ISSN 2175-6295 Rio de Janeiro- Brasil, 12 e 13 de agosto de 2010

UM ALGORITMO VNS APLICADO AO PROBLEMA DE DEFINIÇÃO DE ÁREAS DE PONDERAÇÃO

José André de Moura Brito
Escola Nacional de Ciências Estatísticas - ENCE
Rua André Cavalcanti,106 - sala 403 -Santa Teresa, Rio de Janeiro – RJ
e-mail: jose.m.brito@ibge.gov.br

Flávio Marcelo Tavares Montenegro
IBGE – Instituto Brasileiro de Geografia e Estatística - Diretoria de Pesquisas – DPE/COMEQ
Av. Chile, número 500, 10º Andar, Centro – Rio de Janeiro – RJ.
e-mail: flavio.montenegro@ibge.gov.br

RESUMO

É proposto um novo método de resolução para o problema de definição de áreas de ponderação. Tais áreas são utilizadas como domínios de estimação e de calibração no Censo Demográfico do IBGE. Esse problema pode ser mapeado em um problema clássico de agrupamento com restrições de capacidade e de conexidade. Tendo em vista a sua complexidade intrínseca, propõe-se, neste trabalho, a aplicação de um algoritmo que utiliza o conceito de árvore geradora e procedimentos da metaheurística VNS. De forma a avaliar a aplicabilidade e a eficiência desse algoritmo, são apresentados resultados de experimentos computacionais realizados a partir de dados do Censo.

PALAVRAS CHAVE. Agrupamento, Metaheurísticas, Censo.

ABSTRACT

This paper proposes a new method to solve the problem of defining weighting areas. Such areas are used as estimation and calibration domains in the Demographic Census produced by the IBGE (Brazilian Bureau of Statistics). This problem can be mapped into a classical clustering problem with capacity and contiguity restrictions. In view of its inherent complexity, it is proposed in this paper the application of an algorithm that uses the concept of spanning tree and procedures of the VNS metaheuristic. In order to evaluate the applicability and efficiency of such an algorithm, results of computational experiments carried out using census data are presented.

KEYWORDS. Clustering, Metaheuristics, Census. MH.

1. Introdução

O presente trabalho descreve um novo algoritmo de otimização para a resolução do problema de definição de áreas de ponderação (APONDs) (Silva et al., 2000, Censo, 2001). As APONDs são formadas por um agregado de setores censitários, definidos no decorrer de cada censo demográfico.

Por definição, uma área de ponderação deve ser constituída por um grupo de setores que sejam contígüos e cuja soma dos domicílios seja maior ou igual a um total pré-definido. Ademais, tendo em vista que tais áreas são utilizadas como domínios de estimação e de calibração (Censo, 2001), é desejável que os setores que compõe cada uma das APONDs sejam homogêneos entre si, segundo alguma medida de homogeneidade.

A definição das APONDs corresponde à resolução de um problema de agrupamento com restrições de conexidade e de capacidade (Murtagh, 1985, Gordon, 1996). Em geral, tendo em vista a alta complexidade computacional desse tipo de problema, efetua-se sua resolução utilizando-se heurísticas, dissociadas de um procedimento que permita verificar a otimalidade de suas soluções (Batagelj e Ferligoj, 2000).

Neste trabalho, propõe-se um algoritmo que utiliza os conceitos da metaheurística VNS (Hansen e Mladenovic, 2001) (*Variable Neighborhood Search*) e de árvore geradora (AG). Tal algoritmo constrói as soluções, isto é, os agrupamentos de setores censitários, trabalhando em duas fases, quais sejam: Em uma primeira fase, os agrupamentos são produzidos mediante o particionamento de um conjunto de árvores geradoras, considerando, inclusive, a árvore geradora mínima (AGM), obtida a partir do algoritmo de *Kruskal* (Ahuja, 1993). A AGM está associada a um grafo G que é utilizado para representar a relação de contigüidade entre os setores censitários. Na segunda fase, aplica-se o método VNS nos agrupamentos produzidos, de forma a melhorar a qualidade das soluções no que concerne ao critério de homogeneidade definido previamente. Um tratamento diferente, também baseado em uma árvore geradora mínima, mas sem a utilização de uma metaheurística, é apresentado por Assunção et al. (2002).

O desenvolvimento desse trabalho está dividido em 4 seções. Na segunda seção, descreve-se, em detalhes, o problema de definição de APONDs, bem como a sua modelagem. Na terceira seção, são apresentados os conceitos básicos do VNS e os detalhes do algoritmo. Finalmente, a última seção traz um conjunto de resultados computacionais obtidos a partir de experimentos realizados com os dados do censo demográfico de 2000.

2. Definição do Problema

Uma área de ponderação (APOND) é uma unidade geográfica formada por um agrupamento de setores censitários (formados cada um, em média, por 300 domicílios). As APONDs são utilizadas para se estimar informações para a população. O tamanho dessas áreas, em termos de número de domicílios e de população, não pode ser muito reduzido, sob pena de perda de precisão de suas estimativas. As APONDs são definidas considerando esta condição. São, também, os níveis geográficos mais detalhados da base operacional, desenvolvidos como forma de atender às demandas por informações em níveis geográficos menores que os municípios (Silva, 2004, Censo, 2001).

As áreas de ponderação são formadas a partir de k agrupamentos mutuamente exclusivos de setores censitários, observando-se, obrigatoriamente, os critérios de contigüidade e total de domicílios (critérios de viabilidade) e um critério de homogeneidade:

(1) Contigüidade - Os setores agregados em cada uma das APONDs devem ser vizinhos (possuir fronteira em comum) ou deve ser possível sair de um setor A e chegar em um setor B, ambos em uma mesma APOND, passando apenas por setores que também estejam alocados nessa mesma APOND.

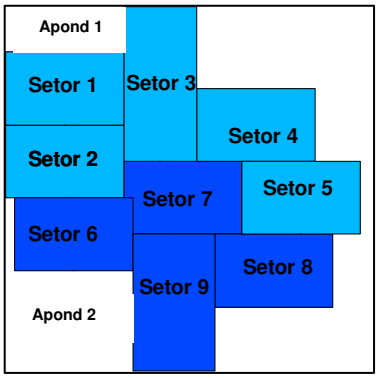

Figura 1 – Exemplo de duas áreas de ponderação definidas pelo agrupamento de setores

(2) Total de Domicílios: A soma dos domicílios associados aos setores que definem cada uma das APONDS deve ser maior ou igual a um total C pré-estabelecido.

Além desses dois critérios de viabilidade, como em qualquer outro problema de agrupamento, torna-se necessária a definição de uma função objetivo para mensurar a qualidade dos grupos formados (critério de homogeneidade).

Inicialmente, considerando um conjunto de p variáveis X_s $(s=1,...,p)$ associadas às características populacionais e de infra-estrutura (Reis, 2002), são calculadas as distâncias d_{ij} entre todos os setores (tomados dois a dois), segundo a equação abaixo:

$$d_{ij} = \sqrt{\sum_{s=1}^{p} (X_s^i - X_s^j)^2} \qquad (1)$$

As distâncias d_{ij} representam o grau de homogeneidade entre as variáveis X_s^i e X_s^j associadas aos setores censitários i e j a serem agregados. Tais distâncias não têm caráter geográfico.

A partir das distâncias calculadas de acordo com a equação (1), define-se uma função objetivo baseada em *medoids*. De acordo com a literatura, os algoritmos que trabalham com o conceito de k-medoids em sua função objetivo são mais robustos à existência de ruídos ou a *outliers* e geralmente produzem agrupamentos de alta qualidade (Kaufman & Rousseeuw, 1989, Han & NG, 2002).

Ademais, os medoids são de uso mais geral para a construção de agrupamentos, sendo aplicáveis em situações onde os objetos que serão agrupados não podem ser representados por atributos quantitativos ou cujas médias não estejam definidas.

Os medoids utilizados no cálculo da função objetivo são definidos da seguinte forma: Dado um conjunto X formado por n objetos ($X = \{o_1, o_2, ..., o_n\}$) com p atributos cada, deve-se selecionar, a partir de X, k objetos que definem um conjunto $M = \{med_1, med_2, ..., med_k\}$ de medoids, de forma a minimizar a soma das distâncias de cada um dos $(n-k)$ objetos restantes ao seu medoid, $med_i \in M, i \in \{1,...,k\}$, mais próximo. Ou seja, procura-se minimizar a soma das distâncias d_{ij} de todos os objetos $o_j \in med_i$, $i = 1,...,k$, aos seus respectivos medoids:

$$\sum_{i=1}^{k} \sum_{\forall o_j \in med_i} d_{ij} \qquad (2)$$

A equação (2) define a função objetivo utilizada para avaliar a qualidade dos agrupamentos formados.

A solução exata (ótimo global) do problema dos k-medoids, com ou sem a restrição de capacidade, pode ser obtida através de uma formulação de programação matemática devida a Kaufman e Rousseeuw (1989). Todavia, mesmo para um número de objetos apenas moderado, a resolução dessa formulação, ou seja, a obtenção do ótimo global, pode levar a um consumo expressivo de tempo computacional, ou até mesmo à não convergência, resultando em uma solução que corresponde a um ótimo apenas local.

Tendo em vista que esse problema pertence a classe de problemas NP-difíceis (Kariv e Hakimi, 1979), encontram-se na literatura alguns algoritmos heurísticos bem conhecidos para o problema dos k-medoids. Tais algoritmos têm uma maior ou menor capacidade de produzir boas soluções (ótimos locais) em um tempo computacional bem pequeno, quando comparado ao tempo consumido pela formulação exata de Kaufman e Rousseeuw ou por métodos exatos.

Um dos métodos heurísticos, que permaneceu como um dos mais populares, é o algoritmo proposto por Kaufman e Rousseeuw no mesmo trabalho (Kaufman e Rousseeuw, 1989) e que foi denominado PAM *(Partitioning Around Medoids*). Basicamente, o algoritmo PAM inicia o processo de formação dos agrupamentos selecionando aleatoriamente k-medoids, dentre os n objetos de uma base de dados. Em seguida, em cada iteração, efetua-se uma troca entre um objeto que corresponde a um medoid e um objeto qualquer, de forma a reduzir o valor da função definida na equação (2).

Kaufman e Rousseeuw (1989), também desenvolveram uma versão modificada do algoritmo PAM, denominada CLARA (*Clustering Large Applications*). Tal algoritmo pode ser aplicado em bases de dados de dimensão elevada, considerando a combinação de uma técnica de amostragem e do algoritmo PAM. Em vez de determinar os k-medoids considerando toda a base de dados, o algoritmo CLARA seleciona m amostras (Kaufman e Rousseeuw sugerem que se trabalhe com pelo menos cinco amostras. Além disso, propõem uma expressão para determinar o número de objetos que se deve selecionar para cada amostra) compostas por n' objetos da base dados ($n' < n$), aplicando, em seguida, o algoritmo PAM em cada uma dessas amostras.

Mais recentemente, Han e Ng (2002) propuseram uma variação do algoritmo CLARA denominado CLARANS. Tal algoritmo utiliza uma técnica de computação mais intensiva para a determinação dos medoids.

Observa-se que todos os algoritmos supracitados consideram, exclusivamente, o critério de homogeneidade. Ou seja, as restrições de capacidade e contigüidade não são levadas em conta.

Entre as várias abordagens encontradas na literatura para a resolução do problema de agrupamento com a restrição de contigüidade, destaca-se o algoritmo AZP (*Automatic Zoning Procedure*) apresentado em Neves (2003) e um algoritmo baseado na utilização da *AGM* (Árvore Geradora Mínima) (Assunção et al, 2002). Tais algoritmos trabalham com uma função objetivo diferente daquela usada para a obtenção dos *k*-medoids.

Os algoritmos acima destacados tendem a convergir para pontos de ótimo local que apresentam, em geral, valores que podem estar distantes do valor no ótimo global (ou seja, o valor no ponto de ótimo, correspondente à melhor solução obtida caso fossem verificadas todas as soluções possíveis para o problema).

Finalmente, considerando apenas a restrição de capacidade e o critério de homogeneidade baseado em *k*-medoids, Brito et al. (2008) implementaram um algoritmo a partir do estudo da metaheurística ILS (Baxter, 1981 e Glover e Kochenberger, 2002).

Com a expectativa de produzir soluções de boa qualidade e, simultaneamente, contemplar as duas restrições do problema de definição de áreas de ponderação e o critério de homogeneidade baseado no conceito de *k*-medoids, propõe-se um novo algoritmo que trabalha em duas fases e que é baseado na metaheurística VNS. Em uma primeira fase, o algoritmo produz uma solução viável considerando a construção e o particionamento de um conjunto de árvores geradoras (inclusive da árvore geradora mínima). Os elementos de uma

partição assim definida correspondem a APONDs iniciais e satisfazem simultaneamente as restrições de capacidade e de contigüidade. Em seguida, visando reduzir o valor da função objetivo, efetuam-se, na segunda fase, várias tentativas de trocas de objetos entre as partições, ou seja, de realocação de setores entre as APONDs definidas na primeira fase.

2.1 Modelagem do Problema

Um primeiro passo para aplicação do algoritmo consiste na definição de uma estrutura matemática capaz de representar o problema, ou seja, agregar tanto as informações de contigüidade e do total de domicílios quanto a relação de homogeneidade entre os setores. Tal estrutura corresponde a um grafo $G = (V, A)$.

Um grafo $G = (V, A)$ consiste de um conjunto $V = \{v_1, ..., v_i, ..., v_j, ..., v_n\}$ de elementos, chamados de vértices, e de uma relação binária A entre eles, que corresponde a um conjunto cujos elementos são chamados arestas. Cada elemento de A é, portanto, um par de vértices $[v_i, v_j]$ pertencentes ao conjunto V. Por definição, $|V|$ e $|A|$ correspondem, respectivamente, ao número de nós e ao número de arestas de G.

Utilizando um grafo G, pode-se associar cada um de seus vértices v_i em V a um setor censitário e associar a cada vértice v_i o valor correspondente ao total de domicílios do setor e também expressar a relação de vizinhança entre dois setores (contigüidade) através de uma aresta $[v_i, v_j] \in A$.

De forma a possibilitar a aplicação do algoritmo VNS, define-se um conjunto maior de arestas $A'' = A \cup A'$ e atribui-se a todas as arestas $[v_i, v_j] \in A''$ as distâncias d_{ij}. Cabe observar que as arestas em A' não expressam a relação de contigüidade entre setores vizinhos, ou seja, apenas têm os valores das distâncias d_{ij} entre os setores não vizinhos. Tais arestas, quando unidas com as arestas de A, produzem um grafo completo.

Exemplificando, considere o grafo $\boldsymbol{G}$ (Figura 2) associado a um conjunto de quatro setores. A existência de contigüidade entre dois setores $\boldsymbol{i}$ e $\boldsymbol{j}$ é, portanto, representada em G através de cada aresta A ligando os vértices v_i e v_j.

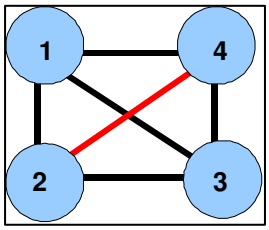

Figura 2 – Grafo representando os quatro setores

Considerando as definições acima, vemos na Figura 2 que os conjuntos A e A' são formados, respectivamente, pelas arestas {[1,2],[1,3],[1,4],[2,3],[3,4]} e {[2,4]}, sendo atribuído a cada uma dessas arestas o valor da distância d_{ij} (calculada através da equação (1)).

3. Algoritmo

Em função da modelagem proposta para o problema, uma boa alternativa para se definir um conjunto de k áreas de ponderação consiste em determinar uma árvore geradora (AG) $T = (V, A^* \subset A)$ associada ao grafo $\boldsymbol{G}$ ($G = (V, A)$), onde T é um grafo que contém todos os vértices de G e $|V|-1$ arestas de A denotadas por A^* (ver Figura 3). A partir de uma árvore geradora T qualquer, pode-se construir k subárvores (subgrafos conexos) a partir da remoção de exatamente $(k-1)$ arestas de T. Assim sendo, a propriedade de conexidade, observada para cada uma das subárvores, possibilita o cumprimento imediato da restrição de contigüidade entre os setores que compõem cada APOND.

A obtenção de uma árvore geradora qualquer é relativamente simples, podendo ser realizada com pouco esforço computacional a partir, por exemplo, de pequenas alterações em um algoritmo de obtenção de árvores geradoras mínimas (AGM) tal como o de *Kruskal* (Ahuja, 1993). Contudo, tal como ocorre para o problema sem a restrição de contigüidade, a obtenção de um ótimo global para o problema contígüo – ou seja, o conjunto de k subárvores que, satisfazendo as restrições de contigüidade, produz o menor valor possível para a função expressa em (2) – tenderia a exigir, também nesse caso, um tempo de processamento excessivo ou mesmo impraticável com o aumento do tamanho (número de setores) das instâncias consideradas no problema, pois o número de árvores geradoras possíveis a serem necessariamente enumeradas tende a crescer rapidamente com esse tamanho.

Alternativamente, pode-se buscar soluções viáveis de boa qualidade, às expensas de um tempo computacional factível, através do uso de uma quantidade limitada de árvores geradoras e da aplicação de um método heurístico de otimização de uso geral, ou seja, uma metaheurística, tal como o VNS.

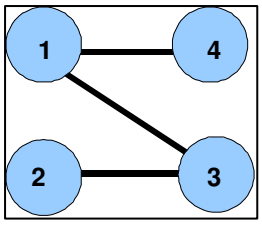

Figura 3 – Árvore Geradora obtida a partir do grafo da Figura 2

3.1 Metaheurística VNS

O enfoque explorado pela metaheurística VNS (*Variable Neighborhood Search*), proposta por Hansen e Mladenovic (2001), está baseado em uma troca sistemática de estruturas de vizinhança associadas a um algoritmo de busca local. De maneira geral, uma vizinhança $V(x_0)$ de uma solução x_0 é um conjunto de soluções que podem ser obtidas a partir de x_0 através da inserção, mudança de posição ou remoção de um elemento de x_0.

Diferentemente de outras metaheurísticas, o VNS não segue uma trajetória, mas explora, uma a uma, vizinhanças previamente definidas e distantes daquela associada à solução x_0, de forma a obter soluções melhores. Ou seja, utiliza-se um conjunto finito de vizinhanças V_s $(s = 1,...,s_{\max})$ previamente selecionadas, sendo $V_s(x_0)$ o conjunto de soluções na *s*-ésima vizinhança de x_0.

O algoritmo VNS básico trabalha considerando a escolha de uma solução $x_0^{'} \in V_s(x_0)$, a qual se submete a uma busca local. Se a busca fracassar na obtenção de uma solução melhor que x_0, incrementa-se a ordem da vizinhança corrente, ou seja, são exploradas as soluções em uma nova vizinhança $V_{s+1}(x_0)$. Caso contrário, se a busca local encontrar uma solução melhor do que a solução x_0 na vizinhança V_s, atualiza-se a solução e a ordem da vizinhança volta a ser 1.

3.2 Algoritmo VNS

O algoritmo apresentado a seguir é composto de duas fases. Na primeira fase (passos de 1 até 3), são construídas q árvores geradoras e cada uma dessas árvores é particionada em $\boldsymbol{k}$ subárvores que satisfaçam à restrição de capacidade. A segunda fase é caracterizada pela aplicação de um procedimento VNS na solução com menor valor de função objetivo (equação 2) obtida na fase anterior. Basicamente, esse procedimento é caracterizado por movimentos de troca de vértices entre as subárvores. Tais movimentos devem preservar as restrições de capacidade e de contigüidade e, ao mesmo tempo, buscar reduzir o valor da função objetivo.

Passo 1: Defina os grafos G e $G*$considerando as informações dos setores censitários. Onde cada grafo $G*$ é construído efetuando-se uma pequena perturbação sobre G. Efetua-se a seleção aleatória de um subconjunto de arestas de G e a substituição das distâncias d_{ij} associadas com essas arestas por valores iguais a $10.d_{\max}$. Dessa forma, produz-se um grafo $G*$ com distâncias d_{ij} **e** $d_{\max}$, sendo $d_{\max}$ a maior distância d_{ij} observada para as arestas de G.

Passo 2: Aplique o algoritmo de *Kruskal* em G e $G*$, para construir, respectivamente, a ***AGM*** e um conjunto formado por q árvores geradoras ***AG*** obtidas a partir dos grafos $G*$

Mediante a aplicação desse procedimento de perturbação, busca-se introduzir uma maior variabilidade para as soluções iniciais construídas, não se restringindo a apenas um ótimo local, possivelmente único, como no caso da ***AGM***. A geração de um conjunto de soluções iniciais possibilita ao método *VNS* (aplicado na 2ª fase do algoritmo) a construção de soluções de melhor qualidade, quando comparadas com uma solução inicial obtida a partir do particionamento de uma única árvore geradora.

Cada árvore geradora (incluindo AGM) corresponde a um ótimo local para o problema, tendo em vista que o valor da função objetivo (2) é calculado em cada subárvore, escolhendo-se em cada uma o nó que corresponde ao *medoid* e calculando-se a soma das distâncias deste aos seus nós restantes. Então, diferentes árvores geradoras produzirão diferentes partições (subárvores), uma vez que as partições dependem das arestas presentes em cada árvore geradora ***T***.

Passo 3: Para cada árvore geradora ***AG*** obtida no passo (2), incluindo a ***AGM***, serão removidas $(k-1)$ arestas de forma a produzir ***k*** partições (subárvores) que satisfaçam à restrição de capacidade. Ou seja, o total de domicílios contidos nos setores associados a cada uma das partições deve ser maior ou igual ao valor mínimo ***C*** pré-estabelecido.

A construção das ***k*** subárvores, ou seja, a definição das áreas de ponderação, consiste na aplicação de um procedimento guloso que pode ser resumido da seguinte forma: Inicialmente, cada ***AG*** corresponde a um único grupo (partição). Então, na primeira iteração $(i=1)$, retira-se uma a uma as arestas da ***AG*** . Cada aresta retirada produz duas novas partições. Avalia-se, entre as duas novas partições, qual é a que tem a maior capacidade residual $Max(C-c_i^1, C-c_i^2)$ e se $c_i^1 - C > 0$ e $c_i^2 - C > 0$. Os termos c_i^1 e c_i^2 estão associados com a soma das capacidades (total de domicílios) dos setores (vértices) que estão em cada uma partições obtidas a partir da ***i***-ésima partição.

Seguindo o mesmo raciocínio, na segunda iteração retira-se cada uma das arestas da partição com maior capacidade residual obtida na 1ª iteração, e assim sucessivamente. Então, para uma iteração ***i*** qualquer, retira-se cada uma das arestas da partição com maior capacidade obtida na iteração ***(i-1)***. O processo de particionamento termina, quando não for possível encontrar uma aresta (entre as restantes) que, ao ser retirada, produza duas novas partições com capacidades maiores do que ***C*.**

Ao final desse passo, são produzidas $q+1$ soluções x_0 viáveis para o problema. Dentre estas, seleciona-se a solução (obtida a partir do particionamento da AGM ou da AG) com menor valor associado à função objetivo dos medoids.

Cabe observar que, em função da restrição de capacidade, algumas ou todas as soluções podem ter um número de grupos (APONDs) inferior ao número *k* previamente estabelecido.

Passo 4: Considerando a melhor solução (conjunto de k subárvores) obtida no passo 3, calcular as distâncias de cada um dos vértices v_j que compõem uma subárvore ao seu respectivo ***medoid (med$_i$) i=1,...,k***. Nesse passo serão utilizadas as distâncias d_{ij} que foram calculadas previamente entre todos os vértices (setores) tomados dois a dois.

Passo 5: Para cada ***medoid***, colocar em ordem crescente as distâncias obtidas no passo 4.
Passo 6: Dividir cada grupo de vértices (subárvore) em quartis, segundo os valores das distâncias ordenadas.
Passo 7: Definir o número de vizinhanças do ***VNS*** igual a quatro.
Passo 8: Associar as vizinhanças V_1, V_2, V_3 e V_4, respectivamente, ao quarto, terceiro, segundo e primeiro quartis definidos no Passo 6.
Passo 9: Sortear um grupo doador dentre os k grupos (subárvores), atribuindo maior chance de seleção aos grupos cujas somas das distâncias dos objetos aos respectivos *medoids* são maiores. Em seguida, selecionar t objetos do grupo doador na vizinhança N_1 e tentar realocá-los a um dos $(k-1)$ grupos receptores, de forma a produzir a maior redução possível na função objetivo. Se houver redução no valor da função, continua-se em N_1, caso contrário, muda-se para N_2 , e assim sucessivamente.
Passo 10: Atualizar os *medoids* e os quartis, repetindo-se os passos 4 a 10 até que uma condição de parada seja satisfeita.

Nos passos de 4 a 8 é definida a vizinhança sobre a qual o método *VNS* será aplicado. A idéia é possibilitar aos objetos mais "distanciados" do seu respectivo *medoid* uma maior chance de mudar de grupo. No passo 9 são efetuadas m tentativas de troca dos objetos entre os grupos, ou seja, de realocações dos vértices a outras subárvores. A finalidade deste passo é produzir soluções $x_0^{'} \in V_s(x_0)$ de qualidade superior a uma solução $\boldsymbol{x}_0$. Após as trocas, os *medoids* e os quartis são recalculados e a busca prossegue. Os passos de 4 a 10 são executados um certo número de vezes, de forma a observar uma redução no valor da função objetivo.

4. Resultados Computacionais

A presente seção traz um conjunto de resultados computacionais obtidos a partir da aplicação do algoritmo VNS. Tal algoritmo foi implementado em linguagem JAVA e foi executado em um computador *DELL-OPTIPLEX* &760 dotado de dois processadores de 3.33 GHz e 4 MB de memória RAM. Os experimentos foram realizados com base nos dados de dezoito municípios selecionados em algumas das unidades da federação, com as áreas de ponderação sendo definidas dentro de cada um desses municípios. Ou seja, foram criados agrupamentos de setores censitários (APONDs) em cada um dos municípios.

As variáveis consideradas para formar os agrupamentos foram as mesmas usadas para definir as áreas de ponderação do censo de demográfico de 2000, quais sejam (Reis, 2002): **(1)** proporção de domicílios particulares permanentes do tipo casa, **(2)** proporção de domicílios particulares permanentes ligados à rede geral de água, **(3)** proporção de domicílios particulares permanentes ligados à rede geral de esgoto ou pluvial, **(4)** proporção de domicílios particulares permanentes ou improvisados com apenas um morador, **(5)** número médio de pessoas por domicílio particular permanente, **(6)** proporção de pessoas com idade entre 0 e 4 anos, **(7)** proporção de pessoas com 65 anos ou mais de idade, **(8)** proporção de pessoas com 15 anos ou mais de idade e que sabem ler e escrever, **(9)** renda média dos responsáveis pelos domicílios, **(10)** proporção de domicílios particulares permanentes com mais de um banheiro, **(11)** proporção de domicílios com o lixo coletado por serviço de limpeza ou colocado em caçamba de serviço de limpeza e **(12)** número médio de moradores por banheiro em domicílios particulares permanentes onde exista ao menos um banheiro.

Em relação ao algoritmo VNS, foram construídas, além da AGM, $\boldsymbol{q}$=10 árvores geradoras e os passos de quatro a dez do algoritmo foram repetidos 800 vezes.

A tabela 1 apresentada a seguir traz os resultados obtidos a partir da aplicação do algoritmo VNS a um conjunto de dezoito instâncias (municípios). A primeira coluna apresenta o nome do município, e nas colunas dois e três tem-se o número de setores e o número de áreas de ponderação que foram construídas mediante a aplicação do algoritmo. Observe-se que foram consideradas instâncias de tamanho variado, com o número de setores (objetos) variando entre 200 e 2500, de forma a avaliar a robustez do algoritmo. Nas colunas

de quatro e cinco, tem-se, respectivamente, o melhor valor da função objetivo após a construção (particionamento das $q+1$ árvores geradoras) e após a aplicação do VNS. Finalmente, a coluna seis apresenta o ***gap*** entre a solução obtida na fase da construção e a solução obtida na fase de busca local (VNS). Os ***gaps*** foram calculados através da seguinte expressão: $100.(Fobj_{Construção} - Fobj_{VNS}) / Fobj_{VNS}$.

Com a aplicação do VNS (passos de 4 até 10) a partir da melhor solução inicial produzida na fase de construção, foi possível obter ganhos – isto é, redução no valor da função objetivo – da ordem de 3,9% a até 12,4%. Os *gaps* médio e mediano foram da ordem de 7,6%, com pouca dependência do tamanho das instâncias utilizadas, o que indica um comportamento razoavelmente robusto do método em relação a esses tamanhos. Os tempos de processamento consumidos pelo algoritmo variaram de 10 segundos (municípios com 200 setores) a até uma hora (municípios com mais de 2000 setores).

Complementando os resultados, são apresentadas nas figuras 4, 5 e 6 as áreas de ponderação que o algoritmo produziu para os municípios de Belo Horizonte, São Gonçalo, Niterói, Vitória, Fortaleza e Florianópolis.

Como continuação desse trabalho, pretende-se implementar novos procedimentos de construção e de busca local baseados no estudo de outras metaheurísticas, tais como GRASP (Resende, 1995) e ILS (Baxter, 1981 e Glover e Kochenberger, 2002), comparando seus resultados com aqueles obtidos pelo algoritmo VNS descrito acima. Também será considerada a utilização de um conjunto maior de instâncias, de forma a possibilitar uma análise mais ampla dos algoritmos.

Tabela 1 – Resultados do Algoritmo VNS

Município	Setores	APONDs	Fobj Construção	Fobj VNS	Gap
Fortaleza	2181	92	5150,5	4757,5	8,3
Belo Horizonte	2561	115	4782,5	4397,5	8,8
Contagem	598	30	1387,6	1324,7	4,8
Vitória	268	18	574,1	530,9	8,1
Duque de Caxias	1061	41	2793,1	2633,1	6,1
Niterói	704	29	1524,2	1426,5	6,9
São Gonçalo	1218	52	3137,9	2906,6	8,0
Volta Redonda	418	14	1153,8	1110,6	3,9
Rio Claro	238	11	552,2	497,9	10,9
Santos	602	29	1207,9	1074,4	12,4
Florianópolis	454	23	1092,6	1028,9	6,2
Porto Alegre	2148	85	4674,6	4267,9	9,5
Belford Roxo	619	26	1827,4	1725,6	5,9
Aracaju	502	20	1006,6	937,4	7,4
Vila Velha	305	22	678,0	627,6	8,0
Maceió	677	39	1589,0	1455,1	9,2
Nova Iguaçu	1108	50	2772,5	2617,7	5,9
Osasco	821	39	1864,5	1751,5	6,5

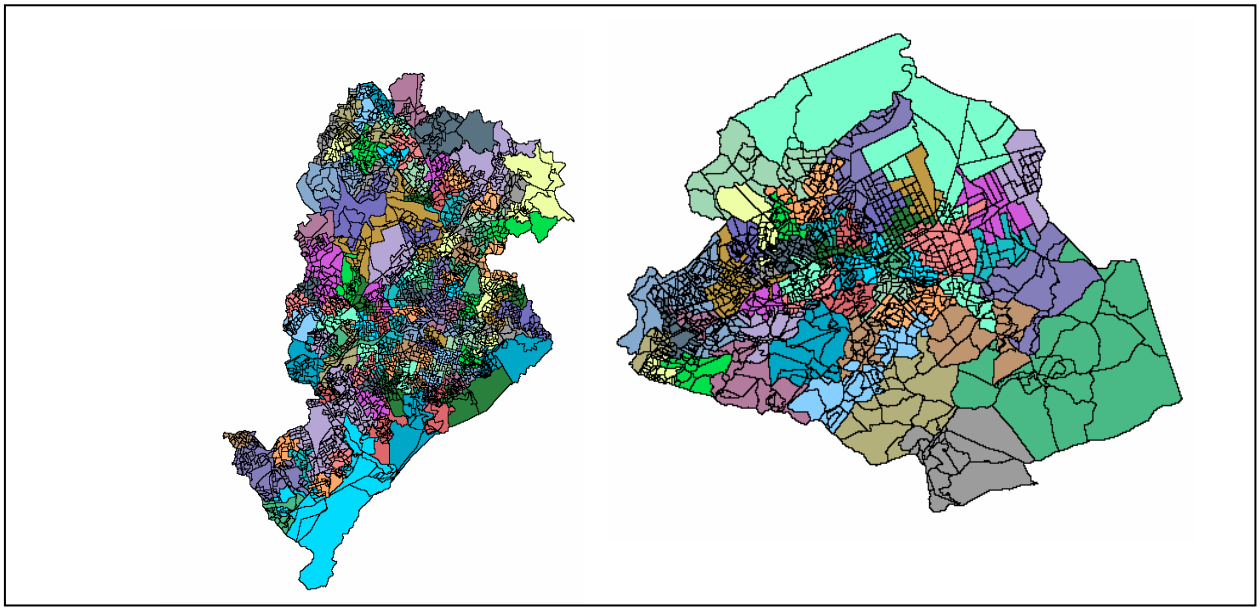

Figura 4 – Áreas de Ponderação dos Municípios de Belo Horizonte e São Gonçalo

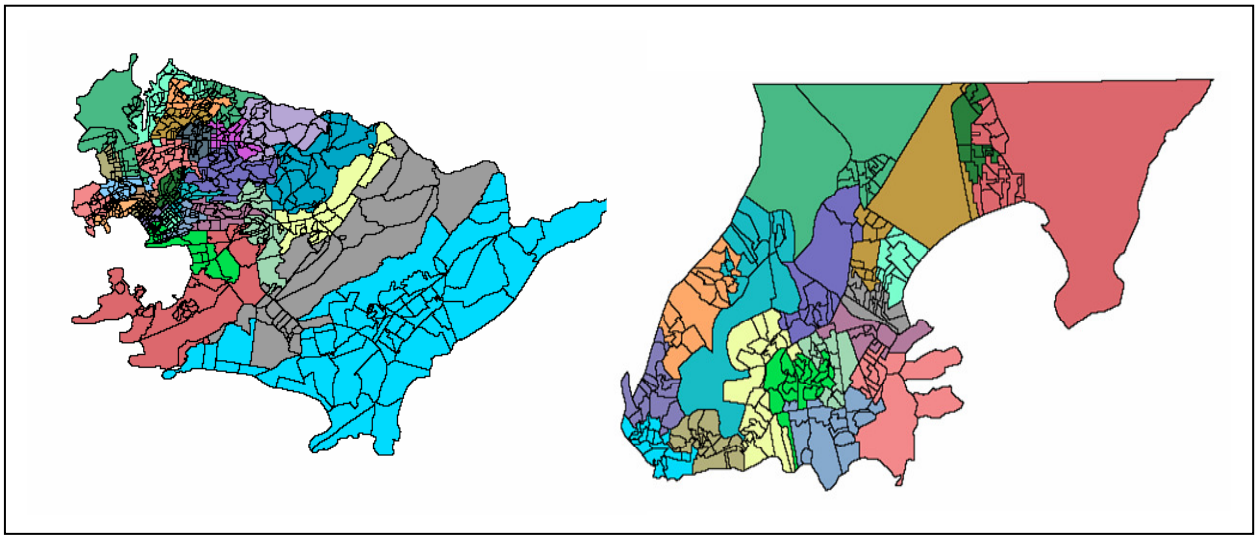

Figura 5 – Áreas de Ponderação dos Municípios de Niterói e Vitória

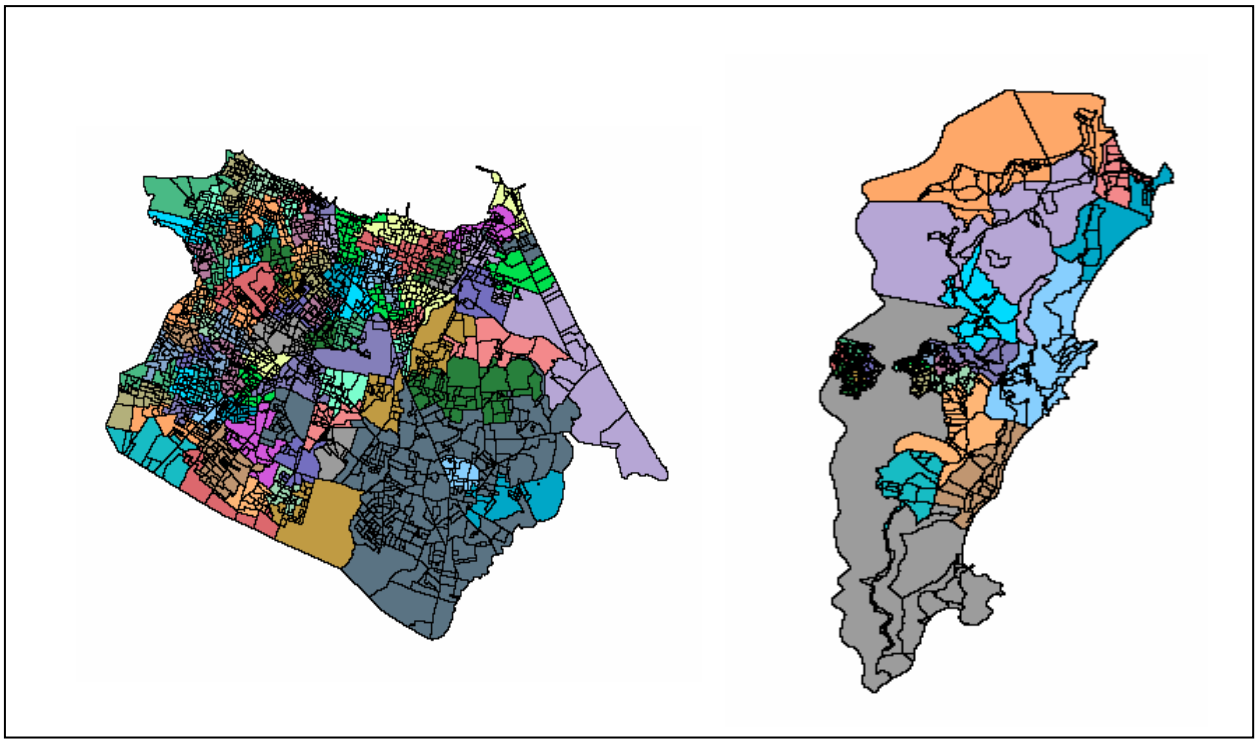

Figura 6 – Áreas de Ponderação dos Municípios de Fortaleza e Florianópolis

5. Bibliografia

Ahuja, **R. K**. (1993). *Network Flows , Theory, Algorithms, and Applications* , Prentice Hall.

Assunção, **R. M**., **Lage**, **J. P., Reis, E. A.** (2004). *Análise de Conglomerados Espaciais Via Árvore Geradora Mínima*, Revista Brasileira de Estatística, vol. 63, n. 220, 7-24.

Batagelj, **V**, **Ferligoj**, **A.** (2000). *Clustering Relational Data*, Data Analysis (ed.: W. Gaul, O. Opitz, M. Schader), Springer, Berlin, 3-15.

Baxter, **J.** (1981). "Local Optima Avoidance in Depot Location", *Journal of the Operational Research Society*, vol. 32, 815-819.

Brito, **J. A. M.**, **Ochi**, **L. S.**, **Montenegro**, F. M. T., **Brito**, L. R. (2008). Algoritmo ILS Aplicado ao Problema das K-Medianas Capacitado. In: Simpósio de Pesquisa Operacional e Logística da Marinha, Rio de Janeiro. Anais do XI SPOLM.

Censo Demográfico 2000 (2001). *Primeiros Resultados da Amostra, Parte I*, IBGE/CDDI.

Glover, **F.**, **Kochenberger**, **G. A.** (2003). *"Handbook of Metaheuristics"*, First Edition Norwell: K1uwer Academic Publishers.

Gordon, **A. D.** (1996). *A Survey of Constrained Classification*, Computational Statistics and Data Analysis, vol. 21, 17-29.

Hansen, **P.** and **Mladenovic**, N.: (2001). Variable Neighborhood Search: Principles and applications, European Journal of Operational Research, vol. 130, n. 3, 449–467.

Han, **J. and NG, R..** (2002). "CLARANS:A Method for Clustering Objects for Spatial Data Mining. IEEE Transactions Knowledge of Data Enginnering,vol. 14, n. 5, 1003-1016.

Johnson A.R. e Wichern D.W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall. Sixth Edition.

Kaufman, **L.** and **Rousseeuw**, P. J. (1989)**.** *Finding Groups in Data – An Introduction to Cluster Analysis*. Wiley-Interscience Publication.

Kariv, O. and Hakimi, L. (1979). An algorithmic approach to network location problems, part ii: The p-medians. SIAM Journal of Applied Mathematics, vol 37, n. 3, 539-560.

Murtagh, **F.** (1985). *A Survey of Algorithms for Contiguity-Constrained Clustering and Related Problems*. The Computer Journal vol. 28, n. 1, 82-88.

Neves,**M.C.** (2003). Procedimentos Eficientes para Regionalização de Unidades Socioeconômicas em Bancos de Dados Geográficos. Tese de Doutorado, INPE, São José dos Campos.

Reis, **A. S.** (2002). *Escolha de variáveis a serem utilizadas na definição das áreas de expansão e de disseminação do Censo Demográfico 2000*, IBGE/DPE/ COMEQ.

Reis, **A. S.** (2002). *Padronização das variáveis a serem usadas na formação das áreas de expansão e de disseminação do Censo Demográfico 2000*, IBGE / DPE / COMEQ.

Resende, M.G.C., Feo, T. A. (1995), Greedy Randomized Adaptative Search Procedures, *Journal of Global Optimization*, vol. 6, 109-133.

Silva, **A. N. Cortez**, **B.F.** e **Matzenbacher**, **L.A.** (2004). *Processamento das Áreas de Expansão e Disseminação da Amostra no Censo Demográfico 2000*, Textos para Discussão, número 17, IBGE/DPE / COMEQ.

Späth, **H.** (1980). Cluster Analysis Algorithms for Data Reduction and Classification of Objetcs. John Wiley & Sons.

Agradecimentos: À FAPERJ pelo financiamento parcial deste estudo.