1

Aplicações de Técnicas de Pesquisa Operacional em Problemas de Agrupamento do IBGE

José André de Moura Brito [*]
Flávio Marcelo Tavares Montenegro[**]

Resumo:

Alguns dos principais problemas de estratificação estatística que aparecem no âmbito do IBGE, como os problemas de definição de áreas de ponderação do censo demográfico, de estratificação por cortes e de estratificação de unidades primárias de amostragem, estão intrinsecamente associados a problemas de agrupamento de alta complexidade computacional. Esta característica resulta, com frequência, no uso de métodos estatísticos e de pesquisa operacional que demandam a realização de experimentos de computação intensiva. Neste trabalho são descritos os problemas supracitados e apresentadas algumas metodologias que podem ser aplicadas a seus equacionamentos. Apresentam-se também algumas reflexões no que refere à dificuldade computacional de resolução destes problemas.

Palavras-Chave: Análise de Agrupamentos; Amostragem; Estratificação; Programação Inteira e Metaheurísticas

[*] Professor da Escola Nacional de Ciências Estatísticas (ENCE/IBGE). Doutor em Engenharia de Sistemas e Computação pela UFRJ e Pós-Doutorado em Otimização na UFF.
[**] Professor da Escola Nacional de Ciências Estatísticas (ENCE/IBGE). Doutor em Engenharia de Sistemas e Computação pela UFRJ.

Introdução

Uma parte significativa das pesquisas realizadas pelos institutos oficiais de estatística considera a adoção de um plano amostral. O levantamento por amostragem possibilita a obtenção de estimativas para parâmetros reais da população, aplicando a pesquisa em apenas um subconjunto dessa população denominado amostra (BOLFARINE; BUSSAB, 2005). Mais especificamente, antes da realização da pesquisa, define-se a população que será investigada, o recorte geográfico, a base de dados (registros) a ser utilizada para seleção da amostra e o esquema de amostragem que será considerado.

Ao aplicar-se um plano amostral, busca-se o equilíbrio entre o orçamento disponível à pesquisa e a necessidade de um bom nível de precisão para as estimativas a serem divulgadas. Por sua vez, esse nível de precisão pode ser alcançado explorando-se de uma forma eficiente as relações de homogeneidade observadas entre elementos da população em estudo, ou seja, mediante a aplicação de uma estratificação estatística (BOLFARINE; BUSSAB, 2005; LOHR, 2010; COCHRAN, 1977).

Neste sentido, como um importante produtor e disseminador de uma grande variedade de informações geográficas e estatísticas oficiais, o IBGE trabalha com diversas pesquisas baseadas em uma amostragem probabilística cujo plano amostral incorpora a estratificação estatística.

Observa-se que a definição desses estratos, de uma forma eficiente, está relacionada com a resolução de problemas de agrupamento (HANSEN; JAMAURD, 1997; MICHAUD, 1997) de difícil solução, sendo essa dificuldade decorrente de uma substancial quantidade de dados que devem ser analisados e agrupados, considerando as restrições particulares da estratificação estatística aplicada em cada pesquisa (BOLFARINE; BUSSAB, 2005; LOHR, 2010; COCHRAN, 1977).

Em geral, a resolução de tais problemas exige a aplicação de métodos estatísticos e de pesquisa operacional mais avançados e/ou voltados à abordagem de características específicas aos problemas considerados, o que demanda, com grande frequência, a realização de experimentos computacionais intensivos. Assim sendo, o auxílio ao estudo e desenvolvimento desses métodos é de grande importância no âmbito do IBGE.

Considerando essas observações, o presente texto se propõe a apresentar uma descrição de alguns dos problemas de estratificação, bem como realizar uma revisão das metodologias que têm sido aplicadas à resolução dos mesmos. Em particular, serão apresentados os seguintes problemas: (i) o problema de definição de áreas de ponderação; (ii) o problema de estratificação por cortes; e (iii) o problema de estratificação de unidades primárias de amostragem.

Inicialmente, com objetivo de facilitar o entendimento desses problemas, bem como estabelecer a sua relação com problemas de agrupamento, a seção dois traz

uma visão geral sobre agrupamentos. A seção três traz uma descrição detalhada dos problemas de amostragem listados acima, bem como das metodologias que podem ser aplicadas ao seu equacionamento. Concluindo o presente trabalho, são feitas algumas considerações finais no que concerne à resolução desses problemas.

O problema de agrupamento

Nos dias atuais, são inúmeras as aplicações reais que podem ser mapeadas em um problema de agrupamento. Essas aplicações, por sua vez, estão associadas aos mais diferentes domínios, quais sejam: Processamento de Imagens, Bioinformática, Mineração de Dados, Estatística, Amostragem, Medicina e Biologia (KAUFMAN; ROUSSEEUW, 1989; ROMESBURG, 2004; TAN et al., 2009).

A tarefa de agrupar os n objetos de uma base de dados associada a uma aplicação real consiste, basicamente, em alocar esses objetos em k grupos, com o objetivo de maximizar a similaridade (homogeneidade) dos objetos de um mesmo grupo (variação *intracluster*) e minimizar a similaridade entre os objetos de grupos distintos (variação *intercluster*), sendo essa medida calculada em função dos p atributos (características) dos n objetos que compõem a base de dados (KAUFMAN; ROUSSEEUW, 1989; TAN et al., 2009; MICHAUD, 1997; JOHNSON; WICHERN, 2002).

Formalmente, o problema de agrupamento (HANSEN; JAMAURD, 1997) pode ser definido da seguinte maneira: Dado um conjunto X formado por n objetos $X=\{x_1,x_2,...,x_i,...,x_n\}$, sendo cada objeto x_i definido em função dos seus p atributos, ou seja, $x_i=(x_{i1},x_{i2},...,x_{ip})$, deve-se alocar esses n objetos em k grupos (ou *clusters*) $G_1,G_2,...,G_i,...,G_k$ de forma que sejam respeitadas as seguintes restrições:

$$(i)\ \bigcup_{i=1}^{k} G_i = X$$
$$(ii)\ G_i \cap G_j = \varnothing, \forall i, j = 1,...,k$$
$$(iii)\ |G_i| > 0, \forall i = 1,...,k$$

A restrição (i) garante que a união dos objetos alocados aos k grupos corresponde ao conjunto X; a restrição (ii) garante que cada objeto será alocado a um único grupo e a restrição (iii) garante que todos os grupos têm pelos menos um objeto, ou seja, não há grupos vazios. Além dessas restrições, os grupos formados devem ser homogêneos internamente (objetos similares) e heterogêneos entre si, considerando alguma medida de similaridade que seja uma função dos p atributos de cada um dos objetos.

Quando o problema contempla essas restrições e o parâmetro k é definido *a priori*, temos um agrupamento de agrupamento clássico (HANSEN; JAMAURD, 1997; NALDI, 2011). Quando k não é definido *a priori*, temos um problema de agrupamento automático, também denominado problema de clusterização automática (CRUZ, 2010; NALDI, 2011; SOARES, 2004).

A determinação do agrupamento ótimo para muitas aplicações reais é muito difícil. Entende-se por agrupamento ótimo a partição de X que produz os *clusters* mais homogêneos, segundo algum critério de homogeneidade (medida de distância). A dificuldade decorre de o número de partições (soluções) possíveis ser fortemente impactado pelo número de objetos da base de dados associada à aplicação. Mais especificamente, o número de soluções possíveis para o problema de agrupamento clássico está associado ao número de *Stirling* de segundo tipo (JOHNSON; WICHERN, 2002), dado pela seguinte equação:

(1)
$$\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{n}$$

Supondo, por exemplo, que $n = 20$ objetos que serão alocados em dois *clusters* ($k=2$), o número de soluções a serem consideradas é de 524 287. Mantendo-se o mesmo número de *clusters* e apenas dobrando o número de objetos, teremos 549 755 813 887 soluções possíveis. Ao considerar um número n maior de objetos, esses valores crescem exponencialmente. Este fato torna limitada a resolução do problema mediante a aplicação de um método de enumeração exaustiva (WOLSEY; NEMHAUSER, 1999).

Não obstante, abrindo-se mão do ótimo global, é possível produzir soluções viáveis de qualidade razoável às expensas de um tempo computacional factível. Essas soluções podem ser produzidas mediante a aplicação de um método hierárquico ou de um método não hierárquico (HAIR et al., 2009; MINGOTI, 2007). Tais métodos encontram soluções (*clusters*) de boa qualidade (razoáveis) no que diz respeito à homogeneidade dos grupos, sem examinar todas as soluções possíveis.

Os métodos hierárquicos, por sua vez, dividem-se em aglomerativos e divisivos. Em um método aglomerativo, inicialmente há n grupos de um objeto cada, sendo efetuada uma série de uniões até que sejam obtidos k grupos. Já no método divisivo, inicialmente há um único grupo formado por n objetos, sendo efetuadas sucessivas divisões dos grupos até que sejam atingidos k grupos.

Existem vários métodos hierárquicos que diferem somente pela escolha do critério de partição. Uma desvantagem desses métodos é a possibilidade de tornarem-se impraticáveis para grandes conjuntos de dados, devido à alta complexidade com-

putacional do problema (HANSEN; JAMAURD, 1997). Os métodos não hierárquicos procuram encontrar uma partição viável dos n objetos sem a necessidade de associações hierárquicas. Primeiramente, uma partição inicial com um determinado número k de *clusters* deve ser considerada. A seguir, seleciona-se uma partição dos n objetos em k grupos, otimizada segundo algum critério de homogeneidade. Dentre os métodos não hierárquicos disponíveis na literatura, dois dos mais conhecidos são o método k-means (JOHNSON; WICHERN, 2002; MINGOTI, 2007; JAIN, 2010) e o método dos k-medoids (KAUFMAN; ROUSSEEUW, 1989).

Em decorrência das limitações desses métodos e das características dos problemas reais, há que se buscar, em muitas situações, o desenvolvimento de métodos específicos (baseados em heurísticas e metaheurísticas) de agrupamento para a resolução dos mesmos. Essa última observação está em consonância com o que será exposto na seção seguinte, ou seja, as três aplicações reais do IBGE apresentadas, a seguir, podem ser resolvidas mediante a aplicação de tais métodos específicos.

Problemas de agrupamento do IBGE

A presente seção traz uma descrição detalhada de alguns problemas do IBGE que estão intrinsecamente associados com um problema agrupamento e comenta, de forma sucinta, algumas metodologias que podem ser aplicadas à resolução destes.

O problema de definição de áreas de ponderação

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 et al., 2004; CENSO, 2010).

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

1. **Contiguidade -** 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 a um setor B, ambos em uma mesma Apond, passando apenas por setores que também estejam alocados nessa mesma Apond.

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 preestabelecido.

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 $X_s\ (s=1,\ldots,\mathrm{p})$ associadas às características populacionais e de infraestrutura (REIS, 2002), são calculadas as distâncias d_{ij} entre todos os setores (tomados dois a dois), segundo a equação abaixo:

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

As distâncias d_{ij} representam o grau de homogeneidade entre as variáveis X_s^i e X_s^j que são associadas aos setores censitários i e j a serem agrupados. Considerando essas distâncias, a função objetivo pode ser a função objetivo utilizada no problema das k-médias (MINGOTI, 2007) ou dos k-medoids (KAUFMAN; ROUSSEEUW, 1989).

A partir da descrição acima, observa-se que o problema de definição de áreas de ponderação corresponde a um problema de agrupamento com restrições de conexidade (contiguidade das Aponds) e de capacidade (Total de domicílios).

A literatura de análise de agrupamentos (*cluster analysis*) disponibiliza vários trabalhos que tratam de problemas de agrupamento capacitados e de problemas de agrupamento com a restrição de conexidade: (MURTAGH, 1985; GORDON, 1996; SOSA, 1996; FURTADO, 1998; SHIEH; MAY, 2001; LIMA; COSTA; OCHI, 2003; DIAS, 2004; AHMADI; OSMAN, 2005; VIEIRA, 2006; DUTRA, 2008).

Não obstante, considerando as duas restrições do problema de definição de Aponds, encontram-se poucos trabalhos na literatura, sejam eles: (ASSUNÇÃO et al., 2002, BRITO; MONTENEGRO, 2010; BRITO et al., 2012, BRITO et al., 2004; SEMAAN et al., 2013; SEMAAN et al., 2009; SEMAAN et al., 2008).

O trabalho de Assunção et al. (2002) traz a proposta de uma heurística que efetua o particionamento de uma árvore geradora mínima (AHUJA, 1993), de forma

a produzir o conjunto de Aponds mais homogêneas e que satisfaçam as restrições de capacidade e de contiguidade. Nos trabalhos de Brito e Montenegro (2010) e Brito et al. (2011 e 2012), foi implementado um algoritmo que realiza a construção de um conjunto de árvores geradoras (AGs) e procede a aplicação da metaheurística VNS (HANSEN; MIADENOVIC, 2001) sobre essas árvores, de forma a produzir uma solução de boa qualidade, ou seja, Aponds homogêneas.

E em 2004, Brito et al. (2004) propuseram uma formulação de programação inteira que efetua o particionamento de uma árvore geradora. E, finalmente, nos trabalhos de Semaan et al. (2011, 2009 e 2008) foram propostos algoritmos heurísticos e metaheurísticas que trabalham com particionamento de grafos e de árvores geradoras.

O problema de estratificação por cortes

Suponha que seja definida uma população de pesquisa identificada por um conjunto P formado por todas as N unidades da população tal que $P=\{1,2,3,...,i,...N\}$. Em seguida, definindo-se uma variável Y de interesse na pesquisa, para a qual será calculada uma estimativa, a população é dividida em um número prefixado de L estratos, denotados por $E_1, E_2,..., E_L$. Considera-se, também, uma variável de tamanho X (AZEVEDO, 2004) que é usada para a estratificação e tem o valor conhecido para cada unidade da população.

Seja $Y_P = \{y_1, y_2,..., y_N\}$ um vetor populacional associado à variável Y e $X_P = \{x_1, x_2,..., x_N\}$ o vetor populacional gerado pela variável auxiliar X, tal que, sem perda de generalidade, se supõe que $x_1 \leq x_2 \leq ... \leq x_N$. As observações populacionais do vetor X_P são distribuídas em L estratos denotados por $E_1, E_2,..., E_h,..., E_L$, sendo tais estratos construídos em função de $L-1$ pontos de corte $b_1 < b_2 < ... < b_h < ... < b_{L-1}$:

$$E_1 = \{i : x_i \leq b_1\}, E_h = \{i : b_{h-1} < x_i \leq b_h\}; h = 2,3..., L-1, E_L = \{i : b_{L-1} < x_i\}$$

Após a construção dos estratos seleciona-se de cada um deles uma amostra aleatória simples de tamanho $n_h, h = 1,..., L$. A amostragem aleatória simples (COCHRAN, 1977) é o método mais simples e mais importante para seleção de uma amostra.

Ele pode ser caracterizado através da seguinte definição operacional: "De uma lista com N unidades elementares, sorteiam-se com igual probabilidade n unidades".

A partir de tais considerações, a resolução do problema de estratificação consistirá em determinar os limites (pontos de corte) $b_1 < b_2 < ... < b_h < ... < b_{L-1}$ de forma a minimizar a variância da variável Y,

(3) $$V_Y = \sum_{h=1}^{L} N_h^2 \frac{S_{yh}^2}{n_h}.\left(1 - \frac{n_h}{N_h}\right)$$

Deve-se observar que os valores de N_h e S^2_{yh} são definidos em função dos limites dos estratos. Todavia, o valor da variância (equação 3) também dependerá do critério adotado para definir o tamanho de amostra n_h alocado em cada um dos estratos, a partir do tamanho da amostra n. Para efetuar tal alocação, pode-se utilizar uma das seguintes expressões a seguir (LOHR, 2010):

(4) $$n_h = \frac{n}{L}$$

(5) $$n_h = \frac{n.N_h}{N}$$

(6) $$n_h = \frac{n.N_h.S_{hy}}{\sum_{h=1}^{L} N_h.S_{hy}}$$

A expressão (4) está associada com a alocação *Uniforme*, que considera a alocação de um mesmo tamanho de amostra para cada estrato. É o esquema de alocação indicado quando se pretende apresentar estimativas separadas para cada estrato. A expressão (5) está associada com a alocação *Proporcional*. Neste caso, a amostra de tamanho n é distribuída proporcionalmente ao tamanho dos estratos, o que corresponde a uma amostra autoponderada, normalmente utilizada quando se deseja fazer numerosas estimativas. Finalmente, a expressão (6) está associada com a alocação de *Neyman (*COCHRAN, 1977).

Neste caso, o número de unidades da amostra a serem observadas no estrato h é proporcional a $N_h \cdot S_{hy}$. Em geral, os tamanhos de amostra obtidos a partir da alocação de *Neyman* produzem uma maior redução do valor da variância (equação 3).

Um fato comum em amostragem é o de substituir-se Y por X na expressão de variância, levando-se em conta a correlação entre as variáveis. Dessa forma, tanto os pontos de corte quanto a expressão da variância serão calculados em função de X. Muitos autores fazem essa mesma substituição, entre os quais, os trabalhos de: Dalenius and Hodges (1959), Ekman (1959), Lavallée and Hidiroglou (1988), Hedlin (1998, 2000).

Uma vez efetuada essa substituição, deve-se minimizar a seguinte expressão de variância:

(7) $$V_X = \sum_{h=1}^{L} N_h^2 \frac{S^2_{xh}}{n_h}.(1 - \frac{n_h}{N_h})$$

A obtenção do mínimo global para a variância definida em (3) ou (7), aplicando um dos esquemas de alocação mencionados anteriormente, corresponde a um problema de difícil resolução tanto analítica quanto computacional, pois S_{xh}^2 é uma função não linear dos valores $b_1, b_2, ..., b_{L-1}$ e o número de possibilidades diferentes de escolha desses valores (para um dado $L>1$ e ao menos duas observações em cada estrato) é, no mínimo, igual ao número de combinações de ($[N/2]$-1) tomados (L-1) a (L-1): $C_{L-1}^{[N/2]-1}$, ou seja, é da ordem de $\Omega(N^{L-1})$.

Observa-se que a alocação de Neyman raramente produz os tamanhos de amostra inteiros, o que implica, por sua vez, apenas em uma solução que é um ótimo local. Ademais, para algumas populações, a aplicação dessa alocação pode produzir tamanhos de amostra maiores que os tamanhos populacionais ($n_h > N_h$). Isso implica, por sua vez, em adotar-se um procedimento que efetue a redistribuição do tamanho de amostra excedente para outros estratos onde $n_h < N_h$, sendo mais uma vez comprometida a questão da otimalidade.

Em decorrência da dificuldade de determinar os limites nos estratos, vários algoritmos heurísticos foram propostos nas últimas décadas. Um algoritmo bem conhecido e antigo foi proposto por Dalenius and Hodges (1959). Esse algoritmo aproxima a distribuição da variável de estratificação X usando um histograma com várias classes, adotando a hipótese de que a variável de estratificação é uniformemente distribuída (LOHR, 2010) dentro de cada classe. Com isto, o problema tem uma solução simples com a aplicação da Regra da Distribuição Cumulativa da Raiz da Frequência, ou regra de Dalenius-Hodges, cuja descrição pode ser encontrada em Cochran (1977) (capítulo 5).

O algoritmo proposto por Hedlin (1998, 2000) está associado com a regra estendida de Ekman (1959), sendo, por esta razão, também chamado de algoritmo de Hedlin alterado. É o primeiro algoritmo a tratar simultaneamente do problema de delimitação dos estratos e de alocação da amostra.

Considerando um nível de precisão prefixado, Lavallée e Hidiroglou (1988) propuseram um algoritmo que constrói os estratos considerando a minimização de uma expressão associada com o tamanho de amostra n que será alocado aos L estratos. Esse estudo também se diferencia dos demais por considerar a alocação potência (AZEVEDO, 2004).

Gunning e Horgan (2004) propuseram um algoritmo, denominado Geométrico, muito simples e prático para a definição dos limites dos estratos. Eles verificaram que para distribuições assimétricas os coeficientes de variação poderiam ser aproximadamente iguais entre os estratos, desde que os limites dos estratos formassem uma progressão geométrica e que a variável de estratificação tivesse uma distribuição aproximadamente uniforme. Kozak (2004) apresentou um algoritmo de estratificação

denominado *Random Search*, que tem certa similaridade com a metaheurística *VNS* (GLOVER; KOCHENBERGER, 2002).

Khan et al. (2008) desenvolveram um algoritmo baseado em programação dinâmica para determinar os limites dos estratos, sendo esse algoritmo aplicado apenas quando X tem uma distribuição normal ou triangular (MEYER, 2009) e a amostragem é feita com reposição (BOLFARINE; BUSSAB, 2005).

Acrescentam-se a tais algoritmos heurísticos mais algumas abordagens baseadas em metaheurísticas, quais sejam: Keskinturk and Sebnem (2007), que propuseram um algoritmo baseado na metaheurística algoritmos genéticos (SIVANANDAM, 2008) para determinar simultaneamente os limites dos estratos e a alocação da amostra, considerando quatro possíveis esquemas de alocação. Além disso, a concepção desse algoritmo permite o número de estratos varie, caracterizando um problema de clusterização automática (CRUZ, 2010). Brito et al. (2007) também desenvolveram um algoritmo genético que determina os limites dos estratos e que utiliza a alocação de *Neyman*. Em um trabalho mais recente, Brito et al. (2010) propuseram um algoritmo que utiliza os conceitos da metaheurística ILS e de *Path Relinking* (GLOVER; KOCHENBERGER, 2002) e, em 2011, Brito et al. (2011) propuseram um algoritmo de estratificação baseado na metaheurística GRASP (FEO; RESENDE, 1995) que produz os pontos de corte dos estratos. Esse algoritmo incorpora um método de programação inteira (BRITO, 2005) que determina os tamanhos de amostra inteiros que serão alocados aos estratos.

O problema de estratificação de unidades primárias de amostragem

Como outro exemplo de aplicação da estratificação estatística, temos o novo sistema de pesquisas domiciliares por amostragem para integração de algumas pesquisas do IBGE, considerando o uso de um mesmo cadastro de seleção e de uma amostra em comum, denominada Amostra Mestra. Segundo Freitas et al. (2007), essa amostra corresponde a um conjunto de unidades de área selecionadas de um cadastro, segundo um método probabilístico de seleção, a partir do qual seja possível selecionar subamostras para atender às diversas pesquisas. Essas subamostras podem ser selecionadas de forma independente ou com certo controle para que tenham ou não algumas unidades coincidentes.

A população-alvo da Amostra Mestra inclui toda a população a ser investigada em todas as pesquisas, sendo constituída pelos moradores residentes em todos os domicílios na área que constitui a abrangência geográfica. A abrangência geográfica da Amostra Mestra considera o âmbito das diversas pesquisas que farão uso dessa

amostra comum. Assim, não se pode deixar de incluir qualquer parte do território que seja contemplado por alguma das pesquisas.

Portanto, a abrangência geográfica da Amostra Mestra é constituída pelos setores censitários da Base Operacional Geográfica de 2010 de todo o Território Nacional.

Um importante aspecto para seleção de uma Amostra Mestra é a definição de suas unidades primárias de amostragem (UPAs). Em pesquisas domiciliares tais unidades são definidas, em geral, por unidades de área com um determinado tamanho mínimo populacional contabilizado em termos de domicílios ou pessoas. As UPAs podem ser definidas como sendo as unidades básicas do Cadastro Mestre, ou podem corresponder a agregações contíguas destas, como, por exemplo, uma divisão administrativa. Como na Base Operacional Geográfica de 2010 há muitos setores censitários pequenos, foi preciso realizar uma agregação de setores censitários para a composição das UPAs, de tal modo que estas possuíssem um número de domicílios suficientes para atender a demanda das pesquisas a serem integradas (reunidas no chamado Sistema Integrado de Pesquisas Domiciliares - SIPD, do IBGE).

Após a avaliação de alguns fatores, concluiu-se que as UPAs deveriam ter no mínimo 60 domicílios particulares permanentes (dpps), incluindo os ocupados, os ocupados sem entrevista realizada e os vagos, de acordo com o Censo Demográfico 2010. Aplicando-se um algoritmo de construção, a agregação foi feita com o objetivo de maximizar o número de grupos, juntando os setores o mínimo possível e tendo como restrições: a contiguidade, o tamanho mínimo e algumas características dos setores, quais sejam: tipo, situação e divisão administrativa (subdistrito).

Do total de 316 574 setores censitários da Base Operacional Geográfica de 2010, 312 090 fazem parte do âmbito da Amostra Mestra, sendo estes agrupados em 296 762 UPAs.

Os setores dentro de cada um dos Municípios foram alocados aos grupos (UPAs) respeitando as restrições de contiguidade, de mínimo de 60 dpps por grupo, a situação do setor e o tipo de setor. Foram gerados seis tipos de "estratificação", quais sejam: (1) TipoSitu x Subdistrito, (2) TipoSitu x Distrito, TipoSitu (situação do domicílio) x Município, TipoSituacao (situação do setor) x Subdistrito, TipoSituacao x Distrito e TipoSituacao x Município. Nesta fase, a melhor estratificação correspondeu ao maior número de grupos e ao maior percentual de grupos viáveis no que concerne às restrições de contiguidade e de mínimo de 60 dpps. A última etapa de estratificação foi a estratificação estatística, na qual as UPAs foram classificadas em grupos homogêneos segundo a renda total dos domicílios e o total de domicílios particulares permanentes (dpps). Nesta etapa, foram construídos de dois até cinco estratos, considerando um número mínimo de 150 UPAS por estrato. Os grupos de UPAs foram estratificados de forma a minimizar a variância do estimador de total da renda domiciliar, considerando o plano amostral normalmente utilizado nas pes-

quisas domiciliares: amostragem conglomerada (COCHRAN, 1977) com seleção de UPAs com probabilidade proporcional a uma medida de tamanho (número de dpps). Considerando esse desenho amostral, busca-se minimizar a seguinte expressão de variância dentro de cada um dos estratos E_h $(h=1,...,L)$:

$$V_h = \sum_{\forall i,j \in E_h} d_j = N_i.N_j\left(\frac{Y_i}{N_i} - \frac{Y_j}{N_j}\right)^2,\ h=1,...,L \quad (8)$$

sendo N_i = número de dpps na i-ésima UPA, N_j = número de dpps na j-ésima UPA, Y_i = renda total domiciliar na i-ésima UPA e Y_j = renda total domiciliar na j-ésima UPA.

Em função desta descrição, observa-se que os estratos estatísticos podem ser definidos mediante a resolução de um problema de agrupamento capacitado. Em particular, no presente problema, a restrição de capacidade está associada ao número mínimo de UPAs por estrato (150) e a expressão de variância (equação 8) corresponderá à função objetivo do problema de agrupamento.

Mais especificamente, o presente problema está associado a um conhecido problema da literatura, denominado problema de clique de soma mínima (HANSEN; JAUMARD, 1997, BASTOS, 2012), sendo esse problema classificado como NP-completo (LEISERSON et al., 2012).

Essa característica restringe a aplicação de algoritmos baseados em métodos de enumeração exaustiva, além de indicar que não existe nenhum algoritmo que produza a solução ótima em tempo polinomial (LEISERSON et al., 2012).

Ainda que seja utilizado um método de enumeração implícita (WOLSEY; NEMHAUSER, 1999), associado a uma formulação de programação inteira como a apresentada em (HANSEN; JAUMARD, 1997), a execução do mesmo poderia consumir um tempo computacional de dias, meses, anos ou até Séculos, até que fosse produzida uma solução viável, não necessariamente o ótimo global.

No que diz respeito ao problema de clique de soma mínima, listamos a seguir algumas referências bibliográficas concernentes às metodologias baseadas em metaheurísticas e em formulações de programa inteiras, como as propostas nos trabalhos de: Brito et al. (2008); Serpa et al. (2009); Nascimento et. al (2010); Dorndorf; Pesch (1994); e Marcotorchiuo; Michaud (1979).

Todavia, nenhuma dessas metodologias contempla a restrição de capacidade, ou seja, não é possível garantir o número mínimo de UPAs por estrato. Em virtude dessa observação e da necessidade de se produzir soluções de alta qualidade, foram desenvolvidos dois algoritmos que consideram a minimização da expressão de variância (equação 8) e contemplam essa restrição de capacidade.

Mais especificamente, Brito et al. (2011) propuseram dois algoritmos baseados, respectivamente, nas metaheurísticas Otimização Microcanônica (MONTENEGRO et al., 2003) e ILS (*Iterated Local Search*) (GLOVER et al., 2002).

Foram processados 72 arquivos contendo, cada um, uma lista de UPAs com seus respectivos números de domicílios e rendas totais. Estes arquivos correspondem a um subconjunto de todos os arquivos da amostra mestra. O menor arquivo processado continha 302 UPAs e o maior 3314 UPAs. Além disso, o número de estratos construídos variou entre dois e cinco, considerando o mínimo de 150 UPAs por estrato.

Uma vez aplicados os dois algoritmos, foi realizada uma avaliação da qualidade dos estratos produzidos, considerando o valor da expressão de variância da equação. A partir desta avaliação, observou-se que o algoritmo ILS produziu soluções ligeiramente melhores para 41 dos 72 arquivos, quando comparado ao Microcanônico, que produziu soluções também ligeiramente melhores que o ILS para 21 dos 72 arquivos.

Considerações finais

O presente trabalho teve como objetivo apresentar algumas das aplicações reais que aparecem no âmbito do IBGE e que estão associadas com problemas de agrupamento de difícil solução computacional.

Neste sentido, foi possível observar que tal característica demanda o estudo de metodologias e o desenvolvimento de métodos de computação intensiva que sejam eficientes e eficazes no que concerne, respectivamente, ao tempo computacional e à qualidade das soluções produzidas para esses problemas.

Considerando essas observações e as metodologias que foram descritas na seção 3, percebe-se que a pesquisa operacional (PO) é uma importante ferramenta computacional. Mais especificamente, a PO disponibiliza uma gama muito rica de metodologias que podem ser aplicadas e/ou adaptadas para o equacionamento de diversos problemas que aparecem no âmbito do IBGE, em particular para aqueles que foram descritos no presente trabalho.

Referências

AHMADI, S.; OSMAN, I.H. Greedy random adaptive memory programming search for the capacitated clustering problem. *European Journal of Operational Research*, 162, 30-44, 2005.

AHUJA, R. K. *Networks Flows – Theory, Algorithms and Applications*. Prentice Hall, 1993.

ASSUNÇÃO, R. M.; LAGE, J. P.; REIS, E. A. Análise de Conglomerados Espaciais Via Árvore Geradora Mínima. *Revista Brasileira de Estatística*, 63 (220), 7-24, 2002.

AZEVEDO, R. V. *Estudo Comparativo de Métodos de Estratificação Ótima de Populações Assimétricas*. Dissertação de Mestrado. IBGE/ENCE, Rio de Janeiro, 2004.

BASTOS, L.O. *Novos Algoritmos e Resultados Teóricos para o Problema de Particionamento de Grafos por Edição de Arestas*. Tese de Doutorado - UFF/IC, Rio de Janeiro, 2012.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de Amostragem*. ABE/Projeto Fisher. Edgard Blucher, 2005.

BRITO, J.A.M.; BRITO, L.R. Algoritmos VNS e Genéticos Aplicados ao Problema de agrupamento com Soma Mínima de Distâncias. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, João Pessoa, Paraíba, Anais, 2008.

BRITO, J. A. M. ; BRITO, L. R. ; PASSINI, M. M. ; MONTENEGRO, F. M. T. . Uma Formulação de Programação Inteira para o Problema de Criação de Áreas de Ponderação Agregadas. In: XXXVI - Simpósio Brasileiro de Pesquisa Operacional, São João Del Rei. Anais do XXXVI SOBRAPO,1, 1662-1672, 2004.

BRITO, J. A. M. Uma Formulação de Programação Inteira para o Problema de Alocação Ótima em Amostras Estratificadas. In: Simpósio Brasileiro de Pesquisa Operacional, Gramado, Rio Grande do Sul, Anais, 2005.

BRITO, J.A.M; AZEVEDO, R.V.; MONTENEGRO, F.M.T. Algoritmos Genéticos Aplicados ao Problema de Estratificação. *Revista Brasileira de Estatatística*, 68 (229), 7-32, 2007.

BRITO, J.A.M; OCHI, L.S.; MONTENEGRO; F.M.T AND MACULAN, N. An iterative local search approach applied to the optimal stratification problem. *International Transactions in Operational Research*, 17, 753-764, 2010.

BRITO, J. A. M.; MACULAN, N.; BRITO, L. R.; MONTENEGRO, F. M. T. Um Algoritmo Grasp Aplicado ao Problema de Estratificação. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, Ubatuba, São Paulo, Anais, 2011.

BRITO, J. A. M., MONTENEGRO, F. M. T. Um Algoritmo VNS Aplicado ao Problema de Definição de Áreas de Ponderação. In: Simpósio de Pesquisa Operacional e Logística da Marinha, Rio de Janeiro, Anais, 2010.

BRITO, J. A. M.; DIAS, A. J. R.; MONTENEGRO, F. M. T.;CORTEZ, B. F. Um Algoritmo de Agrupamento Aplicado à Definição das Áreas de Ponderação do Censo Demográfico 2010. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, João Pessoa, Anais, 2012.

BRITO, J. A. M.; MONTENEGRO, F. M. T.; SOARES, M. P. Algoritmos de Otimização Aplicados à Estratificação da Amostra Mestra. In: ESCOLA DE AMOSTRAGEM E METODOLOGIA DE PESQUISA, Juiz de Fora, Minas Gerais, Anais, 2011.

CENSO Demográfico 2010: Resultados Gerais da Amostra. *Censo Demográfico*, Rio de Janeiro, 2010.

COCHRAN, W. G. *Sampling Techniques*, Third Edition. New York. John Wiley, 1977.

CRUZ, M. D. *O Problema de Clusterização Automática.* Tese de Doutorado, UFRJ/ COPPE, Rio de Janeiro, 2010.

DALENIUS, T; HODGES, J. L. JR. Minimum Variance Stratification. *Skandinavisk Aktuarietidskrift*, 54, 88-101, 1959.

DIAS, C.R. *Algoritmos Evolutivos para o Problema de Clusterização de Grafos Orientados: Desenvolvimento e Análise Experimental*. Dissertação de Mestrado, UFF/IC, Rio de Janeiro, 2004.

DORNDORF, U; PESCH E. Fast Clustering Algorithms. *ORSA J. Computing,* 6, 141-153, 1994.

DUTRA,V.G. *Algoritmo genético aplicado ao problema de p-medianas c*apacitado. Monografia de final de curso, UFOP, Minas Gerais, 2008.

EKMAN, G. An aproximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30, 219–229, 1959.

FEO, T.A; RESENDE, M.G.C. Greedy randomized adaptive search procedures. *Journal of Global Optimization*, 6, 109-133, 1995.

FREITAS, M.P.S; LILA, M.F.; AZEVEDO, R.V.; ANTONACI, G.A. Amostra Mestra Para o Sistema Integrado de Pesquisas Domiciliares, Textos para Discussão (23), Diretoria de Pesquisas, 2007.

FURTADO, J.C. *Algoritmo genético construtivo na otimização de problemas combinatoriais de agrupamentos*. Tese de Doutorado, INPE, São José dos Campos, 1998.

GLOVER, F.; KOCHENBERGER, G. A. H*andbook of Metaheuristics.* First Edition Norwell: K1uwer Academic Publishers, 2002.

GORDON, A. D. A Survey of Constrained Classification. *Computational Statistics and Data Analysis*, 21, 17-29, 1996.

GUNNING, P.; HORGAN, J. A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 159-166, 2004.

HAIR, J.F; BLACK, W.C; BABIN, B.J.; ANDERSON, R.E.; TATHAM, R.L. *Análise Multivariada de Dados*. Bookman, 2009.

HANSEN, P.; JAMAURD B. *Cluster Analysis and Mathematical Programming*. Les Cahiers du GERAD, 1997.

HANSEN, P.; MIADENOVIC, N. Variable Neighborhood Search: Principles and applications. *European Journal of Operational Research*, 130 (3), 449-467, 2001.

HEDLIN, D. *On the stratification of highly skewed populations.* RD Report. Statistics Sweden, Sweden, 1998.

HEDLIN, D. A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 16, 15-29, 2000.

JAIN, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*, 651-666, 2010.

JOHNSON, A.R.; WICHERN, D.W. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.

KAUFMAN L. E ROUSSEEUW P.J. *Finding Groups in Data – An Introduction to Cluster Analysis*. Wiley-Interscience Publication, 1989.

KESKINTURK T.; ER, SEBNEM. A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics and Data Analysis*, 52, 53-67, 2007.

KHAN, M.G.M.; NAND, N; AHMAD, N. Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34, 205-214, 2008.

KOZAK, M. Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806, 2004.

LAVALLÉE, P.; HIDIROGLOU, M. On the stratification of skewed populations. *Survey Methodology* (Statistics Canada), 14, 33-43, 1988.

LEISERSON, C. E.; RONALD, C. S.; RIVEST, L.; CORMEN, T.H. *Algoritmos – Teoria e Prática*. 3ª Edição. Elsevier, 2012.

LIMA, B.B.; COSTA, F.L.P; OCHI, L.S. Melhorando o desempenho de metaheurísticas grasp e algoritmos evolutivos: uma aplicação para o problema de árvore de custo mínimo com grupamentos. In: Simpósio Brasileiro de Pesquisa Operacional, Natal/RN, Anais, 2003.

LOHR, S.L. *Sampling: Design Analysis*. Brooks/Cole. Cengage Learning, 2010.

MARCOTORCHIUO, E.; MICHAUD, P. *Optimisation en Analyse Ordinate des Donmfes,* Masson, Paris, 1979.

MEYER, P. L. *Probabilidade – Aplicações à Estatística.* LTC, 2009.

MICHAUD, P. Clustering techniques. *Future Generation Computer Systems*, 14, 135-147, 1997.

MINGOTI, S.A. *Análise de Dados Através de Métodos de Estatística Multivariada – Uma Abordagem Aplicada*. Editora UFMG, 2007.

MONTENEGRO, F.M.T.; TORREÃO, J.R.A; MACULAN, N. Microcanonical Optimization algorithm for the Euclidean Steiner problem in R^n with application to phylogenetic inference. *Physical Review E*, 68 (5), 567021-567025, 2003.

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

NALDI, C. N. *Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados*. Tese de Doutorado, USP, São Carlos, 2011.

NASCIMENTO, M.C.V.; TOLEDO, F.M.B; CARVALHO, C.P.L.F. Investigation of a new GRASP-based clustering algorithm applied to biological data. *Computers & Operations Research*, 37, 1381-1388, 2010.

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

ROMESBURG, H.C. *Cluster Analysis for Researchers*. Lulu Press, 2004.

SEMAAN, G. S. ; MONTENEGRO, F. M. T. ; BRITO, J. A. M. ; OCHI, L. S. . Um Método Sistemático de Particionamento de Grafos Aplicado ao Problema de Agrupamento Automático. In: X Congreso Chileno de Investigación Operativa, Concepcion. X Congreso Chileno de Investigación Operativa, 2013.

SEMAAN, G. S. ; OCHI, L. S. ; BRITO, J. A. M. . Um Algoritmo Evolutivo Hibrido Aplicado ao Problema de Clusterização em Grafos com Restrições de Capacidade e Contiguidade. In: IX Congresso Brasileiro de Redes Neurais e Inteligência Computacional (IX CBRN), Ouro Preto. Anais do IX CBRN, 2009. v. 1.

SEMAAN, G. S. ; OCHI, L. S. ; BRITO, J. A. M. ; MONTENEGRO, F. M. T. . An efficient evolutionary algorithm for the aggregated weighting areas problem. In: International Conference on Engineering Optimization - EngOpt2008, 2008, Rio de Janeiro. Proc. of the EngOpt2008 - Sponsoring Societies: Mathematical Programming Society (MPS), ISSMO, EUROPT, ABCM. RJ : EngOpt, 1, 2008.

SERPA, D.L.; CHAVES, A. A.; CORRÊA, F.A.; LORENA, L.A.N. Metaheurística VNS Aplicada a Problemas de Agrupamentos. In: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, Bento Gonçalves, Rio Grande do Sul, Anais, 2009.

SIVANANDAM, S.N.; DEEPA S. N. *Introduction to Genetic Algorithms*. Springer, 2008.

SHIEH, H.M AND MAY, M.D. Solving the capacitated clustering problem with genetic algoritm. *Journal of the Chinese Institute of Industrial Engineers*, 18 (3), 1-12, 2001.

SILVA, A. N.; CORTEZ, B.F; MATZENBACHER, L.A. *Processamento das Áreas de Expansão e Disseminação da Amostra no Censo Demográfico 2000*. Textos para Discussão, 17, IBGE/DPE/ COMEQ, 2004.

SOARES, A. S. R. F. *Metaheurísticas para o Problema de Clusterização Automática*. Dissertação de Mestrado, UFF/IC, Niterói, 2004.

SOSA, N.G.M. *Heurísticas e Metaheurísticas para o Problema de Agrupamento Capacitado*. Dissertação de Mestrado, Unicamp, São Paulo, 1996.

TAN, P. N.; STEINBACH, M.; KUMAR V. *Introdução ao Data Mining – Mineração*. Ciência Moderna, 2009.

VIEIRA, C.E.C. *Heurísticas para o Problema das p-Medianas Conectadas*. Tese de Doutorado, PUC, Rio de Janeiro, 2006.

WOLSEY, L.A; NEMHAUSER, G.L. *Integer and Combinatorial Optimization*. Wiley Interscience, 1999.