UM ALGORITMO GENÉTICO APLICADO AO PROBLEMA DE ESTRATIFICAÇÃO UNIVARIADA

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

Flávio Marcelo Tavares Montenegro
Instituto Brasileiro de Geografia e Estatística - IBGE
Av Chile, número 500, 10º andar, centro, Rio de Janeiro – RJ
e-mail: fmtmontenegro@gmail.com

Resumo

O presente trabalho traz uma proposta de resolução para o problema de estratificação univariada. Neste problema, devem ser construídos ***L*** estratos (grupos) populacionais de forma que a soma das variâncias seja a menor possível. Tal problema é de grande importância na área de amostragem probabilística e tem elevada complexidade, sendo esta decorrente da sua associação com um problema agrupamento. De modo a resolver o problema e produzir soluções de boa qualidade em um tempo computacional viável, é proposto um algoritmo de estratificação com base na metaheurística algoritmos genéticos. Ao final, são apresentados resultados computacionais obtidos a partir da aplicação do algoritmo a um conjunto de dados reais.

Palavras-Chave: Amostragem; Estratificação; Análise de Agrupamentos; Metaheurísticas; Algoritmos Genéticos;

Abstract

This paper presents a proposal to solve the stratification problem. This problem consists in to construct L populational strata (groups) such that the sum of the variances is the lowest possible. It is a very important problem in probabilistic sampling. Also, it has high complexity, because of its association with a clustering problem. In order to produce good quality solutions for this problem in feasible computational time, a stratification algorithm based on a genetic algorithms metaheuristic is proposed. At the end, computational results provided by the algorithm from real data set are presented.

Keywords: Sampling; Stratification; Clustering Analysis; Metaheuristics; Genetic Algorithms;

1. INTRODUÇÃO

Os governos têm demandado cada vez mais aos institutos de estatística oficial informações socioeconômicas de suas populações. Estas informações são de suma importância à implementação de políticas públicas concernentes à educação, saúde, etc. Neste sentido, a amostragem aparece como uma ferramenta indispensável para o levantamento destas informações. A amostragem (Bolfarine e Bussab, 2005) é uma técnica estatística que permite a investigação de várias características de uma população, mediante a observação de apenas uma parte do seu universo de estudo, denominado amostra.

A população que será o objeto de investigação, bem como os custos da pesquisa, o tempo necessário para realizá-la e o nível de precisão desejado, são alguns dos fatores que estão intrinsecamente associados com o método de amostragem que será utilizado na pesquisa. Neste sentido, a amostragem fornece várias possibilidades para se efetuar o planejamento amostral (Lohr, 2010) de uma pesquisa. Pode-se pensar, por exemplo, em uma pesquisa realizada mediante a amostragem aleatória simples ou em uma pesquisa cujo plano amostral tenha vários estágios (níveis de conglomeração).

Levando-se em conta esta última observação, muitas das pesquisas realizadas pelos órgãos de estatística oficial trabalham com um plano amostral que agregue a conglomeração e a estratificação (Freitas et al, 2007). Em particular, a utilização de um plano amostral que incorpore a estratificação estatística tende a produzir estimativas mais precisas. Ou seja, as estimativas produzidas a partir da amostra terão um menor erro padrão associado.

Uma vez que estes estratos correspondem a grupos constituídos por elementos de uma população (pessoas, domicílios, setores, etc) que é o objeto de estudo da pesquisa, estes grupos podem ser definidos mediante a aplicação de um algoritmo não hierárquico, como por exemplo, o algoritmo *k*-means (Guojun et al, 2007 e Johnson e Wichern, 2002). Não obstante, com a aplicação deste tipo algoritmo, há a tendência de se produzir estratos não tão homogêneos, implicando, por sua vez, na obtenção de estimativas não tão precisas. Ademais, o critério de homogeneidade (a função objetivo) utilizado neste algoritmo não corresponde ao critério utilizado para avaliar o critério de homogeneidade dos estratos, ou seja, a minimização de uma expressão de variância.

De forma a trabalhar com a expressão de variância associada com o plano amostral estratificado e produzir soluções de boa qualidade, propõe-se no presente trabalho um algoritmo de estratificação baseado na metaheurística algoritmo genéticos (Sivanandam e Deepa, 2008, Linden, 2008). Os algoritmos genéticos têm sido aplicados com sucesso em diversos problemas de otimização, dentre os quais, problemas de agrupamento (Tseng and Yang, 2001, Furtado, 1998, Cruz, 2010).

Este trabalho está dividido da seguinte forma: Na seção dois são apresentados os conceitos básicos de amostragem e uma descrição detalhada do problema de estratificação. Na seção três é feita uma descrição da metaheurística algoritmos genéticos e do algoritmo de estratificação proposto. E finalmente, na seção quatro, são apresentados e discutidos um conjunto de resultados computacionais obtidos a partir da aplicação do algoritmo considerando dados reais.

2. O PROBLEMA DE ESTRATIFICAÇÃO UNIVARIADA

2.1 CONCEITOS BÁSICOS DE AMOSTRAGEM

Atualmente, a técnica de amostragem é utilizada em boa parte das pesquisas realizadas pelos órgãos de estatística oficial. A amostragem (Bolfarine e Bussab, 2005) permite fazer inferências para toda uma população de interesse (pessoas, domicílios, estabelecimentos comerciais ou industriais, etc), mediante a investigação de um subconjunto desta população denominado amostra. Através da aplicação da amostragem é possível

produzir estimativas de totais, médias, proporções ou razões associados à população. Dentre os motivos para o uso da amostragem destacam-se os seguintes:

- Custo: Observa-se, que em muitas pesquisas, a população a ser investigada é substancialmente grande, o que implica, por sua vez, em elevados custos para aplicação de um censo (investigação de toda a população).
- Tempo: Tendo em vista que muitas pesquisas precisam ser realizadas em um intervalo de reduzido e que, em contrapartida, a população a ser investigada está muito espalhada em relação à região de cobertura da pesquisa ou é muito grande, a utilização amostragem facilitaria o processo.
- Controle: Quando se pesquisa um número reduzido de elementos, pode-se dar uma maior atenção aos casos individuais, evitando erros nas respostas.

O tipo de plano amostral, ou seja, método de amostragem que será utilizado na pesquisa está intrinsecamente associado com a população a ser investigada, o nível de precisão desejado (erro da pesquisa), o orçamento disponível para a pesquisa e os dados disponíveis no cadastro (de onde será selecionada uma particular amostra). Sendo assim, a partir da avaliação destes fatores, pode-se adotar um plano amostral baseado em um dos seguintes esquemas (Cochran, 1977, Lohr, 2010): amostragem aleatória simples, amostragem estratificada, amostragem de conglomerados, amostragem sistemática ou uma combinação destes esquemas, considerando, por exemplo, conglomerados e estratos (Freitas et al, 2007). Em particular, nos restringiremos à descrição da amostragem estratificada, tenho em vista a proposta do presente trabalho.

2.2 AMOSTRAGEM ESTRATIFICADA E O PROBLEMA DE ESTRATIFICAÇÃO

A aplicação de um plano amostral com estratificação (Cochran, 1977) consiste na divisão de uma população em um número fixo de grupos mutuamente exclusivos e cuja união corresponde a toda a população. De acordo com Silva (2001) a opção por um plano amostral estratificado é motivada pelas seguintes questões: Deseja-se melhorar a precisão das estimativas; produzir estimativas para diversos segmentos da população (domínios); deseja-se que a amostra tenha a mesma composição da população segundo algumas características básicas e por questões administrativas ou operacionais.

Na amostragem estratificada, uma população com N unidades definidas por pessoas, domicílios, setores censitários (são demarcados pelo IBGE, obedecendo a critérios de operacionalização da coleta de dados, de tal maneira que abranjam uma área que possa ser percorrida por um único recenseador em um mês e que possua em torno de 250 a 350 domicílios), dentre outras unidades, é dividida em L subpopulações com $N_1, N_2, ..., N_h, ..., N_L$ unidades, respectivamente, chamadas de estratos. Essas subpopulações não se superpõem e, juntas, abrangem a totalidade da população, de tal modo que:

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

Depois de definidos os estratos, selecionam-se amostras independentes em cada um deles, considerando algum esquema de amostragem, como por exemplo, a amostragem aleatória simples. Neste caso, temos uma amostragem estratificada simples. Os tamanhos das amostras que serão selecionadas em cada um 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 (previamente definido). Apresenta-se a seguir a notação básica de amostragem estratificada simples utilizada neste trabalho:

N_h : Número total de unidades da população em cada estrato *h* (*h*=1,...,*L*).

n : Número total de unidades na amostra.

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

$\overline{X_h} = \sum_{i=1}^{N_h} X_{hi} / N_h$: Média populacional de X no *h-ésimo* estrato (2)

$S_{hx}^2 = \frac{\sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^2}{N_h - 1}$: Variância de *X* no *h-ésimo* estrato (3)

$V_X = \sum_{h=1}^{L} N_h^2 . \frac{S_{hx}^2}{n_h} . (1 - \frac{n_h}{N_h})$: Variância do total da variável *X* (4)

$cv_X = \sqrt{V_X} / X$: Coeficiente de Variação em relação à variável *X* (5)

Cabe observar que a precisão (um maior ou menor erro padrão) das estimativas produzidas a partir das amostras definidas em cada um dos estratos depende do grau de homogeneidade dos L estratos construídos em relação a uma variável ***X*.** Sendo esta variável denotada como variável da estratificação da pesquisa.

A homogeneidade é avaliada a partir do cálculo da variância total em relação aos estratos, ou seja, aplicando a equação (4) ou a equação (5). Quanto menor for o valor desta variância, mais homogêneos serão considerados os estratos. Esta homogeneidade melhora não apenas a estimativa para *X*, mas também para um conjunto variáveis de interesse $Y_1, Y_2, ..., Y_m$ que sejam investigadas na pesquisa e que tenham um bom grau de correlação com *X*.

Analisando a expressão (4) observa-se que o valor de $\boldsymbol{V_X}$ depende do número total de unidades da população em cada estrato (N_h) e das variâncias dos estratos (S_{hx}^2), sendo ambos definidos em função da distribuição das unidades populacionais pelos estratos. Além disso, $\boldsymbol{V_X}$ é impactada pelo número de unidades amostrais n_h alocadas a cada um dos estratos. Para a determinação de n_h , inicialmente poderia se optar pela aplicação da alocação de *Neyman* (equação 6), que corresponde a um caso especial da alocação ótima, usado quando os custos nos estratos são aproximadamente iguais. Esta alocação tende a produzir um valor de variância menor do que a variância considerando a alocação proporcional ou uniforme.

$$n_h = \frac{n . N_h . S_{hx}}{\sum_{g=1}^{L} N_g . S_{gx}} \quad (6)$$

Todavia, raramente esta alocação produz os tamanhos de amostra inteiros, o que implica, por sua vez, apenas em uma solução que é um ótimo local em relação à expressão (4). Ademais, para algumas populações, a aplicação desta particular alocação pode produzir tamanhos de amostra maiores que os tamanhos populacionais ($n_h > N_h$). E neste caso, deve-se adotar o 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 função disso, optou-se pela utilização de um algoritmo exato proposto por Brito (2005). Este algoritmo garante o cumprimento das restrições $n_h \leq N_h$ e $\sum n_h = n$ e produz os tamanhos de amostra inteiros que minimizam a variância da equação (4).

A partir da descrição do problema de estratificação, observa-se que o mesmo está associado a um problema de agrupamento (Kaufman e Rousseeuw, 1989, Guojun et al, 2007) sem restrições e que agrega uma particular função objetivo (variância).

A literatura de *Cluster Analysis* (Johnson A.R. e Wichern, 2002, Guojun et al, 2007) mostra que a obtenção do ótimo global para os problemas de agrupamento é muito difícil. Sendo esta dificuldade decorrente de que o número de soluções viáveis para este tipo de problema cresce exponencialmente à medida que o número de objetos a serem agrupados aumenta.

Aplicação de um procedimento de busca exaustiva para garantir a obtenção da solução global, implicaria na enumeração de todas as soluções viáveis, isto é, na avaliação de todas as combinações 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). Caso se

tenha, por exemplo, N = 10 elementos a serem alocados em L=3 estratos, o número de soluções a serem consideradas é de 9330. Para o caso em que N = 20 e L=3, o número de soluções viáveis sobe para 580.606.446. Considerando-se um número N maior de elementos, estes valores crescem exponencialmente.

Alternativamente, abrindo-se mão de trabalhar com a função objetivo do problema de estratificação e do ótimo global, os estratos poderiam ser construídos mediante a utilização de um algoritmo de agrupamento não hierárquico, como por exemplo, o algoritmo *k*-means Guojun et al, 2007).

Todavia, este algoritmo tende a produzir soluções viáveis que podem estar muito afastadas do ótimo global. Sendo assim, de forma a contornar a complexidade intrínseca ao problema de estratificação e trabalhar com a função de variância deste problema, propõe-se neste trabalho a aplicação de um algoritmo estratificação baseado na metaheurística algoritmos genéticos (AG) (Linden, 2008, Goldberg, 1989 e Sivanandam e Deepa, 2008). Assim como em outros problemas de otimização, com a aplicação de um AG será possível obter soluções com boa qualidade em um tempo computacional factível.

3. ALGORITMOS GENÉTICOS

3.1. ALGORITMOS GENÉTICOS

Os algoritmos genéticos (AG) (criados por Holland em 1975) é uma metaheurística que tem sido aplicada com êxito em diversos problemas (Glover and Kochenberger, 2002) de otimização. Um algoritmo genético (Linden, 2008) parte de uma população inicial de cromossomos (soluções) gerados aleatoriamente, faz a avaliação de cada um, seleciona os melhores e aplica manipulações genéticas, como cruzamento e mutação, com o objetivo de criar uma nova população. Primeiramente, deve-se determinar uma estrutura para a representação dos cromossomos, o que comumente é feito através de vetores. A população inicial pode ser gerada aleatoriamente ou utilizando alguma heurística de construção, produzindo p soluções $\boldsymbol{S}=(s_1 , s_2 , ..., s_p)$, sendo p o tamanho da população de $\boldsymbol{S}$. Estas soluções correspondem a um pequeno subconjunto do espaço total de soluções viáveis para o problema. Após a geração desta população, seguido da avaliação da função objetivo para cada solução $\boldsymbol{s_i}$ (i=1,...,p), são aplicados operadores genéticos (Linden, 2008, Goldberg, 1989 e Sivanandam e Deepa, 2008) à população na seguinte ordem:

Seleção: Em função dos valores da função objetivo, as melhores soluções são selecionadas e duplicadas em substituição às piores. Neste passo, pode-se considerar, por exemplo, o chamado método da roleta viciada (Linden, 2008), que consiste em fazer com que os cromossomos que possuem uma maior probabilidade de seleção sejam copiados um número maior de vezes do que os cromossomos 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. O número total de cópias realizadas deve ser igual ao tamanho da população.

Cruzamento: Esta operação possibilita a diversificação no espaço das soluções viáveis do problema. O operador de cruzamento padrão escolhe aleatoriamente dois cromossomos e troca partes do seu padrão genético (das soluções).

Mutação: Efetua a modificação de um gen ou de genes de alguns cromossomos. Este procedimento permite a regeneração de uma parte da solução que tenha sido eliminada da população, de forma inesperada, durante o cruzamento ou a seleção.

Após a aplicação destes procedimentos, obtém-se uma nova população. E, sobre esta, deve-se repetir a avaliação da função objetivo para cada solução 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 neste algoritmo: um número máximo gerações por um número de vezes pré-determinado, um tempo máximo de

processamento ou um número máximo gerações sem que ocorra melhoria no valor da função objetivo. A melhor solução produzida durante as gerações do algoritmo corresponderá à solução do problema.

3.2. ALGORITMO PROPOSTO PARA O PROBLEMA DE ESTRATICAÇÃO

3.2.1 Representação da População

Assim como em outros problemas de otimização, um primeiro passo à aplicação de um AG consiste na definição da estrutura que será utilizada na representação de cada cromossomo (solução) que comporá a população.

Em função das características do problema de estratificação, optou-se pela representação baseada no *group-number* (Michalewicz, 1996). Dessa forma, cada cromossomo terá *N* posições (este valor correspondente ao número de elementos da população) e atribui-se a cada posição um valor aleatório entre *1* e *L*. Este valor corresponde ao número do estrato populacional ao qual cada elemento será inicialmente alocado. A tabela abaixo ilustra esta representação para dois cromossomos (possíveis estratificações), considerando *N*=10 e *L*=2.

Tabela 1 – Representação de duas soluções

Solução 1	**1**	**2**	**1**	**2**	**2**	**2**	**1**	**2**	**1**
Solução 2	**2**	**1**	**1**	**2**	**1**	**1**	**1**	**2**	**2**

3.2.2 Geração da População Inicial

Ao aplicar este procedimento foram geradas 10.000 soluções (cromossomos com *N* valores entre 1 e *L* gerados aleatoriamente), selecionando-se, dentre estas, cem soluções (*tamanho_pop*=100) para compor a população inicial. A seleção destas soluções é feita utilizando o método da roleta (Linden, 2008) (avalia o valor da variância). As soluções restantes são armazenadas, sendo utilizadas posteriormente para substituição de soluções inviáveis (eventualmente geradas no cruzamento ou na mutação) e para a substituição de um percentual de 5% das soluções atuais a cada 100 gerações. Mais especificamente, neste último caso, serão substituídas algumas soluções repetidas (mediante seleção aleatória).

Após a definição do conjunto inicial de soluções, ou seja, uma vez efetuada a alocação dos elementos aos estratos, ficam definidos os valores de N_h e S_{hx}^2. Para a definição dos tamanhos de amostra que serão alocados a cada um dos estratos optou-se pela utilização de um algoritmo exato proposto por Brito (2005).

3.2.3 Cruzamento Uniforme

Considerando uma probabilidade de cruzamento definida a priori, são selecionadas *m* (par) soluções dentre as 100, definindo-se um vetor *V* com os índices destas soluções. Em seguida, são selecionadas em sequência, as soluções associadas às posições *2i-1* e *2i* (*i=1,..., m div 2*) de *V*. E, para cada duas soluções é gerado um vetor *M* (uma "máscara") com *N* posições contendo valores 0 e 1 (gerado aleatoriamente segundo uma distribuição uniforme). Para as posições de *M* que tiverem o valor 1, será efetuada a troca de valores entre as duas soluções selecionadas (cromossomos). A Tabela 2 traz um exemplo da aplicação do cruzamento uniforme para *N=10* e L=3.

Tabela 2 - Exemplo de Aplicação do Cruzamento Uniforme

Solução 1 (Antes)	1	2	1	2	2	1	3	3	1	1
Solução 2 (Antes)	3	2	2	3	1	2	2	3	2	1
Vetor M	1	0	0	1	1	0	1	0	1	0
Solução 1 (Depois)	3	2	1	3	1	1	2	3	2	1
Solução 2 (Depois)	1	2	2	2	2	2	3	3	1	1

3.2.3 Cruzamentos de um ponto e de dois pontos

Novamente, considerando a mesma probabilidade de cruzamento, são selecionadas *m* (par) soluções dentre as 100, definindo um vetor *V* com os índices destas soluções. Em seguida, são selecionadas em sequência as soluções associadas às posições *2i-1* e *2i* (*i=1,...,m div 2*) de *V*. E, para cada duas soluções é selecionado aleatoriamente um valor *k* entre 1 e *N*. Este valor corresponderá ao ponto de corte aplicado às duas soluções, tal que os valores entre as posições *k*+1 e *N* nestas duas soluções serão trocados. No caso do cruzamento de dois pontos, a diferença está na seleção de dois valores k_1 e k_2 entre 1 e *N* e a troca dos valores entre estas posições, considerando as duas soluções. As Tabelas três e quatro ilustram, respectivamente, a aplicação do cruzamento de um ponto e de dois pontos para *N=10* e L=3.

Tabela 3 - Exemplo de Aplicação do Cruzamento de um Ponto (k=6)

Solução 1 (Antes)	1	2	1	2	2	1	**3**	**3**	**1**	**1**
Solução 2 (Antes)	3	2	2	3	1	2	**2**	**3**	**2**	**1**
Solução 1 (Depois)	1	2	1	2	2	1	**2**	**3**	**2**	**1**
Solução 2 (Depois)	3	2	2	3	1	2	**3**	**3**	**1**	**1**

Tabela 4 - Exemplo de Aplicação do Cruzamento de Dois Pontos (k_1=3 e k_2=7)

Solução 1 (Antes)	1	2	1	2	2	1	3	3	1	1
Solução 2 (Antes)	3	2	2	3	1	2	2	3	2	1
Solução 1 (Depois)	1	2	**2**	**3**	**1**	**2**	**3**	3	1	1
Solução 2 (Depois)	3	2	**1**	**2**	**2**	**1**	**2**	3	2	1

3.2.4 Mutação por Posição

Considerando uma probabilidade de mutação (*pmut*) definida a priori, são selecionados aleatoriamente *q* (q=inteiro(*pmut* x *N*)) valores *i* entre 1 e (*N* x *tamanho_pop*) definindo um vetor $I=\{i_1, i_2,..., i_q\}$. Neste caso, o número e a posição da solução que será alterada, são respectivamente obtidos através das seguintes operações: (i_j *div N*)+1 (divisão inteira) e (i_j *mod N*)+1 (resto da divisão inteira) (*j*=1,...,q). Por exemplo, supondo *pmut*=0.03, *N*=200 e *tamanho_pop*=100 teremos q=6 valores que serão sorteados entre 1 e 20000. A Tabela cinco ilustra a aplicação deste tipo de mutação.

Tabela 5 - Exemplo de Aplicação da Mutação por Posição

Valores Sorteados (I)	Número da Solução	Posição na Solução
800	5	1
15410	70	11
18970	94	171

Os valores associados a cada uma das posições destas soluções serão substituídos por um valor aleatório selecionado entre 1 e ***L*** (excluindo o valor atual).

3.2.5 Mutação por Solução

São selecionadas aleatoriamente *q* (q=inteiro(*pmut* x *tamanho_pop*)) soluções da população atual e para cada uma destas soluções são sorteadas *p* valores (posições) entre 1 e *N*. Por exemplo, supondo *pmut*=0.03, *N*=200 e *tamanho_pop*=100, teremos q=3 valores entre 1 e 200 que serão sorteados para cada solução. A Tabela seis ilustra a aplicação deste tipo de mutação.

Tabela 6 - Exemplo de Aplicação da Mutação por Solução

Soluções sorteadas	Posições sorteadas
1	10, 88,193
17	23, 57, 91
82	20, 40, 71

E, tal qual no caso da mutação por posição, os valores associados a cada uma destas posições serão substituídos por um valor aleatório selecionado entre 1 e *L* (excluindo o valor atual).

3.3 SELEÇÃO

Para a seleção das soluções que serão promovidas para a próxima geração, optou-se pela aplicação do método da roleta (Linden, 2008). Neste caso, avaliou-se a expressão da variância total (equação 4). Além disso, foi adotada uma estratégia de elitismo que consistiu na seleção das quatro melhores soluções obtidas até a geração atual e a sua promoção para a geração seguinte.

4. RESULTADOS COMPUTACIONAIS

O algoritmo genético de estratificação e o algoritmo exato (Brito, 2005) foram desenvolvidos em linguagem ***R*** (versão 2.14)**.** Em particular, no caso do algoritmo de estratificação baseado no AG, foram desenvolvidas seis versões a partir da combinação dos procedimentos de cruzamento (três), mutação (dois) e seleção (um) (Tabela 7).

Tabela 7 - Versões do Algoritmo de Estratificação

Operador\ versão AG	AG1	AG2	AG3	AG4	AG5	AG6
Cruzamento Uniforme	X	X				
Cruzamento de Um Ponto			X	X		
Cruzamento de Dois Pontos					X	X
Mutação por Posição	X		X		X	
Mutação por Solução		X		X		X

De forma a avaliar o desempenho de cada uma destas versões, foi realizado um conjunto de experimentos computacionais com dados reais do universo do censo demográfico de 2000. Mais especificamente, foram selecionados 24 municípios do Brasil (instâncias) e para cada um destes, disponibilizou-se por setor censitário a média dos rendimentos nominais das pessoas responsáveis pelos domicílios contidos no setor. Esta variável foi utilizada no cálculo da variância total (equação 4), e por consequência, para a definição de estratos constituídos pelos setores censitários. Todos os experimentos computacionais foram efetuados em um computador com 16GB de memória RAM e dotado de oito processadores de 2.93 GHz (I7-870).

A Tabela oito traz os nomes dos municípios escolhidos, o número de setores censitários (*N*) e algumas medidas estatísticas associadas com a variável de estratificação (média dos rendimentos nominais).

Tabela 8 – Informações sobre os Municípios Selecionados

Município	Setores	Média	Q_1	Mediana	Q_3
Aparecida	300	639,5	447,1	558,5	720,5
Campinas	1297	1538,8	712,3	1106,8	1853,8
Campos	537	722,4	387,4	529,1	824,6
Caruaru	271	649,6	335,1	473,5	802,8
Contagem	593	801,9	559,2	746,8	962,8
Curitiba	2081	1581,4	741,4	1137,8	1952,8
Duque de Caxias	1054	662,4	491,8	602,3	743,3
Fortaleza	2172	1025,7	432,3	658,4	1104,6
Guarulhos	1322	965,4	584,8	781,7	1137,5
Itacaquacetuba	249	696,9	529,1	630,2	771,0
Londrina	432	1194,6	605,9	826,5	1347,6
Maceió	672	1020,1	405,1	617,8	1081,2
Manaus	1556	1017,1	504,4	688,5	1111,8
Niterói	695	1850,2	860,0	1554,9	2641,7
Petrópolis	377	1138,0	640,9	927,3	1420,6
Rio Claro	210	1184,0	681,4	1001,2	1410,8
Santos	597	1670,7	872,2	1468,2	2155,6
São Bernardo do Campo	847	1339,1	719,4	1136,9	1715,9
São Gonçalo	1211	753,0	553,8	695,8	889,2
São José dos Campos	750	1390,5	726,4	969,2	1477,5
São Luís	773	967,0	424,8	606,1	1084,0
Viamão	280	802,5	615,3	716,8	907,2
Vitória	263	1780,1	715,1	1142,3	2352,1
Volta Redonda	416	808,8	453,5	655,7	948,3

* Por motivos de sigilo da informação, os dados das vinte e quatro populações (valores de X por setor) não foram disponibilizados pelo autor.

Além disso, no que concerne aos parâmetros do algoritmo genético, em todos os experimentos realizados, o tamanho da população foi definido como 100, o número de gerações foi fixado em 1000 e as probabilidades de cruzamento e de mutação foram definidas, respectivamente, em 50% e 5%. Esses parâmetros foram ajustados previamente, considerando

a aplicação dos seis algoritmos em seis instâncias da Tabela 8, quais sejam: Aparecida, Duque de Caxias, Guarulhos, Londrina, Rio Claro e Volta Redonda.

Acrescenta-se, ainda, que o algoritmo foi aplicado em cada uma das 24 instâncias considerando o número de estratos variando entre três e sete e um tamanho de amostra *n* (utilizado no algoritmo de alocação) correspondente a 20% do total da população (*N*). Observa-se que o intervalo de variação dos estratos foi definido levando em conta questões da aplicação real, ou seja, um número de estratos maior não produz ganhos substanciais (Azevedo, 2004) em relação ao valor da variância ou, equivalentemente, do coeficiente de variação. Nesta seção, optou-se em apresentar os resultados sobre a homogeneidade utilizando o coeficiente de variação, tendo em vista a sua utilização nos trabalhos sobre estratificação.

A Figura 1 traz os gráficos de caixa (*boxplots*) (Bussab e Morettin, 2011) construídos a partir dos coeficientes de variação (cvs) observados para cada uma das vinte e quatro instâncias, considerando a aplicação das seis versões do AG. A partir da análise destes gráficos observa-se que as versões 1, 3 e 5 tiveram um melhor desempenho em relação às outras versões. Em função desta observação e pelo fato dos algoritmos 2, 4 e 6 não terem produzido nenhuma solução superior às soluções dos outros algoritmos, os gráficos e as tabelas seguintes foram baseadas nos resultados obtidos em relação às versões 1, 3 e 5.

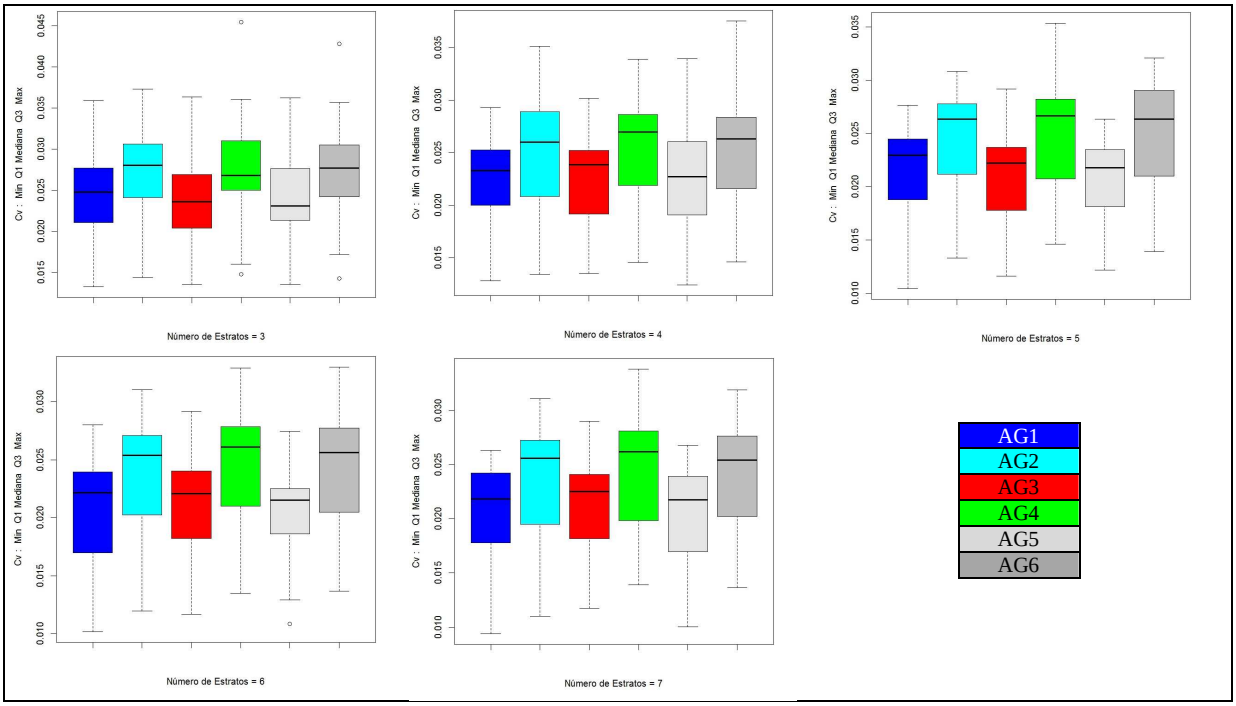

Figura 1- Boxplots dos Cvs Obtidos pelas seis versões do AG para as 24 Instâncias

Doravante, denotaremos por solução ótima, a melhor solução obtida (menor coeficiente de variação) para cada uma das vinte e quatro instâncias, considerando as versões 1, 3 e 5 do AG. A Tabela nove e o Gráfico 1 possibilitam uma comparação entre as performances destas três versões, no que concerne ao número de soluções ótimas. Neste caso, o algoritmo que teve o melhor desempenho foi o AG5 com 48 soluções ótimas dentre as 120 produzidas (número de estratos x número de instâncias), ou seja, em 40% dos casos o AG5 produziu a melhor solução.

Tabela 9 – Número de Soluções Ótimas por Algoritmo e por Estrato

Estratos	AG1	AG3	AG5
3	5	11	8
4	5	9	10
5	6	7	11
6	10	4	10
7	9	6	9
Total	**35**	**37**	**48**

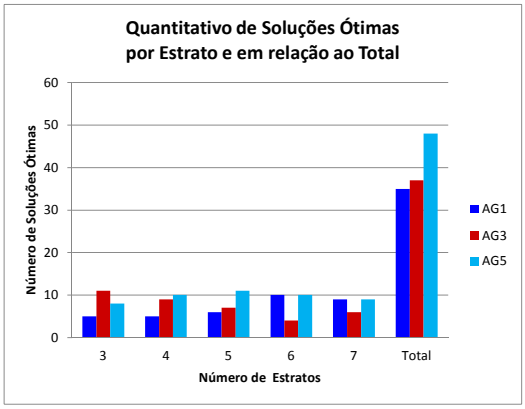

Gráfico 1 – Soluções Ótimas por Algoritmo e por Estrato

A Tabela 10 traz (por estrato) as medidas-resumo (Bussab e Morretin, 2011) em relação aos gaps (em percentual) calculados entre as soluções produzidas pelos algoritmos 1, 3 e 5. Cada gap foi calculado utilizando a seguinte expressão:

$$Gap_a = 100*(cv_i - cv_j)/cv_j \quad (i=1,3,5 \text{ tq } i \neq j)$$

Onde cv_j é o coeficiente de variação do algoritmo j (1,3 ou 5) que produziu a solução ótima e cv_i corresponde ao coeficiente de variação dos outros algoritmos. Por exemplo, no caso do Gap_{31}, o valor médio de 2,3 (1ª linha) corresponde à média dos gaps entre os cvs de todas as instâncias nas quais a solução do AG1 foi melhor que a solução do AG3.

Prosseguindo com a análise desta tabela, e particularmente, considerando os casos em que o AG1 foi melhor do que o AG3 e o AG5, a média, a mediana e o 3º quartil dos gaps deste algoritmo foram maiores (excetuando-se três estratos) do que a média, a mediana e o 3º quartil dos outros dois algoritmos. Este fato indica que as soluções ótimas deste algoritmo foram melhores que as dos algoritmos AG3 e AG5.

Tabela 10 – Medidas-Resumo dos Gaps entre os Algoritmos

Estratos	Algoritmo	Gap	Min	Q1	Mediano	Médio	Q3	Máximo
	AG1	Gap31	0,4	0,7	1,3	2,3	2,0	7,1
		Gap51	0,4	1,0	2,0	3,6	5,0	9,9
3	AG3	Gap13	1,1	4,3	6,7	7,1	9,5	15,3
		Gap53	0,8	2,4	3,6	5,5	6,9	13,7
	AG5	Gap15	0,3	1,2	5,0	6,0	8,0	19,1
		Gap35	0,1	0,9	1,2	2,7	3,0	11,4
	AG1	Gap31	5,7	6,8	**9,7**	**10,8**	**13,1**	18,7
		Gap51	1,9	3,0	**5,4**	**6,2**	**9,1**	11,4
4	AG3	Gap13	0,2	2,8	5,9	6,4	9,3	13,5
		Gap53	0,7	4,0	5,9	8,0	10,3	22,5
	AG5	Gap15	0,5	2,5	4,5	4,4	5,3	9,5
		Gap35	0,3	4,1	5,9	7,8	11,6	19,3
	AG1	Gap31	2,9	5,4	**8,4**	**8,4**	**11,4**	14,2
		Gap51	1,0	8,0	**11,8**	**10,3**	**13,4**	16,6
5	AG3	Gap13	0,6	4,7	6,8	6,9	9,3	12,9
		Gap53	0,3	2,0	3,9	5,0	7,6	11,6
	AG5	Gap15	2,0	4,7	5,9	7,0	8,3	18,1
		Gap35	0,5	1,3	3,9	4,7	8,1	10,7
	AG1	Gap31	0,9	3,4	6,0	**9,0**	**11,2**	32,7
		Gap51	1,0	2,2	**6,0**	**6,6**	**9,8**	18,1
6	AG3	Gap13	3,6	5,3	7,0	6,9	8,5	10,0
		Gap53	0,6	1,1	2,2	2,2	3,3	3,8
	AG5	Gap15	0,3	1,5	2,5	5,3	7,3	17,4
		Gap35	0,2	2,5	7,7	6,5	10,7	11,7
	AG1	Gap31	0,3	6,6	**13,5**	**13,2**	**19,9**	25,1
		Gap51	2,0	4,4	**7,1**	**8,9**	**11,4**	23,8
7	AG3	Gap13	0,6	2,7	5,3	5,1	7,2	9,7
		Gap53	1,1	2,6	4,2	3,6	4,7	5,4
	AG5	Gap15	0,4	1,0	3,2	5,3	10,9	13,1
		Gap35	0,8	1,0	6,1	6,2	7,8	14,5

A Tabela 11 traz as medidas-resumo em relação aos tempos de processamento dos três algoritmos. Uma análise desta tabela indica um equilíbrio entre o AG1, AG3 e AG5, tendo em vista que os seus tempos mediano e médio ficaram próximos em relação às três versões, excetuando-se o experimento com quatro estratos, onde houve uma diferença razoável entre o AG1 e os algoritmos AG3 e AG5 no que diz respeito a estas medidas.

Tabela 11 – Medidas-Resumo dos Tempos de Processamento dos Três Algoritmos considerando as 24 instâncias

Estratos	Algoritmo	Min	Q1	Mediano	Médio	Q3	Máximo
3	AG1	8	12	21	34	39	148
	AG3	8	13	18	35	35	171
	AG5	8	11	19	30	31	133
4	AG1	14	23	49	60	75	174
	AG3	8	14	24	41	44	236
	AG5	8	11	20	39	44	194
5	AG1	10	18	31	62	86	366
	AG3	13	18	36	57	74	201
	AG5	11	17	38	58	65	232
6	AG1	10	24	45	75	121	266
	AG3	9	24	32	72	71	336
	AG5	9	25	36	73	52	426
7	AG1	14	26	41	75	81	321
	AG3	13	23	38	67	98	320
	AG5	17	29	50	76	83	275

Tempos de processamento em minutos

As Figuras 2, 3, 4 e 5 a seguir correspondem às soluções (estratificação dos setores censitários) produzidas pelo AG5 para os municípios de Campinas (3 estratos), Curitiba (4 estratos), Guarulhos (5 estratos) e São Gonçalo (6 estratos).

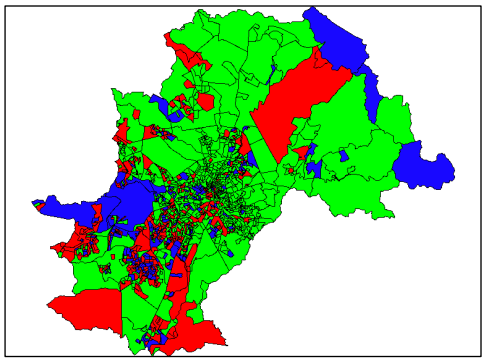

Figura 2 – Campinas

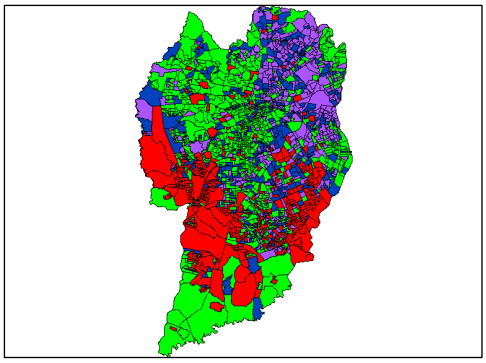

Figura 3 - Curitiba

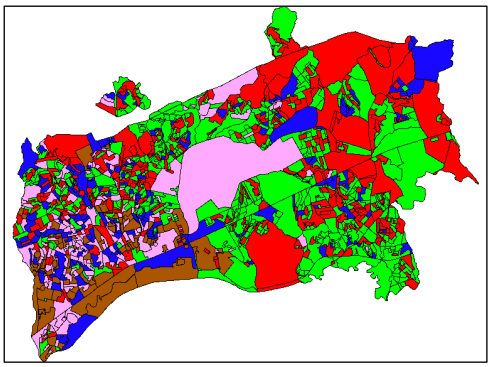

Figura 4 – Guarulhos

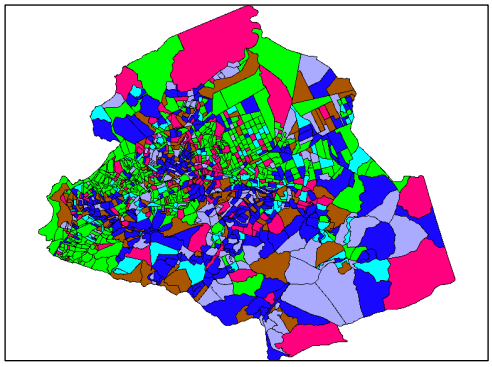

Figura 5 – São Gonçalo

Os comentários e as análises efetuadas nesta seção indicam que os algoritmos apresentados neste trabalho são uma boa alternativa para a resolução do problema de estratificação univariada. Em trabalhos futuros, pretende-se desenvolver versões mais eficientes do AG para a estratificação, considerando as probabilidades de cruzamento e mutação variáveis (Linden, 2008), outros tipos de cruzamento (Sivanandam e Deepa, 2008) e procedimentos construtivos para a geração da população inicial.

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Azevedo, R.V, Estudo Comparativo de Métodos de Estratificação Ótima de Populações Assimétricas, Dissertação de Mestrado, Escola Nacional de Ciências e Estatísticas (ENCE/IBGE), 2004.

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

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

[4] Bussab, W.O e Morettin, P.A, Estatística Básica – 7ª Edição, Editora Saraiva, 2011.

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

[6] Cruz, M.D. e Ochi, L.S. Um Algoritmo Evolutivo com Memória Adaptativa para o Problema de Clusterização Automática. Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Neural Network Society, v. 8, Iss. 4, pp. 227-239, 2010.

[7] Freitas, M.P.S, Lila, M.F., Azevedo, R.V. e Antonaci, Giuseppe de Abreu. Amostra Mestra para o Sistema Integrado de Pesquisas Domiciliares, Textos para Discussão, 23, Diretoria de Pesquisas, IBGE, 2007.

[8] Furtado, J.C, Algoritmo Genético Construtivo na Otimização de Problemas Combinatoriais de Agrupamentos, Tese de Doutorado, INPE, 1998.

[9] Glover, F. e Kochenberger, G. A. Handbook of Metaheuristics, 1ª ed., Norwell: K1uwer Academic Publishers, 2002.

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

[11] Guojun G., Chaogun M and Jianhong W. (2007). Data Clustering: Theory, Algorithms, and Applications. (ASA-SIAM Series on Statistics and Applied Probability).

[12] Holland, J. H., Adaptation in natural artificial systems. University of Michigan Press, 1975.

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

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

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

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

[17] Michalewicz, Z. Genetic Algorithms + Data Structures = Evolution Programs. Springer, Third, Revised and Extended, 1996.

[18] Silva, N. N. Amostragem Probabilística, Editora Edusp, 2001.

[19] Sivanandam, S.N. and Deepa S. N. Introduction to Genetic Algorithms, Springer, 2008.

[20] Tseng, L.Y. and Yang, S. B. A genetic approach to the automatic clustering problem, Pattern Recognition, 34, p. 415-424, 2001.

Agradecimentos: A Viviane Quintaes (da Comeq/DPE/IBGE) pelo suporte nos mapas.