ALGORITMOS DE OTIMIZAÇÃO APLICADOS AO PROBLEMA DE ESTRATIFICAÇÃO MULTIVARIADA

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

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

Luciana Roque Brito
UERJ/FFP
e-mail: luroque@hotmail.com

RESUMO

O presente trabalho traz uma proposta para a resolução do problema de estratificação multivariada, mediante a aplicação de dois algoritmos de otimização. O problema em questão consiste em dividir uma população composta por N unidades (objetos) em um conjunto de L subpopulações, chamadas de estratos (agrupamentos), tendo em conta um conjunto de $\boldsymbol{p}$ variáveis da pesquisa. De forma a produzir estratos mais homogêneos, os algoritmos visam a minimização de uma expressão de variância. Resultados computacionais obtidos a partir de um conjunto de dados reais são apresentados e discutidos.

PALAVRAS CHAVE. Estratificação, Metaheurísticas, Amostragem. MH.

ABSTRACT

This work intends to solve the multivariate stratification problem by using two optimization algorithms. The problem in question consists in divide a population composed by N objects into L sub-populations, called strata, taking in account a set of p study variables. In order to produce homogeneous strata, the algorithms aim to minimize a variance expression. Computational results obtained from real data are provided and discussed.

KEYWORDS. Stratification, Metaheuristics, Sampling. MH.

1. Introdução

O presente trabalho concentra-se, em particular, no estudo e resolução do problema de estratificação multivariada (Jarque, 1981 e Jewett and Judkins, 1988). Tal problema consiste em dividir uma população composta por N unidades (objetos) em um conjunto de L subpopulações, chamadas de estratos. A construção destes estratos é efetuada utilizando um conjunto de $\boldsymbol{p}$ variáveis da pesquisa, cujas observações são conhecidas para todas as unidades da população.

Para cada uma das $\boldsymbol{p}$ variáveis é calculada uma variância (Cochran, 1977), sendo a soma destas variâncias correspondente a uma variância total. Com a finalidade de produzir boas estimativas (Bolfarine e Bussab, 2005) a partir de amostras selecionadas nos estratos e reduzir os erros associados com tais estimativas, busca-se a minimização da variância total, o que resulta na construção de estratos mais homogêneos.

O problema de estratificação multivariada corresponde a um problema clássico de agrupamento (Kaufman e Rousseeuw**,** 1989) com uma função objetivo não linear. Assim como em outros problemas de agrupamento, a obtenção de um ótimo global, mediante a aplicação de algum método de enumeração, torna-se computacionalmente inviável à medida que o número de objetos a serem agrupados aumenta (Johnson A.R. e Wichern, 2002). Como alternativa, poder-se-ia utilizar um método clássico de agrupamento como ***k***-means (Späth, 1980), que produz soluções viáveis em um tempo computacional razoavelmente baixo e que trabalha com uma função objetivo similar à utilizada no problema de estratificação multivariada. Não obstante, tal método tende a produzir soluções viáveis de qualidade não muito boa, afastadas do ótimo global.

Dessa forma, objetivando produzir soluções de boa qualidade, e simultaneamente trabalhar com a função objetivo (variâncias) do problema de estratificação, propõe-se neste trabalho dois algoritmos, desenvolvidos a partir do estudo das metaheurísticas GRASP (Resende e Sousa, 2004) e Algoritmos Genéticos (Holland, 1975, Goldberg, 1989 e Linden, 2008).

Este trabalho está dividido da seguinte forma: Na seção seguinte, são apresentados de forma concisa os principais conceitos sobre amostragem. Na seção 3, apresenta-se o problema de estratificação multivariada. A seção 4 traz uma descrição dos procedimentos GRASP e Algoritmos Genéticos e apresenta detalhadamente os dois algoritmos propostos. A última seção traz um conjunto de resultados computacionais obtidos a partir da aplicação dos dois algoritmos, considerando um conjunto de dados do Censo Demográfico de 2000.

2. Conceitos sobre Amostragem

A amostragem faz parte não apenas de procedimentos técnicos, mas é realizada também, com freqüência, em eventos comuns da vida diária. Por exemplo, para avaliar o estado clínico de um paciente, um médico pode solicitar um exame de sangue (amostra). Por outro lado, para verificar o tempero de um alimento em preparação, podemos provar (observar) apenas uma porção. Neste caso, estaremos também realizando uma amostragem.

A partir da utilização da amostragem é possível obter informações sobre uma população investigando apenas uma parte dessa população, denominada amostra. A população é definida como um conjunto de elementos que possuem pelo menos um atributo (parâmetro) em comum na população. Os resultados observados na amostra constituem estimativas (valores aproximados) para os parâmetros de interesse.

Segundo Barbetta (2008), a utilização do levantamento por amostragem torna-se interessante pelos seguintes motivos:

- Economia: Em geral, é bem mais econômico a investigação de apenas uma parte da população.
- 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 que houvesse recursos financeiros em abundância.

- Confiabilidade dos dados: Quando se pesquisa um número reduzido de elementos, pode-se dar mais atenção aos casos individuais, evitando erros nas respostas.
- Operacionalidade: É mais fácil realizar operações de pequena escala. Um dos problemas típicos nos grandes censos (pesquisas de toda a população) é justamente o controle dos entrevistadores.

Considerando as informações que se pretende obter para a população, as restrições de custo da pesquisa e o nível de precisão com o qual se pretende trabalhar, pode-se optar por um dos seguintes esquemas clássicos de amostragem, quais sejam: amostragem aleatória simples, amostragem de conglomerados, amostragem sistemática e amostragem estratificada.

Neste trabalho, em particular, será feita uma descrição da amostragem estratificada, tendo em vista que tal esquema está associado ao problema em estudo.

2.1 Amostragem Estratificada

Entende-se por amostragem estratificada (Bolfarine e Bussab, 2005), o processo de dividir uma população U formada por N unidades (ou elementos) em L grupos mutuamente exclusivos, denominados estratos. Tal divisão é feita segundo alguma(s) característica(s) (variáveis) conhecida(s) na população. A estratificação é usada principalmente para resolver problemas como: melhoria da precisão das estimativas (menor erro padrão associado); produção de estimativas para a população e para subpopulações; por questões administrativas, etc.

Sendo $N_1, N_2, ..., N_h, ..., N_L$ o total de elementos (ou unidades) da população alocados a cada um dos estratos, tem-se naturalmente a seguinte equação:

$$N_1 + N_2 + ... + N_h + ... + N_L = N \quad . \tag{1}$$

Para que sejam obtidos todos os proveitos da estratificação, os valores N_h devem ser conhecidos. Definidos os estratos, a partir do conhecimento de uma ou mais características da população, seleciona-se uma amostra em cada um deles, sendo as seleções feitas de maneira independente em cada um dos estratos, considerando algum esquema de amostragem (amostragem aleatória simples, amostragem sistemática, etc).

Os tamanhos das amostras dentro dos estratos são denotados respectivamente por $n_1, n_2, ..., n_h, ..., n_L$, sendo a soma desses valores correspondente ao tamanho total n da amostra.

Apresenta-se a seguir a notação básica, estratificada utilizada neste trabalho, associada à amostragem estratificada.

N : Número total de unidades da população.

N_h : Número total de unidades da população em cada estrato *h*, considerando os L estratos.

n : Número total de unidades na amostra.

X_{hi} : Valor associado a uma variável X, para a *i-ésima* unidade do *h-ésimo* estrato, na população.

$$\overline{X_h} = \sum_{i=1}^{N_h} X_{hi} / N_h \quad : \text{ Valor médio de } X \text{ na população do } h\text{-}\acute{e}simo \text{ estrato} \tag{2}$$

$$S_{hx}^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^2}{N_h - 1} \quad : \text{ Medida da variância de } X \text{ na população do } h\text{-}\acute{e}simo \text{ estrato} \tag{3}$$

$$V_X = \sum_{h=1}^{L} N_h^2 . \frac{S_{hx}^2}{n_h} . (1 - \frac{n_h}{N_h}) : \text{ Variância do total populacional associada com uma variável X} \tag{4}$$

$$CV_X = \sqrt{V_X} / X : \text{ Coeficiente de Variação de uma variável X} \tag{5}$$

3. O Problema de Estratificação

Em tal problema (Cochran, 1977), definida uma amostra de tamanho n e um conjunto L de estratos $E_1, E_2, ..., E_L$, deve-se alocar cada uma das N observações (elementos) de uma população (unidades) definida por U aos L estratos considerando as seguintes restrições:

$$\bigcup_{h=1}^{L} E_h = U \tag{6}$$

$$E_h \neq \varnothing, 1 \leq h \leq L \tag{7}$$

$$E_i \cap E_j = \varnothing, 1 \leq i, j \leq L, i \neq j \tag{8}$$

Além disso, em cada um dos L estratos construídos em função da alocação, os elementos populacionais devem ser homogêneos. Inicialmente, considerando-se apenas uma variável de interesse X, ou seja, um problema de estratificação univariada, calcula-se a variância em cada um dos estratos através da equação (3) e toma-se a variância total dos estratos através da equação (4). Quanto menor for o valor de (4), mais homogêneos serão considerados os estratos.

Todavia, muitas das pesquisas realizadas por amostragem trabalham com um conjunto de variáveis, ou seja, o planejamento da pesquisa deve considerar a escolha de algumas potenciais variáveis de interesse. Enquanto cada variável separadamente pode ser útil para um certo propósito, muitas pesquisas têm propósitos múltiplos. Conseqüentemente, deve-se considerar algum esquema de estratificação multivariada que contemple tais propósitos.

Ademais, espera-se que, aplicando a estratificação multivariada, haja a possibilidade de se produzir estimativas mais precisas (com menor erro associado) para o conjunto de variáveis.

Dessa forma, na estratificação multivariada, os ***L*** estratos serão construídos levando em conta um conjunto de p variáveis denotadas por $X_1, X_2, ..., X_p$. De forma análoga ao caso univariado, avalia-se a homogeneidade dos estratos considerando a soma das variâncias (equação 4) das p variáveis:

$$V_T = V_{X_1} + V_{X_2} + ... + V_{X_p} = \sum_{h=1}^{L} N_h^2 . \frac{S_{hx_1}^2}{n_h} . (1 - \frac{n_h}{N_h}) + \sum_{h=1}^{L} N_h^2 . \frac{S_{hx_2}^2}{n_h} . (1 - \frac{n_h}{N_h}) + ... + \sum_{h=1}^{L} N_h^2 . \frac{S_{hx_p}^2}{n_h} . (1 - \frac{n_h}{N_h}) \tag{9}$$

Observa-se que os valores de N_h e $S_{hx_i}^2$ são definidos em função da distribuição das unidades populacionais pelos estratos, enquanto que o tamanho n da amostra é fixado previamente. Todavia, o valor obtido para no cálculo da variância (equação 9) depende, também, do critério adotado para definir o tamanho de amostra n_h que será alocado a cada um dos estratos, dado o valor de n. Para a determinação de n_h, necessário para o cálculo da variância, pode-se utilizar a equação abaixo:

$$n_h = \frac{n.N_h}{N} \tag{10}$$

A equação (10) está associada à alocação *Proporcional*. Nesse tipo de alocação, uma amostra de tamanho n (autoponderada) é distribuída proporcionalmente ao tamanho dos estratos. Quando há necessidade de se fazer várias estimativas, tal amostra poupa tempo.

Em particular, substituindo-se os valores de n_h na equação (9) pela equação (10), para cada uma das p variáveis, têm-se, respectivamente, a variância total (V_T) (equação 11), a variância por estrato (V_{T_h}) (equação 12) e o coeficiente de variação total (equação 13):

$$V_T = \frac{N-n}{n} (\sum_{h=1}^{L} N_h . S_{hx_1}^2 + \sum_{h=1}^{L} N_h . S_{hx_2}^2 + ... + \sum_{h=1}^{L} N_h . S_{hx_p}^2) \tag{11}$$

$$V_{T_h} = \frac{N-n}{n}(N_h.S^2_{hx_1} + N_h.S^2_{hx_2} + ... + N_h.S^2_{hx_p}) \quad (12)$$

$$CV_T = \frac{\sqrt{\frac{N-n}{n}(\sum_{h=1}^{L} N_h.S^2_{hx_1})}}{X_1} + \frac{\sqrt{\frac{N-n}{n}(\sum_{h=1}^{L} N_h.S^2_{hx_2})}}{X_2} + ... + \frac{\sqrt{\frac{N-n}{n}(\sum_{h=1}^{L} N_h.S^2_{hx_p})}}{X_p} \quad (13)$$

A partir dessa descrição, pode-se observar que o problema de estratificação multivariada corresponde a um problema de agrupamento sem restrições e que agrega uma particular função objetivo.

É fato conhecido, na literatura de *Cluster Analysis* (Johnson A.R. e Wichern, 2002), que a resolução exata de tal problema, ou seja, a obtenção do ótimo global, torna-se impraticável à medida que se aumenta o número de grupos e o número de objetos a serem agrupados.

Na verdade, caso fosse aplicado algum processo de busca exaustiva para garantir a obtenção da solução global, seria necessário enumerar todas as soluções, ou seja, todas as possibilidades de combinação de N elementos em L estratos. O número de possibilidades, neste caso, está associado ao número de Stirling de segundo tipo (ver Johnson e Wichern, 2002), dado por $\frac{1}{k!}\sum_{j=0}^{k}(-1)^{L-j}\binom{k}{j}j^N$. Caso se tenha, por exemplo, N = 16 elementos a serem alocados em 2 estratos, o número de soluções a serem consideradas é de 32.767. Para o caso de 3 estratos, este número cresce para 7.141.686 soluções possíveis. Considerando-se um número N maior de elementos, estes valores crescem exponencialmente.

Alternativamente, poder-se-ia utilizar um método clássico de agrupamento não hierárquico como o ***k-médias*** (Späth, 1980), que produz soluções viáveis em um tempo computacional razoavelmente baixo e que trabalha com uma função objetivo similar à utilizada no problema de estratificação.

Existem também trabalhos na literatura que resolvem especificamente esse problema, considerando a aplicação de técnicas de análise multivariada.

Hagood *et al*. (1945) faz o uso de análise de componentes principais para definir estratos mais homogêneos em um problema de estratificação multivariada. Jarque (1981) resolve o problema de estratificação multivariada combinando um método de estratificação univariada proposto por Dalenius (1957) com o método das ***k-médias***. Ainda considerando a aplicação desse método, Jewett e Judkins (1988) também resolveram um problema de estratificação multivariada com restrições. Skinner et al. (1994) propuseram um método de estratificação considerando duas ou três variáveis de interesse.

Não obstante, tais métodos podem produzir soluções viáveis de baixa qualidade, isto é, afastadas do ótimo global. Sendo assim, de forma a contornar a complexidade intrínseca ao problema de estratificação multivariada, possibilitando a obtenção de soluções de boa qualidade em um tempo computacional viável, propõe-se no presente trabalho dois algoritmos baseados nas metaheurísticas GRASP (Resende e Sousa, 2004) e Algoritmos Genéticos (Linden, 2008, Goldberg, 1989 e Holland, 1975).

4.Metodologia Proposta

A presente seção traz uma descrição concisa das metaheurísticas GRASP e Algoritmos Genéticos e, em seguida, os detalhes da implementação dos dois algoritmos ao problema tratado neste artigo.

4.1 GRASP

A metaheurística GRASP (Resende e Sousa, 2004) é um procedimento *multistart* no qual cada iteração consiste de duas fases: construção e busca local. Na fase de construção, produz-se uma solução viável x_o. Na segunda fase, de busca local, investiga-se uma vizinhança de x_o até que seja encontrada uma solução viável de qualidade superior a x_o. O resultado do GRASP

é dado pela melhor solução produzida após um número *n* predeterminado de iterações nas quais são aplicados os procedimentos de construção e de busca local.

Na fase de construção, uma solução é iterativamente construída, elemento por elemento. Em cada iteração dessa fase, define-se uma lista de candidatos (*LC*) formada por todos os elementos que podem ser incorporados na solução parcial x_o sem provocar a inviabilidade do problema. 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 $t \in LC$ em x_o.

Para conferir maior variabilidade às soluções obtidas na fase de construção, define-se uma lista de candidatos restrita (*LCR*). A *LCR* é formada pelos melhores elementos avaliados na *LC* através da função ***g(.)***, quais sejam, aqueles que, quando incorporados na solução x_o em construção, produzem um acréscimo mínimo (caso de minimização). Esse processo representa o aspecto guloso do GRASP, pois têm-se sempre os melhores elementos para serem incorporados na solução. Em Resende e Feo (1995) são propostos dois 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 $g_{\min}$ e $g_{\max}$ o menor e o maior acréscimos provocados pela inserção de um elemento $t \in LC$ na solução, segundo a função ***g(.)***. A partir da utilização dessa função e dos valores mínimo e máximo de ***g(i)***, a *LCR* é construída considerando a seguinte expressão:

$$LCR = \{t \in LC \mid g(t) \leq g_{\min} + \alpha(g_{\max} - g_{\min}), \alpha \in [0,1]\} \quad (14)$$

Na fase de busca local, há uma tentativa de se melhorar a solução x_o produzida na fase de construção, tendo em vista que esta ainda não é necessariamente uma solução de boa qualidade para o problema. A busca local consiste na substituição da solução x_o pela melhor solução x^* encontrada na vizinhança de x_o. A solução obtida a partir do GRASP será a melhor das soluções obtidas em todas as iterações.

Atualmente, de forma a melhorar a qualidade da soluções produzidas pelo GRASP, tem-se incorporado à fase de busca local um procedimento denominado *path-relinking*. O *path-relinking* é um procedimento que surgiu como uma estratégia de intensificação de busca para melhorar a qualidade das soluções obtidas a partir da utilização de outras metaheurísticas, sendo inicialmente proposto por Glover *et al* (2000). Tal método é uma generalização da estratégia denominada *Scatter Search*. A idéia principal por trás do *path-relinking* é: dado um conjunto de soluções de melhor qualidade, denominadas soluções elite (construído durante a aplicação da busca local), deve-se construir "caminhos" entre duas delas de forma que seja possível avaliar as soluções intermediárias pertencentes a tais caminhos, pois espera-se que algumas dessas soluções intermediárias possam ser melhores que as soluções armazenadas no conjunto elite. Com isso, o *path-relinking* procura funcionar como uma fase de intensificação sobre as soluções obtidas na fase de busca local, melhorando a qualidade dessas soluções.

4.2 Algoritmos Genéticos

Os algoritmos genéticos (AG) (Linden, 2008) são um ramo dos algoritmos evolucionários, e como tal, podem ser definidos como uma técnica de busca baseada na teoria biológica de evolução natural. Os algoritmos genéticos, criados por Holland em 1975, são técnicas heurísticas de uso geral para otimização global, diferenciando-se de métodos tradicionais de busca local, como o gradiente, que utiliza-se da informação derivada da função de forma a atingir o máximo. Tais métodos ficam facilmente retidos em pontos de máximo (ou mínimo) local que muitas vezes são de qualidade ruim.

Um algoritmo genético utiliza uma população inicial de cromossomos (soluções viáveis para o problema) gerados aleatoriamente, faz a avaliação de cada um, seleciona os melhores e promove manipulações genéticas, como cruzamento e mutação, com a finalidade de criar uma nova população. Este processo pode ser usado para resolver qualquer problema de otimização.

Primeiramente, deve-se determinar como os cromossomos serão representados, o que comumente é feito através de *strings* ou vetores. A população inicial pode ser gerada aleatoriamente ou utilizando alguma heurística de construção, produzindo *q* configurações (C_1 , C_2 , ..., C_q), onde *q* é o tamanho da população, que representa um pequeno subconjunto do

espaço total de configurações, ou seja, de soluções viáveis do problema. Para cada cromossomo na população, deve-se obter $f_i = f(C_i), \forall i = 1,...,q$, que é a avaliação da função objetivo. Finalmente, operadores genéticos devem ser aplicados à população na seguinte ordem:

(1) Reprodução dos melhores cromossomos (*seleção*). Com base nos valores f_i, as melhores seqüências são selecionadas e duplicadas em substituição às piores. Neste passo, pode-se considerar, por exemplo, o chamado método da roleta viciada ou do torneio, que consiste em fazer com que os cromossomos com melhor aptidão (valor da função objetivo) sejam copiados um número maior de vezes do que aqueles que não possuem uma possibilidade de seleção tão alta. Com isso, os cromossomos com baixa probabilidade de seleção tendem a ser eliminados, pois é pouco provável que alguma cópia destes cromossomos seja realizada, visto ser o número total de cópias igual ao tamanho da população.
(2) Recombinação das estruturas genéticas da população, ou seja, o cruzamento (*crossover*). Esta operação possibilita a diversificação no espaço das soluções. O operador *crossover* (tradicional) escolhe aleatoriamente dois cromossomos e troca partes do seu padrão genético.
(3) Perturbação de uma configuração (*mutação*), isto é, a modificação de partes (genes) de alguns cromossomos. Este procedimento permite a regeneração de algum gene que possa ter sido eliminado da população, de forma inesperada, em alguma etapa do processo.

Após a execução de todos esses procedimentos, obtém-se uma nova população. Sobre esta, deve-se repetir a avaliação da função objetivo para cada cromossomo e a aplicação dos operadores genéticos, e assim sucessivamente, em um processo iterativo de geração de novas populações. Normalmente, são considerados os seguintes critérios de parada nesse algoritmo: (1) Número máximo gerações - Repetir os passos (1, 2, 3) por um número de vezes pré-determinado; (2) Tempo máximo de processamento; (3) Número máximo iterações em que não há melhoria no valor da função objetivo.

A saída é representada pela última população gerada, da qual o melhor elemento, isto é, o cromossomo que tem associado o menor (maior) valor da função objetivo no problema de minimização (maximização) é usualmente escolhido como a solução do problema.

4.3 Algoritmos Propostos

Neste trabalho, ambos os algoritmos GRASP e AG utilizaram como função objetivo a equação (13). Além disso, cada um desses algoritmos produz como solução ***L*** estratos, formados por conjuntos de setores censitários. Em linhas gerais, os procedimentos implementados nesses dois algoritmos efetuam vários movimentos de troca de setores entre os estratos, objetivando uma redução da função objetivo e a conseqüente obtenção de estratos mais homogêneos.

4.3.1 Algoritmo Grasp

Procedimento de Construção:

(1) Selecionar aleatoriamente ***L*** setores dentre os ***N*** setores disponíveis.
(2) Alocar cada um dos ***L*** setores aos estratos $E_1, E_2,..., E_h,...,E_L$.
(3) Definir a lista de candidatos ***LC*** a partir dos ***(N-L)*** setores restantes.
Enquanto (LC ≠ ∅) Faça
(4) Selecionar aleatoriamente um dos ***L*** estratos, definido por E_h.
(5) Calcular o aumento da variância (equação (12)) em E_h, considerando a inclusão de cada um dos setores que compõem a ***LC*** no estrato selecionado (corresponderá à função ***g(i)***).
(6) Determinar g_{min} e g_{max}.
(7) Inserir na ***LCR*** todos os setores ***i*** da ***LC*** tais que $g(i) \leq g_{min} + 0.5\,(g_{max} - g_{min})$.
(8) Selecionar aleatoriamente um setor da ***LCR*** e incluí-lo no estrato E_h.
(9) Atualizar a ***LC*** excluindo o setor que foi adicionado no estrato E_h.
Fim-Enquanto

Os passos de (4) a (9) são repetidos até que todos os setores censitários tenham sido alocados a um dos ***L*** estratos $E_1, E_2,..., E_h,...,E_L$. Ainda considerando o procedimento de construção, introduziu-se um procedimento de filtro. Ou seja, em vez de se construir apenas uma solução inicial, são construídas *p* soluções iniciais, selecionando-se a de melhor qualidade para a aplicação da busca local. A aplicação sistemática de tal procedimento tende a eliminar de soluções iniciais de baixa qualidade.

Procedimento de Busca Local:

Buscas=0 e Sbest (melhor solução na construção)

Enquanto (Buscas < Tot_Buscas) Faça

(1) Selecione aleatoriamente dentre os $\boldsymbol{L}$ estratos um estrato $\boldsymbol{E_i}$ e um estrato $\boldsymbol{E_j}$.

(2) $V_{best}=V_{Ti}+V_{Tj}$ (equação 12), Melhorias=0 (V_{Ti} e V_{Tj} estão associadas com Sbest) , t=0

Enquanto (Melhorias < Tot_Melhorias) E (t<Tentativas) Faça

(3) Selecionar aleatoriamente de $\boldsymbol{E_j}$ um setor $\boldsymbol{s_k}$ **(k=1,...,|$\boldsymbol{E_j}$|) e** selecionar aleatoriamente de $\boldsymbol{E_i}$ um setor $\boldsymbol{s_r}$ **(r=1,...,|$\boldsymbol{E_i}$|).**

(4) Atribuir o setor $\boldsymbol{s_r}$ ao estrato $\boldsymbol{E_j}$ e atribuir o setor $\boldsymbol{s_k}$ ao estrato $\boldsymbol{E_i}$

(5) Recalcular a variância utilizando a equação (12) para os estratos E_i e E_j, obtendo $V_{Ti}+V_{Tj}$

(6) **Se** ($V_{Ti}+V_{Tj} < V_{best}$) **Então**

(7) Guarda Nova_Solução e Retorna à Solução Anterior

(8) $V_{best} = V_{Ti}+V_{Tj}$, Melhorias=Melhorias + 1

Fim-se

(9) Atribuir o setor $\boldsymbol{s_r}$ ao estrato $\boldsymbol{E_j}$

(10) Recalcular a variância para o estrato $\boldsymbol{E_j}$, obtendo V_{Tj}

(11) **Se** ($V_{Ti}+V_{Tj} < V_{best}$) **Então**

(12) Guarda Nova_Solução e Retorna à Solução Anterior

(13) $V_{best} = V_{Ti}+V_{Tj}$, Melhorias=Melhorias + 1

Fim-se

(14) Atribuir o setor $\boldsymbol{s_k}$ ao estrato $\boldsymbol{E_i}$

(15) Recalcular a variância para o estrato $\boldsymbol{E_i}$, obtendo V_{Ti}

(16) **Se** ($V_{Ti}+V_{Tj} < V_{best}$) **Então**

(17) Guarda Nova_Solução e Retorna à Solução Anterior

(18) $V_{best} = V_{Ti}+V_{Tj}$, Melhorias=Melhorias + 1

Fim-se

(19) t=t+1

Fim-Enquanto

(20) Atualizar S_{best} com a Nova_Solução

Fim-Enquanto

Em cada iteração do "*loop*" mais interno (passos de (3) a (19)), são aplicados dois tipos de movimentos de troca. A idéia é considerar o movimento que produz a maior redução possível no valor da função objetivo. Após um certo número tentativas ou melhorias ser atingido, atualiza-se (passo (20)) a solução S_{best} com a Nova_Solução obtida após as trocas.

Procedimento de Path-Relinking:

A aplicação desse procedimento considera a combinação de cada solução $\boldsymbol{S_{elite}}$ de um conjunto elite (atualizado após a aplicação da busca local) com a melhor solução $\boldsymbol{S_{best}}$ obtida até o momento. Em cada movimento desse procedimento, substitui-se cada um dos setores dos estratos definidos em $\boldsymbol{S_{best}}$ pelos setores dos estratos definidos em $\boldsymbol{S_{elite}}$. Por conseguinte, a cada substituição, produz-se uma solução intermediária $\boldsymbol{S_i}$ entre $\boldsymbol{S_{elite}}$ e $\boldsymbol{S_{best}}$. Ao final das substituições, toma-se a melhor solução intermediária $\boldsymbol{S_i}$ obtida, caso ela seja melhor do que a solução atual. Inicialmente, define-se a solução intermediária inicial $\boldsymbol{S_i}$ como a própria solução $\boldsymbol{S_{best}}$.

Ao substituir-se um setor s_a pertencente a um estrato $\boldsymbol{E_h}$ da solução $\boldsymbol{S_i}$ por um setor $\boldsymbol{s_b}$ pertencente a um estrato $\boldsymbol{E_h}$ de solução $\boldsymbol{S_{elite}}$, verifica-se, previamente, se s_b já não faz parte do estrato $\boldsymbol{E_h}$ de $\boldsymbol{S_i}$. Caso $\boldsymbol{s_b}$ não pertença ao estrato $\boldsymbol{E_h}$ de $\boldsymbol{S_i}$, ou seja, $\boldsymbol{s_b}$ está em outro estrato $\boldsymbol{E_l}$ de $\boldsymbol{S_i}$, efetua-se a substituição de $\boldsymbol{s_a}$ por $\boldsymbol{s_b}$.

Todavia, efetuando-se apenas essa substituição, o setor $\boldsymbol{s_a}$ ficaria automaticamente fora da solução (não associado a nenhum estrato), produzindo uma solução não viável. Então, de forma a garantir que a cada movimento do *path-relinking* a solução continue viável, busca-se o setor $\boldsymbol{s_b}$ em $\boldsymbol{E_l}$, substituindo esse setor pelo pelo setor $\boldsymbol{s_a}$.

Considerando o exemplo abaixo, com ***L***=2 e ***N***=9, e aplicando o procedimento de *path-relinking* no 1º estrato (***h=1***) ocorrem as seguintes substituições:

	Estrato 1				Estrato 2				
S_{elite}	5	6	1	2	4	7	3	8	9
Si (inicial) = S_{best}	4	9	2	7	1	5	3	8	6
Si (1)	5	9	2	7	1	4	3	8	6
Si (2)	5	6	2	7	1	4	3	8	9
Si (3)	5	6	1	7	2	4	3	8	9
Si=S_{elite}	5	6	1	2	7	4	3	8	9

Tabela 1 – Aplicação do Path Relinking

- O setor 5 (s_b) de S_{elite} substitui o setor 4 (s_a) de S_i no 1º estrato (E_h). Em seguida, substitui-se 5 (s_b) no 2º estrato (E_l) de S_i por 4 (s_a), produzindo a primeira solução intermediária .
- O setor 6 (s_b) de S_{elite} substitui o setor 9 (s_a) de S_i no 1º estrato. Em seguida, substitui-se o setor 6 (s_b) no 2º estrato de S_i pelo setor 9 (s_a), produzindo a segunda solução intermediária.
- O setor 1 (s_b) de S_{elite} substitui o setor 2 (s_a) de S_i no 1º estrato. Em seguida, substitui-se 1 (s_b) no 2º estrato de S_i por 2 (s_a), produzindo a terceira e última solução intermediária tal que $S_i \neq S_{elite}$.

Tal procedimento será aplicado em cada um dos estratos, com a expectativa de produzir soluções intermediárias melhores do que a solução atual.

4.3.2 Algoritmo Genético

Procedimento de Geração:

É construída uma população formada por ***q*** vetores (soluções viáveis) com ***N*** posições cada, correspondentes aos setores censitários que serão alocados aos ***L*** estratos. Em seguida, atribui-se a cada uma das posições desses vetores um valor aleatório entre 1 e ***L***. Tal valor define a qual estrato o setor será inicialmente alocado. Após a geração da população, calcula-se o valor da função objetivo $f_i, i = 1,...,q$ utilizando-se a equação (11).

Procedimento de Cruzamento:

Inicialmente, seleciona-se um valor real aleatório no intervalo (0,1) para cada par de cromossomos C_i e C_j $(i = 2k-1, j = 2k, k = 1,..., N/2)$ da população. Caso esse valor seja inferior ao valor de uma probabilidade (p_{cruza}) de cruzamento definida previamente, aplica-se o cruzamento em C_i e C_j , considerando dois pontos de corte. Ou seja, são selecionados dois valores aleatórios entre 1 e ***N***, verificando-se o mínimo e o máximo entre esses valores. Tais valores corresponderão aos pontos inicial (p_i) e final (p_f) de corte. Dessa forma, trocar um conjunto de valores entre C_i e C_j em um intervalo entre (p_i) e (p_f) implica em mudar os setores de estratos. A figura abaixo ilustra a aplicação desse procedimento, considerando ***L***=3, ***N***=10, p_i=4 e p_f=7.

C_i (antes)	1	3	3	2	1	2	2	3	1	1	C_i (depois)	1	3	3	2	3	1	1	3	1	1
C_j (antes)	2	1	3	2	3	1	1	2	2	3	C_j (depois)	2	1	3	2	1	2	2	2	2	3

Tabela 2 – Aplicação do Cruzamento de Dois Pontos

Procedimento de Mutação:

Definida uma probabilidade de mutação (p_{mut}), serão selecionados ***s***=(***N****q*p_{mut}), valores entre 1 e ***N*q*** , ou seja, as posições dos ***q*** vetores que terão seus valores substituídos por um valor entre 1 e ***L*** (excluindo o próprio valor original). Isso corresponde a colocar os cromossomos em fila e selecionar ***s*** valores aleatoriamente. Considere, por exemplo, que tenham sido gerados ***q***=10 cromossomos, com 10 posições cada (10 setores por cromossomo). Para um percentual de mutação de 5% (p_{mut}=0.05), devem ser selecionados ***s=5*** posições (gens) entre 1 e 100. Supondo que os valores tenham sido 8, 21, 29, 47, 92 temos que o primeiro cromossomo terá o valor da sua oitava posição alterado, o terceiro cromossomo terá os valores da 1ª e 9ª posições alterados e assim sucessivamente.

Espera-se, naturalmente, que nas primeiras gerações do algoritmo genético haja uma necessidade de realizar muitos cruzamentos e poucas mutações na população, tendo em vista a grande diversidade genética e a necessidade de explorar o espaço de soluções o máximo possível.

Dessa forma, foram consideradas, no presente trabalho, probabilidades de cruzamento e de mutação variáveis. De acordo com Linden (2008), precisa-se, então, que a probabilidade do

operador de cruzamento decaia progressivamente, ao passo que, concomitantemente, a probabilidade do operador de mutação aumente. Para tanto, pode-se interpolar as probabilidades dos dois operadores de forma a produzir tal efeito. Um exemplo seria definir a probabilidade de cruzamento inicial como 80% e fazê-la decair no decorrer das gerações até atingir o valor de 20%. Simultaneamente, poder-se-ia iniciar com uma probabilidade de mutação da ordem 1% e aumentá-la durante as gerações até atingir o valor de 5%.

Procedimento de Reprodução:

Para reprodução da população, optou-se pela aplicação do método do torneio (Linden, 2008). Isto é, são selecionados aleatoriamente ***t*** cromossomos da população (***t<q***), e em seguida, seleciona-se o cromossomo com o menor valor da função objetivo (equação 11). Considerando que em cada seleção é escolhido um cromossomo para compor a nova população, tal procedimento foi aplicado ***q*** vezes.

Além disso, de forma a garantir que a melhor solução (cromossomo) da população não seja perdida a partir da aplicação do método do torneio, adotou-se uma estratégia de elitismo, qual seja, após construção da nova população, verifica-se se a melhor solução faz parte da população. Caso tal solução não esteja na população, seleciona-se aleatoriamente um cromossomo entre o primeiro e o ***N***-ésimo cromossomo e efetua-se a sua substituição pelo melhor cromossomo.

5. Resultados Computacionais

A presente seção contém os resultados de alguns experimentos computacionais obtidos a partir da aplicação dos algoritmos propostos. Mais especificamente, de forma a avaliar a eficiência desses algoritmos, no que concerne à homogeneidade dos estratos construídos, foram selecionados 20 municípios do Brasil (instâncias), considerando os dados da amostra do censo demográfico de 2000. Em seguida, para um dos setores censitários pertencentes a esses municípios, foram disponibilizadas os valores das seguintes variáveis: média dos anos de estudo dos responsáveis pelos domicílios, média do rendimento nominal dos responsáveis pelos domicílios e número médio de banheiros. Tais variáveis foram utilizadas para o cálculo das variâncias (equação 11) no problema de estratificação multivariada.

Acrescenta-se, ainda, que todos os experimentos foram realizados em computador *DELL*, dotado de dois processadores de 3.33 GHz e com 4MB de memória de RAM. Os dois algoritmos foram implementados utilizando a linguagem DELPHI.

5.1 Resultados

Para a execução do ***GRASP***, isto é, a aplicação dos procedimentos de construção e de busca local, o total de iterações foi fixado em 200. O número soluções iniciais geradas na fase de construção foi igual a oito (número de filtros). O total de buscas (***Tot_Buscas***) foi fixado em 200, o número de melhorias (***Tot_Melhorias***) foi fixado em 8 e o número de tentativas (***Tentativas***) foi fixado em 100. O procedimento de *Path Relinking* foi aplicado a cada 15 iterações do ***GRASP***, com o total de soluções do conjunto elite igual a 10.

Em relação ao algoritmo genético, definiu-se o tamanho da população como sendo igual a 100, o número de gerações igual a 300 e o total de iterações sem melhoria igual a 100. Caso o algoritmo alcance 100 iterações sem produzir uma solução melhor que a melhor solução atual, o algoritmo termina. As probabilidades iniciais de cruzamento e de mutação foram respectivamente de 0.85 e 0.01, e as probabilidades finais foram de 0.60 e 0.05. Todos esses parâmetros foram ajustados a partir de experimentos previamente realizados em um subconjunto de municípios

A seguir, apresenta-se uma tabela com os resultados obtidos para os 20 municípios, mediante a aplicação dos algoritmos. Nesse experimento, o número de estratos **(*L*)** variou entre três e cinco, o que implicou, por sua vez, em seis possíveis estratificações (três do ***GRASP*** e três do ***AG***) para cada um dos municípios.

As colunas 1 e 2 apresentam, para cada município, respectivamente, o seu nome e o número de setores censitários (número de objetos a serem agrupados). Em seguida, na terceira coluna, tem-se o tamanho de amostra considerado para a realização da alocação proporcional. As colunas de 4 a 6 (três estratos), 7 a 9 (quatro estratos) e 10 a 11 (cinco estratos) contêm a o valor

da melhor solução produzida pelo ***GRASP***, o valor da melhor solução produzida pelo ***AG*** e o ***gap*** entre a solução do ***GRASP*** e do ***AG***. As melhores soluções encontradas para cada instância, considerando L=3,...,5 estão assinaladas em uma caixa cinza.

Seguindo o padrão dos trabalhos que tratam sobre a problemática de estratificação, a medida da homogeneidade (valor da função objetivo) dos estratos é apresentada em função do coeficiente de variação (equação 13), que é uma medida de precisão admensional mais fácil de ser compreendida.

A partir de uma análise inicial da tabela, pode-se observar que em 77% dos casos (46 dos 60 resultados) o ***GRASP*** produziu soluções melhores do que as do ***AG***. Além disso, os gaps médio e mediano do ***GRASP*** em relação ao ***AG*** (46 instâncias onde ***GRASP*** foi superior) foram de respectivamente: 22,64% e 32,75% (3 estratos), 23,66% e 30,75% (4 estratos) e 21,12% e 29,90% (5 estratos). Observa-se, ainda, que os tempos de processamento dos dois algoritmos foram bem similares, variando de 10 segundos (instâncias menores) até 2 horas (instâncias maiores). Tais resultados indicam que a utilização de metaheurísticas pode ser uma boa alternativa para a resolução do problema de estratificação multivariada.

Tabela 1 – Resultados do ***Grasp*** e do ***AG***

Instância (Município)	Setores (N)	n	Nº Estratos								
			3			4			5		
			Grasp	AG	Gap	Grasp	AG	Gap	Grasp	AG	Gap
Barra Mansa	258	100	**7,23**	8,58	18,67	**6,46**	7,28	12,55	5,62	**5,53**	1,72
Bauru	457	150	**6,50**	7,97	22,48	**5,87**	7,28	24,14	**5,00**	5,04	0,76
Belo Horizonte	2551	400	**5,02**	7,04	40,13	**4,24**	6,89	62,43	**3,77**	6,68	77,09
Campinas	1266	200	**6,63**	8,26	24,66	**5,92**	7,63	28,85	**5,50**	6,97	26,73
Cariacica	291	120	**3,08**	7,83	154,34	**2,80**	6,54	133,99	**2,32**	4,26	83,98
Contagem	589	200	**3,20**	3,78	18,19	**2,94**	3,47	17,85	**2,80**	3,34	19,06
Duque Caxias	1041	300	**2,92**	3,52	20,58	**2,74**	3,17	15,60	**2,59**	3,20	23,43
Florianópolis	423	150	**5,49**	6,74	22,80	**4,96**	5,37	8,23	**4,74**	5,15	8,59
Magé	221	100	**3,93**	4,36	10,73	**3,72**	3,89	4,62	**3,49**	3,77	8,06
Niterói	696	200	3,69	**3,61**	2,41	3,34	**3,10**	7,71	3,03	**2,79**	8,34
Ponta Grossa	325	100	**7,29**	8,89	21,89	**6,87**	7,56	10,14	**6,08**	6,41	5,33
Porto Alegre	2079	400	**3,67**	5,41	47,41	**3,31**	4,68	41,26	**3,06**	5,02	64,27
Rio Claro	197	50	7,32	**6,33**	15,63	6,00	**5,22**	14,95	5,90	**5,14**	14,75
Santo André	683	200	**4,93**	6,23	26,34	**4,48**	5,52	23,18	**4,20**	4,87	15,99
Santos	594	200	3,76	**3,64**	3,35	**3,29**	3,37	2,46	**3,02**	3,33	9,99
São Carlos	243	80	**7,30**	8,18	12,08	6,52	**5,57**	17,09	5,72	**4,73**	21,12
São Gonçalo	1214	300	**2,46**	3,37	36,79	**2,26**	3,23	42,95	**2,13**	3,10	45,08
São José Cam.	731	200	**6,74**	8,60	27,51	**6,23**	8,38	34,44	**5,76**	8,01	39,02
Vitória	268	100	6,87	**6,30**	8,93	5,93	**5,50**	7,81	5,44	**5,00**	8,90
Volta Redonda	414	100	**7,89**	9,43	19,48	**6,52**	8,44	29,35	5,77	**5,19**	11,10

A análise do gráfico 1 (*box plot*) reforça a observação de que o ***GRASP*** produziu estratos mais homogêneos do que o algoritmo genético. Além disso, os maiores ganhos de precisão (maior redução da função objetivo) foram obtidos com número de estratos igual a cinco.

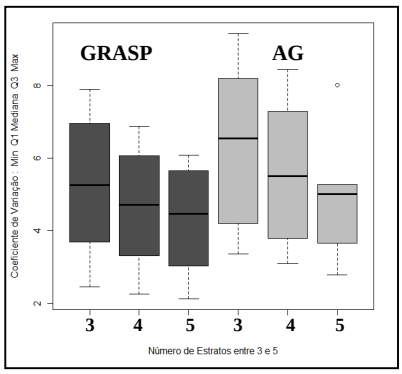

Gráfico 1 – Cv´s do GRASP e do AG

Uma possível explicação para a diferença de performance do ***GRASP*** em relação ao ***AG***, no que concerne à qualidade das soluções produzidas, reside na combinação do procedimento de busca local com o *path relinking* no ***GRASP***, o que contrasta com a utilização apenas de procedimentos ortodoxos de cruzamento e de mutação no ***AG***. Isso sugere que, para tornar o AG mais competitivo, uma possibilidade a ser considerada é a implementação de procedimentos de cruzamento e mutação mais sofisticados.

Em trabalhos futuros, pretende-se avaliar a eficiência dos algoritmos GRASP e Genético considerando novas variáveis de estudo e outros esquemas de alocação (*Neyman* e Potência), além da alocação proporcional, o que implicará, por sua vez, em minimizar outras expressões de variância. Pretende-se, também, implementar procedimentos de busca local (GRASP) e de cruzamento e mutação (AG) mais sofisticados, objetivando ganhos de precisão (redução da função objetivo) ainda maiores que os obtidos nesse estudo. Pretende-se, também, comparar os estratos produzidos por esses algoritmos com os estratos produzidos pelo algoritmo das ***k-médias***.

Bibliografia

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.

Cochran, Willian G., Sampling Techniques, Third Edition – New York, John Wiley, 1997.

Jarque, C.M. (1981), A Solution to the Problem of Optimum Stratification in Multivariate Sampling. *Applied Statistics*, vol 30, n. 2 , 163-169.

Jewett R.S. e Judkins D.R. (1988), Multivariate Stratification with Size Constraints. *Siam Journal of Science Statistical Computing*, vol 9, n. 6, 1091-1097.

Johnson A.R. e Wichern D.W, *Applied Multivariate Statistical Analysis*, Prentice Hall, Fifth Edition, 2002.

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

Linden, R, Algoritmos Genéticos – Uma Importante Ferramenta da Inteligência Computacional, Editora Brasport, 2008.

Goldberg, D. E, *Genetic Algorithms in Search, Optimization and Machine Learning*, Boston, MA: Addison-Wesley, 1989.

F. Glover, M. Laguna, and R. Martí (2000). "Fundamentals of Scatter Search and Path Relinking", *Control and Cybernetics*, vol. 29, n.3, pp. 653–684.

Haggod, M.J. and Bernert, E. H. (1945). Component Indexes as a Basis for Stratification in Sampling, Journal of American Statistical Association 40, 330-341.

Holland, J. H, Adaptation in natural and artificial systems. University of Michigan Press, Ann Arbor, 1975.

Resende, M.G.C e Sousa, J.P., *Metaheuristics : Computer Decision-Making*. Kluwer Academic Publishers, 2004.

Resende, M.G.C., Feo, T. A. (1995), Greedy Randomized Adaptative Search Procedures, *Journal of Global Optimization*, pgs. 1-27.

Skinner, C.J., Holmes, D.J. and Holt, D. (1994), Multiple Frame Sampling for Multivariate Stratification, International Statistical Review, 62, 3, 333-347.

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