LUCIANO RIOS SCHERRER

DETECÇÃO DE CONGLOMERADOS ESPAÇO-TEMPORAIS COM GEOMETRIA CILÍNDRICA E NÃO-CILÍNDRICA

Belo Horizonte
24 de outubro de 2007

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

DETECÇÃO DE CONGLOMERADOS ESPAÇO-TEMPORAIS COM GEOMETRIA CILÍNDRICA E NÃO-CILÍNDRICA

Proposta de dissertação apresentada ao Curso de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Estatística.

LUCIANO RIOS SCHERRER

Belo Horizonte
24 de outubro de 2007

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Detecção de Conglomerados Espaço-Temporais com Geometria Cilíndrica e Não-Cilíndrica

LUCIANO RIOS SCHERRER

Proposta de dissertação defendida perante banca examinadora constituída por:

Doutor. Marcelo Azevedo Costa – Orientador
Universidade Federal de Minas Gerais

Ph. D. Renato M. Assunção – Co-orientador
Universidade Federal de Minas Gerais

Doutor. Michel Ferreira da Silva
Universidade Federal de Minas Gerais

Doutor. Marcos Antônio da Cunha Santos
Universidade Federal de Minas Gerais

Doutor. Vitor Ozaki
ESALQ/USP

Belo Horizonte, 24 de outubro de 2007

Resumo

A detecção de conglomerados no espaço-tempo tem como objetivo a delimitação de uma região geográfica em determinado tempo na qual a hipótese de ocorrência aleatória de um evento pontual é rejeitada. Tal informação é de extrema relevância em estudos epidemiológicos. Este estudo apresenta dois métodos de detecção de conglomerados espaço-temporais nos quais a estrutura de vizinhança espaço-temporal é definida a partir de um grafo interconectado e agregada ao processo de crescimento e busca de conglomerados. Esse procedimento possibilita a detecção de conglomerados espaço-temporais de geometria arbitrária (não-cilíndrica). Os métodos tradicionais de varredura espaço-temporal se restringem à geometria de busca a conglomerados com geometria cilíndrica, resultando em uma detecção parcial ou superestimação do conglomerado e apresenta alto poder nesse contexto. Uma avaliação do poder de detecção dos métodos abordados neste estudo para conglomerados espaço-temporais com geometria cilíndrica e arbitrária é realizada via dados simulados e reais. Nos dados reais, utilizou-se a ocorrência de casos de câncer cerebral no Novo México e registrado pelo Departamento de Saúde do Novo México no ano de 1973 a 1991. Também utilizou-se a ocorrência de casos reais de dengue em Vitória, Espírito Santo, registrados pela Vigilância Epidemiológica de Vitória nos meses de janeiro a dezembro de 2006. Restrições durante o processo de crescimento dos conglomerados são sugeridas para evitar tamanho excessivo e geometria muito irregular.

Abstract

Space-Time cluster detection aims at detection a particular region in certain time intervals and time period analyses in which the hypothesis of random occurrence of event is rejected. This information is extreme relevance in epidemiology studies. Many different methods have been proposed to test for spatial-time clustering of any type of observations. This report presents two space-time cluster detection methods that aggregate the spatial-time neighbor structure into the cluster growing process. The procedure allows the detection of arbitrarily shaped space-time clusters. Standard spatial-time scan statistics confine the cluster geometry shape to cylinder shaped clusters, resulting in partial or over sized clusters and show high power in this context. Restrictions during the growth process are suggested in order to prevent over sized clusters with odd geometries. We use simulation data set and real data set to compare the power of the methods. Results of space-time cluster detection in brain cancer data are presented for New Mexico and data for case of dengue are presented in Vitória city.

Em 1998, tive a oportunidade de iniciar um relacionamento real e pessoal com Deus, que me concedeu pela graça a convicção da minha salvação, mediante a fé em Jesus Cristo. Continuo, desde então, a aprofundá-lo, cada vez mais, através da leitura da Bíblia. Posso citar dois versículos, dentre vários, que impactaram minha vida.

" Porque sou eu que conheço os planos que tenho para vocês ", diz o SENHOR, " planos de fazê-los prosperar e não de lhes causar dano, planos de dar-lhes esperança e um futuro. Então vocês clamarão a mim, virão orar a mim, e eu os ouvirei. Vocês me procurarão e me acharão quando me procurarem de todo coração ". (Jeremias 29:11-13)

Respondeu Jesus: " Eu sou o caminho, a verdade e a vida. Ninguém vem ao Pai, a não ser por mim ". (João 14:6)

Enfim, a capacidade de crer nessas verdades depende da boa vontade do Deus vivo em se revelar a nós.

Agradecimentos

- A Jesus pela força, determinação e sabedoria que me concedeu.
- Aos meus pais e irmãos que sempre me apoiaram e me ensinaram a valorizar os estudos desde criança.
- A minha querida namorada Márcia que me aguentou nos momentos de estresse.
- Aos meus orientadores, Marcelo e Renato, que me incentivaram a conduzir, de maneira admirável, este estudo.
- Aos meus colegas e amigos de mestrado, Marcos Prates, Elias T. Krainski, Danilo Lopes e Alexandre Elias pelo companheirismo, pelas brincadeiras e pela contribuição significativa de cada um no desenvolvimento deste estudo.
- Ao laboratório LESTE pela ótima estrutura de trabalho que me disponibilizou e pelo apoio financeiro.
- Ao Departamento de Estatística da UFMG pela oportunidade e aos professores que contribuiram na minha formação acadêmica.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Aspectos Gerais

Pode-se definir a Estatística como um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar *dados* oriundos de estudos ou experimentos realizados em qualquer área do conhecimento. Denominam-se dados um (ou mais) conjunto de valores, numéricos ou não, obtidos pela variável (ou variáveis) de interesse do pesquisador[32].

Nesse contexto, a Estatística Espacial procura mensurar propriedades e relacionamentos, levando-se em conta a localização da variável em estudo de forma explícita. Ou seja, a ênfase é incorporar a componente espacial à análise que se deseja fazer.

A razão de utilizar-se de técnicas espaciais nesta dissertação (ao invés de técnicas que não utilizam informação geográfica) é que, quando dados espaciais são envolvidos, existe a possibilidade de uma dependência entre as observações da variável do estudo, podendo produzir resultados diferentes e, na maioria das vezes, mais significativos e realísticos do que aqueles obtidos usando a análise tradicional, já que essa assume independência entre as observações. Um exemplo, para entender melhor a importância da informação geográfica associada ao conjunto de valores das variáveis de interesse, será mostrado a seguir.

Escolhemos a cidade de Vitória, Espírito Santo, para ser a região geográfica na qual deseja-se avaliar um modelo estatístico espacial para a detecção de conglomerados. O objetivo do exemplo será modelar a variação de casos de dengue nos meses de janeiro a dezembro de 2006. A Figura 1.1 mostra o mapa de Vitória dividido em 79 bairros. Segundo o Censo demográfico de 2000 (IBGE), a população urbana de Vitória, estimada em 282.611 habitantes, foi estratificada por bairro. No período estudado, foi registrado pela Vigilância Epidemiológica de Vitória um total de 2.172 casos de dengue, porém houve uma perda de 5% dos dados. A quantidade total de casos de dengue para cada mês de 2006 foi: $C_1 = 163$, $C_2 = 238$, $C_3 = 414$, $C_4 = 589$, $C_5 = 538$, $C_6 = 88$, $C_7 = 15$, $C_8 = 1$, $C_9 = 7$, $C_{10} = 2$, $C_{11} = 2$ e $C_{12} = 2$, onde, para $i = 1$(jan),...,12(dez), C_i denota o total de casos. Para cada ocorrência de caso de dengue (*variável resposta*), tem-se associado o mês de ocorrência da doença, a latitude e a longitude do centro do bairro onde a pessoa reside e a informação que o município fornece ou não a água tratada nesse bairro. Tal informação é relevante visto que, em lugares sem abastecimento de

água, as pessoas são incentivadas a armazenar água em baldes ou tambores, gerando, com isso, um potencial criador de *Aedes aegypti*. Essas são possíveis *covariáveis* que talvez explique a variação de casos de dengue.

É razoável assumir que a variável resposta, casos de dengue, tem *distribuição Poisson*, pois é uma contagem de números inteiros. Com isso, com a ajuda do *software* R, ajustamos um *Modelo Linear Generalizado* para os dados. O modelo proposto foi: $\log\mu = \eta = \beta_0 + \beta_1(\text{Localização}) + \beta_2(\text{Tempo}) + \beta_3(\text{Água})$, em que β_0 denota a presença do intercepto no modelo, Localização é um fator com 79 níveis (cada nível representando um bairro), Tempo é um fator com 12 níveis (cada nível representando um mês), e Água é um fator com 2 níveis (cada nível representado a informação de fornecimento ou não de água tratada). Pode-se dizer que pelo diagnóstico dos resíduos (ver Figura 7.1 do Anexo) e também pela comparação do valor observado da função desvio com os percentis da distribuição qui-quadrado da qual obtemos um p-valor de 6%, o modelo é razoavelmente adequado. Seria interessante adicionar mais covariáveis nesse modelo para explicar melhor a média de casos de dengue, mas, como se trata de uma ilustração, assumimos que essas variáveis explicativas são suficientes. Descrevemos na Tabela 1.1 os resultados dos três fatores. Note-se que, quando testamos a inclusão do fator localização e depois o fator tempo, estes influenciam na média de casos de dengue significativamente. Já o fator água não influencia em nada, ou seja, esse modelo está saturado.

Tabela 1.1: Análise do desvio referente ao exemplo sobre casos de dengue em Vitória.

Modelo	Desvio	Diferença	G.L	Testando	Chi
Constante	6002.4	-	-	-	-
+ Localização	4175.7	1826.7	78	Localização	0.00001
+ Tempo	924.0	3251.7	11	Tempo|Localização	0.00001
+ Água	924.0	0	1	Água|Localização+Tempo	1

Parece ser útil, através do exemplo dado, a relevância da componente espacial. Por esse motivo, as técnicas de análise de dados espaciais devem ser escolhidas de acordo com o tipo de problema e os dados envolvidos. A seção 1.2 procura identificar os tipos de dados.

1.2 Tipos de dados em Análise Espacial

Para caracterizar os problemas de análise espacial, a taxonomia mais utilizada considera quatro tipos de dados: *Eventos ou Padrões Pontuais, Superfícies Contínuas, Áreas com Contagem ou Taxas Agregadas e Análise de dados de Interação Espacial.*

O objeto de interesse, no caso da *análise de Padrões Pontuais*, é a própria localização geográfica dos eventos, na região em estudo, que poderiam ser casos de diversas doenças, tipos de crimes, distribuição de plantas, localizações do centro de vulcões, etc. O objetivo é, pois, estudar a distribuição geográfica desses pontos, testando hipóteses sobre o padrão observado tais como: se é aleatório, apresenta-se em aglomerados ou são regularmente distribuídos.

Figura 1.1: Mapa de Vitória dividido por bairros.

Vejamos: se um aglomerado é detectado, investiga-se se tal achado tem associação com algum fator de risco. Para ilustrar podemos pensar que um aglomerado de doenças foi localizado ao redor de usinas nucleares. Com o intuito de verificar a causa das mesmas, testa-se, em outras regiões, a hipótese de que as usinas nucleares são responsáveis pelo elevado risco da doença.

Um conjunto de pontos localizados na região em estudo e sua componente espacial são dados importantes para a *análise de superfícies*. O objeto de interesse são as medidas das características e não o padrão dos pontos em si mesmos. Ou seja, o interesse está em entender o padrão das características medidas, em cada um dos pontos observado na região em estudo. Esses pontos são fixos e considerados como amostra de todos os outros possíveis. Com base nessa amostra, tenta-se então, modelar as medidas das características e estimar outros valores em locais onde a amostragem não foi realizada. O objetivo é, pois, reconstruir a superfície da qual se retirou e mediu as amostras. Um exemplo seria a predição das precipitações pluviométricas em toda a região em estudo baseando-se em alguns pontos escolhidos para receber as estações de monitoração responsáveis pelas coletas das medidas das variáveis de interesse.

Na análise de *Áreas com Contagem* e *Taxas Agregadas*, a característica de interesse não varia continuamente sob toda a região. O que se tem é uma realização dos valores das características de interesse medidas em sub-regiões disjuntas que cobrem a região em estudo. Esses locais poderiam ser os municípios, setores censitários, bairros, distritos ou quaisquer outras divisões político-administrativas usuais. O grande interesse não é a predição dos valores das características em algumas áreas, já que todas elas têm suas próprias medições e o interesse

não está voltado somente para o padrão da localização destas áreas. O que é relevante é a detecção e possível explicação do padrão espacial das características de interesse em termos das covariáveis medidas no mesmo conjunto de áreas e da configuração espacial das mesmas. Taxa de homicídio por área em centros urbanos, número de acidentes de trânsito, número de infrações cometidas por menores, taxa de mortalidade infantil e as medidas socioeconômicas associadas são exemplos de informações que podem ser medidas em nível de áreas (setor censitário, microrregiões, distritos, etc).

Na *análise de dados de interação espacial*, a atenção esta voltada para dados envolvendo pares de pontos ou pares de áreas que originam um padrão de fluxo, o qual deve ser modelado para possíveis predições e, com isso, salientar as necessidades de novos serviços ou melhorar os já existentes. Um exemplo é a trajetória dos pais de pacientes infantis que, necessitando de tratamentos especiais, são obrigados a percorrer longas distâncias entre a residência e um hospital especializado localizado em uma região mais afastada. A análise de tais fluxos pode servir de apoio para que políticas públicas sejam tomadas visando a melhorar a oferta deste serviço especializado.

1.3 Representação Computacional de Dados Geográficos

Mais recentemente, a análise estatística espacial tem despertado interesse e progresso devido ao desenvolvimento do *Sistema de Informação Geográfica* (SIG) que é um conjunto de ferramentas computacionais que coleta, edita, armazena a geometria, integra, analisa e mostra dados que estão *georeferenciados*, isto é, localizados na superfície terrestre e representados numa projeção cartográfica.

A proliferação de dados georeferenciados junto a necessidade de criar funções analíticas que sejam implementadas em pacotes do tipo SIG tem servido como fator de motivação para publicação de artigos. Em particular, novas funções têm sido desenvolvidas para análise estatística espacial (mencionadas na seção seguinte). Essas técnicas já estão disponíveis como bibliotecas no computador que podem ser chamadas de funções internas de alguns pacotes de análise espacial (TerraView, Satscan, ARC/INFO, R, etc). Elas fornecem uma variedade de ferramentas para visualização, exploração e modelagem para todos os tipos de dados em análise espacial, obtendo de maneira mais rápida e fácil os resultados.

1.4 Técnicas Usadas para Análise de Conglomerado

Há muitas ferramentas utilizadas na análise espacial para os quatro tipos de dados, mas vamos concentrar-nos em uma específica que é a *detecção de conglomerados*. Para isso, é essencial ter dados pontuais ou agregados em área para o uso correto. O objetivo deste estudo é, pois, identificar áreas ou agrupamentos de áreas geográficas com um risco, significativamente, elevado de um evento de interesse. Encontra-se, na literatura, uma extensiva referência para esse tipo de estudo e uma revisão dos vários métodos existentes pode ser encontrada em Marshall[33], Lawson e Kulldorff[30] e, mais recentemente, em Bailey[4].

Um conglomerado pode ser definido, segundo o dicionário em inglês[13](cluster), como *qualquer configuração de elementos que ocorra muito próxima.* Pode-se fazer uma inferência sobre os *elementos* referidos como sendo quaisquer eventos de interesse dos pesquisadores, tais como crimes violentos, mortalidade (homicídios, neonatal, pós-natal e etc), espécie de planta, animais raros, e, com uma freqüência maior, as doenças (AIDS, câncer, leucemia, dengue, febre amarela, leishmaniose, etc). Também, pode-se definir um conglomerado, segundo Knox[18], como *um grupo de ocorrências geograficamente limitado de tamanho e concentração tal que seja impossível de ocorrer por mero acaso.* Um fator que deve ser considerado na definição de conglomerado relaciona-se com a escala de medida. O conglomerado poderia ser definido como sendo o espaço de uma simples casa ou, até mesmo, uma região de centenas de quilômetros quadrados. A grande variação de escala se deve à diversidade da população e do ambiente geográfico. Além disso, um conglomerado caracteriza-se como temporal (a taxa de incidência de um evento é mais elevada durante um intervalo de tempo do que em outros), espacial (a taxa de incidência de um evento é mais elevada em algumas áreas do que em outras) e espaço-temporal (a taxa de incidência da doença é temporariamente maior em algum local do que em outros com maiores taxas e que variam com o tempo).

Os estudos de conglomerados apresentam várias abordagens que são feitas através de testes de hipóteses, e os métodos para detecção são classificados de acordo com as características e hipóteses feitas sobre o conglomerado. Besag e Newell[6] primeiro classificaram esse estudo como testes: Gerais e Focados. Lawson e Kulldorff[30] subdividiram os testes Gerais em Globais e testes para conglomerado Localizado. Segundo a classificação de Lawson e Kulldorff[30], os testes Globais investigam uma tendência geral dos eventos em formar um conglomerado (sem pré-especificar uma ou mais região do mapa). Os testes Focados pré-especificam o conglomerado na região em estudo e, em seguida, verifica-se se o risco de ocorrência do evento é elevado devido a uma fonte suspeita próxima. Os testes Gerais para conglomerado Localizados investigam uma tendência geral dos eventos em formar um conglomerado (sem pré-especificar uma ou mais região do mapa). Se essa tendência existir, então o teste estima a área mais provável. Nesse caso, a hipótese nula de risco constante de ocorrência de um evento na região, do ponto de vista estatístico, significa dizer que o valor esperado de casos é proporcional à população que está em risco naquele local.

Em função da proliferação de métodos para identificação de conglomerado, refina-se o critério para escolher o melhor deles em determinada situação. Para isso, um fator relevante é avaliar o poder de detecção em relação à hipótese alternativa de interesse. Nesse contexto, o poder é interpretado como a probabilidade do teste detectar um conglomerado quando este realmente existe. Maiores detalhes sobre o poder pode ser encontrada em Wartenberg e Greenberg[47] e alguns estudos envolvendo comparações entre métodos que pode ser estudado em Waller e Lawson[46], Alexander e Boyle[1] e Kulldorff e Tango[24].

Dentre os métodos de detecção de conglomerados pertencentes à classe dos Localizados, na prática, o que mais se destaca, atualmente, é o *Scan spatial* de Kulldorff e Nagarwalla[19] fundamentado na estatística de varredura espacial tem sido amplamente utilizado em virtude do poder de detecção (Kulldorff[24], Costa e Assunção[10]) e da habilidade de conceder um

nível de significância à estatística de teste via simulação Monte Carlo, reduzindo o erro tipo I. Entretanto, em sua formulação original, o método é condicionado à busca de conglomerados que apresentam geometria circular. Tal característica reduz, substancialmente, o custo computacional. Apesar dessa vantagem, o método apresenta limitações quando o conglomerado real passa a apresentar uma geometria irregular, detectando nenhuma ou pequenas áreas do mesmo. O tratamento da irregularidade do conglomerado tem sido abordado a partir de técnicas heurísticas computacionais, como o método de *Simulated Annealing* (Duczmal e Assunção[14]) ou delimitando uma região circular de tamanho fixo menor que a região de estudo e realizando uma busca exaustiva nas áreas contidas em seu interior (Tango e Takahashi[44]). Sob suposição de que as regiões que definem o conglomerado compartilham fronteira geográfica, foram propostos os métodos *dMST* e *Doubly* desenvolvido por Costa, Scherrer e Assunção[11] que promove o crescimento de conglomerados agregando-lhes as áreas vizinhas que favorecem a maximização da verossimilhança do conglomerado.

A próxima seção pretende esclarecer melhor a importância do tema abordado na dissertação.

1.5 Relevância das Técnicas de Detecção de Conglomerados

Estudos de detecção de conglomerados espaciais são procedimentos importantes na área de saúde pública. O diagnóstico preciso sobre a característica aleatória ou não de um determinado evento espacial como, por exemplo, uma doença contagiosa, e a delimitação da região geográfica de sua ocorrência possibilitam aos órgãos competentes a elaboração de políticas eficientes de controle e combate. Como resultado, procura-se identificar áreas geográficas com um risco significativamente elevado da região, a princípio, sem o conhecimento de quais e quantas áreas são.

Como já foi mencionado, o método *Scan* tem sido considerado o melhor para o estudo de conglomerados localizados. Ele se propõe a detectar a presença de conglomerados no tempo, no espaço ou no espaço-tempo em uma determinada região de estudo. A estatística *Scan* temporal utiliza-se de uma janela que move ao longo do tempo (que é uma dimensão), visitando cada período com o objetivo de encontrar em algum uma maior incidência de caso. Se avaliarmos uma região apenas com dados espaciais (que são duas dimensões), essa técnica encontrará um conglomerado em forma de um círculo, de tal forma que todos os indivíduos, pertecentes a essa zona, possuem uma chance maior de ser um caso do que os indivíduos vivendo fora da respectiva zona. Analogamente, se tivermos dados no espaço-tempo (que são três dimensões), essa técnica encontrará um conglomerado em forma de um cilindro, onde os indivíduos vivendo nessa área geográfica e também em um determinado período de tempo têm uma chance maior de ser um caso do que os indivíduos que vivem fora desse espaço e (ou) nesse período determinado.

Em algumas situações, o modo como o padrão de eventos (casos) cresce ao longo do tempo pode gerar conglomerados no espaço-tempo que possuam uma forma geométrica diferente de um cilindro. Nesse contexto, a aplicação da técnica *Scan* espaço-tempo não seria capaz

de identificar o conglomerado real, por inteiro. Isso se deve ao fato de que a geometria espacial do conglomerado real é arbitrária e o *Scan* apenas detecta um conglomerado de forma cilíndrica. Logo, a diferença entre um cilindro e uma geometria espacial irregular reduziria consideravelmente a capacidade de detectar o conglomerado real.

Tal suposição sobre o crescimento dos eventos (casos), ao longo do tempo, é razoável visto que, por exemplo, a incidência de uma epidemia pode ter seu ponto de partida em um determinado lugar, tempo e numa sub-região específica. Com o passar do tempo, a doença é transmitida a mais pessoas e mais áreas são afetadas e as medidas de combate só podem ser feitas quando se tem o conhecimento das regiões afetadas, ainda que a doença continue a crescer em outras regiões. Sendo assim, o conglomerado pode tomar formas arbitrárias e fragmentar-se no decorrer do tempo.

O presente trabalho é importante porque a utilização do uso da estatística *Scan* espaço-tempo é limitada em casos nos quais o conglomerado real apresenta o tipo de comportamento descrito anteriormente. Acredita-se que, nessas condições, o poder de detecção seja reduzido significativamente.

1.6 Objetivos

Diante da limitação do método *Scan* espaço-tempo proposto por Kulldorff e Nagarwalla[19] para casos nos quais o conglomerado real apresenta uma geometria arbitrária (diferente de um cilíndro), a pesquisa propõe desenvolver uma metodologia inédita que se baseia na variação da estrutura de vizinhança espacial e avalia dados no espaço e no tempo. O método consiste na definição de uma estrutura de vizinhança espaço-temporal representada sob a forma de um grafo. Uma vez definida a estrutura, criam-se novas técnicas heurísticas para detecção de conglomerados no espaço-tempo cuja geometria de busca é não-cilíndrica. Em seguida, comparam-se essas heurísticas com o método tradicional *Scan* espaço-tempo.

Como objetivos específicos, temos:

- Apresentar os principias métodos de detecção para conglomerados localizados.
- Propor uma generalização do método espacial de geometria arbitrária para o método *Scan* espaço-tempo de detecção de conglomerados espaciais.
- Avaliar e comparar o poder de detecção dos métodos: *Scan, dMST* e *Doubly* na versão espaço-tempo para dados simulados utilizando-se como cenário o *Novo México*.
- Apresentar resultados de detecção de conglomerados espaço-tempo dos métodos: *Scan, dMST* e *Doubly* em dados de casos de doenças para as regiões de Vitória e Novo México (casos reais).

1.7 Estrutura da Dissertação

No Capítulo 2, abordaremos os pertinentes aspectos à análise do conglomerado: razões para examiná-lo, formulação dos modelos adequados, suas hipóteses e exemplos de testes. No

Capítulo 3, demonstraremos a teoria estatística puramente geográfica que envolve os métodos *Scan*, *dMST* e *Doubly*. No Capítulo 4, apresentaremos a teoria estatística espaço-temporal do método *Scan* e, em seguida, a proposta de varredura arbitrária que avalia dados no espaço-tempo (grafo) que será usada pelos métodos *dMST* e *Doubly*. No Capítulo 5, mostraremos a forma de simulação dos casos, as medidas de avaliação do poder de detecção dos métodos *Scan*, *dMST* e *Doubly* na versão espaço-tempo e os resultados obtidos no estudo. No Capítulo 6, apresentaremos duas aplicações reais de dados epidemiológicos para os métodos estudados. Finalmente, o Capítulo 7 mostrará as conclusões.

Capítulo 2

Revisão dos Métodos de Detecção de Conglomerados

2.1 Introdução

A análise de conglomerados de doenças é de grande interesse nas áreas de Epidemiologia e Saúde Pública. Desde a década de oitenta estudam-se os efeitos ambientais que prejudicam a saúde das pessoas e, por isso, vários métodos de conglomerados foram desenvolvidos a fim de melhor avaliar o comportamento de uma doença em determinada área e tempo. Neste capítulo, falaremos sobre os pertinentes aspectos à análise de conglomerado. Primeiramente, a justificativa para o interesse em examiná-lo. Em segundo, os modelos apropriados de conglomerados que procuram explicar a distribuição dos eventos em determinada região. Em seguida, a estrutura geral dos testes e exemplos dos mesmos.

2.2 Interesse no Estudo de Conglomerados de Doença

Há três situações nas quais a análise estatística de conglomerados de doenças mostra ser relevante para a área de saúde:

- Em pesquisa epidemiológica, no estudo da etiologia de uma doença.
- Em saúde pública, como parte de um sistema de vigilância geográfico de uma doença.
- Em resposta a alarmes de conglomerados de doenças para avaliar se uma investigação epidemiológica mais apurada seria ou não necessária.

2.2.1 Encontrar a etiologia de uma doença

Existem muitas maneiras nas quais a hipótese sobre a etiologia de uma doença é generalizada. Uma delas seria utilizar-se de métodos de detecção de conglomerados para varrer sistematicamente, uma grande área à procura de conglomerados sem, a princípio, ter nenhum conhecimento onde eles possam estar localizados, esperando, assim, detectá-los. Espera-se,

desse modo, obter idéias sobre alguma etiologia desconhecida. Ou seja, um conglomerado de doença é visto como uma consequência de uma aglomeração de fenômenos que causam a doença.

2.2.2 Avaliação de Alarmes de Conglomerados de doença

Alarmes de conglomerados de doença é fato comum. Wartenberg e Greenberg[47] descobriram que, somente em 1989, receberam aproximadamente 1.500 requerimentos para investigar conglomerados de câncer. Muitos desses alarmes foram facilmente tratados através de informações recebidas pelo telefone, mas outros precisaram investigações mais extensivas. Com o crescente interesse da comunidade envolvida pelos problemas e a preocupação de órgãos públicos com os recursos gastos, os métodos de investigação de conglomerados passam a ser úteis para confirmar ou rejeitar alarmes de doenças. Esse procedimento ajuda na decisão do órgão competente de necessitar ou não de gastos públicos para a referida doença.

2.2.3 Sistema de Vigilância em Saúde Pública

Para muitas doenças, há fatores de risco bem estabelecidos e, assim, é preciso investigá-los. Nesse contexto, o objetivo é ficar, constantemente, em alerta para quaisquer alarmes de doenças. Caso isso ocorra, pode-se de uma forma bastante rápida, avaliar o ocorrido antes mesmo que a mídia ou a população fiquem cientes do acontecimento. Se o alarme não indicar um conglomerado significante, então, pode ser descartado ou, dependendo da natureza do alarme, as autoridades podem ser avisadas para que seja tomada decisão sobre uma possível investigação adicional. Caso o conglomerado detectado seja significativo, é natural olhar primeiro se há presença de fatores de risco conhecidos e comunicá-los aos responsáveis pela vigilância sanitária para que tomem medidas compatíveis com o caso.

2.3 Modelos de Conglomerados

Segundo Wartenberg e Greenberg[47], existem basicamente duas classes de conglomerados: *Hot-spot* e *Clinal*. No *Hot-spot*, o risco é elevado e constante nas áreas (ou área) que constituem o conglomerado (ver Figura 2.1). No *Clinal*, o risco não se mantém constante entre as áreas do conglomerado: este é elevado no centro, mas decresce à medida que vai se distanciando, para outras áreas, o risco adicional é desprezível (ver Figura 2.2).

Em resumo, considerando o centro do conglomerado como foco de risco à saúde, se existe uma suspeita de que há sub-região ao redor desse foco com uma taxa elevada e constante de doença, então neste caso, o melhor modelo a ser escolhido será o *Hot-spot*. Se, por outro lado, há uma suposição de que a taxa de doença é elevada apenas em uma pequena região ao redor do foco, seguida de um declínio no restante do mapa, então deve ser utilizado o modelo *Clinal*. Essas duas particularidades são usadas para desenvolver modelos estatísticos que investigam os conglomerados de eventos através dos testes de hipóteses.

Um modelo estatístico para o conglomerado que aborda as questões mencionadas pode ser apresentado por (Lawson e Clark[29]):

$$E(Y_j) = \mu_j = \beta n_j g_j(\varrho, \xi_1, ..., \xi_\varrho, \nu), \tag{2.1}$$

onde Y_j é a variável aleatória que desempenha o volume de casos na área j, μ_j e n_j representam, respectivamente, o valor esperado de casos e a população em risco nesta área. β é a taxa global de incidência da doença, estimada por $\beta = \frac{\sum_{j=1}^{k} y_j}{\sum_{j=1}^{k} n_j}$, onde k representa o número de áreas no mapa. O risco relativo é determinado por uma função $g(.)$ que representa alguma medida de exposição ao foco para cada indivíduo pertencente à área A_j. Esta função é parametrizada em termos do número ϱ desconhecido de conglomerado, do conjunto de localizações $(\xi_1, ..., \xi_\varrho)$ com ξ_j representando o centróide de A_j e do parâmetro ν que está relacionado ao decaimento do risco em volta do conglomerado. O centróide, nesse caso, é o centro da massa do polígono que delineia à área e é representado pelas suas coordenadas geográficas.

2.3.1 O Modelo Hot-Spot

Se o objetivo de busca é encontrar pequenos conglomerados localizados, o modelo *Hot-spot* é o mais conveniente. Presume-se que a população seja particionada em dois grupos formados por expostos e não-expostos, que também é conhecido como estudo caso-controle. Para essa situação uma possível especificação para $g_j(\varrho, \xi_1, ..., \xi_\varrho, \nu)$ pode ser dada por:

$$g_j(\varrho, \xi_1, ..., \xi_\varrho, \nu) = \begin{cases} \rho_j, & se \quad \xi_j \quad \text{faz parte do hot-spot,} \\ 1, & se \quad \xi_j \quad \text{não faz parte do hot-spot,} \end{cases} \tag{2.2}$$

desta forma, considerando áreas, somente, pertencentes ao conglomerado *Hot-spot*, $E(Y_j) = \beta n_j \rho_j$, as mesmas possuem o valor esperado de casos maior do que o valor βn_j.

2.3.2 O Modelo Clinal

No modelo *Clinal*, a informação de exposição é difícil de ser obtida, pois o aumento (ou diminuição) no volume de casos é proporcional à exposição entre a área A_j e o foco. Nesse caso, o mais apropriado é assumirmos $g_j(\varrho, \xi_1, ..., \xi_\varrho, \nu)$ como uma função não-crescente. Uma possível especificação para $g_j(.)$ é dada por (Tango[41]):

$$g_j(\varrho, \xi_1, ..., \xi_\varrho, \nu) = 1 + exp\left(\frac{-d_j}{\nu}\right), \tag{2.3}$$

onde d_j é a distância euclidiana entre o centróide (ξ_j) da área A_j e o centro do conglomerado.

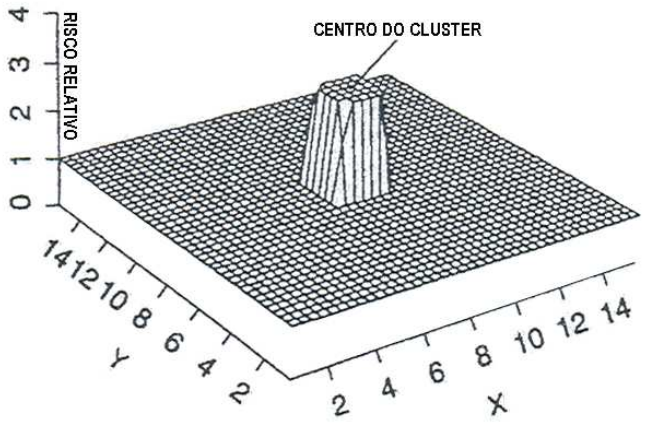

Figura 2.1: Superfície de risco relativo para o conglomerado Hot-spot. Os centros de cada área são representados pelas coordenadas geográficas (x,y).

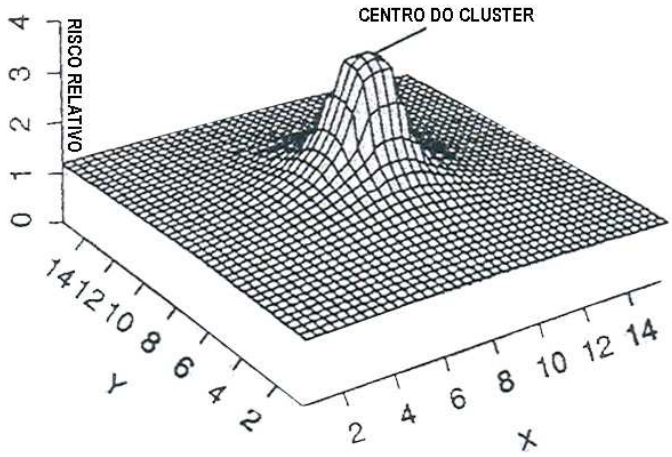

Figura 2.2: Superfície de risco relativo para o conglomerado Clinal. Os centros de cada área são representados pelas coordenadas geográficas (x,y).

O modelo estatístico do conglomerado definido por Lawson e Clark[29] demonstra que o risco relativo pode ser especificado por inúmeras funções para representar a medida de exposição ao foco. Sendo assim, o pesquisador pode escolher a função que mais se adaptar ao caso em estudo ou, também, criar outra função.

2.4 Hipóteses sobre o Conglomerado

O estudo de conglomerados espacialmente distribuídos é abordado através de testes de hipóteses e, por isso, muitos métodos estatísticos são desenvolvidos para detectá-los incorporando-lhe a variação geográfica da população em estudo. Nesses testes o importante é saber quando um padrão de eventos ou casos de uma doença, em uma ou mais áreas, ocorrem por mero acaso

ou não. Para isso, os modelos de probabilidade de Poisson (Besag e Newell[6]) ou Bernoulli (Cuzick e Edwards[12]) são, em geral, utilizados para avaliações dos testes para detecção de conglomerados ainda que outros se utilizam de ambos os modelos (Kulldorff[20]). As taxas ou contagens de casos de uma doença na população em estudo dependem do modelo selecionado. Explicando: o modelo Poisson é escolhido quando, o valor esperado de casos em cada sub-área sob a hipótese nula é proporcional ao tamanho da população em risco na região. Já o modelo Bernoulli é utilizado quando dividimos a população em risco em dois grupos casos e não-casos (controle). É importante esclarecer que, no caso de estudos de doenças raras como leucemia, por exemplo, ambos os modelos se aproximam.

Para a formulação da hipótese nula em ambos os modelos, considera-se que existem m áreas e que Y_j seja a variável aleatória que representa a quantidade observada de casos na área j, n_j a população em risco e μ_j o valor de casos esperados nesta área. A hipótese nula para o modelo de Poisson é que $Y_j \sim Poisson\ (\beta n_j)$ isto é:

$$H_0 : E(Y_j) = \mu_j = \beta n_j,$$

onde β é a taxa global da região que é estimada pela divisão entre o total de casos e a população total.

Para o modelo Bernoulli, supõe-se que a probabilidade de um indivíduo ser um caso em uma estabelecida sub-região é p, e a de ser um caso fora dessa sub-região é q. A hipótese nula para o modelo admite que as probabilidades de ser infectado são as mesmas para todos indivíduos na região em estudo, ou seja, $H_0 : p = q$.

Conclui-se que, nos modelos de Poisson e Bernoulli, os casos de doenças ocorrem aleatoriamente em toda a região sob a hipótese nula.

2.5 Testes Disponíveis para Detecção de Conglomerados

A proposta desta seção é descrever, resumidamente, alguns testes que podem ser usados para investigar o padrão espacial de eventos. Não será preocupação nossa o modelo estatístico utilizado para cada método descrito. Maiores detalhes sobre a aplicação de tais métodos podem ser obtidas nas referências mencionadas nesta seção. Lembrando que os testes para detecção de conglomerados podem ser classificados em: Testes para conglomerados Globais, conglomerados Localizados e Focados.

2.5.1 Método fundamentado em quadrante

O polonês Choynowsky[9] propôs o primeiro método fundamentado em quadrantes para a detecção de conglomerados espaciais, interessado que estava em estudar a distribuição geográfica de casos de tumores no cérebro de pessoas em sessenta municípios na Polônia. A idéia básica era construir um mapa de probabilidades de ocorrência da doença sob a hipótese de que a verdadeira ocorrência em todas áreas fosse a mesma. Esse método que se baseia em taxas brutas para o cálculo da probabilidade em cada área, apesar de simples, não leva em

conta a distinção das taxas (função da proporção da população de cada área). Assim, áreas com pequenas populações tinham taxas com grande variabilidade.

O teste, proposto por Choynowsky, avalia cada uma das áreas separadamente para determinar se o volume de casos é significativamente alto, atribuindo uma medida de significância α. Em cada quadrante testado, individualmente, a abordagem introduz o problema de múltiplos testes e era incapaz de detectar conglomerados que não seguissem os limites geográficas dos municípios da região em estudo.

2.5.2 Máquina de Análise Geográfica(GAM)

Openshaw[36] implementou o GAM -Geographical Analysis Machine- no procedimento de avaliar cada uma das áreas separadamente a fim de detectar se o volume de casos da doença é significativamente alto, atribuindo uma medida α. Para isso, não se utilizou do método de quadrantes, e sim, em zonas circulares onde cada círculo é posicionado no centróide de cada área. Esse método usa múltiplos círculos de raio R, constante, sobrepostos permitindo, assim, que os conglomerados possam ter formas distintas daquelas impostas pelas delimitações geográficas dos municípios da região em estudo. Para testar a significância, calcula-se o ponto crítico c^*_{iR} para cada possível zona circular, de centro i e raio R. Esse ponto é igual ao percentil 99,8 da distribuição da variável aleatória C_{iR} (volume de casos na zona circular sob a hipótese de que os casos são distribuidos de forma aleatória sobre a região em estudo). Um determinado círculo que possuir uma quantidade de casos que excede o ponto crítico (c^*_{iR}) da distribuição C_{iR}, é considerado como uma área de alta incidência de casos e, então, o método traça o emaranhado de círculos significativos no mapa. O GAM mostra-se mais útil como método descritivo, indicando os vários possíveis conglomerados em uma região.

2.5.3 Procedimento de Permutação para Avaliação de Conglomerado(CEPP)

Turbull[45], desenvolveu o CEPP (Cluster Evaluation Permutation Procedure), método que se fundamenta no estudo de zonas circulares sobrepostas. Os círculos são traçados de tal maneira que tenham o mesmo tamanho populacional P, ou seja, as regiões vizinhas a um dado ponto são agregadas até se obter o tamanho P desejado. Sob a hipótese nula de que os casos se distribuem aleatoriamente na população, as variáveis aleatórias C_{KP} (volume de casos na zona k com tamanho da população fixa em P) têm a mesma distribuição de probabilidade. Sob estas condições, o CEEP apura-se a zona com maior volume de casos e então testa-se sua significância usando a simulação de Monte Carlo para obter amostras da distribuição sob a hipótese nula dessa zona.

2.5.4 TBN-(Test of Besag and Newell)

Besag e Newell[6], criaram o teste *TBN*, em que o volume de casos K é determinado como o tamanho do aglomerado a ser buscado. O ponto principal é, fixado o tamanho K do conglomerado, posicionar o círculo em um ponto no mapa e ir aumentando o seu raio e

agregando os centróides vizinhos até que o círculo tenha agregado a menor quantidade de centróides necessários para que o volume de casos, dentro do círculo, tenha no mínimo K casos. Um valor pequeno para o número de centróides vizinhos indica um potencial aglomerado, e sua significância é obtida pelo cálculo da probabilidade de observar o menor número de centróides.

2.5.5 Tango (C_λ)

Para finalizar esta seção, o teste C_λ proposto por Tango[41] é baseado mais para teste global de conglomerado, mas poderia ser usado para estimar a localização de um provável conglomerado. A estatística C_λ, da mesma forma que outros testes, necessita definir, a princípio, o tamanho do conglomerado que se deseja procurar por meio do parâmetro λ que, nesse caso, é escala de uma função que mede as vizinhanças entre as áreas que fazem parte do conglomerado.

2.6 Discussão

A definição, a priori, dos parâmetros que caracterizam o tamanho do conglomerado: no GAM, o raio R do conglomerado; no CEPP o raio P populacional; no TBN o raio K de casos; no C_λ o parâmetro λ não é precisa. Isto nos leva a repetir os testes usando valores diferentes para os parâmetros, pois as características do conglomerado pesquisado não são conhecidas. Consequentemente, além do vício de pré-seleção, a maioria dos métodos estatísticos para análise de conglomerados espaciais são descritivos, no sentido de que eles podem detectar a localização do conglomerado, mas não fazem nenhuma inferência sobre a descoberta, ou eles fazem a inferência, mas não têm a capacidade de detectar a localização do conglomerado. Uma importante característica dos testes fundamentados na razão de verossimilhança é que fazem as duas funções citadas, ou seja, quando a hipótese nula é rejeitada, pode-se localizar a especifica área do mapa que causa a rejeição. O Capítulo 3 apresenta, com maior detalhe, três métodos de detecção espacial que utlilizam a estatística de varredura.

Capítulo 3

Métodos de Detecção de Conglomerados Espaciais Fundamentados na Estatística de Varredura

3.1 Introdução

A fim de detectar e testar a significância do conglomerado local, sem o conhecimento, a priori, de seu tamanho e localização, foram empregados métodos fundamentados na estatística de varredura, conforme alguns exemplos registrados por Kulldorff e Nagarwalla[19] e em Costa, Scherrer e Assunção[11]. A metodologia busca solucionar o problema de ajuste de testes múltiplos, pois é fundamentada na razão de verossimilhança, que é inerente a outros métodos concorrentes, anteriormente implementados. Na terminologia matemática, dizemos que os métodos varrem o mapa em estudo, impondo-lhe uma janela que pode apresentar qualquer forma geométrica (Kulldorff[20]). Neste estudo, será abordado e discutido o desenvolvimento e utilização dos métodos usando uma geometria circular (Circular Spatial *Scan*) e uma geometria arbitrária (*dMST* e *Doubly*).

Esses métodos de varredura possuem três propriedades básicas:

- Geometria da área que é varrida.
- Distribuição de probabilidade que gera os casos sob a hipótese nula, Bernoulli ou Poisson.
- Tamanho e forma da janela de varredura.

Nas próximas seções será apresentado cada método de detecção de conglomerado espacial contendo sua descrição, teoria, suposições sobre as distribuições na construção do teste e seu algoritmo.

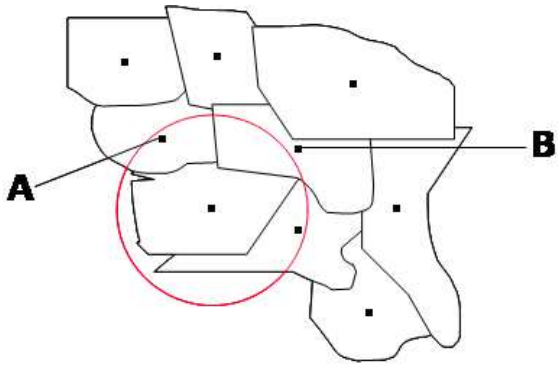

Figura 3.1: Limite geográfico de uma zona para dados agregados em área.

3.2 Scan Circular

O método de varredura *Scan* circular foi, inicialmente, formulado para detecção de conglomerados espaciais. Nesse contexto, o método se restringe à busca de conglomerados que apresentam geometria circular. Uma janela circular de raio variável é utilizada para varrer a região em estudo. Explicando: dada uma partição em sub-áreas da região de interesse, a janela circular é posicionada em cada um dos respectivos centróides e o seu raio é, continuamente, modificado, partindo de um valor nulo, que representa um conglomerado formado por uma única área, até um limite superior especificado pelo usuário. Esse limite pode ser representado pela porcentagem máxima da população permitida no conglomerado. O número total de círculos construído pelo método é infinito e, cada um deles, pode conter os mesmos conjuntos de áreas, incluindo áreas vizinhas. Para cada conjunto de áreas distinto que pertence a um determinado círculo é chamado de zona, onde cada zona é um possível candidato a conglomerado. Além disso, uma zona é formada por todos os indivíduos pertencentes a uma área na qual o centróide se encontra dentro do círculo. Sendo assim, apesar de o número de círculos ser infinito, as zonas serão finitas. Para dados pontuais, a zona é perfeitamente circular, ou seja, os indivíduos dentro da zona são exatamente aqueles localizados dentro do círculo definido. Para dados agregados em área, como por exemplo setor censitário ou bairros, uma zona pode apresentar um limite geográfico irregular, pois depende do tamanho e forma das áreas presente na região de estudo. As Figuras 3.1 e 3.2 mostram o limite geográfico e a varredura de uma zona para dados agregados em área respectivamente. Nota-se (Figura 3.1) que os indivíduos em "A" fora do círculo que define a zona, mas que pertencem a uma área cujo centróide esteja dentro do círculo, serão também incluídos na zona. Por outro lado, os indivíduos em "B", que estão dentro do círculo que define a zona, mas que pertencem a uma área cujo centróide não esteja dentro, não serão incluídos na zona.

Para cada possível zona, o método registra a quantidade de casos observados e esperados dentro e fora de cada círculo e em seguida calcula-se a verossimilhança assumida pelo modelo (Bernoulli ou Poisson). A zona que obteve o maior valor para a estatística calculada, dentre todas as zonas calculadas variando os centróides, atribui-se um teste baseado na razão entre verossimilhança sobre os modelos nulos e alternativo. A hipótese nula desse teste é de alea-

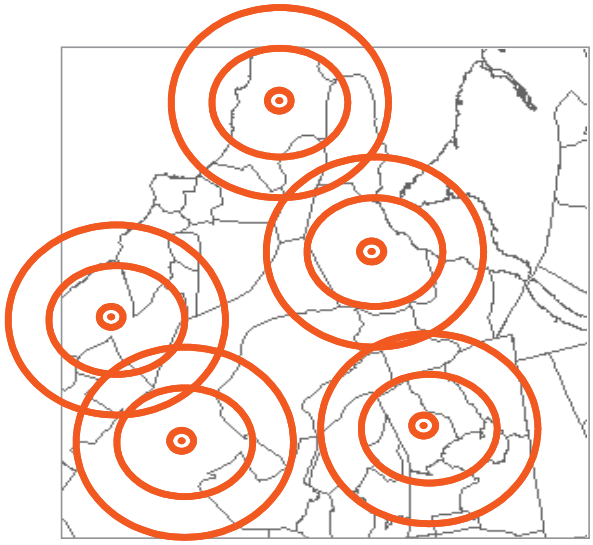

Figura 3.2: Varredura de uma região. Os círculos são centrados no centróide (.) de cada área e para cada centróide o raio cresce continuamente.

toriedade de casos na região ou ausência de conglomerado, e a hipótese alternativa é de que existe um conglomerado de ocorrências de casos (eventos) dentro do círculo que não ocorre por mero acaso. Para se testar essas hipóteses, simula-se a distribuição da estatística calculada condicionada ao número total de casos de acordo com o modelo probabilístico multinomial e proporcional a população de cada área com o propósito de obter o p-valor associado ao conglomerado. Caso o p-valor seja pequeno (menor que 0,05), pode-se dizer que a ocorrência do conglomerado detectado não é meramente aleatória. Esses cálculos são realizados usando o *software* SatScan (http://www.satscan.org/download.html) que implementa o método. A restrição circular para a geometria de busca reduz, significativamente, o número de candidatos a conglomerados e, consequentemente, o custo computacional.

3.2.1 Scan Circular para os modelos Bernoulli e Poisson

Denota-se C_j a variável aleatória que desempenha o volume de casos na j-ésima área A_j para j = 1,2,...,J , n_j é a população em risco desta área, C e N são o total de casos e a população no mapa de interesse. Defina-se também $\pounds$ o conjunto de zonas Z traçadas pelo método. Z será usada para representar a zona. Para o modelo Bernoulli, há, exatamente, uma zona Z (um único conglomerado) para o qual cada indivíduo passa a ter uma probabilidade p de vir a ser um caso, enquanto a probabilidade para os indivíduos fora de Z é q. Essas probabilidades são independentes para todos os indivíduos. A hipótese nula é $H_0 : p = q$. A alternativa é $H_1 : p > q, Z \in \pounds$. Sob H_0, $C_j \sim Bin(n_j, p)$ para todo A_j. Sob H_1, $C_j \sim Bin(n_j, p)$ para todo $A_j \in Z$ e $C_j \sim Bin(n_j, q)$ para todo $A_j \in Z^C$.

No modelo Poisson, há uma zona Z tal que $C_j \sim Poisson(pn_j)$ para todo $A_j \in Z$, e para todo $A_j \in Z^C$, $C_j \sim Poisson(qn_j)$. Nesse caso, p é compreendido como uma medida de fator de risco para a doença. As hipóteses nula e alternativa são as mesmas do modelo Bernoulli.

Este modelo possui a seguinte vantagem: ele pode ser ajustado para a população heterogênea e qualquer número de covariáveis.

Kulldorff[20] provou que, para a hipótese alternativa de apenas um conglomerado na região, o método é Uniformemente Mais Poderoso (UMP), entretanto para hipóteses alternativas de pequenos conglomerados dispersos em sub-regiões, esse método tem baixo poder de detecção de uma doença. Uma outra limitação é que algumas vezes, ele pode encontrar um conglomerado maior do que o real caso tenha formato muito diferente de um círculo. Nessa situação, a busca executada pelo SatScan, centrada num círculo, identifica conglomerados constituídos por áreas compactas englobando muitas áreas que, de fato, não pertencem ao conglomerado. Pode acontecer também que o algoritmo SatScan escolha um conglomerado pequeno que inclua poucas regiões do conglomerado real.

Modelo Bernoulli

Denota-se c_z o valor observado da variável aleatória C_Z que desempenha o volume de casos em Z. Agora, suponha-se que o modelo Bernoulli seja adequado para os dados. Então a expressão de verossimilhança referente ao modelo Bernoulli é dada por

$$L(Z,p,q) = p^{c_z}(1-p)^{n_z-c_z}q^{C-c_z}(1-q)^{(N-n_z)-(C-c_z)}. \tag{3.1}$$

O valor de p que aumenta a verossimilhança, não é necessariamente aquele do conglomerado que corresponde a maior taxa nem ao maior volume de casos c_z. Para achar a zona, dentre todas as possíveis, como sendo a mais provável, o teste desenvolvido por Kulldorff e Nagarwalla[18] utiliza a razão de verossimilhança,

$$\lambda = \frac{sup_{Z\in\pounds, p>q}L(z,p,q)}{sup_{p=q}L(z,p,q)} \qquad \text{com p,q} \in (0,1). \tag{3.2}$$

Sob H_0 os estimadores de máxima verossimilhança de p e q são dados por $\hat{p} = \hat{q} = \frac{C}{N}$, logo o denominador da equação 3.2 é reduzido por

$$sup_{p\in(0,1)}p^C(1-p)^{N-C} = \frac{C^C(N-C)^{N-C}}{N^N} = L_0. \tag{3.3}$$

L_0 depende somente do volume total de casos e, não, da sua distribuição espacial e é uma constante, pois foi condicionada em C. Sob a hipótese alternativa, os valores do numerador de p e q da equação 3.2 que aumentam a função de verossimilhança, para uma zona fixa z sobre todos os possíveis valores de $0 < q < p < 1$, são:

$$\hat{p}(z) = \frac{c_z}{n_z} \quad \text{e} \quad \hat{q}(z) = \frac{C-c_z}{N-n_z}, \quad \text{se} \quad \frac{c_z}{n_z} > \frac{C-c_z}{N-n_z}.$$

Logo

$$L(Z) = \left(\frac{c_z}{n_z}\right)^{c_z}\left(\frac{n_z-c_z}{n_z}\right)^{n_z-c_z}\left(\frac{C-c_z}{N-n_z}\right)^{C-c_z}\left(\frac{(N-n_z)-(C-c_z)}{N-n_z}\right)^{(N-n_z)-(C-c_z)},$$

quando $\frac{c_z}{n_z} > \frac{C-c_z}{N-n_z}$, caso contrário

$$L(Z) = \frac{C^C(N-C)^{N-C}}{N^N}.$$

Dessa maneira, a equação 3.2 pode ser expressa como:

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

Para encontrar o conglomerado verossímil é selecionada a zona $\hat{Z}$ na qual $L(Z)$ é maximizada, ou seja, $\hat{Z} = \{Z : L(Z) \geq L(Z^*) \forall Z^* \in \pounds\}$.

A distribuição de λ sob H_0 é muito difícil de ser obtida analiticamente e neste caso a aproximação assintótica para uma qui-quadrado pode não ser satisfatória (Tango[42]). Mas, a distribuição exata de λ condicionada ao volume total de casos observados C pode ser obtida utilizando um procedimento de simulação Monte Carlo, através do seguinte algoritmo:

1. Gerar B conjuntos de dados independentes, possuindo o mesmo número de casos C que o conjunto original, obtidos como realizações de uma distribuição multinomial e proporcional a população de cada área.

2. Para cada conjunto, calcula-se a estatística do teste da razão de verossimilhança obtendo $\lambda_1, \lambda_2, ..., \lambda_B$.

3. A partir da ordenação dos valores de λ para os B conjuntos simulados, compara-se o valor de λ, associado ao conjunto de dados original. Se este estiver entre os maiores $100\alpha\%$ valores, rejeite a hipótese nula ao nível de significância α.

4. Uma vez rejeitada a hipótese nula, então a zona $\hat{Z}$ associada com a máxima verossimilhança do modelo alternativo é o conglomerado mais verossímil.

Além do conglomerado mais provável (conglomerado primário), o método também executa conglomerados secundários que são os que possuem valores calculados de λ maiores que o percentil (1- α) da distribuição de λ sob H_0. Entretanto, os p-valores associados a esses conglomerados tendem a ser maiores que o p-valor do conglomerado primário.

Modelo Poisson

A função de verossimilhança para o modelo de Poisson é um pouco mais complexa que o modelo Bernoulli e é definida como:

$$L(Z,p,q) = \frac{e^{-pn_z - q(N-n_z)}}{C!} p^{c_z} q^{(C-c_z)} \prod_{j=1}^{J} n_j. \tag{3.4}$$

A estatística λ do teste da razão de verossimilhança pode ser escrita por:

$$\lambda = sup_{Z \in \pounds} \frac{\left(\frac{c_z}{n_z}\right)^{c_z} \left(\frac{C-c_z}{N-n_z}\right)^{C-c_z}}{\left(\frac{C}{N}\right)^{C}} I\left(\frac{c_z}{n_z} > \frac{C-c_z}{N-n_z}\right), \tag{3.5}$$

se existe pelo menos uma zona Z tal que $\left(\frac{c_z}{n_z} > \frac{C-c_z}{N-n_z}\right)$, ou $\lambda = 1$ caso contrário. Onde $I(.)$ é a função indicadora.

Maiores detalhes sobre a maximização da função de verossimilhança para o respectivo caso podem ser encontrados no artigo do Kulldorff[20].

Para se escolher entre o modelo Bernoulli ou Poisson depende-se dos dados em estudo. Se temos um estudo de caso-controle é aconselhável usar o modelo Bernoulli, e caso exista alguma covariável relevante é o modelo Poisson. Por outro lado, se o volume de casos é pequeno da ordem de 10% ou menos da população em risco, ambos os modelos se aproximam.

O *Scan* circular se tornou bastante popular por ter um *software* (SatScan) que executa o ajuste de ambos os modelos. Esse *software* está disponível aos interessados, gratuitamente, e analisa dados no espaço, tempo e interação espaço-tempo usando uma versão multidimensional desta estatística de varredura. O algoritmo usado pelo SatScan é o seguinte:

1. Selecionar um ponto (centróide) no mapa em estudo.
2. Calcular as distâncias entre o ponto escolhido e os demais pontos, ordenando-as de forma crescente. Armazená-las em um vetor.
3. Repetir os passos 1 e 2 até que todos os vetores de distâncias pertencentes a região de estudo estejam armazenados, caso o ponto escolhido já existir anteriormente, interrompa o algoritmo, iniciando-o a partir de um novo ponto do mapa.
4. Selecionar novamente um ponto no mapa.
5. Traçar um círculo de raio zero centrado no ponto escolhido no passo 4 e aumentar gradativamente o seu raio de acordo com as distâncias encontradas no passo 2. Para cada novo círculo inserido, atualizar o número de casos c_z e a população n_z dentro desse.
6. Calcular a estatística λ usando um dos modelos Bernoulli ou Poisson para cada círculo do passo 5.
7. Repetir os passos 4 a 6 até que todos os pontos sejam escolhidos, caso o ponto escolhido já existir anteriormente, interrompa o algoritmo, iniciando-o a partir de um novo ponto do mapa.
8. Registrar o círculo com maior estatística λ dentre todos os círculos calculados anteriormente no passo 7.
9. Utilizar simulações de Monte Carlo para avaliar o erro tipo I do círculo de maior verossimilhança.

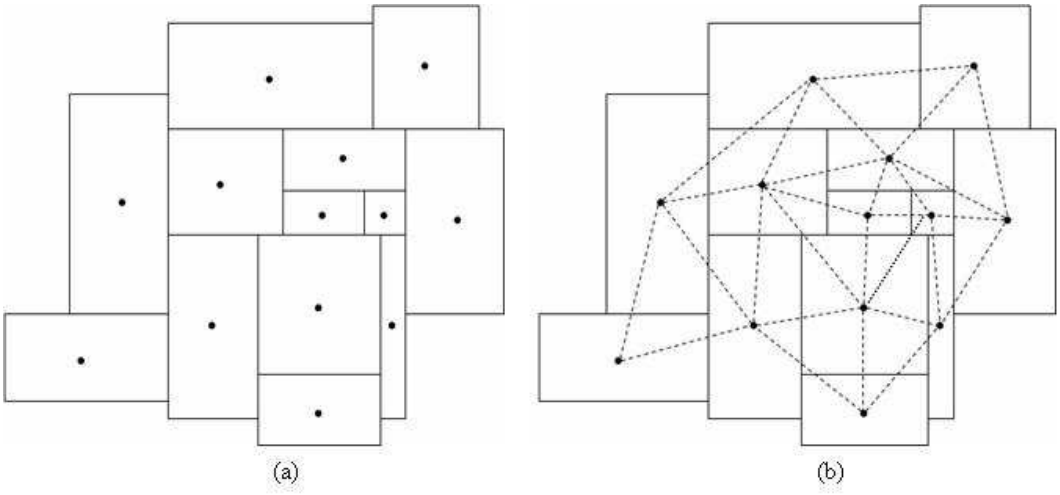

Figura 3.3: Estrutura de Grafo Interconectado. A informação de vizinhança é representada sob forma de um grafo interconectando os vizinhos que compartilham a mesma fronteira geográfica.

3.3 Método dMST(dynamic Minimum Spanning Tree)

O método de Crescimento de Árvore gera conglomerados a partir da informação de vizinhança geográfica das áreas em estudo e, não, com referência aos centróides das mesmas. Explicando da seguinte maneira: cada sub-área possui vizinhos que, por sua vez, também possuem outros vizinhos e assim, sucessivamente, de forma que, para uma particular sub-área i, exista, pelo menos, uma outra sub-área j que possua fronteira geográfica comum. É importante esclarecer que o termo vizinho diz respeito a, pelo menos, duas áreas disjuntas que tenham um ponto em comum. Pode-se expressar essa informação sob a forma de um grafo interconectando os centróides das sub-áreas aos seus vizinhos, conforme ilustra a Figura 3.3.

O algoritmo para construção de conglomerados, denominado *dMST (dynamic Minimum Spanning Tree)*[11], opera de forma semelhante ao algoritmo *Scan circular* sendo a direção de crescimento da árvore orientada segundo a disposição dos grafos. De forma suscinta, o algoritmo inicia o crescimento da árvore a partir de cada uma das regiões. A verossimilhança é então calculada, de acordo com o modelo escolhido (Bernoulli ou Poisson), para cada um dos vizinhos. A área vizinha que, quando agregada ao conglomerado, favorece a maximização da verossimilhança é definitivamente agregada ao conglomerado e a vizinhança do mesmo é então atualizada. O processo de crescimento é interrompido quando não existe, entre os vizinhos do conglomerado, uma área que, ao ser agregada, resulte em um conglomerado com a estatística λ maior que o conglomerado anterior, ou quando o conglomerado atinge um tamanho máximo. A Figura 3.4 mostra o modo de operação do algoritmo *dMST*.

Em sua proposta original, o critério de parada do algoritmo *dMST* consiste no valor do tamanho máximo do conglomerado especificado pelo usuário. Os resultados demonstram que este critério tende a gerar conglomerados com geometria muito arbitrária, semelhante aos encontrados pelo método de *Simmulated Annealing*[14]. Este fenômeno é denominado efeito

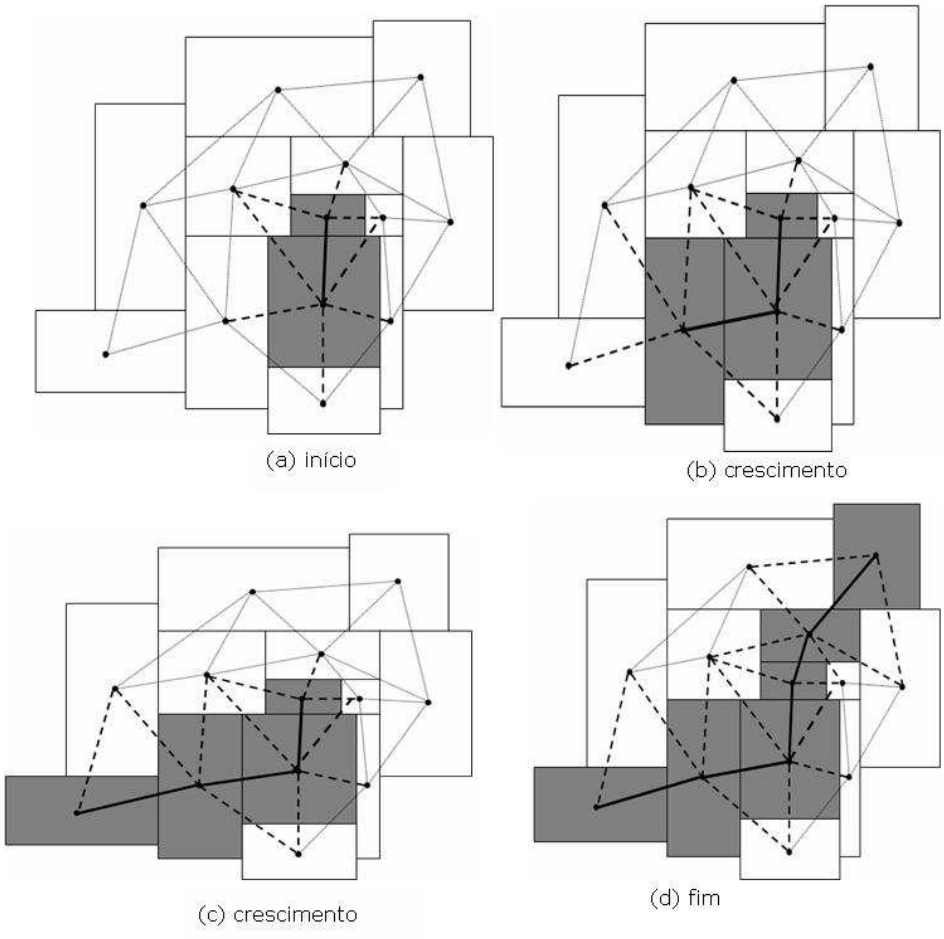

Figura 3.4: Modo de Operação do algoritmo dMST.

polvo (ver Figura 3.5) pois é representado por um conglomerado com alto valor da estatística de teste, mas com uma geometria extremamente arbitrária. A parada prematura garante conglomerados mais compactos, mantendo a característica de arbitrariedade do formato.

O Algoritmo *dMST* ainda está em fase de desenvolvimento, mas já existe um *software* (Spatial-Scan) que implementa o ajuste de ambos os modelos (Bernoulli e Poisson). Ele já está disponível, gratuitamente, no endereço http://www.est.ufmg.br/leste/spatialscan.htm e analisa dados no espaço usando uma versão bidimensional dessa estatística de varredura. O algoritmo para a construção de geometrias arbitrárias tem como objetivo a construção de árvores geradoras mínimas na qual o custo de agregação de uma área à árvore está associado à verossimilhança da árvore resultante. O algoritmo de crescimento da árvore geradora mínima utilizando a equação de verossimilhança do modelo Bernoulli ou Poisson é descrito a seguir:

1. Escolhendo uma sub-área da região, calcule a verossimilhança usando um dos modelos Bernoulli ou Poisson considerando os seus vizinhos como possíveis candidatos a aderir ao conglomerado.

2. Inclua na árvore o vizinho capaz de produzir o maior aumento na verossimilhança do

Figura 3.5: Efeito polvo.

conglomerado, caso não exista nenhum vizinho capaz de aumentar a verossimilhança, interrompa o algoritmo, iniciando-o a partir de uma nova área.

3. Atualize os vizinhos da nova árvore.

4. Retorne à etapa 2 e repita o procedimento até que todas as sub-áreas estejam incluídas na árvore geradora mínima ou até que a árvore alcance um tamanho máximo pré-definido.

5. Repetir os passos 1 a 4 até que todas as sub-áreas sejam escolhidas, caso a sub-área já estiver escolhida anteriormente, interrompa o algoritmo, iniciando-o a partir de uma nova sub-área do mapa.

6. Registrar a árvore geradora com maior estatística λ dentre todas as árvores calculadas anteriormente da etapa 5.

7. Em seqüência, o método de simulação de Monte Carlo é utilizado para o cálculo do nível descritivo associado à estatística da razão de verossimilhança λ, sob H_0, de forma semelhante ao método *Scan* circular.

3.4 Método Doubly

Várias restrições para o crescimento de conglomerados podem ser impostas ao grafo para limitar o tamanho e a geometria de busca. Dentre essas soluções, destaca-se o algoritmo *Doubly*[26] de crescimento de conglomerados. O método utiliza um critério simples de restrição do formato do conglomerado, mas mantém a dependência da estrutura de vizinhança. De

forma sucinta, o método inicia o crescimento do conglomerado de forma semelhante ao método *dMST*. Uma vez que o conglomerado atinge um tamanho pré-estabelecido, o algoritmo direciona o crescimento do conglomerado considerando somente as áreas conectadas a pelo menos duas áreas do conglomerado. Essa abordagem torna a geometria dos conglomerados mais compactada, gerando soluções intermediárias entre os métodos *Scan* circular e *dMST*. O cálculo do p-valor também é obtido via simulação de Monte Carlo. A Figura 3.6 apresenta o comportamento do algoritmo.

O *Doubly*, em fase de desenvolvimento, está disponível no mesmo *software* do método *dMST* (Spatial-Scan), que também implementa o ajuste de ambos os modelos (Bernoulli e Poisson). O algoritmo do *Doubly* utilizando a equação de verossimilhança do modelo Bernoulli ou Poisson é descrito a seguir:

1. Escolhendo uma sub-área da região, calcule a verossimilhança usando um dos modelos Bernoulli ou Poisson considerando inicialmente todos os vizinhos como possíveis candidatos a aderir ao conglomerado.

2. Inclua na árvore o vizinho que resulta na maior verossimilhança, caso não exista vizinhos capazes de gerar um conglomerado com maior verossimilhança, interromper o crescimento iniciando o mesmo a partir de uma nova área.

3. Atualize os vizinhos da nova árvore considerando os elementos que estejam conectados a pelo menos duas áreas do conglomerado.

4. Inclua na árvore o vizinho capaz de produzir o maior aumento na verossimilhança do conglomerado e que esteja conectado a pelo menos duas áreas do conglomerado, caso não exista nenhum vizinho capaz de aumentar a verossimilhança ou duplamente conectado, interrompa o algoritmo, iniciando-o a partir de uma nova área.

5. Retorne à etapa 3 e repita o procedimento até que não exista nenhuma região vizinha duplamente conectada ou até que a árvore alcance um tamanho máximo pré-definido.

6. Repetir os passos 1 a 5 até que todas as sub-áreas sejam escolhidas, caso a sub-área já estiver escolhida anteriormente, interrompa o algoritmo, iniciando-o a partir de uma nova sub-área do mapa.

7. Registrar a árvore geradora com maior estatística λ dentre todas as árvores calculadas anteriormente do passo 6.

8. Em seqüência, o método de simulação de Monte Carlo é utilizado para o cálculo do nível descritivo associado à estatística da razão de verossimilhança λ, sob H_0, de forma semelhante ao método *Scan* circular.

3.5 Discussão

É importante entender que quando os métodos de detecção de conglomerados espaciais: *Scan* circular, *dMST* e *Doubly* identificam uma região que é a mais verossímil, não, necessariamente,

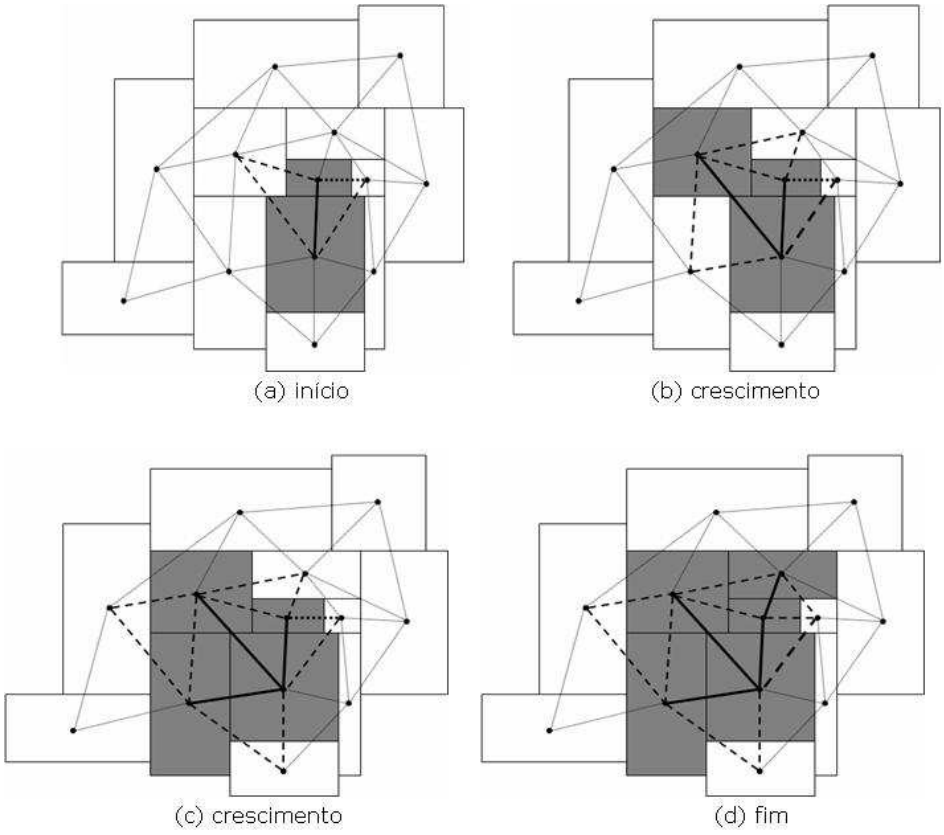

Figura 3.6: Modo de Operação do algoritmo Doubly.

coincidirá com o conglomerado real devido a uma possível diferença entre as geometrias dos conglomerados detectado e a do real.

Capítulo 4

Métodos de Detecção de Conglomerados Espaço-Tempo Fundamentados na Estatística de Varredura

4.1 Introdução

Em algumas situações, a aplicação de técnicas de detecção de conglomerados puramente espaciais, particularmente, no contexto epidemiológico, torna-se de pouco interesse ao pesquisador. Isso se deve ao fato de que o tempo de ocorrência dos eventos é registrado. Com isso, o padrão de eventos (casos) em um período de tempo fixado não seria tão informativo sobre o modo que esse padrão cresce ao longo do tempo. Queremos verificar, portanto, se existe uma interação entre o conglomerado espacial e o tempo.

Se repetirmos, com freqüência, a análise puramente espacial como parte de um período de tempo de um sistema de vigilância, teremos um baixo poder para detectar, recentemente, um conglomerado emergente. Também teremos o problema de ajuste de testes múltiplos devido à repetição de análises em cada parte do período.

A abordagem *Scan* espaço-tempo e também uma abordagem com geometria arbitrária é apresentada, nas seções 4.2 e 4.3, respectivamente, para resolver os problemas mencionados anteriormente.

4.2 Scan Circular Espaço-Tempo

Uma solução para a defasagem abordada na seção 4.1 é o uso da estatística *Scan* no espaço-tempo. Esse método já existe na literatura[22] e está implementado no *software* SatScan (http://www.satscan.org/download.html).

Ao invés de utilizar uma janela circular em duas dimensões, o *Scan* circular espaço-tempo impõe um cilindro em três dimensões. A base do cilindro representa o espaço, exatamente

como o *Scan* circular, e a altura corresponde ao tempo. O cilindro é flexível tanto na sua base circular geográfica quanto no período inicial e final do tempo, sendo um independente do outro. Isso significa que, para cada tamanho e localização de um possível círculo, consideramos cada período inicial e final para o conglomerado e vice-versa. Na notação matemática, denotamos $[T_1, T_2]$ como o intervalo de tempo no qual os dados existem e seja s e t o início e o final das datas do cilindro respectivamente. Então, consideramos todos os cilindros que estão no intervalo $T_1 \leq s \leq t = T_2$. Com isso, obtemos um número infinito de sobreposição de cilindros de tamanho e forma diferentes, cobrindo conjuntamente toda a região do estudo, onde cada cilindro reflete um possível conglomerado (ver Figura 4.1). Apesar de o número de cilindros ser infinito, os dados epidemiológicos contêm um conjunto finito de indivíduos, de tal maneira que muitos dos cilindros irão conter exatamente as mesmas pessoas. Essa situação nos leva a um conjunto finito de cilindros no qual a verossimilhança tem que ser calculada.

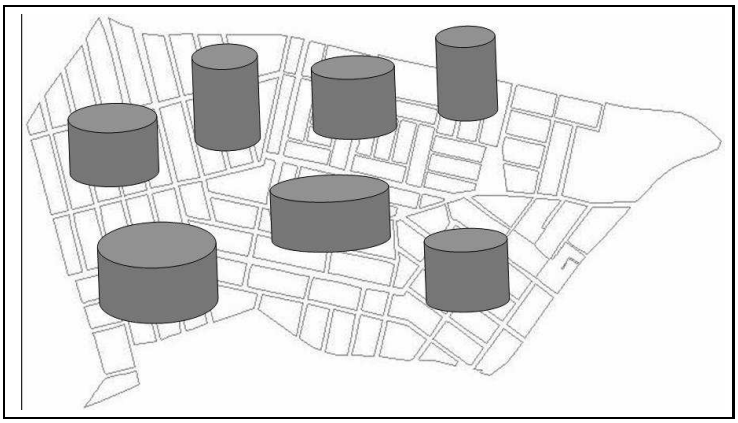

Figura 4.1: Alguns exemplos de cilindros possíveis para varredura de uma região. Os cilindros são centrados no centróide de cada sub-área e para cada centróide o raio e a altura crescem independentemente.

Assumimos que os eventos (casos de uma doença, por exemplo) ocorrem aleatoriamente no espaço e no tempo sob a hipótese nula, ou seja, os casos têm distribuição de Poisson ou distribuição Binomial com risco constante na região e no tempo, contra a hipótese alternativa que existe, pelo menos, um conglomerado espaço-tempo com elevado risco de ocorrência de casos dentro do cilindro quando comparado com os casos fora do cilindro. Para cada cilindro, o número de casos de doença dentro e fora é registrado, junto com o valor esperado da população sob o risco. Com base nesses números, o teste da razão de verossimilhança é construído da mesma maneira que o método *Scan* circular, puramente espacial, usando a Equação 3.1 para o modelo Bernoulli ou a Equação 3.4 para o modelo Poisson para cada cilindro. No cilindro com máxima verossimilhança e com maior número de casos observados do que o esperado é denotado o conglomerado mais verossímil.

A distribuição sobre a hipótese nula e o p-valor associado ao teste são obtidos via simulação de Monte Carlo (Dwass[15]). Com um nível de significância definido por 5%, p-valor menor

ou igual a 5% significa que o conglomerado mais verossímil é uma região de maior risco de incidência da doença.

4.3 Proposta do Estudo: Varredura Arbitrária no Espaço-Tempo

A proposta do presente trabalho é apresentar o estudo feito sobre varredura arbitrária em conglomerados levando-se em conta o espaço-tempo, uma vez que a técnica *Scan* circular espaço-tempo, então utilizada, tem limitações devido à sua geometria cilíndrica. Essa metodologia -variação da abordagem arbitrária puramente espacial- se apresenta como uma proposta inovadora.

Sendo assim, o uso da varredura arbitrária no espaço-tempo seria a solução para os problemas abordados na seção 4.1.

A varredura arbitrária baseia-se na definição de uma estrutura de vizinhança espaço-temporal onde as sub-áreas que compartilham fronteira geográfica, pelo menos duas áreas disjuntas que tenham um ponto de interseção, e fronteira temporal, a menor variação das mesmas áreas em tempos distintos, em comum são representadas sob a forma de um grafo interconectando os centróides das sub-áreas aos vizinhos que se encontram no espaço e no tempo; e também na codificação dos eventos (casos) e da população que estão no espaço-tempo para o espaço. Uma vez definida essa estrutura de vizinhaça e a codificação, pode-se utilizar, normalmente, os métodos de detecção puramente espacial *Doubly* e *dMST* para identificar o conglomerado arbitrário. Porém, o conglomerado detectado leva em conta, além de sua localização, o período de tempo onde surgiu.

Para melhor entendermos a metodologia proposta, usaremos um exemplo simples, e depois o expandimos para um caso mais geral. Vamos supor que exista uma região geográfica delimitada, sub-dividida em quatro sub-áreas. Os centróides de cada sub-região estão conectados aos seus vizinhos da forma como mostra a Tabela 4.1.

Tabela 4.1: Vizinhança da Região.

sub-área	**vizinha da sub-área**
1	2
1	3
2	1
2	4
3	1
3	4
4	2
4	3

Em tempos discretos conhecidos desta região geográfica, além da estrutura de vizinhança espacial (Figura 4.2), temos a população e o número de eventos (casos) registrados de cada

sub-área (Tabela 4.2). Deseja-se, então, identificar um grafo espaço-temporal para essa região.

Tabela 4.2: Matriz de Casos e de População da Região Geográfica.

Casos				**População**			
	tempo				tempo		
Sub-area	$t=0$	$t=1$	$t=2$	Sub-area	$t=0$	$t=1$	$t=2$
1	$C_{1,0}$	$C_{1,1}$	$C_{1,2}$	1	$P_{1,0}$	$P_{1,1}$	$P_{1,2}$
2	$C_{2,0}$	$C_{2,1}$	$C_{2,2}$	2	$P_{2,0}$	$P_{2,1}$	$P_{2,2}$
3	$C_{3,0}$	$C_{3,1}$	$C_{3,2}$	3	$P_{3,0}$	$P_{3,1}$	$P_{3,2}$
4	$C_{4,0}$	$C_{4,1}$	$C_{4,2}$	4	$P_{4,0}$	$P_{4,1}$	$P_{4,2}$

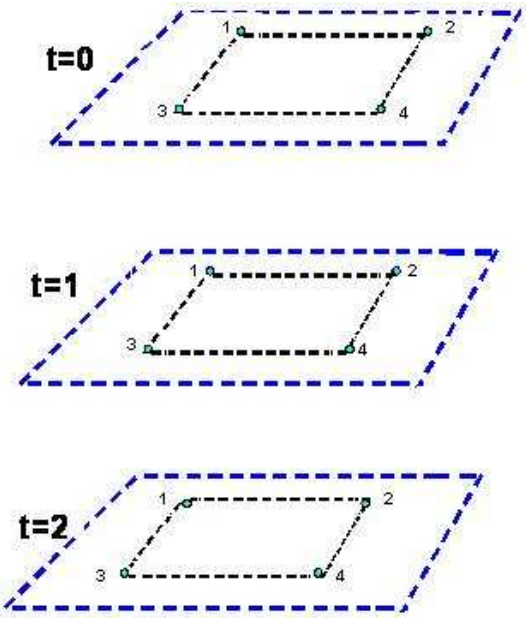

Figura 4.2: Estrutura de vizinhança da região geográfica em cada período de tempo.

Como visto anteriormente no capítulo 3, para o cálculo da verossimilhança de um candidato a conglomerado no espaço (Equação 3.1 para o modelo Bernoulli e Equação 3.4 para o modelo Poisson), necessita-se da população e do número de eventos associado em cada sub-região. Entretanto, para dados no espaço-tempo, a população e os eventos associados em cada sub-região variam de tempo a tempo. Portanto, para utilizar a equação de verossimilhança no espaço, nesse caso, deve-se fazer uma transformação da matriz de casos e da matriz de população para vetor casos e vetor população, respectivamente. Essa transformação é bastante simples, basta codificar os índices da matriz para vetor. A Tabela 4.3 mostra a codificação feita e a Figura 4.3 a estrutura de vizinhança espacial em cada tempo onde as sub-áreas foram codificadas.

Tabela 4.3: Vetor de Casos e de População Codificados da Região Geográfica.

Casos			**População**		
Sub-área		Sub-área-Tempo	Sub-área		Sub-área-Tempo
C_1	=	$C_{1,0}$	P_1	=	$P_{1,0}$
C_2	=	$C_{2,0}$	P_2	=	$P_{2,0}$
C_3	=	$C_{3,0}$	P_3	=	$P_{3,0}$
C_4	=	$C_{4,0}$	P_4	=	$P_{4,0}$
C_5	=	$C_{1,1}$	P_5	=	$P_{1,1}$
C_6	=	$C_{2,1}$	P_6	=	$P_{2,1}$
C_7	=	$C_{3,1}$	P_7	=	$P_{3,1}$
C_8	=	$C_{4,1}$	P_8	=	$P_{4,1}$
C_9	=	$C_{1,2}$	P_9	=	$P_{1,2}$
C_{10}	=	$C_{2,2}$	P_{10}	=	$P_{2,2}$
C_{11}	=	$C_{3,2}$	P_{11}	=	$P_{3,2}$
C_{12}	=	$C_{4,2}$	P_{12}	=	$P_{4,2}$

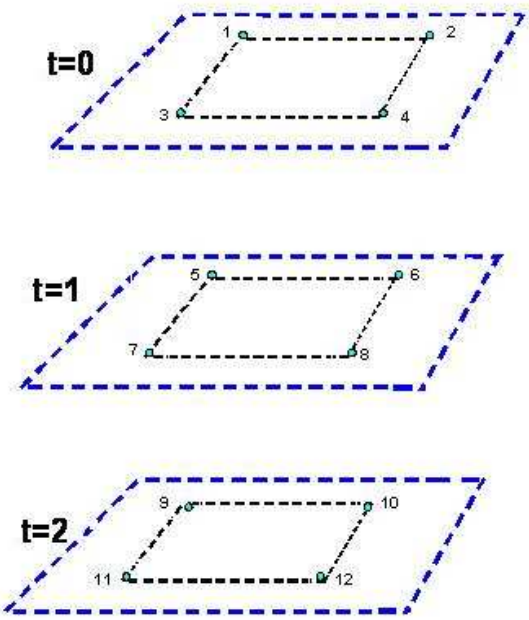

Figura 4.3: Estrutura de vizinhança da região geográfica codificada em cada período de tempo.

Uma vez definida a transformação, partimos do princípio que cada sub-área que varia ao longo do tempo apresenta característica semelhante a sua mesma sub-área somente no período seguinte e assim sucessivamente (fronteira temporal). Isso significa que originando sempre no tempo inicial ($t = 0$), a sub-área 1, por exemplo, tem característica semelhante a sua mesma sub-área (5) no tempo $t = 1$ e esta tem característica semelhante a sua mesma sub-área (9) no tempo $t = 2$. Baseado nessa idéia, conectamos os centróides de cada sub-região a sua mesma

sub-região somente no período seguinte e, assim, sucessivamente como mostra a Figura 4.4. Surge, assim, uma nova matriz da estrutura de vizinhança em que o tempo foi extinto (Tabela 4.4).

Tabela 4.4: Vizinhança da Região com o tempo extinto.

sub-area	**vizinha da sub-área**
1	2
1	3
2	1
2	4
3	1
3	4
4	2
4	3
5	6
5	7
6	5
6	8
7	5
7	8
8	6
8	7
9	10
9	11
10	9
10	12
11	9
11	12
12	11
12	10
1	5
2	6
3	7
4	8
5	9
6	10
7	11
8	12

Uma alternativa para obter-se a nova metodologia de varredura arbitrária no espaço-tempo seria acrescentar mais informação à estrutura de vizinhança espaço-temporal da região geográfica proposta na Tabela 4.4 e Figura 4.4. Essa informação consiste na conexão dos centróides de cada sub-região a seus vizinhos que estão somente no período seguinte e, assim, sucessivamente como mostra a Figura 4.5. Como resultado tem-se a matriz de estrutura de vizinhança espaço-temporal da Tabela 4.5. Isso significa que, originando sempre no tempo inicial ($t = 0$), a sub-área 1, por exemplo, está conectada a seus vizinhos (6 e 7) no tempo

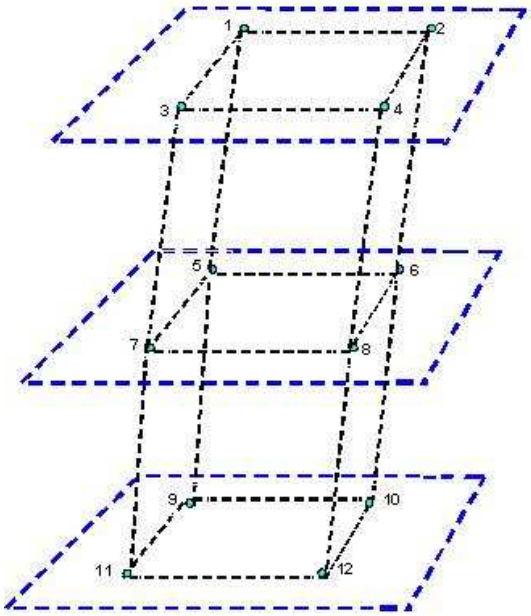

Figura 4.4: Nova Estrutura de vizinhança da região geográfica onde o tempo foi extinto.

$t = 1$ e essa sub-área, nesse tempo, está conectada a seus vizinhos (10 e 11) no tempo $t = 2$.

A partir da codificação dos dados no espaço-tempo para dados no espaço e da estrutura de vizinhança espaço-temporal criada, pode-se calcular a verossimilhança no espaço (Equação 3.1 para o modelo Bernoulli ou Equação 3.4 para o modelo Poisson) para os métodos *Doubly* e *dMST* utilizando a abordagem puramente espacial. O conglomerado espacial com máxima verossimilhança encontrado deve ser, então, decodificado para o espaço-tempo para verificarmos sua geometria espacial.

A distribuição sobre a hipótese nula e o p-valor associado aos testes *dMST* e *Doubly* no espaço-tempo são obtidos via simulação de Monte Carlo (Dwass[15]) dessa forma:

- Gerar 10.000 simulações.
- Para cada simulação: distribuir o número de casos total de cada ano entre as sub-regiões utilizando como referência a população do respectivo ano. Em seguida, utilizar a metodologia de conversão espaço-tempo-grafo.
- Calcular a estatística dos métodos *dMST* e *Doubly* para cada simulação.
- Ordenar as estatística simuladas, comparando-as com a observada.

4.4 Discussão

Apresentou-se uma nova abordagem para detectar conglomerados no espaço-tempo que utiliza a estrutura de vizinhaça espaço-tempo ao invés das coordenadas geográficas. A primeira

vantagem dessa abordagem é que pode ser aplicada em dois métodos de detecção puramente espacial (*dMST* e *Doubly*). Com isso, é possível com uma geometria arbitrária detectar um conglomerado emergente ao longo do tempo.

Tabela 4.5: Vizinhança Mais Informativa da Região com o tempo extinto e os vizinhos estão interconectados no tempo.

sub-area	**vizinha da sub-área**
1	2
1	3
2	1
2	4
3	1
3	4
4	2
4	3
5	6
5	7
6	5
6	8
7	5
7	8
8	6
8	7
9	10
9	11
10	9
10	12
11	9
11	12
12	11
12	10
1	5
2	6
3	7
4	8
5	9
6	10
7	11
8	12
1	6
1	7
2	5
2	6
3	5
3	8
4	6
4	7
5	11
5	10
6	9
6	12
7	9
7	12
8	11
8	10

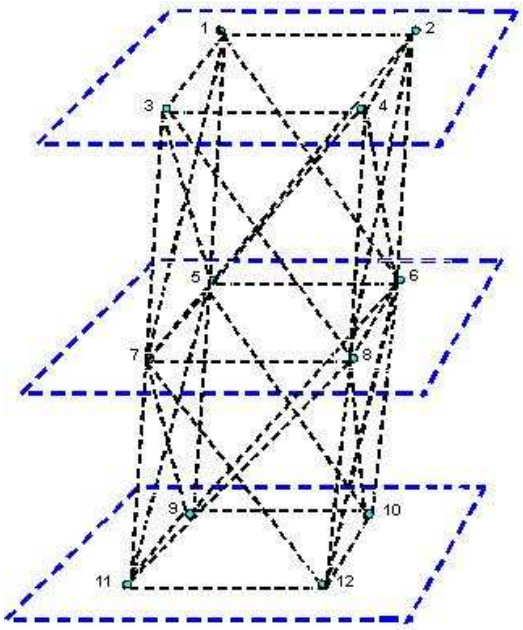

Figura 4.5: Nova Estrutura de vizinhança mais informativa da região geográfica onde o tempo foi extinto e os vizinhos estão interconectados no tempo.

Capítulo 5

Descrição da Simulação e Medidas de Avaliação do Poder

5.1 Introdução

A avaliação do poder estatístico de um teste é uma abordagem muito comum na comparação de diferentes métodos propostos para a análise estatística de conglomerados espaciais ou espaço-temporal de doenças. O poder estatístico é definido como a probabilidade do teste rejeitar corretamente a hipótese nula. Em termos de análise de conglomerados espaciais, o poder seria entendido como sendo a probabilidade do teste detectar um conglomerado espacial, quando realmente o conglomerado existe. Dados simulados podem ser usados para validar ou estabelecer propriedades do poder estatístico de métodos quando um modelo alternativo é assumido. Neste capítulo será descrito o mapa de interesse e a forma de distribuir os casos no mapa de tal maneira que exista no banco de dados simulados um conglomerado na região. Em seguida, medidas de avaliação do poder para os métodos *Scan*, *dMST* e *Doubly* serão adotadas para se apresentar os resultados.

5.2 Mapa de Interesse

O mapa de interesse é representado pelo estado do Novo México nos Estados Unidos subdividido em 32 áreas (condados). Utiliza-se, neste mapa, dados reais para a população. Em cada condado tem-se registrado os habitantes residentes que representam a população em risco. Esses dados são típicos de estudos epidemiológicos com a população variando do período de 1973 a 1991. A população total no início do estudo era de 1.104.347 habitantes no ano de 1973 e no final do estudo de 1.548.640 habitantes no ano de 1991. Como o condado de Cibola foi separado do condado de Valência em 1981 e nesse conjunto de dados a população de ambos está listada sob o condado de Valência, consideraremos os dois condados como se fosse apenas um. Cada uma destas 32 áreas (ver Figura 5.1) é geograficamente representada pelas coordenadas de seu centróide. Os dados de população e as coordenadas geográficas fazem parte de um banco de dados disponível na internet pelo endereço http://dcp.nci.nih.gov/bb/datasets.html.

Os dados simulados foram gerados a partir de dois cenários distintos para os conglomerados no espaço-tempo, apresentados na Figura 5.1, o primeiro cenário, e na Figura 5.2, o segundo cenário. O primeiro cenário especificou-se um conglomerado espaço-temporal com geometria cilíndrica constituído por 35 *condados-tempo*, sub-áreas que estão associadas a um determinado comprimento de tempo de ocorrência dos eventos (casos) de 5 anos, onde para cada ano tem-se as mesmas 7 sub-áreas. Já no segundo, especificou-se um conglomerado espaço-temporal com geometria arbitrária constituído por 39 *condados-tempo*, sub-áreas que estão associadas a um determinado comprimento de tempo de ocorrência dos eventos (casos) de 7 anos, onde para cada ano tem-se um total de sub-áreas diferente. Para este estudo foi atribuído um risco elevado e uniforme nas áreas pertencentes aos dois conglomerados no espaço-tempo referidos anteriormente (baseado no modelo de conglomerado *Hot-spot*), e em seguida, foi distribuído sobre cada população anual a quantidade total de casos de câncer no cérebro, de cada ano, associado a sua respectiva população de acordo com uma distribuição multinomial na qual as probabilidades referentes aos condados-anos do conglomerado espaço-temporal foram ajustadas a partir da especificação de um risco relativo, favorecendo a rejeição da hipótese nula com probabilidade 0,999 (Kulldorff e Tango[24]). A Tabela 5.1 mostra o total real de casos de câncer no cérebro ocorridos no Novo México para cada ano.

Tabela 5.1: Total Real de Casos de Câncer no Cérebro no Novo México para cada ano.

Ano	**Total de Casos**
1973	49
1974	55
1975	48
1976	39
1977	49
1978	53
1979	47
1980	58
1981	44
1982	61
1983	66
1984	54
1985	81
1986	81
1987	70
1988	77
1989	89
1990	69
1991	85

Uma vez definidos os parâmetros de simulação, foram geradas 10.000 simulações para cada cenário. Este processo será explicado na subseção seguinte.

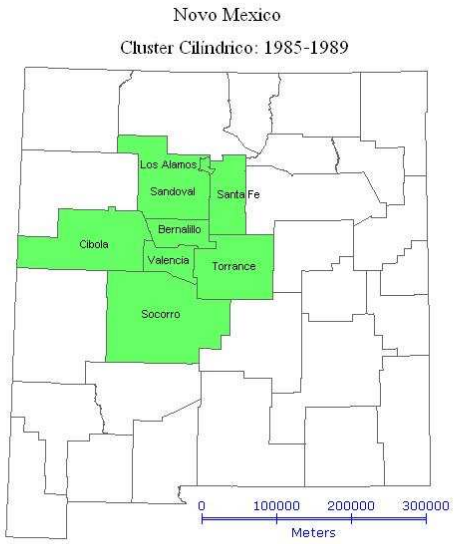

Figura 5.1: Primeiro Cenário de Conglomerado Espaço-Tempo entre os anos de 1985 a 1989 para dados simulados nos condados do Novo México.

5.2.1 Simulação dos Casos

Seja uma região sub-dividida em J áreas e denota-se C_j a variável aleatória que desempenha o volume de casos na j-ésima área e n_j o tamanho da população sob risco na área j, para $j = 1, 2, ..., J$. Suponha-se que $C_j \sim Poisson(\beta n_j)$. Então a hipótese nula de que não existe conglomerado no mapa é dada por:

$$H_0 : E(C_j) = \beta n_j,$$

em que $\hat{\beta} = \frac{TotalCasos}{Pop.Total}$. Significa dizer que o mecanismo gerador dos casos é um Processo Espacial de Poisson em que o valor esperado de casos em uma área j é o produto da taxa global β de casos e n_j que representa o tamanho da população sob risco na área j. Assuma que o volume total de casos observados C seja apresentado desta forma: $C = C_1 + C_2 + ... + C_J$, e, ao mesmo tempo, conhecido. Então, condicional a C, a distribuição conjunta é $(C_1, C_2, ..., C_J | C = c) \sim \mathbf{M}(c, \tau_1, ..., \tau_J)$ onde $\mathbf{M}$ representa o modelo estatístico multinomial e $\tau_j = \frac{n_j}{N}$ representa a probabilidade de um indivíduo vir a ser um caso na área j. N representa a população total da região, calculada por $N = \sum_{i=1}^{J} n_i$. Repare que tendo a informação do valor de C, C_j possui distribuição independente de β. Assim, a hipótese nula pode ser reescrita como:

$$H_0 : (C_1, C_2, ..., C_J | C = c) \sim \mathbf{M}(c, \tau_1, ..., \tau_J),$$

com $E(C_j) = c\tau_j$. Nessa situação, simular sob a hipótese nula é equivalente a gerar conjuntos independentes de vetores de casos, cuja soma dos elementos de cada vetor seja C, a partir de realizações de um modelo estatístico multinomial com os τ_j proporcionais à população de

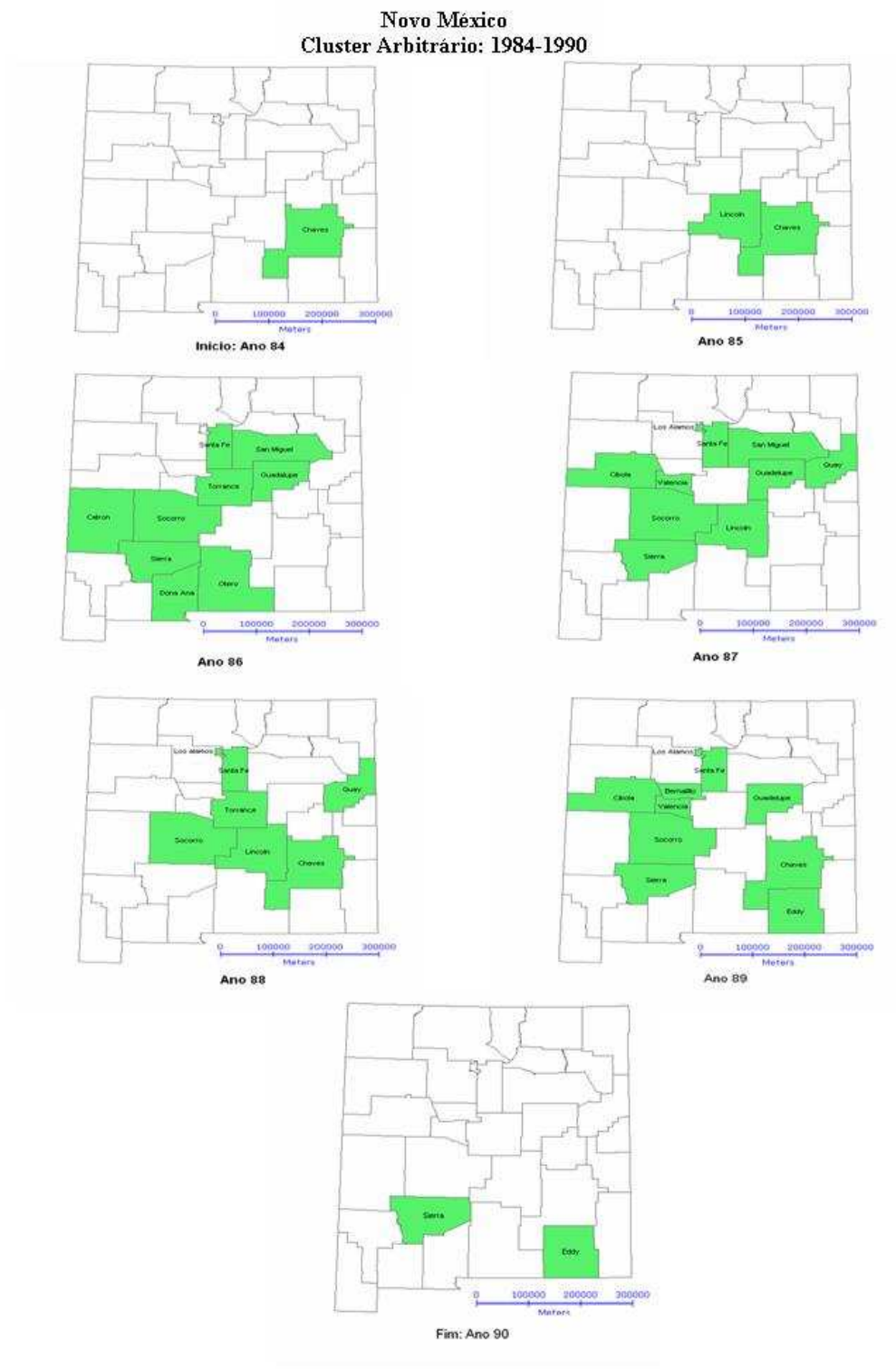

Figura 5.2: Segundo Cenário de Conglomerado Espaço-Tempo entre os anos de 1984 a 1990 para dados simulados nos condados do Novo México.

cada área.

Para simulação dos casos sob a hipótese alternativa, seja $s_1, s_2, ..., s_J$ os centróides de toda

região do mapa e suponha que exista uma zona Z constituída por um grupo de áreas. O risco relativo em cada área da região é determinado por:

$$\rho_j = \begin{cases} \rho_z, & se \quad s_j \in Z, \\ 1, & se \quad s_j \notin Z, \end{cases}$$

onde ρ_j representa o risco relativo para cada área. Nas sub-áreas que pertencem a zona Z, temos que $\rho_j > 1$. Então, a hipótese alternativa pode ser escrita como:

$$H_1 : E(C_z) = \beta n_z \rho_z, \tag{5.1}$$

onde $n_z = \sum_{j=1}^{J} n_j I(s_j \in Z)$ representa a população em risco na zona Z, $C_z = \sum_{j=1}^{J} C_j I(s_j \in Z)$ e $I(.)$ é a função indicadora. Então, por 5.1, o valor esperado de casos no interior do conglomerado (ou dos conglomerados) aumenta de acordo com o risco relativo ρ_z, supostamente, maior que 1. Logo, um teste equivalente é dado por:

$$\begin{cases} H_0 : & \rho_z & = & 1; \\ H_1 : & \rho_z & > & 1 \text{ para uma zona Z.} \end{cases} \tag{5.2}$$

Utilizando o mesmo raciocínio feito sob a hipótese nula, os casos são distribuidos condicionalmente a C e nos centróides de cada área. Dessa maneira, temos que o modelo estatístico é $(C_1, C_2, ..., C_J | C = c) \sim \mathbf{M}(c, \tau_1, ..., \tau_J)$, onde $\tau_j = \frac{n_j \rho_j}{\sum_{i=1}^{J} n_i \rho_i}$. Note que nesse caso τ_j passou a ser ponderado por ρ_j.

Com isso, a alocação dos casos no espaço-tempo neste estudo foi obtida usando simultaneamente simulações sob a hipótese nula e hipótese alternativa. Significa dizer que para períodos em que não se encontra o conglomerado espaço-tempo realizaram-se simulações sob a hipótese nula e para períodos onde o conglomerado espaço-tempo está contido, utilizam-se simulações sob a hipótese alternativa.

5.2.2 Cálculo do Risco Relativo Para os Conglomerados

Seja $A = \{s_1, s_2, ..., s_J\}$ o grupo constituído pelas áreas da região em estudo representadas pelos seus respectivos centróides. Considere também Z a região contendo as áreas que pertencem ao conglomerado existente no mapa. Assuma que este conglomerado seja do tipo *Hot-spot*, ou seja, o risco é elevado e constante nas áreas que constituem o conglomerado e fora dessas não há elevação do risco. Se para todo s_j pertencente a Z, faça $\rho_j > 1$ e $\rho_j = 1$ caso contrário. Então, sob o modelo alternativo, temos a probabilidade referente a cada área que é:

$$\tau_j = \frac{n_j \rho_j}{\sum_{i=1}^{J} n_i \rho_i}.$$

Nas áreas que formam o conglomerado (ou aos conglomerados) são concedidos riscos relativos maiores que 1 favorecendo a rejeição da hipótese nula com probabilidade 0,999 quando utiliza-se de um banco de dados simulados onde o conglomerado é especificado a priori. Para

definir este risco, Kulldorff e Tango[24] consideraram a situação descrita a seguir.

Seja n_z a população em risco do conglomerado na região, e $N = \sum_{j=1}^{J} n_j$ a população total de toda região. Condicionado ao volume total de casos C, o número observado C_z de casos na região do conglomerado sob a hipótese nula possui um modelo binomial com parâmetros (C, τ_z), onde $\tau_z = \frac{n_z}{N}$. A média e variância desse modelo são definidas, respectivamente, por:

$$m_0 = \frac{n_z C}{N} \quad e \quad v_0 = n_z \frac{C(N - n_z)}{N^2}. \tag{5.3}$$

Aproximando-se o modelo normal ao modelo binomial (ver figura 5.3), o valor crítico de casos k para que o teste unilateral rejeite a hipótese nula com nível de significância α é tal que:

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

onde $\Phi(.)$ é a função de distribuição acumulada da normal padrão. Se $\alpha = 0,05$ temos que $\Phi^{-1}(\alpha) = 1,645$, logo o valor crítico k é tal que:

$$\left(\frac{k - m_0}{\sqrt{v_0}}\right) = 1,645 \implies k = 1,645\sqrt{v_0} + m_0.$$

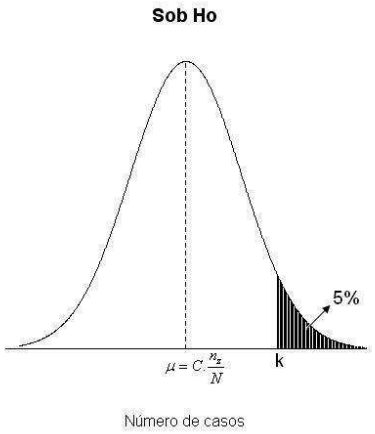

Figura 5.3: Aproximação normal para a distribuição binomial do número observado C_z de casos na região do conglomerado sob a hipótese nula.

Sob a hipótese alternativa, com risco relativo $\rho_j > 1$ para a região do conglomerado, o volume de casos nesta região possui um modelo binomial com média e variância definidas, respectivamente, por:

$$m_a = C\frac{n_z\rho_j}{N - n_z + n_z\rho_j} \quad e \quad v_a = C\frac{n_z\rho_j}{N - n_z + n_z\rho_j}\frac{N - n_z}{N - n_z + n_z\rho_j}. \tag{5.4}$$

Nota-se, neste caso, que $\tau_z = \frac{n_z\rho_j}{(N - n_z + n_z\rho_j)}$. Realizando, novamente, uma aproximação para o modelo normal (ver figura 5.4), encontre o risco relativo ρ_j do conglomerado tal que $\frac{(k - m_a)}{\sqrt{v_a}} =$

$\Phi^{-1}(\theta)$ seja solucionada. Dessa forma, o risco relativo é escolhido de modo que o poder atingido por qualquer teste para conglomerado espacial tem um limite superior igual a θ. Neste trabalho foi escolhido $\theta = 0,999$.

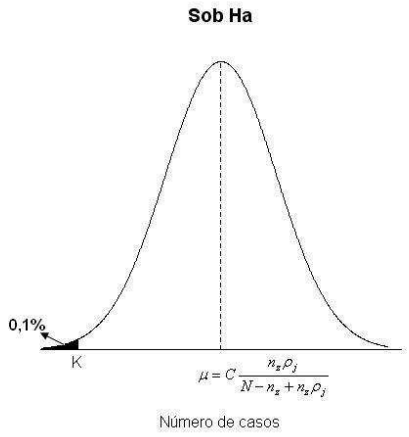

Figura 5.4: Aproximação normal para a distribuição binomial do número observado de casos na região do conglomerado sob a hipótese alternativa.

Para estabelecer o risco relativo para o conglomerado devemos seguir os seguintes passos:

- Condicionado ao volume total de casos na região C, ache o valor crítico de casos K em Z para que o teste unilateral rejeite a hipótese nula com o nível $\alpha = 0,05$ (neste trabalho). Considere $\tau_z = \frac{n_z}{N}$.
- Conhecido o valor k, junto com a expressão da média e da variância sob a hipótese alternativa, substitua em

$$\frac{(k - m_a)}{\sqrt{v_a}} = \phi^{-1}(\theta) = -3,09, \tag{5.5}$$

  onde o valor de $\theta = 0,999$.
- Encontre o valor ideal para ρ_j de tal maneira que a equação 5.5 seja solucionada.

A Tabela 5.2 apresenta os resultados dos riscos relativos estabelecidos para o conglomerado espaço-tempo no cenário 1 em cada ano com $\theta = 0,999$. Para esse conglomerado, obtivemos no ano de 1985 um risco relativo para cada condado de 2,85 para uma população em risco de 686.669 habitantes. Eram esperados naquele ano, 38,67 casos nesse aglomerado sob a hipótese nula e 58,52 casos nesse mesmo aglomerado sob a hipótese alternativa. Essa mesma análise encontra-se na tabela 5.1 para os outros anos em que o conglomerado exista.

Analogamente, a Tabela 5.3 apresenta os resultados dos riscos relativos, população sob risco, casos sob hipótese nula e alternativa para o conglomerado espaço-tempo de geometria arbitrária (cenário 2) em cada ano.

Observe que, em ambas as Tabelas (5.2 e 5.3), os riscos são diferentes em cada ano, e isto é devido à diferença entre as populações e as quantidades de casos observados anualmente.

Tabela 5.2: Resultado da simulação do conglomerado espaço-temporal Hot-spot que possui 35 áreas associadas à um comprimento de tempo de ocorrência dos eventos de 5 anos para o cenário 1, população do conglomerado por ano, quantidade total de casos da região, número esperado de casos sob a hipótese nula, sob a hipótese alternativa e risco relativo com $\theta = 0,999$ e $\alpha = 0,05$.

áreas do Cluster	**Ano**	$\mathbf{n_z}$	**Casos Simulados**	$\mathbf{E(c/H_0)}$	$\mathbf{E(c/H_a)}$	**Risco**
Bernalillo Sandoval Valencia Torrance Santa Fe Los Alamos Socorro	85	686.669	81	38,669	58,5211	2,85
Bernalillo Sandoval Valencia Torrance Santa Fe Los Alamos Socorro	86	704.028	81	38,9853	58,7747	2,85
Bernalillo Sandoval Valencia Torrance Santa Fe Los Alamos Socorro	87	722.764	70	34,2193	52,344	3,1
Bernalillo Sandoval Valencia Torrance Santa Fe Los Alamos Socorro	88	735.178	77	37,9833	57,063	2,94
Bernalillo Sandoval Valencia Torrance Santa Fe Los Alamos Socorro	89	747.821	89	44,254	64,9463	2,73

Essa diferença entre os riscos relativos estabelecidos para cada ano se torna importante quando avalia-se o poder do teste em situação na qual o método utilizado *(Scan,dMST* ou *Doubly)* pode identificar por inteiro ou parcialmente o período do conglomerado especificado, a priori, do banco de dados simulados.

Tabela 5.3: Resultado da simulação do conglomerado espaço-temporal Hot-spot que possui 39 áreas associadas à um comprimento de tempo de ocorrência dos eventos de 7 anos para o cenário 2, população do conglomerado por ano, quantidade total de casos da região, número esperado de casos sob a hipótese nula, sob a hipótese alternativa e risco relativo com $\theta = 0,999$ e $\alpha = 0,05$.

áreas do Cluster	**Ano**	$\mathbf{n_z}$	**Casos Simulados**	$\mathbf{E(c/H_0)}$	$\mathbf{E(c/H_a)}$	**Risco**
Chaves	84	56.458	54	2,1519	14,602	8,93
Chaves Lincoln	85	69.877	81	3,935	18,874	5,95
Sierra Otero Dona Ana Catron Socorro Torrance Guadelupe San Miguel Santa Fe	86	326.224	81	18,065	38,138	3,1
Quay Los Alamos Valencia San Miguel Guadelupe Socorro Lincoln Sierra Santa Fe	87	252.726	70	11,965	29,916	3,62
Quay Torrance Socorro Lincoln Chaves Santa Fe Los Alamos	88	217.123	77	11,2177	29,49	3,64
Bernalillo Guadelupe Valencia Socorro Chaves Sierra Eddy Santa Fe Los Alamos	89	795.550	89	47,08	67,35	2,77
Sierra Eddy	90	58.517	69	2,665	16,17	7,62

A partir dos riscos calculados para cada cenário iremos avaliar o desempenho do método *Scan* circular no espaço-tempo, *Doubly* espaço-temporal e *dMST* espaço-temporal.

5.3 Medidas de Avaliação para o Poder dos Testes Scan Circular, dMST e Doubly no Espaço-Tempo

O poder do teste dos métodos *Scan* Circular, *dMST* e *Doubly* no espaço-tempo para os modelos alternativos de conglomerados descritos na seção 5.2 é obtido via simulação de Monte Carlo. Matematicamente, esse poder é determinado por:

$$P_{H_1}(\lambda \in R),$$

onde λ é a estatística de teste encontrada, descrita na seção 3.2; R representa a região de rejeição do teste e P_{H_1} representa a probabilidade do teste rejeitar corretamente a hipótese nula.

Computacionalmente, estima-se essa probabilidade gerando B simulações independentes em que existe um conglomerado espaço-tempo real com risco relativo $\rho_z > 1$ para alguma zona Z dentro da região e calcula-se a estatística de teste em cada simulação:$\lambda_1, \lambda_2, ..., \lambda_B$, usando cada um dos métodos. Nesse caso, o poder é aproximado por:

$$P_{H_1}(\lambda \in R) \approx \sum_{j=1}^{B} \frac{I(\lambda_j \in R)}{B}.$$

Isto representa a freqüência de vezes em que os métodos (*Scan* Circular, *dMST* e *Doubly*) rejeitam a hipótese nula.

Neste estudo o *Scan* Circular espaço-tempo varrerá o mapa ao longo do tempo em busca de conglomerados que contenham no máximo 50% do total da população da região e para os métodos *Doubly* e *dMST*, no espaço-tempo, varrerão o mapa em busca de conglomerados que contenham no máximo 60 áreas associadas ao tempo de ocorrência dos eventos, considerando o banco de dados simulado e os bancos de dados reais (Novo México e Vitória). A razão dessas condições sobre a busca do conglomerado (50% do total da população ou 60 áreas associadas ao tempo de ocorrência) é que, na prática, seria, na maioria das vezes, inviável aos órgãos competentes elaborar políticas eficientes de controle e combate para situações em que as áreas a serem tratadas englobam grande parte da região em estudo. Também o número de amostras independentes sob hipótese nula para o cálculo de p-valor em cada simulação sob hipótese alternativa será 9.999.

Para avaliar os métodos de conglomerados localizados, além do poder, considera-se outro ponto importante: a habilidade do teste em encontrar, corretamente, todo ou pelo menos uma parte relevante do conglomerado real na região e ao longo do tempo. Isso porque, aleatoriamente, o método pode identificar conglomerados que contenham áreas e (ou) tempos situados fora do conglomerado real ou deixe de incluir áreas e (ou) tempos dentro do referido conglomerado. Nas próximas subseções, serão exibidas medidas básicas que se usam para avaliar

a habilidade e o poder dos métodos levando em conta esses aspectos. Antes de descrevê-las, julgamos importante estes aspectos: a informação do tamanho do conglomerado encontrado, tamanho da interseção e o comprimento do tempo do conglomerado. A primeira medida é a proporção de áreas detectadas no conglomerado real. A segunda, mais elaborada, usa a proporção da população detectada no conglomerado real e a terceira leva em conta o erro de classificação da população do conglomerado real. Estas medidas chamam-se de proporção de detecção do teste.

5.3.1 Proporção de Áreas Detectadas no Conglomerado Real

A proporção de áreas detectadas no conglomerado real é uma função do conjunto $\hat{Z} = \{s_j : j = 1, ..., \hat{M}\}$, em que $\hat{Z}$ é a região formada pelas sub-regiões do mapa, ao longo do tempo, que faz parte do conglomerado encontrado pelo método e $\hat{M}$ é o número de áreas detectadas.

Seja M o número de áreas que faz parte do conglomerado real Z e m o número de áreas que faz parte do conjunto $(Z \cap \hat{Z})$ que representa a interseção entre o conglomerado real e o detectado pelo método. A proporção de áreas detectadas no conglomerado real é determinada por:

$$P_A = \begin{cases} \frac{m}{M}, & se \quad M \geq \hat{M}; \\ \frac{m}{\hat{M}}, & se \quad M < \hat{M}. \end{cases}$$

Nota-se que P_A é, simplesmente, o mínimo entre $\frac{m}{M}$ e $\frac{m}{\hat{M}}$. Essa medida procura resolver a situação quando $M > \hat{M}$ e também quando $M < \hat{M}$. Observe que P_A=1 se o conglomerado detectado for igual ao conglomerado real.

Como ilustração, imagine que, em uma específica situação do estudo, o único conglomerado espaço-temporal pertinente no mapa seja $Z = \{s_7, s_{20}, s_{85}, s_{100}, s_{121}, s_{128}, s_{132}, s_{144}\}$ e que o método tenha detectado $\hat{Z} = \{s_{85}, s_{100}, s_{121}, s_{128}\}$, então $M = 8$, $\hat{M} = 4$ e $m = 4$, logo $P_A = 0,5$. Isso significa que o método encontrou corretamente 50% do conglomerado real.

Se em 10.000 repetições desse evento forem obtidas 6.000 aspectos diferentes com quatro áreas, em que estas áreas, sempre, fazem parte do conglomerado real, então o poder computado é de 0,6 para o teste detectar corretamente 50% do conglomerado real.

Essa medida apesar de ser calculada, facilmente, tem a desvantagem de não informar se o conglomerado detectado foi maior ou menor que o conglomerado real.

5.3.2 Proporção da População Detectada no Conglomerado Real

Em casos em que as populações envolvidas no conglomerado forem homogêneas, a medida da sub-seção 5.3.1 é eficaz, pois considera as áreas como igualmente influentes. Entretanto, se a distribuição espacial da população for heterogênea esse critério pode ocasionar problemas. Exemplificando: uma área com uma população pequena associada à população do conglomerado real pode causar uma mudança, considerável, na medida de P_A, caso ela não seja encontrada pelo método, o que não é realista. Uma forma de evitar esse problema é atribuir pesos nas áreas da região. Isso é feito através da proporção da população detectada no conglomerado real. Esta proporção é uma função da população heterogênea em $\hat{Z}$ e em sua

vizinhança. A população estimada no conglomerado encontrado é calculada por:

$$\hat{n}_{\hat{z}} = \sum_{j=1}^{\hat{M}} n_j, \ \ \forall s_j \in \hat{Z},$$

e a população do conglomerado real é n_z. A população, na região, de interseção entre o conglomerado real e o estimado é desempenhado por:

$$n\left(Z \cap \hat{Z}\right) = \sum_{j=1}^{J} n_j I\left(s_j \in Z \cap \hat{Z}\right),$$

ou seja, verificamos se cada sub-área faz parte da interseção entre o conglomerado real e o detectado. Em seguida, somam-se todas as populações referentes as sub-áreas que pertencem à interseção.

A proporção de detecção da população detectada no conglomerado real é determinada por:

$$P_P = \begin{cases} \frac{n(Z \cap \hat{Z})}{n_z}, & se \quad n_z \geq \hat{n}_{\hat{z}}; \\ \frac{n(Z \cap \hat{Z})}{\hat{n}_{\hat{z}}}, & se \quad n_z < \hat{n}_{\hat{z}}. \end{cases}$$

A distinção dessa proporção para a primeira é que esta é fundamentada na razão populacional que cada área representa para o conglomerado. A adição ou não de uma área com pequena população em relação ao conglomerado real pouco modifica esta proporção. O propósito nesse caso é, também, computar o peso da distribuição da população do conglomerado no poder do teste. A interpretação dessa proporção é a mesma que a proporção da sub-seção 5.3.1.

5.3.3 Proporção de Erro na Classificação da População do Conglomerado Real

Através dessa proporção temos uma idéia de quanto os métodos erram em identificar a população do conglomerado real. Suponhamos que exista um conglomerado real e um detectado, sendo que no segundo a importância é dada à população das áreas que não aparecem no conglomerado real e também àquelas cuja população não aparece no conglomerado detectado. Para estimar essas populações temos que primeiro definir algumas variáveis. Sejam $\hat{M}_+$ e $\hat{M}_-$ o número de áreas detectadas que não aparecem no conglomerado real e o número de áreas do conglomerado real que não aparecem no conglomerado detectado, respectivamente. Também $\hat{Z}_+$ e $\hat{Z}_-$ representam a região do mapa formada por áreas detectadas que não aparecem no conglomerado real e a região do mapa formada por áreas do conglomerado real que não aparecem no conglomerado detectado, respectivamente. Portanto, a população no conglomerado encontrado que não aparece no conglomerado real é dado por

$$\hat{n}_{\hat{z}_+} = \sum_{j=1}^{\hat{M}_+} n_j, \quad \forall s_j \in \hat{Z}_+,$$

e a população no conglomerado real que não aparece no conglomerado encontrado é dado por

$$\hat{n}_{\hat{z}_-} = \sum_{j=1}^{\hat{M}_-} n_j, \quad \forall s_j \in \hat{Z}_-.$$

Logo a proporção do erro de classificação da população do conglomerado real é definida por:

$$P_E = \frac{(\hat{n}_{\hat{z}_+} + \hat{n}_{\hat{z}_-})}{N},$$

onde N é a população total da região.

Essa medida representa o erro em percentual da população total da região espaço-tempo. Entretanto, o raciocínio apresenta a desvantagem de não informar se esse erro detectado foi para mais populações ou para menos populações que o conglomerado real.

5.4 Resultados do Estudo do Poder dos Testes Scan Circular, dMST e Doubly no Espaço-Tempo

Nesta seção serão apresentados os resultados obtidos através do critério de avaliação para o poder dos testes *Scan* Circular, *dMST* e *Doubly* no espaço-tempo fundamentados nos modelos alternativos de conglomerados descritos na seção 5.2. Consideraremos a verossimilhança de Bernoulli e Poisson para cada método e também o cenário cilíndrico e arbitrário. Todos os resultados são para 10.000 simulações.

Optamos, neste estudo, pelo tipo de estrutura de vizinhança da Tabela 4.5 para os métodos *dMST* e *Doubly* espaço-tempo na aplicação de dados reais e simulados, pois acreditamos que fornecerá mais informação ao conglomerado identificado.

Pela Tabela 5.4 calculamos as estimativas do poder de detecção do *Scan*, *dMST* e *Doubly* para o cenário cilíndrico, considerando a verossimilhança de Bernoulli, que são: 100%; 96,26% e 99,86%, respectivamente. No cenário arbitrário, o poder dos testes foi: 100%; 99,92% e 97,10%, respectivamente.

Tabela 5.4: Contagem do número de simulações, em 10.000, nas quais a hipótese nula foi rejeitada considerando a verossimilhança Binomial para os cenários cilíndrico e arbitrário.

Método	**Poder do Teste**	**Cenário Cilíndrico**	**Cenário Arbitrário**
Scan-Binomial	p-valor $< 0,05$	10.000	10.000
dMST-Binomial	p-valor $< 0,05$	9.626	9.992
Doubly-Binomial	p-valor $< 0,05$	9.986	9.710

Através da Tabela 5.5 calculamos as estimativas do poder de detecção desses mesmos métodos, porém utilizando a verossimilhança de Poisson. Para o cenário cilíndrico encontramos um poder de 100%, 96,26% e 99,86% no método *Scan*, *dMST* e *Doubly*, respectivamente. Para o cenário arbitrário tem-se um poder de 100%, 99,92% e 97,10%, respectivamente.

Tabela 5.5: Contagem do número de simulações, em 10.000, nas quais a hipótese nula foi rejeitada considerando a verossimilhança Poisson para os cenários cilíndrico e arbitrário.

Método	**Poder do Teste**	**Cenário Cilíndrico**	**Cenário Arbitrário**
Scan-Poisson	p-valor$< 0,05$	10.000	10.000
dMST-Poisson	p-valor $< 0,05$	9.626	9.992
Doubly-Poisson	p-valor $< 0,05$	9.986	9.710

Os três métodos estudados apresentam um ótimo desempenho em relação ao poder de detecção ($P_{H_1}(\lambda) > 0,96$) tanto no cenário cilíndrico quanto no arbitrário. Vale ressaltar que o método *Scan* é o que tem maior poder entre eles. O método *dMST* tem maior poder no cenário arbitrário do que no cilíndrico. Já o método *Doubly* tem maior poder no cenário cilíndrico do que no arbitrário. Tais afirmações são semelhantes em ambos os modelos com verossimilhança de Bernoulli e Poisson.

Apresentaremos os resultados de outros fatores que devem ser considerados no poder do teste primeiro para a verossimilhança de Bernoulli e depois para a Poisson.

Modelo Bernoulli

A Figura 5.5 e Tabela 5.6 mostram uma análise simultânea dos métodos, considerando a verossimilhança Bernoulli e o cenário cilíndrico, levando-se em conta o comprimento do tempo encontrado pelo conglomerado, seu tamanho identificado e a contagem da interseção com o conglomerado real para cada método. Como o comprimento do tempo total dos dados é 19 anos, padronizamos a escala de: 0-20 no eixo x que representa o comprimento do tempo e o eixo y que representa o número de simulações para os gráficos da distribuição do comprimento do tempo do conglomerado identificado. Para os gráficos da distribuição do tamanho do conglomerado identificado padronizamos a escala de: 0-100 ou 0-60 no eixo x que representa o tamanho do conglomerado identificado no método *Scan* e nos métodos *dMST* e *Doubly*, respectivamente. Também padronizamos a escala de: 0-35 no eixo x que representa a interseção entre o conglomerado encontrado e o conglomerado real para os gráficos da distribuição do tamanho da contagem da interseção, pois o conglomerado real cilíndrico tem 35 sub-áreas associadas a um determinado comprimento de 5 anos, onde para cada ano tem-se as mesmas 7 sub-áreas.

Pela análise cilíndrica simultânea, o método *Scan* apresentou um excelente desempenho. Mais de 70% das simulações desse método identificou o comprimento do tempo exato de 5 anos do conglomerado real e também identificou um tamanho de 30 ou 35 áreas para o conglomerado detectado. Além disso, dessas áreas encontradas, 30 ou 35 pertencem ao conglomerado real, lembrando-se, ainda, de que o mesmo tem 35 áreas associadas a um comprimento de 5 anos.

O método *dMST* identificou um comprimento de tempo entre 7 a 9 anos na maior parte das simulações, consequentemente, o conglomerado encontrado apresenta dimensões elevadas, concentrado próximo do limite de tamanho máximo (60). A distribuição da interseção para o

Tabela 5.6: Estatística Descritiva das 10.000 simulações para cada método avaliando o comprimento do tempo, seu tamanho identificado e a interseção com o conglomerado real de geometria cilíndrica e utilizando o modelo de verossimilhança de Bernoulli.

Distribuição do Comprimento do Tempo do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	5	5	5	5,16	5	9	0,54
dMST-Binom	5	7	8	8,46	9	19	1,64
Doubly-Binom	4	8	9	9,41	11	18	1,73
Distribuição do Tamanho do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	10	30	35	32,83	35	63	5,62
dMST-Binom	25	56	58	56,74	59	59	3,11
Doubly-Binom	9	41	48	47,15	55	59	9,01
Distribuição do Tamanho da Interseção							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	10	30	30	31,56	35	35	3,99
dMST-Binom	0	21	24	23,32	25	32	2,84
Doubly-Binom	5	22	24	23,88	26	33	2,81

dMST foi bem regular com média inferior ao tamanho real do conglomerado (média = 23,32). Apenas uma simulação detectou um conglomerado sem nenhuma área que pertencesse ao conglomerado real nesse método.

O método *Doubly* identificou um comprimento de tempo entre 8 a 10 anos na maior parte das simulações, porém a distribuição do tamanho do conglomerado encontrado foi bem irregular e superestima o tamanho real. A distribuição da interseção para o *Doubly* evidencia que esse método detecta parcialmente o conglomerado real.

Além dessas análises, foi proposta uma outra mais criteriosa na avaliação do poder dos testes. O conglomerado encontrado só seria aceito se coincidisse, pelo menos, em parte com o conglomerado real. A definição dessa análise permite que se verifique, experimentalmente, qual é a proporção ideal da interseção entre os dois conglomerados para que o poder do teste seja mais útil. A primeira variante dessa análise é fundamentada na contagem do número de sub-áreas em comum do conglomerado real e do conglomerado encontrado. A segunda é fundamentada na contagem do número de sub-áreas ponderada pela razão da população de cada sub-área. A terceira é baseada no erro percentual da população total da região espaço-tempo. Assim, através dessas medidas, temos um melhor entendimento de cada método. A Figura 5.6 e Tabela 5.7 mostram essa análise simultânea das proporções (mencionadas na seção 5.3) dos métodos considerando o cenário cilíndrico. Padronizamos a escala de: 0-1 no eixo x que representa a proporção da contagem de área, a proporção da contagem de população e a proporção de erro na classificação das áreas ponderados por suas populações do conglomerado identificado nas linhas 1,2 e 3 respectivamente da Figura 5.6 e o eixo y que representa o número de simulações para todos os gráficos.

Pela análise cilíndrica simultânea das proporções (Tabela 5.8), o método *Scan* apresentou

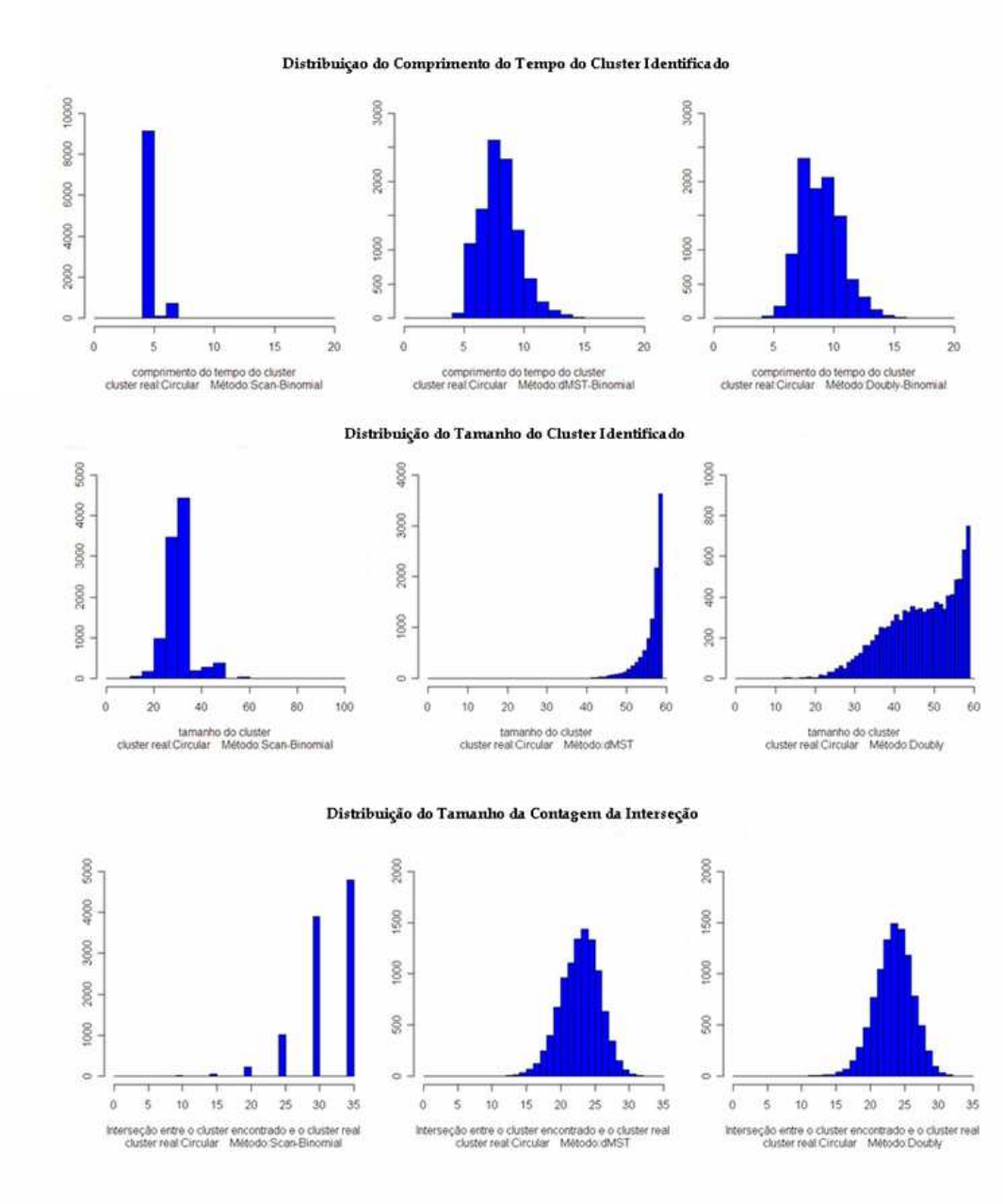

Figura 5.5: Análise Simultânea das Simulações para os métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Bernoulli e o cenário cilíndrico.

um excelente desempenho. Verificou-se que o *Scan* detectou, corretamente, 85% a 100% das áreas do conglomerado real com poder de 0,8, aproximadamente. Percebe-se que a distribuição geográfica da população é heterogênea, pois o método *Scan* detectou uma proporção de 97% a 100% da população do conglomerado real (P_P) com poder de 0,8, aproximadamente. Além disso, o *Scan* errou (P_E) 0% a 0,0035% da população total da região com poder de, aproximadamente, 0,8 na tentativa de encontrar o conglomerado real. O método *dMST* identificou, corretamente, 38% a 45% das áreas do conglomerado real com poder de 0,9, aproximadamente.

Tabela 5.7: Estatística Descritiva das 10.000 simulações para cada método avaliando a proporção de áreas, a proporção de população e a proporção de erro na classificação das áreas ponderadas por suas populações do conglomerado real com geometria cilíndrica e utilizando o modelo de verossimilhança de Bernoulli.

Distribuição da Proporção de Áreas do Conglomerado Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	0,285	0,857	0,857	0,883	1	1	0,122
dMST-Binom	0	0,379	0,411	0,411	0,440	0,6	0,048
Doubly-Binom	0,142	0,448	0,5	0,512	0,571	0,805	0,085
Distribuição da Proporção da População do Congl. Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	0,537	0,974	0,979	0,954	1	1	0,082
dMST-Binom	0	0,675	0,72	0,714	0,758	0,892	0,061
Doubly-Binom	0,295	0,700	0,762	0,76	0,825	0,957	0,084
Distribuição da Proporção de Erro da População do Congl.							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	0	0	0,0028	0,008	0,0035	0,121	0,016
dMST-Binom	0,025	0,055	0,066	0,068	0,079	0,171	0,017
Doubly-Binom	0,008	0,040	0,053	0,055	0,068	0,158	0,020

Considerando a segunda variante (P_P) para o método *dMST*, a proporção foi de 67% a 75% com poder de 0,9, aproximadamente. O *dMST* errou 0,055% a 0,08% da população total da região com poder de, aproximadamente, 0,8 na tentativa de encontrar o conglomerado real. O método *Doubly* identificou, corretamente, 45% a 57% das áreas do conglomerado real com poder de 0,7, aproximadamente. Para a proporção da população do conglomerado real (P_P), esse método detectou 76% a 82,6% com poder 0,7, aproximadamente. O método *Doubly* errou 0,04% a 0,069% da população total da região com poder 0,75, aproximadamente, na tentativa de encontrar o conglomerado real.

A Figura 5.7 e tabela 5.9 mostram o mesmo tipo de análise simultânea dos métodos, porém considerando agora o cenário arbitrário. Para os gráficos da distribuição do tamanho do conglomerado identificado, apenas uma modificação foi feita. Padronizamos a escala de: 0-150 ou 0-60 no eixo x que representa o tamanho do conglomerado identificado no método *Scan* e nos métodos *dMST* e *Doubly*, respectivamente. Também modificamos a padronização da escala de: 0-40 no eixo x que representa a interseção entre o conglomerado encontrado e o conglomerado real para os gráficos da distribuição do tamanho da contagem da interseção, pois o conglomerado real arbitrário tem 39 sub-áreas associadas a um determinado comprimento de 7 anos.

Pela análise arbitrária simultânea, verificou-se que o método *Scan* apresentou um baixo desempenho. Aproximadamente 60% das simulações desse método identificou um conglomerado de 2 anos de comprimento de tempo e com 2 áreas de tamanho. Além disso, dessas áreas encontradas, 2 pertenciam ao conglomerado real, lembrando-se de que o mesmo tem um comprimento de tempo de 7 anos e possui 39 áreas. O método *dMST* apresentou um bom

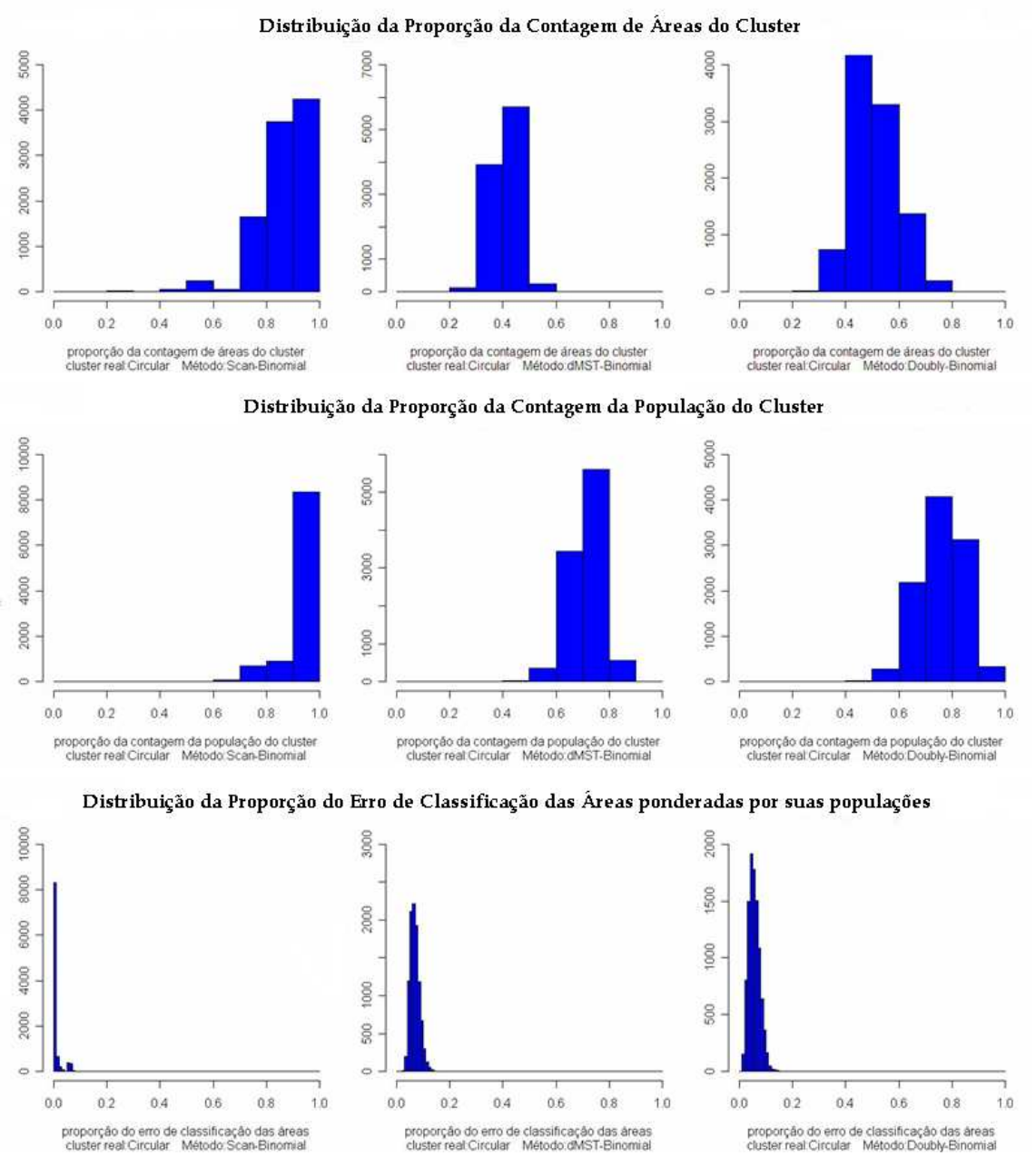

Figura 5.6: Análise Simultânea das Proporções das Simulações nos métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Bernoulli e o cenário cilíndrico.

desempenho. Identificou um comprimento de tempo entre 9 a 11 anos em 50% das simulações e em grande parte das simulações, o conglomerado encontrado apresentou dimensões elevadas, concentrado próximo do limite de tamanho máximo (60). A distribuição da interseção para o *dMST* foi bem regular com média inferior ao tamanho real do conglomerado (média = 25,96). O método *Doubly* identificou um comprimento de tempo entre 6 a 10 anos em 70% das simulações, logo a distribuição do tamanho do conglomerado encontrado foi bem irregular. A distribuição da interseção para o *Doubly* foi quase regular com média inferior ao tamanho

Tabela 5.8: Análise Cilíndrica Simultânea das Proporções dos Métodos de detecção de conglomerado espaço-tempo aplicado no banco de dados simulado para o modelo Bernoulli.

Método	Proporções Ideais	Poder
Scan Circular	P_A=85% a 100%	0,8
	P_P=97% a 100%	0,8
	P_E=0% a 0,0035%	0,8
dMST	P_A=38% a 45%	0,9
	P_P=67% a 75%	0,9
	P_E=0,055% a 0,08%	0,8
Doubly	P_A=45% a 57%	0,7
	P_P=76% a 82,6%	0,7
	P_E=0,04% a 0,069%	0,75

Tabela 5.9: Estatística Descritiva das 10.000 simulações para cada método avaliando o comprimento do tempo, seu tamanho identificado e a interseção com conglomerado real de geometria arbitrária e utilizando o modelo de verossimilhança de Bernoulli.

Distribuição do Comprimento do Tempo do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	1	2	2	3,253	5	9	1,94
dMST-Binom	5	9	10	10,12	11	19	2,275
Doubly-Binom	2	6	8	7,958	10	18	2,458
Distribuição do Tamanho do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	1	2	2	17,13	16	153	27,783
dMST-Binom	13	55	58	55,77	59	59	4,641
Doubly-Binom	2	27	37	35,96	46	59	13,519
Distribuição do Tamanho da Interseção							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	1	2	2	5,712	7	24	6,146
dMST-Binom	8	24	26	25,96	28	35	2,959
Doubly-Binom	2	17	21	20,13	25	36	6,349

real do conglomerado (média = 20,13).

Além dessas análises, a Figura 5.8 e Tabela 5.10 mostram uma análise simultânea das proporções (citadas na Seção 5.3) dos métodos considerando o cenário arbitrário agora. Utilizamos a mesma padronização de escala feita para a análise simultânea cilíndrica.

Pela análise simultânea arbitrária das proporções (Tabela 5.11), o método *Scan* apresentou um baixo desempenho. Verificou-se que o *Scan* detectou, corretamente, 5,12% das áreas do conglomerado real com poder de 0,6, aproximadamente. Para a segunda variante (P_P), o método *Scan* detectou, corretamente, 6,56% da população do conglomerado real com poder de 0,6, aproximadamente. Além disso, o método *Scan* foi o que apresentou maior erro nos métodos estudados no cenário arbitrário. Esse errou 0,0628% a 0,0702% da população total da região com poder de, aproximadamente, 0,75 na tentativa de encontrar o conglomerado

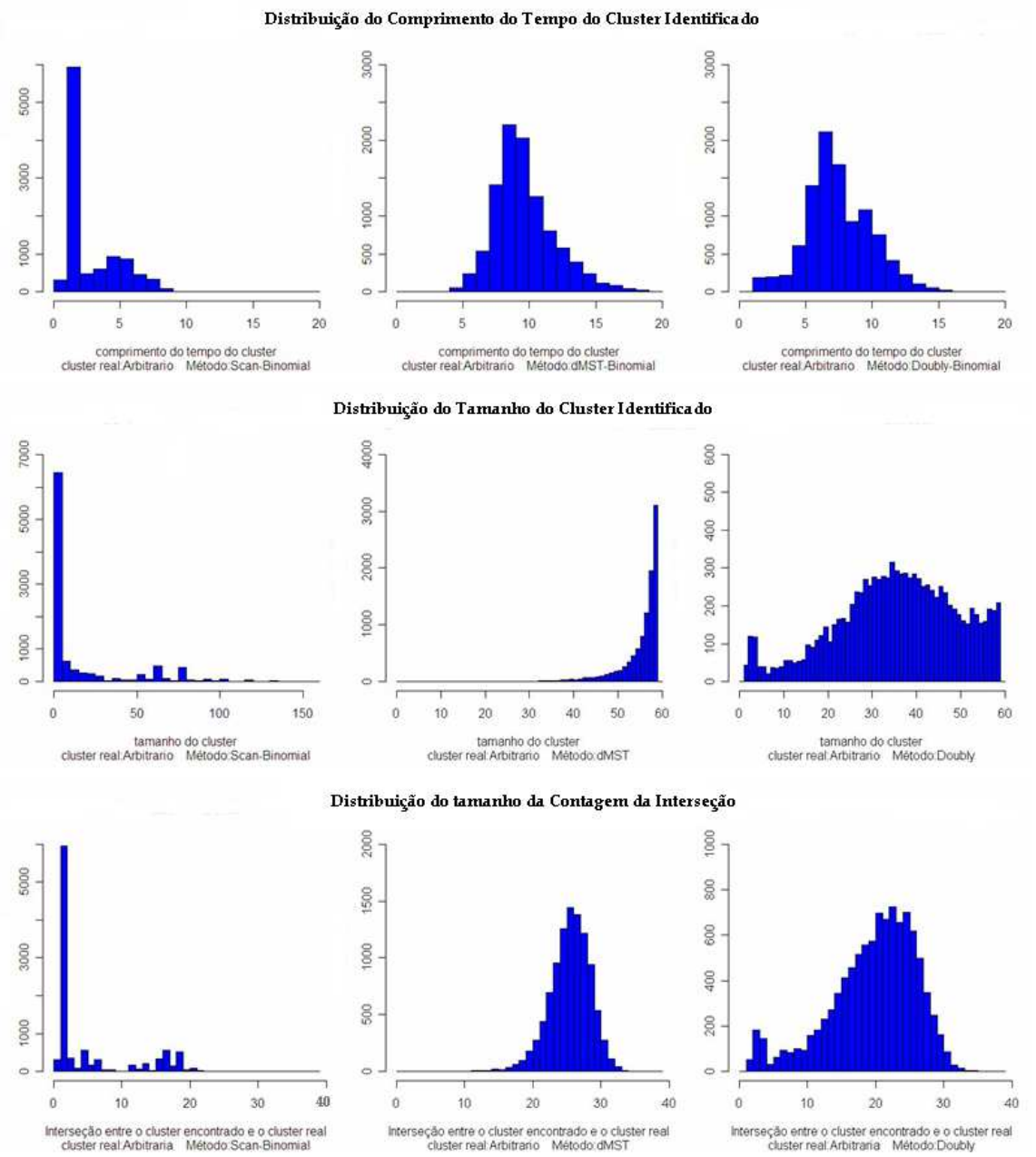

Figura 5.7: Análise Simultânea das Simulações para os métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Bernoulli e o cenário arbitrário.

real. O método *dMST* identificou, corretamente, 43% a 50% das áreas do conglomerado real com poder de 0,8, aproximadamente. Para a segunda variante (P_P), método *dMST* detectou, corretamente, 57% a 75% da população do conglomerado real com poder de 0,9, aproximadamente. O *dMST* errou 0,036% a 0,051% da população total da região com poder de 0,8, aproximadamente, na tentativa de encontrar o conglomerado real. O método *Doubly* identificou, corretamente, 41% a 56% das áreas do conglomerado real com poder de 0,6, aproximadamente. Para a segunda variante (P_P), o método *Doubly* detectou corretamente 59,95% a 75% da população do conglomerado real com poder de 0,6, aproximadamente. O

Tabela 5.10: Estatística Descritiva das 10.000 simulações para cada método avaliando a proporção de áreas, a proporção da população e a proporção de erro na classificação das áreas ponderadas por suas populações do conglomerado real com geometria arbitrária e utilizando o modelo de verossimilhança de Bernoulli.

Distribuição da Proporção de Áreas do Conglomerado Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	0,025	0,05128	0,0512	0,111	0,175	0,4103	0,094
dMST-Binom	0,156	0,431	0,465	0,466	0,5	0,743	0,056
Doubly-Binom	0,051	0,41	0,487	0,466	0,564	0,794	0,136
Distribuição da Proporção da População do Congl. Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	0,0057	0,0656	0,0656	0,132	0,187	0,398	0,106
dMST-Binom	0,136	0,574	0,631	0,621	0,679	0,848	0,084
Doubly-Binom	0,063	0,417	0,599	0,547	0,7005	0,935	0,199
Distribuição da Proporção de Erro da População do Congl.							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Binom	0,045	0,06284	0,06284	0,0678	0,07019	0,1675	0,011
dMST-Binom	0,014	0,03686	0,04349	0,04464	0,05156	0,097	0,0107
Doubly-Binom	0,008	0,0326	0,0435	0,0453	0,0565	0,1478	0,016

método *Doubly* errou 0,032% a 0,056% da população total da região com poder de 0,75, aproximadamente, na tentativa de encontrar o conglomerado real. A seguir, mostraremos os resultados do cenário cilíndrico e arbitrário considerando a verossimilhança de Poisson para os modelos abordados.

Modelo Poisson

A Figura 5.9 e Tabela 5.12 mostram o mesmo tipo de análise simultânea dos métodos feita anteriormente no cenário cilíndrico considerando a verossimilhança Poisson que também apresenta os mesmos resultados encontrados na análise simultânea dos métodos com o modelo de Bernoulli. Por exemplo, os métodos *dMST* e *Doubly* apresentam resultados idênticos em todas as estatísticas descritivas da Tabela 5.6 e Tabela 5.12, com exceção apenas do desvio-padrão que tem uma ínfima diferença. O método *Scan* apresenta diferença apenas nos valores discrepantes (mínimo e máximo) dessas tabelas. Em virtude disso, a relevância está na diferença entre as verossimilhanças dos métodos do que os comentários das análises do modelo Poisson nos dois cenários, uma vez que os comentários serão os mesmos do modelo Bernoulli.

A variabilidade dos resultados é maior no método *Scan* usando a verossimilhança de Poisson do que a de Bernoulli devido aos valores discrepantes. Entretanto, temos que o número de simulações que identifica o conglomerado real, por inteiro, no modelo poisson é

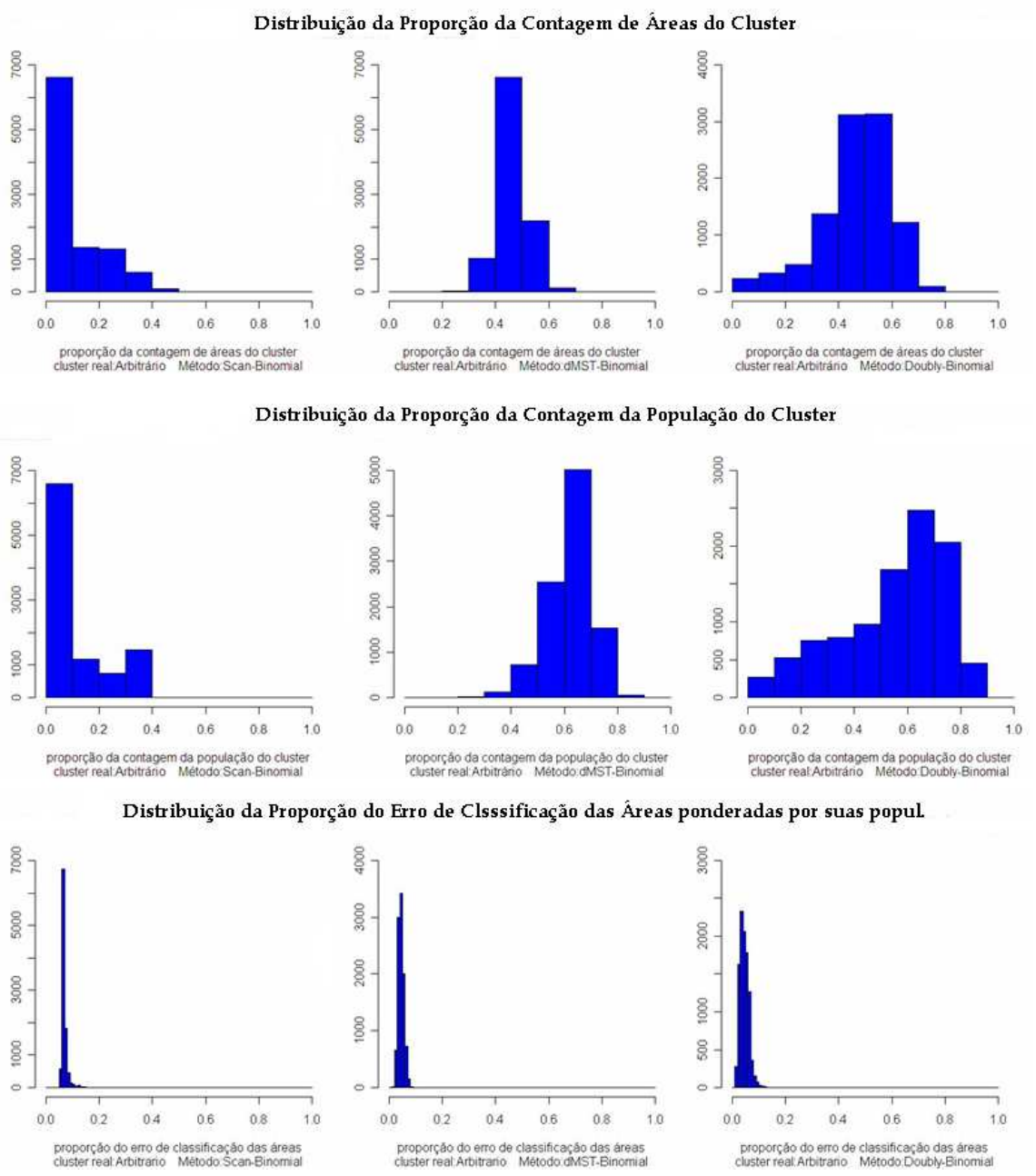

Figura 5.8: Análise Simultânea das Proporções das Simulações nos métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Bernoulli e o cenário arbitrário.

maior do que no Bernoulli, ou seja, temos um pouco mais de precisão no modelo Poisson, o que comprova, no fato, que a mediana do tamanho da interseção com o conglomerado real do modelo Bernoulli é menor do que o modelo Poisson. Já nos métodos *dMST* e *Doubly* pode-se dizer apenas, visualmente, nos gráficos da interseção que o modelo Poisson tem uma precisão ligeiramente superior do que no Bernoulli, mas essa diferença é difícil de ser identificada nos gráficos.

Além dessas análises, a Figura 5.10 e Tabela 5.13 mostram o mesmo tipo de análise

Tabela 5.11: Análise Simultânea Arbitrária das Proporções dos Métodos de detecção de conglomerado espaço-tempo aplicado no banco de dados simulado para o modelo Bernoulli.

Método	Proporções Ideais	Poder
Scan Circular	P_A=5,12%	0,6
	P_P=6,56%	0,6
	P_E=0,0628% a 0,0702%	0,75
dMST	P_A=43% a 50%	0,8
	P_P=57% a 75%	0,9
	P_E=0,036% a 0,051%	0,8
Doubly	P_A=41% a 56%	0,6
	P_P=59,95% a 75%	0,6
	P_E=0,032% a 0,056%	0,75

simultânea das proporções feitas anteriormente no cenário cilíndrico, mas considerando a verossimilhança Poisson e, novamente, apresenta os mesmos resultados encontrados na análise simultânea das proporções dos métodos, considerando a verossimilhança de Bernoulli.

No método *Scan*, a proporção de erro no modelo Poisson é maior que no Bernoulli. Apenas com um olhar minucioso pode-se distinguir a diferença mínima entre os dois modelos nos métodos *dMST* e *Doubly* considerando as análises das proporções. Destaca-se o melhor desempenho no modelo Poisson para esses métodos.

Para o cenário arbitrário, utilizando o modelo Poisson, a Figura 5.11 e Tabela 5.14 mostram as análises simultâneas dos métodos *Scan*, *dMST* e *Doubly*. Pode-se dizer que nesse cenário os métodos *dMST* e *Doubly* apresentam resultados idênticos em todas as estatísticas descritivas da Tabela 5.9 (modelo Bernoulli) e Tabela 5.14 (modelo Poisson), com exceção do desvio-padrão que tem uma ínfima diferença.

A variabilidade dos resultados é maior no método *Scan* usando a verossimilhança de Poisson do que a de Bernoulli no cenário arbitrário, e o modelo Poisson tem uma precisão melhor para identificar o conglomerado real do que o modelo Bernoulli nesse cenário. Isso se comprova no fato de que a mediana do tamanho do conglomerado detectado no modelo Bernoulli foi menor do que a mediana do modelo Poisson, e, também, a mediana da interseção com o conglomerado real foi menor no modelo Bernoulli do que o modelo Poisson. Já nos métodos *dMST* e *Doubly* pode-se dizer que, apenas, visualmente no gráfico da interseção o modelo Poisson tem um pouco mais de precisão do que no Bernoulli.

A Figura 5.12 e Tabela 5.15 mostram o mesmo tipo de análise simultânea das proporções feita anteriormente no cenário arbitrário considerando a verossimilhança Poisson. Esta análise apresenta, praticamente, de novo, os mesmo resultados encontrados na análise simultânea das proporções dos métodos considerando a verossimilhança de Bernoulli.

No método *Scan* com modelo Poisson, reforça-se o fato interessante que aconteceu de haver precisão melhor nesse cenário. Isto se comprova no fato de que a mediana da proporção de áreas do conglomerado real no modelo Bernoulli foi menor do que a mediana do modelo Poisson, e também, a mediana da proporção da população do conglomerado real foi menor no

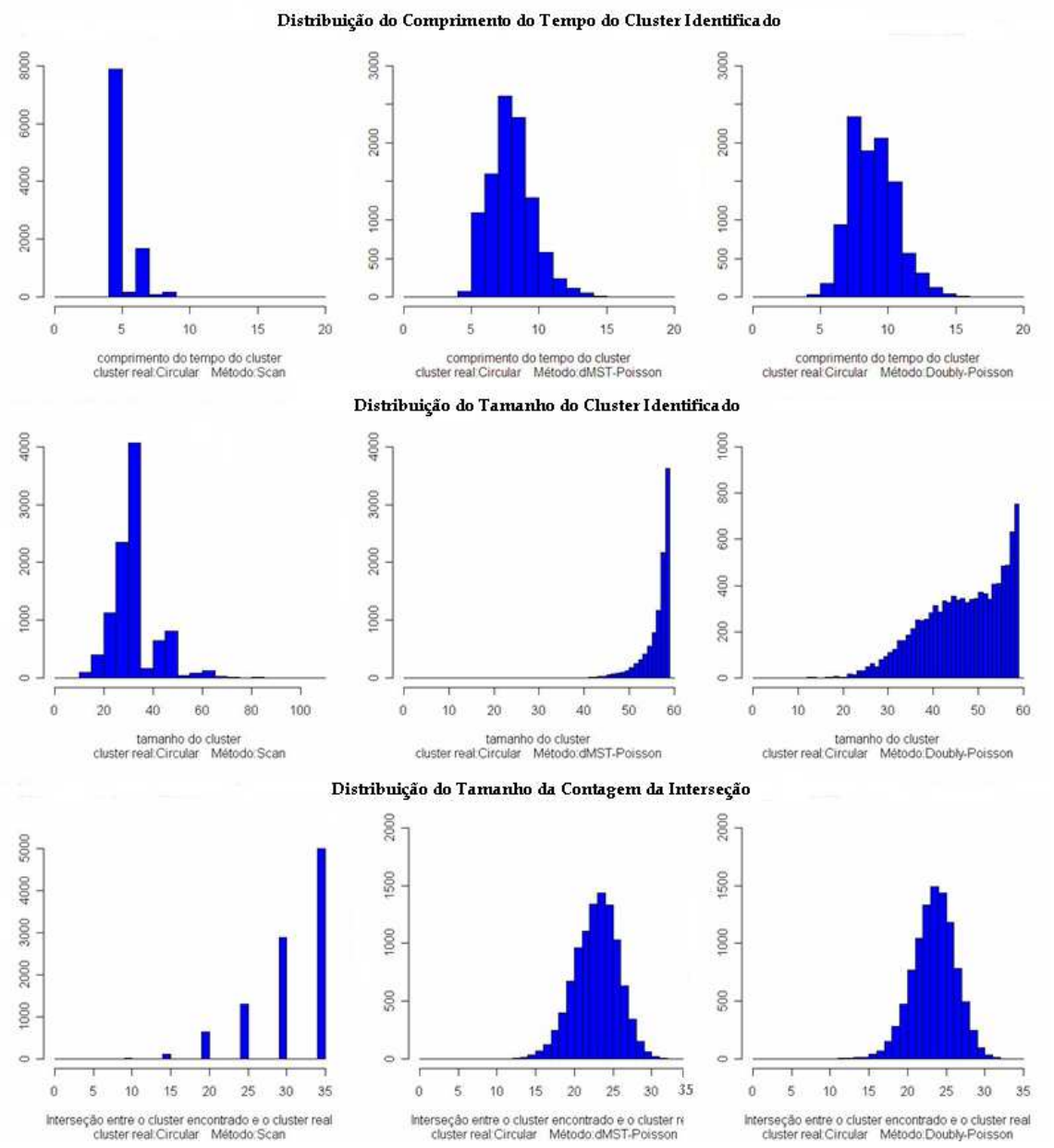

Figura 5.9: Análise Simultânea das Simulações para os métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Poisson e o cenário cilíndrico.

modelo Bernoulli do que o modelo Poisson. Quanto à proporção de erro na classificação das áreas no modelo Poisson, pode-se dizer que é maior que no Bernoulli. Reforça-se, também, que apenas com um olhar minucioso pode-se distinguir a diferença mínima entre o modelo Poisson e o modelo Bernoulli nos métodos *dMST* e *Doubly* nas análises das proporções para o cenário arbitrário, destacando-se melhor desempenho no modelo Poisson para estes métodos.

Tabela 5.12: Estatística Descritiva das 10.000 simulações para cada método avaliando o comprimento do tempo, seu tamanho identificado e a interseção com o conglomerado real de geometria cilíndrica e utilizando o modelo de verossimilhança de Poisson.

Distribuição do Comprimento do Tempo do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	5	5	5	5,447	5	9	0,918
dMST-Pois	5	7	8	8,465	9	19	1,641
Doubly-Pois	4	8	9	9,414	11	18	1,736
Distribuição do Tamanho do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	5	30	35	34,44	35	108	8,894
dMST-Pois	25	56	58	56,74	59	59	3,112
Doubly-Pois	9	41	48	47,15	55	59	9,019
Distribuição do Tamanho da Interseção							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	5	30	35	31	35	35	4,957
dMST-Pois	0	21	24	23,32	25	32	2,841
Doubly-Pois	5	22	24	23,88	26	33	2,813

5.5 Discussão

De maneira geral, pode-se afirmar que quanto a capacidade de detecção, o método *Scan* circular apresenta resultado superior aos métodos *dMST* e *Doubly* para o cenário cilíndrico e inferior no cenário arbitrário. O método *Doubly* apresentou um bom desempenho de detecção em ambos os cenários. A avaliação de desempenho das metodologias propostas de detecção em cenários simulados fornece medidas de sensibilidade quando os mesmos são aplicados a cenários reais. Nesse contexto, usando as mesmas quantidades de casos, por ano, na região do Novo México da Seção 5.2, aplicaremos no Capítulo 6 esses métodos para casos de câncer cerebral reais para facilitar na comparação com dados simulados.

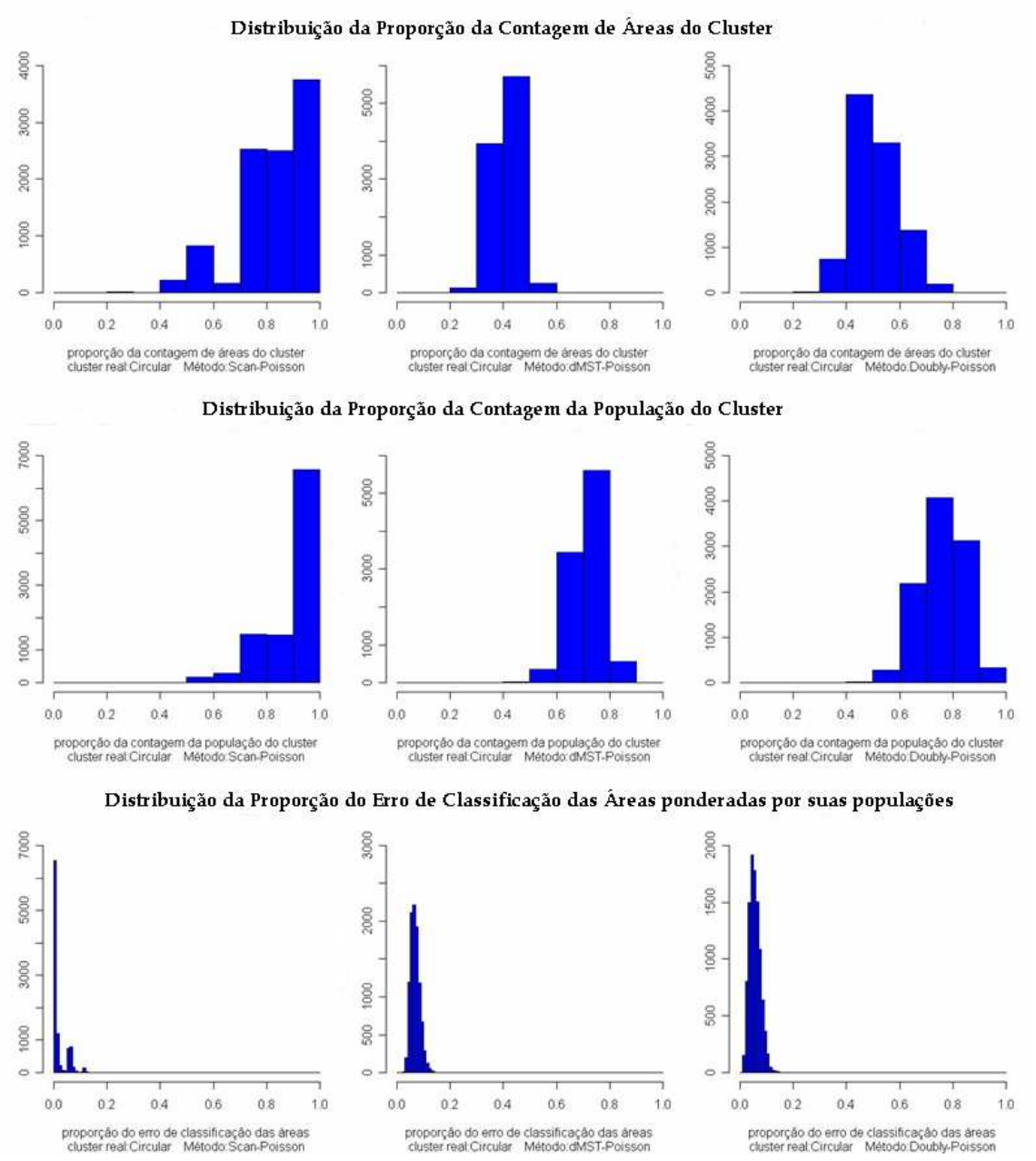

Figura 5.10: Análise Simultânea das Proporções das Simulações nos métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Poisson e o cenário cilíndrico.

Tabela 5.13: Estatística Descritiva das 10.000 simulações para cada método avaliando a proporção de áreas, a proporção da população e a proporção de erro na classificação das áreas ponderadas por suas populações do conglomerado real com geometria cilíndrica e utilizando o modelo de verossimilhança de Poisson.

Distribuição da Proporção de Áreas do Conglomerado Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	0,143	0,7143	0,8571	0,8375	1	1	0,156
dMST-Pois	0	0,379	0,4118	0,4114	0,4407	0,6	0,048
Doubly-Pois	0,142	0,4483	0,5	0,512	0,5714	0,8056	0,085
Distribuição da Proporção da População do Congl. Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	0,5029	0,8773	0,9798	0,9119	1	1	0,121
dMST-Pois	0	0,6756	0,72	0,714	0,7585	0,8926	0,061
Doubly-Pois	0,2956	0,7008	0,7626	0,7601	0,8259	0,9578	0,084
Distribuição da Proporção de Erro da População do Congl.							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad
Scan-Pois	0	0	0,0028	0,01721	0,01723	0,139	0,0269
dMST-Pois	0,025	0,05536	0,06658	0,06845	0,07904	0,1714	0,017
Doubly-Pois	0,0086	0,04026	0,05351	0,0556	0,06893	0,15860	0,0206

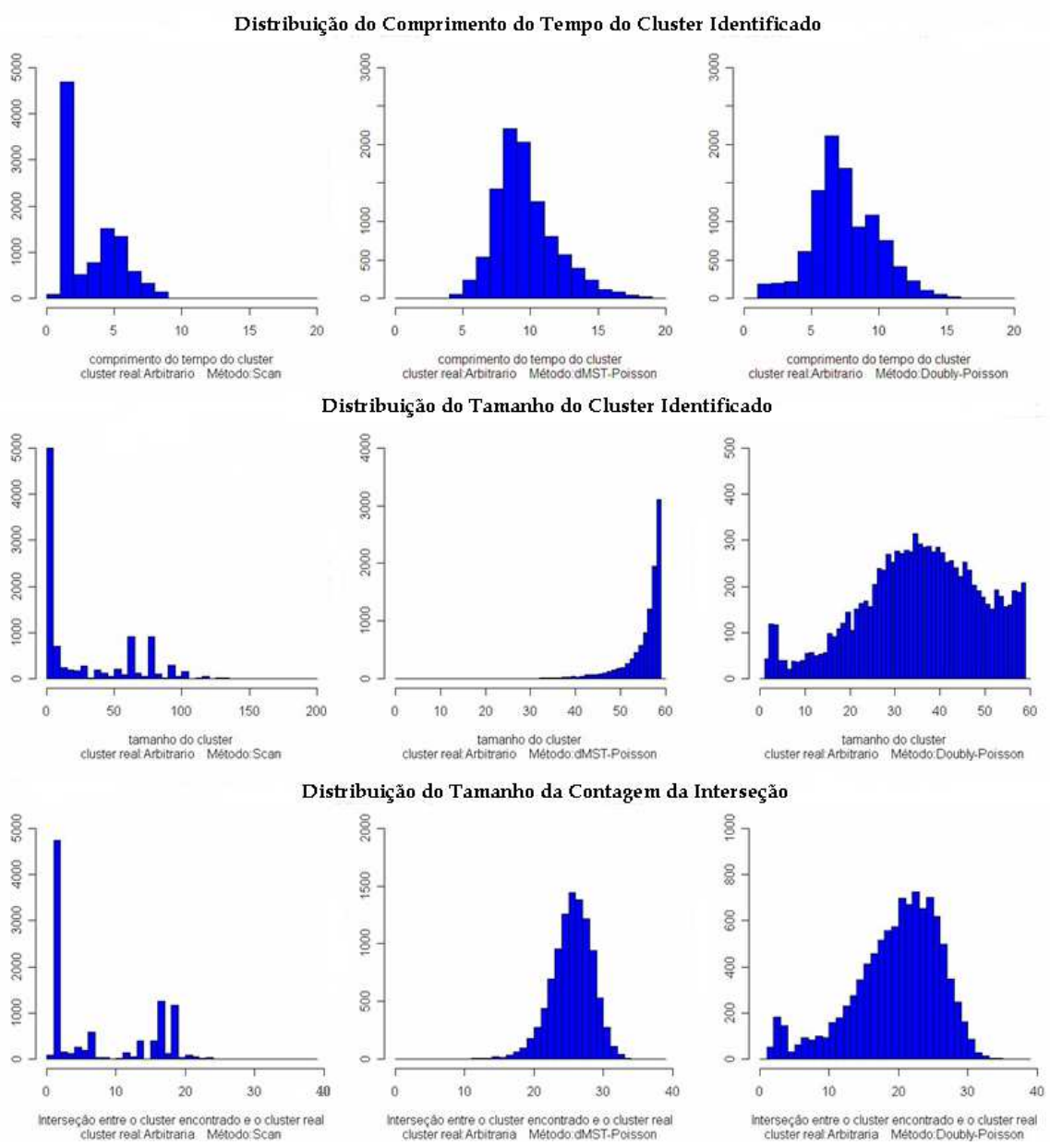

Figura 5.11: Análise Simultânea das Simulações para os métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Poisson e o cenário arbitrário.

Tabela 5.14: Estatística Descritiva das 10.000 simulações para cada método avaliando o comprimento do tempo, seu tamanho identificado e a interseção com o conglomerado real de geometria arbitrária e utilizando o modelo de verossimilhança de Poisson.

Distribuição do Comprimento do Tempo do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	1	2	3	3,783	5	9	2,017
dMST-Pois	5	9	10	10,12	11	19	2,275
Doubly-Pois	2	6	8	7,958	10	18	2,458
Distribuição do Tamanho do Conglomerado							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	1	2	6	28,14	65	162	34,101
dMST-Pois	13	55	58	55,77	59	59	4,642
Doubly-Pois	2	27	37	35,96	46	59	13,516
Distribuição do Tamanho da Interseção							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	1	2	4	8,286	17	24	7,251
dMST-Pois	8	24	26	25,96	28	35	2,959
Doubly-Pois	2	17	21	20,13	25	36	6,348

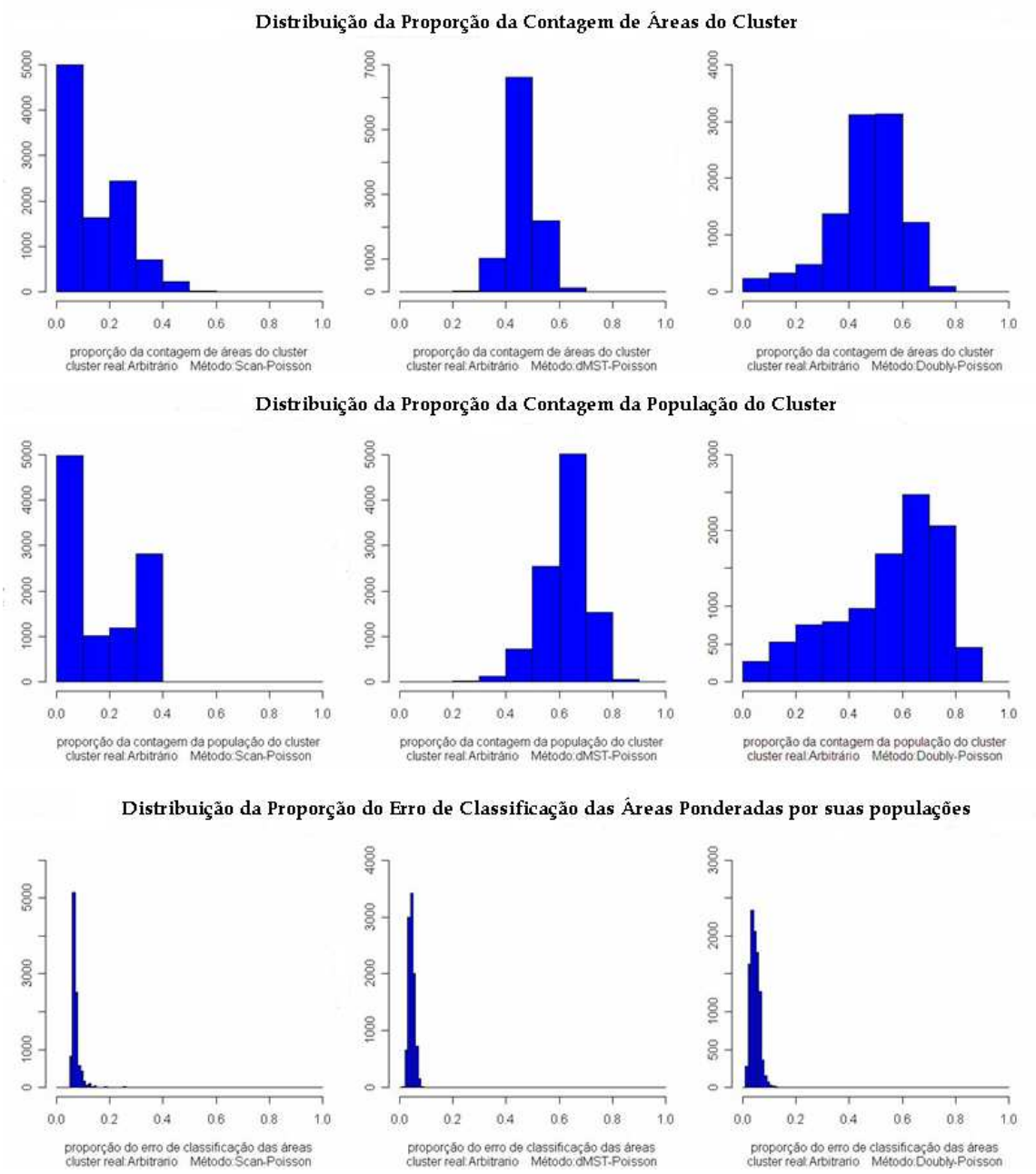

Figura 5.12: Análise Simultânea das Proporções das Simulações nos métodos Scan (coluna 1), dMST (coluna 2) e Doubly (coluna 3) considerando a verossimilhança Poisson e o cenário arbitrário.

Tabela 5.15: Estatística Descritiva das 10.000 simulações para cada método avaliando a proporção de áreas, a proporção da população e a proporção de erro na classificação das áreas ponderadas por suas populações do conglomerado real com geometria arbitrária e utilizando o modelo de verossimilhança de Poisson.

Distribuição da Proporção de Áreas do Conglomerado Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	0,0256	0,0512	0,1026	0,1457	0,2436	0,5227	0,107
dMST-Pois	0,1569	0,431	0,4655	0,4668	0,5	0,7436	0,056
Doubly-Pois	0,0512	0,41	0,4872	0,4669	0,5641	0,7949	0,136
Distribuição da Proporção da População do Congl. Real							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des-Pad.
Scan-Pois	0,0057	0,06561	0,1223	0,1779	0,3083	0,4543	0,124
dMST-Pois	0,136	0,5746	0,6319	0,6217	0,6792	0,8487	0,084
Doubly-Pois	0,0638	0,4178	0,5996	0,547	0,7006	0,9354	0,199
Distribuição da Proporção de Erro da População do Congl.							
Método	Mínimo	1quartil	Mediana	Média	3quartil	Máximo	Des.-Pad.
Scan-Pois	0,045	0,0628	0,06284	0,0711	0,07642	0,2541	0,017
dMST-Pois	0,014	0,0368	0,04351	0,04464	0,05156	0,0969	0,0107
Doubly-Pois	0,0084	0,0326	0,04352	0,04536	0,05653	0,1478	0,016

Capítulo 6

Aplicação dos Métodos de Detecção de Conglomerados Espaço-Tempo

6.1 Introdução

No capítulo anterior, foi demonstrado o poder dos testes quando o conglomerado real, a priori, tem a forma cilíndrica ou arbitrária e, também, em que tipo de cenário cada método é o melhor a ser empregado. Aplicaremos esses métodos de detecção para conjuntos de dados epidemiológicos reais dos quais não sabemos o formato no conglomerado real. Entretanto, existem outras áreas potenciais de aplicação dessas técnicas.

O primeiro conjunto de dados epidemiológicos será definido pelos casos de câncer cerebral ocorridos nas pessoas que vivem nos condados do Novo México, Estados Unidos, quando, no período de estudo, o total de casos por ano, a população e as coordenadas geográficas de cada condado já foram descritas na Seção 5.2. Escolhemos esse conjunto de dados, por ter sido analisado anteriormente na literatura por Kulldorff[22] usando seu próprio método (*Scan* circular).

O segundo conjunto será definido pelos casos de dengue ocorridos nas pessoas que vivem na cidade de Vitória, Espírito Santo, quando, no período de estudo, o total de casos por mês, a população e as coordenadas geográficas de cada bairro foram descritas na Seção 1.1. Escolhemos esse conjunto de dados, pois a dengue é uma arbovirose que se tornou um grave problema de saúde pública no Brasil, assim como em outras regiões tropicais do mundo. É de transmissão essencialmente urbana, ambiente no qual encontram-se todos os fatores fundamentais para sua ocorrência: o homem, o vírus, o vetor e principalmente as condições políticas, econômicas e culturais que formam a estrutura que permite o estabelecimento da cadeia de transmissão (Marzochi[34]).

6.2 Novo México

Os dados de incidência de câncer cerebral em pessoas que vivem nos condados do Novo México foram obtidos através do programa SEER (em português: Vigilância, Epidemiologia e

Resultados Finais) do Instituto Nacional de Câncer do Novo México. De 1973 a 1991 tinham 1.175 casos dessa doença registrados.

Antes de realizar uma análise espaço-tempo, fizemos uma análise puramente espacial para os métodos estudados em cada ano do período. O único conglomerado, puramente, espacial de casos de câncer cerebral que apresentou um p-valor menor que 5% foi detectado no ano de 1987 para todos os métodos. Veja na Figura 6.1 o conglomerado detectado por cada método e a Tabela 6.1 com o p-valor de cada método neste ano.

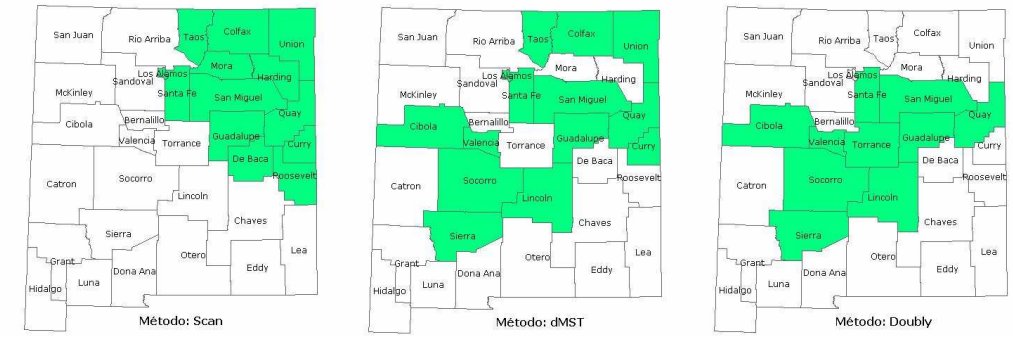

Figura 6.1: Conglomerado espacial de casos de câncer cerebral em pessoas que vivem nos condados do Novo México detectado por cada método no ano de 1987 com p-valor menor que 5%.

Tabela 6.1: P-valor do conglomerado encontrado para cada método de detecção espacial, no ano de 1987, considerando os casos de câncer cerebral de pessoas que vivem nos condados do Novo México.

Método	**p-valor**
Scan circular	0,034
dMST	0,022
Doubly	0,045

6.2.1 Modelo Bernoulli

O modelo Bernoulli é o mais natural e indicado para esse conjunto de dados, pois usamos o estudo de caso-controle. Temos, separadamente, a contagem de casos e não-casos em cada condado para cada ano.

As Figuras 6.2, 6.3 e 6.4 apresentam os resultados para os dados de câncer cerebral em Novo México durante os anos de 1973 a 1991 para os métodos de detecção estudados. O conglomerado de casos de câncer cerebral detectado pelo método *Scan* circular (Figura 6.4) abrange 40 condados no período de 1985 a 1989 e apresentou um p-valor significativo (1,1%), mostrando que o risco de uma pessoa ter câncer cerebral é maior nas áreas que pertencem ao conglomerado do que em outras. O método *Doubly* (Figura 6.3) obteve o maior conglomerado (59 condados no período de 1983 a 1991) e esse apresentou um p-valor significativo (0,6%), ou

seja, a incidência de casos de câncer não acontece por mero acaso como mostrou esse método. O método *dMST* (Figura 6.2) identificou um conglomerado de tamanho 56 no período de 1985 a 1991 com um p-valor não significativo (10%), demonstrando que os casos de câncer cerebral ocorrem de forma aleatória sobre toda a região. A Tabela 6.2 mostra os valores observados da estatística de teste para os dados de câncer cerebral nos condados do Novo México e o p-valor considerando 10.000 simulações sob hipótese nula para cada método.

Figura 6.2: Conglomerado de casos de câncer cerebral detectado entre os anos de 1985 a 1991 pelo método dMST espaço-tempo considerando as pessoas que vivem nos condados do Novo México .

Figura 6.3: Conglomerado de casos de câncer cerebral detectado entre os anos de 1983 a 1991 pelo método Doubly espaço-tempo considerando as pessoas que vivem nos condados do Novo México.

6.2.2 Modelo Poisson

O modelo Poisson, diante dos resultados do Capítulo 5, é o que nos dá a maior precisão para encontrar o conglomerado real e se aproxima bem do modelo Bernoulli. Temos a população total e essa estratificada por condado para cada ano. Também os casos de câncer cerebral para cada condado.

Como os conglomerados encontrados nos métodos *dMST* e *Doubly* são os mesmos do modelo Bernoulli, as Figuras 6.2 e 6.3 mostram os resultados do modelo Poisson respectivamente.

Figura 6.4: Conglomerado de casos de câncer cerebral detectado entre os anos de 1985 a 1989 pelo método Scan Circular espaço-tempo considerando as pessoas que vivem nos condados do Novo México.

Tabela 6.2: Tamanho, P-valor e a Estatística de Teste dos Métodos de detecção espaço-tempo aplicado aos casos de câncer cerebral considerando as pessoas que vivem nos condados do Novo México para o modelo Bernoulli e Poisson.

	Método	**Tamanho**	**Estatística de Teste**	**P-valor**
	Scan Circular	40	11,710082	0,011000
	dMST	56	49,799500	0,102690
Modelo Bernoulli	Doubly	59	44,418600	0,006099
	Scan Circular	35	8,867596	0,081000
	dMST	56	49,796567	0,102690
Modelo Poisson	Doubly	59	44,416193	0,005999

O método *dMST* identificou um conglomerado com p-valor não significativo (10%), o que implica casos de doença que ocorrem de forma aleatória. Para o método *Doubly* encontrou-se um conglomerado com p-valor significativo (0,5%), a distribuição de casos de câncer cerebral não acontece por mero acaso. O conglomerado detectado pelo método *Scan* circular (Figura 5.1) abrange 35 condados no período de 1985 a 1989. Observamos que houve um decréscimo de um condado no conglomerado espaço-tempo detectado nesse método considerando o modelo de Poisson comparado ao conglomerado do modelo Bernoulli. O conglomerado identificado pelo *Scan* circular apresentou um p-valor não significativo (8%), sendo que o risco de uma pessoa ter a doença é constante em toda região. A Tabela 6.2 mostra os valores observados da estatística de teste e o p-valor para cada método.

Apesar da discrepância em relação à irregularidade da geometria dos conglomerados detectados pelos métodos *dMST* e *Doubly* e também da variabilidade do p-valor dos três métodos

abordados, é evidente a formação de um conglomerado arbitrário com p-valor menor que 5% no ano 1986 a 1989 a partir da interseção das áreas encontradas por cada método. Com base nos resultados da análise puramente espacial do ano 1987 e na característica de interseção dos resultados, pode ser considerado como conglomerado final do processo de busca, as áreas: Los Alamos, Santa Fé, San Miguel, Cibola, Valência, Socorro, Torrance e Bernalillo.

6.3 Vitória, Espírito Santo

Como foi mencionado anteriormente na Seção 1.1, existe evidência de um conglomerado espaço-tempo nesse conjunto de dados, pois encontramos no modelo ajustado que as covariáveis coordenadas e tempo influenciam significativamente nos casos de dengue. Porém, apenas com a aplicação dos métodos de detecção temos a capacidade de localizar especificamente a região e o tempo em que esse conglomerado se encontra.

6.3.1 Modelo Bernoulli

As Figuras 6.5, 6.6 e 6.7 apresentam os resultados para os dados de casos de dengue em Vitória durante os meses de janeiro a dezembro de 2006 para os métodos de detecção estudados. O método *Scan* circular (Figura 6.7) obteve o maior conglomerado de casos de dengue (141 bairros no período de março a maio). Esse conglomerado detectado apresentou um p-valor significativo (0,1%), mostrando que o risco de uma pessoa ter dengue foi maior nas áreas que pertencem ao conglomerado do que em outras. O conglomerado detectado pelo método *Doubly* (Figura 6.5) abrangeu 29 bairros no período de fevereiro a maio e este apresentou um p-valor significativo (0,01%), ou seja, a incidência de casos de dengue não aconteceu por mero acaso neste método. O método *dMST* (Figura 6.6) identificou um conglomerado de tamanho 42 no período de fevereiro a maio com um p-valor significativo (0,01%), implicando que os casos de dengue ocorreram de forma não aleatória sobre toda a região. A Tabela 6.3 mostra os valores observados da estatística de teste para estes dados e o p-valor considerando 10.000 simulações sob hipótese nula para cada método.

6.3.2 Modelo Poisson

Como os conglomerados encontrados nos métodos *Scan* circular e *Doubly* são os mesmos do modelo Bernoulli, as Figuras 6.7 e 6.5 mostram os resultados do modelo Poisson respectivamente. O método *Scan* circular identificou um conglomerado com p-valor significativo (0,1%), isso implica que os casos de dengue ocorreram de forma não aleatória em toda a região. Para o método *Doubly* encontrou-se um conglomerado com p-valor significativo (0,01%), a distribuição de casos de dengue não acontece por mero acaso neste método. O conglomerado detectado pelo método *dMST* (Figura 6.8) abrange 42 bairros no período de fevereiro a maio. Observamos que houve um deslocamento no tempo de um bairro no conglomerado espaço-tempo detectado pelo método *dMST* considerando o modelo de Poisson comparado ao conglomerado do modelo Bernoulli, ou seja, o bairro Enseada do Suá encontrava-se no conglomerado

Figura 6.5: Conglomerado de casos de dengue detectado entre os meses de fevereiro a maio de 2006 pelo método *Doubly* espaço-tempo aplicado no banco de dados de Vitória.

do modelo Bernoulli no mês de abril e já para o conglomerado do modelo Poisson este bairro estava no mês de março. O conglomerado identificado pelo *dMST* apresentou um p-valor significativo (0,01%), sendo que o risco de uma pessoa ter a doença é maior nas áreas que pertencem ao conglomerado. A tabela 6.3 mostra os valores observados da estatística de teste e o p-valor para cada método.

É notório a super-estimação do método *Scan* circular nos casos de dengue em Vitória. A razão disto é a existência de um vasto maciço rochoso dentro da cidade que fez com que a distribuição dos casos tivessem uma forma geométrica arbitrária. Esse maciço rochoso encontra-se na parte central do conglomerado detectado pelo método *Scan* circular da Figura 6.7, motivo pelo qual não foi incluido como área ao conglomerado. Conseqüentemente, para o método *Scan* circular detectar o conglomerado real arbitrário, precisou aumentar o raio exageradamente incorporando áreas onde nenhum caso foi registrado.

Apesar da discrepância em relação à irregularidade da geometria dos conglomerados detectados pelos métodos *dMST* e *Doubly*, é evidente a formação de um conglomerado arbitrário

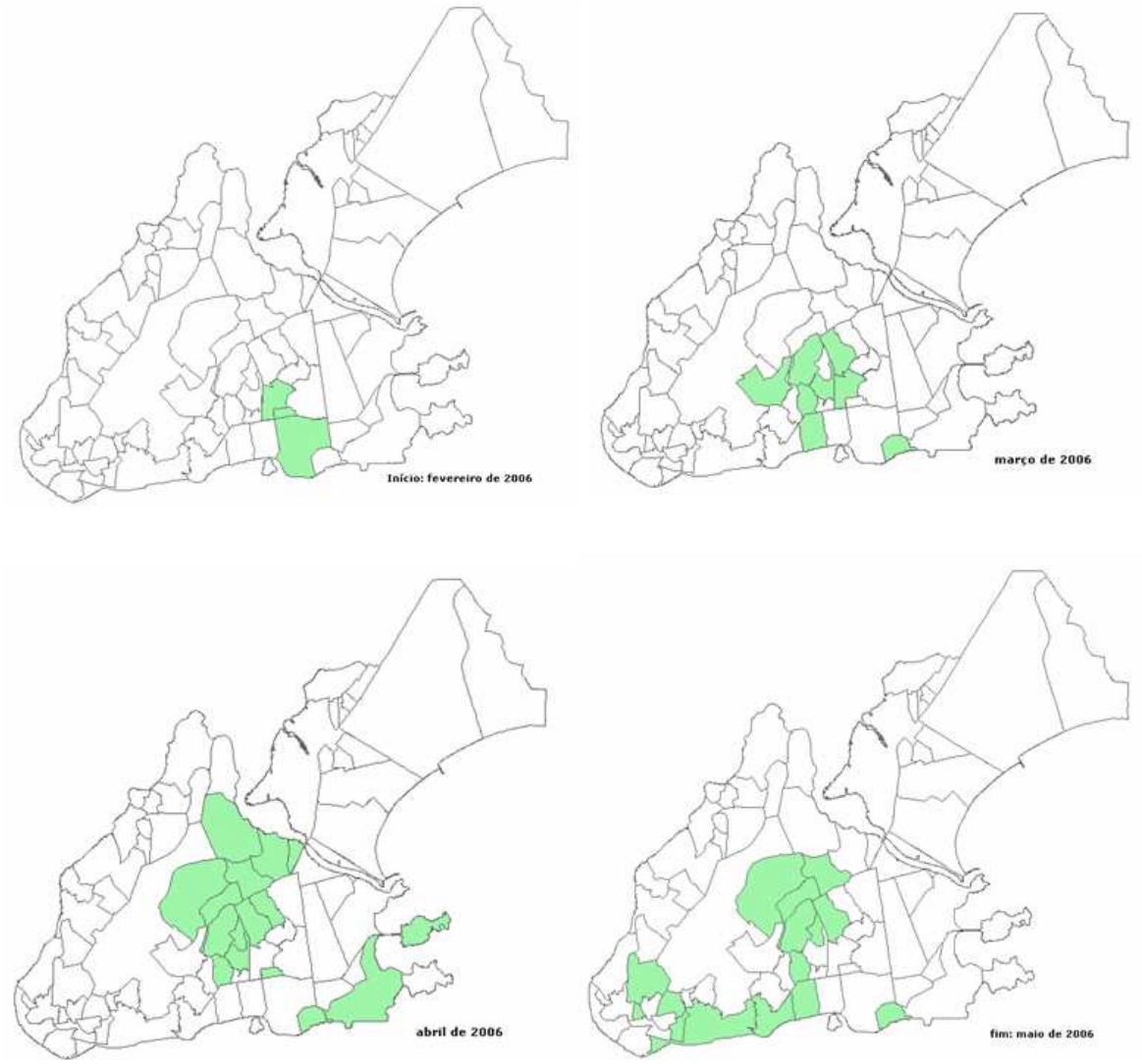

Figura 6.6: Conglomerado de casos de dengue detectado entre os meses de fevereiro a maio de 2006 pelo método *dMST* espaço-tempo aplicado no banco de dados de Vitória.

com p-valor menor que 5% nos meses de março a maio a partir da interseção das áreas encontradas por cada método. Com base na característica de interseção dos resultados, pode ser considerado como conglomerado final do processo de busca, as áreas: Forte São João, Ilha de Santa Maria, Jucutuquara, Bairro da Penha, Bairro de Lourdes, Bonfim, Consolação, Joana D'arc, Maruípe, Santa Cecília, Santa Martha, Santos Dumont, São Cristóvão e Tabuazeiro. Esse conglomerado final de casos de dengue foi detectado em um período de chuvas, fator crítico na reprodução e proliferação do agente transmisor.

6.4 Discussão

Utilizar apenas um dos métodos de detecção espaço-tempo estudados na aplicação de dados reais, poderia ocasionar, apenas, uma detecção de sub-áreas associada ao tempo de ocorrência

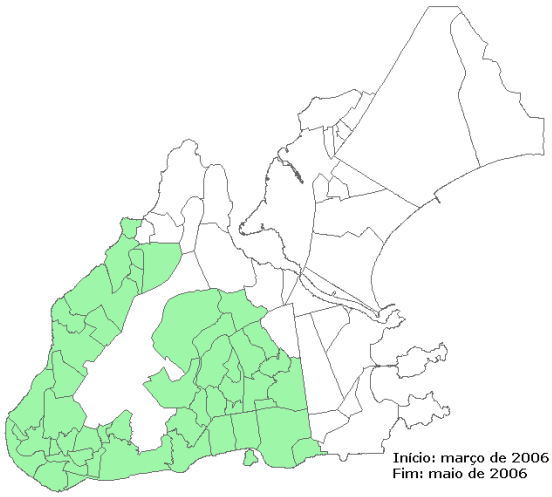

Figura 6.7: Conglomerado de casos de dengue detectado entre os meses de março a maio de 2006 pelo método Scan circular espaço-tempo aplicado no banco de dados de Vitória.

da doença que não pertencesse ao conglomerado real. Nesse contexto, para reduzir o número dessas sub-áreas, uma recomendação seria a *análise da interseção* de todos os métodos para identificá-lo. Essa análise consiste em selecionar as áreas em comuns de cada conglomerado encontrado.

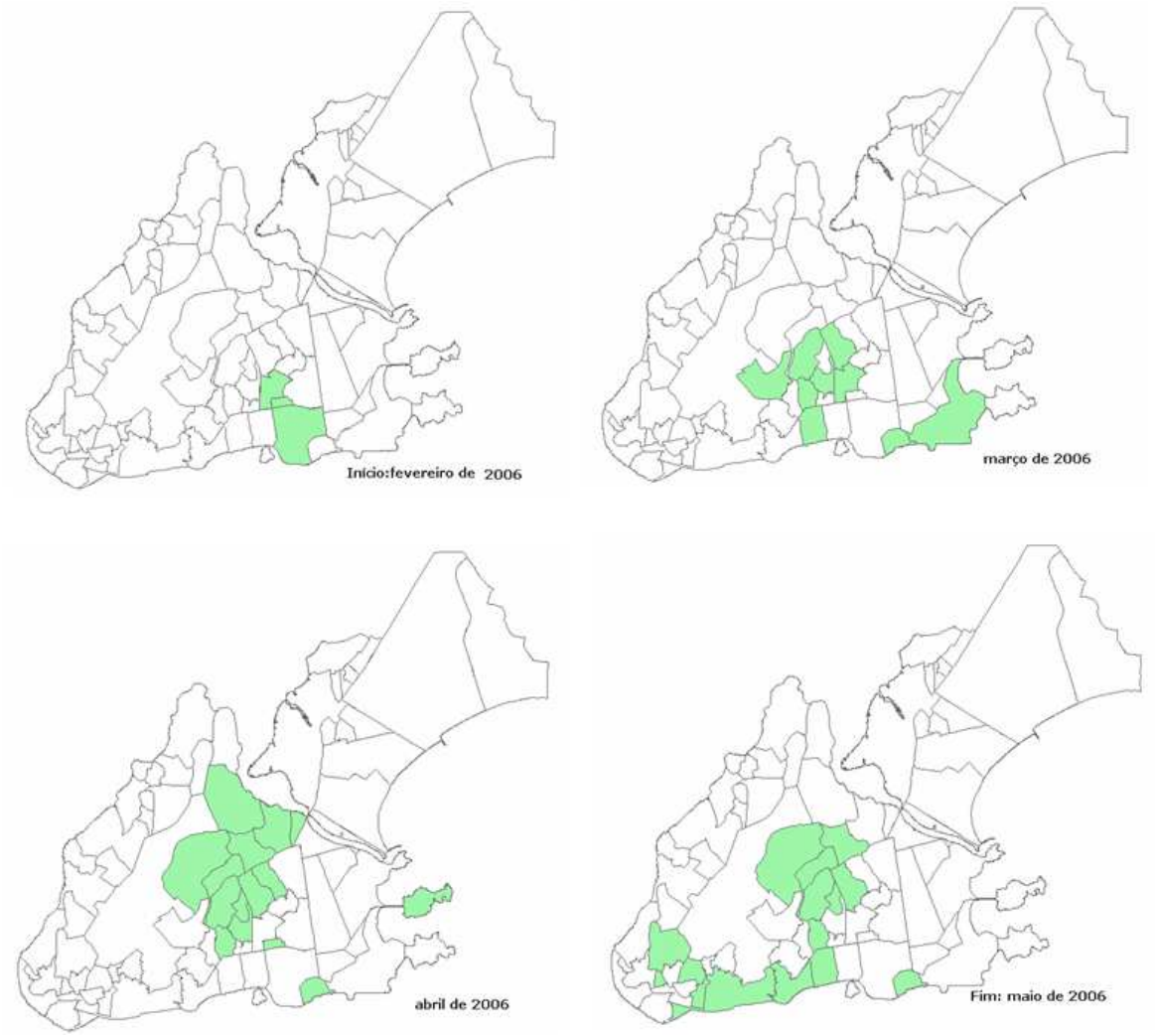

Figura 6.8: Conglomerado de casos de dengue detectado entre os meses de fevereiro a maio de 2006 pelo método *dMST* espaço-tempo aplicado no banco de dados de Vitória.

Tabela 6.3: Tamanho, P-valor e a Estatística de Teste dos Métodos de detecção espaço-tempo aplicado aos casos de dengue no banco de dados de Vitória para o modelo Bernoulli e Poisson.

	Método	**Tamanho**	**Estatística de Teste**	**P-valor**
	Scan Circular	141	668,929812	0,001
	dMST	42	710,8246	0,0001
Modelo Bernoulli	Doubly	29	474,796	0,0001
	Scan Circular	141	669,380437	0,001
	dMST	42	709,8339	0,0001
Modelo Poisson	Doubly	29	474,134625	0,0001

Capítulo 7

Considerações Finais

O poder do teste, que mede a habilidade de detecção do método sob a hipótese alternativa da existência de um conglomerado real, depende do número de casos no conjunto de dados, do tamanho da área do conglomerado medido em termos do número esperado de casos, sob a hipótese nula, do risco relativo utilizado dentro e fora do conglomerado, da geometria de busca escolhida e, principalmente, do conjunto escolhido de dados. Para o presente estudo nossas conclusões sobre o poder dos testes de detecção espaço-tempo foram baseadas nos dados do Novo México, onde já especificamos, anteriormente, todos os parâmetros necessários para avaliar o poder do teste.

O poder do teste é aproximado pela freqüência de vezes que o método pesquisado rejeita a hipótese nula. A partir das 10.000 simulações avaliou-se o poder de detecção dos métodos *Scan* circular, *dMST* e *Doubly* no espaço-tempo. Porém, existe a possibilidade de que nessa avaliação o método estudado encontre algum conglomerado espúrio que é completamente distinto do conglomerado real. Sendo assim, a capacidade verdadeira de detectar o conglomerado real é menor do que o poder estimado. Concluímos que existem outros critérios que devem ser levados em consideração na avaliação do poder dessas simulações. Os critérios da proporção de áreas do conglomerado real e da proporção da população do conglomerado real, fizeram com que identificássemos, experimentalmente, a proporção ideal de interseção entre os dois conglomerados para que o poder dos testes fosse estimado de forma sensata. Do ponto de vista de freqüência de detecção do conglomerado real nas simulações, os resultados são muito interessantes, pois verificamos, experimentalmente, que o método *Scan* circular tem um poder de 0,8, aproximadamente, para identificar 97% a 100% da população do conglomerado real para o cenário cilíndrico. Verificamos também que a proporção da população do conglomerado real (P_P) encontrado pelo o método *dMST* no cenário cilíndrico foi de 67% a 75% com um poder de 0,9, aproximadamente. E também para o método *Doubly*, a proporção da população do conglomerado real (P_P) foi um intervalo de 76% a 82,6% com poder de 0,7, aproximadamente. Para o caso do cenário arbitrário, o método *Scan* circular identificou uma proporção de 6,56% da população do conglomerado real com poder de 0,6, aproximadamente. O método *dMST* identificou um P_P que está no intervalo de 57% a 75% com poder de 0,9, aproximadamente. O método *Doubly* identificou um intervalo de proporção da população do conglomerado real

(P_P) de 59,95% a 75% com poder de 0,6, aproximadamente.

Assim, os resultados desse estudo nos mostram, claramente, que os testes de detecção de conglomerado espaço-tempo são úteis para diferentes tipos de hipóteses alternativas de interesse. Se estivermos interessados em detectar conglomerado com geometria cilíndrica, é melhor usar o método *Scan* circular, enquanto para um conglomerado de geometria arbitrária é aconselhável usar os métodos *Doubly* ou *dMST*. Se não tivermos nenhum conhecimento, a priori, da geometria do conglomerado real, o método *Doubly* é uma boa opção, entretanto, este apresenta menor poder de detecção comparado com o *Scan* circular no cenário cilíndrico e também menor poder comparado com o *dMST* no cenário arbitrário.

Adicionar flexibilidade à geometria de busca pode gerar o efeito *polvo* que pode ser minimizado a partir de heurísticas de parada prematura, como implementado no método *dMST*, ou restrições no crescimento do conglomerado, como abordado pelo algoritmo *Doubly*. A característica principal dos métodos de crescimento arbitrário é a sub-estimação do tamanho da interseção do conglomerado real com o conglomerado detectado, conforme observado nos resultados. Mesmo programando um conglomerado com tamanho mínimo especificado pelo usuário, não há evidência de uma melhora significativa na estimação do mesmo.

Na aplicação aos dados reais do Novo México, a análise da interseção de todos os métodos indica um conglomerado espaço-tempo de casos de câncer cerebral com geometria arbitrária estatisticamente significativo (p-valor menor que 5%) em Los Alamos, Santa Fé, Torrance, San Miguel, Socorro, Bernalillo, Valência e Cibola emergente em 1986 e detectado em 1989.

Na aplicação aos dados reais de Vitória, a análise da interseção de todos os métodos indica um conglomerado espaço-tempo de casos de dengue com geometria arbitrária e estatisticamente significativo (p-valor menor que 5%) em Forte São João, Ilha de Santa Maria, Jucutuquara, Bairro da Penha, Bairro de Lourdes, Bonfim, Consolação, Joana D'arc, Maruípe, Santa Cecília, Santa Martha, Santos Dumont, São Cristóvão e Tabuazeiro emergente em março e detectado em maio de 2006, o que confirma a manutenção do padrão de sazonalidade da dengue no Brasil, que acompanha a estação chuvosa (verão).

Do ponto de vista da detecção de conglomerados nos dados reais, comparando com outros métodos existentes na literatura, os métodos de varredura abordados nesta dissertação possuem algumas características importantes, pois levam em conta as diferenças espaço-tempo da população de risco, soluciona o problema de ajuste de testes múltiplos, localiza no mapa um candidato a conglomerado e ainda não formula hipótese, a priori, sobre o conglomerado real tais como localização ou comprimento do tempo. Apesar disso, os métodos apresentam algumas deficiências. A principal delas é a tendência a identificar um conglomerado maior do que o conglomerado real devido a diferença entre a forma geométrica desses. A deficiência dos métodos de detecção pesquisado mostra que os resultados da análise estatística devem ser usados apenas como guia para a formulação de hipóteses sobre a real situação do conglomerado espaço-tempo.

Para esta pesquisa utilizamos um típico conjunto de dados epidemiológicos para aplicação, em que a população e os casos são agregados em áreas. Seria de maior informação nas conclusões deste estudo elaborar um conjunto de dados em que cada caso tivesse sua própria

coordenada geográfica e população.

Os métodos estudados não incluíram covariáveis nas observações, e seria interessante estudar modelos incorporando fatores de risco. Já existe para o método *Scan* circular modelos que levam em conta fatores de risco, porém para o *dMST* e *Doubly* ainda não. Também como estudos futuros sugerimos programar para o método *Doubly* e *dMST* a verossimilhança exponencial para aplicar em conjuntos de dados de sobrevivência que são caracterizados pelos tempos de falha e, muito frequentemente, pelas censuras (presença de observações incompletas ou parciais). Esses dois componentes constituem a variável resposta.

Anexo

Diagnóstico dos Resíduos

No geral, pode-se dizer que praticamente não existe pontos suspeitos de serem aberrantes e/ou influentes pelos gráficos abaixo.

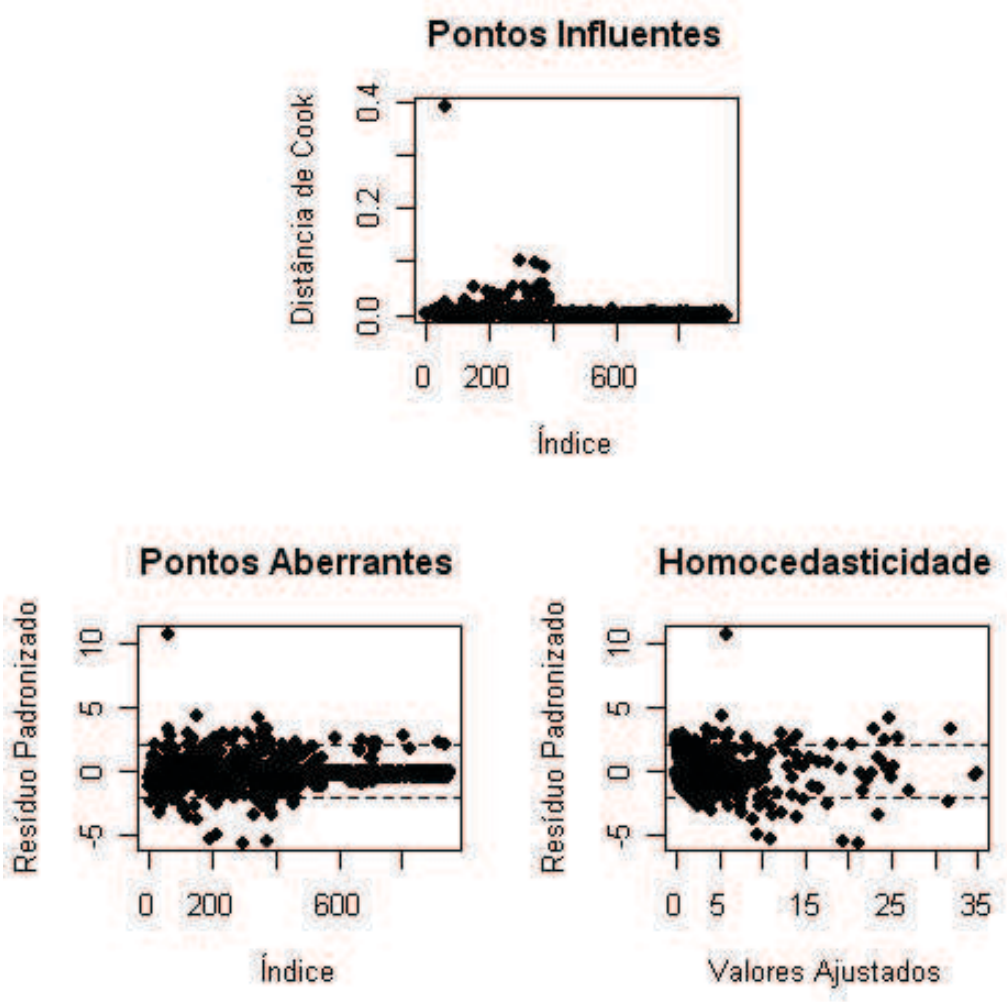

Figura 7.1: Gráfico de diagnóstico dos resíduos para o exemplo sobre casos de dengue no banco de dados de Vitória.

Referências Bibliográficas

[1] Alexander, F.E e Boyle, P. (1996). *Methods for Investigating Localized Clustering of Disease*. IARC Scientific Publication 135. Lyon. France. (100,176).

[2] Anderson,N.H e Titterington,D.M. (1997). *Some Methods for Investigating Spatial Clustering, with Epidemiological Applications*. Journal of the Royal Statistics Society, Séries A, **160**, 87-105.

[3] Bailey, T.C. e Gatrell A. C. (1995). *Interactive Spatial Data Analysis*.Longman Scientific Technical, England.

[4] Bailey, T.C. (2001). *Spatial statistical methods in health*. Manuscript.

[5] Besag, J. e Diggle, P.J. (1977). *Simple Monte Carlo tests for spatial pattern*. Applied Statistics, **26**, 327-333.

[6] Besag, J e Newell, J. (1991). *The detection of clusters in rare disease*.Jornal of the Royal Statistical Society, Series A,**154**, 143-155.

[7] Bonetti,M. e Pagano, M (2001). *On detecting clustering*. Proceedings of the Biometrics Section, American Statistical Association,pp, 24-33.

[8] Câmera G, Monteiro M, Fucks S e Carvalho M. (2000). *Geoprocessamento: Teoria e Aplicações - Análise Espacial de Dados Geográficos*. Um dos quatro livros eletrônicos disponíveis para para download no site do Gilberto Câmera do INPE.

[9] Choynowski,M. (1959). *Maps based on probabilities*. Journal of the American Statistical Association,**54**, 385-388.

[10] Costa MA e Assunção RM.(2005). *A fair comparison between the spatial scan and Besag-Newell disease clustering tests*. Environmental and Ecological Statistics, **12**, 301-319.

[11] Costa MA, Assunção RM e Scherrer L.(2006). *Detecção de Conglomerados Espaciais com Geometria Arbitrária*. In: Informática Pública, v.8, n.1, pp. 23-34.

[12] Cuzick, J. e Edwards, R. (1990). *Spatial clustering for inhomogeneous populations*. Journal of the Royal Statistics Society, Séries B, **52**, 73-104.

[13] Davies,P. (1976). *The American Heritage Dictionary of the English Language*. New York: Dell Publishing Co., Inc.

[14] Duczmal, L. e Assunção R.M. (2003). *A simulated annealing strategy for the detection or arbitrarily shaped spatial clusters.* Computational Statistics and Data Analysis,**45**, 269-286.

[15] Dwass,M. (1957). *Modified randomization test for nonparametric hypotheses.* Annals of Mathematical Statistics, **28**,181-187.

[16] Fonseca, J.A.(2001).*Avaliando Poder Estatístico do Teste da Razão de Verossimilhança Para Detecção de Aglomerados (Clusters) sob Modelo Alternativo Mal Especificado.* Dissertação de Mestrado,Icex-UFMG.

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

[18] Knox,G. (1989). *Detection of clusters.* In Methodology of Enquiries into Disease Clustering, London, Small Area Health Statistics Unit, 17-22.

[19] Kulldorff M e Nagarwalla N.(1995). *Spatial disease clusters: Detection and Inference.* Statistics in Medicine, **14**, 799-810.

[20] Kulldorff M.(1997). *A spatial scan statistic.* Communications in Statistics: Theory and Methods, **26**, 1481-1496.

[21] Kulldorff M, Feuer EJ, Miller BA e Freedman LS.(1997). *Breast cancer in northeastern United States: A geographical analysis.* American Journal of Epidemiology, **146**, 161-170.

[22] Kulldorff M, Athas W, Feuer E, Miller B e Key C.(1998). *Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos.* American Journal of Public Health, **88**, 1377-1380.

[23] Kulldorff M.(2001). *Prospective time-periodic geographical disease surveillance using a scan statistic.* Journal of the Royal Statistical Society **A,164**, 61-72.

[24] Kulldorff M, Tango T e Park P. (2003). *Power comparisons for disease clustering tests.* Computational Statistics and Data Analysis, **42**, 665-684.

[25] Kulldorff M, Heffernan R, Hartman J, Assunção RM e Mostashari F.(2005) *A space-time permutation scan statistic for the early detection of disease outbreaks.* PLoS Medicine, **2**, 216-224.

[26] Kulldorff M,Costa MA e Assunção RM.(2007). *Constrained Spanning Tree Algorithms for Irregular Spatial Clustering.* Journal of Computational and Graphical Statistics(submitted).

[27] Kulldorff M, Huang L, Pickle L e Duczmal L.(2006). *An elliptic spatial scan statistic.* Statistics in Medicine.

[28] Lawson,A.B. (1993). *On the analysis of mortality events around a prespecified fixed point.* Journal of the Royal Statistics Society, Series A,**156**, 363-377.

[29] Lawson,A.B e Clark,A. (1999). *Markov Chain Monte Carlo Methods for Putative Source of Hazard and General Clustering.* In Disease Mapping and Risk Ressessment for Public Health, pp. 120-142, chichester: Editora John Wiley and Sons.

[30] Lawson, A.B e Kulldorff, M. (1999). *A review of cluster detection methods.* In: Disease Mapping and Risk Assessment for Public Health, pp. 99-110, chichester: Editora John Wiley and Sons.

[31] Lima, M.S.(2004).*Avaliação do Poder do Teste da Estatística Scan para Múltiplos Clusters.* Dissertação de Mestrado,Icex-UFMG.

[32] Magalhães MN e Lima ACP. (2001).*Noções de Probabilidade e Estatística.* 3 edição, São Paulo: Edusp.

[33] Marshal, R.C. (1991). *A review of the statistical analysis of spatial patterns of disease.* Journal of the Royal Statistics Society, Series A, **154**, 421-441.

[34] Marzochi,K.B.F. (1994). *Dengue in Brazil:situation,transmission and control - a proposal for ecological control.* Mem.Inst.Oswaldo Cruz,**34**, 235-245.

[35] Molenaar, W. (1973). *Approximations to the Poisson, Binomial and Hipergeometric distribution.* Matematisch Centrum Amsterdam, **31**, 403-407.

[36] Openshaw, S., Charlton, M. e Craft, A. W. (1987). *A Mark 1 Geographical Analisys Machine for the automated analysis of point data set.* International Journal of Geographical Systems,**1**, 335-358.

[37] Paula, G.A. (1994).*Modelos de Regressão com apoio Computacional.*IME-USP.

[38] Ronald, E.G e Murray, K.C. (2001). *A weighted average likelihood ratio test for spatial clustering of disease.* Statistics in Medicine,**20**, 2977-2987.

[39] Rothman,K.J. (1990). *A sobering start to the cluster buster conference.* American Journal of Epidemiology,**132**, 6-13.

[40] Stone, R.A. (1988). *Investigation of excess enviornmental risk around putative source: statistical problems and a proposal test.* Statistics in Medicine,**7**, 649-660.

[41] Tango, T. (1995). *A class of test for detecting general and focused clustering of rare diseases.* Statistics im Medicine, **14**, 2323-2324.

[42] Tango,T. (1999). *Comparison of General Test for Spatial Clustering.*In: Disease mapping and risk assessment for public health, pp.120-142,chichester:Editora John Wiley and Sons.

[43] Tango, T. (2000). *A test spatial disease clustering adjusted for multiple testing.* Statistics im Medicine, **19**, 191-204.

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

[45] Turnbull, B. W., Iwano, E.J., Burnett, W. S., Howe, H.L. e Clark, L.C. (1990). *Monitoring for clusters of disease: application to leukemia incidence in upstate New York.* American Journal of Epidemiology,**132**, 136-143.

[46] Waller,L.A. e Lawson,A.B. (1995). *The power of focused test to detect disease clustering.* Statistics in Medicine,**14**, 2291-2308.

[47] Wartenberg,D. e Greenberg,M. (1990). *Detecting diseases cluster: the importance of statistical power.* American Journal of Epidemiology,**132**, 156-166.

[48] Whittemore,A.S., Friend, N., Brown J., e Holly, E.A. (1987). *A test to detect clusters of disease.* Biometrika,**74**, 631-635.