação, para o *cluster* com o menor p-valor.

### 7.1.4 Avaliação dos resultados

**Poder**

O método multiobjetivo foi superior aos três outros métodos.O poder de detecção de todos os métodos decresce à medida que a interseção entre os *clusters* dos dois conjuntos de dados torna-se menor. Entre os métodos mencionados na capítulo 3 não é possível indicar um melhor método, porém há que se destacar que o desempenho da norma da soma cai muito quando os *clusters* estão completamente separados. Isto se deve ao fato de que quando os conjuntos de dados são adicionados tomando a soma das contagens da população em risco e dos casos, um *cluster* presente em um conjunto de dados pode não ser detectado devido ao erro aleatório presente nos outros conjuntos de dados.

**Sensibilidade**

Quanto a sensibilidade método multiobjetivo mostrou uma bom resultado, sendo, novamente, melhor que todos os outros métodos. À medida que a interseção vai diminuindo entre os *clusters*, o método multiobjetivo mantém a sua boa capacidade de detecção ficando próximo de valores teóricos ótimos, como por exemplo, o valor um para a total interseção entre os *clusters* e 0.5 quando os *clusters* estão completamente separados.

*VPP*

Para o *VPP* pode-se observar que o algoritmo multiobjetivo apresenta uma probabilidade de que um *cluster* detectado venha a ser um *cluster* real próximo do que seria o teoricamente esperado. Além deste fato, o algoritmo multiobjetivo foi, novamente, superior aos outros três métodos aqui estudados. Os métodos apresentados na capítulo 3 têm desempenho semelhante quanto ao *VPP*, porém quando há pouca ou nenhuma interseção, a norma da soma se mostra o pior destes métodos.

Podemos concluir que o algoritmo multiobjetivo tem várias vantagens: a representação da função de avaliação para cada fonte de dados é muito clara, e não sofre de uma artificialidade, e, possivelmente, de uma confusão provocada pela mistura com as outras funções de avaliações dos diversos conjuntos de dados; é possível atribuir, de forma rigorosa, a significância estatística de cada *cluster* candidato e é possível analisar e selecionar a melhor solução de *cluster*, dentre aqueles fornecidos pelo conjunto não-dominado.

## 7.2 *Clusters* com formato não circular

Para este conjunto de simulações foram construídos clusters artificiais de formato não circular, para cada uma, das duas fontes de dados, em um mapa formado por 203 regiões, com um total 2.030 casos.

Em cada uma das 20.000 simulações, sob a hipótese alternativa, os 2.030 casos foram distribuídos, em cada fonte de dados, fixado um risco relativo esperado igual para todas as regiões que não fazem parte do *cluster*, e um risco relativo maior do que um para as regiões que formam um *cluster*. A população sob risco é a mesma nos dois conjuntos de dados: para cada uma das 203 regiões do conjunto de dados a população é igual a 1.000 indivíduos. O risco relativo, em cada um dos bancos de dados, foi escolhido de forma que se tenha uma probabilidade de 0,99 de que em cada distribuição aleatória se forme um *cluster* exatamente nas regiões com risco $r_j,\ j = 1, 2$.

Para cada uma destas simulações foram produzidos conjuntos de pontos não dominados. Estes conjuntos foram comparados com a isolinha de p-valor 0,05, obtida por meio de 20.000 simulações Monte Carlo sob a hipótese nula, e a proporção de conjuntos não dominados que tem pelo menos um ponto localizado acima da isolinha de p-valor de 0,05 foi usada com uma estimativa do poder do algoritmo. As medidas de sensibilidade e *VPP* foram obtidas para o *cluster* com menor p-valor dentre os pontos não dominados em cada simulação.

Para a estimação do poder, da sensibilidade e do *VPP* do algoritmo multiobjetivo e dos três outros métodos, utilizaram-se quatro situações diferentes de acordo com a Figura 7.5, para três diferentes cenários, que serão descritos nas próximas seções.

Estes *clusters* serão denotados por *clusters* reais e os *clusters* detectados são os *clusters* encontrados pelos algoritmos aqui considerados.

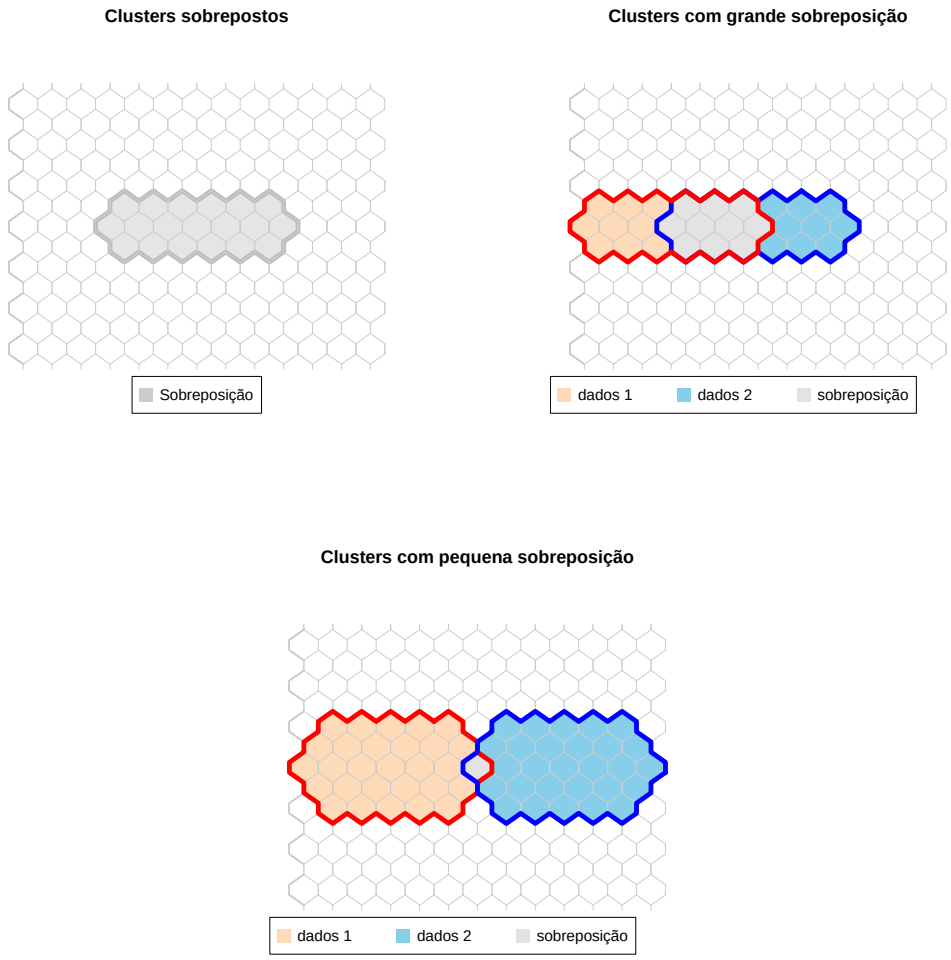



Figura 7.5: Da esquerda para direita e de cima para baixo: dois *clusters* sobrepostos; *clusters* com uma grande região de interseção; *clusters* com uma interseção moderada.

1. Os *clusters* artificiais que foram construídos por meio dos bancos de dados coincidem;
2. Os *clusters* artificiais foram construídos com uma grande interseção entre eles;
3. Os *clusters* artificias apresentam uma pequena interseção entre eles.

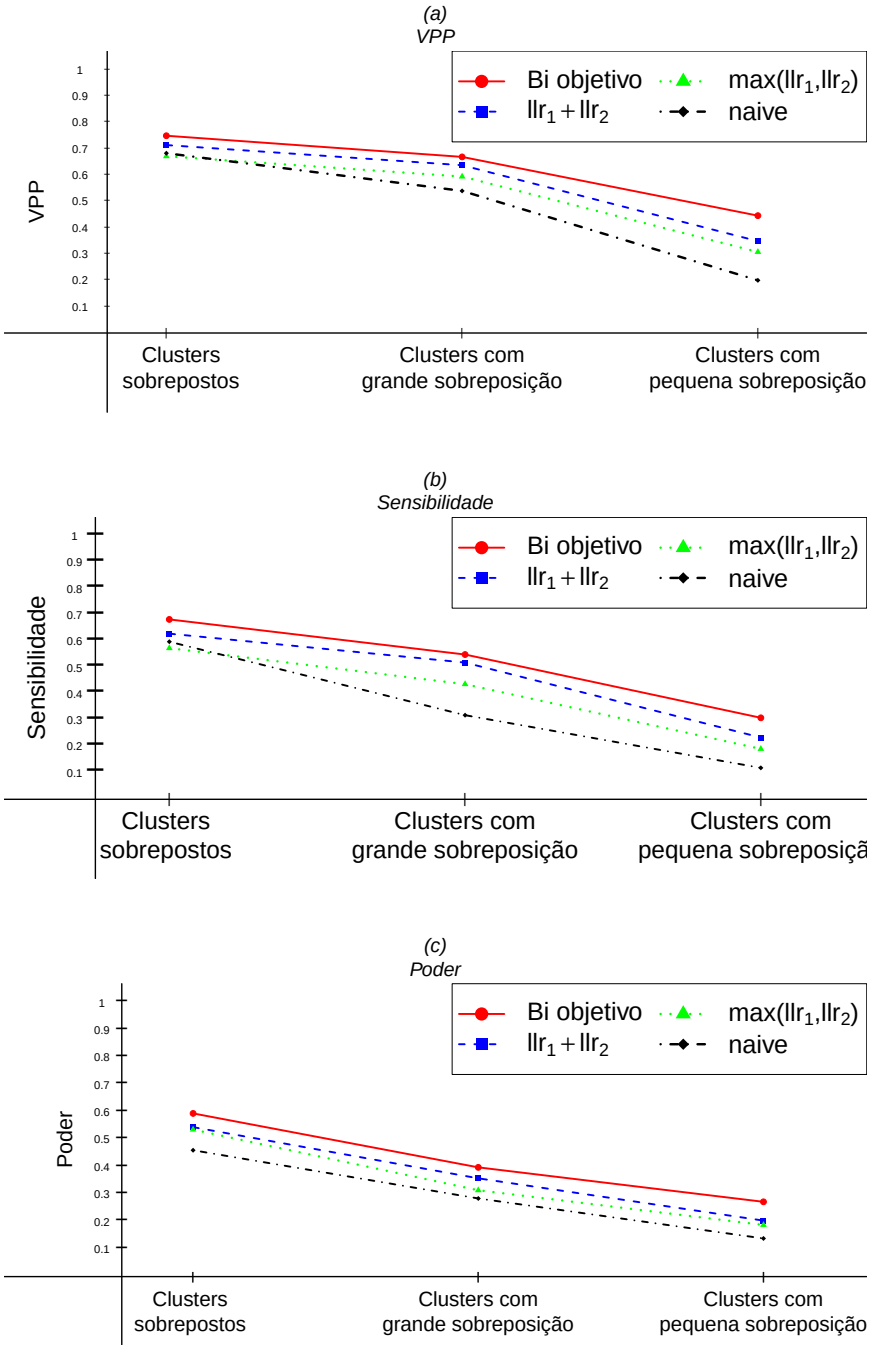



Figura 7.6: (a) Comparação do *VPP* entre os métodos para as combinações de *clusters*; (b) Comparação da sensibilidade entre os métodos para as combinações de *clusters* e (c) Comparação do poder do teste entre os métodos para as combinações de *clusters*;.

### 7.2.1 Avaliação dos resultados

Nestas simulações o método multiobjetivo, também foi superior aos três outros métodos em todas as situações. Novamente, o poder de detecção de todos os métodos

vai caindo à medida que a interseção entre os *clusters* dos dois conjunto de dados vai diminuindo e entre todos os métodos descritos na capítulo 3 não é possível indicar um melhor método, porém há que se destacar que o desempenho da norma da soma (*naive*) cai muito o seu desempenho quando os *clusters* estão completamente separados.

O método multiobjetivo demonstrou que tem uma excelente capacidade de detectar um *cluster* quando ele de fato existe no mapa em estudo, apresentando um melhor desempenho em relação a todos os outros métodos. À medida que a interseção vai diminuindo entre os *clusters*, o método multiobjetivo mantém a sua boa capacidade de detecção ficando próximo de valores teóricos.

O algoritmo multiobjetivo apresenta uma boa probabilidade de que um *cluster* detectado venha a ser um *cluster* real perto do que seria teoricamente esperado, obtendo valor superior de *VPP* em relação aos outros três métodos aqui estudados. Os métodos apresentados na capítulo 3 tem desempenho semelhante quanto ao *VPP*, porém quando há pouca ou nenhuma interseção a norma da soma (*naive*) se mostra o pior destes métodos.

# Capítulo 8

# Aplicações

## 8.1 Óbitos por Dengue/Febre Amarela e Óbitos por Malária na região Norte do Brasil

Diversos órgãos públicos brasileiros têm tido uma política de abertura no sentido de disponibilizar dados. Talvez isso tenha ocorrido devido à força do argumento de que dados coletados com recursos públicos também deveriam ser de domínio público, ou seja, dados públicos deveriam ser publicados e acessíveis livremente. A importância de se ter dados públicos é a de que qualquer pessoa ou órgão pode utilizá-los em pesquisas e para subsídio de políticas públicas.

O Ministério da Saúde também tem disponibilizado sistematicamente uma variedade enorme de informações de saúde com referência geográfica, (*DATASUS*). São disponibilizados micro dados do registro de nascimento e morte, micro dados de atendimento ambulatorial, internações, etc. Nos dados de nascimento e morte há referência de município.

O método proposto nesta tese será aplicado em dois conjuntos de dados de mortalidade obtidos no site do *DATASUS* para os municípios da região norte do Brasil no período de 2006 a 2011: Um dos conjuntos de dados é sobre mortalidade por dengue ou febre amarela, sendo apresentados, ao todo, 132 casos. O outro conjunto de dados é sobre mortalidade por malária sendo, ao todo, 352 casos.

A motivação para a escolha dos bancos de dados utilizados nesta seção é que tanto a malária quanto a dengue e a febre amarela são transmitidas por mosquitos infectados. No caso da malária, o mosquito *Anopheles* é infectado por um protozoário, o *Plasmodium*, e o mosquito que transmite a dengue, o *Aedes aegypti* é infectado por um vírus.

A malária é endêmica em algumas regiões da África e também na região norte do Brasil e em relação aos sintomas, as duas doenças apresentam febre, dor no corpo, dor de cabeça e tremores. Uma particularidade da malária é que a febre vem por períodos, conhecida como "febre terçã" ou "quartã", que é quando o protozoário rompe a hemácia parasitada e a pessoa sente tremores, febre e calafrios. Isto faz com que na apresentação clínica, os sintomas das duas doenças possam ser confundidos pois a febre terçã pode não ocorrer em períodos tão certos de tempo.

A Figura 8.1 mostra o conjunto Pareto-ótimo obtido pelo algoritmo multiobjetivo, constituído por 18 soluções. Um resumo destas soluções é apresentado na Tabela 8.1, em que são encontrados os valores de $LLR$ relativo aos casos de mortalidade por dengue e

febre amarela e os valores de $LLR$ relativos aos casos de mortalidade por malária.

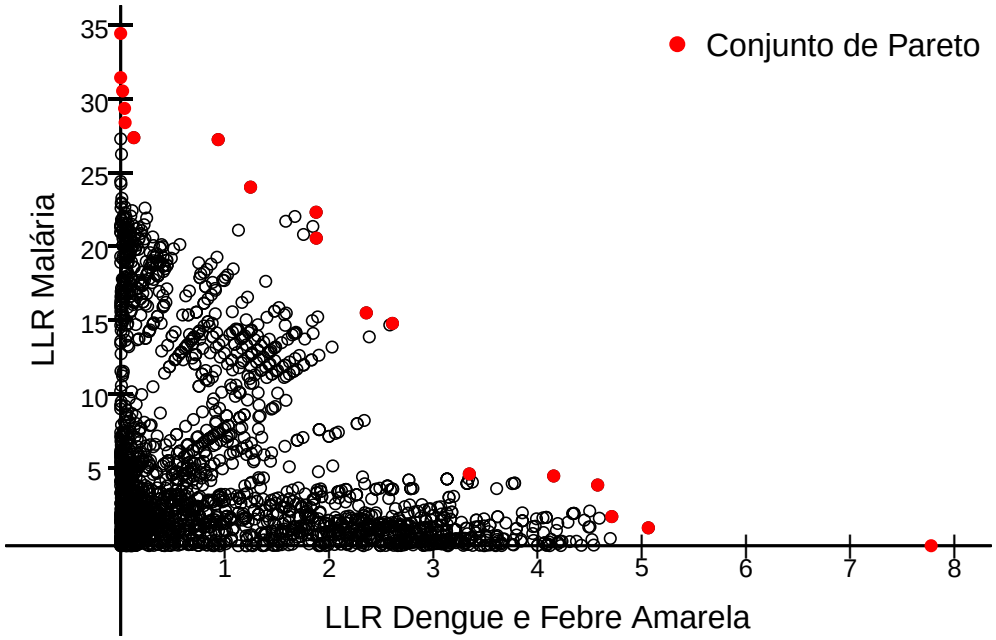



Figura 8.1: Conjunto Pareto-ótimo encontrado para os casos de dengue/febre amarela e os casos de malária na região norte do Brasil.

Tabela 8.1: Resumo dos *clusters* para os casos de dengue/febre amarela e malária para o Norte do Brasil.

| Cluster | Casos dengue | LLR Dengue | Casos malaria | LLR malaria | População |
|---|---|---|---|---|---|
| 1 | 5 | 0 | 56 | 34,7053 | 345428 |
| 2 | 2 | 0,0000068428 | 32 | 31,71304 | 120972 |
| 3 | 4 | 0,01888125 | 42 | 30,80548 | 220077 |
| 4 | 8 | 0,0376027 | 60 | 29,62827 | 440888 |
| 5 | 8 | 0,04298807 | 59 | 28,66471 | 437923 |
| 6 | 9 | 0,1262316 | 60 | 27,64623 | 461940 |
| 7 | 16 | 0,935867 | 76 | 27,51359 | 688146 |
| 8 | 13 | 1,246272 | 60 | 24,29199 | 501064 |
| 9 | 18 | 1,875297 | 71 | 22,59385 | 684750 |
| 10 | 17 | 1,87685 | 66 | 20,83815 | 635985 |
| 11 | 17 | 2,356761 | 58 | 15,77664 | 596127 |
| 12 | 17 | 2,607384 | 56 | 15,06224 | 577520 |
| 13 | 4 | 3,345318 | 8 | 4,86506 | 47500 |
| 14 | 5 | 4,154287 | 9 | 4,725589 | 59965 |
| 15 | 5 | 4,576274 | 8 | 4,111954 | 54086 |
| 16 | 10 | 4,711731 | 15 | 1,975031 | 195891 |
| 17 | 10 | 5,06269 | 13 | 1,214041 | 186334 |
| 18 | 11 | 7,775771 | 2 | 0 | 157107 |

Foram feitas 10.000 réplicas sob a hipótese nula para se obter as isolinhas de p-valor que estão na Figura 8.2. Os pontos numerados de 1 a 12 na Figura 8.2 estão localizados acima da isolinha 1 e têm p-valor menor 0,0000999, os pontos numerados 13, 14, 15 e 18 estão localizados entre a isolinha 1 e a isolinha 2 e têm p-valor entre 0,05 e 0,0000999 e os pontos 16 e 17 estão localizados entre a isolinha 2 e a isolinha 3 e têm p-valor entre 0,05 e 0,1.

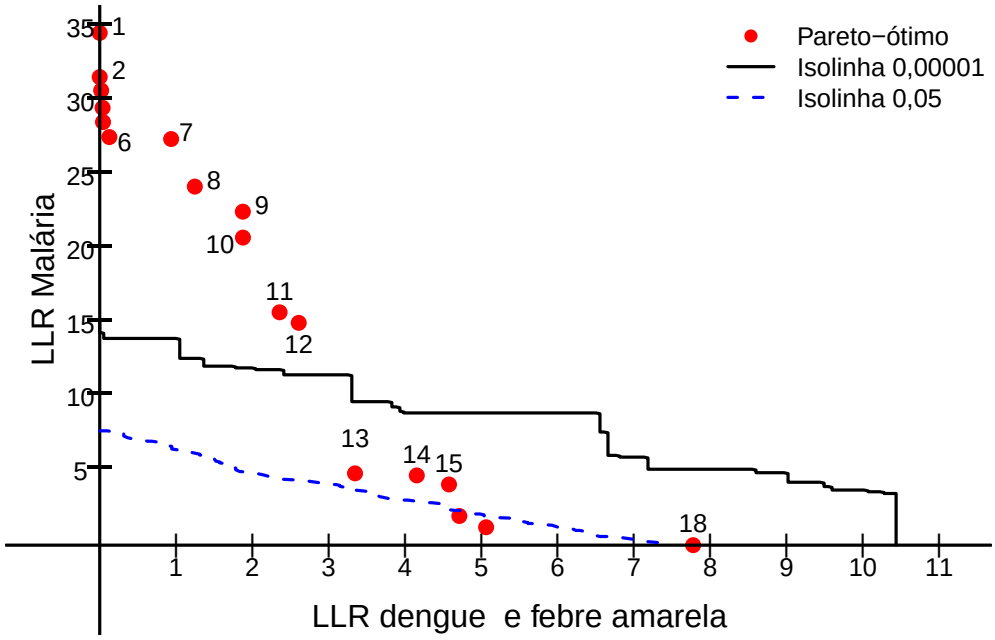



Figura 8.2: Isolinhas de p-valor.

As Figuras 8.3 e 8.4 mostram 8 *clusters* selecionados dentre os 18. Como dito anteriormente os *clusters* numerados de 1 a 12 na Figura 8.2 estão localizados acima da isolinha 1 e têm p-valor menor que 0,0000999, nas Figuras 8.3 e 8.4 podemos observar que os *clusters* 1 e 8 contêm municípios dos estados do Amazonas, Rondônia e do Pará, os *clusters* 7 e 9 contém municípios dos estados do Amazonas, Rondônia e do Amapá e o *cluster* 11 contêm municípios dos estados do Amazonas e de Rondônia.

Os *clusters* 13, 14, 15 e 18 têm p-valor entre 0,05 e 0,0000999. Na Figura 8.4 podemos observar que o *cluster* 13 contém apenas municípios do estado de Roraima e o *cluster* 13 contém apenas municípios do estado de Rondônia.

Os *clusters* 16 e 17 têm p-valor maior que 0,1 e, como pode ser visto na figura 8.4, o *cluster* 16 é composto por apenas 5 municípios do Amapá.

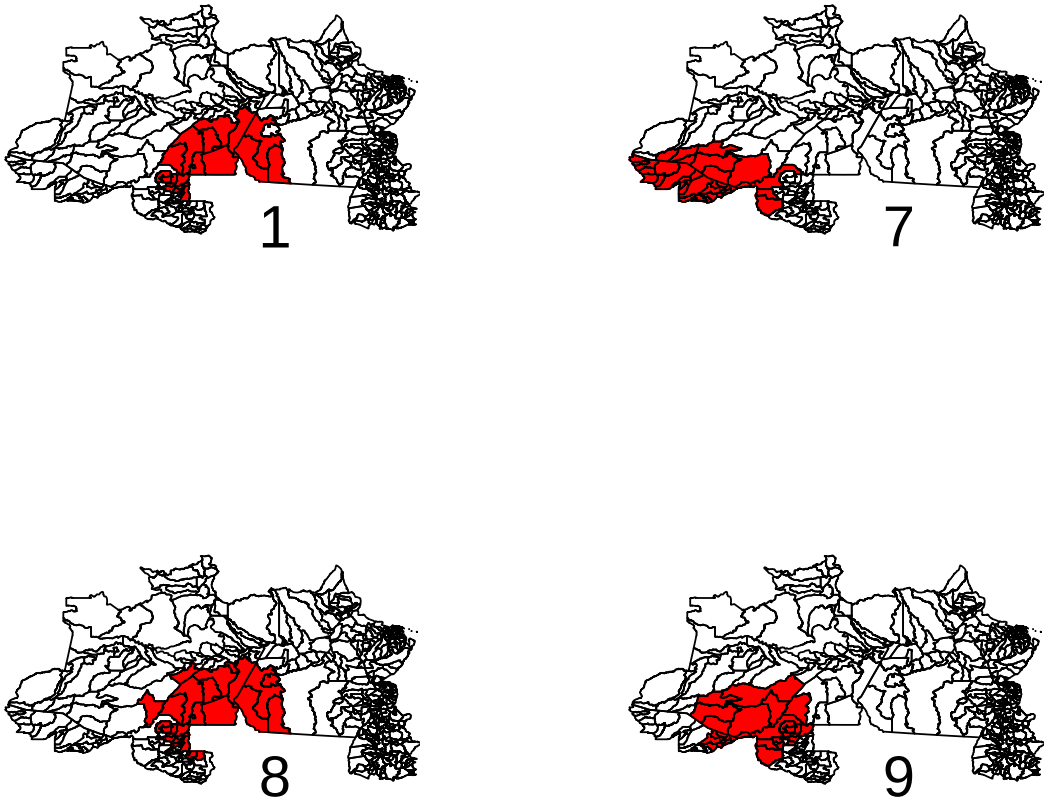



Figura 8.3: *Clusters* Detectados 1

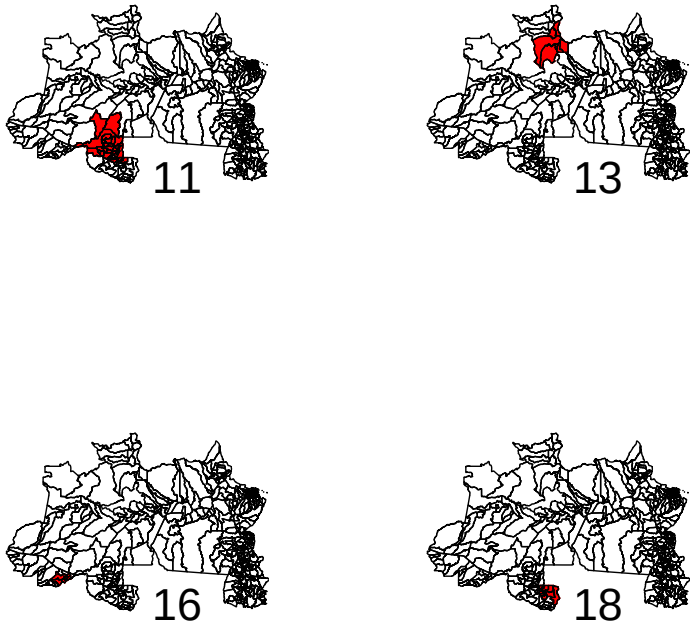



Figura 8.4: *Clusters* Detectados 2

## 8.2 Mortalidade por câncer de cérebro para adultos de 48 estados dos Estados Unidos da América

Nesta aplicação usaremos dois conjuntos de dados referentes às taxas de mortalidade por câncer de cérebro padronizados para adultos do sexo masculino e feminino para cada um dos 3.111 municípios em 48 estados contíguos dos Estados Unidos da América, de 1986 a 1995.

O conjunto não-dominado é inspecionado para observar possíveis correlações entre os dois mapas do agrupamento câncer no cérebro (Figura 8.5).

A Figura 8.6 mostra o conjunto Pareto-ótimo obtido pelo algoritmo multiobjetivo, constituído por 39 soluções. Um resumo de algumas destas soluções é apresentado na Tabela 8.2.

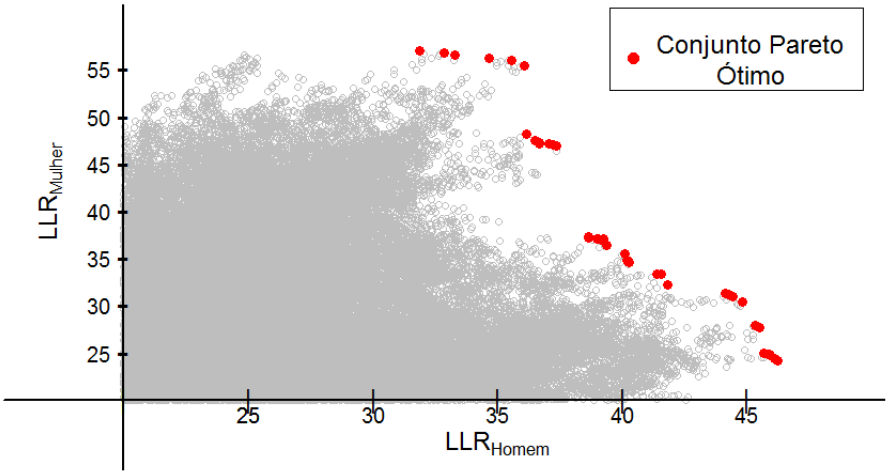



Figura 8.5: Parte da solução definida no espaço $LLR_{Homem} \times LLR_{Mulher}$ dos conjuntos de dados de câncer de cérebro para homens e mulheres de condados dos EUA. *Clusters* são indicadas por pontos cinza, com as soluções não-dominadas representados por círculos vermelhos.

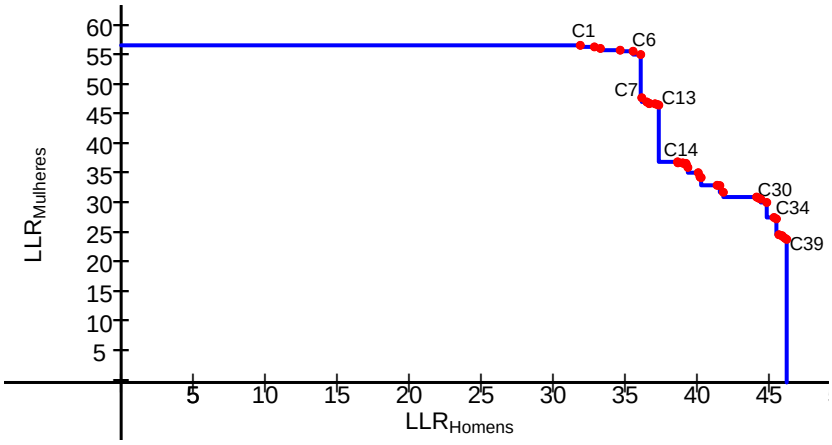



Figura 8.6: Conjunto Pareto-ótimo encontrado para os casos de câncer de cérebro para homens e mulheres, respectivamente.

A Figura 8.7 mostra 2 *clusters* selecionados: o mapa da esquerda, é referente ao *cluster* C5 na Tabela 8.2 e, podemos perceber, que este *cluster* tem um valor alto de $LLR$ no mapa

Tabela 8.2: Resumo dos *clusters* para os casos de câncer de cérebro de homens e mulheres adultos para condados dos Estados Unidos da América.

| *Cluster* | Homens | | | | Mulheres | | | |
|---|---|---|---|---|---|---|---|---|
| | *LLR* | Casos | População | Casos esperados | *LLR* | Casos | População | Casos esperados |
| C1 | 31,88 | 4281 | 53954822 | 3796,24 | 56,99 | 3865 | 61132087 | 3259,87 |
| C2 | 32,86 | 4212 | 52934738 | 3723,98 | 56,73 | 3801 | 60022011 | 3202,12 |
| C3 | 33,28 | 4246 | 53312340 | 3753,10 | 56,46 | 3825 | 60427150 | 3225,58 |
| C4 | 34,65 | 4194 | 52483967 | 3694,42 | 56,17 | 3770 | 59514394 | 3176,31 |
| C5 | 35,55 | 4185 | 52271363 | 3679,77 | 55,99 | 3757 | 59286739 | 3165,20 |
| C6 | 36,08 | 4136 | 51596355 | 3630,09 | 55,43 | 3709 | 58539771 | 3123,73 |
| C7 | 36,15 | 5888 | 77580398 | 5288,74 | 48,14 | 5147 | 87254309 | 4506,40 |
| C8 | 36,50 | 5858 | 77116370 | 5257,42 | 47,43 | 5114 | 86793104 | 4479,87 |
| C9 | 36,65 | 5861 | 77135036 | 5259,07 | 47,22 | 5114 | 86812570 | 4481,21 |
| C10 | 36,68 | 5863 | 77154553 | 5260,75 | 47,14 | 5115 | 86834816 | 4482,68 |
| C11 | 37,07 | 5813 | 76196987 | 5210,01 | 47,11 | 5074 | 85832416 | 4444,21 |
| C12 | 37,24 | 5813 | 76179912 | 5208,69 | 46,97 | 5072 | 85814758 | 4443,23 |
| C13 | 37,33 | 5812 | 76158967 | 5207,01 | 46,89 | 5070 | 85791750 | 4441,88 |
| C14 | 38,64 | 5526 | 73341544 | 4925,29 | 37,29 | 4783 | 82744233 | 4235,54 |
| C15 | 38,66 | 5528 | 73375445 | 4927,04 | 37,24 | 4784 | 82773222 | 4236,90 |
| C16 | 38,66 | 5539 | 73515782 | 4937,45 | 37,13 | 4793 | 82934883 | 4246,15 |
| C17 | 38,99 | 5512 | 73125072 | 4909,39 | 37,12 | 4767 | 82495121 | 4221,63 |
| C18 | 39,03 | 5506 | 73055673 | 4903,34 | 37,06 | 4761 | 82416824 | 4216,36 |
| C19 | 39,22 | 5510 | 73081309 | 4905,73 | 37,04 | 4763 | 82445772 | 4218,41 |
| C20 | 39,24 | 5497 | 72934809 | 4893,27 | 36,85 | 4750 | 82277253 | 4207,49 |
| C21 | 39,35 | 5494 | 72883664 | 4889,59 | 36,35 | 4743 | 82221187 | 4204,39 |
| C22 | 40,08 | 5386 | 71272101 | 4781,96 | 35,46 | 4636 | 80357042 | 4109,47 |
| C23 | 40,19 | 5395 | 71376306 | 4789,62 | 34,78 | 4638 | 80484201 | 4116,28 |
| C24 | 40,23 | 5396 | 71387380 | 4790,30 | 34,69 | 4638 | 80497764 | 4116,93 |
| C25 | 40,27 | 5336 | 70657717 | 4733,18 | 34,63 | 4584 | 79643573 | 4066,19 |
| C26 | 41,38 | 5472 | 72111463 | 4853,93 | 33,33 | 4691 | 81237972 | 4177,10 |
| C27 | 41,57 | 5481 | 72194534 | 4861,08 | 33,28 | 4697 | 81330677 | 4183,17 |
| C28 | 41,82 | 5457 | 71842127 | 4836,63 | 32,15 | 4666 | 80930958 | 4162,23 |
| C29 | 44,13 | 5464 | 72138962 | 4826,98 | 31,35 | 4646 | 81354259 | 4149,35 |
| C30 | 44,27 | 5468 | 72177582 | 4829,77 | 31,17 | 4647 | 81399359 | 4151,69 |
| C31 | 44,41 | 5462 | 72084956 | 4823,11 | 31,01 | 4640 | 81295983 | 4146,23 |
| C32 | 44,83 | 5439 | 71744089 | 4798,49 | 30,43 | 4615 | 80927159 | 4126,93 |
| C33 | 45,34 | 5595 | 73762684 | 4942,36 | 27,90 | 4721 | 83096105 | 4248,04 |
| C34 | 45,50 | 5610 | 73939824 | 4955,39 | 27,65 | 4730 | 83288037 | 4258,60 |
| C35 | 45,67 | 5705 | 75962272 | 5043,99 | 24,98 | 4776 | 85261225 | 4325,21 |
| C36 | 45,89 | 5697 | 75839669 | 5034,92 | 24,83 | 4767 | 85135580 | 4317,96 |
| C37 | 45,94 | 5703 | 75909244 | 5040,26 | 24,69 | 4770 | 85205103 | 4322,04 |
| C38 | 46,09 | 5693 | 75753298 | 5029,72 | 24,39 | 4759 | 85046842 | 4314,14 |
| C39 | 46,22 | 5686 | 75650819 | 5022,26 | 24,19 | 4751 | 84938690 | 4308,25 |

das mulheres, mas não em homens, o inverso acontece com a mapa a direita, referente ao *cluster* C39, que tem um valor alto de $LLR$ para os homens e um valor não tão alto de $LLR$ para as mulheres.

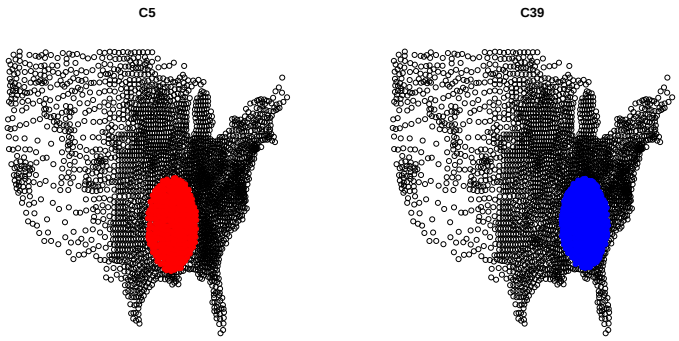



Figura 8.7: *Clusters* Detectados

## 8.3 Casos de Câncer de Laringe para Homens e Mulheres em Kentucky (EUA)

O câncer de laringe é um dos mais comuns a atingir a região da cabeça e do pescoço. Os tumores malignos podem surgir em qualquer região da laringe, mas a maioria deles se desenvolvem na glote (ROTHMAN *et al.*, 1980; De Stefani *et al.*, 1987).

Nesta seção iremos analisar *clusters* de casos de câncer laringe para homens e mulheres no ano de 2005 nos condados do estado americano de Kentucky (*Age-Adjusted Invasive Cancer Incidence Rates by in , 2005-2011*, 2014).

A Figura 8.8 mostra o conjunto Pareto-ótimo obtido pelo algoritmo multiobjetivo, constituído por 4 soluções. Um resumo de algumas destas soluções é apresentado na Tabela 8.3, em que são encontrados os valores de $LLR$ relativo aos casos de câncer para homens, o número de casos de câncer para homens, a população em risco de homens e, os valores de $LLR$ relativo aos casos de câncer de laringe para mulheres, o número de casos de câncer para mulheres e a população em risco de mulheres. .

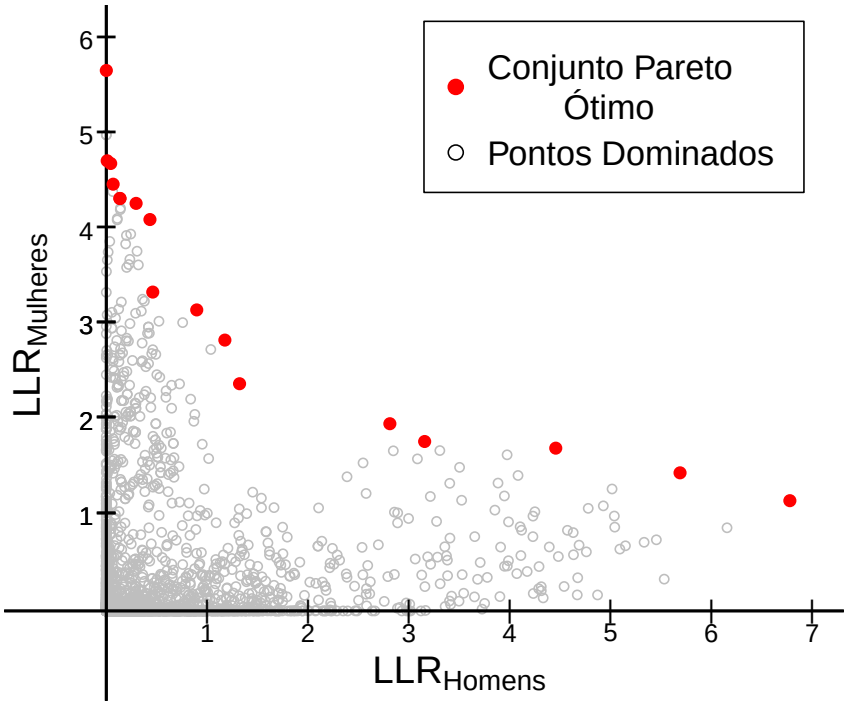



Figura 8.8: Conjunto Pareto-ótimo encontrado para os casos de câncer de Laringe para homens e mulheres para os condados do estado de Kentucky no ano de 2005.

Tabela 8.3: Resumo dos *clusters* para os casos de câncer de Laringe para homens e mulheres para os condados do estado de Kentucky em 2005.

| *Cluster* | Homens | | | Mulheres | | |
|---|---|---|---|---|---|---|
| | Casos | LLR | População | Casos | LLR | População |
| 1 | 32 | 0,00 | 217229 | 10 | 5,68 | 225175 |
| 2 | 17 | 0,01 | 97576 | 6 | 4,73 | 101473 |
| 3 | 3 | 0,04 | 10542 | 2 | 4,71 | 10691 |
| 4 | 24 | 0,07 | 133413 | 7 | 4,49 | 137078 |
| 5 | 18 | 0,14 | 153713 | 10 | 4,34 | 152856 |
| 6 | 27 | 0,14 | 143655 | 7 | 4,34 | 146656 |
| 7 | 2 | 0,30 | 6934 | 2 | 4,29 | 6944 |
| 8 | 30 | 0,43 | 242340 | 12 | 4,12 | 245094 |
| 9 | 17 | 0,46 | 150105 | 10 | 3,35 | 149109 |
| 10 | 11 | 0,90 | 119080 | 10 | 3,17 | 122771 |
| 11 | 20 | 1,18 | 171797 | 10 | 2,85 | 171593 |
| 12 | 11 | 1,32 | 112404 | 9 | 2,39 | 116114 |
| 13 | 30 | 2,81 | 155873 | 7 | 1,97 | 159180 |
| 14 | 14 | 3,16 | 137714 | 10 | 1,78 | 140422 |
| 15 | 20 | 4,46 | 196265 | 12 | 1,71 | 196768 |
| 16 | 21 | 5,69 | 199873 | 12 | 1,45 | 200515 |
| 17 | 21 | 6,78 | 199873 | 12 | 1,16 | 200515 |

Foram feitas 20.000 réplicas sob a hipótese nula para se obter as isolinhas de p-valor que estão na Figura 8.9. O ponto 17 (correspondentes ao *cluster* 17 da Tabela 8.3) na Figura 8.9 esta localizado entre as isolinhas 1 e 2 e têm p-valor entre 0,001 e 0,05, já o ponto 13 esta localizado entre as isolinhas 2 e 3 e os pontos numerados 1 e 5 estão localizados abaixo da isolinha 3 e têm p-valor inferior a 0,1.

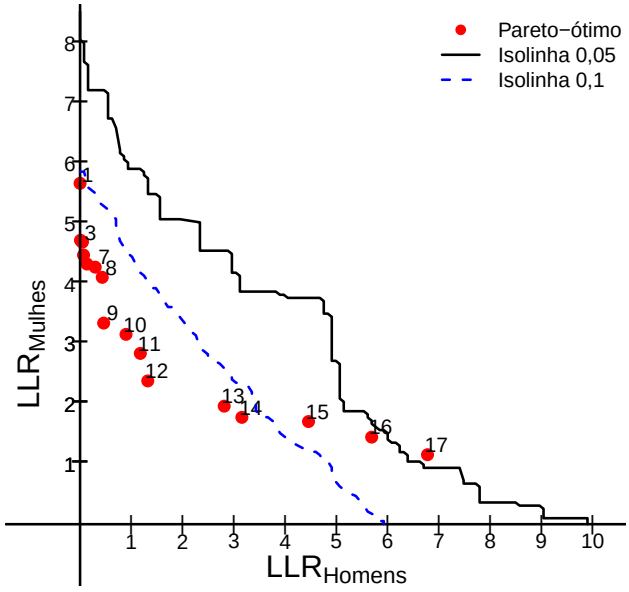
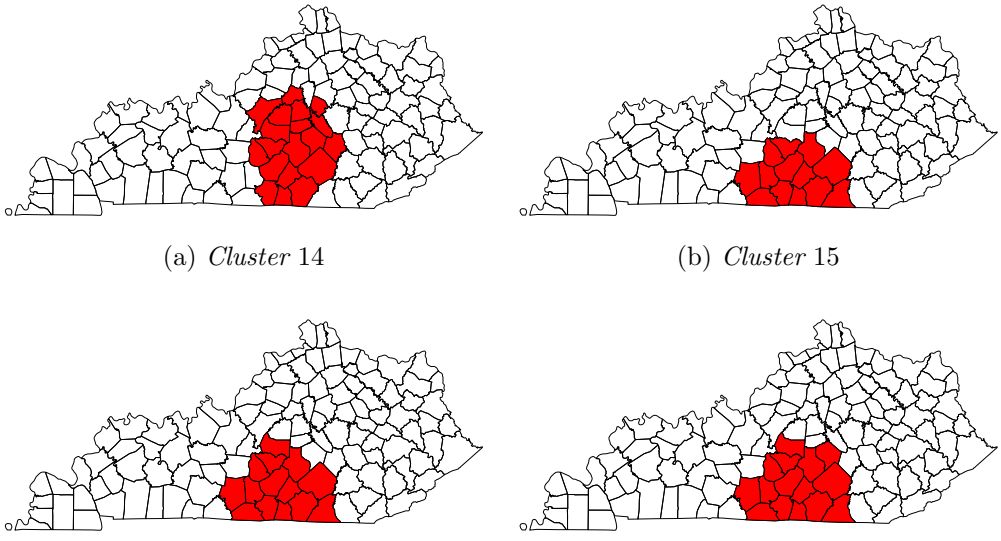



Figura 8.9: Isolinhas de p-valor.

As Figuras 8.10 (a)-(d) mostram os *clusters* detectados. Da esquerda para a direita, de cima para baixo temos os *clusters* detectados de 14 a 17.



(a) *Cluster* 14

(b) *Cluster* 15

(c) *Cluster* 16

(d) *Cluster* 17

Figura 8.10: *Clusters* Detectados

## 8.4 Casos de Câncer de Ovário e Colo do Útero em Kentucky (EUA)

Nesta seção iremos aplicar o procedimento produzido nesta Tese aos casos de câncer no ovário e aos casos de câncer no colo do útero de mulheres nos condados do estado americano de Kentucky do ano de 2009 ao ano de 2011 (*Age-Adjusted Invasive Cancer Incidence Rates by in , 2005-2011*, 2014).

Segundo Sensu (2008) o câncer do colo do útero é a principal causa de morte por câncer entre mulheres que vivem em países em vias de desenvolvimento e é causado pela infecção persistente por alguns tipos (chamados oncogênicos) do Papilomavírus Humano - HPV. A infecção genital por este vírus é muito frequente e não causa doença na maioria das vezes. Entretanto, em alguns casos, podem ocorrer alterações celulares que poderão evoluir para o câncer, Estas alterações das células são descobertas facilmente no exame preventivo (conhecido também como Papanicolau), e são curáveis na quase totalidade dos casos. Por isso é importante a realização periódica deste exame.

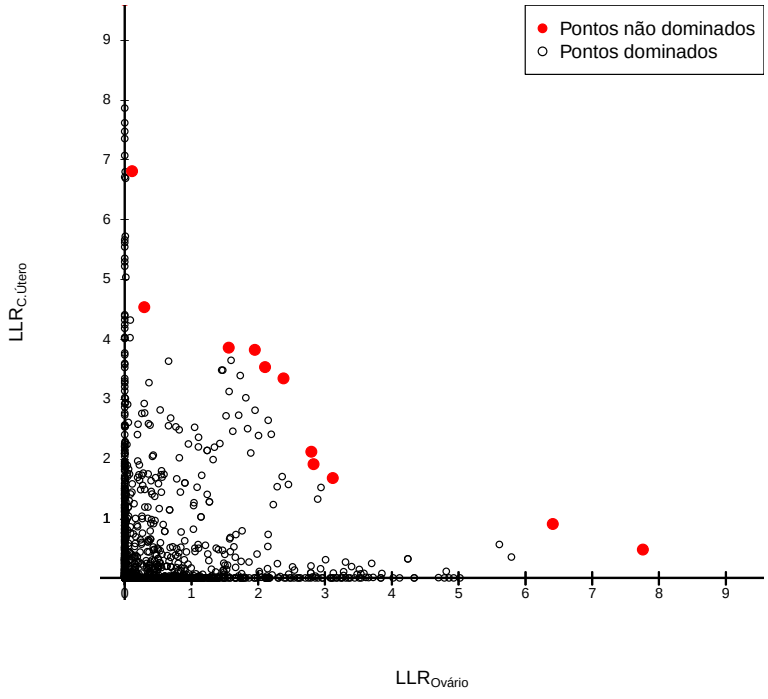



Figura 8.11: Conjunto Pareto-ótimo encontrado para os casos de câncer de ovário e os casos de câncer de colo de útero para os condados do estado de Kentucky do ano de 2009 ao ano de 2011.

A Figura 8.11 mostra o conjunto Pareto-ótimo obtido pelo algoritmo multiobjetivo, constituído por 12 soluções. Um resumo de algumas destas soluções é apresentado na Tabela 8.4, em que são encontrados os valores de $LLR$ relativo aos casos de câncer de ovário, o número de casos de câncer de ovário, os valores de $LLR$ relativo aos casos de câncer de colo de útero, o número de casos de câncer de colo de útero e a população em risco.

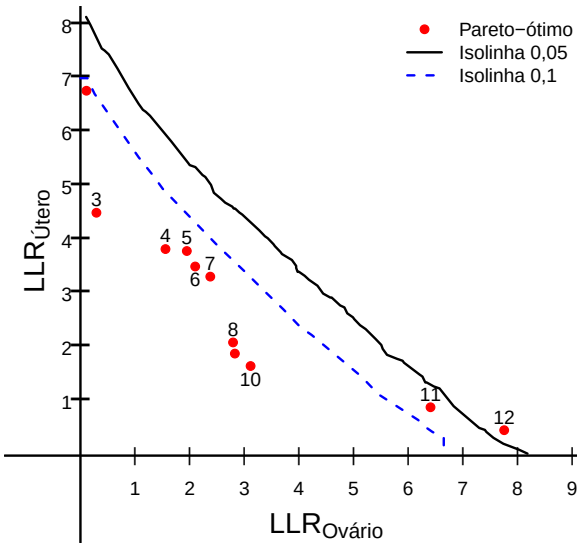


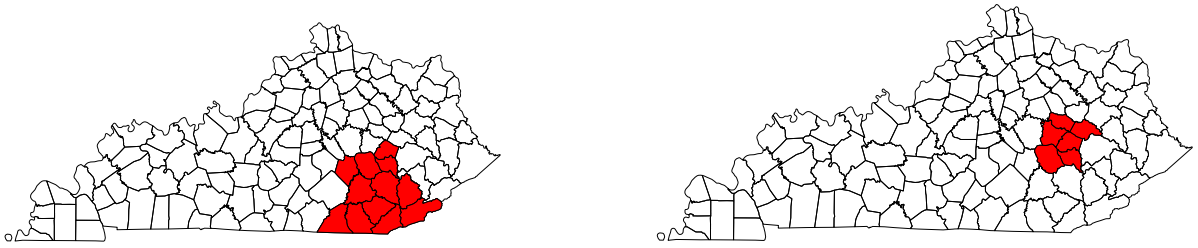

Figura 8.12: Isolinhas de p-valor.

Foram feitas 20.000 réplicas, sob a hipótese nula para se obter as isolinhas de p-valor que estão na Figura 8.12. Os pontos 1 e 12 (correspondentes aos *clusters* 1 e 12 da Tabela 8.4) na Figura 8.12 estão localizados entre a isolinha 1 e a isolinha 2 e têm p-valor entre 0,05 a 0,000999, já o ponto 11 está localizado entre a isolinha 2 e a isolinha 1 e têm p-valor entre 0,05 e 0,1 e os pontos numerados de 2 a 10 estão localizados abaixo da isolinha 3 e têm p-valor maior do que 0,1.

As Figuras 8.13 (a)-(b) mostram os *clusters* detectados.

Tabela 8.4: Resumo dos *clusters* para os casos de câncer de ovário e os casos de câncer de colo de útero para os condados do estado de Kentucky do ano de 2009 ao ano de 2011.

| *Cluster* | Câncer de ovário | | | Câncer de colo de útero | | |
|---|---|---|---|---|---|---|
| | Casos | LLR | População | Casos | LLR | População |
| 1 | 53 | 0,00 | 368308 | 47 | 9,65 | 368308 |
| 2 | 52 | 0,11 | 348371 | 44 | 6,80 | 348371 |
| 3 | 61 | 0,29 | 421219 | 47 | 4,52 | 421219 |
| 4 | 84 | 1,56 | 590122 | 62 | 3,85 | 590122 |
| 5 | 86 | 1,95 | 601326 | 64 | 3,81 | 601326 |
| 6 | 86 | 2,10 | 601326 | 64 | 3,52 | 601326 |
| 7 | 56 | 2,38 | 447267 | 47 | 3,33 | 447267 |
| 8 | 58 | 2,79 | 458471 | 49 | 2,11 | 458471 |
| 9 | 64 | 2,83 | 514531 | 51 | 1,90 | 514531 |
| 10 | 12 | 3,11 | 20874 | 9 | 1,67 | 20874 |
| 11 | 112 | 6,41 | 735708 | 72 | 0,90 | 735708 |
| 12 | 31 | 7,76 | 32078 | 11 | 0,47 | 32078 |



(a) *Cluster* 1 (b) *Cluster* 12

Figura 8.13: *Clusters* Detectados

## 8.5 Casos de Câncer de estômago para homens e mulheres em Kentucky (EUA)

O câncer de estômago é um dos tumores malignos mais frequente do mundo. O estômago é um órgão que faz parte do sistema digestivo, cuja responsabilidade é processar os alimentos ingeridos, extraindo deles nutrientes (vitaminas, minerais, carboidratos, gorduras, proteínas e água).

O câncer de estômago é o crescimento de células anormais no órgão desse sistema digestivo e pode ocorrer em qualquer local de sua extensão. Grande parte desse tipo de tumor ocorre na camada mucosa (a camada de revestimento interna), surgindo na forma de irregulares pequenas lesões com ulcerações (rompimento do tecido mucoso) - características de cânceres ou tumores malignos (Hirayama, 1971).

Nesta seção iremos analisar *clusters* de casos de câncer de estômago para homens e mulheres entre os anos de 2009 e 2011 nos condados do estado americano de Kentucky (*Age-Adjusted Invasive Cancer Incidence Rates by in , 2005-2011*, 2014).

A Figura 8.14 mostra o conjunto Pareto-ótimo obtido pelo algoritmo multiobjetivo, constituído por 4 soluções. Um resumo de algumas destas soluções é apresentado na Tabela 8.5, em que são encontrados os valores de $LLR$ relativo aos casos de câncer de estômago para homens, o número de casos de câncer de estômago para homens, a população em risco de homens e, os valores de $LLR$ relativo aos casos de câncer de estômago para mulheres, o número de casos de câncer de estômago para mulheres e a população em risco de mulheres.

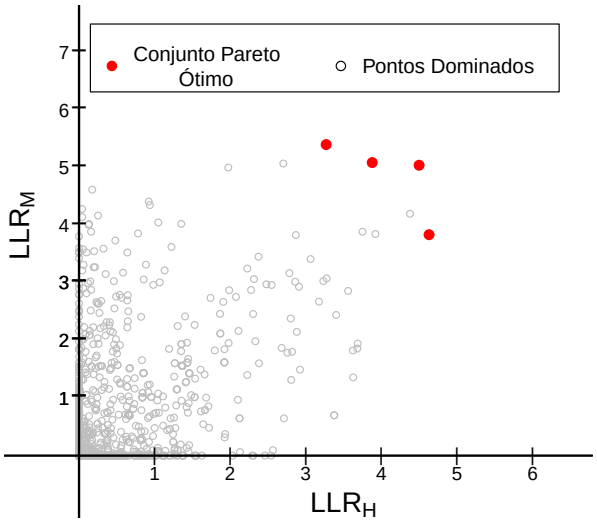



Figura 8.14: Conjunto Pareto-ótimo encontrado para os casos de câncer de estômago para homens e para mulheres nos condados do estado americano de Kentucky entre os anos 2009 e 2011.

Tabela 8.5: Resumo dos *clusters* para os casos de câncer de estômago para homens e para mulheres nos condados do estado americano de Kentucky entre os anos de 2009 e 2011.

| *Cluster* | Homens | | | Mulheres | | |
|---|---|---|---|---|---|---|
| | Casos | LLR | População | Casos | LLR | População |
| 1 | 54 | 3,27 | 467877 | 41 | 5,41 | 478006 |
| 2 | 52 | 3,88 | 431794 | 38 | 5,09 | 439959 |
| 3 | 55 | 4,50 | 448063 | 39 | 5,05 | 456572 |
| 4 | 58 | 4,63 | 475569 | 39 | 3,84 | 491564 |

Foram feitas 20.000 réplicas sob a hipótese nula para se obter as isolinhas de p-valor que estão na Figura 8.15. Os pontos 2 e 3 (correspondentes aos *clusters* 2 e 3 da Tabela 8.5) na Figura 8.15 estão localizados acima da isolinha 1 e têm p-valor igual a 0,000999, já os pontos 1 e 4 estão localizados entre a isolinha 2 e a isolinha 1 e têm p-valor entre 0,05 e 0,000999.

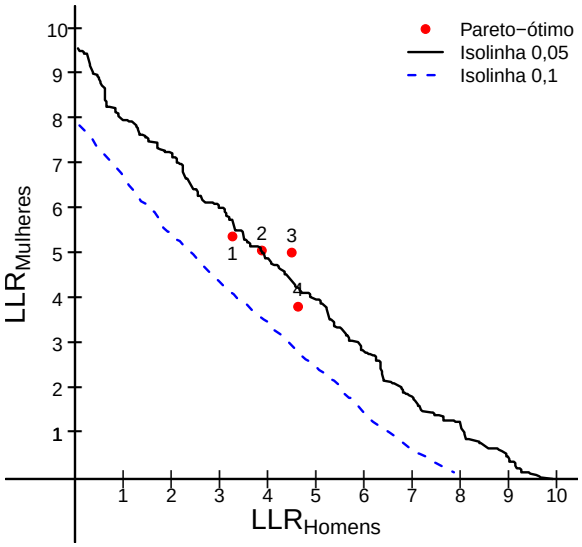
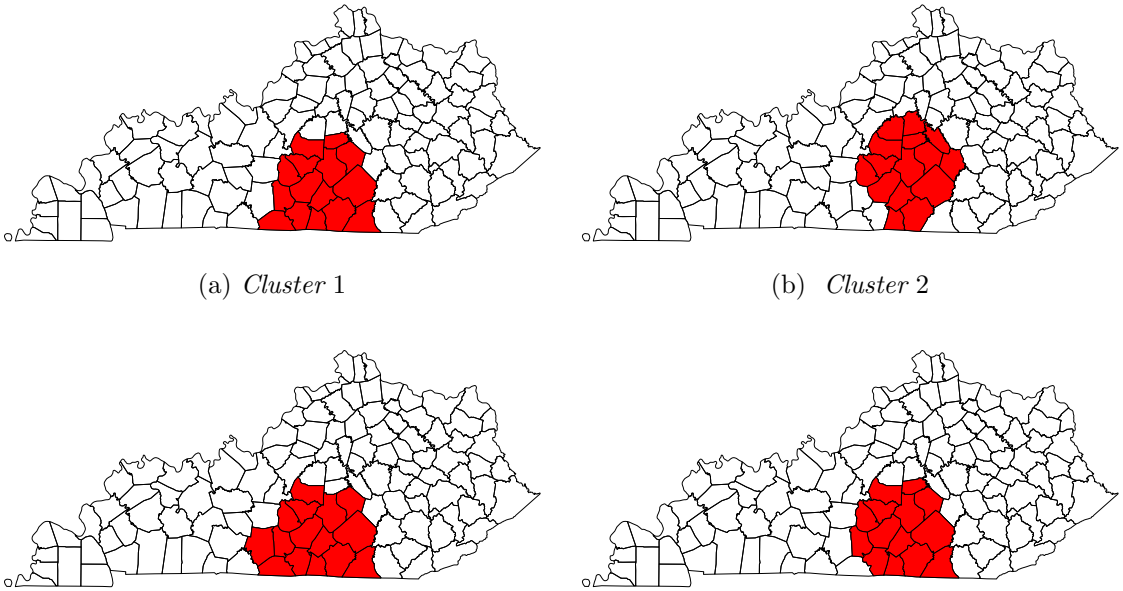



Figura 8.15: Isolinhas de p-valor.

A Figuras 8.16 (a)-(d) mostram os *clusters* detectados.



(a) *Cluster* 1

(b) *Cluster* 2

(c) *Cluster* 3

(d) *Cluster* 4

Figura 8.16: *Clusters* Detectados

# Capítulo 9

# Considerações Finais

Nesta tese desenvolveu-se um método para o problema de detecção de *clusters* espaciais com o uso de duas fontes de dados, através da implementação de um procedimento multiobjetivo para a detecção e inferência de *clusters*. Como discutido no capítulo 2, a estatística *scan* espacial (Kulldorff, 1997) é um método estatístico poderoso para detecção de *clusters*. Para cada conjunto de dados avaliou-se as zonas construídas no mapa utilizando a estatística espacial *scan*. Desenvolvemos um algoritmo multiobjetivo que encontra o conjunto de soluções eficientes através da maximização de vários objetivos, que são a estatística *scan* calculada para cada um dos bancos de dados. Usou-se o conceito de soluções pareto ótimas para detectar os *clusters* com potencial. E, valendo-se de simulações Monte-Carlo (Dwass *et al.*, 1957) sob a hipótese nula, encontrou-se as isolinhas de p-valor (Cançado *et al.*, 2010; Knowles, 2005) que possibilitam a avaliação dos *clusters* candidatos.

As tentativas anteriores de lidar com o problema levavam em conta apenas procedimentos *ad-hoc* que eram baseados na estatística *scan* e se valiam de uma otimização mono-objetivo. O uso do conceito de conjunto de Pareto nesse problema, seguido da escolha da solução mais significativa, permite que a escolha da melhor solução seja rigorosa, com a representação das funções objetivo, para cada conjunto de dados, e não sofre de um possível confundimento, ou pequeno poder de detecção, tal como acontece nesses procedimentos *ad-hoc*.

Estendemos o conceito usual de significância de forma natural para o problema multiobjetivo através do conceito de isolinhas de p-valor. Essas isolinhas são calculadas através da função de aproveitamento, que se mostrou a ferramenta mais adequada para tratar esse problema.

A análise feita neste trabalho permite comparar a posição relativa entre o conjunto de soluções não-dominadas e as isolinhas. Assim, é possível analisar e selecionar as melhores soluções de *cluster*, dadas pelo conjunto não-dominado. Este algoritmo, de acordo com os estudos, é uma ferramenta apropriada e altamente eficiente, obtendo excelentes resultados. Testes numéricos mostram que o poder de detecção, a sensibilidade e o valor positivo de predição do algoritmo multiobjetivo foi melhor que os procedimentos usados na literatura.

No capítulo 7 o método proposto nesta tese foi aplicado a exemplos, com o objetivo de encontrar *clusters* espaciais de casos. Os resultados demonstram que os nossos métodos podem detectar rapidamente os *clusters* relevantes do ponto de vista real. Também temos o interesse de estender esse trabalho em uma série de outras técnicas, e estas extensões

são discutidos na seção seguinte.

## 9.1 Os trabalhos futuros

Vamos agora considerar brevemente três direções importantes para trabalhos futuros: a extensão do método proposto para mais de duas fontes de dados, detecção de *clusters* no espaço e no tempo e a detecção de *clusters* irregulares. Cada uma destas extensões tem o potencial de melhorar drasticamente a generalidade, como discutiremos nas subseções a seguir.

### 9.1.1 Extensão para múltiplas fontes de dados

Sistemas de vigilância em saúde pública coletam e analisam automaticamente vários tipos de dados na busca de possíveis sinais de um surto de uma doença. Por exemplo, um sistema pode obter, a partir de um subconjunto de departamentos de emergência em uma região, a contagem diária do número de visitas de emergência para cada uma de várias categorias de doenças. O sistema de vigilância pode, também, obter dados de laboratórios, vendas de remédios em farmácias, e outras fontes fornecedoras de dados.

Como discutido nesta tese, a contribuição mais importante do nosso método é a detecção de *clusters*, com informações de duas fontes de dados. É natural pensar em uma extensão do método de detecção de *clusters* multiobjetivo que nos permite considerar muitas outras fontes de dados.

Acreditamos que a extensão para múltiplos fontes de dados fará este método valioso para uma ampla variedade de aplicações e acreditamos que o nosso método permitirá realizar a detecção de *cluster* para enormes conjuntos de dados sendo capaz de detectar *clusters* em tempo real.

### 9.1.2 Detecção de *clusters* irregulares

A busca de *clusters* de forma irregular nos daria maior poder de detecção de *clusters* com as áreas que não podem ser aproximadas por áreas circulares. O delineamento geográfico de *clusters* irregulares apresenta algumas dificuldades. A liberdade geométrica ilimitada para a forma do *cluster* diminui o poder de detecção (Duczmal e Buckeridge, 2006). Isto acontece porque o conjunto de todas as soluções conexas, independente da forma, é muito grande. O máximo da função objetivo tende a estar associado a um *cluster* com forma de árvore, que simplesmente liga as regiões do mapa com maior verossimilhança, sem contribuir para a descoberta de soluções que fazem o delineamento correto do *cluster* verdadeiro. Este é um problema que ocorre em todos os métodos de detecção de *clusters* irregulares e pode ser contornado, em parte, limitando o número máximo de regiões que podem constituir cada solução.

Duczmal e Buckeridge (2006); Duczmal *et al.* (2007) aplicaram uma penalização usando o conceito de compacidade penalizando a estatística *scan* de acordo com a irregularidade da forma da solução.

Cançado *et al.* (2010) desenvolveram um algoritmo genético que encontra o conjunto de soluções de *clusters* eficientes através da maximização de dois objetivos, a estatística

*scan* e a regularidade da forma, ou compacidade. Este algoritmo multiobjetivo disponibiliza um conjunto de soluções que são ordenadas pelo critério de significância estatística. Dado um conjunto de soluções ótimas obtidas pelo algoritmo de busca de *clusters*, o problema foi reduzido à escolha da solução mais significante entre elas.

### 9.1.3 Detecção de *clusters* no espaço e no tempo

As técnicas de detecção de *clusters* espaciais necessitam da fixação de um período de tempo para a agregação dos casos que ocorreram dentro deste período. Este período pode ser de dias até anos, e a escolha do período utilizado pode ser questionável. A especificação desse valor pode gerar dois tipos de problemas. Incluindo-se poucos períodos, o teste pode não ter poder suficiente para detectar uma doença de risco baixo a moderado que ocorre há um tempo considerável. Caso se inclua muitos períodos, o teste pode não ter poder suficiente para detectar um risco elevado que ocorreu em um período curto.

A extensão da estatística *scan* de Kulldorff (1997) do espaço para o espaço-tempo ocorreu através da ampliação da estatística de varredura com formato circular para um formato cilíndrico. A base circular corresponde à dimensão geográfica e a altura, ao intervalo de tempo.

Sob $H_0$ assume-se que o número de casos, seja distribuído segundo uma *Poisson* com risco constante no espaço e no tempo e sob $H_a$ assume-se que o risco seja distinto dentro e fora de pelo menos um cilindro. Os cálculos realizados em Kulldorff (1997) são replicados em Kulldorff *et al.* (1998), porém, onde havia uma janela circular, agora obtém-se uma janela de formato cilíndrico, que irá varrer a região de estudo no espaço e no tempo.

# Referências Bibliográficas


Abrams, A. M., Kleinman, K. e Kulldorff, M. (2010). Gumbel based p-value approximations for spatial scan statistics, *International journal of health geographics* **9**(1): 61. *Age-Adjusted Invasive Cancer Incidence Rates by in , 2005-2011*

Age-Adjusted Invasive Cancer Incidence Rates by in , 2005-2011 *(2014).* `http://cancer-rates.info/ky`. *Accessed: 2014/09/09.*

*Almeida, A. C., Duarte, A. R., Duczmal, L. H., Oliveira, F. L. e Takahashi, R. H. (2011). Data-driven inference for the spatial scan statistic,* International journal of health geographics **10***(1): 47.*

*Balakrishnan, N. e Koutras, M. V. (2011).* Runs and scans with applications*, Vol. 764, John Wiley & Sons.*

*Boicey, C. (2013). Innovations in social media: The mappyhealth experience,* Nursing management **44***(3): 10–11.*

*Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R. e Moore, A. W. (2005). Algorithms for rapid outbreak detection: a research synthesis,* Journal of biomedical informatics **38***(2): 99–113.*

*Burkom, H. S. (2003). Biosurveillance applying scan statistics with multiple, disparate data sources,* Journal of Urban Health **80***(1): i57–i65.*

*Cançado, A. L., Duarte, A. R., Duczmal, L., Ferreira, S. J., Fonseca, C. M. e Gontijo, E. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters,* International journal of health geographics **9***(1): 55.*

*Chakma, J., Calcagno, J. L., Behbahani, A. e Mojtahedian, S. (2009). The power of social networking in medicine,* NATURE BIoTECHNoloGy **27***(10): 889.*

*Chary, M., Genes, N., McKenzie, A. e Manini, A. F. (2013). Leveraging social networks for toxicovigilance,* Journal of Medical Toxicology **9***(2): 184–191.*

*Chawla, N. V. e Davis, D. A. (2013). Bringing big data to personalized healthcare: a patient-centered framework,* Journal of general internal medicine **28***(3): 660–665.*

*Chunara, R., Bouton, L., Ayers, J. W. e Brownstein, J. S. (2013). Assessing the online social environment for surveillance of obesity prevalence,* PloS one **8***(4): e61373.*

*da Fonseca, V. G., Fonseca, C. M. e Hall, A. O. (2001). Inferential performance assessment of stochastic optimisers and the attainment function,* Evolutionary Multi-Criterion Optimization*, Springer, pp. 213–225.*

*Dailey, L., Watkins, R. E. e Plant, A. J. (2007). Timeliness of data sources used for influenza surveillance,* Journal of the American Medical Informatics Association **14***(5): 626–631.*

*De Stefani, E., Correa, P., Oreggia, F., Leiva, J., Rivero, S., Fernandez, G., Deneo-Pellegrini, H., Zavala, D. e Fontham, E. (1987). Risk factors for laryngeal cancer,* Cancer **60***(12): 3087–3091.*

*Duczmal, L., Cançado, A. L. F. e Takahashi, R. H. C. (2008). Delineation of irregularly shaped disease clusters through multiobjective optimization,* Journal of Computational and Graphical Statistics **17***(1): 243–262.*

*Duczmal, L., Cancado, A. L., Takahashi, R. H. e Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics,* Computational Statistics & Data Analysis **52***(1): 43–52.*

*Duczmal, L., Duarte, A. R. e Tavares, R. (2009). Extensions of the scan statistic for the detection and inference of spatialclusters,* Scan Statistics*, Springer, pp. 153–177.*

*Duczmal, L. e Buckeridge, D. L. (2006). A workflow spatial scan statistic,* Statistics in Medicine **25***(5): 743–754.*

*Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T. e Rothman, R. E. (2013). Influenza forecasting with google flu trends,* PloS one **8***(2): e56176.*

*Dwass, M. et al. (1957). Modified randomization tests for nonparametric hypotheses,* The Annals of Mathematical Statistics **28***(1): 181–187.*

*Edgeworth, F. Y. (1881).* Mathematical psychics: An essay on the application of mathematics to the moral sciences*, C. Keagann Paul.*

*Elliott, P., Martuzzi, M. e Shaddick, G. (1995). Spatial statistical methods in environmental epidemiology: a critique,* Statistical Methods in Medical Research **4***(2): 137–159.*

*Fonseca, C. M., Da Fonseca, V. G. e Paquete, L. (2005). Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function,* Evolutionary Multi-Criterion Optimization*, Springer, pp. 250–264.*

*Gittelman, S. H., Trimarchi, E. R. e Lange, V. W. (n.d.). A new source of data for public health surveillance: Facebook likes.*

*Hay, S. I., George, D. B., Moyes, C. L. e Brownstein, J. S. (2013). Big data opportunities for global infectious disease surveillance,* PLoS medicine **10***(4): e1001413.*

*Hingle, M., Yoon, D., Fowler, J., Kobourov, S., Schneider, M. L., Falk, D. e Burd, R. (2013). Collection and visualization of dietary behavior and reasons for eating using twitter,* Journal of medical Internet research.

*Hirayama, T. (1971). Epidemiology of stomach cancer,* Gann Monogr **11***: 3–19.*

*Jonsson, M. E., Heier, B. T., Norström, M. e Hofshagen, M. (2010). Analysis of simultaneous space-time clusters of campylobacter spp. in humans and in broiler flocks using a multiple dataset approach,* International journal of health geographics **9***(1): 48.*

*Josseran, L., Fouillet, A., Caillère, N., Brun-Ney, D., Ilef, D., Brucker, G., Medeiros, H. e Astagneau, P. (2010). Assessment of a syndromic surveillance system based on morbidity data: results from the oscour® network during a heat wave,* PLoS One **5***(8): e11984.*

*Kara, E., Elliot, A., Bagnall, H., Foord, D., Pnaiser, R., Osman, H., Smith, G. e Olowokure, B. (2012). Absenteeism in schools during the 2009 influenza a (h1n1) pandemic: a useful tool for early detection of influenza activity in the community?,* Epidemiology and infection **140***(7): 1328–1336.*

*Khoury, M. J., Lam, T. K., Ioannidis, J. P., Hartge, P., Spitz, M. R., Buring, J. E., Chanock, S. J., Croyle, R. T., Goddard, K. A., Ginsburg, G. S. et al. (2013). Transforming epidemiology for 21st century medicine and public health,* Cancer Epidemiology Biomarkers & Prevention **22***(4): 508–516.*

*Knowles, J. (2005). A summary-attainment-surface plotting method for visualizing the performance of stochastic multiobjective optimizers,* Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*, IEEE, pp. 552–557.*

*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,* Scan statistics and applications*, Springer, pp. 303–322.*

*Kulldorff, M. (2011). Satscan user guide for version 9.0.*

*Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A. e Key, C. R. (1998). Evaluating cluster alarms: a space-time scan statistic and brain cancer in los alamos, new mexico.,* American journal of public health **88***(9): 1377–1380.*

*Kulldorff, M. e Nagarwalla, N. (1995). Spatial disease clusters: detection and inference,* Statistics in medicine **14***(8): 799–810.*

*Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R. e Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection,* PLoS medicine **2***(3): e59.*

*Kulldorff, M., Huang, L., Pickle, L. e Duczmal, L. (2006). An elliptic spatial scan statistic,* Statistics in medicine **25***(22): 3929–3943.*

*Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. e Platt, R. (2007). Multivariate scan statistics for disease surveillance,* Statistics in Medicine **26***(8): 1824–1833.*

*Kulldorff, M., Tango, T. e Park, P. J. (2003). Power comparisons for disease clustering tests,* Computational Statistics & Data Analysis **42***(4): 665–684.*

*Lawson, A. B. (2013).* Statistical methods in spatial epidemiology*, John Wiley & Sons.*

*Lawson, A. B. e Kulldorff, M. (1999). A review of cluster detection methods,* Disease Mapping and Risk Assessment for Public Health *pp. 99–110.*

*Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., Bertollini, R. et al. (1999).* Disease mapping and risk assessment for public health.*, John Wiley & Sons.*

*Lee, K., Agrawal, A. e Choudhary, A. (2013). Real-time disease surveillance using twitter data: demonstration on flu and cancer,* Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1474–1477.*

*Li, J. e Cardie, C. (2013). Early stage influenza detection from twitter,* arXiv preprint arXiv:1309.7340.

*Liu, T. Y., Sanders, J. L., Tsui, F.-C., Espino, J. U., Dato, V. M. e Suyama, J. (2013). Association of over-the-counter pharmaceutical sales with influenza-like-illnesses to patient volume in an urgent care setting,* PloS one **8***(3): e59273.*

*Minniear, T. D., McIntosh, E. B., Alexander, N., Weidle, P. J. e Fulton, J. (2013). Using electronic surveys to gather information on physician practices during a response to a local epidemic-rhode island, 2011,* Annals of epidemiology **23***(8): 521–523.*

*Morse, S. S. (2012). Public health surveillance and infectious disease detection,* Biosecurity and bioterrorism: biodefense strategy, practice, and science **10***(1): 6–16.*

*Mostashari, F., Fine, A., Das, D., Adams, J. e Layton, M. (2003). Use of ambulance dispatch data as an early warning system for communitywide influenzalike illness, new york city,* Journal of Urban Health **80***(1): i43–i49.*

*Neill, D. B. (2009). An empirical comparison of spatial scan statistics for outbreak detection,* International Journal of Health Geographics **8***(1): 20.*

*Ortiz, J. R., Sotomayor, V., Uez, O. C., Oliva, O., Bettels, D., McCarron, M., Bresee, J. S., Mounts, A. W. et al. (2009). Strategy to enhance influenza surveillance worldwide,* Emerg Infect Dis **15***(8): 1271–1278.*

*Pareto, V. (1964).* Cours d'economie politique*, Librairie Droz.*

*Perez, A. M., Ward, M. P., Torres, P. e Ritacco, V. (2002). Use of spatial statistics and monitoring data to identify clustering of bovine tuberculosis in argentina,* Preventive veterinary medicine **56***(1): 63–74.*

*Pivette, M., Mueller, J. E., Crépey, P. e Bar-Hen, A. (2014). Drug sales data analysis for outbreak detection of infectious diseases: a systematic literature review,* BMC infectious diseases **14***(1): 604.*

*Rodman, J. S., Frost, F. e Jakubowski, W. (1998). Using nurse hot line calls for disease surveillance.,* Emerging infectious diseases **4***(2): 329.*

*Rolland, E., Moore, K. M., Robinson, V. A. e McGuinness, D. (2006). Using ontario's "telehealth"health telephone helpline as an early-warning system: a study protocol.,* BMC health services research **6***(1): 10.*

*ROTHMAN, K. J., CANN, C. I., FLANDERS, D. e FRIED, M. P. (1980). Epidemiology of laryngeal cancer,* Epidemiologic reviews **2***(1): 195–209.*

*Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L. et al. (2012). Digital epidemiology,* PLoS computational biology **8***(7): e1002616.*

*Sankoh, O. A., Yé, Y., Sauerborn, R., Müller, O. e Becher, H. (2001). Clustering of childhood mortality in rural burkina faso,* International Journal of Epidemiology **30***(3): 485–492.*

*Sensu, S. (2008). Mortalidade por câncer do colo do útero no brasil,* Rev Bras Ginecol Obstet **30***(5): 216–8.*

*Sočan, M., Erčulj, V. e Lajovic, J. (2012). Early detection of influenza-like illness through medication sales,* Cent Eur J Public Health **20***(2): 156–162.*

*Song, C. e Kulldorff, M. (2003). Power evaluation of disease clustering tests,* International Journal of Health Geographics **2***(1): 9.*

*Sugawara, T., Ohkusa, Y., Ibuka, Y., Kawanohara, H., Taniguchi, K. e Okabe, N. (2012). Real-time prescription surveillance and its application to monitoring seasonal influenza activity in japan,* Journal of medical Internet research **14***(1): e14.*

*Tian, L., Tan, L., Fan, Y., Wang, Y., Zhang, J., Cheng, L., Wei, S., Liu, L., Yan, W., Xu, B. et al. (2013). The application of integrated surveillance system for symptoms in surveillance of influenza among children.,* Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine] **47***(12): 1095–1099.*

*Todd, S., Diggle, P. J., White, P. J., Fearne, A. e Read, J. M. (2014). The spatiotemporal association of non-prescription retail sales with cases during the 2009 influenza pandemic in great britain,* BMJ open **4***(4): e004869.*

*Van Boeckel, T. P., Gandra, S., Ashok, A., Caudron, Q., Grenfell, B. T., Levin, S. A. e Laxminarayan, R. (2014). Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data,* The Lancet infectious diseases **14***(8): 742–750.*

*Viel, J.-F., Arveux, P., Baverel, J. e Cahn, J.-Y. (2000). Soft-tissue sarcoma and non-hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels,* American journal of epidemiology **152***(1): 13–19.*

*Wagner, M. M., Tsui, F.-C., Espino, J., Hogan, W., Hutman, J., Hersh, J., Neill, D., Moore, A., Parks, G., Lewis, C. et al. (2004). National retail data monitor for public health surveillance,* Morbidity and Mortality Weekly Report *pp. 40–42.*

*Weitzman, E. R., Kelemen, S. e Mandl, K. D. (2011). Surveillance of an online social network to assess population-level diabetes health status and healthcare quality,* Online journal of public health informatics.

*Wu, W. e Gruenwald, L. (2010). Research issues in mining multiple data streams,* Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, ACM, pp. 56–60.*

*Yoon, S., Elhadad, N. e Bakken, S. (2013). A practical approach for content mining of tweets,* American journal of preventive medicine **45***(1): 122–129.*