Um Algoritmo GRASP Aplicado ao Problema de Estratificação

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

Nelson Maculan
COPPE/UFRJ – Programa de Engenharia de Sistemas e Computação
Caixa Postal 68511 , CEP: 21941-972, Rio de Janeiro, RJ, Brasil
e-mail: nelson.maculan@gmail.com

Luciana Roque Brito
Centro Universitário Plínio Leite - UNIPLI
Av. Visconde do Rio Branco, 123, Centro – Niterói – RJ.
e-mail: lubritoroque@gmail.com

Flávio Marcelo Tavares Montenegro
Instituto Brasileiro de Geografia e Estatística – IBGE
Av. República do Chile, 500, 10º andar, Centro – Rio de Janeiro – RJ.
e-mail: flavio.montenegro@ibge.gov.br

RESUMO

O presente trabalho traz uma nova proposta para a resolução do problema de estratificação. Em tal problema, deve-se determinar ***L*** estratos populacionais (grupos) de forma que os elementos em cada um dos estratos sejam mais homogêneos entre si. Para a construção de estratos homogêneos utiliza-se função objetivo que corresponde a uma expressão de variância. A utilização da estratificação possibilita a obtenção de estimativas mais precisas, ou seja, com um menor erro padrão associado. Com o objetivo de produzir estratos mais homogêneos, propõe-se no presente trabalho um novo algoritmo que usa os conceitos da metaheurística GRASP. A última seção traz alguns resultados computacionais para um conjunto de instâncias, considerando a aplicação do GRASP e de quatro algoritmos da literatura.

PALAVRAS CHAVE: Estratificação. Amostragem. Metaheurística GRASP. MH.

ABSTRACT

This paper presents a new approach to solve the stratification problem. In this problem, we must determine ***L*** strata (clusters) so that the elements in each stratum are more homogeneous among themselves. For the construction of homogeneous strata, an objective function that corresponds to an expression of variance should be considered. The application of stratification makes it possible to obtain more precise estimates, i.e., with a smaller associated standard error. Aiming at producing more homogeneous strata, this paper proposes a new algorithm that uses GRASP metaheuristic concepts. Computational results are presented for a set of instances and consider the application of the GRASP algorithm as well as four algorithms found in the literature.

KEYWORDS: Stratification. Sampling. GRASP Metaheuristic. MH.

1. Introdução

Atualmente, boa parte das pesquisas realizadas pelos institutos oficiais de estatística considera a adoção de um plano amostral. Ou seja, antes da realização da pesquisa define-se a população que será investigada, o recorte geográfico, a base de dados a ser utilizada para seleção e o esquema de amostragem que será considerado. O levantamento por amostragem possibilita a obtenção de estimativas (valores aproximados) para parâmetros reais da população entrevistando apenas um subconjunto dessa população denominado amostra (Bolfarine e Bussab, 2005).

Acrescenta-se, ainda, que ao aplicar-se um plano amostral, busca-se o equilíbrio entre o orçamento disponível para a pesquisa e a necessidade de um bom nível de precisão para as estimativas a serem divulgadas. Tal nível de precisão pode ser alcançado explorando de uma forma eficiente a relação de homogeneidade observada entre elementos da população em estudo.

A consideração essas duas questões implica em incorporar ao plano amostral a amostragem estratificada. Com a utilização da estratificação, ou seja, com a reunião dos elementos da população em $\boldsymbol{L}$ grupos, ou *estratos*, mais homogêneos, possibilita-se a produção de estimativas com um maior nível de precisão (menor erro padrão associado). O "grau" de homogeneidade dos estratos é medido a partir da avaliação de uma expressão variância associada com uma variável de estratificação previamente escolhida da pesquisa. Consequentemente, quanto menor o valor da variância, mais homogêneos serão os estratos.

A partir da descrição acima, observa-se que o problema de definição dos estratos pode ser mapeado em um problema de agrupamento com uma função objetivo não linear (variância). De acordo com a literatura, a elevada complexidade dos problemas de agrupamento torna atrativa e necessária a utilização de abordagens baseadas em heurísticas e/ou metaheurísticas, em detrimento de abordagens exatas. Em particular, propõe-se no presente trabalho um algoritmo de estratificação baseado na metaheurística GRASP.

Este trabalho está dividido da seguinte forma: na seção dois são apresentados os conceitos básicos de amostragem e de estratificação. A seção três traz uma descrição detalhada do problema de estratificação e comenta os principais trabalhos da literatura. A seção quatro tem uma descrição da metaheurística GRASP e do novo algoritmo de estratificação proposto. Finalmente, na seção cinco, há uma apresentação de um conjunto de resultados computacionais obtidos a partir da aplicação do algoritmo GRASP e de quatro algoritmos da literatura. Tais algoritmos foram aplicados em instâncias artificiais e reais.

2. Conceitos sobre Amostragem e Estratificação

Segundo Barbetta (2008) "A amostragem é naturalmente usada em nossa vida diária. Por exemplo, para verificar o tempero de um alimento em preparação, podemos provar uma pequena porção. Estamos fazendo uma amostragem, ou seja, extraindo do todo (população) uma parte (amostra), com o propósito de termos uma idéia (inferirmos) sobre a qualidade de tempero de todo o alimento".

A aplicação da amostragem permite a obtenção de informações a respeito de parâmetros populacionais desconhecidos, mediante a observação de apenas uma parte (amostra) do seu universo de estudo (população). Os elementos de uma população são definidos como as unidades de observação. Em termos matemáticos, uma população corresponde a um conjunto de $\boldsymbol{N}$ elementos que possuem pelo menos uma característica em comum. A população pode ser formada por pessoas, famílias, estabelecimentos, domicílios, setores censitários, ou qualquer outro tipo de elementos, dependendo basicamente dos objetivos da pesquisa. Entre os motivos para o uso da amostragem, destacam-se os seguintes: (1) Economia: Em geral, torna-se muito mais econômico o levantamento de somente uma parte da população; (2) Tempo: Em uma pesquisa eleitoral, a três dias de uma eleição presidencial, não haveria tempo suficiente para pesquisar toda a população de eleitores do país, mesmo com a disponibilidade de grandes recursos financeiros; (3) Operacionalidade: É mais

fácil realizar operações de pequena escala. Um dos problemas que tipicamente ocorrem nos grandes censos é o de controlar os entrevistadores, ou seja, garantir que: (i) o entrevistador fará as perguntas de forma correta, evitando entendimento incorreto, (ii) o entrevistador realizará todas as entrevistas previamente estabelecidas, etc.

Levando-se em conta as informações que se pretende inferir para a população, as restrições de custo da pesquisa e o nível de precisão desejado, pode-se optar por vários esquemas clássicos de amostragem (Lohr, 2010), quais sejam: amostragem aleatória simples, amostragem de conglomerados, amostragem sistemática e amostragem estratificada.

Em particular, descreve-se a seguir os conceitos de amostragem aleatória simples estratificada, tendo em vista que tal esquema está associado ao problema que foi o objeto de estudo deste trabalho.

Segundo Bolfarine e Bussab (2005), na amostragem estratificada divide-se uma população em grupos, chamados de estratos, segundo algumas características (variáveis) conhecidas na população da pesquisa. A utilização da amostragem estratificada tende a ser mais conveniente para o administrador, resultando em um menor custo para pesquisa e pode produzir estimativas (valores obtidos a partir da amostra) mais precisas para a população, ou seja, com um menor erro padrão associado. Quando a variabilidade interna dos estratos formados for pequena, a estratificação foi bem sucedida. Apresenta-se a seguir a notação básica associada à amostragem estratificada e que foi considerada nesse trabalho.

L	Número estratos
N	Número total de unidades da população
n	Número total de unidades na amostra
N_h	Número total de unidades da população em cada estrato $(h = 1,..., L)$
n_h	Número total de unidades da amostra em cada estrato $(h = 1,..., L)$
y_{hi}	Valor de uma variável Y, para a *i-ésima* unidade do *h-ésimo* estrato
$\overline{Y_h} = \sum_{i=1}^{N_h} y_{hi} / N_h$	Valor médio de Y na população do *h-ésimo* estrato
$S_{yh}^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \overline{y_h})^2}{N_h - 1}$	Variância de Y na população do *h-ésimo* estrato
T_X	Total associado com uma variável de estratificação $\boldsymbol{X}$

A estratificação é uma técnica bastante usada em pesquisas por amostragem probabilística, sendo a parte essencial do planejamento amostral em pesquisas econômicas e de produção. Com aplicação da amostragem estratificada, uma população U com N unidades é dividida em L subpopulações, com $N_1, N_2,..., N_h,..., N_L$ unidades, respectivamente, chamadas de estratos (Lohr 2010). Essas subpopulações não se superpõem e, tomadas em conjunto, abrangem a totalidade da população, de tal modo que:

$$N_1 + N_2 + ... + N_h + ... + N_L = N \qquad (1)$$

Para que sejam obtidos todos os proveitos da estratificação, os valores de $N_h, h = 1,..., L$, devem ser conhecidos, o que implica, consequentemente, na definição dos estratos. Depois de definidos os estratos, a partir do conhecimento de uma ou mais características (variáveis) da população e considerando um tamanho de amostra n, selecionam-se, de forma independente, L amostras aleatórias de tamanho n_h $(h = 1,..., L)$ que serão alocadas aos L estratos, de forma que:

$$n_1 + n_2 + ... + n_L = n \qquad (2)$$

A aplicação de um plano amostral estratificado requer, concomitantemente, a solução de dois tipos de problemas, quais sejam: a construção dos estratos e a escolha dos métodos de seleção e alocação das amostras nos estratos. Os fatores (Lohr, 2010) que influenciam a eficiência da estratificação estatística no que diz respeito à construção dos estratos são: a escolha da variável de estratificação, o número de estratos a serem formados e como os estratos devem ser delimitados. A escolha do método de alocação da amostra entre os estratos também tem um impacto direto na eficiência da estratificação, ou seja, no valor da variância.

3. Problema de Estratificação

Considere 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 pré-fixado 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 $x_1 \le x_2 \le ... \le x_N$. 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 $b_1 < b_2 < ... < b_h < ... < b_{L-1}$:

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

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 (*AAS*) 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,

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

Deve-se observar que os valores de N_h e S_{yh}^2 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 um das seguintes expressões abaixo (Lohr, 2010):

$$n_h = \frac{n}{L} \qquad (3.3) \qquad n_h = \frac{n.N_h}{N} \qquad (3.4) \qquad n_h = \frac{n.N_h.S_{hy}}{\sum_{h=1}^{L} N_h.S_{hy}} \qquad (3.5)$$

A expressão (3.3) está associada com a alocação *Uniforme (AUn)*, 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 (3.4) está associada com a alocação *Proporcional* (*APr*). 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 (3.5) está associada com a alocação de *Neyman* (*ANe*). Neste caso, o número de unidades da amostra a serem observadas no estrato h é proporcional a $N_h.S_{hy}$. Em geral, os tamanhos de amostra obtidos a partir da alocação de *Neyman* produzem uma maior redução do valor da variância (equação três).

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: Dalenius e Hodges (1959), Ekman (1959), Lavallée e Hidiroglou (1988), Hedlin (1998, 2000).

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

$$V_X = \sum_{h=1}^{L} N_h^2 \frac{S_{xh}^2}{n_h}.(1-\frac{n_h}{N_h}) \qquad (4)$$

A obtenção do mínimo global para a variância definida em (3) ou (4), 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 $(\lfloor N/2 \rfloor - 1)$ tomados $(L\text{-}1)$ a $(L\text{-}1)$: $C_{L-1}^{\lfloor N/2 \rfloor - 1}$, 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 de tal alocação pode produzir tamanhos de amostra maiores que os tamanhos populacionais ($n_h > N_h$). Adotando-se, nesse caso, um procedimento de redistribuir o tamanho de amostra excedente para outros estratos onde $n_h < N_h$, sendo mais uma vez comprometida a questão da otimalidade.

Em decorrência, principalmente, 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). Tal 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 Freqüência, 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 extendida 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 pré-fixado, Lavallée e Hidiroglou (1988) proporam um algoritmo que constrói os estratos considerando a minimização 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. Um algoritmo muito simples e prático para a definição dos limites dos estratos. Eles verificaram que para distribuições assimétricas (Barbetta, 2008) 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*. Tal

método tem certa similaridade com a metaheurística *VNS* (Glover and Kochenberger, 2002). Khan *et al*. (2008) desenvolveram um algoritmo baseado em programação dinâmica para determinar os limites dos estratos. Tal algoritmo pode ser aplicado apenas quando $\boldsymbol{X}$ tem uma distribuição normal ou triangular e a amostragem é feita com reposição (Bolfarine e Bussab, 2005).

Acrescentam-se a tais algoritmos heurísticos três abordagens baseadas em metaheurísticas, quais sejam: Keskintürk e Er (2007) proporam um algoritmo genético que determina simultaneamente os limites dos estratos e 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) proporam um algoritmo que utiliza os conceitos da metaheurística ILS e do Path Relinking (Glover and Kochenberger, 2002).

4. Método de Resolução

Tendo em vista as dificuldades de resolução do problema e a limitação dos métodos heurísticos propostos na literatura, no que concerne à qualidade das soluções viáveis produzidas, propõe-se no presente trabalho um algoritmo de estratificação que utiliza os conceitos da metaheurística *GRASP* (Feo e Resende, 1995; Glover e Kochenberger, 2002). Tal algoritmo apresentar-se-á como mais uma alternativa em relação às abordagens propostas por Keskintürk e Er (2007) e Brito *et al*. (2007, 2010).

4.1 Metaheurística GRASP

A metaheurística *GRASP* (*Greedy Randomized Adaptive Search Procedure*) é um procedimento iterativo utilizado para resolver problemas de otimização combinatória. Cada iteração do GRASP consiste de duas fases: construção e busca local. A fase de construção é caracterizada pela obtenção de uma solução viável cuja vizinhança é investigada na fase busca local. A melhor solução obtida após todas as iterações será a solução do problema.

Para construção da solução $\boldsymbol{S_o}$ é considerada uma lista de candidatos *(LC)* formada por todos os elementos que se incorporados em $\boldsymbol{S_o}$ não a tornam inviável. Definida a ***LC***, deve-se avaliar todos os seus elementos através de uma função gulosa ***g(.)***, que representa o custo de se adicionar um novo elemento $t \in \boldsymbol{LC}$ na solução $\boldsymbol{S_o}$. Uma forma de possibilitar uma maior variabilidade nas soluções obtidas consiste em definir uma lista de candidatos restrita ***(LCR)***, formada pelos melhores elementos avaliados na *LC* através da função ***g.*** Ou seja, aqueles que se incorporados à solução em construção produzem um acréscimo mínimo (problema de minimização). Este processo representa o aspecto *guloso* do *GRASP*, pois se considera sistematicamente os melhores elementos para serem inseridos na solução.

O trabalho de Feo e Resende (1995) descreve dois possíveis esquemas para a construção da ***LCR***: (i) um inteiro ***k*** é fixado e os ***k*** melhores candidatos, ordenados na ***LC*** segundo algum critério, são selecionados para compor a ***LCR*** e (ii) em cada iteração da fase de construção, denota-se respectivamente por $\mathbf{g_{min}}$ e $\mathbf{g_{max}}$ os menores e maiores acréscimos provocados pela inserção de um elemento $t \in \boldsymbol{LC}$ na solução, segundo a avaliação de um função gulosa ***g(.)*** . A partir da aplicação dessa função e da utilização dos valores $\mathbf{g_{min}}$ e $\mathbf{g_{max}}$, pode-se definir:

$$\mathbf{LCR}=\{ t \in \boldsymbol{LC} \,|\, g(t) \leq g_{min} + \alpha(g_{max} - g_{min}),\ \alpha \in [0,1]\}.$$

Quando $\alpha=0$, o procedimento de construção torna-se *guloso*, pois a ***LCR*** tem apenas um elemento, e quando $\alpha=1$, produz-se uma solução aleatória, tendo em vista que a ***LCR*** terá todos os elementos da ***LC***. Tomando-se um valor intermediário para α o procedimento torna-se semi-guloso.

A busca local é aplicada com o objetivo de melhorar a solução inicial S_o obtida na primeira fase, tendo em vista que esta solução não é necessariamente uma solução ótima para o problema. Ou seja, a busca local consiste na substituição da solução S_o pela melhor solução S' encontrada em uma vizinhança de S_o.

Alternativamente, pode-se trabalhar com um conjunto discreto de valores para α em vez de considerar apenas um valor fixo. Dessa forma, em cada iteração seleciona-se um valor α_i de um conjunto de valores. Prais e Ribeiro (2000) mostraram que ao utilizar-se um valor fixo para o parâmetro α, pode-se retardar a obtenção de soluções de alta qualidade, as quais eventualmente seriam encontradas caso fosse utilizado outro valor para α. Sendo assim, foi proposta uma extensão do procedimento GRASP básico, denominado GRASP reativo, na qual o parâmetro α é selecionado aleatoriamente a cada iteração a partir de um conjunto discreto de possíveis valores.

4.2 Algoritmo GRASP

Antes da aplicação do algoritmo, efetua-se uma modificação no vetor populacional X_P. Tendo em vista que as N observações de X_P estão em ordem crescente, é possível agrupá-las considerando-se apenas os seus valores distintos (as observações com mesmo valor devem obrigatoriamente ficar em um mesmo estrato). Ou seja, constrói-se um conjunto $B = \{b_1, b_2, ..., b_k\}$ que contém ***k*** valores distintos de X_P que correspondem aos possíveis limites para a estratificação da população. Uma vez definidos o número L de estratos e o conjunto ***B***, deve-se escolher ***(L-1)*** limites de ***B*** de forma a obter a menor variância possível considerando a expressão (4). Tais limites corresponderão aos pontos de corte $b_1 < b_2 < ... < b_h < ... < b_{L-1}$ da solução corrente.

4.2.1 Procedimento de construção:

Define-se uma lista de candidatos LC cujos elementos correspondem aos limites (pontos de corte) de ***B*** que podem ser adicionados à solução, e consequentemente, definir os estratos populacionais. Cada limite b_i da ***LC*** tem o seu custo de sua inserção g_i avaliado através a seguinte expressão: $N_h^2.S_{xh}^2$. Tal expressão, denominada variância parcial, é um limite superior para o valor da variância total definida na equação (4). Dessa forma, para o primeiro limite b_i inserido na solução, calcula-se a variância parcial no 1º estrato (h=1), para o segundo limite b_i selecionado, calcula-se a variância parcial para o 2º estrato (h=2), e assim sucessivamente.

. Após o cálculo do custo de inserção de cada ponto de corte na solução, também se calcula o mínimo e o máximo de g_i, ou seja, a menor e a maior variância parcial. A partir destes valores, define-se uma lista de candidatos restrita ***LCR*** que é formada por todos os pontos de corte de ***LC*** tais que $LCR = \{ b_i \in LC \mid g_i \leq g_{min} + \alpha(g_{max} - g_{min}) \}$. O ponto de corte que será adicionado na solução é selecionado aleatoriamente da ***LCR***.

A condição de parada do procedimento de construção é definida pelo número de estratos que devem ser formados. Sendo assim, na 1ª iteração temos a escolha do 1º ponto de corte, na 2ª iteração temos a escolha do 2º de corte, e assim sucessivamente, até a escolha do ***(L-1)***-ésimo ponto de corte. Após a escolha de um ponto de corte da ***LCR***, atualiza-se a ***LC*** e são recalculados os valores de g_{mi} e g_{max}.

Considerando o procedimento descrito acima, são construídas ***p*** soluções iniciais em cada iteração do GRASP, selecionando-se a de melhor qualidade (menor variância) para a aplicação da busca local**.** Tais soluções correspondem a algumas das possíveis estratificações para o problema. A aplicação sistemática de tal procedimento tende a eliminar as soluções iniciais de baixa qualidade.

Uma vez definidos os pontos de corte e, por conseguinte, os termos N_h e S_{xh}^2, deve-se calcular os tamanhos de amostra n_h que serão alocados aos ***L*** estratos. Todavia, em vez da aplicação

da alocação de *Neyman*, optou-se pela utilização de uma formulação de programação inteira (proposta por Brito em 2005) que utiliza variáveis binárias que assumem o valor um se um determinado tamanho de amostra $\mathbf{n_h}$ for alocado a um estrato $\boldsymbol{h}$ **(h=1,...L)**. Para resolução desta formulação, programou-se um algoritmo em linguagem ***R*** que faz uso de um algoritmo de programação inteira disponível no pacote *LpSolve* (também do R). Com a aplicação deste algoritmo garante-se: o cumprimento das restrições $\mathbf{n_h \leq N_h}$ e $\mathbf{\sum n_h = n}$ e tamanhos de amostra que minimizam a variância da equação (4). Após a aplicação desse algoritmo, temos construída uma solução inicial S^o.

4.2.2 Procedimento de Busca Local:

A busca local é aplicada em cada ponto de corte de S^o visando uma redução do valor da variância. Mais especificamente, a busca local é caracterizada pela aplicação de dois tipos de buscas binárias, sendo a primeira entre o primeiro ponto de corte de ***B*** e cada um dos ***(L-1)*** pontos de corte de S^o e a segunda entre o último ponto de corte ***B*** e os ***(L-1)*** pontos de corte de S^o. A partir da aplicação dessas buscas, são gerados pontos de corte intermediários que redefinem os estratos, ou seja, há uma alteração nos valores de N_h, S_{xh}^2, n_h e, consequentemente, no valor da variância.

Busca (1): Inicialmente, define-se um limite inferior L_i igual a 1 (1º ponto de corte de ***B***) e um limite superior L_s igual a um valor ***q***, que corresponde à posição que o 1º ponto de corte de S^o ocupa em ***B***. Em seguida, substitui-se o 1º ponto de corte de S^o pelo ponto de corte que ocupa a ***m***-ésima posição em ***B*** , tal que $m = (L_i + L_s)/2$. Tal troca implica na redefinição dos estratos. Consequentemente, calculam-se novos valores de n_h utilizando o algoritmo exato (Brito, 2005) e calcula-se o valor da variância (eq. (4)). Havendo redução no valor da variância, atualiza-se a melhor solução obtida até o momento (denominada $\boldsymbol{S^1}$), L_i permanece inalterado e define-se $L_s = m-1$. Em caso contrário, $L_i = m+1$ e L_s permanece inalterado. Considerando tal procedimento, a busca é aplicada em cada um dos pontos de corte da solução corrente enquanto $L_i < L_s$.

Busca (2): De forma análoga, define-se um limite inferior L_i igual ***q*** (posição do 1º ponto de corte em S^o) e um limite superior L_s igual a um valor ***k*** que corresponde à posição do último ponto de corte em ***B***. Em seguida, substitui-se o 1º ponto de corte de S^o pelo ponto de corte que ocupa a ***m***-ésima posição em ***B***, tal que $m = (L_i + L_s)/2$. E novamente, havendo redução no valor da variância, atualiza-se a melhor solução obtida até o momento (denominada $\boldsymbol{S^2}$), L_i permanece inalterado e $L_s = m-1$. Em caso contrário, $L_i = m+1$ e L_s permanece inalterado. A busca é realizada em cada ponto de corte considerando a mesma condição de parada citada anteriormente.

Observa-se que as duas buscas partem da mesma solução inicial S^o e ao longo das divisões (redefinição dos estratos) as soluções $\boldsymbol{S^1}$ e $\boldsymbol{S^2}$ são atualizadas, caso ocorra uma redução no valor da variância. Ao final da aplicação das duas buscas, toma-se a melhor solução entre $\boldsymbol{S^1}$ e $\boldsymbol{S^2}$**.** A tabela a seguir ilustra a aplicação da busca local binária (algumas iterações) considerando L=4, o primeiro ponto de corte de S^o e |B|=31.

Tabela 1- Exemplo da Aplicação da Busca Binária

Solução	**Pontos de Corte**		
$\mathbf{S^o}$	$\mathbf{b_7}$	$\mathbf{b_{15}}$	$\mathbf{b_{21}}$
S^1 (L_i=1,L_s=7,m=4)	$\mathbf{b_4}$	$\mathbf{b_{15}}$	$\mathbf{b_{21}}$
S^1 (L_i=1,L_s=3,m=2)	$\mathbf{b_2}$	$\mathbf{b_{15}}$	$\mathbf{b_{21}}$
S^2 (L_i=7,L_s=31,m=19)	$\mathbf{b_{19}}$	$\mathbf{b_{15}}$	$\mathbf{b_{21}}$
S^2 (L_i=20,L_s=31,m=25)	$\mathbf{b_{25}}$	$\mathbf{b_{19}}$	$\mathbf{b_{21}}$

A partir da tabela 1 é possível observar, que em algumas situações, a busca binária pode gerar uma solução com os pontos de corte fora ordem. Para tais situações, efetua-se uma simples reordenação desses pontos.

5. Resultados Computacionais

O algoritmo GRASP e o algoritmo exato (Brito, 2005) foram implementados em linguagem ***R***. De forma a avaliar a eficiência do algoritmo GRASP, foi realizado um conjunto de experimentos computacionais com dados artificiais e dados reais, totalizando 27 instâncias. Todos os experimentos foram efetuados em um computador com 4GB de memória RAM e dotado de dois processadores de 3.33 GHz (*Core* 2 Duo). Os dados artificiais (disponibilizados em http://www.britomopt.net/13.html link **Instâncias GRASP**) foram construídos utilizando um procedimento descrito no trabalho de Hedlin (2000) e os dados reais correspondem a quatros populações. A primeira população está associada com um conjunto de estabelecimentos agropecuários produtores de café (IBGE), sendo o efetivo de pés de café a variável de estratificação considerada. As outras três populações foram extraídas da pesquisa industrial anual de 2004 (IBGE), sendo o número de pessoas ocupadas a variável de estratificação considerada. Por motivos de sigilo da informação, os dados das quatro populações supracitadas não foram disponibilizados.

5.1 Aplicação do Algoritmo GRASP

Em todos os experimentos realizados com o GRASP o número de iterações foi fixado em 100 e o número de filtros foi igual a cinco. Acrescenta-se, ainda, que o algoritmo foi aplicado em cada uma das 27 instâncias considerando o número de estratos variando entre três e seis.

Além da aplicação do algoritmo GRASP, foram utilizados, para efeito de comparação, quatro algoritmos da literatura, quais sejam: Distribuição Cumulativa da Raiz da Frequência (Cochran 1977, capítulo 5), Geométrico (Gunning e Horgan 2004), Genético (Brito *et al*, 2007) e o *Random Search* (Kozak, 2004). Os dois primeiros algoritmos estão disponibilizados em uma biblioteca do software ***R*** e o algoritmo Genético está disponibilizado em linguagem *Delphi*. O único algoritmo implementado pelos autores (em ***R***) foi o *Random Search.* Em função dos algoritmos da Distribuição Cumulativa e do Geométrico não serem iterativos, e do algoritmo genético ter sido implementado em *Delphi*, a avaliação da eficiência dos algoritmos ficou restrita a uma comparação entre os coeficientes de variação (cv's) produzidos pelos mesmos. O coeficiente de variação é uma medida adimensional que pode ser definida por $cv = 100.\sqrt{V_X} / T_X$. Ou seja, quanto menor o ***cv***, menor é a variância (e vice-versa).

As tabelas 2 e 3 a seguir contêm alguns resultados da aplicação do algoritmo GRASP e dos quatros algoritmos listados acima. Em particular, as três primeiras colunas dessas tabelas têm a identificação da instância (onde número após a letra "r" indica o tamanho da população ***N*** e o * indica que os dados da instância são reais) e as colunas dois e três trazem, respectivamente, o tamanho da amostra considerada e o número de pontos de corte do conjunto ***B***. E as colunas restantes trazem os valores dos coeficientes de variação obtidos mediante a aplicação do algoritmo GRASP e dos algoritmos Geométrico (GEO), da Distribuição Cumulativa da Raiz da Frequência (DCRF), do *Random Search* (RS) e do algoritmo Genético (AG). Os valores sublinhados e em negrito indicam a melhor solução encontrada para cada instância.

Uma análise das tabelas dois e três permite observar que o GRASP produziu soluções melhores do que os demais algoritmos para a maioria das instâncias, sendo essas de tamanho variado. Além disso, o algoritmo Genético (AG) seguido do *Random Search* (RS) foram os que produziram as soluções mais próximas das soluções GRASP. Uma explicação plausível para esta última observação é a de que esses algoritmos incorporam procedimentos de busca local e perturbação que exploram de uma forma mais eficiente do espaço de soluções. Tais procedimentos inexistem no caso do algoritmo Geométrico e da Distribuição Cumulativa.

Tabela 2 – Resultados dos Algoritmos (Três e Quatro Estratos)

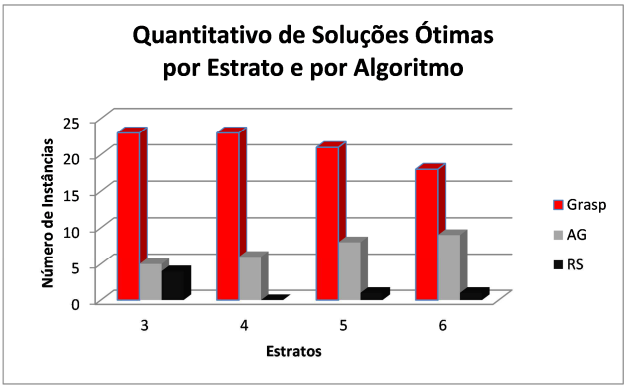
Nº de Estratos = 3			CV					Nº de Estratos = 4			CV				
Instância	n	|B|	Grasp	GEO	DCRF	RS	AG	Instância	n	|B|	Grasp	GEO	DCRF	RS	AG
r800	100	402	**1,555**	3,129	1,669	1,917	1,581	r800	100	402	**0,997**	2,747	1,054	1,475	1,051
r1000	200	469	**0,592**	1,909	0,911	1,091	0,595	r1000	200	469	**0,369**	1,184	0,623	0,581	0,458
r1076*	200	88	**1,045**	2,215	1,065	1,900	1,047	r1076*	200	88	0,692	1,665	0,791	1,475	**0,688**
r1500	140	578	2,055	3,420	2,266	**2,008**	2,056	r1500	140	578	1,290	2,393	1,297	1,439	**1,289**
r1616*	200	165	**1,570**	2,400	1,572	2,165	**1,570**	r1616*	200	165	**1,004**	1,732	1,043	1,439	1,010
r2000	500	730	**0,231**	1,301	0,646	0,548	0,238	r2000	500	730	**0,148**	0,378	0,421	0,243	0,188
r2911*	200	247	**3,224**	3,619	4,117	3,441	**3,224**	r2911*	200	247	2,227	2,525	2,664	2,326	**2,226**
r4000	700	1159	**0,248**	0,735	0,504	0,388	**0,248**	r4000	700	1159	**0,155**	0,511	0,328	0,184	0,156
r5000	800	1331	**0,156**	1,133	0,436	0,388	0,169	r5000	800	1331	**0,099**	0,285	0,418	0,184	0,135
r8000	900	1855	**0,341**	1,366	0,631	0,720	0,344	r8000	900	1855	**0,213**	0,551	0,493	0,335	0,214
r10000	2000	2072	**0,126**	0,563	0,393	0,282	**0,126**	r10000	2000	2072	**0,079**	0,187	0,253	0,123	**0,080**
r12000	800	2410	**0,536**	1,701	0,798	0,917	0,539	r12000	800	2410	**0,322**	1,062	0,657	0,574	0,324
r15000	2000	2688	**0,220**	0,772	0,487	0,420	0,221	r15000	2000	2688	**0,137**	0,410	0,318	0,215	0,138
r16000	1000	2794	**0,603**	1,887	0,927	1,026	0,605	r16000	1000	2794	**0,364**	1,342	0,765	0,676	0,366
r18000	1000	3059	**0,972**	2,305	1,269	1,296	0,977	r18000	1000	3059	**0,594**	1,906	0,904	1,222	**0,596**
r20000	2000	3227	**0,309**	1,038	0,642	0,562	0,310	r20000	2000	3227	**0,191**	0,655	0,388	0,303	0,192
r20742*	734	784	**1,997**	2,195	2,859	2,053	**1,997**	r20742*	439	784	**1,973**	2,200	2,821	2,081	**1,973**
r25000	2000	3685	**0,419**	1,073	0,651	0,613	0,421	r25000	2000	3685	**0,258**	0,846	0,486	0,439	0,259
r28000	1500	3942	**0,658**	1,227	0,885	0,727	0,660	r28000	1500	3942	**0,402**	1,126	0,602	0,720	0,404
r32000	1000	4321	1,509	2,506	1,570	**1,489**	1,513	r32000	1000	4321	**0,709**	1,899	1,028	0,720	0,911
r35000	2800	4486	**0,336**	1,218	0,585	0,656	0,337	r35000	2800	4486	**0,209**	0,662	0,476	0,330	0,212
r40000	3400	4882	**0,350**	0,840	0,596	0,482	0,351	r40000	3400	4882	**0,219**	0,695	0,382	0,373	0,220
r50000	3000	5558	**0,513**	1,291	0,770	0,740	0,516	r50000	3000	5558	**0,316**	1,046	0,525	0,626	0,318
r60000	1000	6218	**1,728**	3,175	1,859	1,813	1,738	r60000	1000	6218	**1,043**	2,474	1,563	1,957	1,046
r70000	1800	6762	1,159	1,971	1,197	**1,157**	1,163	r70000	1800	6762	**0,692**	1,615	0,829	1,153	0,694
r80000	2400	7222	**0,903**	1,907	1,041	1,093	0,907	r80000	2400	7222	**0,546**	1,549	0,900	1,142	0,548
r90000	2800	7925	0,789	1,335	0,974	**0,788**	0,791	r90000	2800	7925	**0,474**	1,127	0,700	0,789	0,480

Tabela 3 – Resultados dos Algoritmos (Cinco e Seis Estratos)

Nº de Estratos = 5			CV					Nº de Estratos = 6			CV				
Instância	n	|B|	Grasp	GEO	DCRF	RS	AG	Instância	n	|B|	Grasp	GEO	DCRF	RS	AG
r800	100	402	**0,704**	1,617	0,883	1,018	0,890	r800	100	402	**0,560**	1,431	0,698	0,799	0,800
r1000	200	469	**0,272**	0,694	0,503	0,435	0,395	r1000	200	469	**0,205**	0,671	0,385	0,362	0,280
r1076*	200	88	**0,517**	1,290	0,644	1,018	0,521	r1076*	200	88	**0,410**	1,069	0,521	0,799	0,429
r1500	140	578	**0,927**	1,958	0,991	1,224	**0,927**	r1500	140	578	**0,749**	1,567	0,824	0,908	0,768
r1616*	200	165	0,735	1,383	0,868	1,224	**0,732**	r1616*	200	165	0,632	1,096	0,705	0,908	**0,630**
r2000	500	730	**0,109**	0,415	0,295	0,223	0,153	r2000	500	730	**0,088**	0,238	0,251	0,154	0,114
r2911*	200	247	**1,617**	2,038	1,713	1,931	1,620	r2911*	200	247	1,272	1,675	1,322	1,602	**1,262**
r4000	700	1159	**0,116**	0,296	0,240	0,152	0,118	r4000	700	1159	**0,092**	0,268	0,212	0,115	0,123
r5000	800	1331	**0,072**	0,323	0,400	0,152	0,106	r5000	800	1331	**0,059**	0,180	0,387	0,115	0,073
r8000	900	1855	**0,154**	0,508	0,456	0,291	0,160	r8000	900	1855	**0,122**	0,363	0,432	0,209	0,136
r10000	2000	2072	**0,059**	0,199	0,193	0,118	0,061	r10000	2000	2072	**0,047**	0,124	0,139	0,078	0,049
r12000	800	2410	**0,239**	0,647	0,611	0,387	0,245	r12000	800	2410	**0,188**	0,572	0,572	0,369	0,206
r15000	2000	2688	**0,101**	0,291	0,236	0,174	**0,101**	r15000	2000	2688	**0,081**	0,238	0,202	0,134	0,090
r16000	1000	2794	**0,270**	0,696	0,697	0,424	0,273	r16000	1000	2794	**0,211**	0,653	0,635	0,447	0,214
r18000	1000	3059	**0,433**	1,016	0,755	0,605	0,437	r18000	1000	3059	0,345	0,885	0,689	0,623	**0,340**
r20000	2000	3227	**0,140**	0,393	0,329	0,232	**0,140**	r20000	2000	3227	**0,110**	0,337	0,302	0,190	0,112
r20742*	327	784	1,888	2,104	2,589	1,979	**1,880**	r20742*	242	784	1,864	2,118	2,595	1,995	**1,860**
r25000	2000	3685	0,188	0,444	0,340	0,264	**0,187**	r25000	2000	3685	**0,148**	0,408	0,312	0,261	0,150
r28000	1500	3942	0,296	0,733	0,450	0,401	**0,295**	r28000	1500	3942	**0,236**	0,523	0,383	0,353	**0,236**
r32000	1000	4321	0,507	1,532	0,790	**0,501**	0,664	r32000	1000	4321	0,406	1,133	0,652	**0,353**	0,525
r35000	2800	4486	0,153	0,436	0,431	0,264	**0,152**	r35000	2800	4486	**0,120**	0,376	0,404	0,211	0,129
r40000	3400	4882	**0,160**	0,367	0,285	0,219	0,163	r40000	3400	4882	**0,124**	0,330	0,246	0,206	0,135
r50000	3000	5558	**0,231**	0,544	0,455	0,324	0,231	r50000	3000	5558	0,189	0,499	0,401	0,337	**0,183**
r60000	1000	6218	**0,741**	1,852	1,355	1,359	0,759	r60000	1000	6218	0,585	1,272	1,237	1,015	**0,581**
r70000	1800	6762	**0,499**	1,203	0,686	0,872	0,502	r70000	1800	6762	0,396	0,844	0,611	0,621	**0,393**
r80000	2400	7222	**0,399**	0,929	0,774	0,591	0,402	r80000	2400	7222	0,313	0,753	0,698	0,560	**0,312**
r90000	2800	7925	**0,341**	0,832	0,598	0,587	0,342	r90000	2800	7925	**0,274**	0,575	0,532	0,416	0,277

Além dessas tabelas, construiu-se um gráfico cujos totais no eixo vertical correspondem ao número de instâncias para as quais os algoritmos GRASP, AG e RS produziram a melhor solução (denominada "solução ótima"). Considerando a cada uma das instâncias *versus* o número de estratos, o GRASP produziu a melhor solução em 79% dos casos, seguido respectivamente pelo AG com 28% dos casos e pelo RS com 6% dos casos.

Gráfico 1



E finalmente, foram calculadas (tabela 4) algumas medidas resumo (mínimo, *1°quartil*, mediana, média, *3°quartil* e máximo) relacionadas com o gap ρ=100*(sol_{AG}-sol_{GRASP})/sol_{GRASP}, considerando apenas as instâncias cujas soluções do GRASP foram melhores do que as soluções do AG. Analisando a tabela quatro, e em particular os *gaps* médio e a mediano, observa-se que os maiores ganhos GRASP em relação ao AG ocorrem à medida que o número de estratos aumenta. Em contrapartida, quando se analisa o quantitativo de soluções (gráfico 1) observa-se que à medida que o número de estratos aumenta o AG tende a produzir mais soluções ótimas.

Tabela 4 - Medidas Resumo dos *Gaps* entre o **GRASP** e o **AG**

Estratos	3	4	5	6
Mínimo	0,04	0,22	0,15	0,09
Q1	0,34	0,39	0,62	2,20
Médio	0,93	5,64	11,15	13,70
Mediano	0,44	0,60	2,24	8,40
Q3	0,50	1,29	15,32	25,85
Máximo	8,61	35,51	47,44	42,69

Embora o tempo de processamento não tenha sido avaliado nos experimentos desse estudo, destaca-se que o algoritmo GRASP consumiu um tempo mínimo dez segundos para resolver a menor (***N***=800) instância e um máximo de duas horas para resolver a maior (***N***=9000) instância.Todas essas observações indicam que a combinação do GRASP com o algoritmo exato proposto por Brito (2005) pode constituir-se como mais uma boa alternativa para a resolução do problema de estratificação.

Em trabalhos futuros pretende-se: (i) Incorporar a esse algoritmo novos procedimentos de construção e novas buscas locais que utilizem o *Path Relinking* e o VNS (Glover e Kochenberger, 2002), (ii) Estudar e propor algoritmos para o problema de estratificação multivariada (Khan *et al.*, 2010) e (iii) Resolver o problema de estratificação univariada sem fixar previamente o número de estratos, ou seja, utilizando a ideia de clusterização automática (Cruz, 2010).

Bibliografia

Azevedo,R. V., Estudo Comparativo de Métodos de Estratificação Ótima de Populações Assimétricas. Dissertação de Mestrado. IBGE/ENCE, 2004.

Barbetta, P.A. , Estatística Aplicada às Ciências Sociais, Editora da UFSC, 2008.

Bolfarine, H. e Bussab, Wilton O, Elementos de Amostragem. ABE, Projeto Fisher, Editora Edgard Blücher, 2005.

Brito, J.A.M, Azevedo, R.V., Montenegro, F.M.T, Algoritmos Genéticos Aplicados ao Problema de Estratificação. *Revista Brasileira de Estatatística*, 68, num. 229, p. 7-32, jul/dez 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*, 2010.

Brito, J. A. M, Uma Formulação de Programação Inteira para o Problema de Alocação Ótima em Amostras Estratificadas. In: Simpósio Brasileiro de Pesquisa Operacional - SOBRAPO, Gramado - RS. *Anais do XXXVII do SOBRAPO*, v. 1. p. 1851-1859, 2005.

Cochran, Willian G., Sampling Techniques, Third Edition – New York, John Wiley, 1977.

Cruz, M. D., O problema de clusterização automática, Tese de Doutorado, COPPE/UFRJ, 2010.

Dalenius, T. e Hodges, J. L. Jr , Minimum variance stratification. *Skandinavisk Aktuarietidskrift*, **54**, pp. 88-101, 1959.

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

Feo, T.A. and Resende, M.G.C., Greedy randomized adaptive search procedures, *Journal of Global Optimization*, **6**, pp. 109-133, 1995.

Gunning, P., Horgan, J., A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, **30**, 159–166, 2004.

Glover, F. and Kochenberger, G. A. , "*Handbook of Metaheuristic*", First Edition Norwell: K1uwer Academic Publishers, 2002.

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**, pp. 15–29, 2000.

Keskintürk T. And 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., N. N. and N., A., Determining the optimum strata boundary points using dynamic programming, *Survey Methodology*, **34**, 205–214, 2008.

Khan, M.G.M., Maiti, T., Ahsan, M.J., An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach, *Journal of Official Statistics*, **26**, num. 4, pp. 695-708, 2010.

Kozak, M., Optimal stratification using random search method in agricultural surveys, *Statistics in Transition*, **6**, pp. 797–806, 2004.

Lavallée, P. and Hidiroglou, M., On the stratification of skewed populations, *Survey Methodology* (Statistics Canada) 14, pp. 33–43, 1988.

Linguagem R (versão 2.13). (www.**r**-project.org/).

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

Prais M. and Ribeiro C.C., Reactive GRASP: An application to a matrix decomposition problem in TDMA traffic assignment. INFORMS Journal on Computing, 12:164–176, 2000.

Agradecimentos: À FAPERJ (projeto E-26/111.587/2010) e ao CNPQ (projeto 474051/2010-2) pelo financiamento parcial deste estudo.