Algoritmo heurístico aplicado ao problema de estratificação ótima

Breno Tiago Novello Trotta de Oliveira[1], Leonardo Silva de Lima[2] and José André de Moura Brito[3]

[1] Instituto Brasileiro de Geografia e Estatística (IBGE),
Av. Republica do Chile, 500 - Centro, Rio de Janeiro – RJ
[2] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ),
Av. Maracanã, 229 - Maracanã, Rio de Janeiro – RJ
[3] Escola Nacional de Ciências Estatísticas (ENCE/IBGE),
Rua André Cavalcanti, 106 - Centro, Rio de Janeiro – RJ

Resumo Este trabalho traz a proposta de um algoritmo de otimização para resolução do problema de estratificação ótima univariado. Nesse problema, dado um nível de precisão fixado, busca-se a minimização do tamanho da amostra. O algoritmo é baseado nas metaheurísticas VNDS com Path-Relinking. A parte final desse artigo traz um conjunto de resultados computacionais considerando a comparação do algoritmo proposto com dois algoritmos bem conhecidos da literatura, produzindo a melhor solução para 90% das instâncias consideradas.

Keywords: Estratificação, Amostragem, Otimização, Metaheurísticas

1 Introdução

O problema de estratificação ótima, que está associado à área de amostragem probabilística, pode ser formulado considerando dois objetivos possíveis, a saber: *(i)* minimizar a variância de um estimador (considerando o tamanho de amostra fixo) ou *(ii)* minimizar o tamanho amostral (considerando o nível de precisão fixo). A maioria dos algoritmos propostos na literatura foram desenvolvidos para atender ao primeiro objetivo, enquanto o segundo objetivo tem sido menos estudado.

Os algoritmos de estratificação disponíveis atualmente na literatura não incorporam a restrição de tamanho amostral mínimo por estrato, algo muito importante para algumas pesquisas amostrais realizadas nos institutos de estatística oficial como, por exemplo, o Instituto Brasileiro de Geografia e Estatística (IBGE).

O presente trabalho traz a proposta de um novo algoritmo alternativo aos algoritmos da literatura, com o intuito de atender ao objetivo *(ii)* do problema de estratificação univariado, incorporando a restrição do tamanho de amostra mínimo por estrato.

O algoritmo proposto neste trabalho combina um procedimento de resolução exata e dois procedimentos de resolução baseados nas metaheurísticas Variable

Neighborhood Decomposition Search (VNDS) proposta por [8] e Path-Relinking (PR) proposta originalmente por Fred Glover em [1].

O presente trabalho está dividido da seguinte forma: a seção dois traz os conceitos básicos sobre a amostragem, incluindo uma descrição detalhada sobre o problema de estratificação ótima. A seção três traz a descrição do algoritmo proposto e a seção quatro traz os resultados computacionais obtidos, a partir de um conjunto de instâncias derivadas de base de dados selecionadas da literatura, e comparados aos algoritmos propostos por [10] e[11]. E por fim, na seção cinco são apresentadas as conclusões e os possíveis desdobramentos desse estudo.

2 Amostragem e Problema de Estratificação

A amostragem consiste na seleção de um subconjunto de unidades para representar a população como um todo, para que se possa inferir sobre características de interesse da população. Segundo [6], entre as vantagens de se utilizar amostragem em vez da enumeração completa da população, estão a redução dos custos, a coleta de dados de forma mais rápida e mais abrangente, e maior acurácia na coleta das informações.

2.1 Amostragem Estratificada (AE)

Conforme [6,12], neste tipo de amostragem a população (U) de N unidades é particionada em L subpopulações constituídas por $N_1, N_2, \ldots, N_L$ unidades, respectivamente, de tal forma que essas subpopulações (denominadas estratos) denotadas por $E_1, E_2, \ldots, E_L$ não se sobrepõem e, juntas, abrangem a totalidade da população.

Uma vez determinados os estratos populacionais, seleciona-se uma amostra aleatória simples para cada estrato h, denotada por $\mathbb{S}_h$ para $h = 1, \ldots, L$. Cada amostra $\mathbb{S}_h$ tem um tamanho associado, denotado por n_h, de modo que o tamanho de amostra total é dada pela soma dos tamanhos amostrais de cada estrato, cuja expressão é: $n = \sum_{h=1}^{L} n_h$.

Outra etapa da AE diz respeito ao procedimento que, consiste na distribuição das n unidades da amostra pelos estratos, chamado de alocação da amostra e denotado por a_h, para $h = 1, \ldots, L$. Uma vez conhecido o tamanho amostral n, pode-se calcular os tamanhos amostrais por estrato (n_h), tal que $n_h = n \cdot a_h$. Os principais métodos são: alocação Proporcional, alocação Uniforme, alocação ótima e alocação de Neyman, que podem ser encontradas em [2,6]. Recentemente, [5] desenvolveram um método que garante o ótimo global para a alocação da amostra.

Outro plano amostral que pode ser utilizado é a Amostragem Estratificada por Corte (AEC), alternativamente à Amostragem Estratificada. Entretanto, difere da AE, apenas por um aspecto: no último estrato todas as unidades da população compõem a amostra, tal que $n_L = N_L$. A AEC é utilizada quando a variável de estratificação associada à população apresenta uma alta assimetria. Considerando estas definições, deve-se procurar os pontos de corte ótimos

visando atender a um dos seguintes objetivos: *(i)* minimizar a variância de um estimador de total, ou *(ii)* minimizar o tamanho amostral.

O procedimento de estratificação por corte pode ser formalizado, considerando $X_U = \{x_1, x_2, ..., x_N\}$ o vetor populacional relacionado à variável de estratificação x, e sabendo que $x_1 \leq x_2 \leq ... \leq x_N$. As observações de X_U são alocadas aos L estratos, segundo os pontos de corte $b_1 < b_2 < \ldots < b_{L-1}$. Assim, para a definição de L estratos são necessários $(L-1)$ pontos de corte, se o valor de x_i for menor ou igual que b_1, essa unidade será alocada ao estrato E_1. Por sua vez, se o valor de x_i estiver entre b_1 e b_2 a unidade i será alocada ao estrato E_2, e assim sucessivamente, até que todas as unidades da população tenham sido alocadas em algum estrato.

Para o cálculo dos estimadores, considere que o total populacional associado à variável de estratificação x seja dado por $X = \sum_{i \in U} x_i = \sum_{h=1}^{L} \sum_{i \in E_h} x_i$, e assim, o estimador do total populacional sob AEC é denotado por $\hat{X}_{AEC}$ e definido por

$$\hat{X}_{AEC} = X_L + \sum_{h=1}^{L-1} N_h \overline{x}_h, \tag{1}$$

sendo $X_L = \sum_{i \in E_L} x_i$ o total populacional do estrato certo e $\overline{x}_h = \frac{1}{n_h} \sum_{i \in \mathbb{S}_h} x_i$ a média amostral de x no h-ésimo estrato.

A variância do estimador de total $(\hat{X}_{AEC})$ é dada por

$$V(\hat{X}_{AEC}) = \sum_{h=1}^{L-1} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{hx}^2}{n_h}, \tag{2}$$

sendo que $S_{hx}^2 = \frac{1}{N_h - 1} \sum_{i \in E_h} (x_i - \overline{X}_h)^2$ é a variância populacional de x no h-ésimo estrato e $\overline{X}_h = \frac{1}{N_h} \sum_{i \in E_h} x_i$ é a média populacional de x no h-ésimo estrato.

O coeficiente de variação do estimador de total $(\hat{X}_{AEC})$ é dado por

$$CV(\hat{X}_{AEC}) = \frac{\sqrt{V(\hat{X}_{AEC})}}{X}. \tag{3}$$

Assim, de acordo com o método de alocação(a_h) e conforme [6,11], o cálculo do tamanho de amostra total (n) é dado por

$$n = N_L + \frac{\sum_{h=1}^{L-1} \frac{N_h^2 S_{hx}^2}{a_h}}{X^2 CV^2(\hat{X}_{AEC}) + \sum_{h=1}^{L-1} N_h S_{hx}^2}. \tag{4}$$

2.2 Etapas do Problema de Estratificação

O problema de estratificação é mais usual para populações assimétricas, em que a AEC se justifica. Entretanto, ele também existe para populações com ausência da assimetria. De forma geral, as etapas do problema podem ser resumidas da seguinte forma:

1. Determine o objetivo do problema;
2. Fixe o número de estratos (L);
3. Escolha o método de alocação;
4. Escolha o método de seleção da amostra;
5. Escolha a variável de estratificação;
6. Calcule os $(L-1)$ pontos de corte $(b_1, ..., b_{L-1})$, para poder dividir a população em L estratos;
7. Calcule os tamanhos amostrais $n_1, n_2, ..., n_L$ de acordo com o método de alocação do item (3);
8. Selecione as unidades dentro de cada estrato de acordo com o método de seleção do item (4) e de acordo com o tamanho n_h do item (7).

A etapa (1) consiste em definir o objetivo dentre os dois objetivos possíveis, a saber: *(i)* minimizar a variância do estimador de total, ou seja, maximizar a precisão, considerando o tamanho de amostra fixo; *(ii)* minimizar o tamanho amostral, ou seja, minimizar o custo, considerando a precisão fixada previamente. A maioria dos algoritmos propostos na literatura foram concebidos para atender ao primeiro objetivo, como em [4,7,9], enquanto o segundo objetivo tem sido menos explorado na literatura, como em [10,11]. A etapa (2) consiste em definir a quantidade de estratos populacionais. A etapa (3) consiste em utilizar algum dos métodos de alocação, sendo o de Neyman o mais comum entre os trabalhos. A etapa (4) consiste em escolher um dos métodos de amostragem probabilística, em que o mais usual é a Amostragem Aleatória Simples(AAS). Na etapa (6) utiliza-se a variável de estratificação escolhida no item (5), para determinar os pontos de corte que possibilitam atender o objetivo *(ii)*. Essa etapa corresponde ao primeiro nível do problema de estratificação, enquanto a etapa (7) corresponde ao segundo nível do problema de estratificação.

Os algoritmos propostos na literatura não incorporam a restrição de tamanho amostral mínimo por estrato, representada pela equação (5):

$$min\{5, N_h\} \leq n_h \leq N_h \qquad h = 1, ..., L-1. \tag{5}$$

Ou simplesmente, $n_h \geq 5$, quando possível. Essa restrição é para evitar que determinados estratos sejam representados por poucas unidades devido a não-resposta.[4] Entretanto, ela não é aplicável para o estrato certo, pois $n_L = N_L$ sempre.

Como o segundo nível do problema já foi resolvido por [5], o algoritmo proposto nesse trabalho utiliza a alocação apresentada nesse artigo e se concentra, apenas, no primeiro nível do problema de estratificação, com o intuito de solucionar o objetivo *(ii)*.

3 Metodologia

A ideia central do algoritmo proposto é baseada no processo de discretização proposto em [3]. Mais especificamente, considere o vetor populacional $X_U = \{x_1,$

[4] A não-resposta ocorre quando um questionário foi enviado, mas não foi respondido, devido a inúmeros motivos.

$x_2, ..., x_N\}$, onde cada x_i corresponde ao valor da variável de estratificação para a unidade $i \in U$. O conjunto Q representa todos os pontos de corte distintos possíveis, a partir da retirada das duplicações de X_U. Assim, assumindo que w seja a quantidade de elementos de Q, e que cada ponto de corte seja denotado por q_j para $j = 1, \ldots, w$, tem-se $Q = \{q_1, q_2, ..., q_w\}$.

Conforme já explicitado na etapa (6) descrita na seção 2.2, são necessários $(L-1)$ pontos de corte para definir os estratos. Por exemplo, no caso em que $L = 4$, o conjunto solução que está se buscando corresponde à melhor escolha possível do vetor $\mathbf{b} = \{b_1, b_2, b_3\}$, tal que $\mathbf{b} \subseteq Q$, que leve a um tamanho amostral n mínimo. Essa discretização permite que se calcule a solução ótima global ($\mathbf{b}^*$) para populações com pequenos valores para w, pois é possível testar todas as combinações de pontos de corte de Q, uma vez que o problema se resume a combinação $\binom{w}{L-1}$. Só é possível garantir que essas soluções são ótimas globais, mediante o uso do método de enumeração exaustiva e a alocação de [5], que também é um método exato.

A partir da ideia de enumeração citada acima, propõe-se um algoritmo para resolução do problema de estratificação, que apresenta um procedimento de resolução exata e dois procedimentos de resolução baseada nas metaheurísticas VNDS e Path-Relinking (PR). Em alusão a esses três procedimentos citados (Exato, VNDS, PR), o algoritmo aqui proposto foi denominado **EVP**. A estrutura resumida está apresentada no algoritmo (1): caso o problema seja considerado pequeno, obtém-se a solução ótima global a partir do procedimento de enumeração exaustiva; caso contrário, o algoritmo produzirá a melhor solução, que não é necessariamente um ótimo global, baseando-se nas metaheurísticas VNDS e PR. A partir de experimentos prévios, definiu-se que um problema é pequeno quando: $N \leq 4.000$ e o valor de $\binom{w}{L-1} \leq 1.000.000$.

Algoritmo 1 Estrutura Resumida do Algoritmo EVP

```
Se Problema Pequeno Então
  | Exato;
Senão
  | b^0 =VNDS(Inicialização)
  | Enquanto Critério de parada não é satisfeito Faça
  |   | k = 1
  |   | Enquanto k ≤ k_max Faça
  |   |   | b^i =VNDS(Perturbação)
  |   |   | b^ii =VNDS(Busca Local)
  |   | Fim Enquanto
  |   | b^iii =Path-Relinking
  | Fim Enquanto
Fim Se
```

Na fase inicial do VNDS, define-se um vetor $\mathbf{b}^0$ viável, selecionando aleatoriamente $(L-1)$ valores do conjunto Q correspondentes aos elementos $b_1, b_2, \ldots, b_{L-1}$.

Além disso, define-se a estrutura de vizinhança com os valores q_j ao redor de b_h, tal que $q_{j-r_1} < b_h = q_j < q_{j+r_1}$, sendo r_1 a amplitude do intervalo da perturbação (a mesma estrutura é válida para a amplitude da busca local r_2) e j é a posição de b_h no conjunto Q, tal que $h = 1, 2, \ldots, L-1$ e $j = 1, 2, \ldots, w$.

Para ilustrar o algoritmo, suponha que a variável de estratificação da população de interesse tenha os seguintes valores $X_U = \{1, 1, 1, 2, 2, 3, 3, 4, 4, 5, 7, 7, 8, 8, 10, 10, 15, 31\}$. Portanto, ao desconsiderar as duplicações, chega-se ao conjunto $Q = \{q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8, q_9, q_{10}\} = \{1, 2, 3, 4, 5, 7, 8, 10, 15, 31\}$. Supondo $L = 4$ e um vetor aleatório inicial $\mathbf{b}^0 = \{3, 7, 10\}$.

Na fase de perturbação do VNDS gera-se um novo vetor solução $\mathbf{b}^i$ que difere de $\mathbf{b}^0$ por k elementos, em que sorteia-se, aleatoriamente, quais elementos serão livres e de acordo com a regra de vizinhança descrita acima. Por exemplo, suponha o elemento b_2 (segundo ponto de corte) livre, tal que o novo vetor $\mathbf{b}^i$ será $\{3, b_2', 10\}$, sendo b_2' obtido a partir da regra de vizinhança descrita acima, $q_{6-r_1} < (b_2 = q_6) < q_{6+r_1}$, onde r_1 é o parâmetro da amplitude da perturbação. Para $r_1 = 2$, esse exemplo de vizinhança está ilustrado na Figura 1, o novo valor de b_2' será um elemento qualquer do conjunto $\{4, 5, 8, 10\}$, sendo produzido, por exemplo, um novo vetor solução vizinho $\mathbf{b}^i = \{3, 4, 10\} = \{q_3, q_4, q_8\}$.

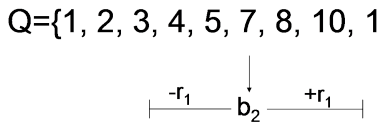

Figura 1. Exemplo de Estrutura de Vizinhança para o Algoritmo Proposto

Na fase de busca local do VNDS, a partir do vetor $\mathbf{b}^i$, produz-se a melhor solução dentre os t_{max} vizinhos: chamada de $\mathbf{b}^{ii}$. Assim como na perturbação, a estrutura de vizinhança nesta fase é a mesma (vide Figura 1), porém a amplitude da busca local é dada pelo parâmetro r_2, em que este é independente de r_1, podendo ser igual ou diferente. Supondo $t_{max} = 2$, produz-se duas soluções candidatas para ser $\mathbf{b}^{ii}$. Então, aplica-se a alocação ótima proposta em [5] e o menor valor de n encontrado dentre eles determinará o novo vetor solução $\mathbf{b}^{ii}$. Se a solução produzida $\mathbf{b}^{ii}$ determina um tamanho de amostra menor ou igual ao tamanho amostral associado à solução corrente, faz-se a atualização da solução e $k = 1$ novamente. Caso contrário, faz-se $k = k + 1$, até atingir k_{max}. Para $k > 1$, permite-se que a busca nos elementos livres não seja na mesma direção.

Após essa etapa, aplica-se o procedimento do Path-Relinking, criando-se um conjunto elite das melhores soluções obtidas (*pool*), que começa vazio e vai sendo adicionada à solução corrente, até atingir seu tamanho máximo (p_{max}). Nessa etapa são consideradas todas as combinações das p_{max} soluções do pool tomadas duas a duas, $\binom{p_{max}}{2}$ combinações, em que trata-se uma solução como a inicial e a outra como a final. A melhor de todas as soluções intermediárias (menor tamanho amostral) é mantida como resposta do algoritmo e chamada de solução

$\mathbf{b}^{iii}$. Se a solução $\mathbf{b}^{iii}$ determina um tamanho de amostra menor ou igual ao tamanho amostral associado à solução $\mathbf{b}$ corrente, faz-se a substituição.

O algoritmo EVP, implementado em linguagem R e disponível no pacote *stratvns*[5], termina quando pelo menos um dos três critérios de parada for satisfeito: número máximo de iterações (i_{max}), número de iterações sem redução no tamanho de amostra ($notBest$) e tempo máximo de processamento ($cpuTime$).

4 Experimentos Numéricos

A maioria dos trabalhos propostos na literatura considerando os dois objetivos para resolução do problema de estratificação, utilizam populações disponíveis em pacotes (instâncias) estatísticos do R ou geradas artificialmente a partir de distribuições estatísticas. Para os experimentos realizados foram utilizadas 25 populações[6], com as mais variadas características, desde muito pequenas com apenas $N = 284$, até muito grandes com $N = 16.057$, também com valores para w variando de 51 a 6.405 e com assimetria variando de $-0,7$ a $22,2$ e ainda, uma população que apresenta valores negativos.

De forma a avaliar os resultados produzidos pelo algoritmo proposto, foram utilizados, para comparação, os dois algoritmos que têm apresentado os melhores resultados na literatura, descritos em [10] e [11]. O primeiro é um algoritmo de busca aleatória em que há dois critérios de parada possíveis: quando atingir um determinado número de iterações ou quando ocorrer um determinado número de iterações sem melhoria da função objetivo. Enquanto o algoritmo de [11] baseia-se e mcálculos numéricos iterativos, considerando um erro pré-determinado para a aproximação.

Contudo, ambos apresentam uma limitação, pois não foram concebidos para contemplar restrições adicionais, como a restrição (5) de tamanho amostral mínimo por estrato e, portanto, uma adaptação foi necessária. Os dois algoritmos estão implementados em linguagem R na função *strata.LH* disponível no pacote stratification. Como essa função não está livre para edição, fez-se uma adaptação, em que optou-se por ignorar essa restrição inicialmente. Para assim, utilizar essa função apenas para gerar o vetor $\mathbf{b}$, e a partir desses resultados, calcular novos tamanhos amostrais que atendam a essa restrição. Todavia isso, pode ocasionar um coeficiente de variação (CV) maior que o fixado.

As abreviações LH88, Ko04, EVP referem-se, respectivamente, aos algoritmos de [11], [10] e ao novo algoritmo proposto. Para os dois primeiros, utilizou-se a alocação de Neyman, enquanto no algoritmo EVP, utilizou-se a alocação proposta por [5], implementada no pacote *MultAlloc*.[7] Os três algoritmos foram

[5] https://cran.r-project.org/web/packages/stratvns/index.html

[6] Essas populações são bem reconhecidas, a saber: BeeFarms, beta103, chi1, chi5, debtors, hhinctot, iso2004,Kozak1, Kozak2, Kozak3, Kozak4, me84, mrts, p100e10, p75, pop800, rev84, SugarCaneFarms, Swiss, TaxableIncome, Usbanks, Uscities, Uscolleges, Rchisq2_30, M101.

[7] http://cran.r-project.org/web/packages/MultAlloc/index.html

executados em um computador dotado de 6GB de memória RAM, com processador i7 de 2.2GHz com 4 núcleos e sistema operacional Windows de 64 bits.

Para cada uma das 25 populações da literatura foram calculados novos limites para os pontos de corte, visando a minimização do tamanho amostral total, a partir da aplicação dos três algoritmos supracitados. Foram testados ainda, os resultados para o número de estratos (L) variando de 3 a 6. E ainda, utilizando a restrição usual ($n_h \geq 1$) e a restrição $min\{5, N_h\} \leq n_h \leq N_h$. Portanto, no total foram produzidos 600 resultados (25 populações x 3 algoritmos x 4 números de estratos x 2 tamanhos para n_h). Para os outros parâmetros do problema de estratificação foram utilizados: CV $\leq 10\%$, $N_h \geq 2$, $n_L = N_L$.

Para o algoritmo proposto neste trabalho, outros parâmetros adicionais também são necessários. Após testes preliminares com algumas combinações de valores para esses parâmetros, optou-se por utilizar os seguintes valores: número de vizinhos máximo $t_{max} = 7$, amplitude da perturbação $r_1 = 30$, amplitude da busca local $r_2 = 20$, tamanho máximo de soluções elite $p_{max} = 5$, número máximo de iterações $i_{max} = 150$, número de iterações sem melhoria $notBest = 25$ e tempo máximo de processamento $cpuTime = 5.000$ segundos.

Nas Tabelas 1 e 2 são apresentados os resultados produzidos por cada algoritmo (LH88, Ko04, EVP) com o número de estratos variando de 3 a 6 das 25 populações consideradas. Em que os resultados da Tabela 1 são para o tamanho amostral mínimo por estrato sem restrição, ou seja, $n_h \geq 1$ e os da Tabela 2 são para a restrição (5), ou simplesmente, $n_h \geq 5$.

Embora, houvesse a possibilidade dos algoritmos de [10] adaptado e de [11] adaptado, pudessem gerar soluções inválidas que ultrapassariam o limite de CV de 10%, esse fato não ocorreu para as populações consideradas desse experimento empírico. Note, ainda, que na população U25, os dois algoritmos da literatura não puderam ser aplicados (indicados nas Tabelas por "N/A"), pois essa população apresenta valores negativos, e portanto, apenas o algoritmo proposto foi capaz de produzir resultados.

Para uma fácil identificação da melhor solução produzida para cada população, optou-se por grifá-la de negrito e denominá-la por *solução vencedora* (menor tamanho amostral de cada população produzida por um dos três algoritmos, que atenda a todas as restrições). Por exemplo, na população U01 da Tabela 1 e $L = 6$, a solução vencedora veio do algoritmo EVP. Note, também, que o empate é permitido, por exemplo, nessa mesma população U01, mas para $L = 3$, todos os três algoritmos produziram a solução vencedora.

A partir das soluções vencedoras, grifadas em negrito nas Tabela 1, observa-se que o algoritmo EVP teve um desempenho melhor do que Ko04 para $L = 3$, pois produziu a solução vencedora em todas populações, enquanto o algoritmo Ko04 produziu a solução vencedora em 24 populações e o algoritmo LH88 apenas em 18 populações. Nos demais estratos, os algoritmos Ko04 e EVP produziram resultados quase equivalentes, com leve vantagem para Ko04.

Por outro lado, na Tabela 2, observa-se que o algoritmo EVP teve um desempenho muito superior aos demais, pois produziu a maior quantidade de soluções vencedora para todos os números de estratos. O algoritmo proposto só não ven-

Tabela 1. Tamanho Amostral Total (n) produzido por cada algoritmo e número de estratos(L) das 25 populações, para $n_h \geq 1$

ID	L=3			L=4			L=5			L=6		
	LH88	Ko04	EVP	LH88	Ko04	EVP	LH88	Ko04	EVP	LH88	Ko04	EVP
U01	**29**	**29**	**29***	21	**20**	**20**	20	**13**	14	15	10	**9**
U02	**4**	**4**	**4***	**5**	**5**	**5**	**6**	**6**	**6**	**7**	**7**	**7**
U03	**37**	**37**	**37***	20	**19**	**19**	13	**12**	**12**	10	**9**	10
U04	**15**	**15**	**15***	10	**9**	10	8	**6**	**6**	**7**	**7**	**7**
U05	**60**	**60**	**60***	33	**32**	33	21	**20**	**20**	15	**13**	15
U06	**23**	**23**	**23**	13	13	**12**	11	9	**8**	8	**7**	**7**
U07	**24**	**24**	**24***	**14**	**14**	**14**	11	**10**	**10**	10	**7**	8
U08	9	**8**	**8***	6	**5**	**5***	7	**6**	**6***	**7**	**7**	**7**
U09	**21**	**21**	**21**	**11**	**11**	13	10	**8**	10	**7**	**7**	8
U10	**31**	**31**	**31***	17	**16**	**16**	**10**	**10**	**10**	9	**7**	**7**
U11	**42**	**42**	**42**	21	**20**	22	14	**13**	**13**	12	**9**	12
U12	18	**17**	**17***	10	**9**	**9**	9	**7**	**7**	9	**8**	**8**
U13	**32**	**32**	**32**	18	**17**	**17**	12	**11**	12	10	**8**	**8**
U14	**4**	**4**	**4***	**5**	**5**	**5**	**6**	**6**	**6**	**7**	**7**	**7**
U15	19	**18**	**18***	**10**	**10**	**10***	9	**7**	**7***	10	**7**	**7**
U16	**22**	**22**	**22***	17	**15**	**15**	**10**	**10**	**10**	10	**8**	**8**
U17	17	**16**	**16***	**10**	**10**	**10**	8	**7**	**7**	8	**7**	**7**
U18	**11**	**11**	**11***	8	**7**	**7***	7	**6**	**6**	**7**	**7**	**7**
U19	**31**	**31**	**31***	17	**16**	**16**	12	**11**	**11**	9	**8**	**8**
U20	**26**	**26**	**26***	16	**15**	**15**	**10**	**10**	**10**	9	**8**	**8**
U21	16	**15**	**15***	**9**	**9**	**9**	9	**6**	**6**	**7**	**7**	**7**
U22	**18**	**18**	**18***	13	**11**	**11***	8	**7**	**7**	9	**7**	**7**
U23	**24**	**24**	**24***	14	**12**	**12**	9	**8**	**8**	8	**7**	**7**
U24	6	**5**	**5***	**5**	**5**	**5**	**6**	**6**	**6**	**7**	**7**	**7**
U25	N/A	N/A	**51**	N/A	N/A	**33**	N/A	N/A	**25**	N/A	N/A	**23**

N/A: Não se Aplica

* Tamanho Mínimo Amostral ótimo

ceu em oito dos 100 resultados produzidos, aparentando ser o algoritmo mais adequado para a resolução do problema de estratificação quando há a restrição (5), pois apresenta resultados de qualidade superior do que os resultados obtidos pelos algoritmos concorrentes.

Considerando as 25 populações literatura e os dois tipos de restrições ($n_h \geq 1$ e $n_h \geq 5$), ou seja, todos os resultados das Tabelas 1 e 2, calculou-se o percentual de soluções vencedoras para cada número de estrato (L) testado, apresentado na Figura 2. Portanto, o algoritmo proposto foi o que apresentou os melhores resultados, variando de 82% a 96% na capacidade de produzir o melhor resultado para a população .

Ainda na mesma figura, tem-se o percentual total com base nos 200 resultados (25 populações x 4 números de estratos x 2 tamanhos mínimo para n_h) que cada algoritmo produziu. Assim, de modo geral, nesse estudo empírico, o algoritmo proposto foi capaz de produzir a solução vencedora em 90% dos casos, enquanto

Tabela 2. Tamanho Amostral Total (n) produzido por cada algoritmo e número de estratos(L) das 25 populações, para $n_h \geq 5$

ID	L=3			L=4			L=5			L=6		
	LH88	Ko04	EVP	LH88	Ko04	EVP	LH88	Ko04	EVP	LH88	Ko04	EVP
U01	**29**	**29**	**29***	28	23	**21**	33	**22**	**22**	30	27	**25**
U02	12	12	**9***	17	17	**14**	22	22	**16**	27	27	**21**
U03	**37**	**37**	**37***	20	**19**	20	22	22	**21**	**27**	**27**	**27**
U04	**15**	**15**	**15***	**17**	**17**	**17**	22	22	**19**	27	27	**21**
U05	**60**	**60**	**60***	33	**32**	33	**22**	**22**	**22**	27	27	**24**
U06	**23**	**23**	**23**	**17**	**17**	**17**	**22**	**22**	**22**	**27**	**27**	**27**
U07	**24**	**24**	**24***	**17**	**17**	**17**	22	22	**19**	27	27	**24**
U08	**12**	**12**	**12***	17	17	**14***	22	22	**16***	27	27	**21**
U09	**21**	**21**	22	**17**	**17**	18	**22**	**22**	23	27	27	**25**
U10	**31**	**31**	**31***	**17**	**17**	**17**	**22**	**22**	**22**	27	27	**24**
U11	**42**	**42**	**42**	21	**20**	21	**22**	**22**	**22**	**27**	**27**	**27**
U12	18	**17**	**17***	18	18	**17**	23	23	**20**	28	28	**22**
U13	**32**	**32**	33	**18**	**18**	19	**22**	**22**	**22**	**27**	**27**	**27**
U14	12	12	**9***	17	17	**14**	**22**	**22**	**22**	**27**	**27**	**27**
U15	19	**18**	**18***	18	18	**17***	23	23	**18***	28	27	**19**
U16	**22**	**22**	**22***	24	**18**	**18**	**23**	**23**	**23**	28	28	**27**
U17	17	**16**	**16***	**18**	**18**	**18**	23	22	**19**	27	27	**21**
U18	**12**	**12**	**12***	17	17	**14***	22	22	**16**	27	27	**20**
U19	**31**	**31**	**31***	**17**	**17**	**17**	**22**	**22**	**22**	**27**	**27**	**27**
U20	**26**	**26**	**26***	19	18	**17**	**22**	**22**	**22**	27	27	**24**
U21	16	**15**	**15***	**17**	**17**	**17**	22	22	**19**	27	27	**21**
U22	**18**	**18**	**18***	**17**	**17**	**17***	22	22	**19**	27	27	**21**
U23	**24**	**24**	**24***	**17**	**17**	**17**	22	22	**19**	27	27	**24**
U24	12	12	**11***	17	17	**14**	22	22	**19**	27	27	**24**
U25	N/A	N/A	**53**	N/A	N/A	**32**	N/A	N/A	**22**	N/A	N/A	**35**

N/A: Não se Aplica

* Tamanho Mínimo Amostral ótimo

o algoritmo Ko04 produziu a solução vencedora em 73% dos casos e o LH88 apenas em 42% dos casos.

Entre as soluções vencedoras produzidas pelo algoritmo EVP, ainda há aquelas que foram produzidas pelo algoritmo exato, assinaladas com um asterisco nas Tabelas 1 e 2, portanto, são soluções ótimas globais. Assim sendo, doravante, esse tipo de solução será denominada apenas por *solução ótima*. Posto isso, a Tabela 3 resume as quantidades de solução ótimas produzidas pelo algoritmo EVP. E como era de se esperar, em função do critério para aplicação do procedimento exato, o número de soluções ótimas produzidas pelo algoritmo proposto foi diminuindo conforme o valor de L foi aumentando. Assim sendo, o algoritmo foi capaz de produzir 20 soluções ótimas dentre as 25 populações para $L = 3$, por outro lado não produziu nenhuma solução ótima para $L = 6$. De modo geral, o algoritmo EVP conseguiu produzir 26 soluções ótimas entre os 100 resultados (25 populações x 4 números de estratos), o que dá um aproveitamento de 26%.

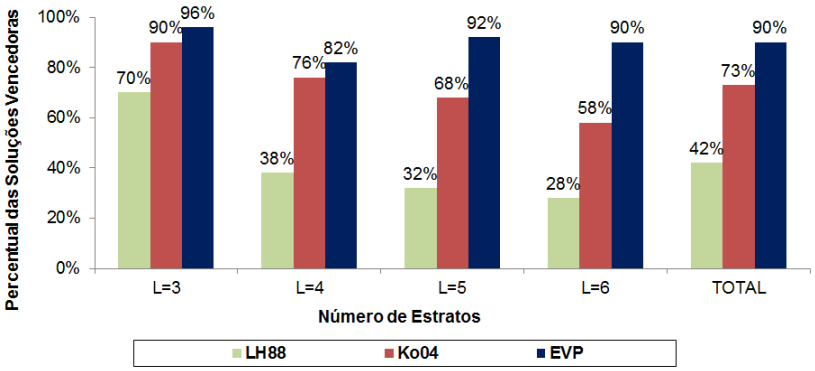

Figura 2. Percentual de soluções vencedoras

Tabela 3. Quantidade de soluções ótimas por número de estratos das 25 populações consideradas

Número de estratos (L)	3	4	5	6	Total
Quantidade de soluções ótimas	20	4	2	0	26

Vale destacar que houve alguns casos em que o algoritmo EVP produziu soluções melhores que a de seus concorrentes, mesmo sem a execução do algoritmo exato. Como, por exemplo, a população U06 para $L = 4$ e $L = 5$ na Tabela 1, e também, a população U02 da Tabela 2.

Em linhas gerais, a qualidade das soluções obtidas aqui foi significativamente melhor, devido à necessidade de uma amostra menor. O custo desta melhoria pode ser expresso pela diferença nos tempos computacionais exigidos pelos algoritmos, os quais foram da ordem de segundos para os algoritmos Ko04 e LH88 e da ordem de minutos ou até horas para o algoritmo EVP.

5 Conclusão

O problema de estratificação é estudado desde a década de 1950 e até hoje é um dos problemas estatísticos que persiste sem solução definitiva. Aqui, conseguiu-se avançar mais um passo em direção à obtenção de soluções de melhor qualidade. O algoritmo EVP foi capaz de produzir a solução vencedora em 90% dos resultados possíveis, sendo que em 26%, ainda foi capaz de garantir que a solução corresponde a um mínimo global. Quando há a restrição para o tamanho amostral mínimo por estrato, o algoritmo proposto apresentou um desempenho superior aos demais, e quando não há essa restrição, o algoritmo EVP produziu resultados superiores ao de [11] e quase equivalentes ao de [10].

Como possíveis extensões, há a possibilidade de se testar outras metaheurísticas, para auxiliar a busca como, por exemplo, GRASP, ILS, GA.

Referências

1. Barr, R.S. and Helgason, R.V. and Kennington, J.L. Interfaces in Computer Science and Operations Research. Kluwer Academic Publishers. (1996)
2. Bolfarine, Heleno and Bussab, Wilton O. Elementos de amostragem, 1st edition. Edgard Blucher. (2005)
3. Brito, J. A. M. and Montenegro, F. M. T. and Maculan, Nelson and Azevedo, Rosemary Vallejo. Propostas para o problema de estratificação em amostras considerando alocação proporcional. XXXIX Simpósio Brasileiro de Pesquisa Operacional (SBPO). (2006)
4. Brito, J. A. M. and Maculan, N. and Montenegro, F. M. T. and Brito, L. R. Um algoritmo GRASP aplicado ao problema de estratificação XLIII Simpósio Brasileiro de Pesquisa Operacional (SBPO). (2011)
5. Brito, J. A. M. and Silva, Pedro Luis Nascimento and Semaan, Gustavo Silva and Maculan, Nelson. Integer programming formulations applied to optimal allocation in stratified sampling. Survey Methodology 41(2), 427–442. (2015)
6. Cochran, William G. Sampling Techniques, 3rd edition. John Wiley & Sons. (1977)
7. Gunning, Patricia and Horgan, Jane. A new algorithm for the construction of stratum boundaries in skewed populations. Statistics Canada 2(30), 159–166. (2004)
8. Hansen, Pierre and Mladenovic, Nenad and Perez-Brito, Dionisio. Variable Neighborhood Decomposition Search. Journal of Heuristics 7(4), 335–350. (2001)
9. Hedlin, Dan. A procedure for stratification by an extended ekman rule. Journal of Official Statistics 1(16), 15–29. (2000)
10. Kozak, Marcin. Optimal stratification using random search method in agricultural surveys. Statistics in Transition 6(5), 797–806. (2004)
11. Lavallee, Pierre and Hidiroglou, Michel A. On the stratification of skewed populations. Survey Methodology 14, 33–43. (1988)
12. Lohr, Sharon L. Sampling: Design and Analysis, 2nd edition. (2010)