

Utilizando Pesquisa Operacional para a Investigação de surtos de doenças em apiários no estado do Rio de Janeiro

Soline Ikeda ¹; Wagner de Souza Tassinari ² ; Maria Cristina Afonso Lorenzon³; Marcelo Dib Cruz⁴

1. Bolsista PROIC, Discente do Curso de Sistemas de Informação, Demat/UFRJ; 2. Professor Demat/ICE/UFRRJ; 3. Professor do DPA/IZ/UFRRJ. 4. Professor do Demat/ICE/UFRRJ.

Palavras-chave: Aglomerados Espaciais, Epidemiologia Veterinária, Apicultura, Programação Inteira.

Introdução

Existem hoje, no mundo perto de 30 mil espécies de abelhas, esta diversidade em consonância com a alta abundância, são responsáveis por mais de 70% da produção agrícola mundial (Gallai et al., 2008; Neumann & Carreck, 2010), além da produção destacada de mel e outros produtos apícolas. Nos últimos dez anos, colônias de abelhas começaram a desaparecer em diferentes países do mundo. Nos Estados Unidos, de acordo com o Departamento de Agricultura dos Estados Unidos (USDA), perto de 30 e 70% das abelhas sumiram nos estados da Califórnia e Texas. Em algumas regiões da Europa, a diminuição chega a 80% (Steinhauer et al., 2014). O desaparecimento das colmeias tem acarretado impacto direto sobre a produção de alimentos (Kevan et al., 2010).

Fato é que a década do ano 2000 assinala, em muitas regiões do mundo, inclusive no Brasil, o momento de exacerbação do perfil de mortalidade das abelhas por fatores diversos como, intoxicação, doenças e parasitoses e, algumas de etiologia desconhecida (Pacheco et al., 2009). O intenso comércio de colmeias e produtos apícolas é um dos principais causadores da propagação das doenças das abelhas melíferas, que aliados à multiplicidade de outros fatores, ampliam o quadro estrutural de sanidade, em déficit, pela falta da aplicação de medidas profiláticas, pelo despreparo técnico, pela ausência de padrões de seleção de colmeias, pelos cuidados com o meio ambiente, entre outros fatores (Allardice et al., 1999).

Na apicultura as perdas anuais de colmeias acima de 20% traduz a relevância do diagnóstico e do controle de doenças, parasitoses e intoxicações. As autoridades responsáveis pela apicultura nacional estão continuamente confrontadas com novos quadros clínicos e agentes patogênicos. Consequentemente, há necessidade de se disponibilizar de exames laboratoriais para se obter diagnósticos das patologias apícolas como medidas essenciais de controle (Lorenzon et al., 2012).

Como ação para um retorno rápido de um estudo para um diagnóstico coletivo a nível populacional, o inquérito epidemiológico é importante ferramenta analítica para a investigação dos fatores de risco presentes na causalidade das doenças e na caracterização de sua distribuição, no espaço e no tempo no âmbito populacional. Desde a década de 80, surge um renovado interesse nos estudos de padrões espaciais e temporais de doenças, conforme salienta a extensa literatura publicada em periódicos de diferentes áreas, incluindo importantes revisões (Knox, 1991; Werneck e Struchiner, 1997). Dentre os desenhos epidemiológicos utilizados neste contexto, destacam-se os estudos de agregados (*clusters*, na língua inglesa). De maneira geral, aglomerados espaciais de doenças podem ser atribuídos aos fatores demográficos, genéticos, ambientais ou, socioculturais superpostos geograficamente ao padrão de ocorrência observado. O estudo de técnicas para detecção de aglomerados espaciais no campo da Epidemiologia recebe importante contribuição científica (Werneck e Struchiner, 1997).

A Pesquisa Operacional (Definição segundo a SOBRAPO – Sociedade Brasileira de Pesquisa Operacional) é uma ciência aplicada voltada para a resolução de problemas reais. Tendo como foco a tomada de decisões, aplica conceitos e métodos de várias áreas científicas na concepção, planejamento ou operação de sistemas. A Pesquisa Operacional surgiu durante a Segunda Guerra Mundial, da necessidade de lidar com problemas de natureza logística, tática e de estratégia militar de grande dimensão e complexidade. Para apoiar os comandos operacionais na resolução desses problemas, foram então criados grupos multidisciplinares de

matemáticos, físicos e engenheiros e cientistas sociais. Esses cientistas não fizeram mais do que aplicar o método científico, que tão bem conheciam, aos problemas que lhes foram sendo colocados. Desenvolveram então a ideia de criar modelos matemáticos, apoiados em dados e fatos, que lhes permitissem perceber os problemas em estudo e simular e avaliar o resultado hipotético de estratégias ou decisões alternativas. Para resolver os modelos matemáticos existem algumas técnicas, dentre as quais podemos citar Programação Linear e métodos heurísticos utilizando Metaheurística.

Programação Linear é uma das técnicas da Pesquisa Operacional das mais utilizadas e que encontra valores exatos para os modelos propostos. Os problemas de Programação Linear (PL) buscam a distribuição eficiente de recursos limitados para atender um determinado objetivo, em geral, maximizar lucros ou minimizar custos. Em se tratando de PL, esse objetivo é expresso através de uma função linear, denominada de "Função Objetivo". É necessário também que se defina quais as atividades que consomem recursos e em que proporções os mesmos são consumidos. Essas informações são apresentadas em forma de equações as inequações lineares, uma para cada recurso. Ao conjunto dessas equações e/ou inequações, denomina-se "Restrições do Modelo". Na Programação Linear Inteira as soluções encontradas possuem valores inteiros. (Bazaraa et al, 1990)

Um algoritmo é considerado um método heurístico quando não há conhecimentos matemáticos completos sobre seu comportamento, ou seja, quando, sem oferecer garantias, o algoritmo objetiva resolver problemas complexos utilizando uma quantidade não muito grande de recursos - especialmente no que diz respeito ao consumo de tempo - para encontrar soluções de boa qualidade. Uma metaheurística é um conjunto de conceitos que pode ser utilizado para definir métodos heurísticos aplicáveis a um extenso conjunto de diferentes problemas. Em outras palavras, uma metaheurística pode ser vista como uma estrutura algorítmica geral que pode ser aplicada a diferentes problemas de otimização com relativamente poucas modificações que possam adaptá-la a um problema específico. Alguns exemplos de metaheurísticas são: simulated annealing, busca tabu, iterated local search, algoritmos evolutivos e ant colony optimization" . (Glover e Kochenberger, 2003)

O problema acima descrito pode ser definido na Pesquisa Operacional como um problema de Clusterização ou Agrupamento (Cruz, 2010; Bastos et al., 2014). Clusterização é o termo genérico para um processo que une objetos similares em um mesmo grupo. Cada grupo é denominado um cluster. O número de clusters pode ser conhecido a priori ou não. Quando o número de clusters é conhecido, a priori, tratamos Problema de K-Clusterização ou, simplesmente Problema de Clusterização (PC). Caso contrário, ou seja, quando não se conhece o número de grupos, o problema é denominado Problema de Clusterização Automática (PCA). Existem várias aplicações relacionadas à clusterização incluindo: Particionamento de Grafos, Problema de Manufatura flexível, Reconhecimento de padrões onde se destacam o processamento de Imagens, Biologia Computacional, Pesquisa de Mercado, Classificação de Documentos, Mineração de Dados, entre outros. O problema de clusterização pode ser resolvido utilizando Programação Linear Inteira ou heurística baseada em metaheurística.

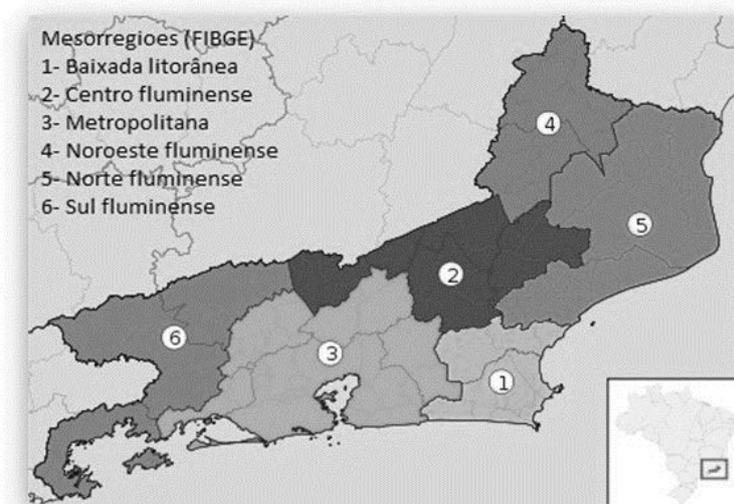
O objetivo deste trabalho é investigar a existência de aglomerados espaciais de doenças em apiários no estado do Rio de Janeiro utilizando pesquisa Operacional, com o intuito de verificar quais são as regiões de maior prioridade para a intervenção na cadeia apícola. Este problema é tratado de duas maneiras diferentes: Na primeira, o objetivo é dividir os casos de doenças em um número pré-definido de grupos (PC). Este processo pode ser utilizado, por exemplo, no caso da existência de um número pré-definido de equipes que investigarão os aglomerados espaciais das doenças, uma equipe para cada aglomerado. Neste caso, dado o número de grupos, o que se deseja é encontrar as configurações dos grupos. Outra interpretação possível é fazer uma simulação dos diâmetros médios dos clusters, pois a proximidade geográfica dos apiários é um fator de grande importância na análise de surtos de doenças. A diminuição ou aumento do diâmetro médio dos clusters é alcançado com o aumento ou diminuição do número de clusters. Foi realizada uma revisão na literatura e encontraram-se dois modelos exatos de Programação Linear Inteira que resolvem o problema. O primeiro proposto por (Rao ,1971) e o segundo proposto por (Vinod ,1969) e conhecido como k-mediana. Na segunda maneira, o que se deseja é encontrar possíveis grupos, sem saber a priori, quantos e quais são (PCA), para investigar os grupos naturalmente formados, e com isso verificar possíveis surtos de doenças. Foi realizada outra revisão na literatura, porém desta vez, não foi encontrado nenhum modelo linear Inteiro para o problema. Por isso, foi construída uma heurística utilizando a metaheurística ILS.

Metodologia

O presente estudo utilizou dados de 38 municípios do estado do Rio de Janeiro (22°54'23"S, 43°10'21"W): Nova Friburgo, Itaocara, Paty do Alferes, Itaguaí, Paraíba do Sul, Teresópolis, São Fidélis, Sumidouro, Rio Claro, Sapucaia, Piraí, Resende, Trajano de Moraes, Barra do Piraí, Miguel Pereira, Petrópolis, Valença, Itatiaia, São José do Vale do Rio Preto, Tanguá, Araruama, Três Rios, Barra Mansa, Cambuci, Engenheiro Paulo de Frontin, Porciúncula, Rio Bonito, Vassouras, Casimiro de Abreu, Paracambi, Silva Jardim, Natividade, Aperibé, Comendador Levy Gasparian, Mendes, Pinheiral, Santa Maria Madalena e Volta Redonda. Este grupo contempla unidades de todas as mesorregiões do estado do Rio de Janeiro (figura 1).

A base de dados utilizada neste estudo é proveniente do levantamento dos déficits da apicultura do estado do Rio de Janeiro, e foram coletados no período de 2009 a 2010, por meio de aplicação de questionários por técnicos dos Núcleos de Defesa Agropecuária (NDA's), vinculados a Secretaria de Estado de Agricultura e Pecuária do Estado do Rio de Janeiro (SEAPEC) (Lorenzon, 2008). Ao todo, 401 apicultores participaram deste levantamento, mas para o presente estudo foram selecionados aqueles que responderam sobre a ocorrência de doenças em seus apiários (n= 269). As coordenadas geográficas dos apiários foram identificadas, via GPS, no momento da aplicação dos questionários e utilizadas em formato *Universal Transversa de Mercator (UTM)*.

Figura 1: Distribuição Geográfica das Mesorregiões do estado do Rio de Janeiro



Para este trabalho, foram utilizados dois modelos exatos de Programação Linear Inteira para o problema de Clusterização. O primeiro proposto por Rao (1971) e o segundo proposto por Vinod (1969) e conhecido como k-medianas, conforme descritos abaixo.

O primeiro modelo (Rao, 1971), é um modelo linear inteiro 0-1. O objetivo deste modelo é minimizar a maior distância interna de todos os clusters, ou ainda, minimizar a maior diagonal de todos os clusters. Cada ponto só pode pertencer a um cluster e o número de clusters k é um dado de entrada. A grande dificuldade deste modelo é que o número de restrições cresce rapidamente, à medida que as quantidades de pontos m e de clusters k crescem. Portanto, o modelo exato Diâmetro é indicado somente para estes valores pequenos. O modelo é descrito abaixo:

$$\text{Min } Z = D_{\max} \quad (1)$$

$$\text{S.A. } D_l \geq d_{ij}(x_{il} - x_{jl} - 1) \quad i, j, l, \quad i = 1, \dots, m, \quad j = 1, \dots, m, \quad l = 1, \dots, k \quad (2)$$

$$\sum_{l=1}^k x_{il} = 1 \quad i, \quad i = 1, \dots, m \quad (3)$$

$$D_{\max} \geq D_l \quad l, \quad l = 1, \dots, k \quad (4)$$

$$x_{il} \in \{0,1\} \quad i, l, \quad i = 1, \dots, m, \quad l = 1, \dots, k$$

$$D_l \geq 0 \quad l, \quad l = 1, \dots, k$$

onde $x_{ij} = 1$, se o ponto i pertence ao cluster j
 0 , caso contrário.

d_{ij} distância entre o ponto i e o ponto j .

D_l diâmetro do cluster l .

D_{\max} o maior diâmetro de todos os clusters.

A restrição (2) garante que a diagonal de cada cluster é maior ou igual a distância máxima entre dois pontos dentro do cluster. A restrição (3) garante que cada ponto só pode ser alocado a um cluster e, finalmente, a restrição (4) garante que D_{\max} seja maior que todas as outras diagonais de todos os clusters.

Já o segundo modelo, o objetivo é selecionar k pontos médios do conjunto de pontos X , um ponto médio para cada cluster, onde a soma das distâncias dos outros pontos do cluster a este ponto médio é mínima. Cada ponto só pode pertencer a um cluster e o número de clusters k é um dado de entrada. O ponto médio de um cluster é definido como sendo aquele ponto que pertence ao cluster e que está mais próximo a todos os outros pontos do cluster.

O modelo K-medianas é mostrado abaixo. A restrição (6) assegura que cada ponto será associado a somente um ponto médio. Restrição (7) assegura que serão escolhidos exatamente k pontos médios e a restrição (8) que cada ponto i está associado a um ponto j se e somente se o ponto j é um ponto médio.

$$\text{Min } \sum_{i=1}^n \sum_{j=1}^n d_{ij} x_{ij} \quad (5)$$

$$\text{S.A. } \sum_{j=1}^n x_{ij} = 1 \quad i, \quad i = 1, \dots, n \quad (6)$$

$$\sum_{j=1}^n x_{jj} = k \quad (7)$$

$$x_{ij} \leq x_{jj} \quad i, j, \quad i = 1, \dots, n, \quad j = 1, \dots, n \quad (8)$$

$$x_{ij} \in \{0,1\} \quad i, j, \quad i = 1, \dots, n, \quad j = 1, \dots, n \quad (9)$$

onde $x_{ij} = 1$, se o ponto i está associado ao ponto médio j
 0 , caso contrário.

$x_{jj} = 1$, significa que o ponto j é o ponto médio.

d_{ij} distância entre o pontos i e o ponto j .

A metaheurística ILS (Iterated Local Search), proposta por Lourenço, Martin e Stützle (Glover e Kochenberger, 2003), consiste, essencialmente, na aplicação iterativa de um procedimento de busca local em uma solução inicial, que é previamente obtida a partir da utilização de um procedimento aleatório de construção ou considerando uma heurística de construção. O procedimento de busca local tem por finalidade melhorar a solução inicial e aquelas produzidas após perturbações de soluções ótimas locais.

O *Algoritmo ILS para o PCA (ILS_PCA)*, aqui proposto, é uma heurística baseada na metaheurística ILS composta por quatro componentes principais: a Fase de Construção, as Buscas Locais, a Perturbação e o Critério de Aceitação.

A Fase de construção do ILS_PCA é uma etapa inicial que tem por objetivos tentar reduzir a cardinalidade dos dados de entrada do problema, bem como gerar uma solução inicial de boa qualidade. A idéia é substituir grupos de objetos da base de dados cuja similaridade é considerada alta, por um único objeto (*cluster inicial*) que represente o grupo.

A redução do tamanho da entrada (pré-processamento) é realizada agrupando-se em um mesmo cluster os pontos pertencentes a uma região densa. Inicialmente, nas linhas 2, 3 e 4, para cada ponto é definido a menor distância a outro ponto qualquer. Depois, na linha 5, é calculada a média destas distâncias, denominada d_{medio} . Então, cada ponto $x_i \in X$ é considerado o centro de um círculo cujo valor do raio é $r = u * d_{medio}$, onde u é um parâmetro de entrada. Logo após, na linha 8, é calculado o conjunto de pontos contidos no círculo de centro x_i e raio r ($N_i = \text{circulo}(x_i, r)$). Estes valores são colocados em uma lista T que é ordenada em ordem decrescente. Os elementos de T são considerados os clusters parciais $B = \{B_1, B_2, \dots, B_i\}$. Para que os clusters não possuam elementos em comum, toda vez que um círculo é selecionado, todos os pontos do seu interior não podem mais entrar em outro círculo. Com este procedimento as regiões mais densas são selecionadas.

Figura 2: Descrição do procedimento construtivo

1. **Procedimento Construtivo (X, u)**
2. **PARA** $i = 1$ **até** n **FAÇA**
3. $d_{\min}(x_i) = \min \|x_i - x_j\|, i \neq j, j = 1, \dots, n$
4. **FIM PARA**
5. $d_{medio} = \frac{1}{n} \sum_{i=1}^n d_{\min}(x_i)$
6. $r = u * d_{medio}$
7. **PARA** $i = 1$ **até** n **FAÇA**
8. $N_i = \text{circulo}(x_i, r)$
9. $T = T \cup N_i$
10. **FIM PARA**
11. ordenar T em ordem decrescente
12. $i = 1$
13. **ENQUANTO** ($T \neq \emptyset$) **FAÇA**
14. $B_i = \text{próximo}(N_j \in T)$
15. $T = T - \{N_j\}$
16. $i = i + 1$
17. **FIM ENQUANTO**
18. retornar $B = \{B_1, B_2, \dots, B_i\}$, os clusters parciais.
19. **FIM construtivo**

Considere os clusters iniciais gerados como $B = \{B_1, B_2, \dots, B_m\}$ e seja $v_i, i = 1, 2, \dots, m$ o centróide de cada cluster B_i . Para representar uma solução é utilizada uma cadeia binária de m posições. Por exemplo, se $m = 7$, então uma cadeia binária poderia ser $\{0110010\}$. Se o valor correspondente ao B_i na cadeia binária for igual a 1, isso significa que o cluster inicial B_i faz parte da solução como cluster pai. Se o valor correspondente ao B_i na cadeia binária for igual a 0, B_i é considerado um cluster filho. Os clusters filhos são unidos aos clusters pais utilizando o critério de menor distância entre os centróides. A cada união, o valor do centróide do cluster pai é recalculado. No final, todos os clusters filhos são unidos aos clusters pais para gerar uma solução completa. Portanto, nesta representação, a cada nova solução podemos ter um

número distinto de clusters pais, que não se alteram depois deste processo. Os clusters gerados após esse processo são denominados clusters finais $C = \{C_1, C_2, \dots, C_k\}$.

Para avaliar a qualidade da solução encontrada, é utilizada uma função de avaliação F (equação 14), proposta por (Kaufman e Rousseeum, 1990), conhecida como *Índice Silhueta*. Esta função admite valores dentro do intervalo $[-1, 1]$ e não é necessário definir o valor de k , pois o valor mais adequado de k é atingido ao *maximizar* esta função. Seja x_i um ponto pertencente ao cluster $C_w \in C$, com $|C_w| = M > 1$. A distância entre os objetos x_i e x_j é definido por d_{ij} . A similaridade média de x_i em relação a todos os pontos $x_j \in C_w$ é dada por $a(x_i)$ onde

$$a(x_i) = \frac{1}{M-1} \sum_{x_j \in C_w, x_j \neq x_i} d_{ij} \quad (10)$$

Nos casos em que C_w possuir um único elemento, definimos $a(x_i) = 0$. Considere ainda, cada um dos clusters $C_t \in C$ com $t \neq w$ e $|C_t| = T > 1$. A similaridade média do ponto x_i em relação a todos os pontos de C_t é

$$d(x_i, C_t) = \frac{1}{T} \sum_{x_j \in C_t} d_{ij} \quad (11)$$

Seja $b(x_i)$ o menor valor dentre todos os $d(x_i, C_t)$. Então,

$$b(x_i) = \min_{C_t \in C, C_t \neq C_w} d(x_i, C_t) \quad (12)$$

O valor de Silhueta do ponto $x_i \in X$ é dado por

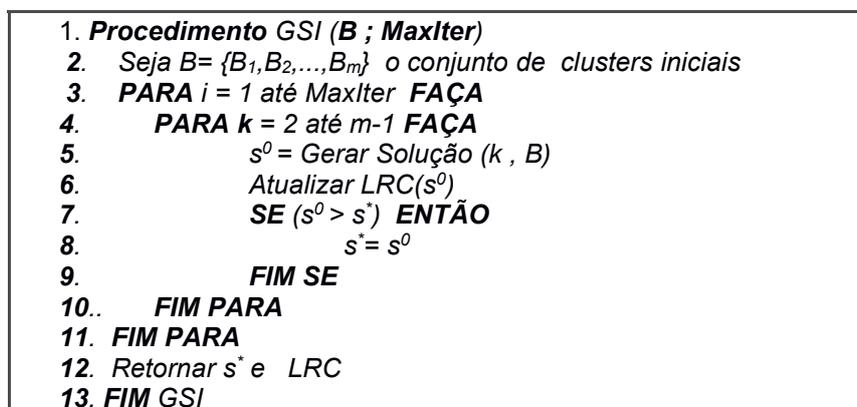
$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))} \quad (13)$$

e a função F , denominada Índice Silhueta, é definida como:

$$F = \frac{1}{n} \sum_{i=1}^n s(x_i) \quad (14)$$

Para gerar uma solução, inicialmente é realizado um procedimento para construir uma lista LRC. Nesta etapa, várias soluções são geradas, e o número de clusters pais das melhores soluções são armazenadas na LRC.

Figura 3: O procedimento GSI



O procedimento denominado *Gerar Solução Inicial (GSI)* é mostrado na figura 3. Na linha 1, os clusters iniciais são dados e são armazenados no conjunto B . Depois é feito um processo iterativo, que a cada momento, gera uma solução s^0 com k clusters pais ($s^0 =$ Gerar Solução (k, B)), $k=2, \dots, m-1$. Cada solução é avaliada e o número de clusters pais das melhores soluções são armazenadas na lista LRC (*Atualizar LRC(s^0)*). Isto está indicado entre as linhas 4 e 6.

Depois, nas linhas 7, 8 e 9, a melhor solução s^* até o momento é atualizada. Este processo iterativo é repetido por $MaxIter$ vezes, como indica a linha 3. O retorno deste procedimento é a solução inicial s^* e a lista *LRC*. A lista *LRC* define o limite inferior (l_i) e o limite superior (l_s) de clusters pais que as soluções do algoritmo podem ter.

Neste trabalho é utilizado um conjunto, denominado *CE* (Conjunto ELITE), que armazena a melhor solução de cada iteração do algoritmo. As soluções armazenadas são diferentes entre si. O *CE* é utilizado no final do algoritmo, para efetuar a busca local *Inversão Individual*, que tem como objetivo tentar melhorar as melhores soluções encontradas ao longo das iterações do algoritmo e armazenadas no *CE*.

A primeira busca local proposta, denominada *Troca Entre Pares*, é uma busca intensiva que troca o status de dois elementos da solução com valores diferentes. Por exemplo, suponha que a solução corrente é $\{10111010\}$. Primeiro é trocado o primeiro e o segundo elemento da solução. Então a nova solução é $\{01111010\}$. Se este novo solução melhorar o valor da função, então ele será aceito e será a nova solução corrente. A próxima troca é feita entre o primeiro e o terceiro elemento da solução corrente. A busca local termina quando todas as trocas entre dois elementos com valores distintos são testadas. O objetivo da busca local *Troca Entre Pares* é investigar possíveis soluções diferentes com o mesmo número de clusters pais.

A idéia básica desta segunda busca local proposta, denominada *Inversão Individual*, é tentar melhorar a solução corrente analisando soluções próximas a ela. Para isso, essa busca permuta o valor de cada elemento da solução (1 por 0, ou 0 por 1), um por vez, e calcula o novo valor da função. Porém, o algoritmo só aceita a mudança, se o novo valor da função for maior que o valor anterior. Por exemplo, imagine que a solução corrente é $\{0101101\}$. Primeiramente, é trocado o primeiro elemento da solução. Então a nova solução é $\{1101101\}$. Se esta solução tem valor da função maior que o anterior, então ele será a nova solução. E então, é trocado o segundo elemento. A nova solução agora é $\{1001101\}$. Se este possuir valor da função maior que o anterior, então a mudança é aceita. Caso contrario, a solução é mantida. A busca acaba quando todos os elementos da solução são testados. A busca local *Inversão Individual* se justifica, pois encontrar o número ideal de clusters é um dos objetivos do problema, e a inclusão ou retirada de um cluster pai pode melhorar uma determinada solução.

O objetivo da Perturbação é alterar o número de clusters pais da solução corrente. Dado uma solução inicial, esta pode ter o número de clusters pais maior ou menor que o ideal. Portanto, pode ser necessário aumentar ou diminuir este número. A perturbação deve aumentar ou diminuir este valor em *uma unidade*, caso o número de clusters pais esteja bem próximo ao ideal, ou ainda, aumentar ou diminuir em *n unidades*, para procurar soluções afastadas da solução inicial. A Perturbação começa incrementando o número de clusters pais em uma unidade. Se a solução obtida após a Perturbação e a busca local for melhor que a anterior, o processo é repetido, ou seja, o número de clusters pais desta solução é incrementado em uma unidade de novo. Caso a solução obtida não melhore a anterior, então a Perturbação aumenta o número de clusters pais em duas unidades. O processo continua enquanto a solução possuir o número de clusters pais dentro do limite superior (l_s) definido pela *LRC*. Depois, o número de clusters pais da solução é decrementado até atingir o limite inferior (l_i) da *LRC*. A Perturbação não permite que um determinado número de clusters pais seja utilizado mais de uma vez, para efetuar a busca local *Troca entre Pares*.

O critério de aceitação utilizado é que a solução gerada é aceita como a solução corrente, se esta melhorar a melhor solução encontrada até o momento.

O algoritmo ILS_PCA gera uma solução inicial e aplica a busca local *Troca entre Pares*. Após cada Perturbação, o algoritmo também aplica a busca local *Troca entre Pares*. No final, o algoritmo aplica a busca local *Inversão Individual*, no conjunto *CE* gerado.

Resultados e Discussão

No Brasil, a sanidade dos apiários tem sido pouco investigada e os apicultores encontram dificuldades para atender seus chamados devido ao despreparo dos técnicos da defesa agropecuária em obter e encaminhar amostras e a falta de laboratórios de patologia apícola. Assim, é incerto o indicador de perdas de colmeias das regiões brasileiras. Em outros países, um patamar acima de 20% é considerado preocupante (Neumann & Carreck, 2010; Steinhauer et al., 2014).

Ao se realizar o diagnóstico apícola do estado do Rio de Janeiro verifica-se alta ocorrência de perdas por doenças de colmeias, dentro de um perfil de incertezas quanto ao seu diagnóstico. Observam-se 33,83% dos apicultores (IC 95%, 28,17%; 39,48%) com relato positivo para doença. Esta cifra pode estar subestimada, ao se analisar as informações sobre as perdas com relatos negativos para doença. Este grupo, que representa 49,36% (IC 95%, 41,51%; 57,20%), apresentaram respostas incertas sobre as fugas dos enxames (Almeida et al., 2013).

Para verificar qual modelo linear Inteiro deve ser utilizado na resolução do problema proposto, foram realizados alguns testes. Tais testes visam verificar o tempo gasto em cada modelo e o comportamento dos modelos à medida que o valor de k cresce. Para isso, foram avaliadas as seguintes instâncias, mostradas na tabela 1: *Ruspini*, *200DATA*, *300p4c1* (Cruz, 2010). As tabelas 2 e 3 mostram estes resultados. Estas tabelas mostram a variação do k , o valor ótimo para cada k e o tempo de execução. Os modelos foram executados utilizando o software XPRESS. Nos testes de contagem de tempo foram utilizados computadores com processadores Intel Xeon Quad core 3.00 Ghz com 16G de memória RAM.

Tabela 1. Descrição das instancias

Nome	Nº pontos	Dimensão
Ruspini	75	R^2
200data	200	R^2
300p4c1	300	R^2

Tabela 2. Testes realizados com o modelo de RAO

Nome	k	ótimo	t(s)
ruspini	K=7	32,984	14163s
	K=6	36,619	1345s
	K=5	40,249	658s
	K=4	47,634	60s
	K=3	88,588	18s
	K=2	102,078	9s
200DATA	K=4	9,257	3688s
	K=3	10,308	344s
	K=2	16,669	243s
300p4c1	K=4	132.966	81356s
	K=3	109,59	1349s
	K=2	365.322	153s

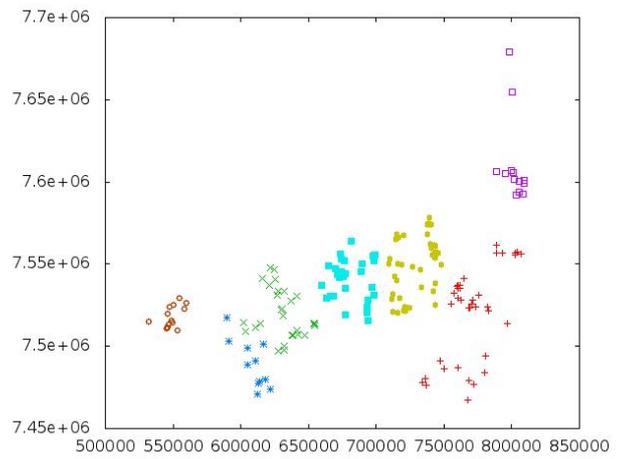
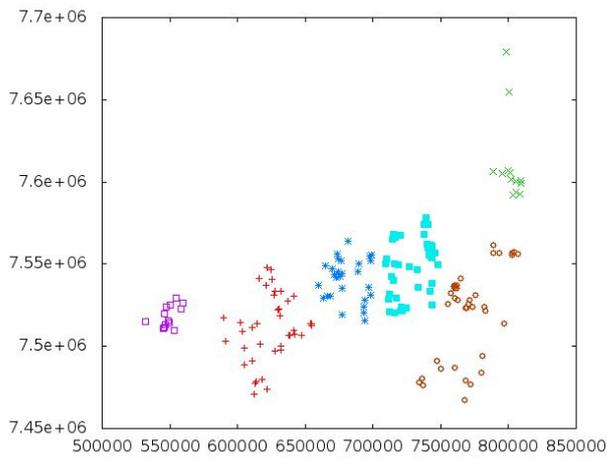
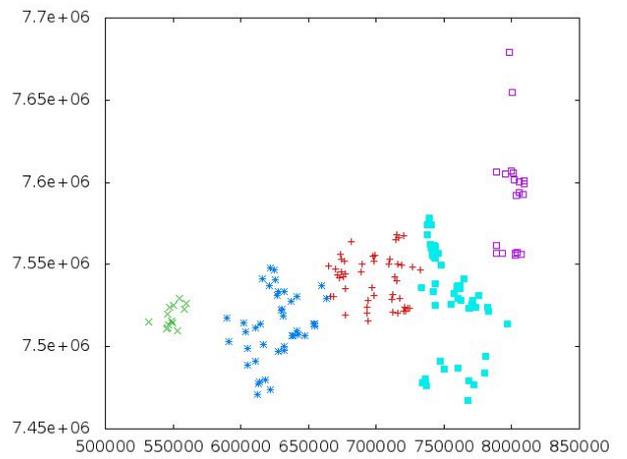
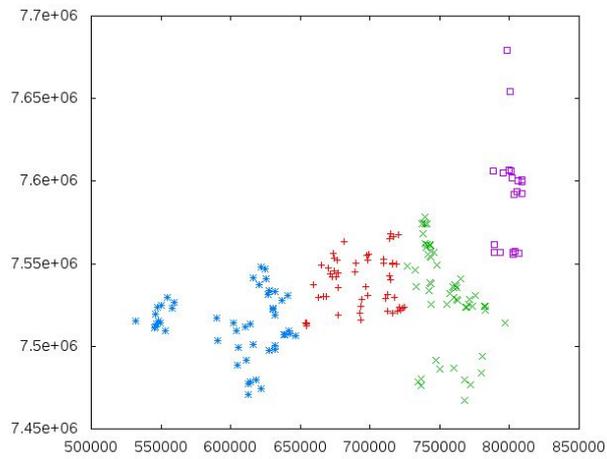
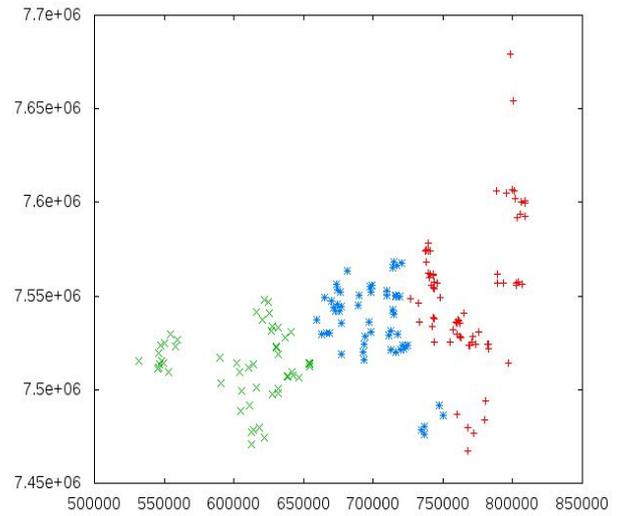
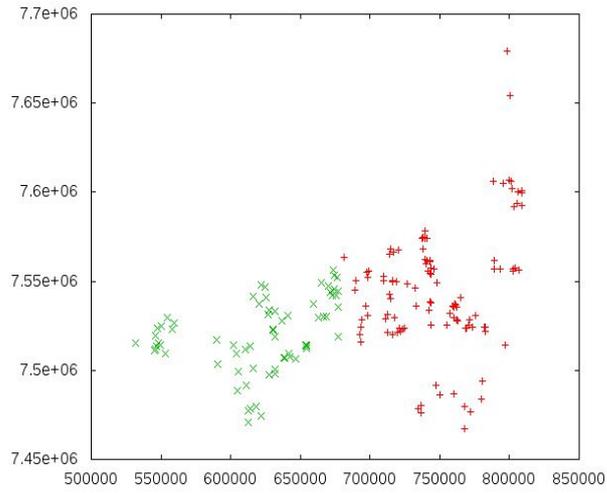
Tabela 3. Testes realizados com o modelo k-medianas

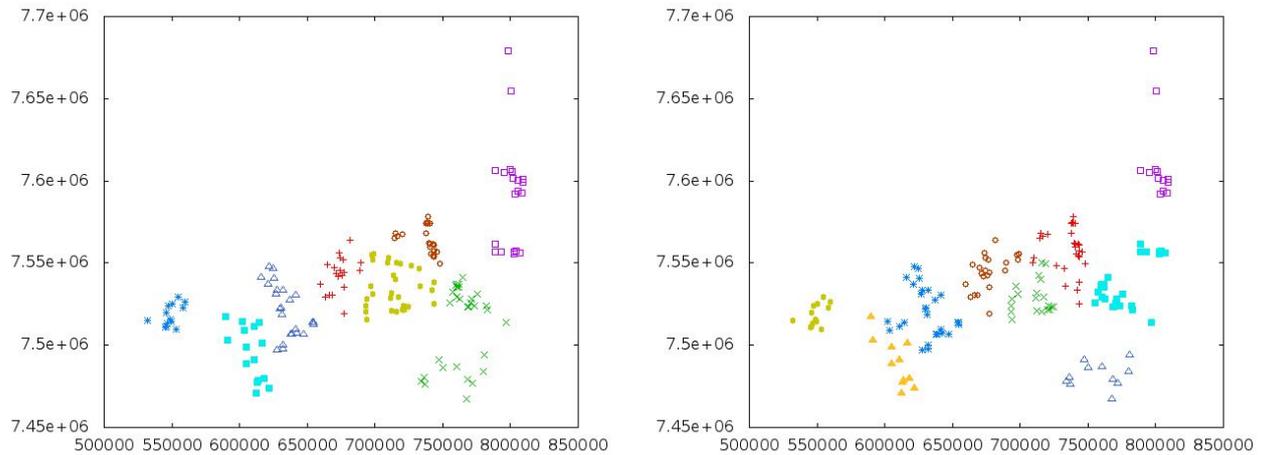
Nome	k	ótimo	t(s)
ruspini	k = 2	2395.8	2s
	k = 3	1619.47	2s

	k = 4	861.47	2s
	k = 5	779.68	2s
	k = 6	714.65	2s
	k = 7	650.84	2s
	k = 8	600.00	2s
	k = 9	555.34	2s
	k=10	512.81	2s
200DATA	k = 2	575.67	4s
	k = 3	337.70	4s
	k = 4	280.21	4s
	k = 5	240.42	3s
	k = 6	217.48	3s
	k = 7	204.00	3s
	k = 8	191.24	3s
	k = 9	179.61	3s
	k =	168.46	3s
300p4c1	k = 2	16063.8	18s
	k = 3	11001.2	18s
	k = 4	8160.05	17s
	k = 5	7344.33	16s
	k = 6	6563.90	16s
	k = 7	5988.01	16s
	k = 8	5579.9	15s
	k = 9	5286.93	15s
	k=10	5019.77	15s
	k=20	3412.00	15s
	k=30	2706.24	15s
	k=40	2274.11	15s
	k=50	1963.32	15s

No modelo de Rao, a partir do k acima dos valores definidos na tabela 3, o software demorou mais de um dia e não convergiu para um resultado, e por isso o software foi interrompido. Portanto, os valores de k da tabela, são os únicos valores possíveis executáveis. Em relação ao modelo k -medianas, os valores de k poderiam ser aumentados, porém não seriam relevantes para a comparação. Pelos testes realizados, verificou-se que o modelo k -medianas é mais rápido e consegue executar instâncias de maior cardinalidade, sendo o escolhido na condução desta pesquisa.

Figura 4: Representação da distribuição espacial dos clusters, com k variando de 2 a 9.

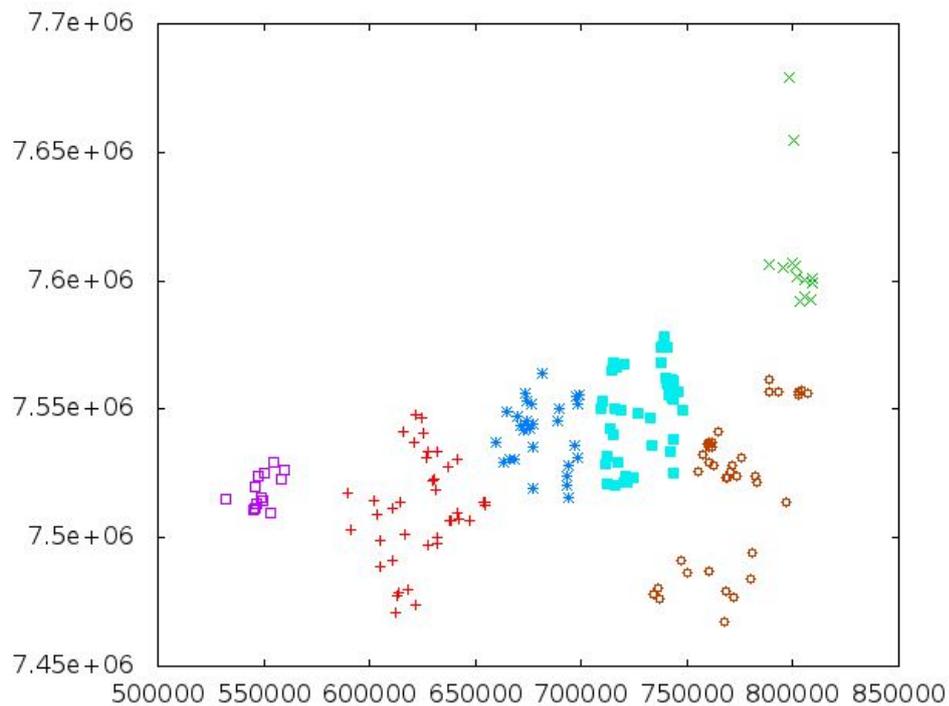




O modelo K-medianas foi executado e o resultado é mostrado na figura 4. O modelo foi executado com vários valores de k, ou seja, os dados foram divididos em grupos variando de 2 a 9.

No caso do algoritmo ILS_PCA, os objetivos do algoritmo são encontrar o número natural de clusters e sua configuração. O ILS_PCA foi executado e o número de clusters encontrado foi 6. Os clusters e a sua configuração são mostrados na figura 5.

Figura 5: Representação da distribuição espacial dos clusters encontrado pelo ILS_PCA.



Analisando os dados gerados na figura 3, outras informações podem ser obtidas. Primeiro, os dados podem ser plotados e associados ao mapa das mesorregiões do Rio de Janeiro, como mostra a figura 3. Neste estudo, os aglomerados foram apresentados segundo um ranqueamento, ou seja, quanto maior a proximidade entre os apiários dentro do cluster estabelecido pelo modelo matemático, maior será sua prioridade por atenção. Tal definição foi estabelecida devido a possibilidade das doenças apícolas serem transmissíveis, já que a

maioria dos apicultores desconhecem os sintomas e há casos de doenças de etiologia duvidosa como, da cria ensacada brasileira (Pacheco et al., 2009). Há ainda fatores que podem favorecer a disseminação de doenças como, a migração de colmeias, a revenda de mel entre regiões e a proximidade entre apiários. Verifica-se que a localização dos aglomerados com indicativo de doença está generalizada, ou seja, ocorre em todas as mesorregiões do estado do Rio de Janeiro. Nas regiões de estudo existem 13 municípios sujeitos à alta prevalência de doenças (acima de 50%): Sapucaia, Mendes, Paracambi, Silva Jardim, Rio Bonito, São José do Vale do Rio Preto, Barra Mansa, Rio Claro, Engenheiro Paulo de Frontin, Casimiro de Abreu, Petrópolis, Araruama e Porciúncula. Esta informação corrobora com Almeida et al. (2013).

Figura 6: Representação da distribuição espacial dos clusters, indicando os possíveis surtos de doenças em apiários no estado do Rio de Janeiro

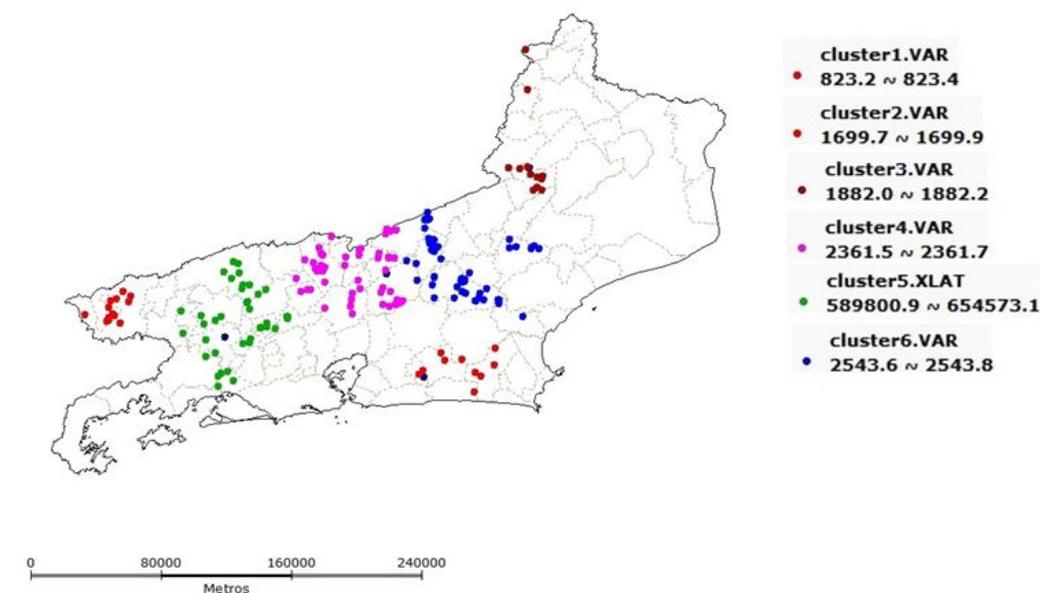


Tabela 4. Identificação dos municípios do estado do Rio de Janeiro e suas respectivas mesorregiões ranqueadas pelos aglomerados surtos de doenças em apiários no período de 2009 a 2010.

Aglomerado	Mesorregiões	Número e municípios mais afetados
1	Sul Fluminense	(2) Itatiaia e Resende
2	Metropolitana e Baixada	(4) Araruama, Rio Bonito, Silva Jardim e Tanguá
3	Noroeste Fluminense	(3) Aperibé, Itaocara e Porciúncula
4	Centro Fluminense e Metropolitana	(9) Com. Levy Gasparian, Paraíba do Sul, Paty dos Alferes, Petrópolis, São José do Vale do Rio Preto, Sapucaia, Teresópolis, Três Rios e Vassouras
5	Sul Fluminense e Metropolitana	(12) Engº Paulo de Frontin, Miguel Pereira, Paracambi, Pinheiral, Piraí, Rio Claro, Barra do Piraí, Barra Mansa, Valença, Vassouras, Volta Redonda e Itaguaí
6	Centro Fluminense	(7) Sapucaia, Casimiro de Abreu, Nova Friburgo, São José do Vale do Rio Preto, Sumidouro, Teresópolis e Trajano de Morais

Tabela 5. Identificação de alguns atributos dos aglomerados detectados com doenças em apiários no período de 2009 a 2010.

Aglomerado	Produção anual mel/colmeia (kg)	Prevalência de apiários doentes (%)	Raio médio do aglomerado (km)	Distâncias médias dos apiários (km)
------------	---------------------------------	-------------------------------------	-------------------------------	-------------------------------------

1	13.723	13,3	17,61	0,82
2	10.662	66,6	27,01	1,70
3	13.317	8,6	46,79	1,88
4	7.105	40,0	42,50	2,36
5	9.407	28,7	40,50	2,47
6	7.175	52,0	11,59	2,54

Na tabela 4 apresentam-se as mesorregiões seguidas dos municípios acometidos pelas doenças em apiários por ordem de prioridade de atenção, sendo o ranqueamento definido segundo a proximidade dos apiários, conforme expostos na tabela 5. Nesta análise verifica-se que as regiões que se deve priorizar atenção, são elas: Metropolitana, Baixada, Sul e Noroeste Fluminense. Tais regiões somam mais que 60% da produção total de mel no estado do Rio de Janeiro. Embora a área do surto não seja tão expressiva, a distância média entre os apiários com doenças é de 1,46 km. A prevalência em si em cada região, não foi tratada como o problema majoritário, mas sim a proximidade geográfica dos apiários que é um fator de grande importância e que é levado em consideração em estudos epidemiológicos (Almeida et al., 2013), haja à vista, que as abelhas migram por abandono e por enxameação (divisão natural) e seu raio de forrageamento médio de 5 km, pode alcançar perto de 25 km, em situação de escassez de alimento.

Conclusão

1. Este trabalho utilizou modelos exatos de Programação Inteira e heurística para a investigação da existência de aglomerados espaciais em apiários, com o intuito de verificar quais são as regiões de maior prioridade para a intervenção na cadeia apícola. O problema pode ser tratado de várias maneiras diferentes. Neste trabalho, utilizou-se duas visões, onde o número de grupos é conhecido a priori e ou não.
2. Os resultados obtidos neste estudo podem auxiliar no estabelecimento de estratégias de prevenção e controle fundamentadas na análise dos fatores associados às doenças no estado do Rio de Janeiro e na prevalência de doenças nos municípios estudados.
3. Outros métodos e algoritmos para detecção de aglomerados devem ser utilizados e difundidos neste contexto, como ferramentas de apoio a tomada de decisões no âmbito da defesa agropecuária.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), pelas bolsas concedidas; aos técnicos da Defesa Agropecuária do Rio de Janeiro pela contribuição na coleta dos dados, a FAPERJ por financiar esse projeto e a UFRRJ por todo apoio institucional oferecido.

Referências Bibliográficas

- ALLARDICE, P; BONE, K; HUTCHISON, F. **Segredos e Virtudes das Plantas Medicinais: Um guia com centenas de plantas nativas e exóticas e seus poderes curativos**. Reader's Digest. 1.ed. 1999.
- ALMEIDA, C.T.; TASSINARI, W.S.; LORENZON M.C.A. Identificação de fatores associados à ocorrência de doenças de abelhas africanizadas (*Apis mellifera* L.) em apiários do estado do Rio de Janeiro. **Revista Brasileira de Medicina Veterinária**, v.35, p. 33-40, 2013.
- BASTOS, L.; OCHI, L.S.; PROTTI, F.; SUBRAMANIAN, A.; MARTINS, I.; PINHEIRO, R. Efficient Algorithms for Cluster Editing. **Journal of Combinatorial Optimization**, v.28 , p.1-26, 2014.
- BAZARAA, M.S ; JARVIS, J.J. E H.D. SHERALI, H.D. Linear Programming and Network Flows. Wiley, 1990.

CRUZ, M. D. **O problema de Clusterização Automática**. Tese de Doutorado. 2010. 120p. Tese (Doutorado) – COPPE – Universidade Federal do Rio de Janeiro, Rio de Janeiro.

CRUZ, M.D.; OCHI, L.S. Um Algoritmo Evolutivo com Memória Adaptativa para o Problema de Clusterização Automática. **Learning and Nonlinear Models**, v.8(4), p.227-239, 2011.

GLOVER, F.; KOCHENBERGER, G. A. Handbook of Metaheuristics. Kluwer Academic Publishers. 2003.

GALLAI, N.; SALLES, J.M.; SETTELE, J.; VAISSIÈRE, B. Economic evaluation of the vulnerability of world agriculture confronted with pollinator decline. **Ecological Economics**, v.68(3), p.810-821, 2008.

KAUFMAN, L.; ROUSSEUM, P. J. Finding Groups in Data : An introduction to cluster Analysis. A Wiley-Interscience publication.1990

KEVAN, P.G.; PHILLIPS, T.P. The economics impacts of pollinator declines: an approach to assessing the consequences. **Conservation Ecology**, v.5(1), p.1-17, 2010.

KNOX, E.G. Spatial and temporal studies in epidemiology. **Oxford Textbook of Public Health: Methods of Public Health**, v.2, p. 95-105, 1991.

LORENZON, M.C.A.; GONÇALVES, E.A.G.B.; PEIXOTO, E.L.T. **Censo Apícola 2006: Análise conjuntural**. Rio de Janeiro: SESCOOP, 2008. 104p.

LORENZON, M.C; TASSINARI, W.S.; KOSHIYAMA, A.S.; ALMEIDA, C.T. **Indicadores e Desafios da Apicultura Fluminense - Um Retrato Brasileiro**. 1. ed. Espírito Santo: ABOVE, v. 1000. 2012, 272p .

NEUMANN, P; CARRECK, N. L. Honey bee colony losses. **Journal of Apicultural Research**, v. 49(1), p.1-6, 2010.

PACHECO, M.R.; BARTH, O.M.; LORENZON, M.C. Tipos polínicos encontrados em colônias de abelhas africanizadas sujeitas à doença cria ensacada brasileira. **Ciência Rural** . v.39(7), p. 2141-214, 2009.

RAO, M.R. Cluster Analysis and Mathematical Programming. **Journal of the American Statistical Association**, v.66, p.622-626, 1971.

STEINHAEUER, N.A.; RENNICH, K; WILSON, E.M.; CARON, D.M; LENGERIC, E.J.; PETTIS, J.S.; ROSE, R.; SKINNER, J.A.; TARPY, D.R.; WILKES, J.T.; VANENGELSDORP, D. A national survey of managed honey bee 2012-2013 annual colony losses in the USA: results from the Bee Informed Partnership. **Journal of Apicultural Research**. V.53(1), p.1-18, 2014.

VINOD, H.D. Integer Programming and Theory of Groups. **Journal of the American Statistical Association**. v.64(326), p. 506-519, 1969.

WERNECK, G.L.; STRUCHINER, C.J. Estudos de agregados de doença no espaço-tempo: conceitos, técnicas e desafios. **Cadernos de Saúde Pública**, v.13(4), p.611-624, 1997.