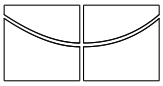


Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Estimação de incertezas no delineamento de clusters espaciais com dados pontuais

por

Wesley de Jesus Silva

Orientador: Prof. Dr. André L.F. Cançado

Brasília/DF, Junho de 2012

Wesley de Jesus Silva

Estimação de incertezas no delineamento de clusters espaciais com dados pontuais

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília/DF, Junho de 2012

Termo de Aprovação

Wesley de Jesus Silva

Estimação de incertezas no delineamento de clusters espaciais com dados pontuais

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidadede de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Data da defesa: 29 de Junho de 2012

Orientador:

Prof. Dr. André L.F. Cançado
Departamento de Estatística, UnB

Comissão Examinadora:

Prof. Dr. Alan Ricardo da Silva
Departamento de Estatística, UnB

Prof. Dr. Luiz Henrique Duczmal
Departamento de Estatística, UFMG

Brasília/DF, Junho de 2012

Ficha Catalográfica

SILVA, WESLEY DE JESUS
Estimação de incertezas no delineamento de clusters espaciais com dados pontuais, (UnB - IE, Mestre em Estatística, 2012).

Dissertação de Mestrado - Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

1. *Cluster* Espacial 2. Estatística Scan de Kulldorff 3. Medidas de Intensidade
4. Dados Pontuais 5. Dados Pontuais 6. *Minimum Spanning Tree - MST*
7 . Diagrama de Voronoi

Aos meus pais, Genilson e Luzia,

ao meu irmão, Webert,

e à minha namorada, Maíra,

que foram fonte de apoio, motivação e inspiração.

Agradecimentos

- Agradeço ao meu orientador, André Caçado, pelo apoio, entusiasmo e solicitude.
- Aos membros da banca avaliadora, por terem aceito o convite e se disponibilizado a avaliar o presente trabalho.
- Ao Departamento de Estatística, pelo empenho do excelente corpo docente e pelo apoio fundamental da Secretaria.
- Aos companheiros e ex companheiros de trabalho do Instituto de Pesquisa Econômica Aplicada, que apoiaram, motivaram e foram compreensivos em muitas das etapas dessa árdua jornada.
- Aos demais mestrandos do Departamento de Estatística da UnB, companheiros de turma com os quais compartilhei momentos difíceis, mas também também muitas conquistas.
- Aos amigos, familiares e companheira, que acreditaram em mim.

Resumo

A preocupação em detectar anomalias em um espaço bidimensional é bastante antiga, e sua importância surgiu a partir de questões de saúde pública envolvendo a detecção de *excessos de ocorrência local* de enfermidades ou indícios de concentração de casos de doenças. Técnicas voltadas à identificação de clusters prováveis foram amplamente empregadas, e grandes avanços foram obtidos com o uso da *Estatística Scan de Kulldorff*, permitindo ao mesmo tempo a detecção e o teste de significância associado ao cluster mais provável. Bem recentemente, outro grande passo foi dado ao se propor *medidas de intensidade*.

As medidas de intensidade estão relacionadas com a importância de cada área como parte da anomalida detectada, além de captar regiões de influencia do cluster mais verossímil. Em suma, tais medidas permitem delinear *incertezas* inerentes ao processo de detecção de conglomerados espaciais.

Essa metodologia estava restrita, até agora, apenas ao caso de dados agregados em regiões delimitadas. O ganho de informação que se tem com dados em referência local, entretanto, não pode ser desprezado, nem tampouco a possibilidade de visualização das incertezas envolvidas em observações pontuais do tipo caso-controle. Essa é a motivação de um esforço ainda não realizado: a implementação de medidas de intensidade associadas a cada ponto em um mapa.

A solução proposta baseia-se na consideração de *vizinhanças* em torno de cada ocorrência: regiões circulares centradas nos casos cujas áreas foram delimitadas com auxílio de uma *Árvore Geradora Mínima*(MST).

Palavras Chave: *Cluster Espacial, Estatística Scan de Kulldorff, Medidas de Intensidade, Dados Pontuais,Minimum Spanning Tree - MST, Diagrama de Voronoi.*

Abstract

The concern on detecting anomalies in a two-dimensional space is quite old, and its importance arose from public health issues involving the observation of *local excess of disease ocurrence*, or signs of disease cases concentration. Techniques aiming the identification of likely clusters have been widely employed, and great advances have been obtained through Kulldorff's *Spatial Scan Statistic*, allowing at the same time the detection and the significance test associated with the most likely cluster. Recently, another big step was taken through the proposition of the *intensity function*.

The intensity function is related to the importance of each area as part of the detected anomaly, and defines a influence region of the most likely cluster. In short, such measures allow the outline of *uncertainty bounds* inherent to the detection process.

This method was restricted, until now, only to aggregated data case. However, the gain of information that arises from local reference data can not be discarded, neither the possibility of viewing uncertainties involved in case-control point observations. This is the motivation of a not performed effort: the application of the intensity function to each point in a map.

The proposed solution is based on *neighborhoods* around each case: circular regions centered in the cases, whose areas was defined by edges of a *Minimum Spanning Tree*(MST).

Keywords: *Spatial Cluster, Kulldorff's Spatial Scan Statistic, Intensity Function, Local Reference Data, Minimum Spanning Tree - MST, Voronoi Diagram.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Na maior parte das aplicações de procedimentos estatísticos, procura-se responder uma questão bem genérica: é possível observar alguma regularidade dentro de um particular sistema ou situação no qual há variabilidade inerente? Com esse objetivo, muitos modelos e testes de hipótese foram desenvolvidos levando-se em conta justamente esse fator aleatório. A Estatística Espacial compreende técnicas dessa natureza, porém levando-se em conta a posição geográfica do elemento observado na tentativa de responder a uma indagação mais restrita: é possível observar alguma regularidade espacial nos elementos observados? Um tema particular da Estatística Espacial procura responder uma questão ainda mais específica: dado um sistema de coordenadas, existe alguma coleção ou *cluster* de regiões contido nesse sistema, onde um particular evento de interesse ocorre com mais frequência que nas demais? Em outras palavras, há algum *cluster* espacial desse evento?

A *Estatística Espacial* abrange todo um universo de técnicas, procedimentos e modelagens estatísticas envolvendo algum tipo de observação definida no espaço, isto é, que está relacionada, direta ou indiretamente, com a *posição* ou *localização* de um ou mais elementos. Cada tipo de abordagem depende do tipo de informação observada. Este trabalho fará várias referências a dois tipos comuns de dados observados nesse campo. Um deles, os dados *agregados*, são aqueles em que o que se observa são *áreas* ou *regiões* às quais estão associadas alguma grandeza relacionada a uma coleção de fenômenos, tais como o índice de criminalidade em Brasília, número de acidentes de trânsito em um estado, etc. Geralmente, a referência à localização, em dados agrega-

dos, é dada pelos *centróides* dessas regiões, isto é, coordenadas de pontos arbitrários dentro de cada uma dessas áreas. Outro tipo de observação a ser abordada é o caso de dados *pontuais*. Nesse tipo de informação, em vez de um conjunto de fenômenos associados a uma área com uma coordenada em comum (o centróide dessa região), sabe-se a localização exata de *cada fenômeno* considerado. Nesse caso, saberíamos não o número de assaltos, mas a localização de cada assalto, de cada ocorrência de Dengue, etc.

A existência de um *cluster* espacial, geralmente investigada através de observações pontuais ou agregadas, é uma questão que tem apresentado relevância há algum tempo, principalmente com relação à saúde pública na detecção de surtos de epidemias e análises de casos de câncer . Em [Rothman, 1987] é citada a importância etiológica dada à detecção de *cluster* de doenças visto que, presumivelmente, se uma enfermidade está concentrada em uma determinada região, então muito provavelmente suas prováveis causas também estão. Porém, citou estudos que até aquela época foram infrutíferos no sentido de descobertas de relações causais. Warner e Aldrich [Warner & Aldrich, 1988] também citam essa improdutividade e listam, entre outras questões, a importância de se definir métodos formais de distinção entre *clusters* efetivos e puramente aleatórios.

A detecção de *cluster*, entretanto, não é essencial apenas em epidemiologia, como também em vários outros campos científicos, exemplificados em [Kulldorff, 1997]. Um geólogo pode necessitar de um estudo sobre o padrão de concentração de determinado tipo de minério em um espaço determinado. Em silvicultura, há potencial interesse em verificar se há *clusters* de um determinado tipo específico de árvore em uma área florestal. Na astronomia, pode haver interesse na determinação de *clusters* de um determinado tipo de estrela.

O histórico de métodos voltados à detecção de *cluster* também é vasto. Em 1965, Naus já havia desenvolvido estudos de detecção de *cluster* em processos pontuais unidimensionais [Naus, 1965b] e bidimensionais [Naus, 1965a], através da estatística Scan. Em [Turnbul *et al.*, 1989] foi implementado um método de detecção de *cluster* baseado em "janelas" sobrepostas no mapa, de maneira que cada conjunto tenha população constante, além de comparar seu método com os propostos por Wittemore [Whittemore *et al.*, 1987] e Openshaw [Openshaw *et al.*, 1988]. Ele não deixa de ci-

tar, entretanto, a dependência entre as variáveis observadas em seu próprio método. A estatística de Whittemore é baseada na distância média entre todos os pares de casos, mas tal método não permite a localização dos *clusters*, além de ser sensível às flutuações da densidade populacional em diferentes locais.

Já o método iterativo GAM (*Geographical Analysis Machine*) de Openshaw é baseado em áreas circulares sobrepostas centradas em cada observação, cujo raio varia em cada iteração. O principal problema nesse método é o fato de que um teste de hipótese é realizado em cada etapa com um determinado nível de significância predeterminado, e assim o nível de significância real do procedimento resulta em um valor muito mais baixo do que o adotado.

Muitos outros métodos além dos exemplificados foram implementados, como em [Anderson & Titterington, 1996], que baseia-se na diferença entre as estimativas de densidades kernel. Mas foi em [Kulldorff, 1997] que Kulldorff desenvolveu não só um método prático e eficiente para detecção de *cluster*, como também um teste poderoso de significância. Várias extensões desse método foram desenvolvidas logo após, como detecção de *clusters* espaço-temporais [Kulldorff, 2001], *clusters* em formatos irregulares [Tango & Takahashi, 2005], *clusters* em dados pontuais [Duczmal *et al.*, 2011], e *clusters* em formato elíptico [Kulldorff *et al.*, 2006].

Os métodos estatísticos não se limitam apenas a testes de hipótese. Além de testar se determinado padrão é efetivo ou puramente um acaso, faz-se necessário observar *limites de incerteza* de estimativas pontuais em intervalos de confiança. Esforços já foram feitos na proposição de métodos de estimação de incertezas em *clusters* espaciais para dados agregados, em [Rosychuk, 2005]. Entretanto, o método proposto limita-se a estimar intervalos de confiança para cada região em estudo, sem uma visualização geral da incerteza no espaço observado. Um grande avanço nesse esforço ocorreu recentemente. Em [Oliveira *et al.*, 2011] foi proposta uma *medida de intensidade*, permitindo uma visão bem mais ampla da incerteza envolvida na detecção do *cluster* mais provável.

Uma questão ainda não abordada é o uso dessas medidas de incerteza em observações não agregadas em regiões delimitadas. O presente trabalho pretende, portanto, propor maneiras de se medir incertezas associadas à detecção de *cluster* conside rando-se dados puramente pontuais. Em um primeiro momento, serão descritas as de-

finições gerais das técnicas desenvolvidas por Kulldorff. O capítulo seguinte abordará a medida de intensidade definida em dados agregados. Em seguida, serão revisados conceitos e definições úteis e largamente utilizadas em trabalhos recentes envolvendo dados pontuais. Finalmente, serão apresentadas mais formalmente as soluções propostas para a obtenção de medidas de intensidade em dados pontuais, bem como sua aplicação em dados reais envolvendo ocorrência de casos de Dengue no município de Lassance, Minas Gerais.

Capítulo 2

A estatística Scan de Kulldorff

2.1 Introdução

A estatística Scan é usada para detectar *clusters* em processos pontuais. Dado um intervalo $[a, b]$ no qual um processo pontual é definido, o procedimento baseia-se na definição de janelas $[t, t+\omega]$, onde ω é o tamanho da janela de modo que $\omega < b-a$. Então, sobre os valores de t possíveis, o número de pontos em cada janela é calculado e comparado com o seu valor esperado sob a hipótese nula. O objetivo é verificar se a distribuição de pontos pelas janelas é resultante de um processo puramente aleatório ou se algum *cluster* pode ser detectado. Em [Naus, 1965b] é obtida a distribuição dessa estatística sob a hipótese nula.

Kuldorff amplifica essa implementação para o caso de um *processo pontual no espaço*, propondo assim uma estatística Scan *espacial* em [Kulldorff, 1997]. Como em [Naus, 1965b], o procedimento também será baseado em janelas, porém com formas pré-determinadas e com o tamanho variando à medida em que "varre" a região em estudo. Outra característica essencial da estatística Scan de Kuldorff é que ela sempre dependerá do número total de pontos observados. O método de Kulldorff pode ser implementado para dados agregados ou pontuais sem maiores distinções entre os dois casos.

Considere um espaço geográfico G em que são observados N pontos, dos quais C possuem alguma caracterítica de interesse (um tipo especial de árvore, algum paciente que sofre de determinada enfermidade, etc). O espaço G pode ser um país, um

estado ou simplesmente um espaço cartesiano bi ou tridimensional, e pode ou não ter subdivisões $R_1, \cdots, R_n$ definidas (mesorregiões, municípios, unidades censitárias, etc.)[1]

De modo geral, o problema de deteção de *cluster* espacial abrange a verificação de "excessos de ocorrência" [Rothman, 1987], definidos em termos de "excessos além do esperado". Sob a hipótese de não existência de *cluster* (hipótese nula), qualquer ponto na região $A \subset G$ terá probabilidade p de ser uma ocorrência do evento de interesse. Nesse caso, o número esperado μ_A de ocorrencia será dado por

$$\mu_A = \frac{C}{N} N_A \tag{2.1}$$

onde N_A é o número total de pontos observados na região A; C e N são, respectivamente, o número total de casos e de observações em todo o espaço G. Ainda sob H_0, $p = \frac{C}{N}$.

O excesso de ocorrência pode ser observado através do risco relativo, a razão entre o número observado e o número esperado de casos sob H_0. Entretanto, o risco relativo por si só não é suficiente para auxiliar na detecção de *cluster*. Considere, por exemplo, o caso de 2 regiões A e B com $N_A = 100$, $N_B = 10.000$, $C_A = 2$, $C_B = 200$, contidas em um espaço geográfico com $N = 10.000.000$ e $C = 100.000$. Pela equação (2.1), $\mu_A = 1$ e $\mu_B = 100$. Nesse caso, $C_A/\mu_A = C_B/\mu_B = 2$. Porém, o risco relativo observado em A é muito mais provável de ser uma flutuação aleatória do que em B. As chances de que o número de casos passe de 100 para 200 é muito menor do que as chances de passar de 1 para 2. Como será visto adiante, a *Estatística Scan de Kulldorff* é capaz de distinguir esses dois casos.

A definição de *cluster* adotada por Kulldorff baseia-se nessa mesma idéia de excesso de ocorrência. Formalmente:

Definição 1. *Um cluster, se existir, é um conjunto ou zona $Z \subset G$ onde a probabilidade de um ponto ser uma ocorrência de caso será p, enquanto pontos fora de Z*

[1] O fato de o espaço G estar ou não particionado em regiões pré-determinadas é o que diferencia a natureza de dados agregados ou pontuais. Apesar da definição bem geral do método de Kuldorff, tal diferenciação muda drasticamente a implementação de medidas associadas à incerteza, e é justamente nessa diferença que o presente trabalho está focado.

terão probabilidade q de modo que $p > q$.

A existência de *cluster* (hipótese alternativa) está relacionada, portanto, à existência de uma zona Z que satisfaça a Definição 1. Assim, o objetivo é encontrar, dentre uma coleção $\mathcal{Z}$ de zonas Z possíveis, aquela em que esse fato ocorre.

Além da detecção, a efetividade do cluster é verificada através de um teste de hipóteses cuja hipótese nula é a de não existência de *cluster*, ou seja, $p = q$, e esta é testada contra a hipótese de existência de alguma região de risco Z tal que $p > q$. De maneira formal, temos:

$$\begin{cases} H_0 & p = q \\ H_1 & p > q \quad \text{para algum } Z \in \mathcal{Z}. \end{cases} \tag{2.2}$$

A estatística Scan proposta por Kulldorff em [Kulldorff, 1997] baseia-se em uma distribuição discreta para a variável aleatória C_A, o número de casos em qualquer região $A \subset G$, que pode ser um modelo Bernoulli ou Poisson dependendo do tipo de contagem adotado. Quando o número total de casos C é pequeno se comparado com N, os dois modelos se aproximam. Caso contrário, o modelo Bernoulli é mais adequado quando a contagem é resultante de uma *contagem binária*, enquanto que o modelo Poisson deve ser usado quando a contagem está relacionada a algum fator de risco contínuo.

2.2 O modelo Bernoulli

No modelo Bernoulli, cada unidade de N_A representa um elemento ou entidade que pode estar em um de dois estados possíveis, caso ou controle. O modelo supõe a existência de uma zona Z tal que cada ponto dentro desta região terá probabilidade p de ser um caso, enquanto cada ponto fora de Z terá probabilidade q. Sob H_0, $p = q$ e $C_A \sim Bin(N_A, p)$ para todo A. Sob a hipótese alternativa, $H_1 : p > q$, $C_A \sim Bin(N_A, p)$ para todo $A \subset Z$ e $C_A \sim Bin(N_A, q)$ para todo $A \subset Z^c$.

A função de verossimilhança baseada no modelo Bernoulli sob a hipótese H_1 é dada por

$$L(Z,p,q) = p^{c_z}(1-p)^{N_Z-c_z}q^{C-c_z}(1-q)^{(N-N_Z)-(C-c_z)}. \tag{2.3}$$

A zona que mais provavelmente define um *cluster* será a zona $\hat{Z}$ que maximize (2.3). Ou seja, $\hat{Z}$ é o estimador de máxima verossimilhança de Z, e sua obtenção é dada em duas etapas. Primeiro, encontra-se o máximo de (2.3) condicionado a Z.

$$\begin{aligned} L(Z) &\stackrel{def}{=} \sup_{p>q} L(Z,p,q) = \left(\frac{c_Z}{N_Z}\right)^{c_Z}\left(\frac{N_Z-c_Z}{N_Z}\right)^{N_Z-c_Z} \\ &\times \left(\frac{C-c_Z}{N-N_Z}\right)^{C-c_Z}\left(\frac{(N-N_Z)-(C-c_Z)}{N-N_Z}\right)^{(N-N_Z)-(C-c_z)} \end{aligned} \tag{2.4}$$

se $\frac{c_Z}{N_Z} > \frac{C-c_Z}{N-N_Z}$. Caso contrário,

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

A segunda etapa consiste em encontrar $\hat{Z}$ tal que (2.4) é maximizada. Sob H_0, a função de verossimilhança é definida por

$$L_0 \stackrel{def}{=} \sup_{p=q} L(Z,p,q) = \left(\frac{C}{N}\right)^{C}\left(\frac{N-C}{N}\right)^{N-C}. \tag{2.5}$$

É interessante notar que L_0 depende apenas do número total de casos e do total de observações.

A estatística Scan proposta por Kulldorff é definida pela razão λ das verossimilhanças definidas acima.

$$\lambda = \frac{\sup\limits_{Z\in\mathcal{Z},p>q} L(Z,p,q)}{\sup\limits_{p=q} L(Z,p,q)} = \frac{L(\hat{Z})}{L_0} \tag{2.6}$$

2.3 O modelo Poisson

O modelo Poisson supõe que os pontos são gerados por um processo de Poisson não-homogêneo. A formulação das hipóteses é a mesma do modelo Bernoulli. Sob

a hipótese alternativa, $C_A \sim Poi(pN_{A\cap Z} + qN_{A\cap Z^c})\ \forall A$. Sob H_0, $p = q$ e $C_A \sim Poi(pN_A)\ \forall A$.

A função de verossimilhança sob H_1 baseia-se na distribuição de probabilidades do número total de casos.

$$P(C) = \frac{e^{-pN_Z - q(N-N_Z)}\left[pN_Z + q\left(N - N_Z\right)\right]^C}{C!} \tag{2.7}$$

A probabilidade de que um caso específico seja observado em um determinado local x é dada por

$$\begin{cases} \mathrm{pN}_x/(pN_Z + q(N - N_Z)) & \text{se } x \in Z \\ \mathrm{qN}_x/(pN_Z + q(N - N_Z)) & \text{se } x \notin Z \end{cases},$$

sendo que N_x é o número total de observações em x. A função de verossimilhança é dada por.

$$\begin{aligned} L(Z,p,q) &= \frac{e^{-pN_Z - q(N-N_Z)}\left[pN_Z + q\left(N - N_Z\right)\right]^c}{c!} \\ &\times \prod_{x_i \in Z} \frac{pN_{x_i}}{(pN_Z + q(N - N_Z))} \prod_{x_i \notin Z} \frac{qN_{x_i}}{(pN_Z + q(N - N_Z))} \\ &= \frac{e^{-pN_Z - q(N-N_Z)}}{c!} p^{c_z} q^{(C-c_z)} \prod_{x_i} N_{x_i} \end{aligned} \tag{2.8}$$

A verossimilhança sob a hipótese nula é dada por.

$$L_0 = \sup_p \frac{e^{-pN} p^C}{C!} \prod_{x_i} N_{x_i} = \frac{e^{-C}}{C!} \left(\frac{C}{N}\right)^C \prod_{x_i} N_{x_i} \tag{2.9}$$

A equação (2.8) é maximizada condicionalmente a Z nos pontos $p = c_Z/N_Z$ e $q = (C - c_z)/(N - N_Z)$. Portanto,

$$L(Z) = \begin{cases} (\mathrm{e}^C/C!)(c_Z/N_Z)^{c_z}\left(\left[C - c_z\right]/\left[N - N_Z\right]\right)^{(C-c_z)} \prod_{x_i} N_{x_i} & \text{se } \frac{c_Z}{N_Z} > \frac{C-c_Z}{N-N_Z} \\ \left(\mathrm{e}^{-C}/C!\right)\left(C/N\right)^C \prod_{x_i} N_{x_i} & \text{c.c} \end{cases} \tag{2.10}$$

A estatística Scan baseada no modelo Poisson, como em (2.6), é a razão entre as verossimilhanças definidas.

$$\lambda = \frac{\sup\limits_{Z\in\mathcal{Z}} L(Z)}{L_0} = \sup_{Z\in\mathcal{Z}} \frac{\left(\frac{c_Z}{N_Z}\right)^{c_Z}\left(\frac{C-c_Z}{N-N_Z}\right)^{C-c_Z}}{\left(\frac{C}{N}\right)^{C}} I\left(\frac{c_Z}{N_Z} > \frac{C-c_Z}{N-N_Z}\right). \tag{2.11}$$

A funçao $I()$ é a função indicadora. Outra maneira de representar (2.11) é através da medida de risco relativo. Ao dividir-se por C tanto o numerador quanto o denominador da razão de verossimilhanças em (2.11) e utilizando relação definida em (2.1), obtem-se

$$\lambda = \sup_{Z\in\mathcal{Z}} \left(\frac{c_Z}{\mu_Z}\right)^{c_Z}\left(\frac{C-c_z}{C-\mu_Z}\right)^{C-c_Z} I(c_z > \mu_Z). \tag{2.12}$$

Os testes baseados em (2.6) ou em (2.12) possuem propriedades importantes que serão descritas a seguir. Métodos para encontrar $\hat{Z}$, tal que (2.6) ou (2.12) seja maximizado, e testar a hipótese $H_1 : p > q$ também serão descritos em tópicos futuros.

2.4 Propriedades da Estatística Scan

O principal ganho que se teve na estatística Scan de Kulldorff, em relação aos métodos de detecção de *cluster* espacial vigentes, é o fato de que tanto a detecção do *cluster* mais provável quanto o teste de significância são obtidos através de um só procedimento. Tal característica é descrita mais formalmente no seguinte teorema. Considere $D = \left\{x_{[j]}, j = 1, \cdots, C\right\}$ o conjunto de coordenadas de casos observados no espaço G, e $\hat{Z}$ o *cluster* mais provável. Seja $D' = \left\{x'_{[j]}, j = 1, \cdots, C\right\}$ um conjunto alternativo de com coordenadas diferentes.

Teorema 1. *Se a hipótese nula é rejeitada sob D, então ela também será rejeitada para todo D' tal que $x'_{[j]} = x_{[j]}, \forall x_j \in \hat{Z}$.*

Em outras palavras, dado que o *cluster* mais provável $\hat{Z}$ foi determinado e a hipótese H_0 rejeitada, qualquer alteração na localização x_j dos casos fora de $\hat{Z}$, mantidas fixas as coordenadas dos casos $x_j \in \hat{Z}$, também resultará na rejeição de H_0.

Outra questão de suma importância diz respeito ao poder de teste da estatística scan. Kulldorff chama a atenção para o fato de que é difícil encontrar um teste uniformemente mais poderoso para o caso de *scan* espacial. Para essa abordagem, é necessário apresentar a definição de um teste "individualmente mais poderoso".

Um teste individualmente mais poderoso resulta da decomposição da Região Crítica em subconjuntos distintos. Suponha que o espaço Θ de parâmetros seja particionado em subconjuntos $\{\Theta_j\}$ disjuntos. Suponha também que a região crítica RC também seja particionada em $\{RC_j\}$, tal que $\cup RC_j = RC$ com os mesmos índices da partição $\{\Theta_j\}$. Considere também uma região crítica alternativa $RC' = \cup RC'_j$.

Definição 2. *Para um particular nível de significância α, um teste é dito* individualmente mais poderoso *com respeito a uma partição $\{\Theta_j\}$ e uma partição $\{RC_j\}$ se, para cada Θ_k não existem conjuntos RC' com partição $\left\{RC'_j\right\}$ tais que*

1. *$RC_j = RC'_j$ para todo $j \neq k$;*
2. *$P(\omega \in RC') = \alpha$;*
3. *$P(\omega \in RC'_k | (Z, p, q)) > P(\omega \in RC_k | (Z, p, q))$ para qualquer $(Z, p, q) \in \Theta_k$.*

De uma maneira mais direta, um teste é individualmente mais poderoso se, ao fixarmos a região crítica RC exceto para uma partição RC_k, o teste será uniformemente mais poderoso com relação a todas as escolhas restantes da região crítica e com relação a todos os parâmetros da partição Θ_k de $Theta$. Tal propriedade é importante para o caso de hipóteses alternativas compostas quando se deseja saber qual parte de Θ causa a rejeição, o que é justamente o caso da detecção de *cluster* espacial.

Teorema 2. *O teste baseado em λ (tanto para o modelo Bernoulli quanto para o modelo Poisson) é individualmente mais poderoso com respeito às partições $\{\Theta_Z\}$ e $\{RC_Z\}$.*

Os teoremas anteriores foram enunciados e demonstrados em [Kulldorff, 1997].

2.5 Métodos de detecção e Simulações de Monte Carlo

Como dito nas seções anteriores, a detecção de um *cluster* envolve a obtenção da zona $\hat{Z}$, entre todas as regiões $Z \in \mathcal{Z}$, para determinar a estatística λ definida em (2.6) ou (2.12). A idéia é obter, para cada zona Z, a razão de verossimilhanças $LR(Z) = L(Z)/L_0$ ou, equivalentemente, o logarítimo da razão, $LLR(Z) = \log(LR(Z))$. A zona correspondente ao maior valor de $LLR(Z)$ será $\hat{Z}$ e $\lambda = LR(\hat{Z})$.

Tal procedimento revela-se evidentemente impraticável. Quando o espaço G é definido por n regiões, como descrito no início do Capítulo, o conjunto $\mathcal{Z}$ será finito, porém terá 2^n elementos (o número de subconjuntos de G). Para contornar esse impasse, a obtenção de $\hat{Z}$ pode ser realizada através da redução do espaço $\mathcal{Z}$ ou através de métodos estocásticos. A primeira alternativa é baseada na definição de "janelas" que podem ter diferentes formatos. Cada tipo diferente de janela define uma classe reduzida $\mathcal{Z}'$. Portanto, $\mathcal{Z}'$ pode ser uma coleção de, por exemplo:

1. Todos os subconjuntos circulares de G;
2. Todos os retângulos de formato e tamanho fixos;
3. Todos os subconjuntos elipticos de G;
4. Quando os dados são agregados, todos os subconjuntos definidos por regiões conexas contendo no máximo K regiões;
5. Em caso de *cluster* espaço-temporal, todos os subconjuntos cilíndricos.

O caso mais simples é o "Scan Circular", baseado em janelas circulares. Considere o caso de dados agregados[2]. Sejam $x_1, \cdots, x_n$ os centróides de cada uma das n regiões e considere d_{ij} a distância entre x_i e x_j. Para cada região R_k, define-se círculos concêntricos de raio r variável cujo centro é x_k. Para cada valor de r, uma nova zona é definida pelas regiões $R_{l_1}, \cdots, R_{l_s}$, com $s < n$ e tal que $d_{kl_1} \leq d_{kl_2} \leq \cdots \leq d_{kl_s} \leq r$. Calcula-se $LLR(Z)$ para cada zona e o processo é repetido para outros valores de k.[3]

[2]Métodos para detecção de *cluster* em dados pontuais serão descritos adiante.

[3]Pode-se definir um número máximo K de regiões dentro de um *cluster*, e nesse caso $s < K$.

Tango [Tango & Takahashi, 2005] define um método de detecção baseando-se em janelas de formato irregular. O procedimento utiliza um número K que representa o número máximo de regiões dentro de um *cluster*. Para cada região k define-se uma janela circular Z_k contendo os $(K-1)$ vizinhos mais próximos. Em seguida, várias janelas definidas por regiões conexas são definidas, de modo que todas estejam contidas em Z_k, e a estatística $LLR(Z)$ é obtida para cada uma delas. O procedimento se repete para outros valores de k.

Outros métodos baseados em diferentes formatos pré-especificados de janelas foram desenvolvidos. Um exemplo é encontrado em [Kulldorff *et al.*, 2006] onde é proposta a detecção de *clusters* através de janélas elípticas.

Com $\hat{Z}$ aproximado por um desses métodos, procede-se com o teste de hipóteses. Outro impasse surge, entretanto, do fato de que a distribuição de λ é muito difícil, se não impossível, de ser definida. O método usual de obter a distribuição de λ sob H_0 consiste em realizar simulações de Monte Carlo baseadas no método de Dwass [Dwass, 1956]. Como os valores sob a hipótese nula μ_{Ri} são conhecidos, obtem-se replicações aleatórias da distribuição de casos entre as regiões R_i, condicionadas ao número total C. Em cada replicação, um dos procedimentos acima descritos é executado e a estatística λ é calculada. Com 10.000 simulações baseadas na hipótese nula e considerando um nível de significância de 5%, a hipótese nula será rejeitada se o valor de $\lambda = LR(\hat{Z})$ estiver entre os 500 maiores valores obtidos pelas simulações.

Capítulo 3

Medidas de intensidade em dados agregados

3.1 Introdução

Após a obtenção do "*cluster* mais provável" $\hat{Z}$, isto é, a zona Z com maior LLR, bem como a obtenção do seu nível de significância via Monte Carlo, é interessante prosseguir com a invetigação da incerteza inerente ao *cluster* detectado.

As primeiras tentativas de abordagem da incerteza em torno da detecção do *cluster* mais provável $\hat{Z}$ foram baseadas na observação dos "*clusters* secundários", que são outras zonas $Z \in Z$ cujas estatísticas $LLR(Z)$ estão entre os valores significativos obtidos via simulação de Monte Carlo. Essas zonas secundárias, além de proverem uma visualização da incerteza envolvida (no caso de *clusters* sobrepostos), permitem que prováveis focos com *excesso de ocorrência* em outros pontos do espaço G sejam investigados (*clusters* disjuntos).

Outro método de investigação de incertezas, baseado em *clusters* secundários e nos riscos relativos, foi proposto em [Boscoe *et al.*, 2003]. Após a simualção de Monte Carlo, as zonas cujas razões de verossimilhança estão acima do quantil de 95% são estratificados em 10 níveis da medida de risco relativo definido em (2.1). Em cada nível, a zona com maior $LLR(Z)$ é mapeada – e outros secundários, desde que não sobreponham a região já detectada. O procedimento é repetido e resulta em um mapeamento de 10 classes de zonas (10 cores diferentes) cujos riscos relativos são

significativos.

Em [Chen *et al.*, 2008], propõe-se a repetição do processo de detecção para S valores diferentes do número máximo K de regiões dentro do cluster. Em geral $S = 8$ e K varia entre 5% e 49% do número total n de regiões no mapa. Realizadas as S detecções, a *confiança* de uma área i é dada pelo número de vezes em que ela aparece no *cluster* $\hat{Z}$ dividido por S.

Os métodos citados abordam a incerteza através da divisão do mapa em "dentro" e "fora" dos *clusters* mais prováveis. Entretanto, é essencial verificar a importância individual de cada área do mapa, refletindo seu potencial de compor um *cluster*, estando ela inserida ou não no *cluster* $\hat{Z}$. Deve-se levar em conta que o número de casos em cada área também está sujeito a variações. Como citado no começo deste trabalho, esforços já foram feitos em [Rosychuk, 2005], onde se propõe a estimativa de intervalos de confiança do risco relativo para cada área, os quais são comparados com os riscos contidos no *cluster* mais provável. Tal método, entretanto, não permite uma visualização ampla da incerteza envolvida.

A *medida de intensidade* proposta recentemente em [Oliveira *et al.*, 2011] mede a importância individual de cada área. O método é semelhante ao teste não-paramétrico da estatística λ, mas aqui as várias simulações de Monte Carlo são baseadas não na hipótese nula, mas sim no número observado de casos. O resultado é uma medida para cada área. Em cada simulação, o $LLR(Z)$ do *cluster* mais provável é obtido. A medida individual dependerá do maior $LLR(Z)$ observado entre os *clusters* mais prováveis contendo essa área.

3.2 Método

Sejam $R_1, \cdots, R_n$ as n áreas que compõem a região G que contém N elementos e C casos. Considere que $c_1, \cdots, c_n$ são as quantidades observadas de casos em cada região, com $\sum_{i=1}^{n} c_i = C$. A medida de intensidade será baseada em m replicações de Monte Carlo baseadas em uma distribuição que tenha como valor esperado os valores c_i observados. Mais precisamente, cada replicação j ($j = 1, \cdots, m$) da simulação irá gerar uma amostra $v_j = (s_1, \cdots, s_n)$ do vetor aleatório $V = (C_1, \cdots, C_n)$, $\sum_{i=1}^{n} s_i = C$, que segue uma distribuição multinomial com pro-

babilidades $(c_1/C, \cdots, c_n/C)$ – de modo que $E(V) = (c_1, \cdots, c_n)$. Em seguida, o algoritmo de detecção de *cluster* é aplicado, gerando o valor LLR_j correspondente ao *cluster* mais provável j (MLC_j). Após m replicações, os valores $LLR_1, \cdots, LLR_n$ são ordenados, formando o conjunto $\left\{LLR_{(1)}, \cdots, LLR_{(m)}\right\}$ correspondente aos *clusters* mais prováveis $\left\{MLC_{(1)}, \cdots, MLC_{(m)}\right\}$. Defindo uma função $f : \{1, \cdots, m\} \to \mathbb{R}$ por $f(j) = LLR_{(j)}$, define-se a *medida de intensidade* da região R_i por

$$q(R_i) = \frac{1}{m} \underset{\left\{1 \leq j \leq m, R_i \in MLC_{(j)}\right\}}{\arg\max} f(j), i = 1, \cdots, n \tag{3.1}$$

Se a área R_i não pertence a nenhum dos conjuntos $MLC_{(1)}, \cdots, MLC_{(n)}$, então $q(R_i) = 0$.

A medida $q(R_i)$ deve ser interpretada como a importância relativa da região R_i como parte da anomalida detectada na região. Esse conceito é mais informativo do que a simples divisão entre *cluster* e não-*cluster*. Grandes valores de $q(R_i)$ $(0 \leq q(R_i) \leq 1)$ significam uma influência maior da i-ésima região na existência do cluster.

Tal medida também está diretamente relacionada à incerteza envolvida no processo de detecção. De fato, quanto maiores as intensidades em regiões pertencentes a uma particular zona Z, maior é a certeza de que essa área configura um *cluster* ([Oliveira *et al.*, 2011]).

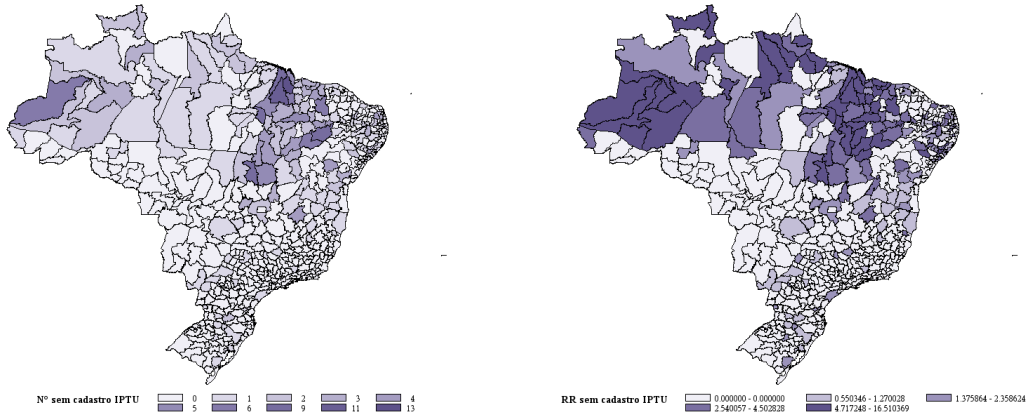
Exemplo 1. *Investigando concentração de municípios sem cadastro de IPTU.*

O Instituto Brasileiro de Geografia e Estatística (IBGE) realizou em 2009 um levantamento a nível municipal, objetivando traçar o perfil dos municípios brasileiros [IBGE, 2010]. As variáveis observadas são categorizadas em vários temas e subtemas.

Uma das variáveis, de relevante interesse nas Finanças Públicas, é o fato de o município ter ou não um cadastro imobiliário. Cerca de 6% dos municípios não possuem esse recurso, e seria interessante investigar, portanto, se existe um *cluster* de municípios sem cadastro imobiliário.

Para implementar essa investigação, os municípios foram agrupados em microrregiões (554, no total). A cada microrregião está associado o número de municípios sem cadastro imobiliário. Essa frequência, bem como os riscos relativos, seguem na Figura 3.1.

O *cluster* detectado é significativo a 95%, como mostra a Tabela 3.1. A detecção foi feita através de janelas circulares contendo até 30% do total de microrregiões. O conjunto possui 102 regiões, correspondendo a 939 municípios, dos quais 203 não possuem cadastro imobiliário. A região detectada e as medidas de intensidade $q(R_i)$ por microrregião estão representadas na Figura 3.2.



(a) Número de casos (b) Risco relativo

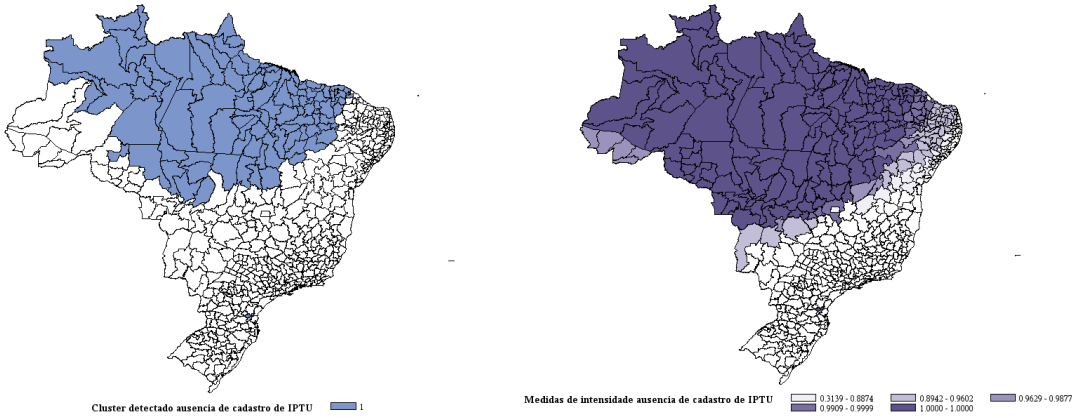
Figura 3.1: Municípios sem cadastro imobiliário, por microrregião

Tabela 3.1: Ausência de cadastro imobiliário: $LLR(\hat{Z})$ observado e quantis de 90%,95% e 99% obtidos das simulações de Monte Carlo

$LLR(\hat{Z})$	p_{99}	p_{95}	p_{90}
159.48	10.41	8.71	7.91

Exemplo 2. *Investigando concentração de municípios com RREOs faltantes.*

A lei complementar n°101/2000, a Lei de Responsabilidade Fiscal (LRF), estabelece normas orientadoras das finanças públicas para todas as esferas do governo: Federal, Estadual e Municipal [Orair *et al.*, 2011]. Essa lei incumbe o poder executivo das três esferas da responsabilidade de elaboração e publicação bimestral das contas públicas. Essas informações são publicadas através dos Relatórios Resumidos de Execução Orçamentária (RREOs). Os RREOs são divulgados no portal da Secretaria do Tesouro Nacional (STN) em STN, em formato *pdf*.



(a) Cluster detectado

(b) Medidas de intensidade

Figura 3.2: Cluster de ausência de cadastro imobiliário e medidas de intensidade, por microrregião

Apesar das normas estabelecidas na LRF, nem todos os municípios divulgam tais informações nos prazos estabelecidos. Na realidade, 42% dos 5.565 municípios deixaram de divulgar RREOs através do STN[1] em pelo menos 1 bimestre entre janeiro de 2006 e dezembro de 2010. Seria interessante verificar, portanto, se existe um *cluster* de municípios com RREOs faltantes.

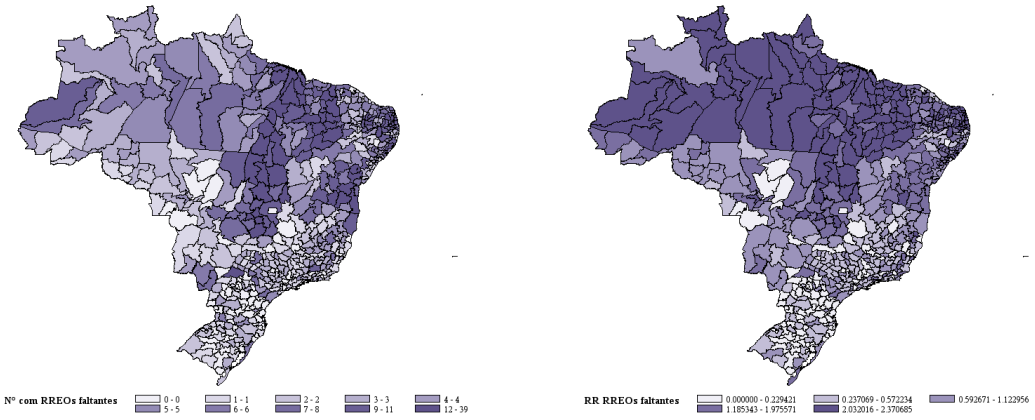
Como no exemplo anterior, os municípios foram agregados em microrregiões. A detecção do *cluster* também foi feita nos moldes do exemplo acima. O *cluster* detectado é significativo a 99% de confiança (Tabela 3.2), possui 163 microrregiões contendo 1.499 municípios, dos quais 1.200 possuem RREOs faltantes. O *cluster* detectado e as medidas de intensidade seguem na Figura 3.4.

Tabela 3.2: RREOs faltantes: $LLR(\hat{Z})$ observado e quantis de 90%,95% e 99% obtidos das simulações de Monte Carlo

$LLR(\hat{Z})$	p_{99}	p_{95}	p_{90}
307.65	57.14	55.16	52.71

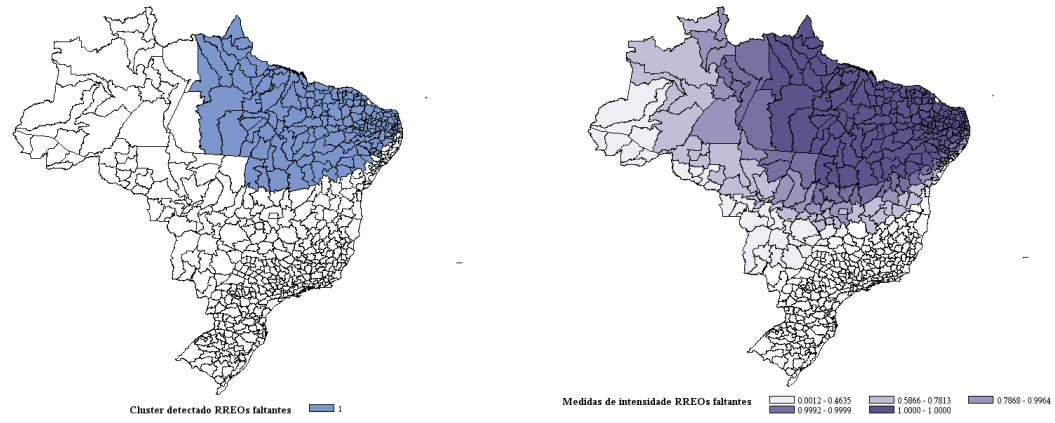
Os exemplos acima ilustram a importância do delineamento de incertezas na detecção de *clusters* espaciais. A simples detecção dos *clusters* sugere maiores riscos em

[1]Um município pode divulgar RREOs em outras fontes que não o STN. Entretanto, para simplificação, este exemplo estará restrito apenas a essa fonte de dados.



(a) Número de municípios (b) Risco relativo

Figura 3.3: Municípios com RREOs faltantes, por microrregião



(a) Cluster detectado (b) Medidas de intensidade

Figura 3.4: Cluster de RREOs faltantes e medidas de intensidade, por microrregião

uma área envolvendo parte das regiões Norte, Nordeste e norte do Centro-Oeste, no caso da ausência de cadastros imobiliários, enquanto que a investigação de municípios com RREOs faltantes resulta em uma área de maior risco mais concentrada na região Nordeste.

Ao se avaliar as incertezas através das medidas definidas em (3.1), entretanto, é possível notar que essa área de risco, nos dois exemplos, pode englobar toda a região Norte e Nordeste, e uma parcela maior da região Centro-Oeste. Ao mesmo tempo, as medidas ainda destacam as regiões dos clusters detectados, mantendo a conclusão inicial de que o risco de um município deixar de apresentar um RREO é maior em uma região mais concentrada no Nordeste.

Outro fato importante das medidas de de intensidade, que não foi visualizado nos exemplos anteriores, é a sua utilidade na detecção de *clusters* secundários. Tal fato será ilustrado adiante nas simulações em dados pontuais. Entretanto, a sensibilidade de tal medida à existência de *clusters* múltiplos em dados agregados é discutida e ilustrada em [Oliveira *et al.*, 2011].

Capítulo 4

Detecção de Clusters em dados pontuais

4.1 Introdução

O método de Kulldorff através do scan circular (ou de qualquer outro formato) aplicado em dados pontuais segue a mesma lógica no caso de dados agregados. Em vez de áreas $R_1, \cdots, R_n$ com centróides $x_1, \cdots, x_n$, considera-se os N pontos com coordenadas $x_1, \cdots, x_N$, $x_i \in \mathbb{R}^2$. A cada ponto é associado o valor C_i que pode ser 1 ou 0 se essa observação for caso ou controle, respectivamente. Janelas em algum formato específico são centradas em cada ponto e são expandidas, gerando assim zonas candidatas a cluster às quais são associadas as estatísticas $LLR(Z)$, e o cluster mais provável é a zona $\hat{Z}$ que maximiza $LLR(Z)$. Sob a hipótese nula, cada ponto tem valor esperado $\mu_{x_i} = p$. As simulações de Monte Carlo para o cálculo do p-valor são baseadas em valores C_i gerados sob a hipótese nula, considerando que $C_i \sim Bern(p)$ para cada ponto e de modo que $\sum_{i=1}^{N} C_i = C$.

Há uma corrente metodológica abordando as limitações, tanto computacionais quanto com relação à *sensibilidade* – ou seja, a capacidade de se detectar um cluster verdadeiro – do método baseado em "janelas" de qualquer formato. O aumento significativo do tempo gasto para a varredura é evidente já que, em vez de $n < N$ regiões nas quais as janelas estão centradas, há N pontos, e então o número de candidatos a cluster aumenta consideravelmente [Cucala *et al.*, 2009]. O problema

de sensibilidade talvez se deva ao fato de que o método deteca apenas clusters em formato regular pré-determinado (circular, elíptico, etc.). Metodologias usuais baseados no *scan* espacial para detecção de clusters em formatos irregulares, como em [Tango & Takahashi, 2005], trabalham apenas em dados agregados. Outro método utilizado na detecção de cluster irregular é baseado em uma *árvore de abrangência mínima* (Minimum Spanning Tree ou MST), e foi proposto originalmente em [Assunção *et al.*, 2006], mas também foi desenvolvido no contexto de dados agregados.

Em dados pontuais, foram propostos recentemente métodos de detecção baseados em *grafos* e *subgrafos* ligando os pontos correspondentes aos casos. Nesse contexto, o conjunto de candidatos a cluster é definido em termos da proximidade (ou densidade local) das ocorrências. Entretano, faz-se necessário corrigir a heterogeneidade existente na distribuição geográfica dos pontos [Duczmal *et al.*, 2011]. Sem essa correção, um raio esférico não é adequado para estimar as densidades populacionais em todas as regiões. Em [Wieland *et al.*, 2007] define-se uma adaptação do método baseado no MST, denominado *Euclidean Minimum Spanning Tree* (EMST). Essa adaptação ocorre através de um *cartograma*, um mapa distorcido baseado no *Diagrama de Voronoi*, de modo que uma novo espaço geográfico seja definido e tenha densidade populacional constante. Esse método não baseia a significância do teste na estatística scan de Kulldorff, mas sim em um *peso total* que depende da densidade local dos casos.

Outro método é baseado não no MST, mas em um grafo completo ligando as coordenadas dos casos [Cucala *et al.*, 2009]. Os candidados a cluster são subgrafos gerados por diferentes níveis δ de distâncias (conectando assim apenas pontos que não estejam distantes entre si além de δ). Em torno de cada região candidata é definida uma região de vizinhança tal que os pontos nela contidos não estejam distantes do cluster candidato além do nível considerado. Logo após, utiliza-se uma das *medidas de concentração* dos casos dentro dessas regiões – entre as quais está a razão de verossimilhanças que define a estatística de Kulldorff. Tal método, entretanto, não atenta para o já referido problema das diferenças de heterogeneidade da população como um todo.

Mais recentemente foi apresentado um método semelhante ao EMST, o *Voronoi Based Scan* (VBScan) [Duczmal *et al.*, 2011]. Em vez das distâncias euclidianas

utilizadas no EMST, utiliza-se o conceito de *Distância de Voronoi* que, como será detalhado nos próximos tópicos, surge como uma solução alternativa no tratamento das densidades locais heterogêneas e dispensa a distorção do mapa original.

Os tópicos seguintes farão uma breve descrição dos métodos baseados no MST para detecção de clusters.

4.2 Algumas definições

4.2.1 Diagrama de Voronoi

Ambos os métodos baseados no MST aqui descritos farão uso do *Diagrama de Voronoi*. Considere o conjunto $E = \{x_1, \cdots, x_N\}$, $x_i \in \mathbb{R}^2$. A cada ponto x_i é definida uma célula $\upsilon(i)$ que consiste no conjunto de pontos em $\mathbb{R}^2$ que são mais próximos de x_i do que de qualquer outro ponto em E. Mais especificamente, $\upsilon(i)$ é definido como

$$\upsilon(i) = \left\{x \in \mathbb{R}^2 | d(x_i, x) \leq d(x_j, x), \forall j \neq i\right\}, \tag{4.1}$$

onde $d(x_i, x_j) = \|x_i - x_j\|^2$. O conjunto de céluluas $\upsilon(i), i = 1, \cdots, N$ constitui o que se chama de *Diagrama de Voronoi* e a célula $\upsilon(i)$ é a *Célula de Voronoi*([Mount, 2012]).

As células de Voronoi são construídas da seguinte maneira: dados os pontos x_i e x_j, traça-se uma reta perperdicular ao segmento que os conecta e que passe exatamente no ponto médio desse segmento. O mesmo é feito para todo par de pontos até que o diagrama de Voronoi esteja completo[1]. A Figura 4.1 ilustra um conjunto de 20 pontos aleatoriamente distribuidos no intervalo bidimensional $[0, 1]X[0, 1]$ e o diagrama de Voronoi correspondente.

4.2.2 MST

Define-se um grafo $Gr(E, L)$ a partir do conjunto E e um conjnto L. Os elementos de E são chamados de vértices, e $L = \{(x_i, x_j) | x_i, x_j \in E\}$ representa um conjunto de arestas interligando pares de pontos pertencentes a E. Tal definição, entretanto,

[1]Um algoritmo eficiente para a construção do Diagrama de Voronoi é detalhado em [Mount, 2012]

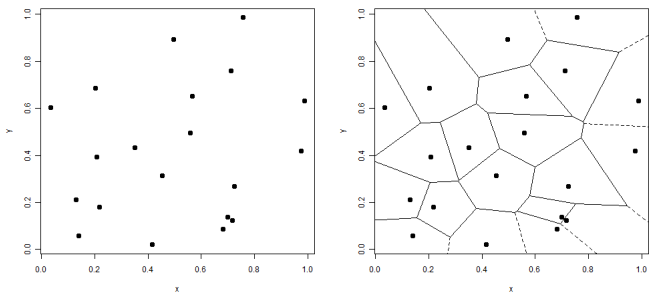


Figura 4.1: Pontos aleatoriamente distribuídos (esquerda) e diagrama de Voronoi do conjunto (direita)

é de interpretação bastante ampla e abstrata para o estudo de detecção de clusters. A maneira usual de abordar cluster em dados pontuais utilizando grafos é através de *subgrafos*.

Uma *árvore* t é um subgrafo de Gr conectado e sem ciclos, ou seja, todos os vértices nele contidos estão conectados a pelo menos um vértice do mesmo subgrafo, sem que uma região "fechada" se forme. Uma *árvore de abrangência* T ("spanning tree") é uma árvore que contém todos os vértices de E. A cada árvore de abrangência T possível é associado uma medida de "peso total", $\omega(T)$, onde

$$\omega(T) = \sum_{l \in L(T)} \omega(l) \tag{4.2}$$

onde $L(T) \in L$ é o conjunto de limites da árvore T e $\omega(l)$ é a distância entre os dois pontos conectados por l.

Com esses termos em mente, pode-ser definir um MST.

Definição 3. *Uma* árvore de abrangência mínima ("Minimum Spanning Tree",MST) *é uma árvore T tal que $\omega(T)$ é mínimo.*

A Figura 4.2 ilustra a construção de um MST baseado nos mesmos 20 pontos gerados aleatoriamente no exemplo anterior. Note que a linha mais espessa representando o MST sobrepõe o grafo regular[2], ilustrando que o MST é um subconjunto de

[2]Um *grafo regular* é um subgrafo cujos vértices possuem o mesmo número de vizinhos, isto é, cada vértice é ligado ao mesmo número de vértices. Esse tipo de grafo produz regiões delimitadas. Se não existem pontos colineares, um grafo regular pode ser obtido através da *Triangulação de Delaunay.*

um grafo.

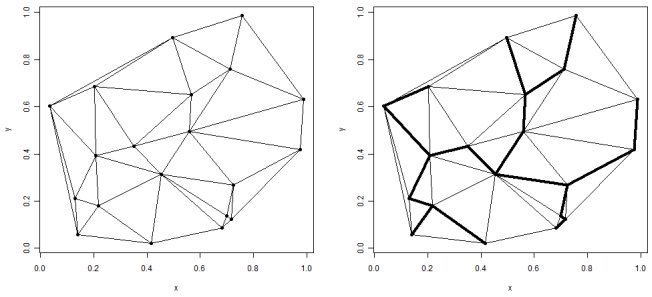


Figura 4.2: Grafo Regular dos pontos (esquerda) e MST do conjunto (direita)

4.2.3 Definição de prováveis clusters

Em [Wieland *et al.*, 2007] é adotada uma definição específica de *cluster* potencial em dados pontuais. Diferentemente da Definição 1 adotada por Kulldorff, a existência de um *cluster* está mais associada à proximidade dos casos entre si: um conjunto de casos S é um cluster em potencial se satisfaz a propriedade de que cada subconjunto de S é mais próximo de um ponto dentro de S do que de outro ponto fora de S.

Formalmente, a distância entre dois conjuntos A e B é definida por

$$d(A,B) = \begin{cases} \min\limits_{\{a \in A, b \in B\}} d(a,b) & \text{se } A \neq \oslash, B \neq \oslash \\ \infty & \text{c.c} \end{cases}$$

A distância interna de um conjunto não vazio S é definida pela distância máxima entre qualquer par de subconjuntos de S, ou seja

$$d(S) = \max_{\{A \subseteq S, B \subseteq S, A \cup B = S\}} d(A,B) \tag{4.3}$$

Com essas distâncias formalmente definidas, é possível agora uma definição formal de um cluster potencial em dados pontuais.

Definição 4. *Considere um espaço geográfico G no qual estão definidas observações. Seja D o conjunto de todos os casos. Um subconjunto $S \subset D$ é um* cluster pontencial *se $d(S) < d(S, D - S)$.*

Assim, por definição, o próprio conjunto D e também cada ponto isolado representam clusters potenciais.

4.3 Métodos baseados no MST

As definições anteriores não deixam claro como localizar outros clusters além dos clusters triviais (isto é, o conjunto de todos os casos e cada ponto individualmente). Uma das maneiras de definí-los é através de uma árvore de abrangência T que seja o MST de D. Cada subgrafo de T define um cluster potencial. Entretanto, mesmo um número razoavelmente pequeno de casos define uma quantidade muito grande de subgrafos possíveis.

Para contornar esse problema, propôs-se um método baseado na remoção iterativa dos limites de $L(T)$. Inicialmente, o limite $l \in L(T)$ com maior $\omega(l)$ é removido, resultando em duas árvores de abrangência disjuntas (dois clusters potenciais). Na etapa seguinte, o limite com maior peso entre os limites restantes é retirado, definido-se mais dois clusters potenciais. O processo se repete até que se tenha o *conjunto trivial de clusters prováveis* (cada ponto isolado). No total, há apenas $2C - 1$ candidatos possíveis. Em [Wieland *et al.*, 2007] há uma demonstração de que esse procedimento define o conjunto total de clusters potenciais. A Figura 4.3 ilustra esse processo de definição dos clusters candidatos considerando os mesmos 20 pontos dos exemplos anteriores.

4.3.1 EMST

Em [Wieland *et al.*, 2007], todo um processo de homogeneização das densidades locais é executado. Primeiro, um diagrama de Voronoi é construído apenas em torno dos pontos representando os controles, de modo que cada caso esteja localizado em uma das células do diagrama. Em seguida, uma transformação não-linear desse diagrama (o *cartograma*) é realizada de modo que cada célula de Voronoi tenha a mesma área. Os casos são então aleatoriamente alocados em suas respectivas células do diagrama original. A Figura 4.4 retirada de [Wieland *et al.*, 2007] ilustra essa construção.

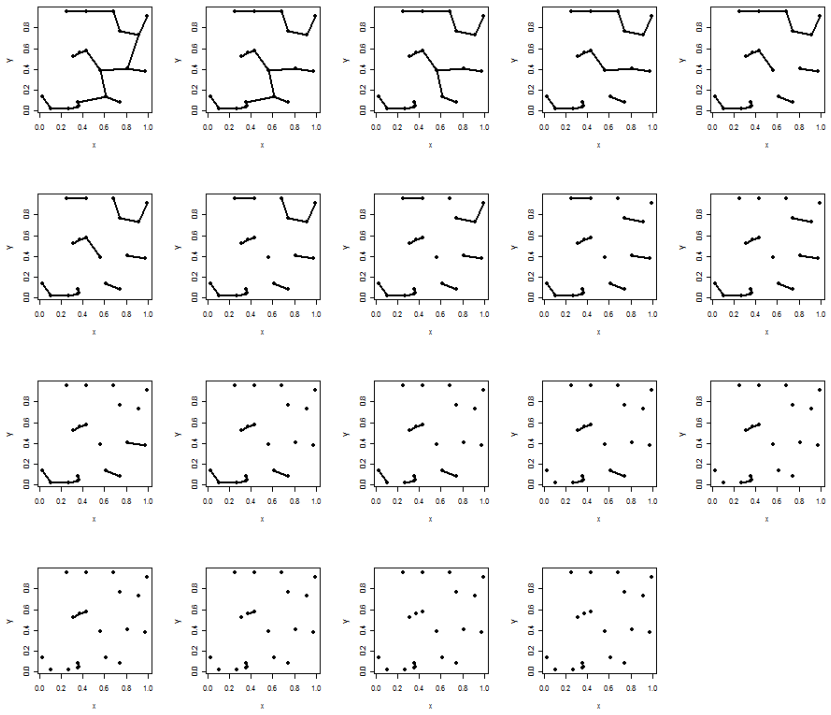


Figura 4.3: Remoção iterativa dos maiores limites do MST

Como mencionado anteriormente, a significância dos clusters é computada baseando-se na proximidade dos pontos dentro de cada cluster pontencial. Sob a hipótese H_0, os casos são uniformemente distribuídos no cartograma. Consiere Z um cluster pontencial gerado sob H_0 e S um cluster potencial observado, ambas definidas de acordo com a Definição 4. O p-valor P_S associado ao cluster S é dado por

$$P_S = P\left(\omega(Z) < \omega(S) | \#Z = \#S\right), \tag{4.4}$$

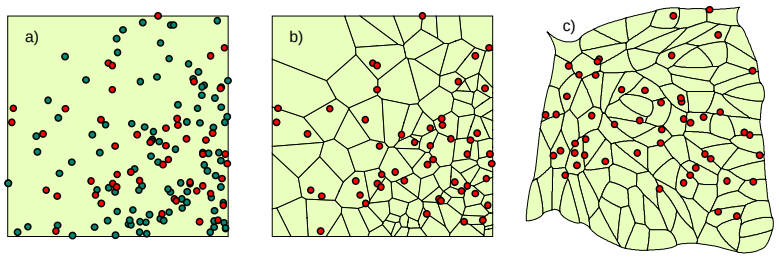


Figura 4.4: Transformação do mapa para homogeneização das densidades locais. Fonte: [Wieland *et al.*, 2007]

onde $\#A$ representa o número de elementos em um conjunto A.

Espera-se que se um conjunto de casos representa um cluster, então é natural imaginar que é mais provável eles estarem mais próximos entre si do que um cluster gerado sob H_0 com o mesmo número de casos. P_S representa, então, o p-valor do cluster candidato S condicionado ao número de casos nele contidos. Define-se então a estatística P como sendo o mínimo P_S observado no conjunto de clusters potenciais não-triviais contendo no máximo metade dos casos. Técnicas de Monte Carlo são usadas para ajustar P_S como função de $\omega(S)$ a uma distribuição Normal para cada valor de $\#(S)$. Como no caso da distribuição de λ no método de Kulldorff, a distribuição de P sob H_0 é obtida também através de técnicas de Monte Carlo. O cluster mais significante é reportado.

No mesmo trabalho em que se propôs tal método, realizou-se um conjunto de simulações para compará-lo aos procedimentos baseados no Scan Espacial. Essa comparação se deu por medidas de *sensibilidade* (Sens), que mede a fração do cluster verdadeiro que foi detectado, e a *acurácia* (PPV, ou Valor Predito Positivo), onde

$$Sens = \frac{\sum_{i=1}^{N} I(x_i \in ClusterDetectado \cap ClusterReal)}{\sum_{i=1}^{N} I(x_i \in ClusterReal)} \tag{4.5}$$

$$PPV = \frac{\sum_{i=1}^{N} I(x_i \in ClusterDetectado \cap ClusterReal)}{\sum_{i=1}^{N} I(x_i \in ClusterDetectado)} \tag{4.6}$$

Após várias simulações, as médias de 4.5 e 4.6 foram tomadas. O métdo EMST revelou-se em média mais sensível que o Scan elíptico, mas perdeu em acurácia, obtendo uma proporção média maior de falsos positivos do que o método usual.

4.3.2 VBScan

Em [Duczmal *et al.*, 2011], propôs-se uma adaptação do método baseado no EMST, o VBScan. O método também utiliza o Diagrama de Voronoi, porém o diagrama é construído não apenas no conjunto de controles, mas sim em todos os pontos observados.

O principal ganho do VBscan com relação ao EMST é a definição da *Distância de Voronoi*. Seja v_{ij} o número de células de Voronoi interceptadas por uma linha reta

para ligar os pontos x_i e x_j, incluindo eles próprios. A *Distância de Voronoi* é dada por $\delta(i,j) = v_{ij} - 1$. Se as células $v(i)$ e $v(j)$ são vizinhas, então $\delta(i,j) = 1$.

O VBScan é baseado no VMST, um MST que minimiza a distância de voronoi total (*peso de Voronoi*) em uma árvore de abrangência. Empates ocorrem com frequência nessa métrica de distâncias. Em casos desse tipo, o limite escolhido para compor o VMST será aquele com maior distância Euclidiana.

O uso da distância de Voronoi dispensa a transformação do espaço G a fim de se homogeneizar as densidades locais. A própria definição de distância de Voronoi corrige essas diferentes densidades. Considere $D = \left\{x_{[j]}, j = 1, \cdots, C\right\}$ o conjunto de coordenadas dos pontos representando os casos. Considere $\mathcal{C}(x_{[j]}, r)$ um círculo de raio r (medido na métrica euclidiana) centrado em $x_{[j]}$ e que a densidade local seja dado por $s_{[j]}$ indivíduos por unidade de área. A população esperada em torno de $\mathcal{C}$ é dada por $s_{[j]}\pi r^2$. Ao substituirmos r por um raio R medido em unidades da distância de Voronoi, então $s_{[j]} = 1$ (já que cada unidade da distância de Voronoi corresponde a um indivíduo), e a população local é dada por πR^2. Em outras palavras, ao se adotar a métrica de Voronoi a população total esperada dentro de um círculo $\mathcal{C}(x_{[j]}, R)$ depende apenas do comprimento de raio R, não da densidade local nas proximidades do ponto $x_{[j]}$.

De maneira semelhante ao método apresentado em [Cucala *et al.*, 2009], o método VBScan define "regiões de influência" em torno dos clusters potenciais. A seguinte proposição é dada em [Duczmal *et al.*, 2011].

Proposição 1. *Considere um conjunto D de casos e seu respectivo VMST dado por T. Seja T_S um subgrafo de T cujos vértices compõem um subconjunto S de D, e seja $f(x)$ a densidade local no ponto x. Para cada caso $x_{[j]} \in S$, seja $\omega_{[j]}$ o peso mínimo dos limites que incidem em $x_{[j]}$ no subgrafo T_S. Considere também o conjunto $\mathcal{B} = \cup\mathcal{C}(x_i^c, \omega_i/2)$. A população local em torno de S pode ser aproximada por $\int_{\mathcal{B}} f(x)dx = \frac{1}{4}\sum_{x_{[j]}\in S} \pi\omega_{[j]}^2$.*

Com as regiões de influência definidas e as populações locais dessas regiões aproximadas, procede-se com a remoção iterativa das arestas do VMST, de modo semelhante à ilustração na Figura 4.3. A cada *cluster* candidato é cálculada a estatística Scan de Kulldorff λ, definida em 2.12, levando-se em conta a proposição acima para aproximar

o número total de observações em torno da região de influência de cada candidado S.

É importante notar que, apesar da utilização do VMST, que está relacionado à idéia de *cluster* como um "conjunto de casos próximos entre si" da Definição 4, o método VBScan, ao considerar a aproximação da população local, não abre mão da noção de zona de risco como uma região com "excessos de ocorrência além do esperado", formalizada na Definição 1.

Simulações foram feitas em [Duczmal *et al.*, 2011] com o intuito de comparar o VBScan com as técnicas usuais de scan no espaço-tempo, para a detecção de vários tipos diferentes de cluster espaço-temporal. As comparações também foram baseadas nas medidas 4.5 e 4.6. As simulações mostraram que o VBScan apresentaram maior Sensibilidade e acurácia que o scan elíptico.

É inevitável também comparar o ganho em eficiência e robustez que se tem no VBScan em comparação com o EMST. Algumas das principais vantagens listadas são:

- O VBScan não necessita de uma tranformação do espaço G em que estão localizados os pontos;
- Por não depender dessa transformação, os clusters potenciais são reportados com as coordenadas originais obervadas dos casos;
- Enquanto o EMST depende do conjunto controle apenas no sentido de se definir as regiões dos casos onde estes serão alocados aleatoriamente, o VBScan aproxima o número de controles em torno de um cluster candidato, aproximando-se mais da idéia de cluster como *excesso de ocorrência álém do esperado*;
- No VBScan, a significância do teste depende diretamente do cluster candidato, enquanto o EMST ainda providencia um ajuste pela normal que dependa do cluster S através de $\omega(S)$;
- O VBScan utiliza a estatística de Kulldorff, que possui as propriedades de poder de teste estabelecidas no teorema 2.

Os métodos acima descritos, embora tenham vantagens evidentes, não serão implementados neste trabalho. O foco será dado no delineamento de incertezas a exemplo

do que foi apresentado no Capítulo 3, porém no contexto de obervações pontuais. As ferramentas acima descritas, entretanto, foram extremamente úteis para o desenvolvimento da solução proposta, como será visto no próximo Capítulo.

Capítulo 5

Medidas de intensidade em dados pontuais

5.1 Introdução

Como foi dito no Capítulo 2, o p-valor da estatística λ para dados agregados em n regiões é obtido através de replicações aleatórias do vetor $V = (C_1, \cdots, C_n)$ sob a hipótese nula. Assim, $E(C_i) = \mu_i = N_i \frac{C}{N}, i = 1, \cdots, n$, onde $C = \sum_{i=1}^{n} c_i$ o total de casos observados, e $N = \sum_{i=1}^{n} N_i$ o total de observações. Em outras palavras, o p-valor é calculado por quantis amostrais, calculados em simulações de Monte Carlo de modelos probabilísticos cujas expectâncias são os valores esperados sob a hipótese de não existência de cluster.

Ainda sob a situação de dados agregados, as medidas de intensidade em (3.1) são obtidas através da realização de simulações de Monte Carlo considerando-se não a hipótese nula, mas a própria estrutura de caso-controle observada na base de dados. Quando dispõe-se de dados agregados, os próprios valores obeservados servem como estimativas das expectâncias $E(C_i)$. Ou seja, $\widehat{E}(C_i) = c_i$, sendo c_i o total de casos observados na região i. Portanto, as simulações são realizadas a partir de amostras da variável aleatória C_i tal que $C_i \sim Poi(ci)$ ou $C_i \sim Bin(ci, ci/C)$

Em dados pontuais, as simulações de Monte Carlo para obtenção do p-valor seguem a mesma lógica dos dados agregados. As replicações aleatórias são baseadas em valores gerados de $C_i \sim Bern(p)$, onde $p = C/N$. Entretanto, a obtenção das medidas

de intensidade em dados pontuais, baseando-se em replicações centradas nos valores observados em cada ponto, não faria muito sentido. Ao i-esimo ponto corresponde a variável aleatória $C_i \sim Bern(pi)$. Considerar $\widehat{E}(C_i) = c_i$, como no caso de dados agregados, resultaria em valores simulados iguais a c_i (0 ou 1), com probabilidade 1. O valor observado c_i, somente, é insuficiente para estimar p_i. Por esse motivo, sendo x_i as coordenadas do i-ésimo ponto, a solução apresentada baseia-se na definição de *regiões de vizinhança* em torno dos $[j]$, ou seja, das coordenadas de cada caso.

Os capítulos anteriores apresentaram alguns recursos bastante promissores e largamente implementados em bases de dados pontuais do tipo caso-controle. Dentre eles, o conceito de Árvore Geradora Mínima (*Minimum Spanning Tree, MST*). Como dito anteriormente, as estimativas das probabilidades p_i serão baseadas em vizinhanças. Mais especificamente, considerou-se uma região circular em torno de cada caso, algo como uma região de influência deste. As probabilidades individuais serão inversamente proporcionais ao número de pontos (controles) dentro desta região de influência de um caso. A proposta é construir um MST ligando os casos. O comprimento de uma das arestas incidentes em cada caso determinará o raio da região circular em torno deste ponto. O presente Capítulo visa detalhar esse método, bem como considerar várias maneiras possíveis de escolha dessas arestas.

Outra importante ferramenta apresentada anteriormente é o Diagrama de Voronoi, bastante útil na detecção de *clusters* irregulares através do VBScan. Entretanto, dado o enfoque deste trabalho no delineamento de incertezas, seu uso será restrito a outra utilidade: a definição de uma malha digital que delimite as áreas "dominadas" por cada ponto, as células de Voronoi. Mais especificamente, enquanto no caso de dados agregados algumas grandezas relacionadas a cada região, contínuas ou não, são representadas por diferentes colorações destas áreas, a representação de grandezas relacionadas aos pontos será feita com diferentes colorações das células de Voronoi. Assim serão representadas as medidas de intensidade para dados pontuais.

O presente Capítulo apresentará o método proposto para a estimativa dos $p_i's$. O cálculo das medidas de intensidade com base em simulações de $C_i \sim Bern(\hat{p}_i)$ será ilustrado e avaliado através de dados simulados em vários cenários prováveis de existência de *cluster*.

5.2 Estimativa de probabilidades individuais

Enfatizando o que foi dito anteriormente, as probabilidades estimadas $\hat{p}_i$ serão baseadas na definição de *regiões de vizinhança* em torno dos casos. O procedimento proposto leva em conta as seguintes idéias:

1. Cada região de vizinhança define um cálculo próprio de probabilidades;
2. As probabilidades estimadas serão diretamente proporcionais ao número de casos e inversamente proporcionais ao número de controles dentro das regiões de vizinhança;
3. A probabilidade assoaciada a um controle será inversamente proporcional à distância que o separa do caso que define a região (ou seja, a distância até o centro do círculo);
4. As estimativas de probabilidades em pontos pertencentes a mais de uma região de influência serão determinadas por uma média ponderada das probabilidades estimadas em cada região, de modo que casos mais próximos tenham peso maior;
5. Controles não inseridos em nenhuma região de influência terão suas probabilidades automaticamente igualadas a zero.

Recaptulando a notação utilizada, considere $D = \left\{x_{[j]}, j = 1, \cdots, C\right\}$, as coordenadas dos casos. Considere também d_{ij} a distância euclidiana entre os pontos com coordenadas x_i e x_j. Com essas notações pré-estabelecidas, o conceito de *regiões de vizinhança* em torno de um caso $x_{[j]}$ pode ser definida.

Definição 5. *A região circular de influência em torno de um caso $x_{[j]}$ será o conjunto $E_{[j]} = \left\{i \in \{1, \cdots, N\} \,|x_i \in \mathcal{C}(x_{[j]}, r_{[j]})\right\}$.*

A escolha do valor de $r_{[j]}$ será discutida mais adiante.

Uma primeira grandeza a ser utilizada é $\theta_{[j]}$, a proporção de casos dentro da região de influência do j-ésimo caso. Considerando-se $Q_{[j]} = \#E_{[j]}$, então

$$\theta_{[j]} = \frac{\sum\limits_{i \in E_{[j]}} c_i}{Q_{[j]}} \tag{5.1}$$

Para cada ponto $x_i, i \in E_{[j]}$ e $x_i \neq x_{[j]}$, necessita-se ponderar $\theta_{[j]}$ de modo que sua probabilidade estimada seja inversamente proporcional à distância que o separa do centro do círculo. Essa ponderação é feita multiplicando-se $\theta_{[j]}$ por $w_{i[j]}$, onde

$$w_{i[j]} = \begin{cases} \frac{1/d_{i[j]}}{\sum\limits_{k \neq [j]} 1/d_{k[j]}} & \text{se } i \neq [j] \\ 1 & \text{c.c} \end{cases} . \tag{5.2}$$

A soma no denominador na primeira parte de (5.2) evita problemas de escala nas coordenadas, além de garantir que $w_{i[j]} \leq 1$.

A probabilidade definida em $E_{[j]}$ para o i-ésimo ponto desta região será representada por $\pi_i^{[j]}$, onde

$$\pi_i^{[j]} = w_{i[j]}\theta_{[j]} \tag{5.3}$$

Portanto, ao centro do círculo que define a região será associada a proporção $\theta_{[j]}$ de casos dentro deste conjunto. Aos pontos restantes serão associados valores de $\theta_{[j]}$ reduzidos a uma razão $w_{i[j]}$, inversamente proporcional à distância deste ponto ao centro da região.

Definem-se, então, C regiões de influência. Um ponto particular x_i pode pertencer a mais de uma região de influência, mesmo este sendo um caso. Quando isso acontece, passa a existir mais de um valor $\pi_i^{[j]}$ para o mesmo ponto. Nesse caso, é plausível que se tome uma média desses valores estimados ponderada pelos inversos das distâncias ao centro de cada círculo ao qual pertence, de modo que casos mais próximos tenham peso maior. Ao mesmo tempo, é plausível que pontos não inseridos em nenhum círculo de influência tenham probabilidades iguais a 0. Mais formalmente, define-se essa média ponderada como π_i, onde

$$\pi_i = \begin{cases} \sum\limits_{j=1}^{C} h_{i[j]}\pi_i^{[j]} & \text{se } \sum\limits_{k=1}^{C} I(i \in E_{[k]}) \geq 1 \\ 0 & \text{c.c} \end{cases} \tag{5.4}$$

onde $I()$ é a função indicadora.

Como em (5.2), os valores $h_{i[j]}$ em (5.4) são componentes de um vetor normalizado de pesos, definidos como

$$h_{i[j]} = \begin{cases} \frac{I(i \in E_{[j]})\left(1/d_{i[j]}\right)}{\sum_{k=1}^{C} I(i \in E_{[k]})\left(1/d_{i[k]}\right)} & \text{se } \sum_{k=1}^{C} I(i \in E_{[k]}) \geq 1 \\ 0 & \text{c.c} \end{cases} \tag{5.5}$$

Quando $x_i = x_{[j]}$, para algum $j \in 1, \cdots, C$, ou seja, quando o ponto i é o centro de uma das C regiões de influência, $\pi_i = \theta_{[j]}$, mesmo que este centro esteja no interior de outro círculo. Isso acontece pela própria definição dos pesos $h_{i[j]}$.

Considere um ponto hipotético com coordenadas x_u. Á medida que x_u se aproxima de $x_{[j]}$, a distância $d_{u[j]}$ se aproxima de zero, fazendo com que $z_{u[j]} = 1/d_{u[j]} \to \infty$. Um cálculo de limite mostra que:

$$\begin{aligned} \lim_{z_{u[j]} \to \infty} (h_{u[j]}) &= \lim_{z_{u[j]} \to \infty} \frac{z_{u[j]}}{z_{u[j]} + \sum_{k \neq [j]} I(u \in E_{[k]}) z_{u[k]}} \\ &= \lim_{z_{u[j]} \to \infty} \frac{z_{u[j]}}{z_{u[j]}} \\ &= 1 \end{aligned} \tag{5.6}$$

$$\begin{aligned} \lim_{z_{u[j]} \to \infty} (h_{u[s]}) &= \lim_{z_{u[j]} \to \infty} \frac{z_{u[s]}}{z_{u[j]} + \sum_{k \neq j} I(u \in E_{[k]}) z_{u[k]}} \\ &= 0. \end{aligned} \tag{5.7}$$

Sucintamente, dado que um ponto é interior a mais de um círculo de influência, quanto mais próximo do centro de um desses círculos, mais a média ponderada das medidas se aproximará daquela definida por este círculo. Quando este ponto é o próprio centro de um círculo, seu peso tende a 1, enquanto o peso de outros centros tende a zero. Assim, caso um centro de um círculo, definido pela coordenada $x_{[j]}$, pertença à região circular de influência de *outros* casos, sua medida de probabilidade será definida por $\theta_{[j]}$, ou seja, nenhuma média ponderada considerando as probabilidades definidas em outras regiões necessita ser calculada.

Finalmente, as estimativas $\hat{p}_i$ são definidas por uma normalização dos valores π_i, de modo que $\sum_{i=1}^{N} \hat{p}_i = C$. Em outras palavras,

$$\hat{p}_i = C \frac{\pi_i}{\sum_{k=1}^{N} \pi_k}. \tag{5.8}$$

5.2.1 Regiões de influência

Tendo em mãos um método de cálculo das probabilidades $\hat{p}_i$ a partir de círculos de influência de raios $r_{[j]}$ e centros $x_{[j]}$, falta definir os raios $r_{[j]}$. Como mencionado no começo deste Capítulo, esse raio será determinado com o auxílio de um MST construído nos casos. Então, $r_{[j]}$ será determinado pelo comprimento l de uma das arestas que incidem em $x_{[j]}$. Definindo-se $l_{[j]}^{max}$ e $l_{[j]}^{min}$ a maior e a menor aresta, respectivamente, incidentes no ponto $x_{[j]}$, considerou-se quatro possibilidades:

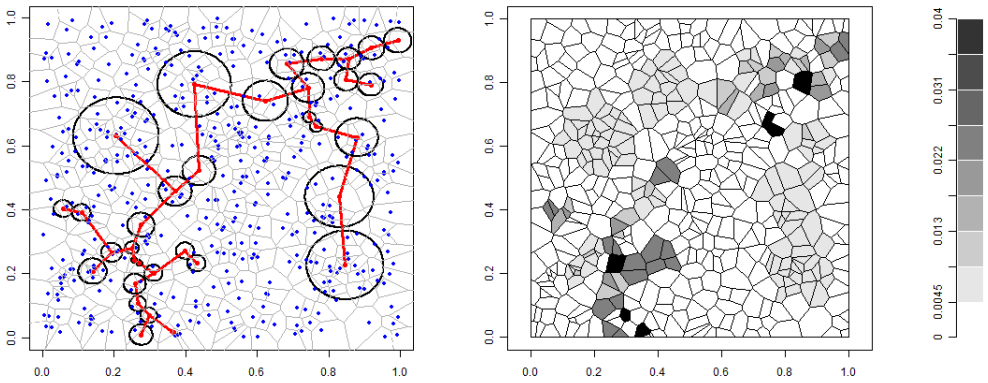
1. Metade da menor aresta: $r_{[j]} = l_{[j]}^{min}/2$;
2. Total da menor aresta: $r_{[j]} = l_{[j]}^{min}$;
3. Metade da maior aresta: $r_{[j]} = l_{[j]}^{max}/2$;
4. Total da maior aresta: $r_{[j]} = l_{[j]}^{max}$.

Para ilustrar esses cenários, as Figuras 5.1(a), 5.2(a), 5.3(a) e 5.4(a) ilustram o MST constuído em um conjunto simulado de pontos, bem como os diferentes círculos centrados nos casos definidos em cada método enumerado acima. A base simulada é composta de 500 observações cujas coordenadas (latitude e longitude) foram amostradas de duas uniformes $U(0, 1)$ independentes. Dentre esses pontos, selecionou-se 35 que seriam definidos como casos. O método de seleção desses pontos será discutido adiante.

As Figuras 5.1(b), 5.2(b), 5.3(b) e 5.4(b)ilustram as probabilidades $\hat{p}_i$ obtidas de acordo com (5.8). É possível notar que à medida que se considera raios maiores (aumentando o número de intersecções entre os círculos), ocorre uma "suavização" das probabilidades estimadas, com menos regiões destacadas com probabilidades maiores.

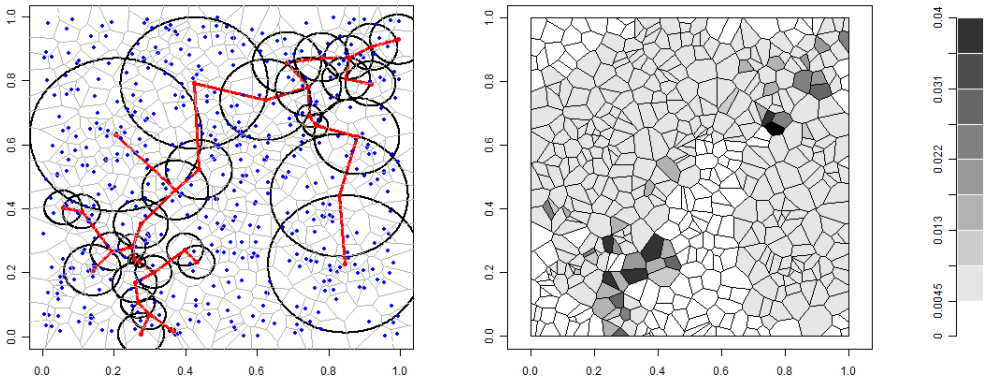
5.3 Propriedades inferenciais dos estimadores das probabilidades individuais

Para assegurar a confiabilidade de um estimador, é crucial que sejam analisadas suas propriedades amostrais, tais como Variância, Viés e Erro Quadrático Médio (EQM). O método de construção de *clusters* artificiais descrito na seção anterior



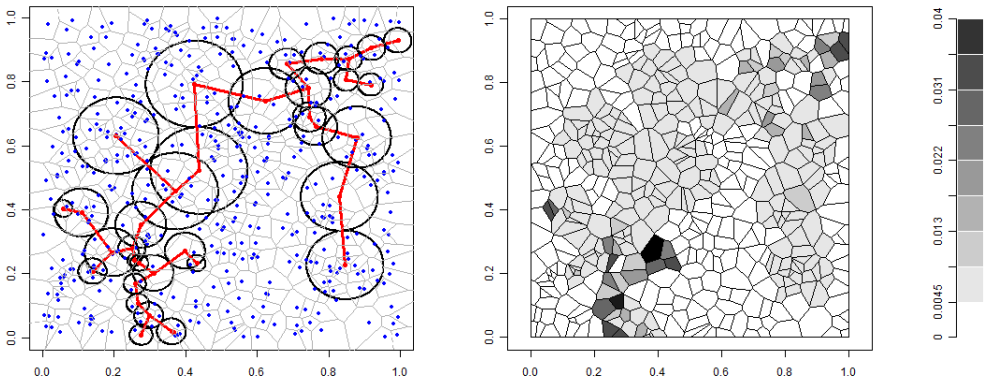
(a) Círculos (b) Probabilidades

Figura 5.1: Círculos de influência: raios iguais à metade da menor aresta do MST



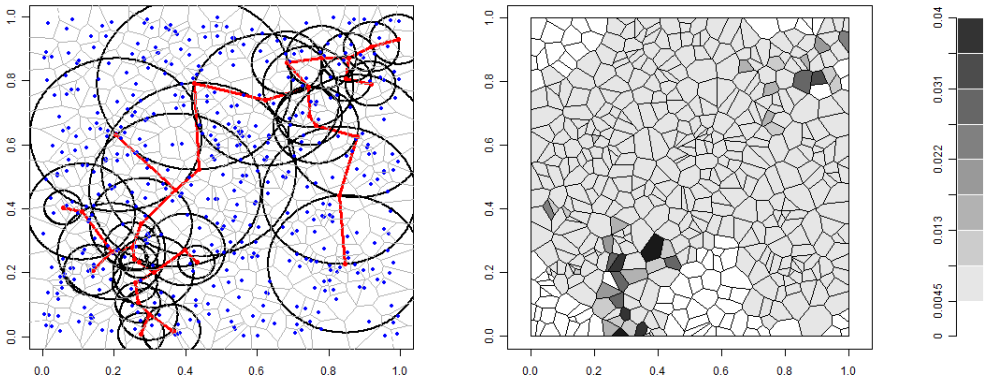
(a) Círculos (b) Probabilidades

Figura 5.2: Probabilidades estimadas: raios iguais ao total da menor aresta do MST



(a) Círculos (b) Probabilidades

Figura 5.3: Círculos de influência: raios iguais à metade da maior aresta do MST



(a) Círculos (b) Probabilidades

Figura 5.4: Círculos de influência: raios iguais ao total da maior aresta do MST

determina os valores p_i verdadeiros usados na simulação, possibilitando a observação de propriedades como EQM e viés.

Considere um espaço paramêtrico d-dimensional $\Theta \subset \mathbb{R}^d$, um parâmetro $\theta = (\theta_1, \cdots, \theta_d) \in \Theta$ e um estimador de θ, digamos $\hat{\theta} = (\hat{\theta_1}, \cdots, \hat{\theta_d})$. O EQM definido para vetores paramétricos com dimensão $d > 1$ é dado por

$$EQM(\hat{\theta}) = E\left(\left\|\hat{\theta} - \theta\right\|^2\right) \tag{5.9}$$

É possível mostrar que, dada a matriz $\Sigma(\hat{\theta})$ de variância e covariância do vetor $\hat{\theta}$

$$\begin{aligned} EQM(\hat{\theta}) &= traço(\Sigma(\hat{\theta})) + \left\|E(\hat{\theta}) - \theta\right\|^2. \\ &= \sum_{i=1}^{d} Var(\hat{\theta_i}) + \sum_{i=1}^{d} (E(\hat{\theta_i}) - \theta_i)^2 \end{aligned} \tag{5.10}$$

Os componentes do lado direito da Equação (5.10) são, respectivamente, a *variância total* e o *viés total* de $\hat{\theta}$.

Na situação em questão, $\theta = p = (p_1, \cdots, p_n)$. A obtenção das características acimas para o estimador (5.8) de forma analítica não é trivial, e portanto simulações de Monte Carlo foram necessárias para visualizar tais valores. A idéia é definir um *cluster* artificial M vezes, gerando M bases simuladas. Com base nessas amostras, a estimativa $\widehat{EQM}(\hat{p})$ é calculada, onde

$$\widehat{EQM}(\hat{p}) = \sum_{i=1}^{N} \widehat{Var}(\hat{p_i}) + \sum_{i=1}^{N} (\widehat{E}(\hat{p_i}) - p_i)^2, \tag{5.11}$$

sendo que

$$\begin{aligned} \widehat{E}(\hat{p_i}) &= \frac{1}{M} \sum_{k=1}^{M} \hat{p_i^k} \\ \widehat{Var}(\hat{p_i}) &= \frac{1}{M} \sum_{k=1}^{M} (\hat{p_i^k} - \widehat{E}(\hat{p_i}))^2 \end{aligned} \tag{5.12}$$

Uma metodologia para de gerar *clusters* artificiais é descrita na próxima seção. As análises das estimativas $\widehat{EQM}(\hat{p})$ serão apresentadas no próximo Capítulo.

5.4 Criação de *clusters* artificiais através de simulação

O Capítulo 2 apresentou o problema de detecção de *clusters* espaciais como sendo um teste construído sob as hipóteses definidas em (2.2). Dessa forma, a criação de *clusters* artificiais envolve a simulação dos modelos probabilísticos adotados sob H_1.

Tendo em mãos um espaço G contendo N pontos e $C < N$ casos, considere $Z \subset G$ uma região pré-estabelecida, contendo $N_z < N$ pontos, como sendo um conjunto candidato a *cluster*. Sob H_0, o risco relativo rr_A em um conjunto A será $rr_A = c_A/\mu_A = 1, \forall A \subset G$. Sob H_1, o risco relativo será $rr_A = r > 1$ se $A = Z$, e $rr_A = 1$ se $A \neq Z$. A simulação sob a hipótese alternativa seguirá a metodologia proposta por Kulldorff em [Kuldorff *et al.*, 2003], que determina o risco relativo rr_Z condicionado a H_1 como função de um poder de teste $(1-\beta)$ e um nível de significância α, ambos definidos no contexto de um teste aproximado da distribuição binomial. A partir desse risco relativo, determina-se os parâmetros de uma distribuição binomial, da qual o número de casos dentro do cluster será gerado. A presente seção visa detalhar esse método.

Kulldorff considera que o número Y de casos, dentre C, que estão no cluster de N_z pontos é uma variável aleatória tal que $Y \sim Bin(C, p_j^*)$, sendo que p_j^* é a probabilidade de um caso qualquer estar em Z, sendo que $j = 0$ sob H_0 e $j = 1$ sob H_1. É importante considerar também o número X de pontos dentre N_Z que são casos, $X \sim Bin(N_Z, p_j)$, onde p_j é a probabilidade de um ponto aleatório dentre N_Z ser um caso, $j = 0$ sob H_0 e $j = a$ sob H_a;. Para entender em detalhes a obtenção de p_j^*, considere os seguintes eventos E_1 e E_2:

1. E_1: o ponto é um caso;
2. E_2: o ponto está no cluster.

Considere também

1. $p_j^* = P_j(E_2|E_1)$ = probabilidade de um caso estar no *cluster*;
2. $p_j = P_j(E_1|E_2)$ = probabilidade de um ponto no *cluster* ser um caso;

Com essas definições em mente, as probabilidades p_j^* e p_j sob as duas hipóteses serão obtidas a seguir.

H_0**: não há** *cluster* Nessa situação, os seguintes fatos ocorrem:

- $P_0(E_1|E_2) = P_0(E_1|\bar{E}_2) = p_0 = C/N$. Em outras palavras, qualquer ponto no espaço tem a mesma probabilidade de ser um caso, o que por definição significa que não há cluster;
- $P(E_2) = \frac{N_z}{N}$;
- Pela regra de Bayes:

$$\begin{aligned} p_0^* = P_0(E_2|E_1) &= \frac{P_0(E_1|E_2)P(E_2)}{P_0(E_1|E_2)P(E_2) + P_0(E_1|\bar{E}_2)P(\bar{E}_2)} \\ &= \frac{p_0 P(E_2)}{p_0 P(E_2) + p_0(1 - P(E_2))} \\ &= P(E_2) \\ &= \frac{N_z}{N}. \end{aligned} \tag{5.13}$$

H_1**: há** *cluster* Nessa situação, observa-se que:

- $P_1(E_1|E_2) = p_1 > P_1(E_1|\bar{E}_2) = q_1$.
- Considerando os riscos relativos esperados dentro e fora do cluster, tem-se:

$$\begin{aligned} E\left(rr_Z\right) &= \frac{\mu_1^Z}{\mu_0^Z} = \frac{N_z p_1}{N_z p_0} = \frac{p_1}{p_0} = r; \\ E\left(rr_{\bar{Z}}\right) &= \frac{\mu_1^{\bar{Z}}}{\mu_0^{\bar{Z}}} = \frac{(N - N_z)q_a}{(N - N_z)p_0} = \frac{q_a}{p_0} = 1. \end{aligned} \tag{5.14}$$

- isso implica que $p_1 = rp_0 = rC/N$ e $q_1 = p_0$;
- Pela regra de Bayes:

$$\begin{aligned} p_1^* = P_1(E_2|E_1) &= \frac{P_1(E_1|E_2)P(E_2)}{P_1(E_1|E_2)P(E_2) + P_1(E_1|\bar{E}_2)P(\bar{E}_2)} \\ &= \frac{rp_0 P(E_2)}{rp_0 P(E_2) + p_0(1 - P(E_2))} \\ &= \frac{rP(E_2)}{rP(E_2) + (1 - P(E_2))} \\ &= \frac{rP(E_2)}{P(E_2)(r-1) + 1}. \end{aligned} \tag{5.15}$$

- Substituindo $P(E_2) = N_z/N$:

$$p_1^* = \frac{N_z r}{(N - N_z + N_z r)} \tag{5.16}$$

Voltando à variável aleatória $Y \sim Bin(C, p_j^*)$ e usando a aproximação da distribuição binomial pela distribuição normal, o primeiro passo é obter um valor k tal que

$$\frac{(k - m_0)}{v_0} = z_{1-\alpha}, \tag{5.17}$$

onde $m_0 = E(Y|H_0) = Cp_0^*$, $v_0 = Var(Y|H_0) = Cp_0^*(1 - p_0^*)$, p_0^* é determinado em (5.13) e $z_{1-\alpha}$ é o quantil de ordem $(1 - \alpha)$ da distribuição $N(0, 1)$. Obtido o valor k, o próximo passo é determinar o valor de r tal que

$$\frac{(k - m_1)}{v_1} = z_{1-\beta}, \tag{5.18}$$

onde $m_1 = E(Y|H_1) = Cp_1^*$, $v_1 = Var(Y|H_1) = Cp_1^*(1 - p_1^*)$, p_1^* é função de r, de acordo com a Equação (5.16), e $z_{1-\beta}$ é o quantil de ordem $(1 - \beta)$ da distribuição $N(0, 1)$.

Assim, o risco relativo teórico r será determinado por α e β considerando-se a distribuição de Y sob H_0 e H_1, respectivamente.

O número de casos c_Z no cluster Z será determinado por uma amostra de $Y \sim Bin(C, p_1^*)$, sendo que, como visto acima, $p_1^* = \frac{N_z r}{(N - N_z + N_z r)}$. Assim, seleciona-se uma amostra aleatória simples de c_z pontos dentre os N_Z pertencentes à região Z, de modo que cada um tenha a mesma probabilidade de ser selecionado. De modo semelhante, seleciona-se $C - c_Z$ pontos dentre os $N - N_Z$ restantes fora do candidato a *cluster*. Esses pontos selecionados serão definidos como sendo casos (ou seja, esses pontos terão $c_i = 1$).

Capítulo 6

Resultados em dados simulados

A avaliação das propriedades inferenciais dos $\hat{p}_i$ requerem, como já foi salientado, o cálculo em bases simuladas de dados. Tais bases também são necessárias para ilustar o cálculo das medidas de intensidade em (3.1), a partir em aleatorizações baseadas nas probabilidades estimadas em (5.8), seguindo a idéia apresentada no Capítulo 3.

Antes de simular a situação de existência de *cluster*, criou-se um mapa hipotético com 500 pontos, cujas coordenadas no intervalo $[0,1] \times [0,1]$ foram geradas de distribuições uniformes independentes. O *cluster* Z foi definido em vários cenários prováveis:

1. *Cluster* circular simples;
2. *Cluster* circular *fraco*, simulado em um cenário com baixo poder de teste.
3. *Cluster* em formato não circular;
4. *Cluster* circular duplo, ou seja, duas regiões circulares no mapa;
5. *Cluster* circular simples em um mapa com diferentes densidades locais;

O conceito de *fraco* no segundo cenário está relacionado com o poder de teste escolhido. Para esse caso, considerou-se um poder relativamente baixo, $(1-\beta) = 0,7$. Nos demais casos, definiu-se um poder de teste maior, $(1-\beta) = 0,999$. Em todos os cenários, foi considerado um nível de significância $\alpha = 0,05$.

Com exceção do caso com diferentes densidades locais, todos os demais cenários foram simulados no mesmo mapa criado.

O número total de casos considerado foi $C = 35$. Os valores C_i indicadores de casos e não-casos observados foram gerados aleatoriamente de acordo com o método descrito anteriormente.

Para a avaliação das propriedades amostrais dos estimadores de $p = (p_i, \cdots, p_N)$, foram geradas $M = 1000$ bases de dados (1000 replicações aleatórias dos valores $C_i, i = 1, \cdots, N$), em cada cenário. Em cada replicação k, obteve-se as estimativas $\hat{p}_i^k$, considerando-se cada uma das quatro alternativas sugeridas para definição do raio $r_{[j]}$ das regiões de influência. Tais replicações possibilitaram a avaliação dos estimadores com base em estimativas do EQM.

A ilustração do cálculo das medidas de intensidade será feita com base em apenas uma dessas replicações aleatórias, não impossibilitando, entretanto, que boas conclusões a respeito do comportamento dessa medida possam ser inferidas.

Esta seção apresentará, através de recursos gráficos e tabulações, os resultados de estimativas de viés, variância e EQM dos estimadores $\hat{p}_i$, além das medidas de intensidade calculadas nos dados simulados em cada cenário. Em cada caso, serão apresentados quatro resultados, cada um relacionado a uma das maneiras possíveis de utilizar comprimentos de arestas do MST para a obtenção de $\hat{p}_i$. O método de detecção adotado foi o *scan* circular, baseado em janelas circulares contendo até $K = 168$ pontos.

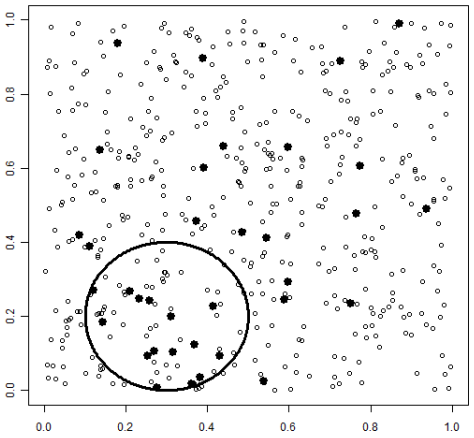
As simulações e resultados foram obtidos através de rotinas próprias programadas através do *software R*. A construção dos diagramas de Voronoi e as representações de valores através de colorações de polígonos contaram, respectivamente, com os pacotes *tripack* e *SpatialEpi*.

6.1 Cenário 1: cluster circular simples

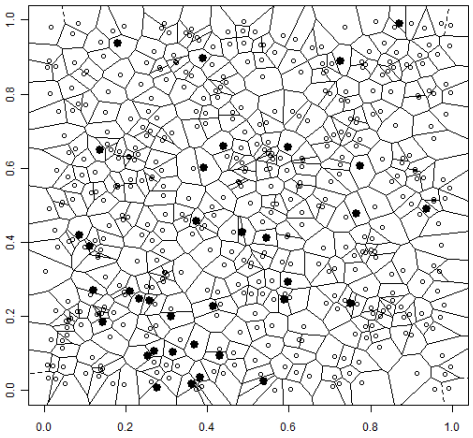
O *cluster* Z foi definido como uma região circular com raio 0.2 e centrado na coordenada $x = (0.3, 0.2)$. Essa região possui 64 pontos e $c_z = 15$ casos. O risco relativo *teórico*, obtido conforme (5.17) e (5.18), é igual a 6.34. O risco relativo observado, a razão $(c_z/N_Z)/p_0$, foi de 3.35. Esses valores estão resumidos na Tabela 6.1. A Figura 6.1 mostra o cluster simulado e o Diagrama de Voronoi com base nos 500 dados da região.

Tabela 6.1: Cenário 1: informações dentro e fora do cluster

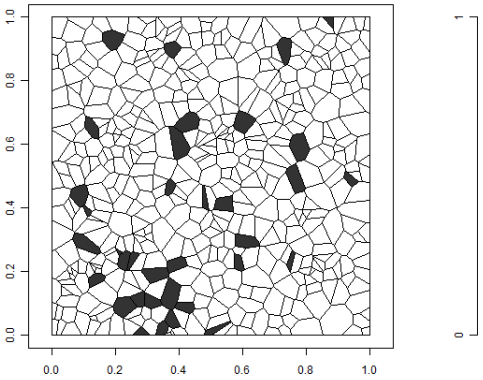
	Cluster	Não-Cluster
N° de Casos	15	20
N° de pontos	64	436
Risco relativo teórico	6.34	1.00
Risco relativo observado	3.35	0.66



(a) Cluster simulado.



(b) Diagrama de Voronoi.



(c) Casos observados no diagrama de Voronoi

Figura 6.1: Cenário 1: mapa com cluster circular simples.

6.1.1 Viés e Expectância dos estimadores $\hat{p}$

As Figuras 6.2, 6.3, 6.4 e 6.5 ilustram as probabilidades estimativas $\widehat{E}(\hat{p}_i)$ nos quatro métodos considerados no cálculo dos $\hat{p}_i$. É possível notar que as expectâncias estimadas dos $\hat{p}_i$ são mais elevadas na região do *cluster* verdadeiro.

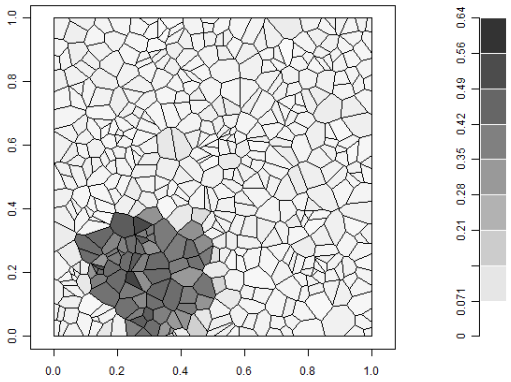


Figura 6.2: Cenário 1: médias das probabilidades, metade da menor aresta.

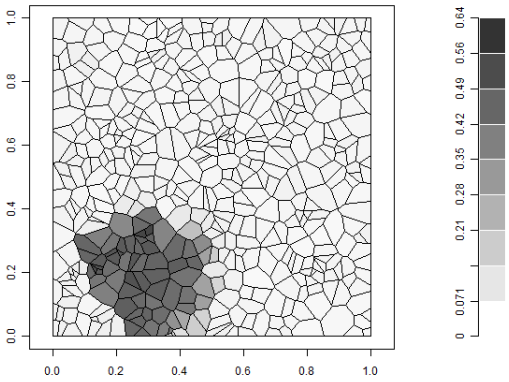


Figura 6.3: Cenário 1: médias das probabilidades, total da menor aresta.

Os *boxplots* presentes nas Figuras 6.6 e 6.7 representam, respectivamente, as variâncias e medidas de viés dos p_i individuais. É possível notar que os métodos baseados na maior aresta apresentam níveis maiores de variabilidade, além de maior heterogeneidade desses valores. Em todos os métodos, nota-se menor variabilidade fora do *cluster*. A metade da menor aresta possui menor mediana de variabilidade; porém, o método considerando o total da menor aresta possue uma amplitude mais reduzida dos valores individuais da variância de $\hat{p}_i$.

O viés se comporta de modo semelhante à variância, tendo níveis e heterogeidade maiores entre indivíduos quando se considera métodos baseados nos maiores pesos do MST. É interessante notar que, em todos os métodos, há uma ocorrência frequente de viés positivo dentro do *cluster*, indicando que as probabilidades nessa região são, em sua maior parte, superestimadas.

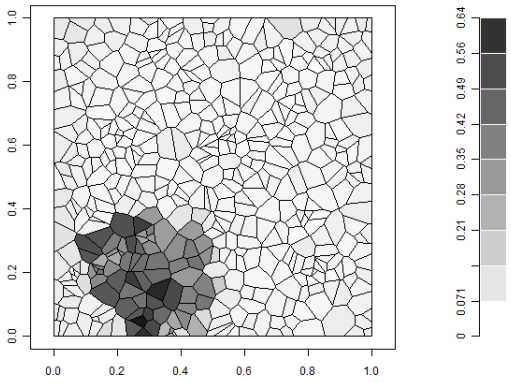


Figura 6.4: Cenário 1: médias das probabilidades, metade da maior aresta.

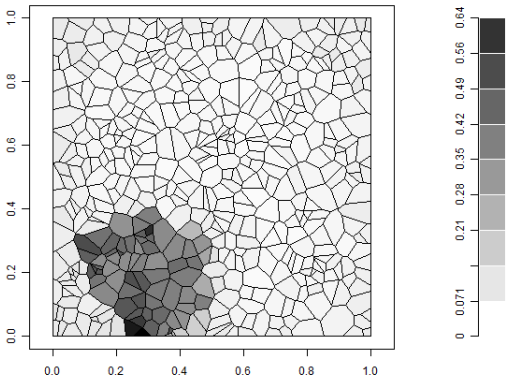


Figura 6.5: Cenário 1: médias das probabilidades, total da maior aresta.

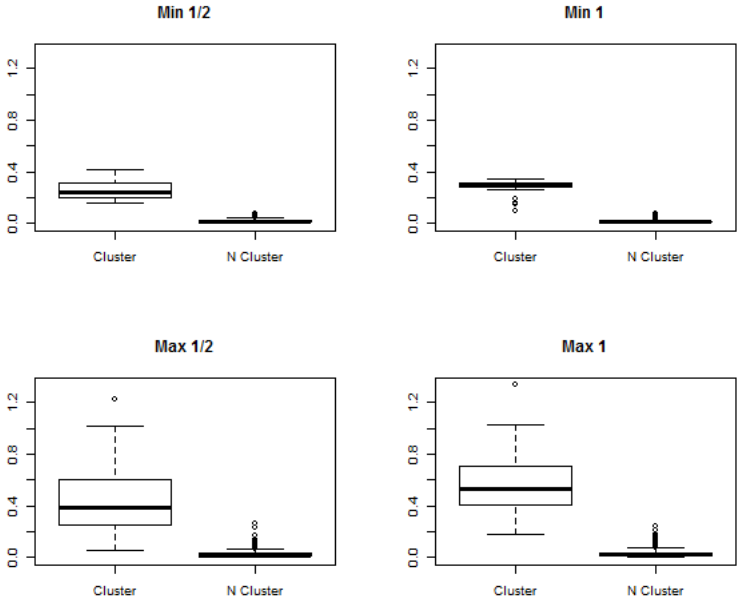


Figura 6.6: Cenário 1: estimativas individuais de variância ($Var(\hat{p}_i)$)

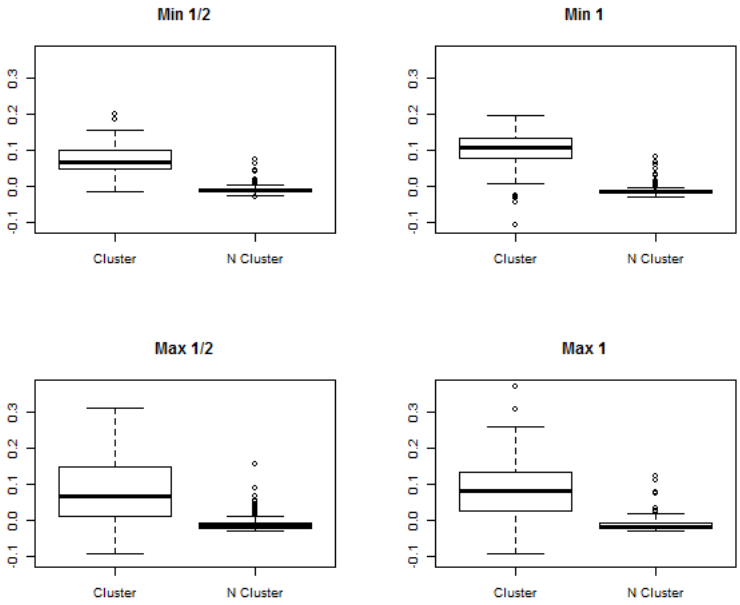


Figura 6.7: Cenário 1: estimativas individuais de viés ($Vies(\hat{p}_i)$)

6.1.2 Medidas de intensidade na base simulada

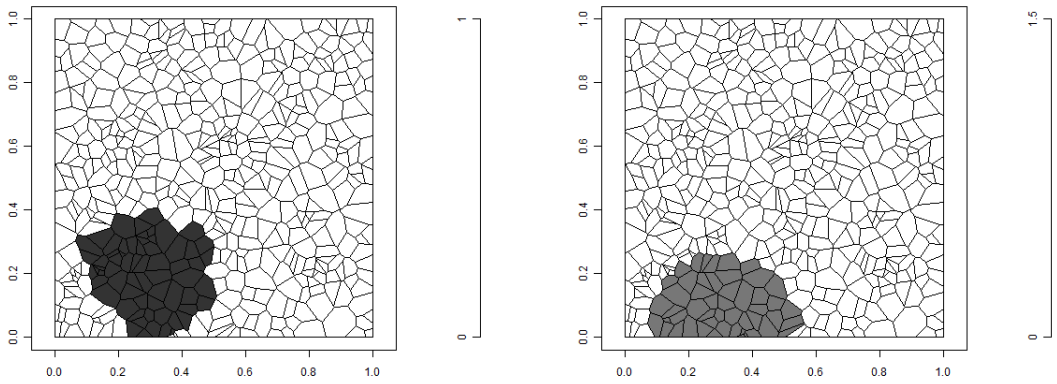
A comparação do *cluster* teórico *versus cluster* detectado segue na Figura 6.8. A Tabela 6.2 mostra que a razão de verossimilhanças associada ao cluster detectado é significativo a 99%.

Tabela 6.2: Cenário 1: razão de verossimilhanças observada × valores críticos.

Detectado	p99	p95	p90
12.31	10.86	8.51	8.10

As Figuras 6.9, 6.10, 6.11 e 6.12 ilustram as medidas de intensidade resultantes dos quatro métodos considerados no cálculo dos $\hat{p}_i$.

Em uma primeira impressão, não fica evidente qual método destaca melhor, através das medidas de intensidade, o *cluster* do restante do mapa. Esse fato pode ser visto também no *boxplot* (Figura 6.13), comparando a distribuição das medidas em Z e $\bar{Z}$. Um fato interessante pode ser observado na Tabela 6.3: através dos métodos baseados nas metades das arestas, todos indivíduos pertencentes ao *cluster* circular



(a) Cluster real.

(b) Cluster detectado.

Figura 6.8: Cenário 1: cluster verdadeiro × cluster detectado.

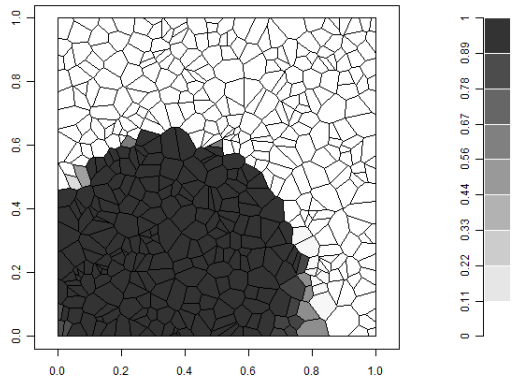


Figura 6.9: Cenário 1: medidas de intensidade, metade da menor aresta.

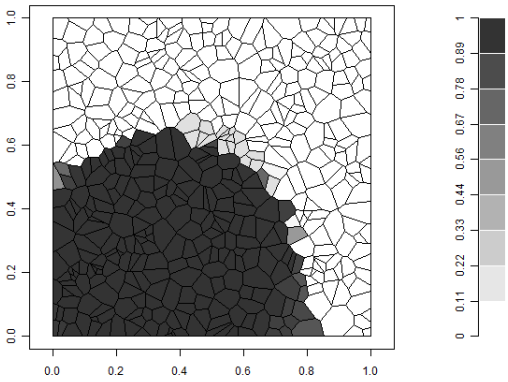


Figura 6.10: Cenário 1: medidas de intensidade, total da menor aresta.

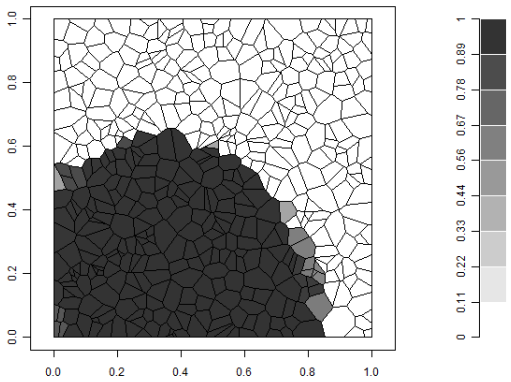


Figura 6.11: Cenário 1: medidas de intensidade, metade da maior aresta.

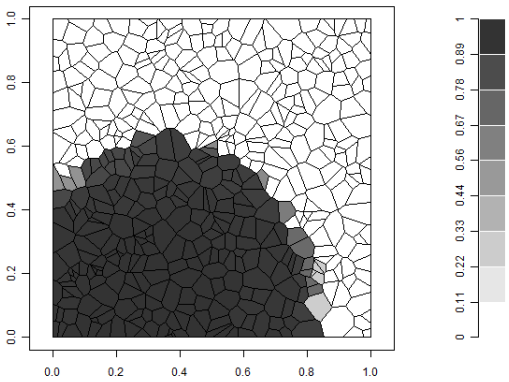


Figura 6.12: Cenário 1: medidas de intensidade, total da maior aresta.

resultaram em medidas iguais a 1, o valor máximo. O método que apresentou a maior diferença entre as médias dentro e fora de Z foi o baseado na metade da menor aresta.

Tabela 6.3: Cenário 1: medidas de intensidade médias dentro e fora do *cluster*.

	Cluster	N Cluster	Diferença
Metade da aresta mínima	1.00000	0.35897	0.64103
Total da aresta mínima	0.99990	0.37994	0.61996
Metade da aresta máxima	1.00000	0.38388	0.61612
Total da aresta máxima	0.99843	0.36931	0.62912

6.2 Cenário 2: cluster circular "fraco"

O *cluster* considerado nesse cenário foi o mesmo do primeiro. A diferença está no poder de teste considerado, $(1-\beta) = 0,7$, que resultou em um $c_z = 8$. O risco relativo observado é baixo se comparado com o caso anterior, mas ainda sim está próximo do teórico. Veja a Tabela 6.4. O *cluster* simulado está representado na Figura 6.14.

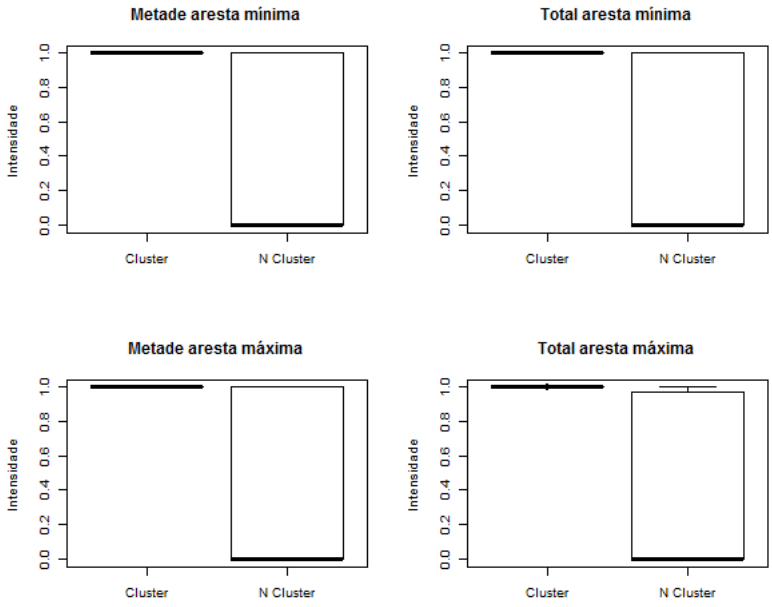
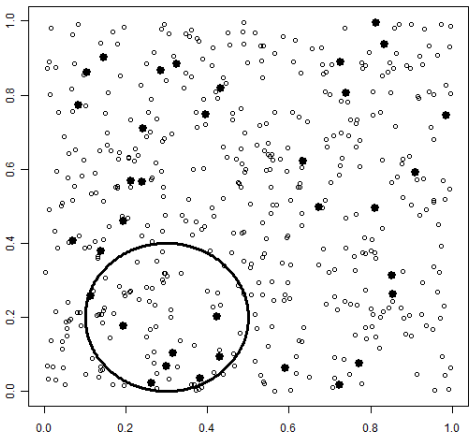


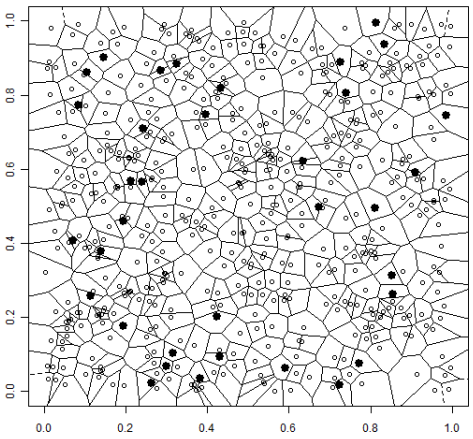
Figura 6.13: Cenário 1: distribuição das medidas de intensidade dentro e fora do cluster verdadeiro.

Tabela 6.4: Cenário 2: informações dentro e fora do cluster

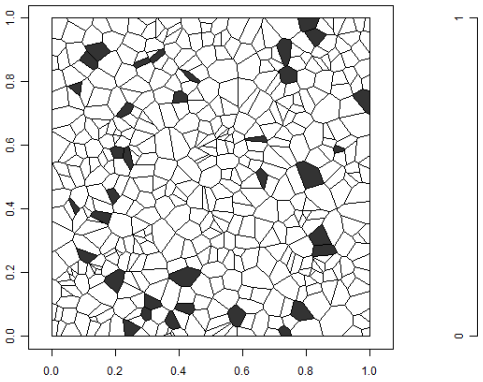
	Cluster	N Cluster
N° de Casos	8	27
N° de pontos	64	436
Risco relativo teórico	2.39	1.00
Risco relativo observado	1.79	0.88



(a) Cluster simulado.



(b) Diagrama de Voronoi.



(c) Casos observados no diagrama de Voronoi

Figura 6.14: Cenário 2: mapa com cluster circular "fraco".

6.2.1 Viés e Expectância dos estimadores $\hat{p}$

As expectâncias estimadas via Monte Carlo estão representadas nas Figuras 6.15, 6.16, 6.17 e 6.18, mostrando mais uma vez que as probabilidades são em média mais elevadas na região do *cluster* verdadeiro. As diferenças das probabilidades dentro e fora de Z, entretanto, são menores nesse caso do que no Cenário 1.

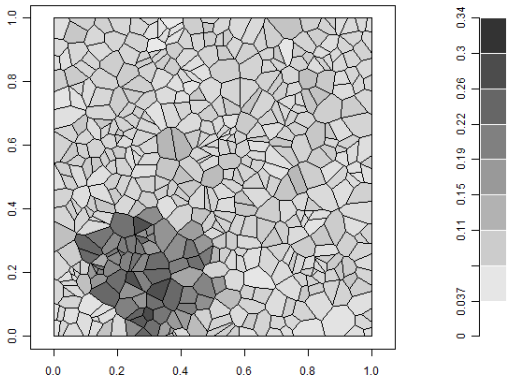


Figura 6.15: Cenário 2: médias das probabilidades, metade da menor aresta.

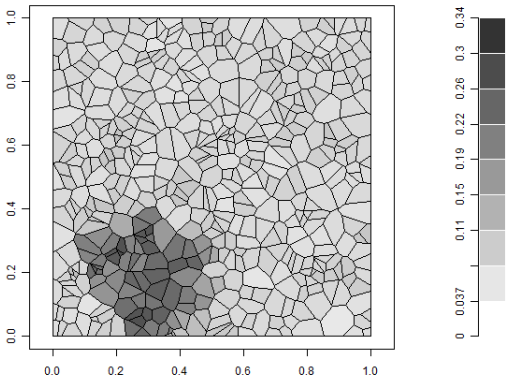


Figura 6.16: Cenário 2: médias das probabilidades, total da menor aresta.

Na Figura 6.19 é possível observar o mesmo padrão do cenário 1: métodos baseados na maior aresta apresentam níveis maiores de variabilidade, e apresentam maior heterogeneidade dos valores $Var(\hat{p}_i)$ entre os indivíduos. Além disso, a variabilidade, em qualquer método, é maior dentro do *cluster*.

O comportamento do viés também é semelhante ao cenário 1(Figura 6.20), isto é, mais componentes $\hat{p}_i$ possuem viés positivo em Z. Os desvios das estimativas individuais são concentrados em medianas menores quando se considera a maior aresta do MST, enquanto com as menores arestas esses erros médios são mais elevados, porém menos diferentes entre os indivíduos.

Vale notar, também, que tanto os valores observados de viés e variância são menores do que no cenário anterior dentro do *cluster*, apesar de que esses valores são mais elevados fora dessa região na situação definida no cenário 2.

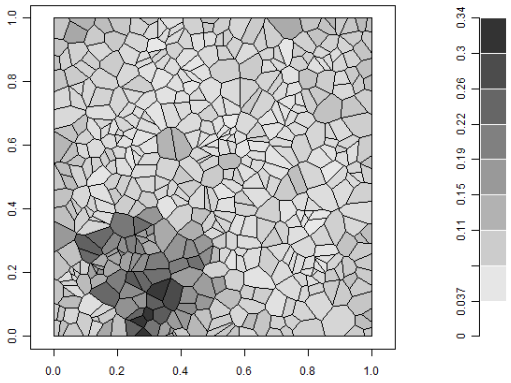


Figura 6.17: Cenário 2: médias das probabilidades, metade da maior aresta.

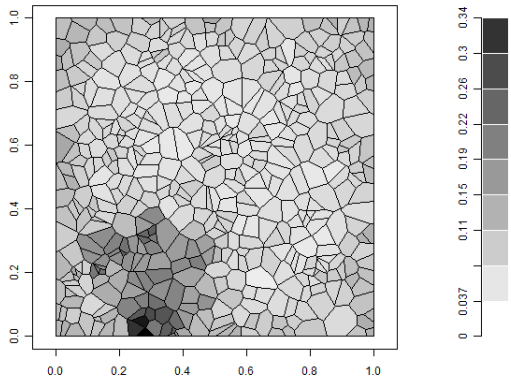


Figura 6.18: Cenário 2: médias das probabilidades, total da maior aresta.

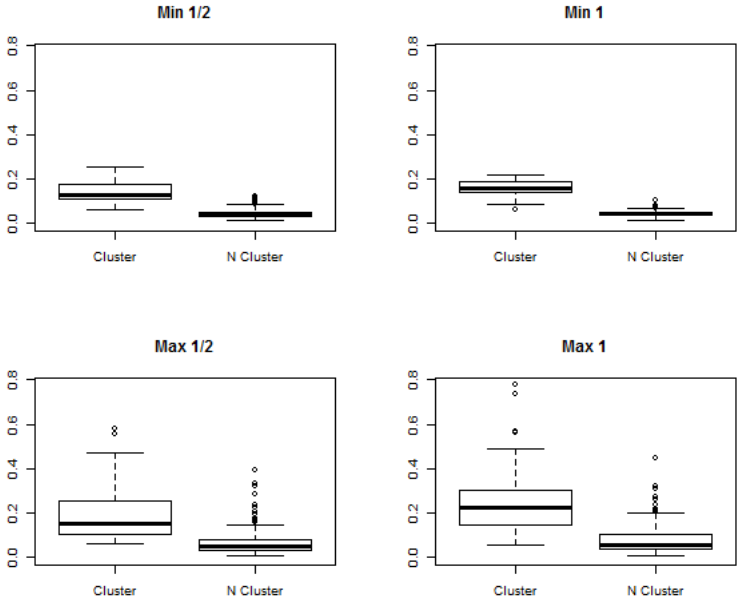


Figura 6.19: Cenário 2: estimativas individuais de variância ($Var(\hat{p}_i)$)

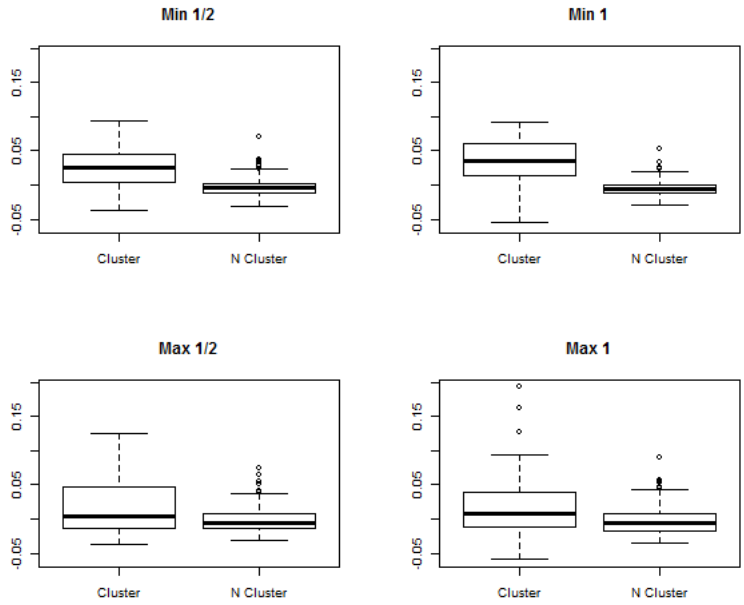


Figura 6.20: Cenário 2: estimativas individuais de viés ($Vies(\hat{p}_i)$)

6.2.2 Medidas de intensidade na base simulada

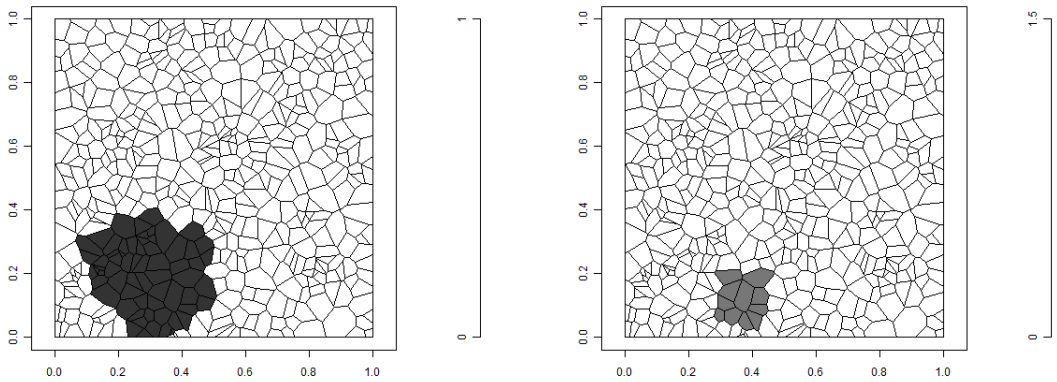
A Figura 6.21 compara o *cluster* detectado com o real. Note que a região detectada coincide menos com a verdadeira se comparado com o cenário 1. Além disso, a Tabela 6.5 mostra que o *cluster* detectado é não significativo.

Tabela 6.5: Cenário 2: razão de verossimilhanças observada × valores críticos.

Detectado	p99	p95	p90
6.45	10.86	8.51	8.10

As medidas de intensidade representadas nas Figuras 6.22, 6.23, 6.24 e 6.25 apresentam "manchas" contendo parte do cluster verdadeiro. Diversas outras "manchas" aparecem em vários lugares, com os mesmos níveis de intensidade da região próxima a Z, e regiões com alguma coloração mais fraca são mais extensas do que no cenário anterior. Os níveis de intensidade dentro do *cluster* verdadeiro foram menores do que no cenário 1 (Tabela 6.6). Em suma, há um grau de incerteza maior nesse cenário.

Pode-se observar algumas diferenças entre os métodos. A metade da maior aresta



(a) Cluster real.

(b) Cluster detectado.

Figura 6.21: Cenário 2: cluster verdadeiro × cluster detectado.

parece "cobrir" melhor o *cluster* verdadeiro, enquanto que a metade da menor aresta parece resultar em um cenário mais preciso. Em termos diferenças entre médias e de distribuição, o método da metade da aresta mínima parece destacar melhor o *cluster* verdadeiro do "não-*cluster*".

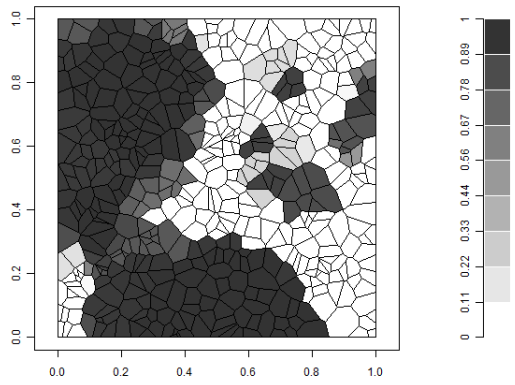


Figura 6.22: Cenário 2: medidas de intensidade, metade da menor aresta.

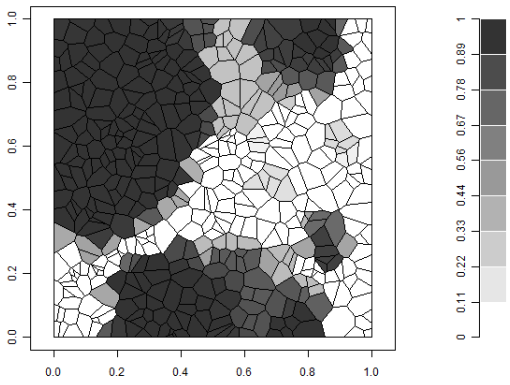


Figura 6.23: Cenário 2: medidas de intensidade, total da menor aresta.

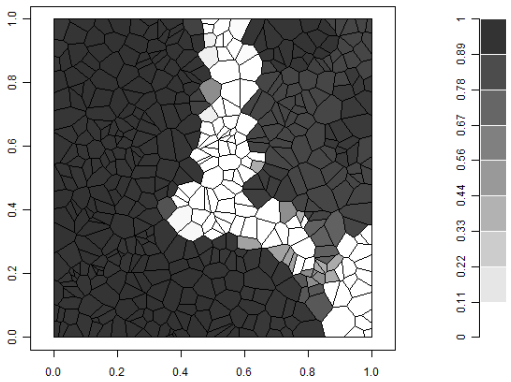


Figura 6.24: Cenário 2: medidas de intensidade, metade da maior aresta.

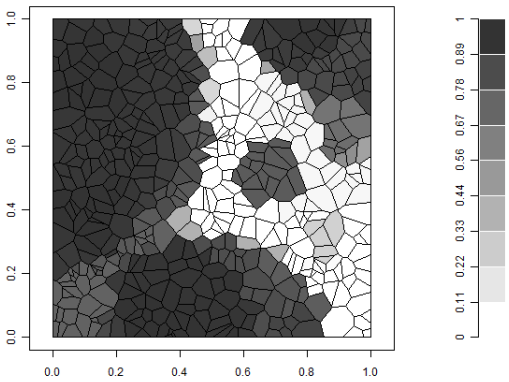


Figura 6.25: Cenário 2: medidas de intensidade, total da maior aresta.

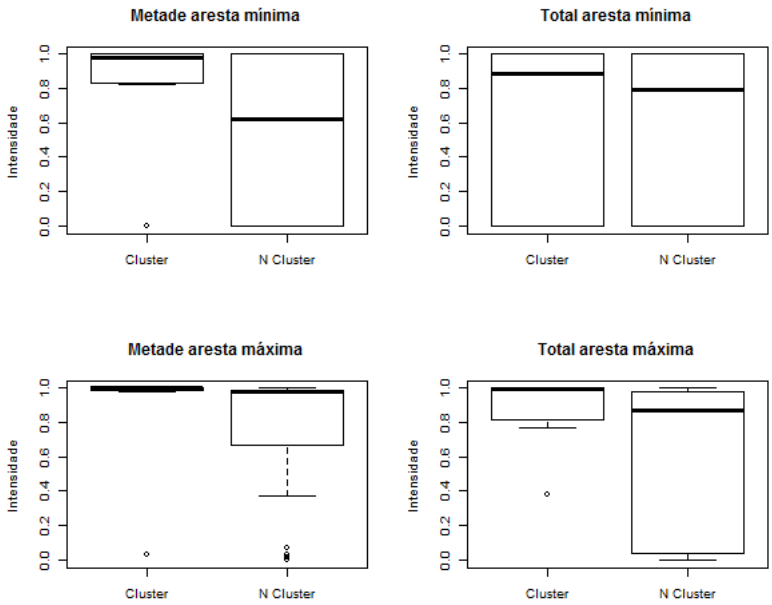


Figura 6.26: Cenário 2: distribuição das medidas de intensidade dentro e fora do cluster verdadeiro.

Tabela 6.6: Cenário 2: medidas de intensidade médias dentro e fora do *cluster*.

	Cluster	N Cluster	Diferença
Metade da aresta mínima	0.90764	0.48795	0.41969
Total da aresta mínima	0.60755	0.53434	0.07321
Metade da aresta máxima	0.97751	0.74008	0.23742
Total da aresta máxima	0.91433	0.62217	0.29216

6.3 Cenário 3: cluster não circular

Considerou-se a possibilidade de haver um *cluster* em formato não circular. Optou-se pela consideração de uma região em formato "L", representado na Figura 6.27. A composição da região em termos de *cluster* e número de casos segue na Tabela 6.7. Note que o risco relativo observado é bem próximo do teórico.

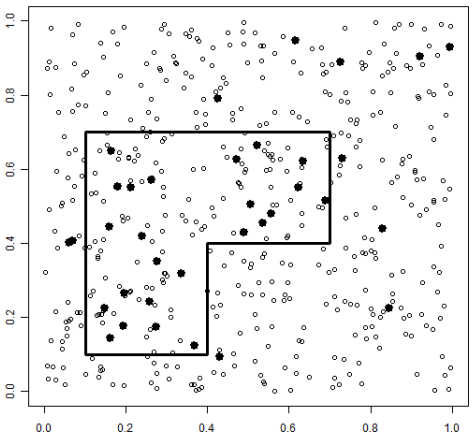
Tabela 6.7: Cenário 3: informações dentro e fora do cluster

	Cluster	N Cluster
N° de Casos	24.00	11.00
N° de pontos	148.00	352.00
Risco relativo teórico	2.39	1.00
Risco relativo observado	2.32	0.45

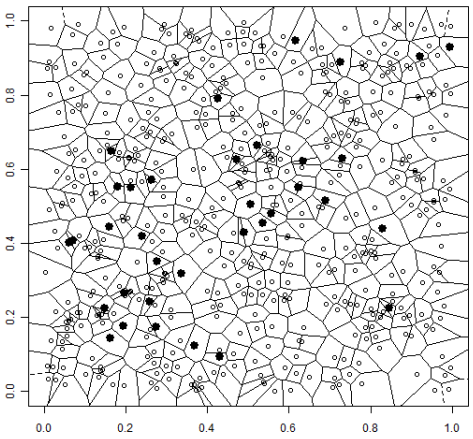
6.3.1 Viés e Expectância dos estimadores $\hat{p}$

Aqui as expectâncias também são maiores em Z (Figuras 6.28, 6.29, 6.30 e 6.31).

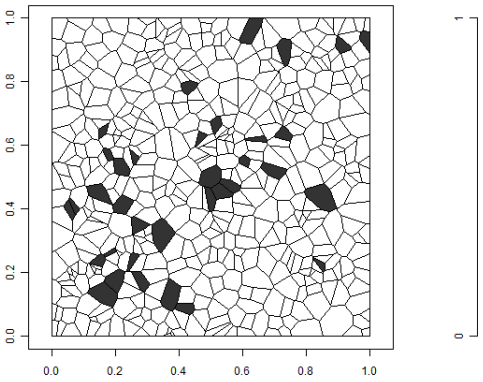
A Figura 6.32 mostra que, como nos cenários anteriores, há mais indivíduos com valores altos de $\widehat{Var}(\hat{p}_i)$ quando se considera as maiores arestas. A configuração do viés (Figura 6.33) é semelhante em todos os métodos, exceto pelo fato de que as menores arestas possuem mais pontos superestimando as probabilidades dentro do *cluster* verdadeiro.



(a) Cluster simulado.



(b) Diagrama de Voronoi.



(c) Casos observados no diagrama de Voronoi

Figura 6.27: Cenário 3: mapa com cluster em formato “L”.

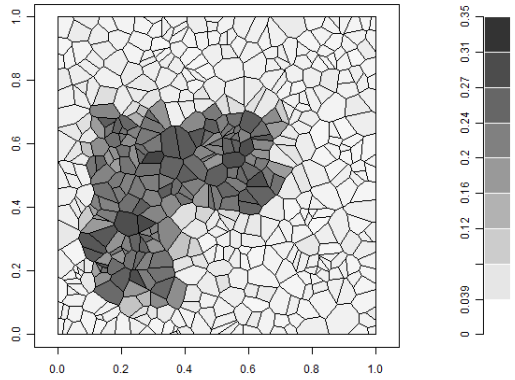


Figura 6.28: Cenário 3: médias das probabilidades, metade da menor aresta.

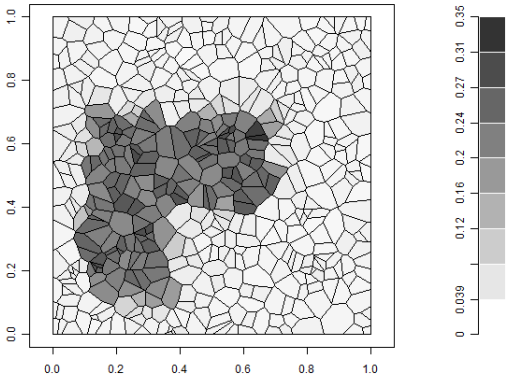


Figura 6.29: Cenário 3: médias das probabilidades, total da menor aresta.

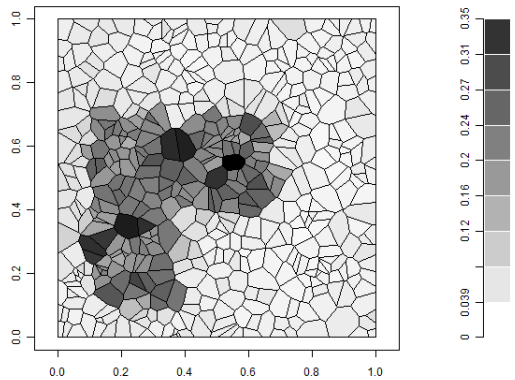


Figura 6.30: Cenário 3: médias das probabilidades, metade da maior aresta.

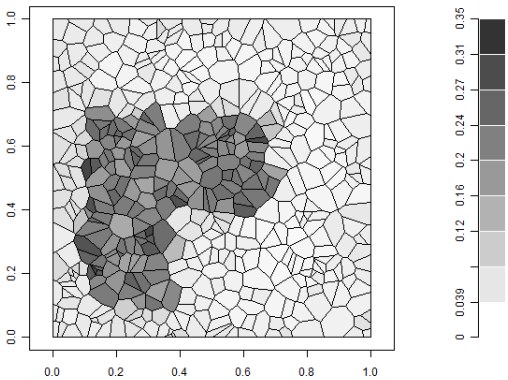


Figura 6.31: Cenário 3: médias das probabilidades, total da maior aresta.

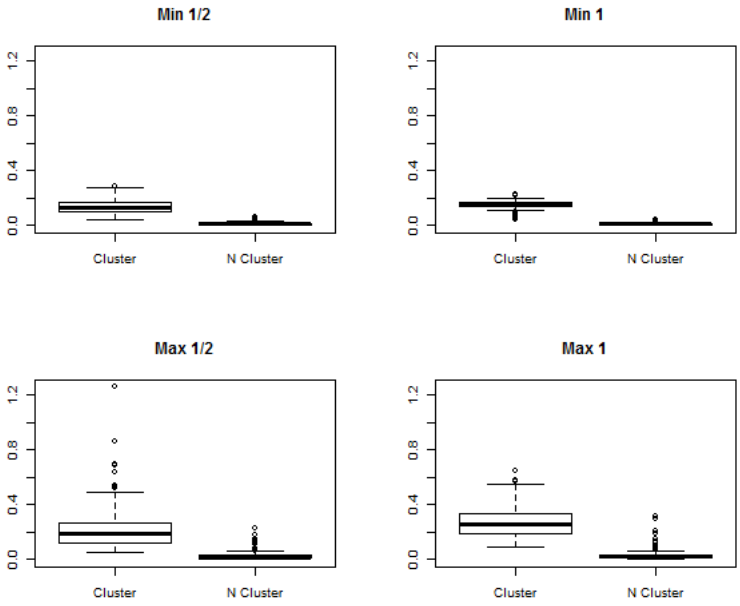


Figura 6.32: Cenário 3: estimativas individuais de variância ($Var(\hat{p}_i)$)

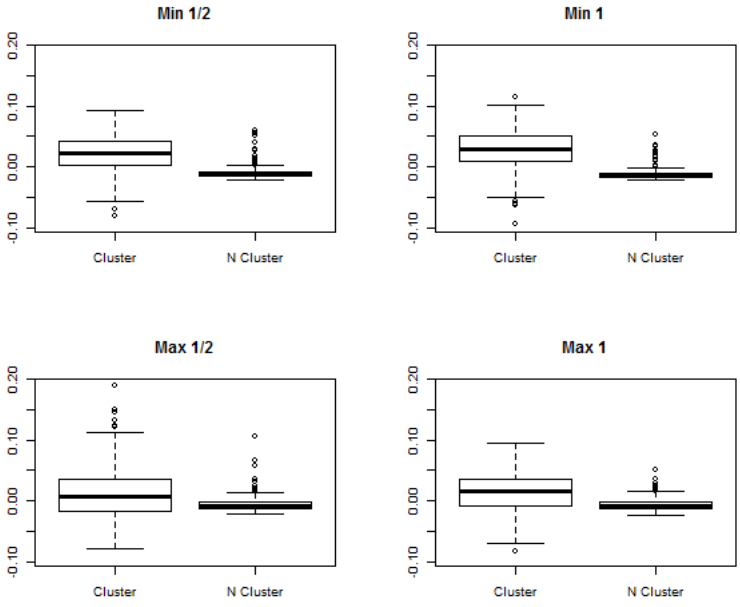
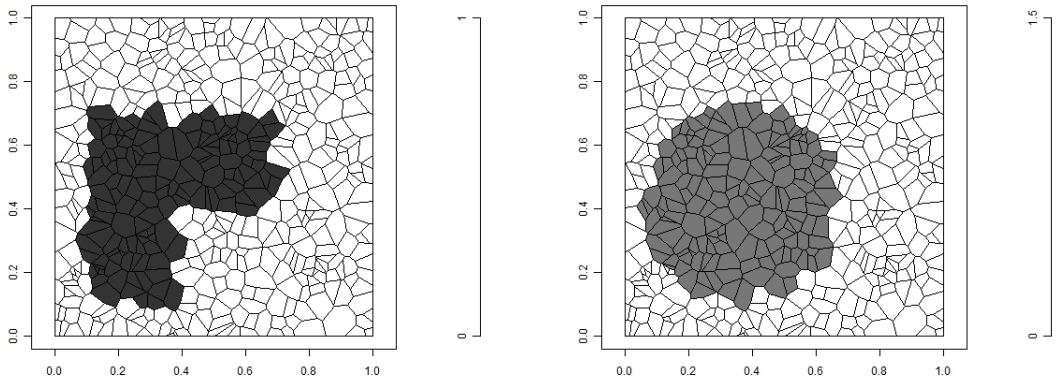


Figura 6.33: Cenário 3: estimativas individuais de viés ($Vies(\hat{p}_i)$)

6.3.2 Medidas de intensidade na base simulada

O *cluster* mais verossímil detectado está na Figura 6.34. Note que a região "L" está totalmente inserida na região com maior razão de verossimilhanças, razão esta significativa a 90% (ver Tabela 6.8).



(a) Cluster real.

(b) Cluster detectado.

Figura 6.34: Cenário 3: cluster verdadeiro × cluster detectado.

Tabela 6.8: Cenário 3: log da razão de verossimilhanças observada × valores críticos.

Detectado	p99	p95	p90
8.68	10.62	8.73	8.27

Os mapas das medidas de intensidade representadas estão nas Figuras 6.35, 6.36, 6.37 e 6.38. Aparentemente, não há diferenças tão notáveis entre os métodos. Repare que, mais uma vez, o nível das intensidades dentro do *cluster* foi menor que o observado no cenário 1 (Tabela 6.9).

A análise da distribuição das medidas de intensidade, através do *box-plot* (Figura 6.39), mostra que há mais indivíduos dentro do cluster verdadeiro com medidas próximas a 1 quando se considera a metade da menor aresta. Com a metade da *maior* aresta, entretanto, há mais *acurárcia*: menos indivíduos fora do *cluster* com grandes medidas de intensidade. Este mesmo método resulta em uma maior diferença entre os níveis das intensidades dentro e fora da região.

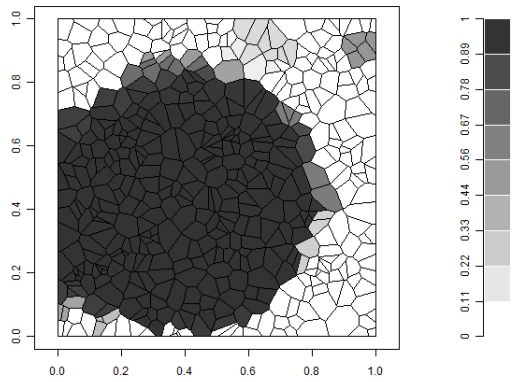


Figura 6.35: Cenário 3: medidas de intensidade, metade da menor aresta.

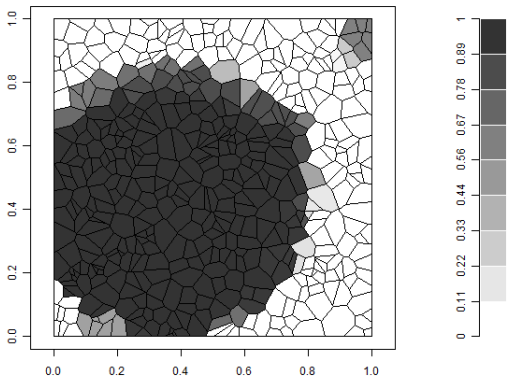


Figura 6.36: Cenário 3: medidas de intensidade, total da menor aresta.

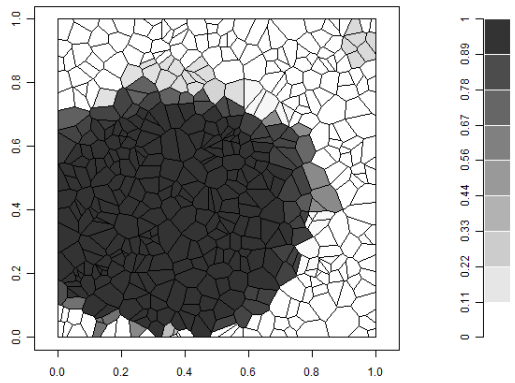


Figura 6.37: Cenário 3: medidas de intensidade, metade da maior aresta.

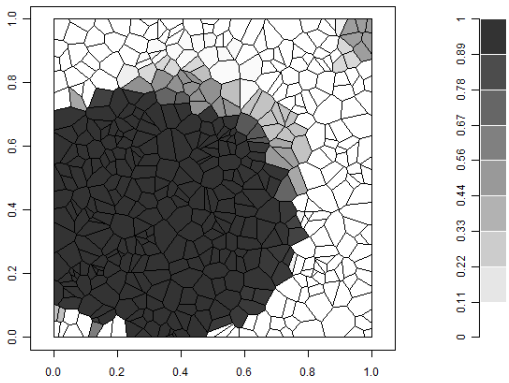


Figura 6.38: Cenário 3: medidas de intensidade, total da maior aresta.

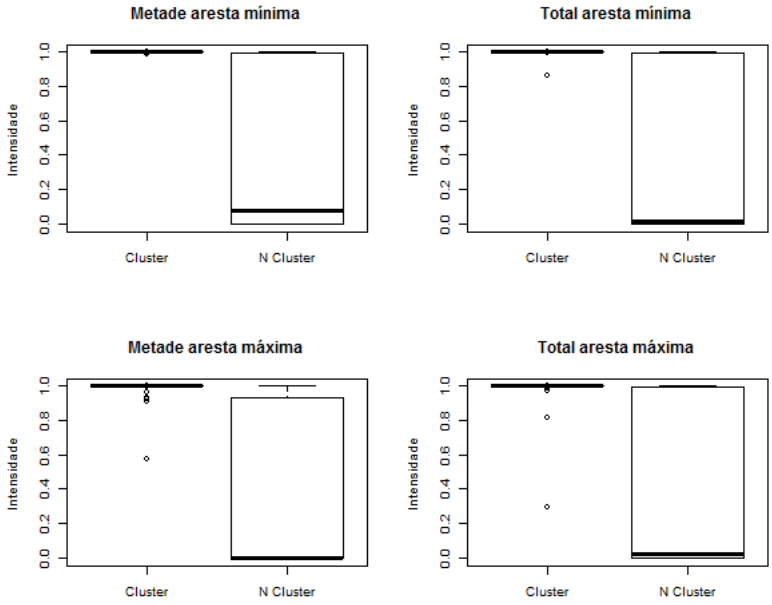


Figura 6.39: Cenário 3: distribuição das medidas de intensidade dentro e fora do cluster verdadeiro.

Tabela 6.9: Cenário 3: medidas de intensidade médias dentro e fora do *cluster*.

	Cluster	N Cluster	Diferença
Metade da aresta mínima	0.99958	0.42660	0.57298
Total da aresta mínima	0.99867	0.43224	0.56643
Metade da aresta máxima	0.99391	0.35463	0.63928
Total da aresta máxima	0.98513	0.37401	0.61112

Além destes resultados, obtidos com janelas contendo até $K = 168$ vizinhos, as medidas de intensidade foram calculadas baseando-se em janelas circulares menores, contendo até $K = 50$ vizinhos (10% do total de pontos). As Figuras 6.40, 6.41, 6.42 e 6.43 mostram um resultado bastante interessante: as "nuvens" de intensidade apresentaram um formato em "L", aproximando-se do formato do *cluster* real.

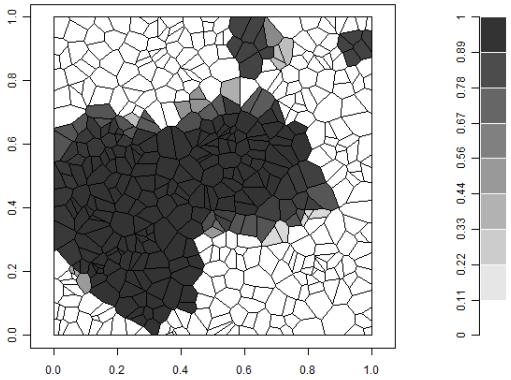


Figura 6.40: Cenário 3: medidas de intensidade com janelas menores, metade da menor aresta.

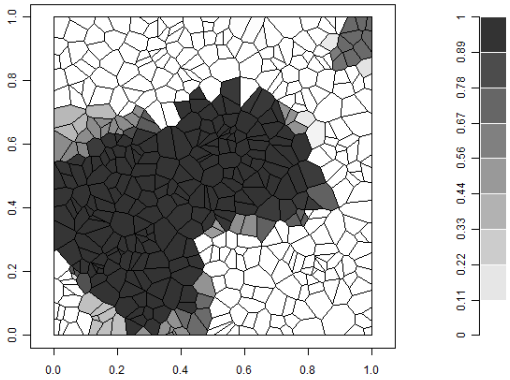


Figura 6.41: Cenário 3: medidas de intensidade com janelas menores, total da menor aresta.

6.4 Cenário 4: cluster circular duplo

O cenário com *cluster* duplo foi construído definindo-se duas regiões circulares com raios diferentes. Na verdade, trata-se de dois *clusters* com diferentes riscos relativos. O procedimento de simulação foi o seguinte: primeiro, definiu-se o conjunto percetencente à primeira região, bem como o primeiro risco relativo e o número c_z simulado de casos. Em seguida, o mesmo procedimento foi realizado para definir o número de casos na segunda região, porém excluindo-se todos os pontos pertencentes ao primeiro. A região definida segue na Figura 6.44, e a descrição do conjunto dentro e fora dos *clusters* segue na Tabela 6.10.

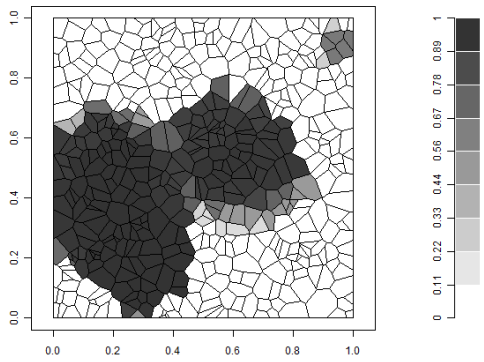


Figura 6.42: Cenário 3: medidas de intensidade com janelas menores, metade da maior aresta.

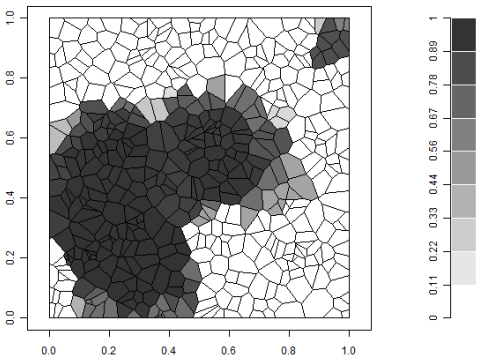
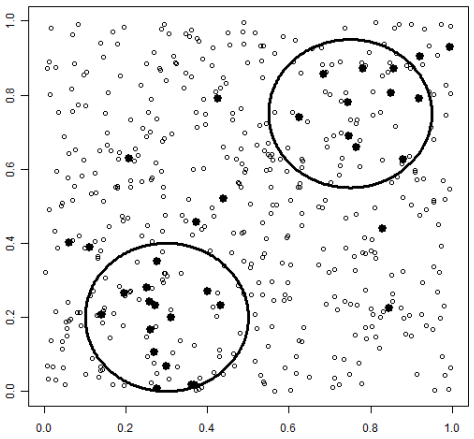


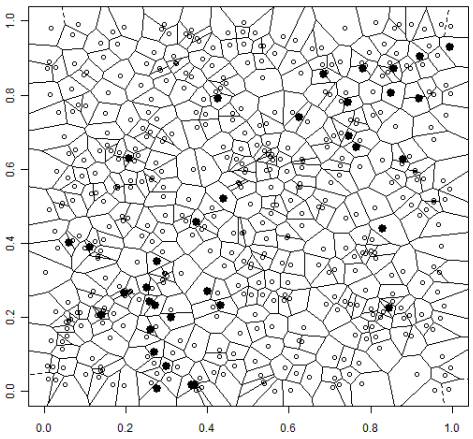
Figura 6.43: Cenário 3: medidas de intensidade com janelas menores, total da maior aresta.

Tabela 6.10: Cenário 4: informações dentro e fora do cluster

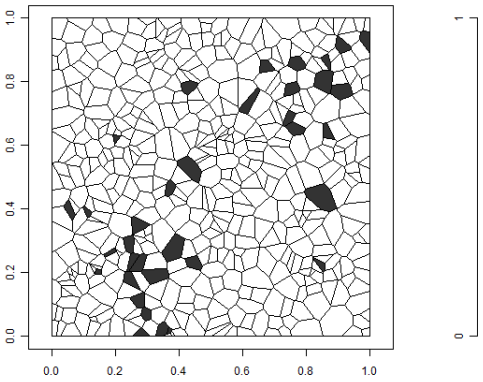
	Cluster 1	Cluster 2	N Cluster
N° de Casos	10.00	15.00	10.00
N° de pontos	59.00	64.00	377.00
Risco relativo teórico	6.34	26.68	1.00
Risco relativo observado	2.42	3.35	0.38



(a) Cluster simulado.



(b) Diagrama de Voronoi.



(c) Casos observados no diagrama de Voronoi

Figura 6.44: Cenário 4: mapa com cluster em formato duplo.

6.4.1 Viés e Expectância dos estimadores $\hat{p}$

O comportamento das espectâncias é semelhante aos outros cenários (Figuras 6.45, 6.46, 6.47 e 6.48).

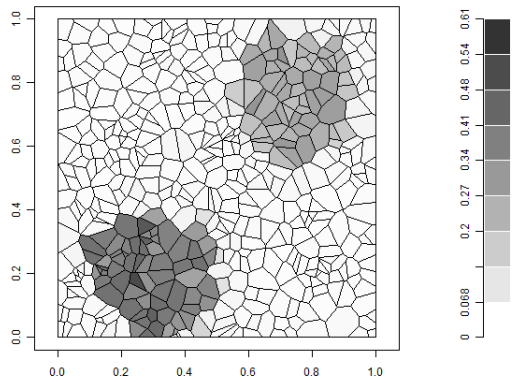


Figura 6.45: Cenário 4: médias das probabilidades, metade da menor aresta.

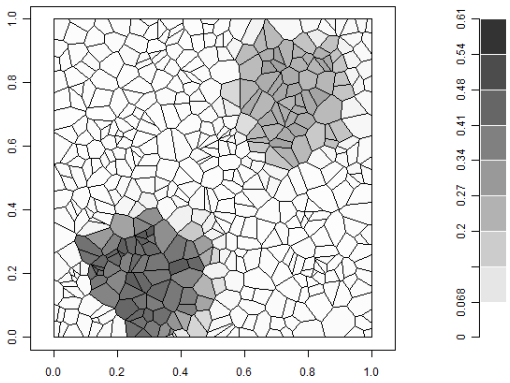


Figura 6.46: Cenário 4: médias das probabilidades, total da menor aresta.

Mais uma vez, as variâncias são maiores quando se considera as maiores arestas (Figura 6.49). Seguindo o padrão dos cenários anteriores, as probabilidades individuais são superestimadas dentros dos *clusters* (Figura 6.50). O viés é maior no *cluster* com maior risco relativo (*cluster* 2). Ao mesmo tempo, a variabilidade nesse *cluster* é menor.

6.4.2 Medidas de intensidade na base simulada

O cluster detectado é bem aproximado de uma das regiões (Figura 6.51). O cluster é estatisticamente significativo a 99% (Tabela 6.11). Foi possível avaliar, também, a significância do segundo *cluster*. Para isso, retirou-se do conjunto de dados aqueles indivíduos pertencentes ao *cluster* mais verossímil detectado. Em seguida, procedeu-se com uma nova detecção que resultou em uma razão de verossimilhanças $LLR = 11.33$, ou seja, também significante. A região está representada na Figura 6.51(c).

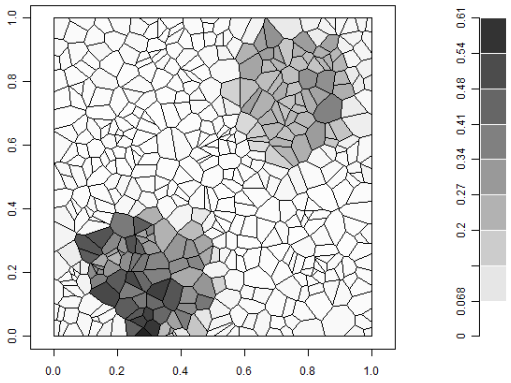


Figura 6.47: Cenário 4: médias das probabilidades, metade da maior aresta.

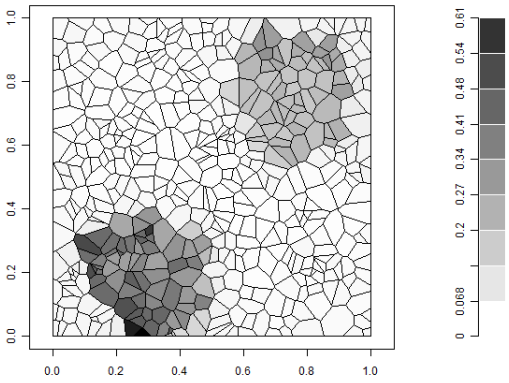


Figura 6.48: Cenário 4: médias das probabilidades, total da maior aresta.

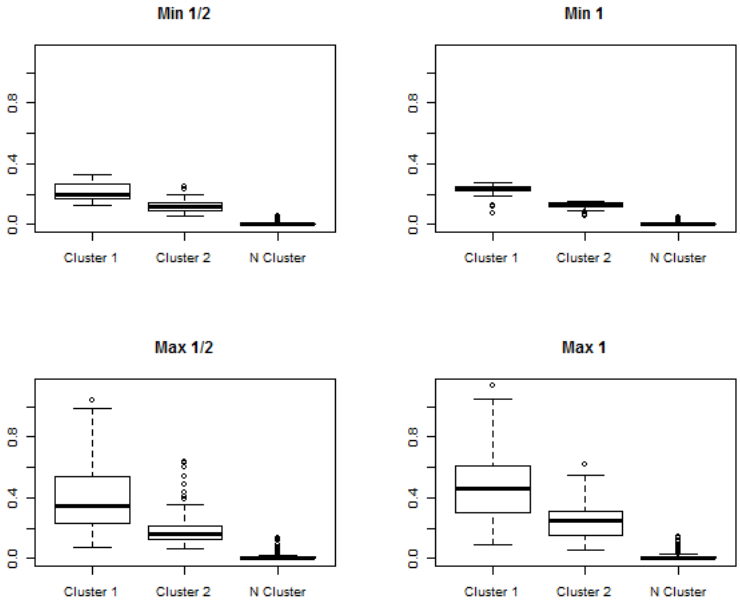


Figura 6.49: Cenário 4: estimativas individuais de variância ($Var(\hat{p}_i)$)

Tabela 6.11: Cenário 4: log da razão de verossimilhanças observada × valores críticos.

Detectado	p99	p95	p90
11.48	10.73	8.88	7.96

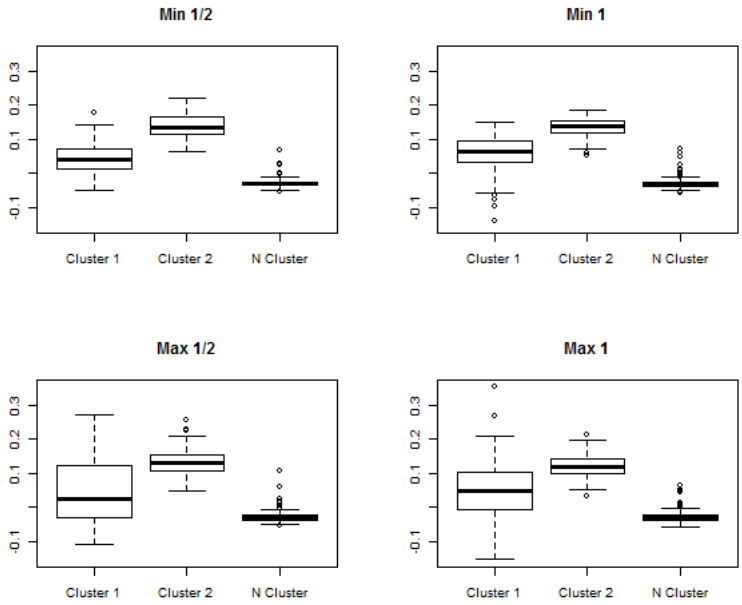
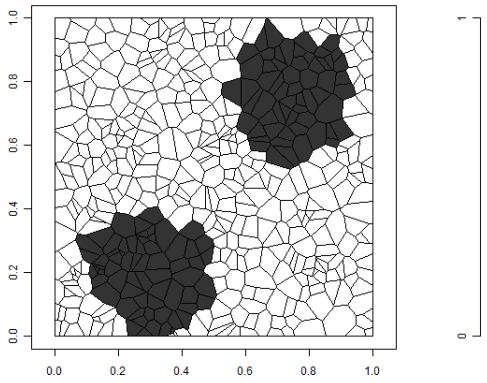


Figura 6.50: Cenário 4: estimativas individuais de viés ($Vies(\hat{p}_i)$)

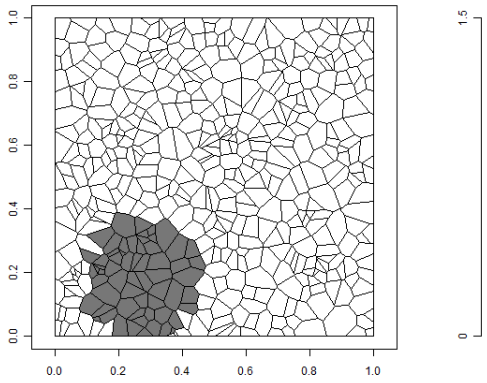
Os níveis de intensidade são maiores dentro dos dois *clusters* em todos os métodos (Figuras 6.52, 6.53, 6.54 e 6.55). Entretanto, o método baseado na metade da aresta máxima pareceu destacar melhor os *clusters* verdadeiros, principalmente o secundário (Figura 6.56 e Tabela 6.12, *cluster* 2 é o *cluster* com menor razão verossimilhanças).

Tabela 6.12: Cenário 4: medidas de intensidade médias por *cluster*.

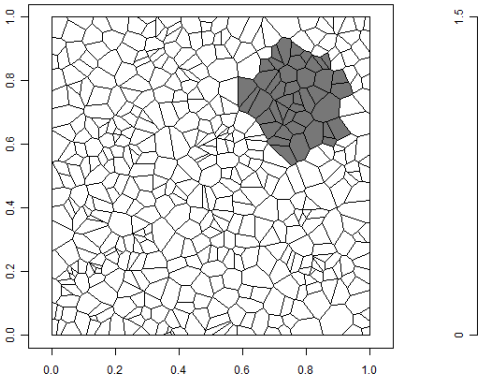
	Cluster 1	Cluster 2	N Cluster	Diferença 1	Diferença 2
Metade da aresta mínima	0.9998	0.7951	0.2920	0.7078	0.5031
Total da aresta mínima	0.9993	0.5759	0.2349	0.7644	0.3410
Metade da aresta máxima	0.9985	0.8399	0.1823	0.8161	0.6575
Total da aresta máxima	0.9993	0.2765	0.1769	0.8225	0.0996



(a) Cluster real.



(b) Cluster detectado.



(c) Cluster 2 detectado.

Figura 6.51: Cenário 4: cluster verdadeiro × cluster detectado.

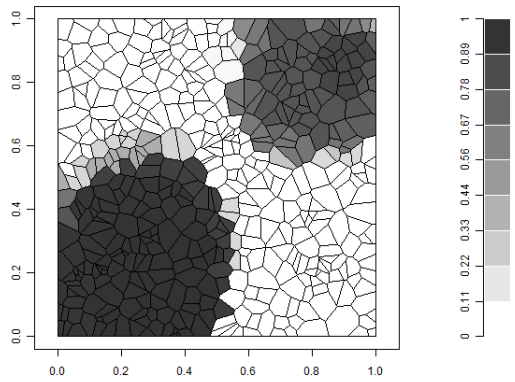


Figura 6.52: Cenário 4: medidas de intensidade, metade da menor aresta.

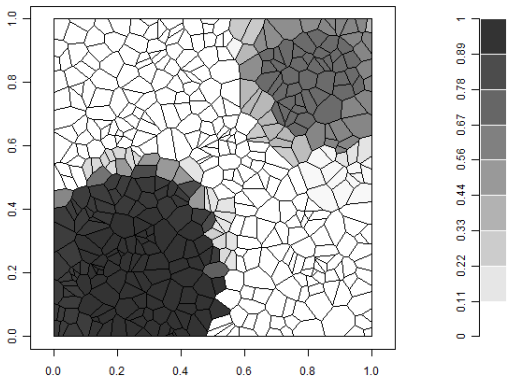


Figura 6.53: Cenário 4: medidas de intensidade, total da menor aresta.

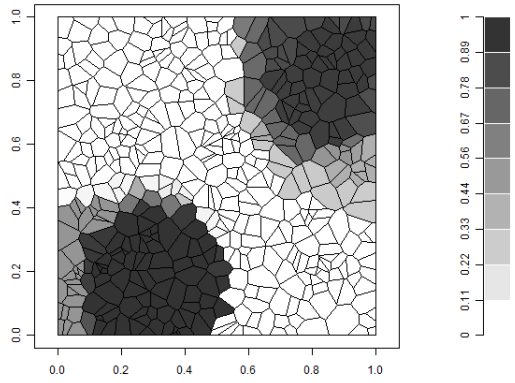


Figura 6.54: Cenário 4: medidas de intensidade, metade da maior aresta.

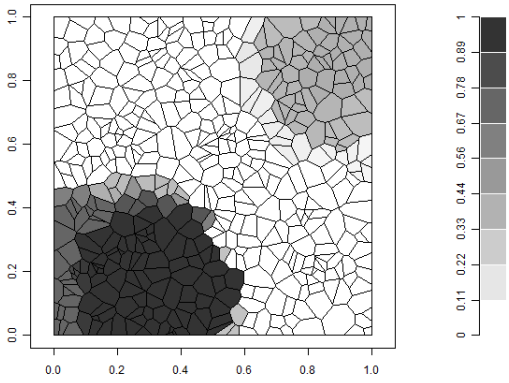


Figura 6.55: Cenário 4: medidas de intensidade, total da maior aresta.

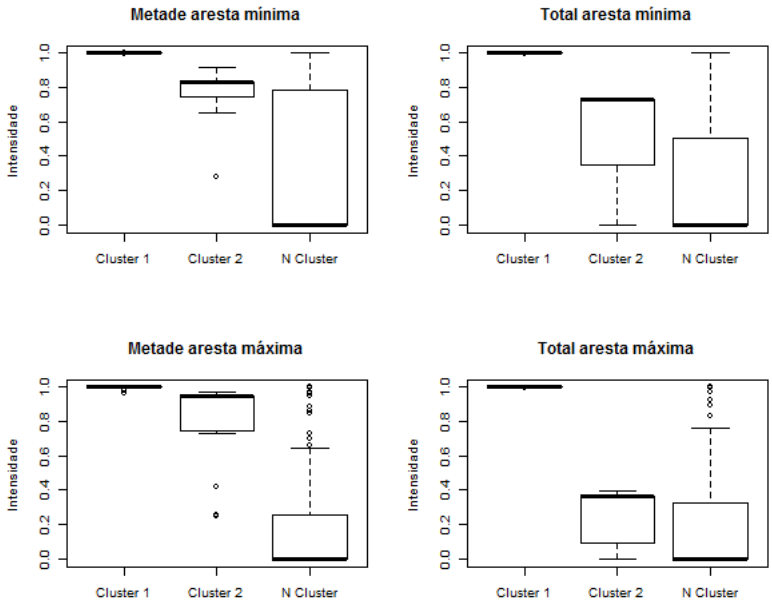


Figura 6.56: Cenário 4: distribuição das medidas de intensidade dentro e fora do cluster verdadeiro.

6.5 Cenário 5: cluster circular em mapa com diferentes densidades

A heteregoneidade de densidades no mapa foi gerada considerando-se distribuições uniformes diferentes na geração das 500 coordenadas. Um quarto dos dados foi gerado de duas uniformes $U(0, 1)$ independentes. O restante foi gerado de duas distribuições $U(0.3, 0.6)$ independentes, criando assim uma região com uma concentração maior de dados, como na Figura 6.57. As estatísticas de número de casos dentro e fora do *cluster* seguem na Tabela 6.13.

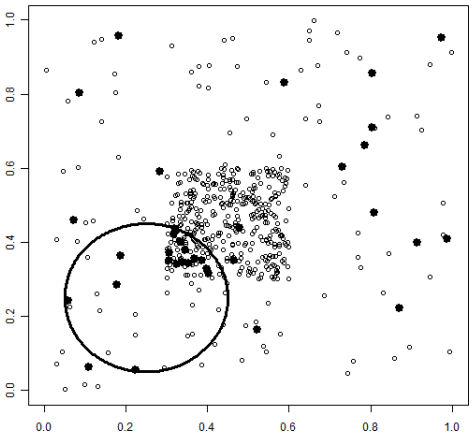
6.5.1 Viés e Expectância dos estimadores $\hat{p}$

As expectâncias seguem o mesmo padrão já observado nos outros cenários (Figuras 6.58, 6.59, 6.60 e 6.61).

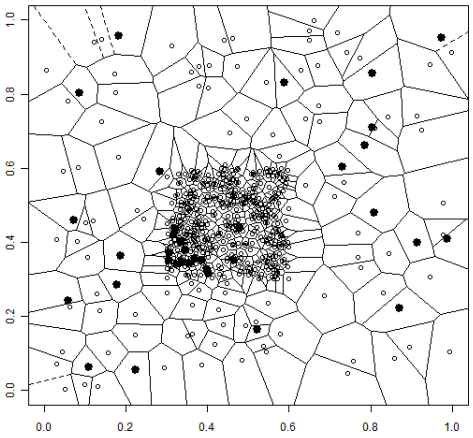
Novamente, há mais pontos com variâncias maiores quando se considera as maiores arestas (Figura 6.62). Aqui, as probabilidades individuais também são superestimadas dentro dos *clusters* (Figura 6.63).

Tabela 6.13: Cenário 5: informações dentro e fora do cluster

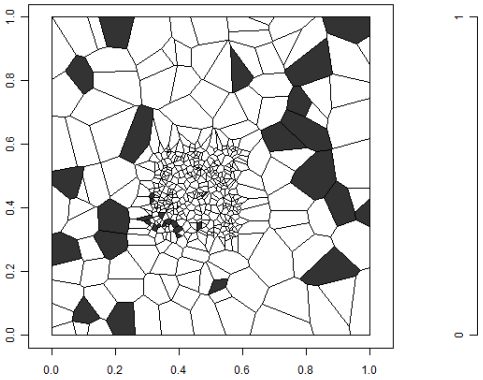
	Cluster	N Cluster
N° de Casos	17.00	18.00
N° de pontos	86.00	414.00
Risco relativo teórico	2.39	1.00
Risco relativo observado	2.82	0.62



(a) Cluster simulado.



(b) Diagrama de Voronoi.



(c) Casos observados no diagrama de Voronoi

Figura 6.57: Cenário 5: cluster em mapa com densidade heterogênea.

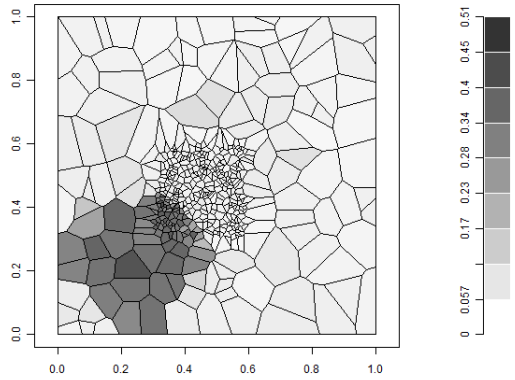


Figura 6.58: Cenário 5: médias das probabilidades, metade da menor aresta.

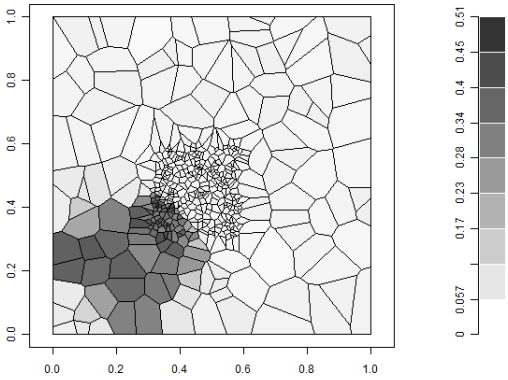


Figura 6.59: Cenário 5: médias das probabilidades, total da menor aresta.

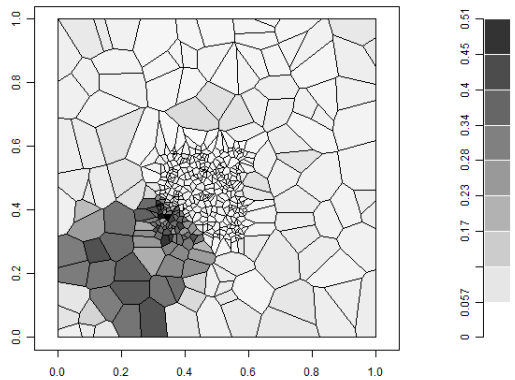


Figura 6.60: Cenário 5: médias das probabilidades, metade da maior aresta.

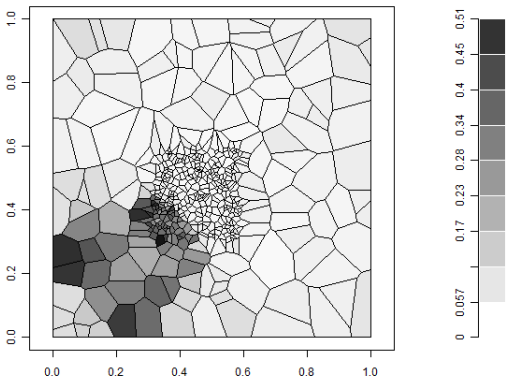


Figura 6.61: Cenário 5: médias das probabilidades, total da maior aresta.

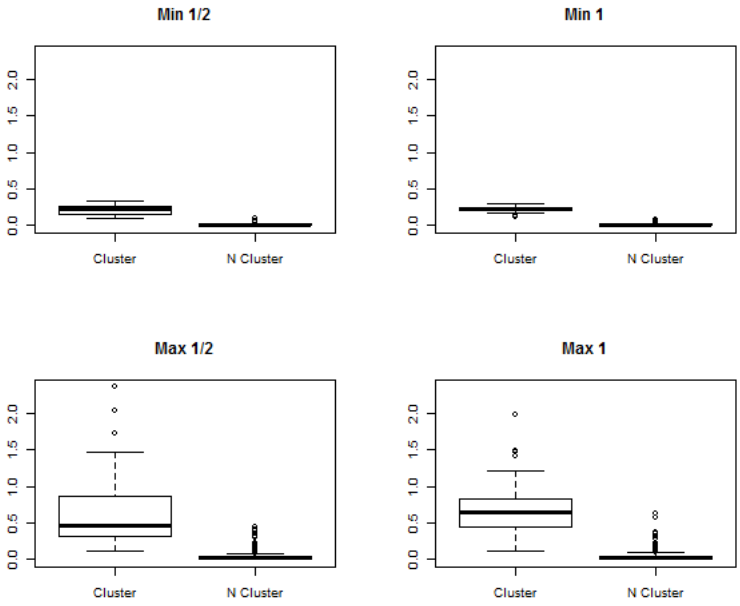


Figura 6.62: Cenário 5: estimativas individuais de variância ($Var(\hat{p}_i)$)

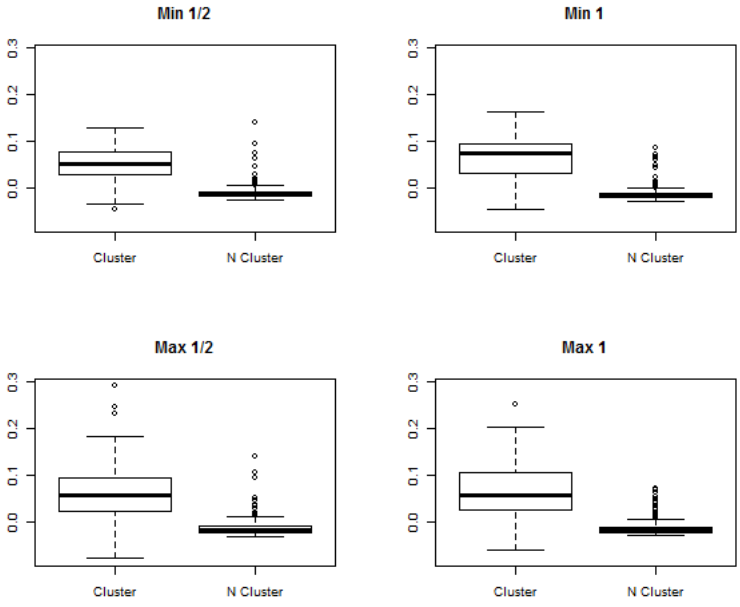
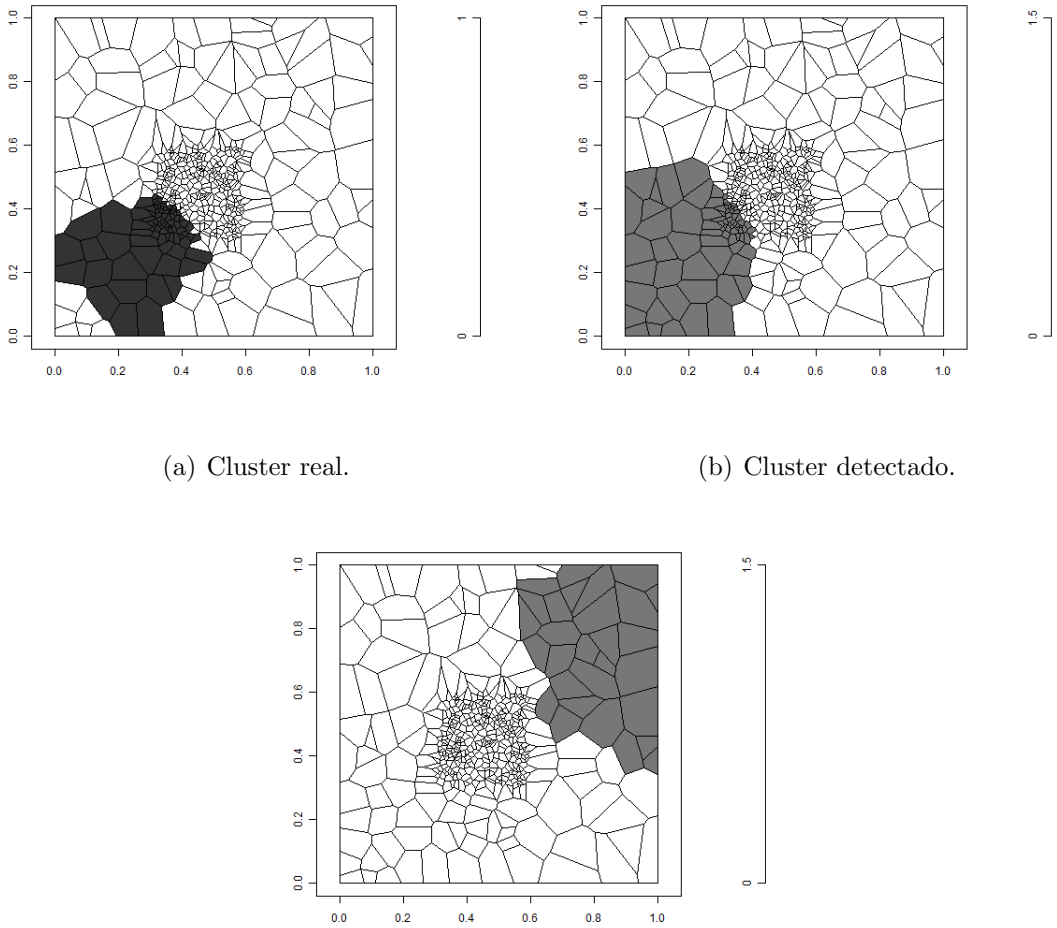


Figura 6.63: Cenário 5: estimativas individuais de viés ($Vies(\hat{p}_i)$)

6.5.2 Medidas de intensidade na base simulada

O *cluster* detectado é bem próximo do verdadeiro, como mostra a Figura 6.64. Além disso, a log-verossimilhança associada à região é significativa a 99% (Tabela 6.14).



(a) Cluster real.

(b) Cluster detectado.

(c) Cluster 2 detectado.

Figura 6.64: Cenário 5: cluster verdadeiro × cluster detectado.

As medidas de intensidade estão representadas nas Figuras 6.65, 6.66, 6.67 e 6.68. Todos os quatro métodos de estimação de probabilidades resultaram em intensidades maiores em uma região particular diferente do *cluster* verdadeiro. Na verdade, por se tratar de uma base gerada aleatoriamente, certas flutuações são sujeitas a acontecer. De fato, essa "nuvem" secundária também é significativa. Seguindo o método da seção anterior, que detectou o segundo *cluster* através de uma varredura em uma nova base sem os indivíduos percentencentes à primeira região detectada, foi possível ver que a

Tabela 6.14: Cenário 5: log da razão de verossimilhanças observada × valores críticos.

Detectado	p99	p95	p90
14.99	10.86	8.51	8.10

região secundária do Cenário 5 é, também, uma região com LLR significativo, onde $LLR = 14.92$. Esse *cluster* secundário está representado na Figura 6.64(c).

O nível de intensidade é semelhante nessas duas reigões quando se considera o método do total da maior aresta. A consideração da metade da maior aresta, entretanto, resulta em níveis maiores de intensidade na região próxima do cluster, além de destacar melhor a diferença com relação as intensidades obtidas fora do cluster (Tabela 6.15 e Figura 6.69).

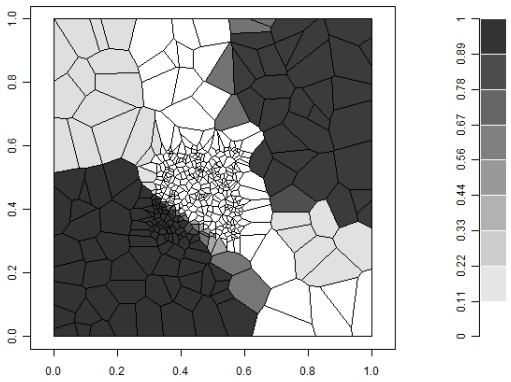


Figura 6.65: Cenário 5: medidas de intensidade, metade da menor aresta.

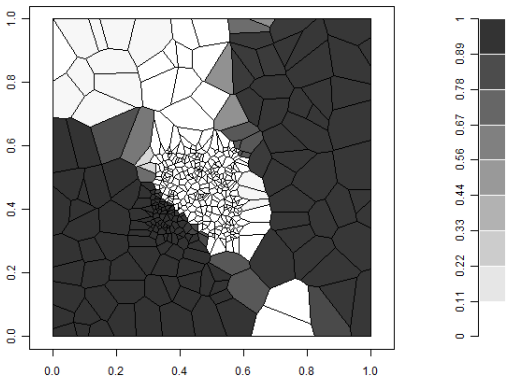


Figura 6.66: Cenário 5: medidas de intensidade, total da menor aresta.

6.6 Conclusões

Para resumir as propriedades inferenciais observadas nos cenários, a Tabela 6.16 informa o Erro Quadrático Médio estimado em cada situação de acordo com (5.11), considerando cada um dos quatro métodos em análise.

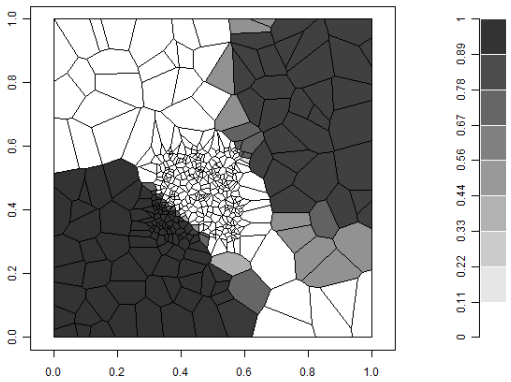


Figura 6.67: Cenário 5: medidas de intensidade, metade da maior aresta.

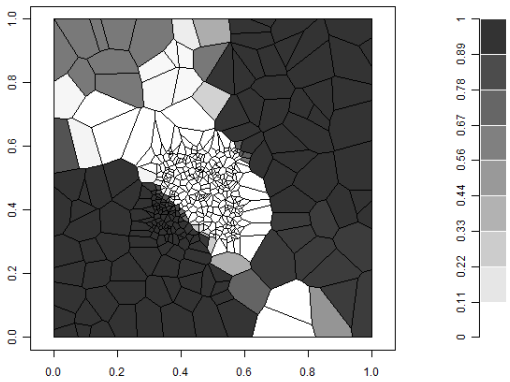


Figura 6.68: Cenário 5: medidas de intensidade, total da maior aresta.

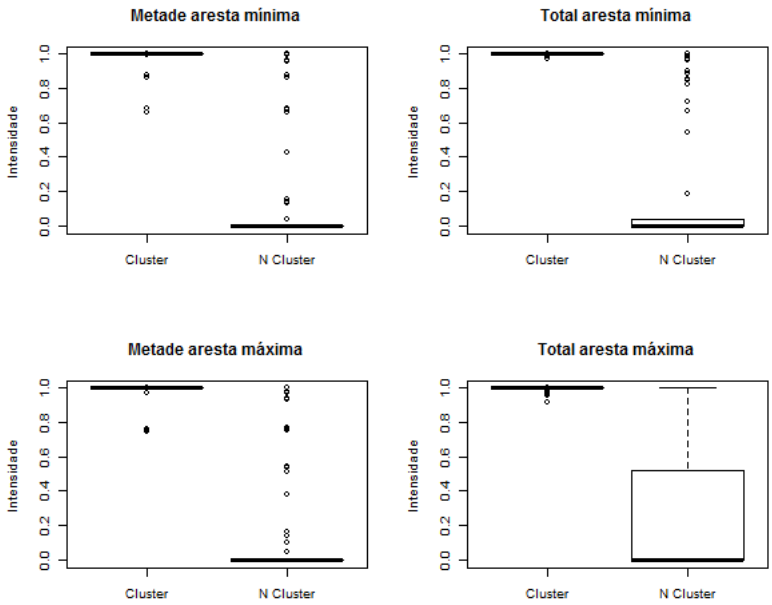


Figura 6.69: Cenário 5: distribuição das medidas de intensidade dentro e fora do cluster verdadeiro.

Tabela 6.15: Cenário 5: medidas de intensidade médias dentro e fora do *cluster*.

	Cluster	N Cluster	Diferença
Metade da aresta mínima	0.97284	0.16511	0.80773
Total da aresta mínima	0.99684	0.21534	0.78149
Metade da aresta máxima	0.98432	0.17148	0.81284
Total da aresta máxima	0.99498	0.24084	0.75414

Tabela 6.16: Medidas de EQM, Variância e Viés totais, por método e cenário.

Medida	Método	Simples	Fraco	Irregular	Duplo	Dens. dif.
Viés	Metade da menor aresta	1.16	0.16	0.28	1.86	0.44
	Total da menor aresta	1.67	0.20	0.36	1.93	0.66
	Metade da maior aresta	1.68	0.25	0.39	2.16	0.87
	Total da maior aresta	1.75	0.34	0.24	2.00	0.88
Variância	Metade da menor aresta	25.01	27.56	25.40	23.38	25.96
	Total da menor aresta	25.56	29.22	26.61	23.72	26.19
	Metade da maior aresta	41.06	39.76	41.03	42.10	69.57
	Total da maior aresta	50.17	49.02	50.64	51.32	75.82
EQM	Metade da menor aresta	26.18	27.72	25.68	25.24	26.40
	Total da menor aresta	27.23	29.42	26.97	25.65	26.85
	Metade da maior aresta	42.74	40.01	41.42	44.26	70.45
	Total da maior aresta	51.92	49.36	50.87	53.32	76.70

A partir dos resultados nos dados simulados em todos os cenários acima, foi possível observar que:

1. As probabilidades estimadas são frequentemente viesadas para cima nos pontos dentro do *cluster*.

2. As estimativas de probabilidades baseadas nas menores arestas são mais precisas e acuradas;

3. Com exceção dos cenários com *cluster* circular simples ("fraco" e "forte"), as medidas de intensidade com base na metade da maior aresta resultaram em maiores diferenças entre os níveis de intensidade dentro e fora do *cluster*;

4. Métodos baseados na metade da menor aresta se mostraram, quase sempre, mais acurados, ou seja, há menos pontos fora de Z com grandes valores da medida de intensidade;

5. Todos os métodos de obtenção de cálculo de intensidades apresentaram alguma sensibilidade nas regiões com maior número de casos concentrados. Tal fato é um indício de que *clusters* secundários são facilmente identificados. Por outro lado, os métodos baseados na metade da *maior* aresta se mostraram mais sensíveis, destacando melhor os *clusters* secundários e apresentando, em quase todos os cenários, maiores níveis de intensidade dentro do *cluster* verdadeiro;

6. O cluster fraco resultou em níveis de intensidade menores, além de uma menor diferenciação entre os níveis dentro e fora da região de anomalia simulada;

7. No cluster em "L", foi possível destacar o formato verdadeiro da região com a restrição do tamanho da janela;

Uma vez que valores *maiores* de intensidade estão associadas a uma incerteza *menor* da existência de *clusters* espaciais, é razoável supor que essas medidas sejam maiores dentro do grupo do que fora deste. Isso motiva o enfoque dado nessa característica, cujos resultados representam indícios de que, em termos de sensibilidade à existência de *clusters* em outras regiões, o método da metade das maiores arestas é a melhor opção em "janelas" circulares.

Há uma explicação intuitiva razoável para o fato de essa metodologia implicar em maiores intensidades dentro do *cluster* verdadeiro. Quando se troca a menor aresta pela maior, pontos mais "afastados", isto é, nos quais incide apenas uma aresta do MST, continuam com o mesmo raio de círculo de influência. Pontos mais próximos entre si, com mais de uma aresta do MST incidindo, passam a ter seus círculos aumentados. Isso diminui suas probabilidades, aumentando entretanto as probabilidades nos não-casos em volta que, na situação anterior, não estavam inseridos em nenhum círculo. Ocorre então um processo de "fechamento" dessa região em torno do *cluster* verdadeiro. Na realidade, pode-se observar na Figura 5.3 que a região com maiores probabilidades coincide com o MST. Esse "fechamento" pode ser responsável também pela maior ocorrência de grandes variabilidades nas probabilidades estimadas através das maiores arestas.

Os níveis menores de intensidade dentro do *cluster*, no cenário 2, levam a crer que *clusters* pouco significativos resultam em regiões com níveis menores de intensidade, bem como regiões de incertezas mais extensas.

O cenário 5 foi construído com o intuito de se verificar a influência de densidades heterogêneas no cálculo das medidas de intensidade. No Capítulo 4 foi mecionada a necessidade de se levar em conta essa diferença de concentrações, principalmente quando se considera um MST com base em distâncias euclidianas. Ainda no referido Capítulo, o VMST foi mencionado como uma medida que leva em conta as densidades desiguais.

O estudo do EQM das probabilidades revelou que as expectâncias continuam mais altas no *cluster* verdadeiro, não sofrendo influência da densidade diferente. As medidas de intensidade com base nas probabilidades estimadas via VMST foram calculadas nos cenários 1 (Figuras 6.70, 6.71, 6.72 e 6.73) e 5 (Figuras 6.74, 6.75, 6.76 e 6.77). Os níveis de intensidade são maiores fora do *cluster* do que quando se considera o MST com distâncias euclidianas, indicando uma precisão maior neste último.

No caso de diferentes densidades locais, o *box-plot* das medidas de intensidade em 6.78 mostra que há menos pontos discrepantes do que em 6.69. Entretanto, o MST euclidiano ainda destaca melhor a diferença entre os valores dentro do cluster e fora do cluster (compare as Tabelas 6.15 e 6.17).

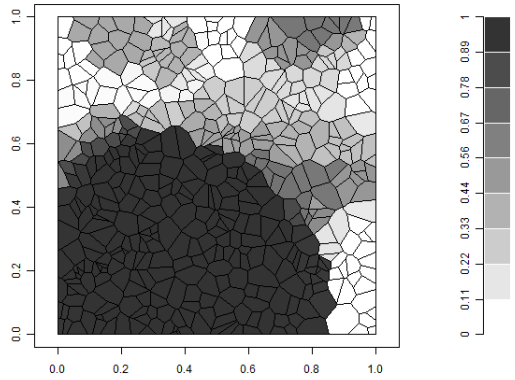


Figura 6.70: Cenário 1: medidas de intensidade, metade da menor aresta do VMST.

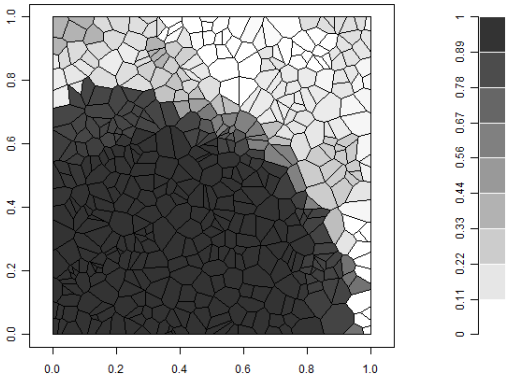


Figura 6.71: Cenário 1: medidas de intensidade, total da menor aresta do VMST.

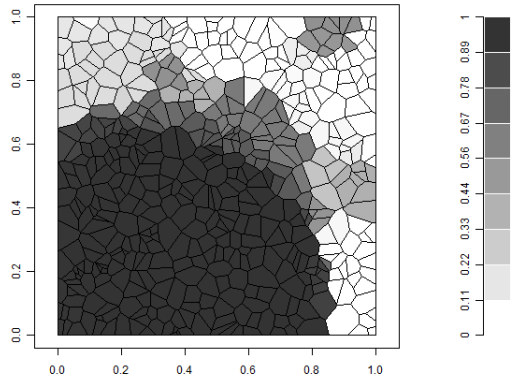


Figura 6.72: Cenário 1: medidas de intensidade, metade da maior aresta do VMST.

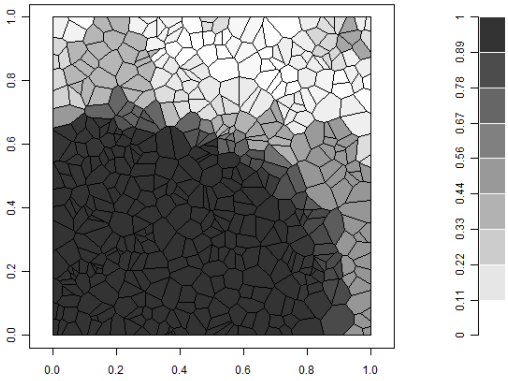


Figura 6.73: Cenário 1: medidas de intensidade, total da maior aresta do VMST.

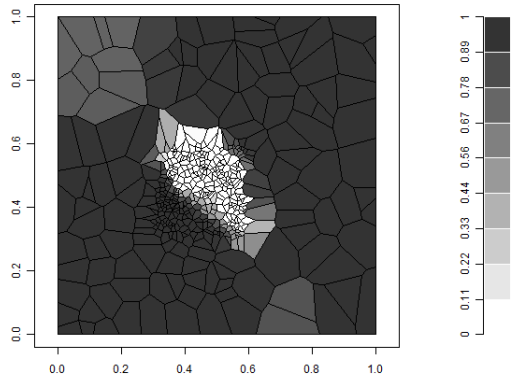


Figura 6.74: Cenário 5: medidas de intensidade, metade da menor aresta do VMST.

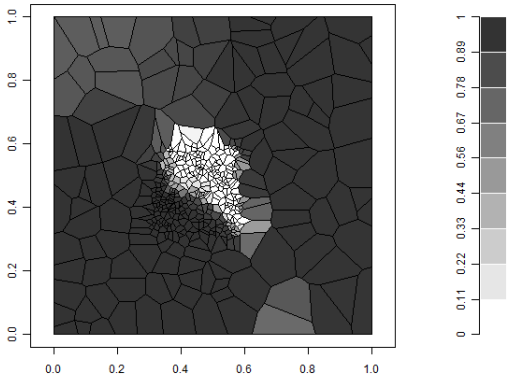


Figura 6.75: Cenário 5: medidas de intensidade, total da menor aresta do VMST.

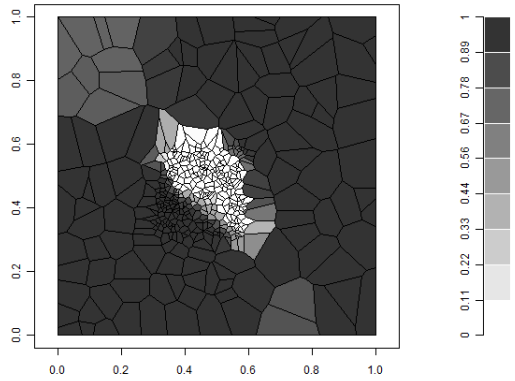


Figura 6.76: Cenário 5: medidas de intensidade, metade da maior aresta do VMST.

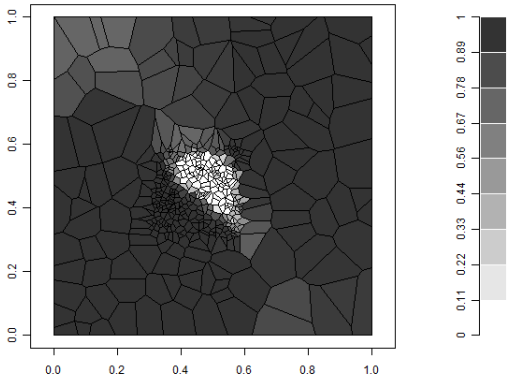


Figura 6.77: Cenário 5: medidas de intensidade, total da maior aresta do VMST.

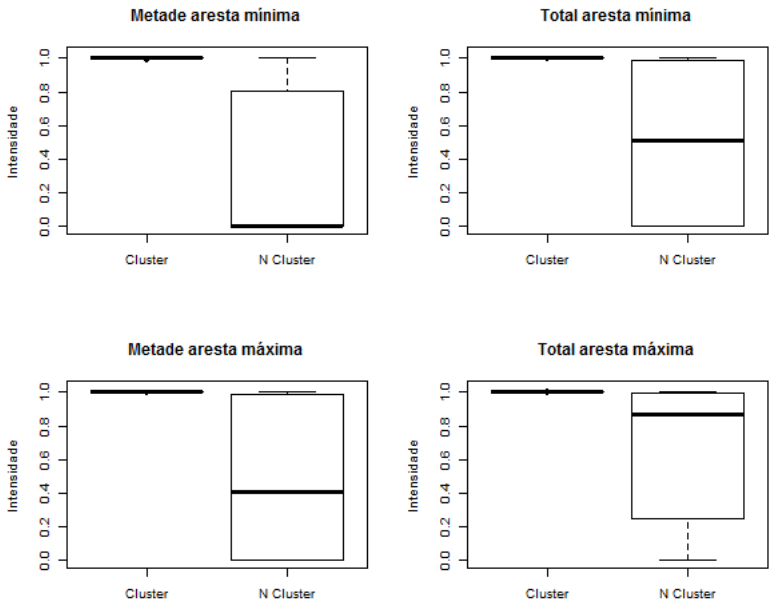


Figura 6.78: Cenário 5: distribuição das medidas de intensidade dentro e fora do cluster verdadeiro (VMST).

Tabela 6.17: Cenário 5: medidas de intensidade médias dentro e fora do *cluster* (VMST).

	Cluster	N Cluster	Diferença
Metade da aresta mínima	0.99912	0.32568	0.67344
Total da aresta mínima	1.00000	0.52011	0.47989
Metade da aresta máxima	0.99987	0.45052	0.54935
Total da aresta máxima	0.99997	0.65167	0.34829

Capítulo 7

Estudo de caso: ocorrências de dengue em Lassance-MG

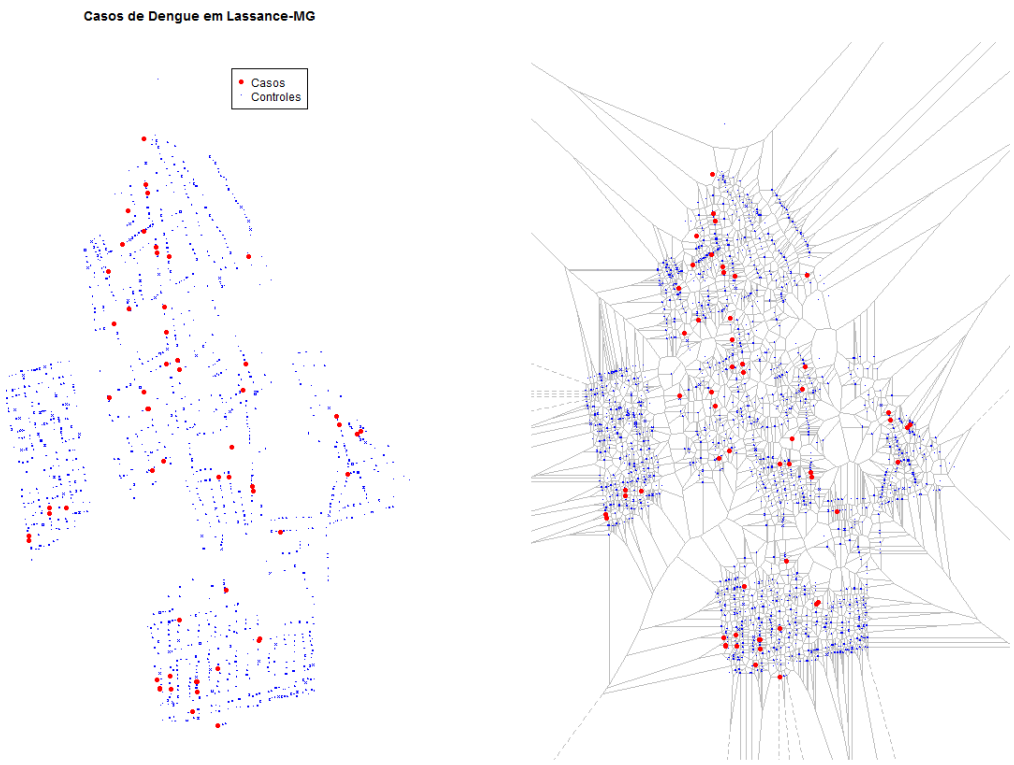
Para ilustar o funcionamento dos métodos descritos no capítulo anterior, utilizou-se a base de dados resultante do trabalho desenvolvido em [Duczmal *et al.*, 2011]. Trata-se de um conjunto de observações do tipo caso-controle de ocorrência casos de dengue. A motivação apresentada para a iniciativa de coleta dessas informações baseou-se na estimativa de que apenas 10% dos casos de dengue são regularmente registrados em hospitais ou em centros de saúde.

A coleta dessas informações fez parte de um projeto piloto, visando a construção de dados confiáveis a nível individual. Os dados foram obtidos com o auxílio de agentes de saúde do Programa Saúde da Família, que realizaram visitas semanais às residências do município entre janeiro e junho de 2010.

A base contém 3.986 indivíduos, dos quais 57 são casos de dengue. Esse conjunto de informações foi disponibilizado através de um *link* citado no referido trabalho.

Os 3.986 pontos estão representados na Figura 7.1, sendo os 57 casos representados em vermelho. O *cluster* detectado através do *Scan* circular, representado na Figura 7.2, possui 5 casos entre 27 pontos. O logaritmo da razão de verossimilhanças, bem como seus valores críticos obtidos via Monte Carlo, seguem na Tabela 7.1. Note que o *cluster* detectado através do *scan* circular não é significativo a 90%. Entretanto, naquele trabalho o *cluster* detectado através do VBScan foi significativo.

As medidas de intensidade, estimadas via método da metade da maior aresta,



(a) Dados observados

(b) Diagrama de Voronoi.

Figura 7.1: Estudo de caso: dados observados.

Tabela 7.1: Estudo de caso: razão de verossimilhanças observada × valores críticos.

Detectado	p99	p95	p90
8.63	12.81	9.95	9.01

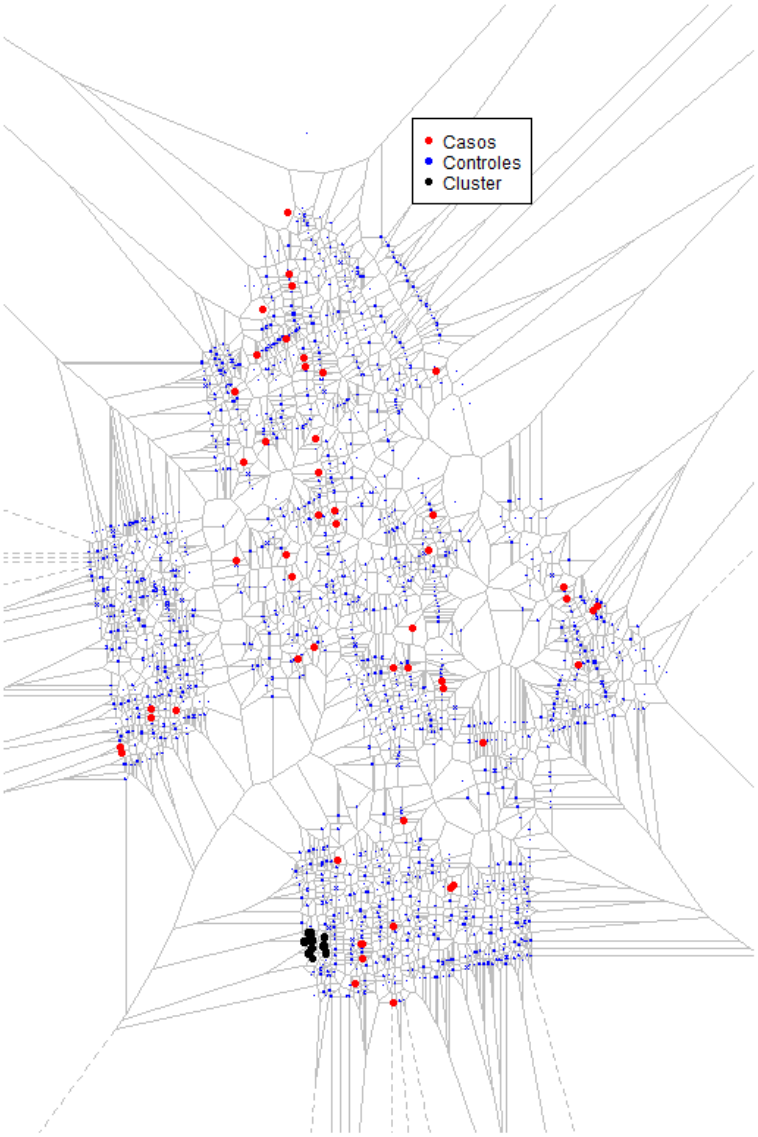


Figura 7.2: Estudo de caso: cluster detectado

seguem representadas na Figura 7.3. Repare que além da região do *cluster* detectado, há outras regiões com níveis de intensidade elevados. São regiões com risco elevado de ocorrência de dengue, que foram detectadas com boa precisão mesmo levando em conta que o método circular resultou em um valor não significativo. Duas dessas regiões são mais ou menos próximas dos dois *clusters* mais verossímeis encontrados em [Duczmal *et al.*, 2011] através do VBScan.

É interessante notar também que, mesmo usando um método mais simples, o *Scan* circular, as medidas de intensidade revelaram um resultado semelhante ao obtido com os métodos mais sofisticados citados, que exigem maior aprimoramento computacional. Tal fato revela o potencial inerente ao uso de tais medidas, tanto na questão de delineamento de incertezas e observação de regiões de influência, quanto na observância de pontos secundários no mapa onde também há "excessos de ocorrência".

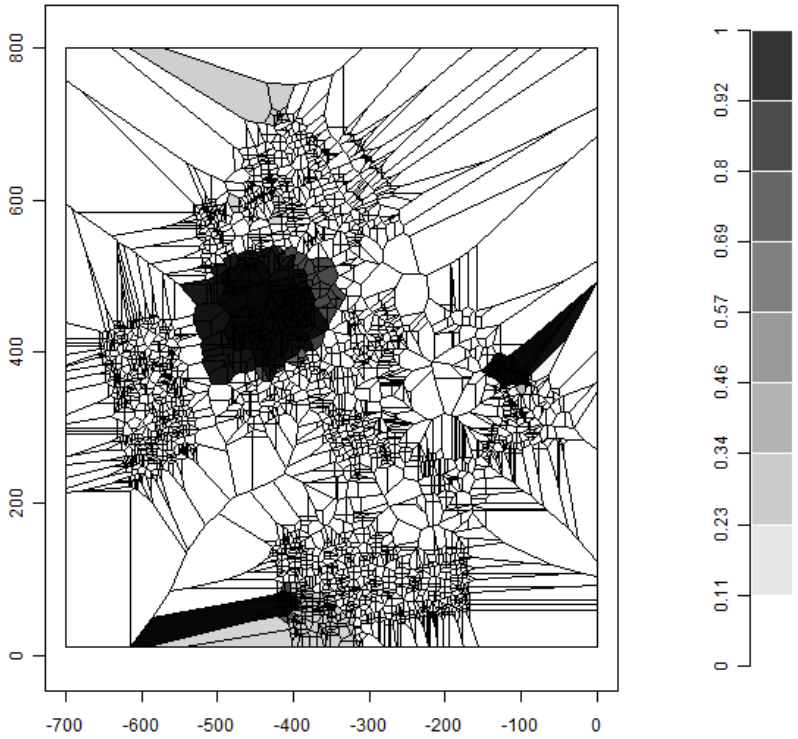


Figura 7.3: Estudo de caso: medidas de intensidade utilizando a metade da maior aresta.

Capítulo 8

Considerações Finais

A estatística λ defnida em (2.6) e (2.12) possue propriedades essenciais desejáveis no teste significância do cluster mais provável identificado, tanto para dados pontuais quanto para dados agregados. O esforço, não trivial, está na busca do clusters Z que maximizem a razão de verossimilhança e que aproximem bem a estatística λ (já que a busca exaustiva entre todos clusters possíveis é computacionalmente impraticável).

Em dados agregados, métodos de *Scan* Espacial bem difundidos já são amplamente empregados para detecção de clusters em diversos formatos. Quando as informações estão em níveis pontuais, entretanto, tem-se um ganho muitíssimo elevado de informação, e o uso de formatos regulares das janelas para detecção de clusters em muitas situações não é adequado. O método VBScan, além de outros, aparece então como um grande avanço nessa questão.

Outro grande avanço bem recente no estudo dos "excessos de ocorrência" reside na definição de incertezas *individuais*, possibilitando então não só a visualização de limites de incerteza em torno do cluster mais provável, como também a contribuição individual de cada região na anomalia detectada. O método auxilia também na visualização de *clusters* secundários, e em alguns casos pode ser utilizado para visualizar possíveis formatos irregulares, mesmo que se utilize uma janela em formato regular. Tal procedimento estava, até agora, definido no contexto de dados agregados, ou seja, nos casos em que o que se observa são regiões às quais estão associadas contagens de casos e controles. A possibilidade de avaliar essas incertezas quando se observa cada ocorrência individualmente resulta em um ganho de informação muito grande,

e permite uma melhor visualização do processo pontual no espaço observado. Permite, inclusive, que através de um método de detecção mais simples seja verificada a necessidade de um método mais aprimorado.

A estrutura de dados na forma caso-controle, ao contrário do que ocorre quando se tem contagens associadas a regiões delimitadas, não fornece prontamente a matéria-prima para o cálculo das medidas de intensidade: valores observados que possam servir como parâmetros de distribuições de probabilidades para as simulações. Essa situação força a abordagem de *vizinhanças* em torno de cada ponto.

O uso do MST, ferramenta que vem se mostrando bastante útil e versátil no delineamento de *clusters* espaciais em dados pontuais, se mostrou bastante frutífero na obtenção desses parâmetros. No caso de métodos baseados em "janelas" (nesse caso, as janelas circulares) revelou-se que, mesmo uma escolha de aresta que implica em estimativas menos precisas das probabilidades individuais, observa-se níveis de intensidade maiores (incertezas menores) dentro do *cluster* e em regiões secundárias com razão de verossimilhança significativas. Tal característica pode não ser verdadeira, entretanto, se o delineamento de incertezas for realizado utilizando-se outro método de detecção de *cluster* (como o VBScan e outros).

8.1 Trabalhos futuros

É necessário enfatizar que, com exceção da análise inferencial das estimativas de probabilidades individuais, com base no EQM estimado via Monte Carlo, as conclusões apresentadas sobre as medidas de intensidade basearam-se em uma amostra, somente. Seria interessante verificar se tais observações mantém-se ao longo de várias simulações.

Obviamente, o MST não é a única maneira de se estimar probabilidades individuais com base em pontos próximos. Uma abordagem interessante, que seria bastante promissora em trabalhos futuros, é a investigação de *Campos Markovianos* [Krainski *et al.*, 2010], que base-ia na interdependência de pontos em um espaço.

Outros trabalhos adicionais, bastante úteis e promissores, podem ser realizados na definição das medidas de intensidade em *clusters* espaço-temporais, tanto no caso de dados agregados quanto no de dados pontuais. Levando em conta o caráter bastante

visual da definição destas medidas, haveria um desafio grande na definição de uma maneira de representá-las.

Em suma, a investigação de incertezas envolvidas em um procedimento estatístico, qualquer que seja, e a metodologia para estimá-las são procedimentos que devem sempre ser aprimorados e implementados, sempre levando em conta as diversas situações e estruturas de dados que podem existir em uma pesquisa.

Referências Bibliográficas

[Anderson & Titterington, 1996] Anderson, NH, & Titterington, DM. 1996. Some methods for investigating spatial clustering with epidemiological applications. *J.R. Statist. Soc.*, **I**(160), 87–105,.

[Assunção *et al.*, 2006] Assunção, R, Costa, M, Tavares, A, & Ferreira, S. 2006. Fast detection of arbitrary shaped disease clusters. *Statistics in Medicine*, **25**, 723–742.

[Boscoe *et al.*, 2003] Boscoe, FP, McLaughling, C, Schymura, MJ, & Kiello, CL. 2003. Visualization of the spatial scan statistic using nested circles. *Health & Place*, **9**, 273–277.

[Chen *et al.*, 2008] Chen, J, Roth, RE, Natio, AT, Lengerich, EJ, & MacEachren, AM. 2008. Geovisual analytics to enhance spatial scan statistic interpretation: an alnalysis of U.S. cervical cancer mortality. *International Journal of Health Geographics*, **7**(57).

[Cucala *et al.*, 2009] Cucala, L, Demattei, C, Lopes, P, & Ribeiro, A. 2009. Spatial scan statistics for case event data based on connected components. *Biometrics*, 1–17.

[Duczmal *et al.*, 2011] Duczmal, LH, Moreira, GJP, Burgarelli, D, Takahashi, RHC, Magalhaes, FCO, & Bodevan, EC. 2011. Voronoi based prospective space-time scans for point data sets: a dengue fever cluster analys in a southeast Brazilian town. *International Journal of Health Geographics*, 10:29.

[Dwass, 1956] Dwass, M. 1956. Modified randomization tests for nonparametric hypoteses. *Northwestern University and Stanford University*, 181–187.

[IBGE, 2010] IBGE. 2010. *Perfil dos municípios brasileiros 2009*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

[Krainski *et al.*, 2010] Krainski, E, Rodriigues, E, & Assunção, R. 2010. Campos Aleatórios de Markov e Distribuições Especificadas Através das Densidades Condicionais. *XIV SINAPE, Simpósio Nacional de Probabilidade e Estatística. ABE, São Paulo-SP*.

[Kuldorff *et al.*, 2003] Kuldorff, M, Toshiro, T, & Peter, JP. 2003. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, **42**, 665–684.

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

[Kulldorff, 2001] Kulldorff, M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of The Royal Statistical Society Series*, 61–72.

[Kulldorff *et al.*, 2006] Kulldorff, M, Huang, L, Picle, L, & Duczmal, L. 2006. An Elliptic Spatial Scan Statistic. *Statistics in Medicine*, **22**(25), 3929–3943.

[Mount, 2012] Mount, D. 2012. Voronoi Diagrams and Fortune's Algorithm. *CMSC*, **754**(11), 2012.

[Naus, 1965a] Naus, JI. 1965a. Clustering of random points in two dimensions. *Biometrika Trust*, 263–267.

[Naus, 1965b] Naus, JI. 1965b. The distribution of the size of the maximun cluster of points on a line. *Journal of the American Statistical Association*, 532–538.

[Oliveira *et al.*, 2011] Oliveira, LP, Duczmal, LH, Cançado, ALF, & Tavares, R. 2011. Nonparametric intensity bounds for the delineation of spatial clusters. *International Journal of Health Geographics*, **10**(1).

[Openshaw *et al.*, 1988] Openshaw, S, Charlton, M, Craft, AW, & Birch, JM. 1988. Investigation of leukemia clusters by use of geographical analysis machine. *Lancet*, 272–273.

[Orair *et al.*, 2011] Orair, R, Santos, CHM, Silva, WJ, Britto, JM, Rocha, W, Ferreira, AS, & Silva, HL. 2011. Uma metodologia de construção de séries de alta frequência das finanças municipais no Brasil com aplicação para o IPTU e o ISS (2004-2010). *Texto para Discussão 1632. IPEA, Brasília.*

[Rosychuk, 2005] Rosychuk, RJ. 2005. Identifying geographic areas with high disease rates: when do confidence intervals for rates and a disease cluster detection method agree? *International Journal of Health Geographics*, 5:46.

[Rothman, 1987] Rothman, KJ. 1987. Clustering of disease. *Am J Public Health*, 13–15.

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

[Turnbul *et al.*, 1989] Turnbul, BW, Iwano, EJ, Burnett, WS, Howe, HL, & Clark, LC. 1989. Monitoring for clusters of desease: application to leukemia incidence in upstate New York. *School of Operations Research and Industrial Engineering College. College of Engineering. Cornell University. Ithaca, NY.*

[Warner & Aldrich, 1988] Warner, SC, & Aldrich, TE. 1988. The status of cancer cluster investigations undertaken by state health departments. *Am J Public Health*, 306–307.

[Whittemore *et al.*, 1987] Whittemore, A, Friend, BW, & Holly, EA. 1987. A test to detect clusters of disease. *Biometrika*, 631–637.

[Wieland *et al.*, 2007] Wieland, SC, John, SB, Berger, B, & Madi, KD. 2007. Density-Equalizing Euclidean spanning trees for the detection of all disease cluster shapes. *Proceedings of the National Academy of Sciences*, **104**(22), 9404–9409.