Revista da Estatística UFOP, Vol III(2), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS

AJUSTE DA DISTRIBUIÇÃO BETA PARA A MEDIDA DE COMPACIDADE GEOMÉTRICA

Anderson Ribeiro Duarte[1], **Ana Carolina Andrade Gonçalves**[2]

Resumo: *Técnicas de detecção e inferência de conglomerados espaciais têm sido bastante abordadas recentemente. Suas utilizações estão associadas à problemas de saúde pública como em casos de epidemiologia e vigilância sindrômica na propagação de doenças infectocontagiosas. A formulação do problema através de uma abordagem multi-objetivo de otimização é notoriamente eficiente. Um dos objetivos é a Estatística Scan Espacial e o outro, em geral, um objetivo associado com a estrutura topológica ou geográfica do conglomerado a ser detectado no mapa em estudo, como por exemplo, a medida de Compacidade Geométrica. Uma estratégia de otimização bem difundida para essa abordagem é a meta-heurística Algoritmo Genético em conjunto com um teste de hipóteses para a confirmação da existência de clusters no mapa em estudo. Os trabalhos anteriores sempre utilizavam a distribuição empírica e a teoria das funções de aproveitamento para o procedimento inferencial. Este estudo revela que para a abordagem multi-objetivo utilizando como funcional objetivo a função de penalização por Compacidade Geométrica pode-se obter um ajuste eficiente considerando uma distribuição bi-variada (X, Y), sendo $X \sim Gumbel(\mu, \sigma)$ e $Y \sim Beta(\alpha, \beta)$, com X e Y sendo independentes.*

Palavras-chave: *Detecção de Conglomerados, Distribuições ajustadas, Estratégias de Otimização, Estatística Scan, Algoritmo Genético.*

Abstract: *Detection of spatial clusters have been widely addressed. A multi-objective optimization is remarkably efficient with the Scan Statistic and the Geometric Compactness. Previous works has used the empirical distribution and attainment functions. This study shows that it can be obtained an efficient fit considering a bi-variate distribution Gumbel and Beta.*

Keywords: *Cluster Detection, Fitted Distributions, Optimization Strategies, Scan Statistic and Genetic Algorithm.*

1 Introdução

Um volume significativo de trabalhos associados às técnicas para detecção e inferência de conglomerados (*clusters*) espaciais e temporais têm surgido recentemente. Em diversas áreas de estudo podem ser encontrados problemas associados à clusters espaciais. Estudos voltados para a saúde pública (vigilância sindrômica e epidomiologia), criminologia, pesquisa de mercados, entre outros. A metodologia para detecção e inferência de clusters, em sua maior parte, se baseia na Estatística Scan Espacial apresentada em [12, 13].

[1]DEEST - UFOP. e-mail: *duarte.andersonr@gmail.com.br*
[2]PPESTBIO - UFV.

A metodologia proposta em [12, 13] se mostra bastante eficaz, entretanto ela se baseia na solução do problema para clusters que possuam uma forma geométrica mais regular (entende-se aqui por regularidade, o aspecto associado a forma geométrica do candidato a cluster). Para situações com clusters de forma muito irregular a metodologia perde um pouco de sua eficiência. Uma possibilidade de recuperar a eficácia da técnica, seria considerar uma busca exaustiva dentre os subconjuntos de áreas no qual o mapa esta subdividido e utilizar a Estatística Scan Espacial sem as restrições impostas pela metodologia original. Por outro lado, essa possibilidade se contrapõe com um problema de inviabilidade computacional devido ao excessivo número de subconjuntos existentes para uma mapa subdivido em uma quantidade de regiões na ordem das centenas.

Visando contornar esse problema, é necessário obter um método que permita analisar somente os candidatos a clusters (subconjuntos conexos de regiões do mapa) mais promissores e descartar os que não parecem muito adequados. Dado que não analisam todos os candidatos, tais métodos não garantem a obtenção de uma solução ótima, mas um bom método tende a encontrar uma boa solução na maioria das suas execuções. Neste sentido, existem alguns algoritmos que propõem estratégias para a detecção de clusters com formatos irregulares. Uma técnica bastante razoável e já utilizada é a incorporação de alguma função de penalização para o formato geométrico associado ao cluster candidato.

Existem métodos algorítmicos que propõem estratégias para a detecção de clusters com formatos irregulares. Muitos deles são heurísticas, portanto não vasculham todas as possíveis soluções. São analisadas apenas algumas das soluções, que seriam as mais promissoras.

Nesse trabalho é utilizada uma dessas metodologias heurísticas através de um algoritmo genético multi-objetivo implementado especificamente para o problema de detecção de clusters. Trata-se de um método que conduz à maximização de dois objetivos, sendo eles: a Estatística Scan Espacial e uma função de penalização associada a forma do cluster detectado. Não é apresentada uma única solução, mas sim um conjunto de soluções não-dominadas, ou seja, que não são inferiores às outras soluções nos dois objetivos simultaneamente.

Nos procedimentos de detecção de clusters, uma segunda e importante etapa precisa ser executada, o procedimento inferencial. Dado que uma ou mais soluções foram fornecidas pelo algoritmo, não necessariamente se tratam de clusters do ponto de vista estatístico. A avaliação quanto à significância estatística é usualmente realizada paralelamente para todos os clusters do conjunto de soluções não-dominadas usando simulações de Monte Carlo, quebrando o laço de dependência entre elas, e determinando a melhor solução no conjunto de soluções não-dominadas. Utiliza-se para a avaliação da significância estatística a teoria de funções de aproveitamento. A utilização da função de aproveitamento no problema específico de detecção de clusters se encontra bem detalhada em [2]. Entretanto, a acurácia na obtenção do $p-$valor associado a uma solução é dependente de um grande volume de simulações de Monte Carlo.

O principal objetivo deste trabalho reside em obter o ajuste de uma distribuição teórica de probabilidades visando a obtenção do $p-$valor associado a uma solução com um volume bem menor de simulações de Monte Carlo. Para a Estatística Scan Espacial, este resultado já é conhecido, o ajuste da distribuição Gumbel se mostra bastante eficaz como pode ser verificado em [1]. Por outro lado, as medidas de penalização utilizadas para metodologias de detecção de clusters irregulares ainda não possuem uma distribuição teórica bem ajustada. Com este intuito, um estudo extensivo de simulações foi executado para obtenção de uma distribuição ajustada para a medida de penalização denominada Compacidade Geométrica.

2 Material e métodos

2.1 Estatística scan espacial

Na primeira versão para a Estatística Scan, em [14], são apresentados estudos para detecção de clusters em processos pontuais unidimensionais. Procurava-se obter qual a probabilidade de se

escolher k pontos independentes de uma distribuição uniforme (a_1, a_2) e existir um subintervalo de $(a_t, a_{t+p}) \subseteq (a_1, a_2)$ com $a_{t+p} - a_t < a_2 - a_1$ que contenha pelo menos n pontos dentre os k pontos observados. Em [15] é apresentada ainda, uma abordagem bidimensional, mas não análoga àquela para uma dimensão, considerando a existência de um sub-retângulo do quadrado unitário, com lados de tamanho u e v orientados paralelamente aos eixos x e y respectivamente, que contenha pelo menos n dos k pontos.

Boa parte dos métodos atuais para detecção e inferência em clusters espaciais utilizam a Estatística Scan Espacial [12, 13] como estatística de teste. A referida estatística se comporta bem, tanto em problemas para dados pontuais, quanto para problemas em dados agregados por regiões no mapa em estudo. Em ambos, um dos problemas reside no grande volume de soluções candidatas a serem analisadas.

Ao avaliar todos os possíveis subconjuntos de regiões existentes no mapa em estudo, o número de possíveis candidatos cresce exponencialmente. Para um mapa dividido em m regiões, existem $2^m - 1$ possíveis subconjuntos de regiões (alguns conexos, outros não), dentre os conexos, qualquer um desses pode ser a solução mais verossímil no mapa em estudo.

Considere um mapa em estudo dividido em m regiões, com população total P e um total de casos C para algum fenômeno de interesse. Considere ainda o conhecimento do volume populacional e de ocorrências de casos para cada uma das regiões que subdividem o mapa. Qualquer subconjunto conexo de regiões no mapa será denominado como uma zona, ou seja, um candidato a cluster. A estatística de teste busca identificar a zona mais verossímil ao longo do mapa.

Considerando um conjunto composto por todas as zonas que serão avaliadas, ora denominado conjunto Z, busca-se determinar as zonas que podem ser considerados de maior relevância quanto ao valor do logaritmo da função de verossimilhança. Vale ressaltar que as zonas mais verossímeis não são necessariamente clusters. Uma zona será dita um cluster quando o valor do logaritmo da função de verossimilhança for considerado significativo do ponto de vista estatístico. Para esta avaliação, executa-se um teste de hipóteses com a Hipótese nula de não existência cluster no mapa contra uma Hipótese alternativa de existência de pelo menos um cluster no mapa.

A estatística de teste Scan Espacial será então definida como o máximo da razão de verossimilhanças. Sob a validade da Hipótese Nula e assumindo o modelo Poisson para a distribuição dos casos, o número de casos esperados em uma possível zona z é dado por $\mu(z) = C\frac{P(z)}{P}$. Desta forma, temos o risco relativo na zona z dado por $I(z) = \frac{C(z)}{\mu(z)}$. Já o risco relativo fora da zona z é dado por $O(z) = \frac{C - C(z)}{C - \mu(z)}$. Seja L_0 a função de verossimilhança sob a Hipótese Nula e $L(z)$ a função de verossimilhança sob a Hipótese Alternativa. Pode-se mostrar que assumindo o modelo Poisson, o logaritmo da razão de verossimilhanças é dado por:

$$LLR(z) = \begin{cases} C(z)\log\left(I(z)\right) + \left(C - C(z)\right)\log\left(O(z)\right) & \text{se } I(z) > 1 \\ 0 & \text{caso contrário} \end{cases} \tag{1}$$

O logaritmo da razão de verossimilhança é então maximizado no conjunto Z, produzindo então a estatística de teste $\max_{z \in Z} LLR(z)$. O formato de escolha das zonas z pertencentes ao conjunto Z definirá o método em uso. Uma forma bastante usual se baseia em janelas circulares e define o Método Scan Circular [12].

Para cada região define-se um centróide, que é um ponto arbitrário em seu interior. Utilizando a metodologia baseada no teste de verossimilhança proposta em [13], o método Scan Circular busca o cluster mais verossímil dentre todas as zonas circunscritas por círculos de raios variados centrados em cada região do mapa.

Uma janela circular sobre a área em estudo define uma zona formada pelas regiões cujos centróides são interiores à janela. Partindo de janelas centradas em cada um dos centróides e de raios variando entre zero e um valor máximo pré-estabelecido, o conjunto Z será formado por todas as zonas obtidas através das janelas circulares. A busca por soluções eficientes seria feita então dentro do conjunto Z.

Um dos problemas dos metódos circulares para detecção de clusters reside nas situações em que existem clusters com formatos bastante irregulares, bastante comum em situações reais. A incidência de uma doença ao longo de um rio, por exemplo daria um formato mais alongado ao cluster. Neste caso, existe a alternativa de utilizar outros formatos de janelas, por exemplo janelas elipticas, como critério para a definição do conjunto Z, como pode ser visto em [5].

Existem outros critérios para a definição do conjunto Z, como por exemplo busca exaustiva sobre todas as possíveis zonas conexas no mapa em estudo. No caso de considerarmos Z como o conjunto de todas as zonas conexas, o problema se tornaria impraticável para mapas com m da ordem de algumas centenas.

Para concluir o teste de hipóteses, a significância estatística de uma possível solução, obtida através da distribuição dos casos observados, em geral, é verificada através de simulações de Monte Carlo, dado o desconhecimento da distribuição exata da estatística de teste. No procedimento de Monte Carlo, casos simulados (sob a validade da Hipótese Nula) são distribuídos aleatoriamente no mapa em estudo, de forma que cada região recebe, em média, um número de casos proporcional à sua população. A significância estatística, de uma solução obtida é considerada sem pré-especificação do número de regiões e/ou da localização do clusters mais verossímel. O processo inferencial compara a solução mais verossímil obtida dos dados observados com as soluções mais verossímeis obtidas de cada distribuição de casos simulada. Essa comparação é feita através da distribuição empírica para a estatística de teste construída através dos dados da simulação de Monte Carlo.

2.2 Detecção de clusters irregulares

Existe um problema que ocorre em diversos cenários e que os métodos citados anteriormente não se preocupam em abordá-lo, que é a ocorrência de possíveis clusters irregulares. Uma vasta revisão sobre os diversos métodos que contemplam clusters irregulares pode ser obtida em [8].

É muito frequente a existência de clusters com formatos bastante irregulares na maioria dos estudos. Os clusters não regulares podem ser observados em problemas de tráfego, poluição, vigilância sindrômica, entre outros. Em muitos destes casos, formatos não regulares se devem às características geográficas do mapa em estudo, tais como rios, regiões litorâneas, regiões montanhosas, entre outras.

Métodos foram desenvolvidos recentemente para detectar clusters de formato irregular, mesmo assim apresentam alguns problemas. Um primeiro problema seria a avaliação de todos os possíveis candidatos (subconjuntos de regiões do mapa), visto que o número destes candidatos cresce exponencialmente a medida que o número de regiões no mapa em estudo aumenta. Um segundo problema é que na possibilidade de avaliarmos todos os candidatos, se avaliando através da razão de verossimilhanças, decorrente da proposta da estatística Espacial Scan Circular, a solução obtida nem sempre seria uma solução viável.

Ao procurar por clusters com liberdade ilimitada de forma geométrica, o poder de detecção é diminuído. Isso acontece porque a coleção de todas as soluções candidatas conexas, independentemente de forma, é muito grande, as soluções mais verossímeis podem se apresentar "em forma de árvore". Nessa forma são candidatos que não podem contribuir para a descoberta de soluções geográficas significativas, que delineiam corretamente o cluster verdadeiro. Em outras palavras, há muito "ruído", contra o qual as soluções legítimas não podem ser distinguidas. Esse problema ocorre para todas as metodologias de detecção de clusters irregulares conhecidas. Este formato de solução tende a não ser muito informativo e em geral não é uma solução de interesse para o problema na prática, a figura 1 ilustra uma situação deste tipo.

Dada a possibilidade de existência de tais soluções, o poder de detecção destes métodos seria reduzido. Neste sentido, existem algoritmos que propõem estratégias para a detecção de clusters com formatos irregulares. Entretanto tais métodos não vasculham todas as possíveis soluções, ou seja, são métodos heurísticos. São analisadas apenas algumas das soluções, que seriam as mais promissoras. Ainda assim, persistiria o problema de soluções não factíveis.

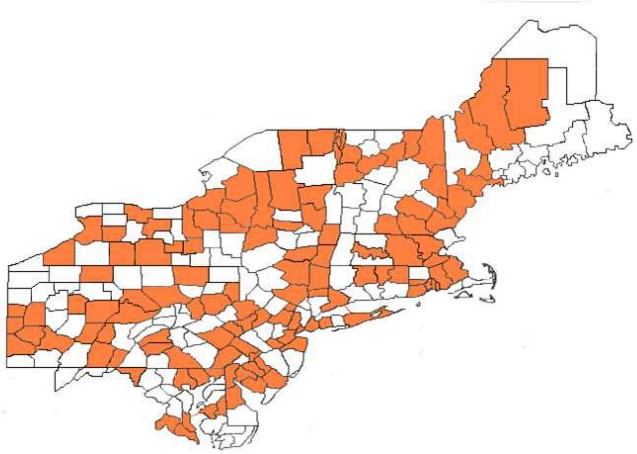

Figura 1: Um cluster conexo com um grande volume de regiões (com 122 regiões, ou 50% da área em estudo) encontrado através de uma busca sem limitação para o tamanho da solução em sem utilização de funções de penalização.

Em [16, 3] são apresentadas propostas para detecção de clusters através da utilização de algoritmos genéticos. Posteriormente em [6] é descrito um algoritmo genético para a detecção e inferência de clusters com forma irregular. Em [6] considera-se um mapa dividido em regiões com populações e números de casos conhecidos considerando como função fitness a Estatística Scan Espacial. Uma função de penalidade para a forma é apresentada em [5], com base no conceito da Compacidade Geométrica, sendo empregada para evitar a irregularidade excessiva das soluções candidatas.

Esta função de penalização tem o objetivo de penalizar as zonas do mapa que possuem formato muito irregular. A Compacidade Geométrica $CG(z)$ de uma zona z é dada pela área da zona z, definida por $A(z)$, dividida pela área do círculo com o mesmo perímetro que o fecho convexo da zona z. O fecho convexo será aqui definido por $H(z)$.

A expressão descrita acima para $CG(z)$ é dada por:

$$CG(z) = \frac{A(z)}{\pi \left(\frac{H(z)}{2\pi} \right)^2} \tag{2}$$

A Compacidade Geométrica é dependente da forma do objeto, mas não do seu tamanho. A Compacidade penaliza a forma que tem área pequena em relação a área da circunferência com perímetro igual ao fecho convexo. O círculo é a forma de maior compacidade ($CG(z) = 1$). Já o quadrado, por exemplo, tem compacidade $CG(z) = 0{,}785$.

Em [7] é apresentada uma estratégia que busca maximizar dois objetivos, sendo eles: a Estatística Scan Espacial e a função associada a forma, denominada Compacidade Geométrica, que avalia a regularidade do formato geométrico do possível cluster. Não é apresentada uma única solução, mas sim um conjunto de soluções não-dominadas, ou seja, que não são inferiores às outras soluções nos dois objetivos simultaneamente. O algoritmo multi-objetivo apresenta uma importante vantagem: todos os clusters potenciais são considerados sem uma classificação de acordo com os valores da penalização. Assim a classificação quanto à qualidade das possíveis soluções é executada somente depois que todos os candidatos são avaliados.

Para as situações de abordagem multi-objetivo utilizando como funções objetivo a Estatística Scan Espacial e a função de Compacidade Geométrica, deve-se observar que a execução do algoritmo não fornece uma única solução, mas sim um conjunto de soluções não-dominadas, ou seja, uma aproximação de um conjunto de Pareto. Busca-se então uma estratégia para verificar para cada solução deste conjunto de soluções não-dominadas sua significância estatística.

De forma similar ao procedimento em [9], através de simulações de Monte Carlo, pode-se executar o algoritmo para diversas distribuições de casos sob a hipótese de não existência de clusters no mapa em estudo. Cada uma destas execuções fornece um conjunto de soluções não-dominadas. O conjunto destas diversas execuções pode ser utilizado para mensurar a significância estatística de uma solução pertencente ao conjunto de soluções não-dominadas. Para tal tarefa é importante definir as funções de aproveitamento que se encontram bem detalhadas em [11, 10, 2].

Para cada uma execução do algoritmo, obtêm-se um conjunto de soluções eficientes. Este conjunto particiona o espaço de objetivos em duas regiões R_1 e R_0: R_1 é a região dos pontos dominados pelo conjunto de soluções eficientes, ou seja, qualquer ponto de R_1 nunca é superior a qualquer dos pontos do conjunto de soluções eficientes se considerando os dois objetivos simultâneamente; já algum ponto que se situasse na região R_0 este seria um ponto não dominado pelos pontos do conjunto de soluções eficientes, ou seja, pontos sempre superiores aos pontos do conjunto de soluções eficientes em pelo menos um dos objetivos (veja Figura 2).

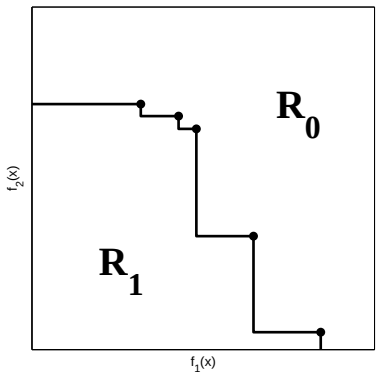

Figura 2: Superfície de aproveitamento dividindo o espaço de objetivos

Para alguma solução x dominada por algum ponto do conjunto de soluções eficientes, ou seja, pertencente a R_1, dizemos que x foi superada por nosso conjunto de soluções eficientes, construindo então um limite para avaliar a significância estatística da solução x.

Pode-se repetir a execução do algoritmo para n alocações distintas de casos no mapa, obtidas de cada uma réplica de Monte Carlo, sob a hipótese de não existência de cluster, obtendo então n conjuntos de soluções eficientes, produzindo n limites distintos (veja Figura 3, lado esquerdo). O conjunto dos n limites pode ser utilizado para dividir o espaço de objetivos em $n + 1$ regiões (veja Figura 3, lado direito).

Uma solução que apresenta um ponto no espaço de objetivos à direita de todas as superfícies de aproveitamento, não foi superada em nenhuma das execuções. Ao passo que uma solução que apresente um ponto à esquerda de alguma das superfícies de aproveitamento, foi superada em algumas das execuções. Um ponto à esquerda de todas as superfícies de aproveitamento foi superado em todas as execuções.

Então o espaço de objetivos está sendo dividido em $n + 1$ regiões. Pode-se com um grande número de execuções sob a hipótese de não existência de cluster no mapa, mensurar a significância estatística de uma solução obtida através dos casos originais distribuídos no mapa, através da proporção de regiões não alcançadas no espaço de objetivos.

O algoritmo genético que é reconhecidamente um estratégia heurística de otimização bastante eficiente. É utilizado o princípio da evolução biológica para procurar as melhores soluções de um problema de otimização, são simulados os mecanismos de variação aleatória e de seleção

adaptativa da evolução natural. O algoritmo genético é constituído por quatro etapas: Geração de População Inicial, Cruzamento, Mutação e Seleção. Através dos operadores de cruzamento, mutação e seleção, é possível melhorar os resultados entre uma geração e outra. São eles que atribuem ao algoritmo a capacidade de evoluir no procedimento de busca por soluções ótimas.

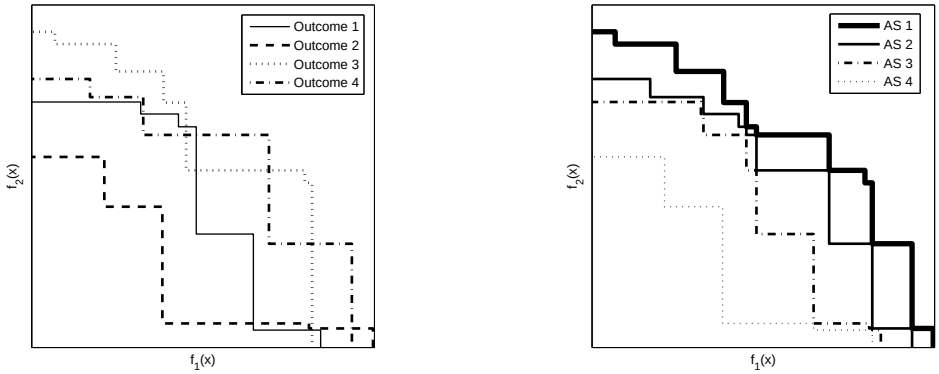

Figura 3: Limites obtidos para diferentes execuções do algoritmo à esquerda e superfícies de aproveitamento para n execuções do algoritmo à direita

Para o problema específico de estudo, a população inicial deve ser capaz de captar as informações do mapa como um todo. Não há razão para iniciar o algoritmo com os indivíduos concentrados em apenas uma parte do mapa, mesmo porque um cluster somente pode ser identificado se possuir valor de LLR discrepante das demais zonas, o que obriga a ter um mínimo de conhecimento sobre zonas espalhadas pelo mapa. Para tanto se utiliza uma estratégia gulosa (*algoritmo guloso*) visando obter zonas com alta LLR, construindo as zonas para a população partindo de cada uma das regiões do mapa em estudo, através da estratégia gulosa.

Lembrando que o método em questão é estocástico, nem todas as possíveis soluções estão sendo avaliadas, portanto não existe garantia que será encontrada a solução ótima. Portanto pode-se ter uma avaliação que subestimasse os p-valores. De fato os p-valores são um pouco menores que os p-valores teóricos.

2.3 Distribuições ajustadas para as funções objetivo

O principal foco desse trabalho reside no ajuste de distribuições teóricas para as funções objetivos do problema de detecção de clusters espaciais através da abordagem multi-objetivo. No caso em estudo as funções objetivo são a Estatística Espacial Scan e a Compacidade Geométrica. É fácil ver que a Estatística Espacial Scan depende exclusivamente da população e do número de casos distribuídos ao longo do mapa em estudo, enquanto a Compacidade Geométrica depende somente da forma das regiões ao longo do mapa. Isto caracteriza uma clara **independência** entre estas duas grandezas.

Considere por X a variável aleatória que modela a função objetivo associada à Estatistica Scan Espacial e por Y a variável aleatória que modela a função objetivo Compacidade Geométrica. Considere ainda, uma possível solução A a ser avaliada. Usando o conceito de dominância já descrito anteriormente, mas agora sem utilizar as funções de aproveitamento, mas sim o conhecimento das distribuições para as variáveis aleatórias X e Y, o p-valor associado a solução A poderia então ser descrito por:

$$P\left\{[X > LLR(A) \cap Y \geq CG(A)] \cup [X \geq LLR(A) \cap Y > CG(A)]\right\} =$$

$$= P[X > LLR(A) \cap Y \geq CG(A)] + P[X \geq LLR(A) \cap Y > CG(A)] - P[X > LLR(A) \cap Y > CG(A)]$$

$$= P[X > LLR(A)]P[Y \geq CG(A)] + P[X \geq LLR(A)]P[Y > CG(A)] - P[X > LLR(A)]P[Y > CG(A)].$$

O resultado anterior foi obtido considerando a independência entre X e Y, agora considerando o fato de se tratarem de variáveis aleatórias contínuas temos:

$$P\left\{[X > LLR(A) \cap Y \geq CG(A)] \cup [X \geq LLR(A) \cap Y > CG(A)]\right\} = P[X > LLR(A)]P[Y > CG(A)] \tag{3}$$

Resta então obter as distribuições adequadas para X e Y.

Quanto a distribuição ajustada para a Estatística Espacial Scan, em [1] é apresentado um vasto estudo que garante alta qualidade de ajuste da Estatística Scan Espacial através da distribui-ção Gumbel, isto se deve principalmente ao bom comportamento da referida distribuição para modelar estatísticas de máximo. Uma variável aleatória X segue distribuição Gumbel com parâmetros μ e σ se possui a seguinte função densidade de probabilidade:

$$f(x|\mu, \sigma) = \frac{1}{\sigma} exp\left\{-\left(\frac{x-\mu}{\sigma}\right) - exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \tag{4}$$

com $\mathbb{E}(X) = \mu - 0.5772\sigma$ e $\text{VAR}(X) = \dfrac{\pi^2\sigma^2}{6}$

Desta forma pode-se obter os estimadores de momentos para os parâmetros por:

$$\hat{\mu} = \bar{X} + 0.5772\hat{\sigma} \tag{5}$$

$$\hat{\sigma} = \frac{1}{\pi}\sqrt{\frac{6\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}} \tag{6}$$

para n suficientemente grande, o estimador em 6 pode ser aproximado por $\hat{\sigma} = \dfrac{\sqrt{6S^2}}{\pi}$.

Já considerando a distribuição ajustada para a Compacidade Geométrica, deve ser considerado que a Compacidade Geométrica retorna um valor no intervalo real $(0, 1)$, em que quanto mais semelhante ao círculo é a figura geométrica, mais próximo de 1 se encontra seu valor. Dentre as variáveis aleatórias contínuas com espaço amostral no intervalo $(0, 1)$, a distribuição Beta de parâmetros α e β se comporta bem, podendo sofrer suficientes alterações no formato da curva de sua função densidade de acordo com os parâmetros α e β. Uma variável aleatória Y segue distribuição Beta com parâmetros α e β se possui a seguinte função densidade de probabilidade:

$$f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} I_{(0,1)}(y) \tag{7}$$

com $\mathbb{E}(X) = \dfrac{\alpha}{\alpha+\beta}$ e $\text{VAR}(X) = \dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Desta forma pode-se obter os estimadores de momentos para os parâmetros por:

$$\hat{\alpha} = \frac{\bar{X}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2\right)}{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2} \tag{8}$$

$$\hat{\beta} = \frac{\left(1 - \bar{X}\right)\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2\right)}{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2} \tag{9}$$

3 Resultados e discussões

Visando atingir o principal alvo do trabalho, foi utilizado um benchmark de dados reais em um mapa composto por 245 condados em 10 estados e no Distrito de Columbia, no Nordeste dos EUA, com 58.943 casos de câncer de mama no período de 1988 a 1992, para uma população de risco de 29.535.210 mulheres em 1990 (em [4] esse benchmark de dados é apresentado com mais detalhes).

Inicialmente, os casos são distribuídos aleatoriamente ao longo do mapa em estudo, com a probabilidade de cada caso ser alocado em cada uma das regiões do mapa sendo a população relativa da respectiva região. Portanto, no cenário de validade da hipótese nula de não existência de cluster no mapa, o algoritmo é executado para 100.000 alocações de casos distintas na validade da hipótese nula através de simulações de Monte Carlo.

Para cada uma alocação de casos, o algoritmo retorna um conjunto Pareto. A primeira análise considera os 100.000 conjuntos Pareto para estimar os parâmetros para as distribuições Gumbel nos valores para a Estatística Scan Espacial e Beta nos valores para a Compacidade Geométrica. Além disso, posteriormente realiza-se uma verificação, através do teste de aderência de Kolmogorov-Smirnov, visando admitir com base estatística a aceitação do ajuste para as distribuições Gumbel e Beta.

Em uma outra análise, os dados são divididos em 100 sub-amostras, cada uma delas com 1.000 conjuntos Pareto. O procedimento de ajuste das distribuições Gumbel e Beta é executado para cada uma das sub-amostras. O intuito deste procedimento é garantir que mesmo com um volume bem menor de simulações de Monte Carlo, as distribuições ajustadas podem ser obtidas com efetiva qualidade nas estimativas. No procedimento usual, através da distribuição empírica, recomenda-se em torno de 10.000 execuções, ou seja, 10.000 conjuntos Pareto para a produção de uma distribuição empírica satisfatória.

3.1 Ajuste para a amostra completa

Utilizando os estimadores das equações 5, 6, 8, 9 e considerando os 100.000 conjuntos Pareto, os valores estimados para os parâmetros foram $\hat{\mu} = 7{,}412916535$ e $\hat{\sigma} = 2{,}742422047$ considerando a distribuição Gumbel. Já considerando a distribuição Beta, os valores estimados para os parâmetros foram $\hat{\alpha} = 2{,}361611983$ e $\hat{\beta} = 3{,}081515271$. A Figura 4 apresenta a distribuição empírica acumulada através de seu histograma e também a curva para a função de distribuição acumulada considerando os parâmetros estimados.

Em uma análise visual da Figura 4, os indícios iniciais levam a acreditar que realmente o ajuste através destas distribuições é adequado. Obviamente este resultado precisa ser corroborado por um teste estatístico para a qualidade do ajuste. Para tanto, foi utilizado o teste de Kolmogorov-Smirnov que confirmou a boa qualidade do ajuste para um nível de significância de 5% ($p-$valores: 0,239073 e 0,409934 respectivamente para as Distribuições Gumbel e Beta).

3.2 Ajuste para as sub-amostras

Considerando as 100 sub-amostras de 1.000 conjuntos Pareto e as equações 5, 6, 8, 9, foram obtidas 100 estimativas para cada um dos parâmetro. A Figura 5 apresenta a distribuição empírica acumulada através de seu histograma para a amostra completa e também as 100 curvas ajustadas para a função de distribuição acumulada considerando os parâmetros estimados. A

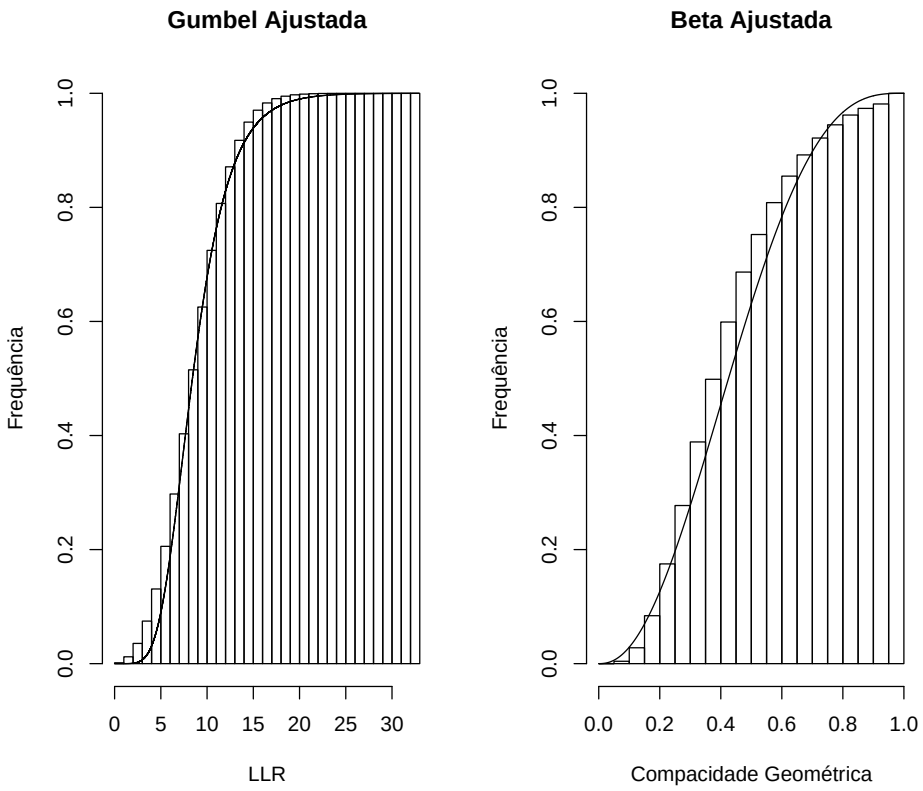

Figura 4: Distribuições Gumbel e Beta ajustadas para Estatística Espacial Scan e para a Compacidade Geométrica respectivamente através da amostra completa.

sobreposição das curvas deixa a impressão de uma curva única representada em traço mais grosso em decorrência da similaridade entre as curvas.

As curvas ajustadas para as 100 sub-amostras revelam em uma análise visual boa qualidade de ajuste. Este resultado também precisa ser confirmado por um teste para a qualidade do ajuste. Na maioria dos 100 casos, o teste de Kolmogorov-Smirnov confirmou a validade do ajuste para um nível de significância de 5%. Em apenas 4 casos para a distribuição Gumbel não ocorreu a aceitação da hipótese nula de que a amostra realmente era proveniente de uma Gumbel. Esse resultado já era esperado em virtude dos resultados amplamente conhecidos sobre a qualidade desse ajuste. O resultado inovador ocorre quando se verifica que em apenas 2 casos não ocorreu a aceitação da hipótese nula de que a amostra realmente era proveniente de uma Beta.

A tabela 1 apresenta os mímimos e os máximos dentre as estimativas obtidas. Obviamente não pode ser encarada como de duas curvas ajustadas, dado que as estimativas mínimas e máximas não são da mesma amostra.

Tabela 1: Mímimo e máximo dentre as estimativas dos parâmetros através das sub-amostras.

	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\alpha}$	$\hat{\beta}$
Mínimo	2,665305	7,308207	2,270916	2,958077
Máximo	2,851280	7,537135	2,437564	3,193104

Ainda sim, se considerarmos as curvas ajustadas com as estimativas mínimas e máximas e compará-las com a curva ajustada através das estimativas da amostra completa (veja Figura 6) nota-se uma grande similaridade. Esse é também um fato que se constata na tentativa

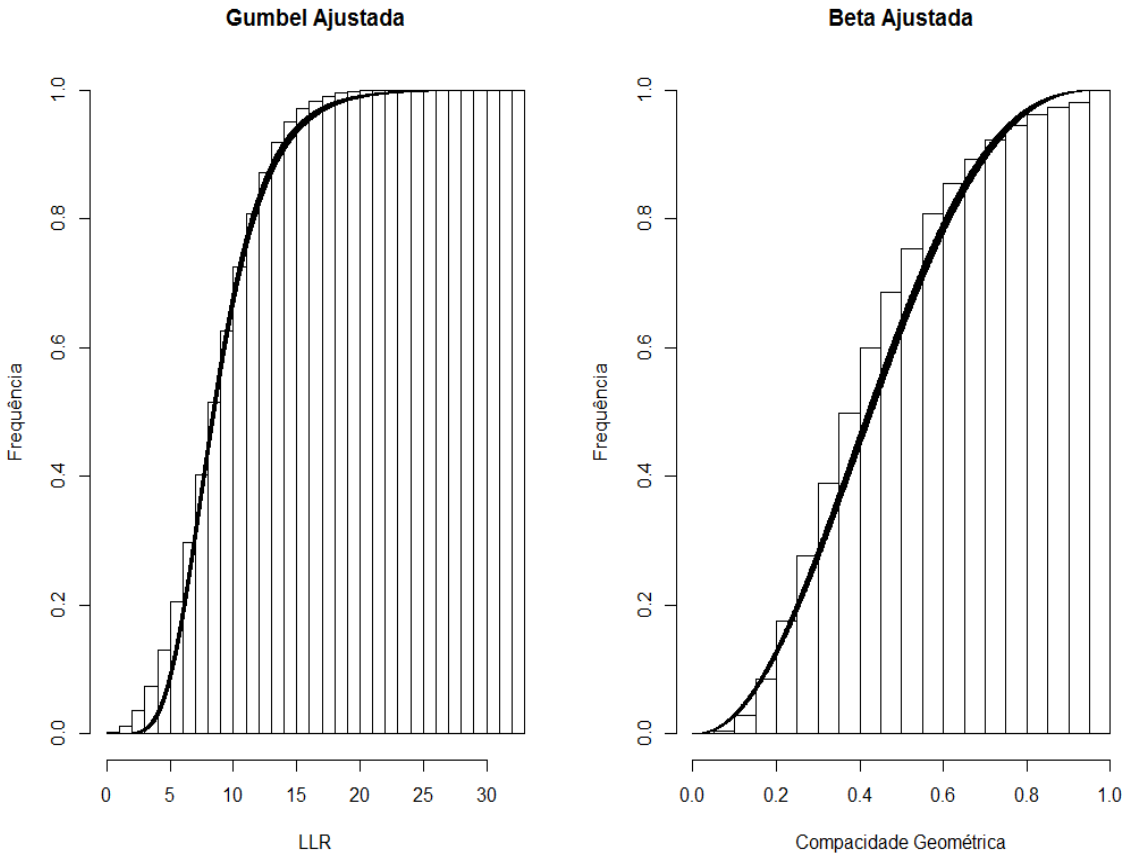

Figura 5: Distribuições Gumbel e Beta ajustadas para Estatística Espacial Scan e para a Compacidade Geométrica respectivamente através da amostra completa.

de assegurar que a distribuição pode ser ajustada através de um volume menor de dados de simulação.

Considerando a Distribuição Gumbel, é possível verificar que a curva ajustada com as estimativas máximas coincide (visualmente) com a curva ajustada com as estimativas da amostra completa, ao passo que a curva ajustada com as estimativas mínimas não coincide, mas apresenta bastante similaridade. Já considerando o caso da Distribuição Beta, a verificação visual aponta para uma coincidência, quase exata, entre a distribuição ajustada com a amostra completa e as distribuições usando as estimativas mínimas e máximas.

Ainda considerando os casos com as estimativas máximas e mínimas, o teste de Kolmogorov-Smirnov confirmou a validade do ajuste para um nível de significância de 5% ($p-$valores: 0,2350035 e 0,408779 respectivamente para as Distribuições Gumbel e Beta considerando as estimativas máximas; $p-$valores: 0,7578463 e 0,411347 respectivamente para as Distribuições Gumbel e Beta considerando as estimativas mínimas).

3.3 Avaliações Numéricas

Considerando a possibilidade de obtenção das distribuições ajustadas para a Estatística Scan Espacial e também para a Compacidade Geométrica se torna possível a obtenção de $p-$valores para soluções obtidas através dos dados observados com maior precisão. Para tanto, os dados reais do benchmark de casos de câncer de mama no Nordeste dos EUA são utilizados.

Será considerado o conjunto Pareto obtido dos dados observados para comparação entre o

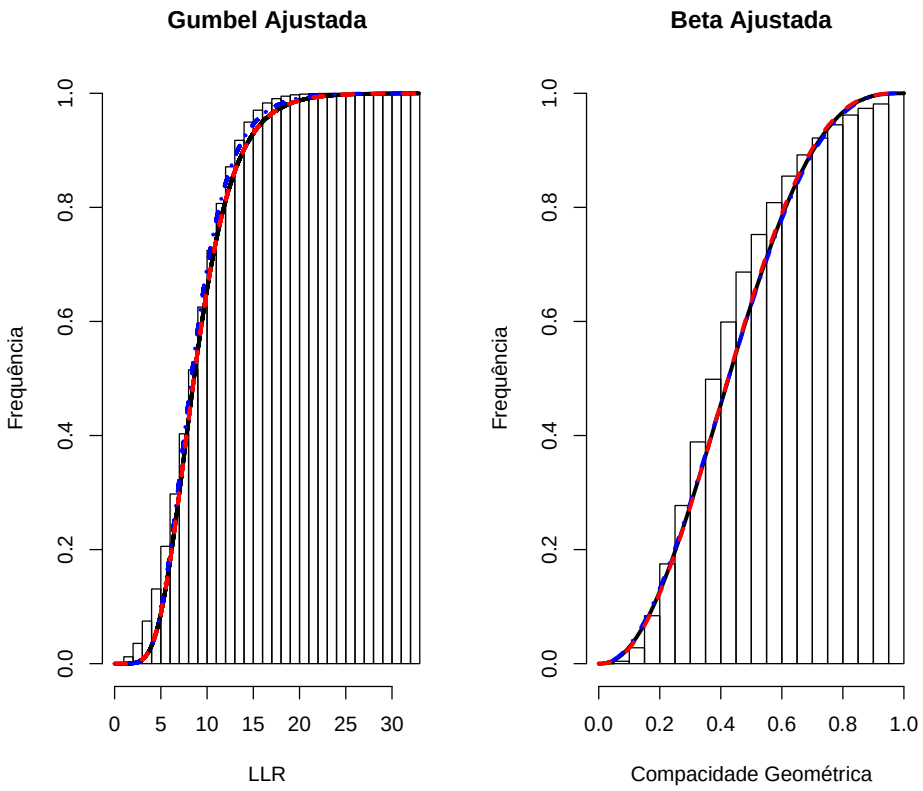

Figura 6: Curvas ajustadas, em azul com as estimativas mínimas, em vermelho com as estimativas máximas e em preto com a amostra completa.

$p-$valor obtido através dos resultados de Simulação com o auxílio das funções de aproveitamento e o $p-$valor obtido através das distribuições ajustadas Gumbel e Beta.

Considerando os conjuntos Pareto obtidos através das simulações de Monte Carlo, o maior valor obtido para a Estatística Scan Espacial sob H_0 foi 32,58047. Já considerando o conjunto Pareto obtido dos dados observados que apresentou 99 soluções, as 98 primeiras possuem Estatística Scan Espacial superior à 32,58047. Dessa forma essas 98 soluções são não-dominadas. A 99^a solução, que foi a única com Estatística Scan Espacial inferior à 32,58047, apresentou Compacidade Geométrica igual a 1. Como todas as soluções nas simulações que possuem Compacidade Geométrica igual a 1, possuem Estatística Scan Espacial inferior ao valor da estatística de teste da 99^a solução (29,978279), pode-se afirmar que também a 99^a solução é não-dominada. Portanto, considerando o procedimento inferencial através das funções de aproveitamento, para todos as soluções pode-se afirmar que o $p-$valor é menor que 1×10^{-5}.

Considerando então a estratégia usual, é possível detectar 99 soluções que são significativas do ponto de vista estatístico. Por outro lado, não é possível estabelecer uma classificação, em ordem de significância estatística, entre elas. Considerando as distribuições ajustadas Gumbel e Beta e a equação 3 para o calculo do $p-$valor, é possível obter um valor específico, associado a cada uma das 99 soluções do conjunto Pareto obtido dos dados observados. A tabela 2 apresenta para as 99 soluções obtidas para os dados observados, os valores para a Estatística Scan Espacial e para a Compacidade Geométrica e também o seu respectivo grau de significância estatística obtido através das distribuições ajustadas Gumbel e Beta. É possível confirmar que em todos os casos realmente o $p-$valor era menor que 1×10^{-5}, mas agora com uma precisão maior e também com a possibilidade de um ranqueamento entre as soluções através dos seus $p-$valores.

Tabela 2: $P-$valores obtidos através das distribuições ajustadas.

LLR	CG	$p-$valor	LLR	CG	$p-$valor
154.899429	0.077502	3.241435×10^{-7}	105.102524	0.300537	2.378840×10^{-7}
154.808563	0.080975	3.235976×10^{-7}	105.102524	0.300538	2.378840×10^{-7}
154.763397	0.092217	3.215843×10^{-7}	104.873077	0.315009	2.297299×10^{-7}
151.052124	0.109380	3.179586×10^{-7}	104.873077	0.315009	2.297299×10^{-7}
151.051758	0.109380	3.179586×10^{-7}	104.872986	0.315009	2.297299×10^{-7}
151.051575	0.109380	3.179586×10^{-7}	104.872810	0.315009	2.297299×10^{-7}
149.988037	0.116757	3.161758×10^{-7}	103.472282	0.325185	2.239580×10^{-7}
147.854294	0.133454	3.116879×10^{-7}	103.231766	0.343616	2.132247×10^{-7}
147.853729	0.133454	3.116879×10^{-7}	103.231590	0.343616	2.132247×10^{-7}
146.793335	0.137204	3.105785×10^{-7}	103.231415	0.343616	2.132247×10^{-7}
144.030548	0.137389	3.105489×10^{-7}	102.190491	0.361771	2.025697×10^{-7}
141.717728	0.138777	3.101317×10^{-7}	102.190407	0.361771	2.025697×10^{-7}
132.520416	0.139623	3.098612×10^{-7}	101.437378	0.362658	2.020370×10^{-7}
132.442139	0.186267	2.935176×10^{-7}	101.437378	0.362658	2.020370×10^{-7}
132.441559	0.186267	2.935176×10^{-7}	101.437210	0.362658	2.020370×10^{-7}
132.211609	0.190924	2.916291×10^{-7}	97.833199	0.375180	1.946157×10^{-7}
132.211609	0.190924	2.916291×10^{-7}	97.098534	0.377016	1.934849×10^{-7}
132.211609	0.190924	2.916291×10^{-7}	97.098373	0.377016	1.934849×10^{-7}
132.211029	0.190924	2.916291×10^{-7}	93.343864	0.379351	1.921151×10^{-7}
131.382446	0.197604	2.888663×10^{-7}	93.343216	0.379351	1.921151×10^{-7}
131.382446	0.197604	2.888663×10^{-7}	92.663437	0.381976	1.905658×10^{-7}
131.012238	0.202758	2.867091×10^{-7}	92.663437	0.381976	1.905658×10^{-7}
131.012238	0.202758	2.867091×10^{-7}	91.948097	0.384730	1.888965×10^{-7}
130.188538	0.213077	2.822144×10^{-7}	91.947609	0.384730	1.888965×10^{-7}
130.188538	0.213077	2.822144×10^{-7}	91.810402	0.385235	1.885983×10^{-7}
128.371628	0.215841	2.809616×10^{-7}	90.797440	0.405865	1.763067×10^{-7}
128.371628	0.215841	2.809616×10^{-7}	90.797188	0.405865	1.763067×10^{-7}
128.370880	0.215841	2.809616×10^{-7}	90.063652	0.408101	1.749353×10^{-7}
128.048096	0.218455	2.797867×10^{-7}	86.564049	0.417457	1.693970×10^{-7}
127.790634	0.223305	2.775429×10^{-7}	82.222076	0.424405	1.652385×10^{-7}
127.790451	0.223305	2.775429×10^{-7}	81.533043	0.428513	1.628085×10^{-7}
127.243324	0.230699	2.741300×10^{-7}	77.688995	0.439825	1.561386×10^{-7}
127.243141	0.230699	2.741300×10^{-7}	73.284073	0.502695	1.202140×10^{-7}
124.821045	0.245230	2.670660×10^{-7}	67.292702	0.611409	6.634051×10^{-8}
124.820892	0.245230	2.670660×10^{-7}	63.482403	0.642655	5.372935×10^{-8}
120.128006	0.249254	2.650781×10^{-7}	62.697060	0.656076	4.871743×10^{-8}
120.127823	0.249254	2.650781×10^{-7}	62.697060	0.656076	4.871743×10^{-8}
119.105347	0.256130	2.615987×10^{-7}	59.049179	0.656799	4.917453×10^{-8}
119.104637	0.256130	2.615987×10^{-7}	58.358135	0.659022	4.859224×10^{-8}
114.565659	0.259439	2.599129×10^{-7}	55.389317	0.700707	3.607768×10^{-8}
114.565659	0.259439	2.599129×10^{-7}	55.388939	0.700707	3.607806×10^{-8}
113.051666	0.263016	2.580584×10^{-7}	54.455631	0.701023	3.701938×10^{-8}
113.051323	0.263016	2.580584×10^{-7}	53.143345	0.828921	3.831640×10^{-8}
108.726585	0.263178	2.580067×10^{-7}	48.082104	0.844206	1.123528×10^{-8}
108.726044	0.263178	2.580067×10^{-7}	44.309002	0.884318	1.198477×10^{-8}
106.207069	0.285048	2.463985×10^{-7}	40.591850	0.893295	3.159937×10^{-8}
106.206985	0.285048	2.463985×10^{-7}	40.591812	0.893295	3.159937×10^{-8}
106.206711	0.285048	2.463985×10^{-7}	36.444969	0.914712	7.020840×10^{-8}
105.980804	0.296693	2.400473×10^{-7}	29.978279	1.000000	6.513719×10^{-14}
105.980629	0.296693	2.400473×10^{-7}			

4 Conclusões

Se encontram presentes na literatura diversos trabalhos propondo abordagens para a detecção e inferência para clusters irregulares. Dentre essas abordagens, as aplicações usando o algoritmo genético multi-objetivo se mostram bem sucedidas na maioria dos casos. Por outro lado, o grande volume de execuções de simulação de Monte Carlo nos trabalhos precursores pode ser visto com um tipo de impeditivo para sua utilização. Considerando a proposição mono-objetivo, o volume de execuções pode ser reduzido de maneira significativa através do ajuste de uma distribuição Gumbel para os valores da estatística de teste para dados simulados sob a validade da hipótese nula de não existência de clusters no mapa em estudo.

Os trabalhos anteriores utilizavam a distribuição empírica e a teoria das funções de apro-

veitamento para o procedimento inferencial. Entretanto, em muitas situações a estimativa de $p-$valor obtida não era suficientemente precisa. Este estudo revelou que para a abordagem multi-objetivo utilizando como funcional objetivo a função de penalização por Compacidade Geométrica pode-se obter um ajuste eficiente considerando uma distribuição bi-variada (X, Y), sendo $X \sim Gumbel(\mu, \sigma)$ e $Y \sim Beta(\alpha, \beta)$, com X e Y sendo independentes. Além da proposição das distibuições ajustadas Gumbel e Beta, o conceito de $p-$valor para distribuições bivariadas é abordado considerando os conceitos de dominância que são usuais para classificação de soluções eficientes nas pesquisas da área de otimização.

Na apresentação original da proposta multi-objetivo com as funções objetivo sendo a Estatística Scan Espacial e a Compacidade Geométrica através do algoritmo genético para estes problemas, a quantidade de execuções das simulações de Monte Carlo sob a hipótese nula para a obtenção de uma distribuição empírica confiável através da função de aproveitamento era calibrada em aproximadamente 10.000 execuções. Foi verificado aqui que através de um volume bem inferior de execuções sob a hipótese nula (1.000 execuções) é possível a obtenção de um ajuste bastante adequado para as distribuições Gumbel para a Estatística Scan Espacial e Beta para a Compacidade Geométrica.

A eficiência no ajuste das distribuições é confirmado através de visualizações gráficas e corroborado através de testes de aderência de Kolmogorov-Smirnov. Em particular, os piores resultados dentre as estimativas obtidas para os parâmetros das distribuições Gumbel e Beta ainda se mostram suficientemente eficientes no momento do ajuste. Esse fato gera uma segurança para o usuário que optar pelo procedimento inferencial através das distribuições ajustadas e não através da técnica de funções de aproveitamento. Os teste com as sub-amostras de tamanho menor confirmam a possibilidade de execução de procedimento inferencial com um volume bem menor de simulações de Monte Carlo.

A utilização do benchmark de casos de câncer de mama no Nordeste dos EUA é adequado para o intuito dos testes, isto se deve ao fato de tal mapa apresentar grande heterogeneidade tanto em sua distribuição populacional ao longo das regiões quanto para a forma geométrica de sua regiões. Considerando esses formatos geométricos e as diferenças populacionais, a possibilidade de surgimento de possíveis clusters de forma irregular se torna mais evidente.

Os resultados atestam que assumindo o ajuste das distribuições Gumbel e Beta permitem a obtenção de um $p-$valor mais preciso para cada uma das soluções obtidas. Isso é o suficiente para um ranqueamento entre as soluções criando a possibilidade de estabelecer um padrão para a alocação de recursos com o intuito de conter os efeitos gerados pela disseminação do fenômeno em estudo (câncer de mama para o benchmark utilizado).

Esse trabalho deixa aberto um leque de opções de estudos futuros, sejam eles no intuito de obter estimadores mais eficientes para os parâmetros das distribuições Gumbel e Beta, ou também buscando o ajuste de distribuições teóricas para outros funcionais objetivos que se encontram em uso para o problema de detecção e inferência em clusters espaciais de forma irregular.

Referências

[1] ABRAMS, A.M.; KLEINMAN, K.; KULLDORFF, M. Gumbel based p-value approximations for spatial scan statistics. **International Journal of Health Geographics**. BioMed Central. v. 9, p. 6 (online version), 2010.

[2] CANÇADO, A.L.F.; DUARTE, A.R.; DUCZMAL, L.; FERREIRA, S.J.; FONSECA, C.M.; GONTIJO, E.C.D.M. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. **International Journal of Health Geographics**. BioMed Central. v. 9, p. 55 (online version), 2010.

[3] CONLEY, J.; GAHEGAN, M.; MACGILL, J. A Genetic Approach to Detecting Clusters in Point Data Sets. **Geographical Analysis**. Wiley. v. 37, p. 286-314, 2005.

[4] DUARTE, A.R.; DUCZMAL, L.; FERREIRA, S.J.; CANÇADO, A.L.F. Internal cohesion and geometric shape of spatial clusters. **Environmental and Ecological Statistics**. Springer. v. 17, p. 203-229, 2010.

[5] DUCZMAL, L.; KULLDORFF, M.; HUANG, L. Evaluation of spatial scan statistics for irregularly shaped disease clusters. **Journal of Computational & Graphical Statistics**. Taylor & Francis. v. 15, p. 428-442, 2006.

[6] DUCZMAL, L.; CANÇADO, A.L.F.; TAKAHASHI, R.H.C.; BESSEGATO, L.F. A genetic algorithm for irregularly shaped spatial scan statistics. **Computational Statistics & Data Analysis**. Elsevier. v. 52, p. 43-52, 2007.

[7] DUCZMAL, L.; CANÇADO, A.L.F.; TAKAHASHI, R.H.C. Geographic Delineation of Disease Clusters through multi-objective Optimization. **Journal of Computational & Graphical Statistics**. Taylor & Francis. v. 17, p. 243-262, 2008.

[8] DUCZMAL, L.; DUARTE, A.R.; TAVARES, R. Extensions of the scan statistic for the detection and inference of spatial clusters. In: BALAKRISHNAN, N.; GLAZ, J. (Ed.) **Scan Statistics**. Birkhäuser. p. 157-182, 2009.

[9] DWASS, M. Modified Randomization Tests for Nonparametric Hypotheses. **Annals of Mathematical Statistics**. Project Euclid. v. 28, p. 181-187, 1957.

[10] FONSECA, C. M.; da FONSECA, V. G.; PAQUETE, L. Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function. In: LECTURE NOTES IN COMPUTER SCIENCE. **Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization**. Springer-Verlag. v. 3410, p. 250-264, 2005.

[11] FONSECA, V. G. da; FONSECA, C. M.; HALL, A. O. Inferential performance assessment of stochastic optimisers and the attainment function. In: LECTURE NOTES IN COMPUTER SCIENCE. **Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization**. Springer-Verlag. v. 1993, p. 213-225, 2001.

[12] KULLDORFF, M. A Spatial Scan Statistic. **Communications in Statistics: Theory and Methods**. Taylor & Francis. v. 26, n. 6, p. 1481-1496, 1997.

[13] KULLDORFF, M.; NAGARWALLA, N. Spatial disease clusters: detection and inference. **Statistics in Medicine**. Wiley. v. 14, p. 799-810, 1995.

[14] NAUS, J.I. The distribution of the size of the maximum cluster of pointes on the line. **Journal of the American Statistical Association**. Taylor & Francis. v. 60, p. 532-538, 1965.

[15] NAUS, J.I. Clustering of random points in two dimensions. **Biometrika**. Oxford Journals. v. 52, p. 263-267, 1965.

[16] SAHAJPAL, R.; RAMARAJU, G.V.; BHATT, V. Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. In: **International Conference on Intelligent Sensing and Information Processing**, 2004.