Um Algoritmo Enxame de Partículas Binário para Detecção de *Clusters* Irregulares

Gladston Moreira
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto, Brasil 35400-000
Email: gladston@iceb.ufop.br

Anderson Duarte
Departamento de Estatística
Universidade Federal de Ouro Preto
Ouro Preto, Brasil 35400-000
Email: duarte.andersonr@gmail.com

David Menotti
Departamento de Computação
Universidade Federal de Ouro Preto
Ouro Preto, Brasil 35400-000
Email: menottid@gmail.com

***Resumo*—Métodos para detecção e inferência de *clusters* espaciais são de grande relevância. Tanto por sua aplicabilidade em problemas de saúde pública, quanto pelo efetivo interesse científico no desenvolvimento desses métodos. A produção de sistemas de vigilância de doenças e o controle epidêmico são bastante relevantes. As principais técnicas se baseiam na estatística espacial Scan e muitas são as aplicações acoplando essa estatística à métodos de otimização já reconhecidamente eficientes. Uma aplicação específica nesse enfoque é apresentada através do algoritmo de Otimização por Enxame de Partículas Binário (*Binary Particle Swarm Optimization* – PSO). Trata-se de uma formulação de simples entendimento, fácil implementação e capaz de fornecer resultados promissores como será discutido. Os resultados obtidos são relevantes tanto em qualidade efetiva das soluções quanto no tempo de processamento necessário para obtê-los.**

I. Introdução

Estudos a respeito da incidência de doenças são, por razões óbvias, de grande interesse no meio científico. Em particular, existem diversos testes estatísticos para monitoramento e vigilância de doenças. As análises podem ser realizadas para incidência no espaço, no tempo e em ambos simultaneamente. O objetivo é verificar se o volume de ocorrências é discrepante em um subconjunto (*cluster*) espacial e/ou temporal quando comparado com toda área em estudo. Trata-se portanto de uma série de métodos com o intuito de gerar sistemas de vigilância para a detecção de surtos de doenças. Os métodos precursores apresentavam avaliações puramente espaciais e posteriormente evoluíram para os métodos de detecção de *clusters* espaço-temporais. Existem três classes bem definidas quando se busca estratificar as metodologias, os testes para interação espaço-temporal, os métodos de somas cumulativas e a estatística Scan.

A estatística espacial Scan proposta por Kulldorff [9] é a metodologia mais difundida, até o presente momento, para procedimentos de detecção e de inferência na busca por *clusters* espaciais, espaço-temporais e temporais. Originalmente, a estatística Scan é estudada por Naus [15], que descreve um método estatístico com muitas aplicações potenciais e particularmente com o objetivo de detectar um excesso significativo de eventos locais. Os métodos locais de análise espacial, em sua maioria, sofrem de problemas de testes múltiplos que levam a produção de valores p não plenamente confiáveis do ponto de vista estatístico. A estatística espacial Scan se mostra capaz de superar esse problema tomando o *cluster* mais verossímil definido pela maximização da razão de verossimilhança. Kulldorff & Nagarwalla [13] apresentam o clássico Scan Circular, um teste que encontra o *cluster* mais verossímil dentre diversos candidatos em análise. Dado um mapa subdivido em regiões, os candidatos são produzidos considerando zonas circunscritas por círculos com raios variados centrados em cada região do mapa.

Kulldorff [10] utiliza uma versão estendida da estatística espacial Scan, na qual as avaliações ocorrem não somente em domínio espacial, mas agora em cilindros espaço-temporais. Os cilindros têm base espacial seguindo o raciocínio desenvolvido previamente através do Scan Circular e com sua altura varrendo os intervalos de tempo em estudo. A análise pode ser feita de forma retrospectiva, com o interesse em detectar *clusters* ainda ativos no fim do intervalo de tempo em estudo e também *clusters* que já deixaram de existir antes do término do período; ou então pode ser uma análise prospectiva, na qual se busca somente *clusters* que ainda estão ativos no final do período (ou seja *clusters* vivos).

As técnicas de computação evolucionária são bem conhecidas por sua capacidade de pesquisa global, e têm sido bastante aplicadas para problemas de detecção de *clusters* espaciais [16], [3], [2]. O interesse está em um esquema que não restrinja diretamente o conjunto das soluções, mas que, através de mecanismos específicos, busque investigar somente algumas das possíveis soluções descartando durante o procedimento algumas soluções menos promissoras.

Um método de detecção de *clusters* espaciais usando o algoritmo *Simulated Annealing* é apresentada por Duczmal & Assunção [1]. É realizada uma busca estocástica tentando se aproximar do que seria o *cluster* verdadeiro, trata-se de uma busca aleatória em diversas zonas enquanto a estatística de teste tem valor baixo, mas a medida que este valor vai aumentando, amplia-se também a chance de o algoritmo realizar uma busca gulosa. O maior problema do algoritmo está na tendência em superestimar a verdadeira solução. Isto é, o *cluster* verdadeiro está incluso na solução encontrada, mas várias outras regiões distintas também são incluídas. Tal fato, deve-se justamente ao método permitir que a solução tenha uma forma qualquer, apenas exigindo conexidade. Em decorrência disso, a solução mais verossímil encontrada, é em alguns casos, uma coleção casual de regiões altamente verossímeis individualmente.

Já os algoritmos genéticos (AGs) abordados por Duczmal

et al. [3], [2] especificamente implementados para o problema de detecção de *clusters* possuem operadores desenvolvidos particularmente para esse problema. Inicialmente uma abordagem utilizando o AG mono-objetivo é proposta [3] e posteriormente uma versão multiobjetivo é discutida [2]. A versão inicial busca, através dos operadores genéticos, promover a procura por soluções que maximizam o valor da estatística de teste, mas sofre de problemas similares aos discutidos para o algoritmo *Simulated Annealing*. A versão multi-objetivo tenta coibir estes problemas através de funcionais penalizadores associados à forma geométrica das soluções em análise, criando um mecanismo que direciona a busca para soluções com formato mais regular, ou pelo menos não tão irregulares, ou seja, coibindo soluções que se espalham em demasia ao longo do mapa. Entretanto, o custo computacional dos procedimentos acoplados com estes funcionais tende se elevar, em geral testes de conexidade das soluções investigadas precisam ser incluídos, criando barreiras na utilização destas metodologias.

Propostas utilizando a técnica *Particle Swarm Optimization* (PSO) (veja Izakian & Pedrycz [7]) também já foram abordadas para o problema de detecção de *clusters*, mas com a avaliação de soluções puramente conexas. Essa restrição faz com que algumas soluções promissoras sejam descartadas precocemente pelo mecanismo de busca, e deixa a qualidade das soluções obtidas inferiorizada. Este trabalho apresentará uma proposta de abordagem eficaz (qualidade) e eficiente (tempo) para o procedimento de detecção de *clusters* usando a técnica PSO, considerando uma abordagem PSO binária e ainda criando um esquema mais flexível para as avaliações ligadas à conexidade das soluções. A abordagem PSO e a formulação específica para o problema em estudo serão discutidas com maior profundidade nas seções seguintes.

II. O PROCEDIMENTO DE DETECÇÃO DE *clusters* ESPACIAIS

Considerando a abordagem puramente espacial, a metodologia de detecção de *clusters* pode ser dividida em procedimentos para dados agregados por área e ainda os métodos para dados pontuais (dados do tipo caso-controle). As técnicas aqui discutidas estão voltadas inicialmente para dados agregados por área, uma proposta de estudos futuros, muito clara, fica definida na extensão desse estudo para dados pontuais. Particularmente considerando os dados agregados, suponha um mapa subdividido em m regiões, cada uma delas com população conhecida e o número de casos observados do fenômeno de interesse (a doença em avaliação) conhecidos. Em outras palavras, um caso pode ser entendido como um indivíduo infectado pela moléstia em estudo. No mapa em estudo, um *cluster* é um conjunto de regiões geograficamente limitadas no qual o risco de ocorrência do fenômeno de interesse é alto ou baixo o suficiente para ser considerado significativo do ponto de vista estatístico se comparado ao restante do mapa.

Considere para cada região do mapa em análise, um ponto arbitrário em seu interior, este ponto será denominado centroide da região. Quaisquer subconjuntos geograficamente limitados de regiões do mapa serão ditos zonas. Denote por Z o conjunto de todas as zonas em avaliação, o conjunto Z pode ser construído por todas as possíveis zonas, ou então como uma coleção das zonas que satisfazem algum conjunto de restrições. Um exemplo para construção do conjunto Z seria considerar as zonas formadas por regiões cujos centroides são interiores a uma determinada janela circular sobrepondo a área em estudo, veja a Figura 1.

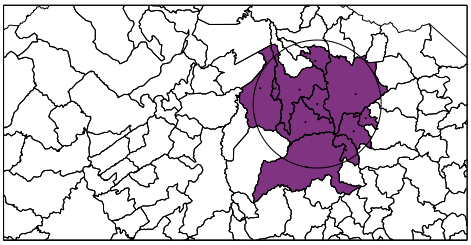

Figura 1. Uma possível zona obtida para uma dada janela circular.

Definido o conjunto Z, seja $z \in Z$ alguma zona em estudo, defina c_z, e p_z como o número de casos e a população na zona z, respectivamente. É fácil ver que o número de casos para uma determinada zona é uma variável aleatória, ou seja, c_z, observado no conjunto de dados em estudo, é apenas uma realização particular dessa variável aleatória. Dentre algumas possibilidades, um modelo bastante adequado para essa variável aleatória é admitir que siga distribuição Poisson com taxa μ_z proporcional ao tamanho populacional da região, em que $\mu_z = C(p_z/P)$. Considere ainda, que o total de casos no mapa em estudo seja representado por C e que a população total no mapa seja representada por P, e admita a validade do modelo Poisson.

Nestes termos, supondo a existência de uma zona z^* que efetivamente seja um *cluster*, defina a função de verossimilhança associada à zona z por $L(z)$. Para a suposição de não existência de *cluster* (hipótese nula para o teste estatístico) a função de verossimilhança será denotada por L_0. Kulldorff [9] mostra que o logaritmo da razão de verossimilhanças $\Lambda(z) = \log\left(L(z)/L_0\right)$ é dado por:

$$\Lambda(z) = \begin{cases} c_z \log\left(\frac{c_z}{\mu_z}\right) + (C - c_z)\log\left(\frac{C - c_z}{C - \mu_z}\right) & \text{se } c_z > \mu_z \\ 0 & \text{caso contrário} \end{cases} \quad (1)$$

A função Λ é maximizada sobre todas as zonas em Z, identificando a zona que constitui o *cluster* mais verossímil, a estatística de teste é definida então por $T = \max_{z \in Z} \Lambda(z)$.

Como mencionado anteriormente, o conjunto Z pode ser composto por todas as possíveis zonas. Por outro lado, tal escolha levaria a uma inviabilidade computacional, dado o alto volume de zonas candidatas a serem avaliadas. Busca-se então técnicas para contornar este problema, como segue:

- Redução do conjunto das soluções candidatas Z para outro conjunto Z' das zonas promissoras ou que permita uma busca exaustiva.
- Utilização de métodos estocásticos de otimização.

Nos dois casos, os métodos não são capazes de garantir a solução ótima, mas em geral na prática tendem a fornecer uma solução de boa qualidade do problema. Uma outra restrição aos métodos de detecção de *clusters* está relacionada com o formato das soluções encontradas. Muitos algoritmos não têm

procedimentos adequados para controlar as formas dos *clusters* encontrados. A solução pode às vezes se espalhar através de diversas regiões do mapa, fazendo com que se torne difícil a avaliação de seu real significado geográfico. Já alguns outros apresentam *clusters* detectados com formatos fixos. Essa seria a situação do Scan Circular [13].

Ainda pensando na construção do conjunto Z, uma proposta alternativa ao Scan Circular, flexibilizando o formato das soluções candidatas é o Scan Elíptico [12]. As premissas básicas envolvidas nessa metodologia são idênticas às discutidas para o Scan Circular. A única alteração reside na construção do conjunto das zonas candidatas a serem avaliadas. As janelas para determinação das zonas não são mais circulares, mas sim janelas elípticas. Uma elipse pode ser definida através das coordenadas x e y de seu centroide, o comprimento de seus eixos maior e menor e o ângulo entre seu eixo maior e o eixo das abscissas. Partindo de cada centroide no mapa em estudo, se forem considerados vários comprimentos para os eixos e ângulos, diversas elipses seriam construídas, para cada elipse será definida uma zona candidata que é composta pelas regiões cujos centroides são interiores à janela elíptica em utilização.

III. Algoritmo de otimização por enxame de partículas

O algoritmo de otimização por enxame de partículas (*PSO - Particle Swarm Optimization*), proposto por Kennedy & Eberhart [4], [6], é uma técnica de otimização estocástica populacional que pode ser comparada ao comportamento de um bando de pássaros ou ao comportamento sociológico de um grupo de pessoas.

É uma técnica de otimização evolucionária baseada em população que evolui de acordo com as iterações. A população é chamada de *swarm* (enxame). Considere um problema de minimização irrestrito de um funcional f. Cada partícula representa uma possível solução para o problema de otimização. Durante cada iteração, cada partícula acelera no sentido de sua melhor posição individual, assim como no sentido da melhor posição global descoberta por algumas das partículas no enxame. Isso significa que se uma partícula descobrir uma nova solução promissora, todas as outras partículas do enxame se moverão naquela direção, explorando esta região mais promissora.

Se s denota o tamanho do enxame, cada indivíduo $1 \leq i \leq s$ tem os seguintes atributos:

- uma posição atual para o $i-$ésimo indivíduo no espaço de busca $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$;
- uma velocidade atual $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$;
- a melhor posição de uma partícula definida como p_i;
- a melhor posição obtida da população como g_i.

Durante cada iteração, o método de otimização procura pela solução ótima atualizando os vetores de velocidade e posição de cada partícula de acordo com as seguintes equações:

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_{1i}(p_{id} - x_{id}^{t}) + c_2 * r_{2i}(g_{id} - x_{id}^{t}) \quad (2)$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (3)$$

em que t denota a t-ésima iteração, $d \in D$ dimensão do espaço de busca, c_1 e c_2 denotam os coeficientes de aceleração, w é chamado peso da inércia em que r_{1i} e r_{2i} é uma sequência de números aleatórios uniformemente no intervalo $(0, 1)$.

O valor de cada componente para o vetor v_i, pode ser limitado ao intervalo $[-v_{max}, v_{max}]$ para reduzir a probabilidade das partículas saírem do espaço de busca. O peso de inércia é tipicamente uma variação linear de 0.4 a 0.9 durante às gerações. Os coeficientes da aceleração c_1 e c_2, que controlam qual distância uma partícula se moverá em uma única iteração, ambos os coeficientes têm um valor de 2.0, (veja detalhes em Eberhart & Yuhui[5]).

O algoritmo de otimização por enxame de partículas foi originalmente proposto para resolver problemas com variáveis reais. No entanto, muitos problemas de otimização, como por exemplo, o problema de detecção de *clusters* espaciais ocorrem em um espaço de busca discreto. Por esta razão, [8] apresenta uma abordagem binária para o método. A equação 2 é ainda aplicada para atualizar a velocidade, em que x_{id}, p_{id} e g_d estão restritos a 1 ou 0. A velocidade na abordagem binária indica a probabilidade de o elemento da posição correspondente assumir valor 1. Uma função sigmoide $s(v_{id})$ é introduzida para transformar v_{id} para o intervalo $(0, 1)$. O algoritmo de otimização por enxame de partículas binário atualiza a posição de cada partícula de acordo com as seguintes fórmulas:

$$x_{id} = \begin{cases} 1, & rand() < s(v_{id}) \\ 0 & \text{caso contrário} \end{cases} \quad (4)$$

em que $s(v_{id}) = \frac{1}{1+\exp^{-v_{id}}}$, e $rand()$ é um gerador de números aleatórios uniformes no intervalo $(0, 1)$.

IV. Algoritmo de otimização por enxame de partículas restrito para *clusters* espaciais

Uma das questões essenciais na concepção de um algoritmo de otimização por enxame de partículas é encontrar um mapeamento adequado entre a solução do problema e a representação das partículas.

Dada uma zona z como qualquer subconjunto de regiões de um mapa com m regiões. Considere as variáveis binárias $\{x_1, x_2, \ldots, x_m\}$, onde $x_i = 1$ se a i-ésima região está presente na zona z e 0 caso contrário, ou seja, o problema de detecção de *clusters* espaciais é um problema de otimização combinatória.

Neste estudo, propomos uma abordagem eficiente, que pode encontrar *clusters* espaciais de forma geral, em particular, aqueles de formato irregular. Para este fim, foi utilizada uma restrição que pode ser definida como uma janela de exploração, para encontrar *clusters* com formas irregulares, [14].

Note que o algoritmo, se proposto sem restrições sobre o tamanho dos subconjuntos (soluções), permite *clusters* com formatos arbitrários, ou seja, pode ser um conjunto de zonas com alta probabilidade de se espalhar de forma aleatória por todo o mapa, que na prática não é uma solução razoável para o problema. A seguir, é definida a restrição geográfica como uma janela exploratória.

A. Janela Exploratória

Considere um mapa subdividido em m regiões e um valor fixo k, com $1 \leq k \leq m$. Para cada região $i \in \{1, \ldots, m\}$, a_i é o centroide da referida região (um ponto arbitrário em seu interior). Seja $d(a_i, a_j)$ a distância Euclidiana entre os centroides a_i e a_j. Para cada região i do mapa de estudo, é definido um sub-mapa S_i constituído da região i e as k regiões mais próximas da região i de acordo com as distâncias euclidianas em relação ao seu centroide. Cada S_i é uma janela exploratória de tamanho $k+1$.

Assim, o método de otimização é usado para explorar os *clusters* nos sub-mapas S_i, $1 \leq k \leq m$. Os *clusters* encontrados, em alguns casos, podem ser desconexos, entretanto isso não configura uma desvantagem, pois não existe uma enorme lacuna entre as suas áreas. Isso permite controlar o número de áreas que compõem o *cluster*, permitindo apenas *clusters* que são formados por regiões de uma zona circular de tamanho máximo moderado.

Neste estudo, a estrutura de cada partícula representa uma zona (*cluster*) arbitrária dado um tamanho de janela $k < m$, para um mapa com m regiões. A estrutura de cada partícula no PSO é um vetor $k-$dimensional binário. O algoritmo foi implementado no ambiente MATLAB.

V. Experimentos e Resultados

Nesta seção, são descritos alguns experimentos numéricos utilizando um *benchmark* de dados reais, a fim de comparar a abordagem proposta utilizando o PSO com o Scan Elíptico [12]. Para essa versão elíptica da estatística espacial Scan, foi utilizada uma janela elíptica de forma variável. A sua excentricidade foi variada considerando os valores $1, 1.5, 2, 3$ e 4. Para cada valor de excentricidade, foram utilizados os ângulos de rotação do sistema de coordenadas de $1, 4, 6, 9$ e 12 radianos. Para o PSO, usamos a seguinte configuração: utilizamos cinco diferentes tamanhos para a janela exploratória, $k \in \{5, 10, 15, 20, 25\}$. O tamanho da população é igual ao tamanho da janela exploratória corrente e o critério de parada com número máximo de gerações igual 20 ou número máximo de estabilização para a função objetivo igual a 7. Os experimentos foram executados em processador Intel Core i5, 2.5 GHz e 8 GB de memória RAM.

Foi utilizado como aplicação um conjunto de dados de referência para dados de câncer de mama ocorridos no Nordeste dos Estados Unidos. A população total de risco é de 29.535.210 mulheres com um total de 58.943 casos no período de 1988 a 1992 em 245 municípios, [11]. A figura 2 ilustra a distribuição populacional (população sob risco) e também a incidência dos casos de câncer de mama ao longo das 245 regiões no mapa em estudo. Em ambos, os tons mais escuros representam maiores valores e os tons mais claros representam menores valores.

As tabelas I e II mostram os resultados obtidos pela abordagem PSO e o Scan Elíptico, para cinco diferentes tamanhos da janela exploratória.

A tabela I mostra uma comparação do valor da estatística de teste encontrados pelos métodos PSO e o Scan Elíptico, além do tempo de execução em segundos deles. Para o PSO o tempo de execução é tomado como uma média de 100 execuções independentes, o mesmo para o valor da estatística de teste. O método PSO é um pouco mais lento que o método Scan Elíptico, entretanto o valor da estatística de teste alcançado através das soluções fornecidas pelo PSO é sistematicamente mais elevado.

Figura 2. Esquerda: Mapa de população de risco; Direita: Mapa de incidência de casos de câncer.

Tabela I. Comparação entre os métodos PSO e Scan Elíptico.

	Janela	Λ		Tempo (s)	
		Média	Desvio Padrão	Média	Desvio Padrão
PSO	5	64,9456	1,4539	1,2939	0,0202
	10	76,3187	1,8345	1,8633	0,0312
	15	76,1193	3,4857	2,4456	0,0582
	20	77,5302	4,5542	3,0473	0.0331
	25	100,0441	6,5000	3,7320	0.0456
Scan Elíptico	5	54,4833		0,8594	
	10	55,2590		1,2813	
	15	55,2590		1,7656	
	20	55,2590		2,1875	
	25	71,0046		2,5156	

Vale ressaltar que quanto ao custo em tempo de processamento, em geral, o algoritmo PSO realizou o procedimento em um tempo 50% maior que o Scan elíptico, mas mesmo com o aumento da dimensão da janela, a diferença de tempo se mantém estável na faixa de 50%, ou seja, não apresentando algum deficiência de crescimento desordenado do custo em tempo de processamento com o aumento do tamanho da janela.

Além disso, para todas as configurações, os dois algoritmo são realizadas 999 execuções considerando dados simulados através de procedimentos de Monte Carlo. Nas simulações, o total de casos C é preservado e os casos são distribuídos considerando a validade da hipótese nula de não existência de *cluster* no mapa em estudo. O p-valor estimado para a solução dos dados observados considera o percentual de soluções sob a hipótese nula que não são ultrapassados pelo valor da estatística de teste nos dados observados.

A tabela II apresenta os resultados obtidos usando a abordagem PSO e o Scan Elíptico. Para o PSO-algoritmo é considerado a solução com melhor valor de estatística de teste dentre 100 execuções independentes realizadas para os casos observados.

As figuras 3 a 7 apresentam as soluções mencionadas na tabela II tanto para a abordagem através do algoritmo PSO quanto para o Scan Elíptico. Inicialmente é possível observar que o PSO admite soluções desconexas, porém não extremamente espalhadas ao longo do mapa, isso ocorre em virtude do mecanismo de divisão do mapa em sub-mapas.

Tabela II. RESULTADOS PARA 999 EXECUÇÕES DE MONTE CARLO OBTIDOS PELOS MÉTODOS PSO E SCAN ELÍPTICO.

	Janela	Λ	Casos	Casos Esperados	p-valor
PSO	5	65.2000	2689	2149,60	0,001
	10	78,3173	3493	2821,30	0,001
	15	83,6474	5646	4767,00	0,001
	20	94,6727	6324	5339,80	0,001
	25	119,5597	9997	8641,60	0,001
Elíptico Scan	5	54,4833	5026	4348,30	0,001
	10	55,2590	4585	3932,50	0.001
	15	55,2590	4585	3932,50	0.001
	20	55,2590	4585	3932,50	0.001
	25	71,0046	6825	5935,80	0.001

Para soluções de tamanho reduzido, as soluções tendem a ser semelhantes para os dos métodos, a figura 3 apresenta soluções obtidas na mesma região do mapa em estudo. Na prática a solução PSO está contida na solução elíptica. Pode estar ocorrendo um caso de superestimação da solução elíptica em relação a solução PSO, ou subestimação da solução PSO em relação a solução elíptica, entretanto o fato da solução PSO apresentar maior valor para a estatística de teste é um efetivo indicativo de qualidade.

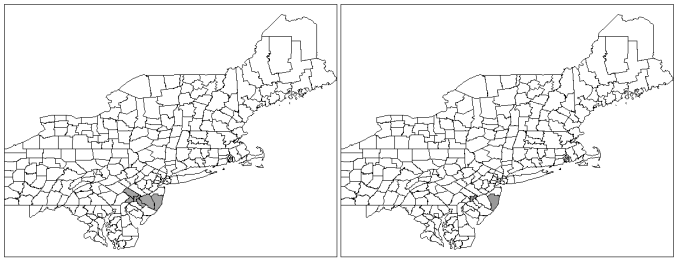

Figura 3. *Clusters* encontrados pelos algoritmos para casos de Câncer de mama, usando uma janela de tamanho 5: Esquerda: Elíptico Scan; Direita: PSO.

A medida que as soluções são investigadas para janelas maiores, a solução obtida pelo Scan elíptico permanece fixa e se desloca em uma direção um pouco diferente do mapa se comparada a solução inicial, para tamanho de janela 5. As soluções do algoritmo PSO sofrem variação pequena mas permanecem na mesma vizinhança do mapa em que foi encontrada a solução para janela de tamanho 5. Em todos os casos a estatística de teste para a solução PSO apresenta valores superiores, ou seja, soluções mais verossímeis. Estes fatos pode ser verificados nas figuras 4 até 6.

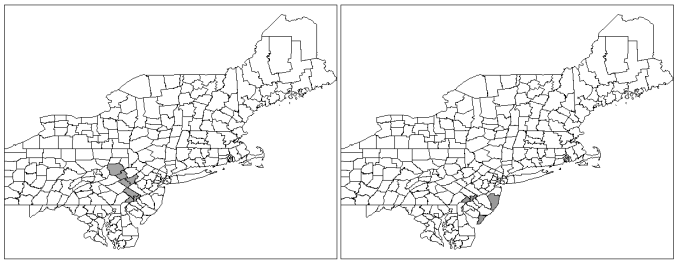

Figura 4. *Clusters* encontrados pelos algoritmos para casos de Câncer de mama, usando uma janela de tamanho 10: Esquerda: Elíptico Scan; Direita: PSO.

Avaliando as soluções com tamanho de janela 25 através da figura 7, nota-se que a solução PSO se mantém concentrada

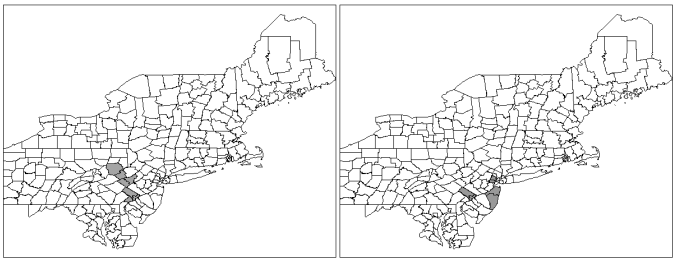

Figura 5. *Clusters* encontrados pelos algoritmos para casos de Câncer de mama, usando uma janela de tamanho 15: Esquerda: Elíptico Scan; Direita: PSO.

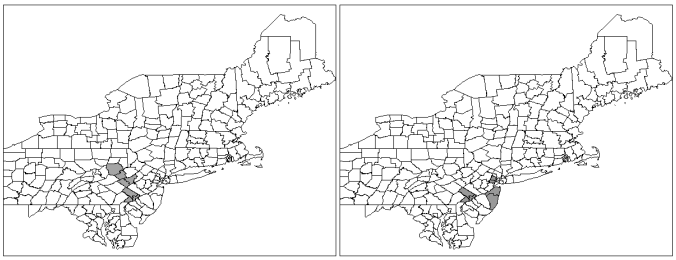

Figura 6. *Clusters* encontrados pelos algoritmos para casos de Câncer de mama, usando uma janela de tamanho 20: Esquerda: Elíptico Scan; Direita: PSO.

na mesma vizinhança do mapa mas eleva ainda mais o valor da estatística de teste. A solução elíptica começa a se espalhar no mapa perdendo um pouco do seu real sentido geográfico. A solução se espalha para conectar áreas com taxa de incidência elevada. Por outro lado, incidência elevada não é suficiente para caracterizar um problema de saúde pública nesse enfoque, a principal evidência pode ser observada através do aumento efetivo do risco relativo (a razão entre os casos observados e esperados) na zona em avaliação.

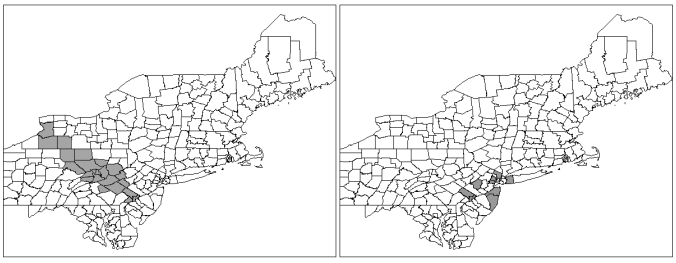

Figura 7. *Clusters* encontrados pelos algoritmos para casos de Câncer de mama, usando uma janela de tamanho 25: Esquerda: Elíptico Scan; Direita: PSO.

VI. CONCLUSÃO

O problema de detecção de *clusters* espaciais foi abordado. A sua relevância é explícita quando considera-se o crescente interesse por dominar técnicas para vigilância de doenças e o controle de surtos epidêmicos. As estratégias mais difundidas são mencionadas e resumidas quanto aos seus propósitos e as suas técnicas de utilização.

O algoritmo PSO foi discutido quanto as suas premissas básicas e posteriormente foi adaptado de forma específica

para o problema de detecção de *clusters* espaciais. A nova metodologia foi delineada e testada para um conjunto de dados já utilizado em outras aplicações de pesquisa nessa área.

A metodologia em estudo apresentou resultados consistentes para o problema, principalmente quando comparados aos resultados obtidos através do clássico Scan Elíptico. A capacidade do PSO-algoritmo fornecer soluções cuja estatística de teste é sistematicamente maior que as das soluções obtidas através do Scan Elíptico ficou expressa na análise dos experimentos.

Os tempos necessários para o processamento, previsivelmente, mostram o PSO-algoritmo um pouco mais lento quando comparado o Scan Elíptico. Essa expectativa se deve ao fato de o Scan Elíptico pré-determinar o conjunto Z das zonas que serão avaliadas previamente e não utilizar estratégias de otimização direcionando as soluções a serem avaliadas.

Entretanto, os tempos de processamento do PSO não foram superiores de forma proibitiva, em todos os casos de estudo (diferentes tamanhos de janela limitante), o tempo foi aproximadamente 50% maior que o tempo de processamento para o Scan Elíptico. Em outras palavras, trata-se de um aumento computacional controlável para a obtenção de soluções sistematicamente superiores.

Os resultados são superiores tanto na efetividade das soluções obtidas (soluções com maiores valores para a estatística de teste), mas também ao oferecer um estratégia de busca das soluções através de um custo computacional sistematicamente competitivo. Vale ressaltar que trata-se de uma discussão que ainda pode evoluir através de propostas de continuidade como a inclusão de mecanismos de regularidade para a forma das soluções obtidas e também a proposição de uma formulação mais robusta através de uma abordagem multiobjetivo.

AGRADECIMENTOS

Os autores agradecem o apoio da UFOP e das agências CAPES, CNPq e FAPEMIG.

REFERÊNCIAS

[1] L. Duczmal and R. Assunção, "A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters," *Computational Statistics & Data Analysis*, vol. 45, no. 2, pp. 269–286, 2004.

[2] L. Duczmal, A. L. F. Cançado, and R. H. C. Takahashi, "Geographic delineation of disease clusters through multi-objective optimization," *Journal of Computational & Graphical Statistics*, vol. 17, pp. 243–262, 2008.

[3] L. Duczmal, A. L. F. Cançado, R. H. C. Takahashi, and L. F. Bessegato, "A genetic algorithm for irregularly shaped spatial scan statistics," *Computational Statistics and Data Analysis*, vol. 52, pp. 43–52, 2007.

[4] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, 1995, pp. 39–43.

[5] R. C. Eberhart and S. Yuhui, "Particle swarm optimization: developments, applications and resources," in *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, vol. 1, 2001, pp. 81–86.

[6] R. Eberhart, P. Simpson, and R. Dobbins, *Computational Intelligence PC Tools*. San Diego, CA, USA: Academic Press Professional, Inc., 1996.

[7] H. Izakian and W. Pedrycz, "A new pso-optimized geometry of spatial and spatio-temporal scan statistics for disease outbreak detection," *Swarm and Evolutionary Computation*, vol. 4, no. 0, pp. 1 – 11, 2012.

[8] J. Kennedy and R. Eberhart, "A discrete binary version of the particle swarm algorithm," in *IEEE International Conference on Systems, Man, and Cybernetics*, 1997, p. 4104?4108.

[9] M. Kulldorff, "A spatial scan statistic." *Communications in Statistics: Theory and Methods*, vol. 26, pp. 1481–1496, 1997.

[10] ——, "Prospective time periodic geographical disease surveillance using a scan statistic." *J. R. Statistic Soc. A*, vol. 164, no. 1, pp. 61–72, 2001.

[11] M. Kulldorff, E. J. Feuer, B. A. Miller, and L. S. Freedman, "Breast cancer clusters in the northeast united states: a geographic analysis." *American Journal of Epidemiology*, vol. 146, p. 161?170, 1997.

[12] M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal, "An elliptic spatial scan statistic." *Statistics in Medicine*, vol. 26, pp. 3929–3943, 2006.

[13] M. Kulldorff and N. Nagarwalla, "Spatial disease clusters: Detection and inference," *Statistics in Medicine*, vol. 14, no. 8, pp. 799–810, 1995.

[14] G. J. P. Moreira, L. Paquete, L. Duczmal, D. Menotti, and R. H. C. Takahashi, "Multi-objective dynamic programming for spatial cluster detection," *Environmental and Ecological Statistics*, vol. 22, no. 2, pp. 369–391, 2015.

[15] J. Naus, "The distribution of the size of maximum cluster of points on the line." *Journal of the American Statistical Association*, vol. 60, pp. 532–538, 1965.

[16] X. Wu and T. Grubesic, "Identifying irregularly shaped crime hot-spots using a multiobjective evolutionary algorithm," *Journal of Geographical Systems*, vol. 12, no. 4, pp. 409–433, 2010.