METAHEURÍSTICA VNDS APLICADA AO PROBLEMA DE ESTRATIFICAÇÃO ÓTIMA

Breno Tiago Novello Trotta de Oliveira
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Av. Maracanã, 229 - Maracanã, Rio de Janeiro – RJ
brenotrotta@yahoo.com.br

Leonardo Silva de Lima
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Av. Maracanã, 229 - Maracanã, Rio de Janeiro – RJ
leolima.geos@gmail.com

José André de Moura Brito
Escola Nacional de Ciências Estatísticas (ENCE/IBGE)
Rua André Cavalcanti, 106 - Centro, Rio de Janeiro – RJ
jambrito@gmail.com

RESUMO

O problema de estratificação ótima está associado à área de amostragem probabilística. Nesse problema deve-se delimitar os estratos populacionais e definir a alocação da amostra, de forma a minimizar ou a variância de um estimador ou o tamanho amostral. Em função de sua relevância prática e complexidade computacional, diversos métodos heurísticos têm sido propostos na literatura. Neste trabalho, foi proposta uma versão da metaheurística VNDS para resolução do problema de estratificação ótima, com o objetivo de minimizar o tamanho da amostra, considerando um nível de precisão fixado. O método proposto foi comparado com dois métodos bem conhecidos da literatura, produzindo o melhor resultado para 83% das populações consideradas.

PALAVRAS CHAVE. Estratificação, VNDS, Amostragem

Tópico (MH - Metaheuristicas)

ABSTRACT

The optimum stratification problem is associated to the probability sampling area. This problem is to delimit the population strata and defining the allocation of sample in order to either minimize a variance of an estimator or minimize the sample size. Due to the practical relevance and the computational complexity of this problem, several heuristic methods have been proposed in the literature. In this paper, we proposed a version of the VNDS metaheuristic to solve the optimum stratification problem in order to minimize the sample size given a fixed level of accuracy. Our method was compared and outperformed two well-known methods of the literature in most of the cases, getting the best results for 83% of the considered populations.

KEYWORDS. Stratification, VNDS, Sampling

Paper topic (MH - Metaheuristics)

1. Introdução

O problema de estratificação ótima está associado à área de amostragem probabilística. Por sua vez, a amostragem corresponde ao processo de seleção de unidades de um subconjunto de uma população, de modo que se possa inferir as características de interesse da população com um certo grau de acurácia, conforme [Lohr 2010].

Os métodos de estratificação propostos na literatura não incorporam a restrição de tamanho amostral mínimo por estrato, algo muito importante para algumas pesquisas amostrais realizadas no âmbito do Instituto Brasileiro de Geografia e Estatística (IBGE). Essa restrição é incorporada às principais pesquisas anuais como, por exemplo, na Pesquisa Anual do Comércio (PAC), na Pesquisa Anual de Serviços (PAS), na Pesquisa Anual da Indústria da Construção (PAIC) etc.

O problema de estratificação pode ser formulado considerando dois objetivos possíveis, a saber: minimizar a variância de um estimador (considerando o tamanho de amostra fixo) e minimizar o tamanho amostral (considerando o nível de precisão fixo). A maioria dos métodos propostos na literatura foram desenvolvidos para atender ao primeiro objetivo, como [Ekman 1959], [Dalenius e Hodges 1959], [Gunning e Horgan 2004], [Keskintürk e Er 2007], [Brito et al. 2011], [Hedlin 2000], enquanto o segundo objetivo foi abordado em [Hidiroglou 1986], [Kozak 2004], [Lavallée e Hidiroglou 1988], [Brito e Montenegro 2007]. Vale ressaltar que o segundo objetivo é bem menos estudado na literatura, o que consiste em uma motivação para o desenvolvimento dessa pesquisa.

Com o objetivo de atender ao segundo objetivo do problema, propõe-se nesse trabalho um novo método baseado na metaheurística Variable Neighborhood Decomposition Search (VNDS) [Hansen et al. 2001] e que incorpora a restrição de tamanho amostral mínimo por estrato, considerando para isso, a alocação proposta por [Brito et al. 2015]. Esse método foi desenvolvido para solucionar o problema de estratificação associado ao planejamento amostral da PAC.

As primeiras abordagens para o problema de estratificação ótima remetem à década de 50, sendo propostas novas abordagens até os dia atuais. A partir da utilização da metaheurística VNDS e da alocação ótima de [Brito et al. 2015], foram produzidas soluções de boa qualidade para esse problema. A contribuição do trabalho está no algoritmo proposto, que se apresenta como uma alternativa aos métodos existentes na literatura à resolução do problema de estratificação, considerando o objetivo de minimizar o tamanho amostral.

O presente trabalho está dividido da seguinte forma: a seção dois traz os conceitos básicos sobre a amostragem e os principais conceitos sobre a amostragem estratificada, incluindo uma descrição detalhada sobre o problema de estratificação ótima e o plano amostral da PAC. A seção três traz uma descrição sucinta da metaheurística VNDS e sua aplicação à PAC. A seção quatro traz os resultados computacionais obtidos nesse trabaho e comparados aos algoritmos propostos por [Lavallée e Hidiroglou 1988] e [Kozak 2004]. E por fim, na seção cinco são apresentadas as conclusões e os possíveis desdobramentos desse estudo.

2. Amostragem e Problema de Estratificação

2.1. Conceitos Básicos de Amostragem

Uma pesquisa tem o objetivo de coletar informações sobre características de interesse das unidades de uma população. O ideal seria a realização de uma pesquisa com cada unidade da população[1], mas na prática, essa forma de realização da pesquisa dificilmente ocorre, seja por problemas geográficos, logísticos ou até mesmo de custo operacional. Assim, utiliza-se da amostragem para que um subconjunto de unidades seja escolhido da melhor forma para representar a população como um todo. E, consequentemente, gerar bons estimadores (totais, médias, proporções) para que se possa inferir sobre características de interesse da população. Segundo [Cochran 1977], dentre as vantagens de se utilizar amostragem em vez da enumeração completa da população, estão a redução dos custos e a coleta de dados de forma mais rápida.

[1] Também chamado de Censo.

Segundo [Lohr 2010], as características desejáveis de uma amostra são: boa representatividade, pois cada unidade da amostra irá representar as características de um número conhecido de unidades na população; inexistência do viés de seleção, que ocorreria se alguma característica da população não fosse encontrada na amostra; e minimização do erro de medição.

Como os parâmetros populacionais, em geral, são desconhecidos, utilizam-se os estimadores (uma função dos dados amostrais) para produzir estimativas a partir das variáveis de interesse da população, tais como: do total, da média e da variância.

A maioria das pesquisas realizadas pelos órgãos de estatística oficial no Brasil e no mundo utilizam a amostragem probabilística. Mais especificamente, nesse tipo de amostragem os elementos da população têm uma probabilidade maior que zero para serem selecionados na amostra e há algum mecanismo de aleatorização para a seleção dos elementos. Os exemplos mais comuns de amostragem probabilística são: amostragem aleatória simples (AAS), amostragem estratificada, amostragem sistemática e amostragem por conglomerados; para mais detalhes ver [Cochran 1977].

2.2. Amostragem Estratificada

Conforme [Cochran 1977] e [Lohr 2010], a Amostragem Estratificada (AE) consiste em particionar a população (U) de N unidades 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, de modo que: $N = \sum_{h=1}^{L} N_h$. Uma vez determinados os estratos populacionais, seleciona-se uma amostra em cada um deles (geralmente por meio de uma AAS), sendo as seleções feitas independentemente nos diferentes estratos. Os tamanhos das amostras dentro dos estratos são denotadas por $n_1, n_2, \ldots, n_L$, respectivamente, de modo que: $n = \sum_{h=1}^{L} n_h$. Algumas motivações para o uso da estratificação são: melhoria da precisão das estimativas; representar os diferentes grupos dentro de uma população, ou em outras palavras, garantir o espalhamento da amostra; questões administrativas.

Para uma variável de interesse Y, o estimador do total populacional considerando esse método de amostragem é denotado por $\hat{Y}_{AE}$. Segundo [Lohr 2010], este estimador é definido por

$$\hat{Y}_{AE} = \sum_{h=1}^{L} N_h \overline{y}_h, \tag{1}$$

sendo que $\overline{y}_h = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$ é a média amostral no h-ésimo estrato e y_{hj} é o valor da variável de interesse Y, para a j-ésima unidade do h-ésimo estrato. A variância do estimador de total é dada por

$$V(\hat{Y}_{AE}) = \sum_{h=1}^{L} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}, \tag{2}$$

sendo que $S_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} (y_{hj} - \overline{Y}_h)^2$ é a variância populacional no h-ésimo estrato e $\overline{Y}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$ é a média populacional no h-ésimo estrato.

Outra etapa da AE é o procedimento que consiste na distribuição das n unidades da amostra nos L estratos, chamado de alocação da amostra e denotado por a_h, tal que $n_h = n \cdot a_h$. A escolha do procedimento de alocação é muito importante porque dela depende a precisão dos estimadores. Os principais tipos são: alocação Proporcional, alocação Uniforme, alocação Ótima e alocação de Neyman, que podem ser encontradas em [Cochran 1977, Bolfarine e Bussab 2005]. Outra alocação bastante usual na literatura é a alocação Potência proposta por [Bankier 1988]. Porém, a alocação mais utilizada é a de Neyman, pois ela tende a produzir a menor variância associada ao estimador de total. Entretanto, assim como os outros métodos, essa alocação pode produzir tamanhos amostrais não inteiros, sendo necessário fazer arredondamentos, o que não é desejável. Além disso, em algumas situações, a alocação de Neyman pode produzir tamanhos amostrais maiores que os tamanhos

populacionais no estrato, ou seja, $n_h > N_h$. Recentemente, [Brito et al. 2015] propuseram uma formulação de programação inteira que produz o ótimo global para o problema de alocação.

Uma vez definido o método de alocação, segundo [Cochran 1977], o cálculo do tamanho de amostra total (n) é dado por

$$n = \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{V(\hat{Y}_{AE}) + \sum_{h=1}^{L} N_h S_h^2}. \tag{3}$$

2.3. Definição do Problema de Estratificação

Seja uma variável auxiliar de medida de tamanho X correlacionada à variável de interesse Y. Com essas duas informações, pode-se utilizar um outro método de amostragem que é uma variação da AE, a Amostragem Estratificada por Cortes (AEC) que difere da primeira, apenas por um ponto: no último estrato todas as unidades da população compõem a amostra, tal que $n_L = N_L$. Nesse método, usualmente, a variável utilizada para a estratificação é a variável auxiliar X, ao invés da variável de interesse Y.

Como no último estrato (L) estão as maiores unidades[2] e todas devem ser incluídas com certeza na amostra, esse estrato é chamado de *estrato certo*, enquanto os demais $(L-1)$ estratos são chamados de *estratos amostrados*, pois será selecionada uma amostra de cada estrato, utilizando-se de algum método de seleção, em geral uma AAS. A AEC é utilizada quando a população de estudo apresenta uma alta assimetria na variável de interesse e/ou na variável auxiliar.

A dificuldade do problema está em definir os pontos de corte que serão utilizados para destacar o estrato certo e delimitar os estratos amostrados. Assim, deve-se procurar os pontos de corte ótimos que minimizem a função objetivo do problema.

O procedimento de estratificação por corte, pode ser definido, matematicamente, conforme [Azevedo 2004]. Considere as definições: seja $Y_U = \{y_1, y_2, \ldots, y_N\}$ o vetor populacional relacionado à variável de interesse Y e $X_U = \{x_1, x_2, \ldots, x_N\}$ o vetor populacional relacionado à variável de estratificação X, sabendo que $x_1 \leq x_2 \leq \ldots \leq x_N$. Essas unidades são alocadas aos L estratos, segundo os pontos de corte $b_1 < b_2 < \ldots < b_{L-1}$. Os estratos são definidos segundo:

$$E_1 = \{i \in U : x_i \leq b_1\}, \tag{4}$$

$$E_h = \{i \in U : b_{h-1} < x_i \leq b_h\}, \quad h = 2, 3, \ldots, L-1 \tag{5}$$

e

$$E_L = \{i \in U : x_i > b_{L-1}\}. \tag{6}$$

Assim, para a construção de L estratos são necessários $(L-1)$ pontos de corte, e toda unidade $i \in U$ que apresentar um valor para x_i menor ou igual que b_1 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. Como comentado anteriormente, utiliza-se a variável auxiliar X como variável de estratificação, tanto para fazer o procedimento acima como para calcular os estimadores, pois, em geral, as informações da variável de interesse Y são desconhecidas.

O estimador do total populacional associado à X é dado por

$$\hat{X}_{AEC} = X_L + \sum_{h=1}^{L-1} N_h \overline{x}_h, \tag{7}$$

[2]Em relação à variavel de estratificação utilizada.

sendo $\overline{x}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} x_{hj}$ a média amostral de X no h-ésimo estrato, tal que, x_{hj} é o valor da variável auxiliar X para a j-ésima unidade amostral do h-ésimo estrato e $X_L = \sum_{j=1}^{N_L} x_{Lj}$ é o total populacional do estrato certo (L). A variância e o coeficiente de variação do estimador de total, respectivamente, são dados por

$$V(\hat{X}_{AEC}) = \sum_{h=1}^{L-1} N_h^2(1 - \frac{n_h}{N_h})\frac{S_h^2}{n_h}, \tag{8}$$

$$CV(\hat{X}_{AEC}) = \frac{\sqrt{V(\hat{X}_{AEC})}}{X}. \tag{9}$$

As etapas associadas à resolução do problema de estratificação ótima podem ser resumidas da seguinte forma:

1. Encontre $(L-1)$ pontos de corte $(b_1, ..., b_{L-1})$, para poder dividir a população em L estratos;
2. Determine o objetivo do problema;
3. Determine a forma de alocar os tamanhos amostrais $n_1, n_2, ..., n_L$;
4. Selecione as unidades dentro de cada estrato de acordo com algum método de seleção e de acordo com o tamanho amostral por estrato (n_h).

Na etapa 1 utiliza-se a variável de estratificação para encontrar os pontos de corte que visem a minimizar a função objetivo. A etapa 2 consiste em definir a função objetivo considerando dois objetivos possíveis: (i) minimizar a variância do estimador de total, dado um tamanho de amostra fixado; e (ii) minimizar o tamanho amostral, considerando a precisão fixada previamente. A maioria dos métodos propostos na literatura foram concebidos para atender ao primeiro objetivo, enquanto o segundo objetivo foi menos explorado na literatura. Observe que os problemas associados a esses dois objetivos estão correlacionados, pois se o objetivo for minimizar a variância, o tamanho amostral tem que ser um dado de entrada do problema, enquanto se o objetivo for minimizar o tamanho amostral, a precisão (a variância ou o coeficiente de variação) tem que ser um dado de entrada do problema. A etapa 3 consiste em utilizar algum dos métodos de alocação, sendo a de Neyman o mais comum entre os trabalhos. E por fim, a etapa 4 consiste em escolher um dos métodos de amostragem probabilística, em que o mais usual é a AAS.[3]

Além disso, a etapa 1 corresponde ao primeiro nível do problema de estratificação, em que ao se definir os pontos de corte, pode-se calcular diretamente os valores para N_h e S_h^2. E a etapa 3 corresponde ao segundo nível do problema de estratificação, em que os tamanhos amostrais são alocados por estratos (n_h) de forma a somar o tamanho amostral total (n), considerando o objetivo especificado na etapa 2.

2.4. O problema de Estratificação da PAC

A Pesquisa Anual de Comércio (PAC), realizada pelo IBGE, tem por finalidade levantar informações referentes às características estruturais básicas do comércio varejista e atacadista do Brasil e visa fornecer estimativas de pessoal ocupado (PO), salários, receita, entre outras variáveis, segundo domínios definidos por níveis geográficos em combinação com classes de atividades da Classificação Nacional de Atividades Econômicas (CNAE).

Todas as empresas com 20 ou mais pessoas ocupadas são consideradas grandes e por isso, são classificadas como pertencentes ao estrato certo (C). As demais empresas são classificadas no estrato amostrado (A), que ainda se subdivide em três, também de acordo com o porte da empresa, a saber: A_1, empresas de 0 a 4 PO; A_2, empresas de 5 a 9 PO; A_3, empresas de 10 a 19 PO. Nesses

[3]Essa quarta etapa está presente em qualquer procedimento de estratificação.

estratos utiliza-se amostragem aleatória simples sem reposição das unidades elementares, que são as empresas. Por haver uma grande concentração de pequenas empresas e menor concentração de grandes empresas, comportamento típico de uma distribuição assimétrica à direita, o uso da AEC na PAC se justifica.

Os tamanhos amostrais da PAC, conforme descrito em [IBGE 2015], são calculados de forma a assegurar que o estimador do total de pessoal ocupado ($\hat{X}_{AEC}$) em cada estrato natural tenha um coeficiente de variação associado menor ou igual a 0,1. A amostra de empresas é obtida por uma AAS em cada estrato final amostrado, A_h $(h = 1, 2, 3)$, e pela inclusão das empresas pertencentes aos estratos finais certos, tal que $n = n_C + n_{A_1} + n_{A_2} + n_{A_3}$, ou resumidamente $n = n_C + n_A$, sendo $n_C = N_C$. Além disso, arbitrou-se um número mínimo[4] de cinco empresas em cada estrato final amostrado, ou seja, $n_{A_h} \geq 5$ $(h = 1, 2, 3)$. Assim, para chegar ao tamanho de amostra do estrato final, basta calcular n_A (pois n_C já é conhecido). Para isso, eleva-se ao quadrado os termos da Equação (9) e utiliza-se a alocação de Neyman ($a_h = N_h S_h / \sum_h N_h S_h$) para substituí-las na Equação (3), após alguns algebrismos, tem-se:

$$n_A = \frac{N_A^2 (\sum_{h=1}^{3} W_h S_h)^2}{CV^2(\hat{X}_{AEC}) \cdot X^2 + N_A \sum_{h=1}^{3} W_h S_h^2}, \tag{10}$$

sendo que $W_h = N_h / N_A$ é a proporção populacional do h-ésimo estrato final amostrado, $X = X_A + X_C$ é o total populacional do pessoal ocupado no estrato natural e $CV(\hat{X}_{AEC})$ é o coeficiente de variação do estimador do total do pessoal ocupado em cada estrato natural – Equação (9).

Conforme comentado anteriormente, é necessário incluir uma restrição que não é contemplada nos algoritmos disponíveis na literatura, pois é preciso definir um limite mínimo para o tamanho amostral por estrato final que, de acordo com a metodologia da PAC, foi fixado em 5. Ou seja, o tamanho amostral do estrato (n_h) deve pertencer ao intervalo fechado $[5, N_h]$, só sendo permitido ser menor que 5, quando $N_h < 5$.[5] O valor prefixado para o CV é de 10% e a função objetivo resume-se ao tamanho amostral total (n).

A amostra da PAC pode ser otimizada visando sua redução, com a definição de novos pontos de corte no primeiro nível do problema de estratificação e no segundo nível, com a utilização da alocação proposta por [Brito et al. 2015] . Assim, o problema discutido nesse trabalho resume-se ao primeiro nível, ou seja, encontrar novos pontos de corte com o intuito de minimizar o tamanho amostral.

Até o momento, ainda não existe um método para encontrar o ótimo global desse problema de minimização, a não ser um método exaustivo, que considera todas as combinações possíveis de soluções. Entretanto, calcular todas as combinações possíveis para os pontos de corte para delimitação dos estratos é algo inviável em populações grandes. Por isso, ao lidar com esse tipo de problema de alta complexidade computacional, deve-se procurar desenvolver algoritmos baseados em metaheurísticas.

3. Metodologia

3.1. Metaheurística VNDS

A metaheurística da Busca Decomposta em Vizinhança Variável (tradução livre de *Variable Neighborhood Decomposition Search* – VNDS) proposta por [Hansen et al. 2001], corresponde

[4]Essa restrição é para evitar que determinados estratos sejam representados por pouquíssimas empresas. E se por acaso, essas poucas empresas não responderem ao questionário, tem-se um estrato sem informação nenhuma. Essa cautela é necessária, pois em alguns estratos específicos a taxa de não-resposta é maior que o esperado. A não-resposta corresponde a um questionário que foi a campo, mas que não foi respondido pela empresa, devido a inúmeros motivos.

[5]Essa restrição é válida apenas para os estratos finais amostrados.

a uma variação do método *Variable Neighborhood Search* (VNS) de [Mladenović e Hansen 1997], em que decompõe o problema de otimização em dois níveis. A decomposição consiste em fixar todos os atributos (ou variáveis), exceto por k atributos. Assim, tanto as estruturas de vizinhanças quanto à busca local são definidas em subproblemas (problemas de tamanhos menores do que o inicial).

Segundo [Hansen et al. 2001], nota-se que a única diferença entre o VNS e o VNDS é na busca local, pois enquanto o primeiro explora todo o espaço de busca $\mathcal{S}$ (partindo de $s_1 \in N_k(s)$), o segundo resolve a cada iteração um subproblema em algum subespaço $V_k \subseteq N_k(s)$, com $s_1 \in V_k$. Assim, uma sequência de subproblemas é gerada a partir de um diferente conjunto de vizinhanças. Se a solução do subproblema não conduz a uma melhoria, a vizinhança é alterada. Caso contrário, a busca é reiniciada da primeira vizinhança, conforme Algoritmo 1.

```
1  Entrada:
2  Selecione o conjunto de estruturas de vizinhanças N_k (k = 1, ..., k_max) que serão usadas na
     busca;
3  s = s_0;                                          < Encontre uma solução inicial >
4  Enquanto Não Satisfizer o Critério de Parada faça
5      k = 1;
6      Enquanto k ≤ k_max faça
7          s_1 = Perturbação(s);                     < s_1 difere de s por k atributos >
8          s_2 = Busca Local(s_1, k atributos);      < Busca somente nos k atributos permitidos >
9          Se f(s_2) < f(s) então
10             s = s_2;
11             k = 1;
12         Senão
13             k = k + 1
14         fim Se
15     fim Enquanto
16 fim Enquanto
17 Saída: Melhor Solução
```

Algoritmo 1: Pseudo Código do VNDS

Em relação ao Algoritmo VNS a diferença está nas linhas 7 e 8. Como só é permitido mexer em k variáveis, a perturbação de s produzirá uma solução s_1 muito similar a s, diferindo apenas por k atributos. E a busca local segue o processo descrito no parágrafo anterior, em que fará a busca em subespaços de $\mathcal{S}$. Conforme [Hansen et al. 2001] e [Hansen e Mladenović 2001], esse processo de decompor em subproblemas tende a ser mais efetivo (economizando tempo computacional) e para problemas muito grandes esse método tende a produzir melhores resultados, em relação ao VNS.

3.2. Algoritmo VNDS aplicado ao problema de estratificação da PAC

Considere o vetor populacional $X_U = \{x_1, x_2, ..., x_N\}$ onde cada x_i representa o total de pessoal ocupado na empresa i, ou seja, X é a variável de estratificação. Define-se o conjunto Q, a partir da retirada das duplicações de X_U, como os pontos de corte distintos possíveis. Assim, assuma que w seja a quantidade de elementos de Q e que cada ponto de corte seja denotado por q_j para $j = 1, \ldots, w$. Portanto, tem-se $Q = \{q_1, q_2, ..., q_w\}$. Conforme já explicitado na Etapa 1 do procedimento de estratificação são necessários $(L-1)$ pontos de corte. Em particular, como no caso da PAC tem-se $L = 4$, o conjunto solução que está se buscando corresponde à melhor escolha possível do vetor $s = \{b_1, b_2, b_3\}$, tal que $s \subseteq Q$, e que o tamanho amostral n seja mínimo. Observe que o número total de soluções possíveis é dado por $\binom{w}{L-1}$. Por exemplo, para uma população em que $w = 6$ e para $L - 1 = 3$, têm-se 20 combinações possíveis de soluções para serem testadas e, nesse caso, a solução ótima pode ser obtida a partir da enumeração explícita. Entretanto, para valores grandes de w o problema é exponencial.

Para o caso da PAC, o algoritmo VNDS inicia-se com um vetor s_0, com valores aleatórios para $b_1, b_2, b_3 \in Q$ e a estrutura de vizinhança é dada pelos valores q_j ao redor de b_h, tal que $q_{j-r} < b_h = q_j < q_{j+r}$, sendo r a amplitude do intervalo a ser definido posteriormente e j é a posição de b_h no conjunto Q, tal que $h = 1, 2, 3$ e $j = 1, 2, \ldots, w$.

No VNDS a decomposição consiste em fixar todos os atributos, exceto por k atributos. No método aqui proposto, k corresponde à quantidade de elementos do vetor s que não serão fixos. Esses elementos livres serão modificados nos passos da perturbação e da busca local.

Para ilustrar o método, suponha que a variável de estratificação 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 $Q = \{1, 2, 3, 4, 5, 7, 8, 10, 15, 31\}$. Conforme o Algoritmo 1, gera-se um vetor aleatório inicial $s = s_0 = \{3, 7, 10\}$, com isso, baseando-se nas Equações (4), (5), (6), os quatro estratos formados são: $E_1 = \{1, 1, 1, 2, 2, 3, 3\}$; $E_2 = \{4, 4, 5, 7, 7\}$; $E_3 = \{8, 8, 10, 10\}$; $E_4 = \{15, 31\}$. Logo, $N_1 = 7, N_2 = 5, N_3 = 4$ e $N_4 = 2$, portanto $N_A = 16$. Então, consequentemente, $W_1 = 0,44; W_2 = 0,31; W_3 = 0,25$ e por fim, $S_1^2 = 0,8; S_2^2 = 2,3; S_3^2 = 1,3$. Com esses valores pode-se calcular o tamanho amostral, conforme Equação (10), ou seja, o valor da função objetivo aplicado em s que é dado por $f(s)$.

A perturbação gera um novo vetor solução s_1 que difere de s por k elementos. Para $k = 1$, apenas um elemento deve ser modificado. Então, sorteia-se qual dos três elementos será o escolhido, por exemplo para b_1, o novo vetor de s_1 será $\{b_1', 7, 10\}$. Em que b_1' segue a regra de vizinhança descrita acima, $q_{3-r_1} < (b_1 = q_3) < q_{3+r_1}$, sendo r_1 o parâmetro da amplitude da perturbação. Para $r_1 = 2$, o novo valor de b_1' será um elemento qualquer do conjunto $\{1, 2, 4, 5\}$. Surgindo assim, por exemplo, um novo vetor solução $s_1 = \{4, 7, 10\}$.

Com esse vetor s_1, faz-se uma busca local em t vizinhos na vizinhança de apenas k elementos, assim como na perturbação, mas utilizando r_2 como parâmetro para a amplitude da busca local, para assim chegar a um novo vetor s_2. Por exemplo, se $t = 2$ e $r_2 = 2$, irá se testar dois vizinhos para b_1'', lembrando que agora, b_1'' deverá pertencer ao conjunto $\{2, 3, 5, 7\}$. Gerando, assim, duas soluções s_2, tais como, $s_2 = \{2, 7, 10\}$ e $s_2 = \{5, 7, 10\}$.

Se o tamanho amostral calculado, utilizando-se alguma das duas soluções, para os pontos de corte de s_2 for menor que utilizando-se os pontos de corte de s, ou seja, $f(s_2) < f(s)$, atualiza-se o vetor s e faz-se $k = 1$ novamente. Caso contrário faz-se $k = k + 1$, até atingir k_{max}. O algoritmo termina quando pelo menos um dos três critérios de parada for satisfeito: número máximo de iterações, número máximo de iterações sem melhoria da função objetivo e tempo máximo de processamento.

4. Resultados e Discussão

A PAC abrange todo o território nacional e representa quase 2 milhões de estabelecimentos comerciais. Por ser tão abrangente, a aplicação de um censo é inviável e, por isso, utiliza-se amostragem para produzir estimativas confiáveis desse ramo da economia. Porém, optou-se aqui por limitar a análise apenas ao Estado de São Paulo, por ser o mais representativo economicamente e por ter uma população de empresas suficientemente grande, com mais de 600.000 empresas no ano de 2014. Nessa Unidade da Federação, a população de empresas é subdivida em 75 estratos naturais de acordo com a CNAE[6].

Todos os 75 estratos naturais foram considerados nos experimentos computacionais, mas por motivos de espaço e uma melhor apresentação, serão reportados neste trabalho apenas os resultados para 35 estratos naturais, sendo cada um deles tratado como uma população independente. Totalizando, assim, 35 populações para aplicação dos métodos de minimização do tamanho amostral total. As características básicas dessas populações estão sumarizadas na Tabela 1, em que a primeira coluna corresponde ao código de identificação da população, a segunda coluna é o tamanho populacional (N), a terceira coluna é a quantidade de valores distintos (w) para a variável

[6]Para mais detalhes, ver metodologia do plano amostral em [IBGE 2015].

auxiliar e a quarta coluna é o coeficiente de assimetria (ASSI). Assim, têm-se os mais variados tipos de populações, desde muito pequenas com apenas $N = 54$ empresas, a muito grandes com 63.565 unidades, também com valores para w variando de 6 a 171 e todas populações com assimetria positiva variando de $1,2$ até $123,3$.

Para o algoritmo proposto aqui, implementado em linguagem R[7], foram utilizados os seguintes critérios de parada: alcançar 100 iterações ou ficar 15 iterações sem melhoria, ou ainda, o algoritmo atingir 6 horas de processamento. Além disso, após testes preliminares, definiu-se $t = 5$ vizinhos, $k_{max} = 3$, e as amplitudes da vizinhança em $r_1 = 30$ e $r_2 = 15$, para a perturbação e para a busca local, respectivamente. O método de alocação proposto por [Brito et al. 2015] está implementado em linguagem R e disponível no pacote *MultAlloc*[8].

De forma a avaliar o algoritmo proposto nesse trabalho também foram utilizados os dois métodos mais usuais da literatura, descritos em [Kozak 2004] e [Lavallée e Hidiroglou 1988]. Contudo, eles apresentam uma limitação, pois não estão programados para receber restrições adicionais, como a restrição de tamanho amostral mínimo por estrato e, portanto, algumas adaptações foram necessárias. O método de [Kozak 2004] está implementado em código fechado em linguagem R dentro do pacote *stratification*[9]. Por isso, optou-se por ignorar essa restrição inicialmente, para assim, utilizar o algoritmo apenas para gerar o vetor s, e a partir desses resultados, calcular novos tamanhos amostrais que atendam a essa restrição, o que pode ocasionar um CV maior que 10%. Já o método de [Lavallée e Hidiroglou 1988] não precisou de ajuste, pois foi implementado em código aberto em linguagem SAS em [Azevedo 2004], o que permitiu a inclusão dessa restrição.

Os três algoritmos supracitados foram executados em um computador dotado de 4GB de memória RAM, 4 processadores i5 de 3.2GHz em sistema operacional Windows de 32 bits. Os resultados obtidos para os tamanhos amostrais totais n, sendo que $n = n_1 + n_2 + n_3 + n_C$, estão apresentados na Tabela 1 da quinta até a sétima coluna, de forma que as abreviações LH, Ko e VNDS referem-se, respectivamente, aos algoritmos de [Lavallée e Hidiroglou 1988], [Kozak 2004] e o proposto.

Na Tabela 1, as 35 populações selecionadas estão ordenadas em relação à quantidade de valores distintos de w. Para as três primeiras populações que têm baixos valores para w, foi possível testar todas as combinações possíveis para encontrar o ótimo global, a partir da combinação explicitada na seção 3.2 e da utilização da alocação ótima de [Brito et al. 2015]. Caso a solução encontrada pelos algoritmos seja igual ao ótimo global, elas são marcadas com um asterisco. Portanto, na primeira população, os três algoritmos convergiram para o ótimo global. Na segunda população, nenhum dos algoritmos produziu o ótimo global ($n^* = 17$), mas o que mais se aproximou foi o VNDS. E por fim, na terceira população, somente o algoritmo VNDS produziu o ótimo global.

Como citado anteriormente, havia a possibilidade do algoritmo de [Kozak 2004] adaptado, gerar soluções inválidas que ultrapassariam o limite de CV de 10%. Esse fato ocorreu em cinco populações, onde os resultados do tamanho amostral desse método estão sublinhados na Tabela 1. Assim, essas soluções não podem ser consideradas para análise.

Para uma fácil identificação da melhor solução encontrada 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 última população da tabela, a solução vencedora veio do algoritmo de [Lavallée e Hidiroglou 1988]. Embora a solução encontrada pelo algoritmo de [Kozak 2004] seja menor, essa solução foi descartada, pois não atende à restrição de CV $\leq 10\%$.

A partir das soluções vencedoras, grifadas em negrito na Tabela 1, observa-se que o algoritmo VNDS teve um desempenho superior aos demais, no que concerne à qualidade das soluções. Destacando que das 35 populações consideradas no estudo, em 29 delas o melhor re-

[7] http://www.r-project.org/
[8] https://cran.r-project.org/web/packages/MultAlloc/index.html
[9] https://cran.r-project.org/web/packages/stratification/

Tabela 1: Informações sobre as Populações e Resultados dos Algoritmos

Código da População	N	w	ASSI	Tamanho amostral (n) LH	Ko	$VNDS$
354542	112	6	1,2	**17***	**17***	**17***
354512	540	14	8,0	19	19	**18**
354682	54	21	5,1	16	13	**11***
354636	149	24	4,0	23	18	**17**
354785	1439	25	6,5	19	**18**	**18**
354773	2327	41	11,2	30	23	**22**
354784	5686	48	70,0	23	**17**	**17**
354631	463	51	6,8	22	22	**20**
354743	4167	56	19,5	**21**	**21**	22
354683	421	57	9,1	24	20	**19**
354751	15059	66	21	35	**26**	27
354652	609	68	9,9	28	21	**20**
354759	9214	68	22,8	28	**27**	29
354753	7405	69	65,1	22	22	**20**
354634	912	73	4,5	24	**23**	**23**
354755	11033	74	46,6	31	30	**29**
354541	4598	77	10,4	31	**30**	**30**
35462	1941	79	12,3	**28**	29	**28**
354684	708	80	9,2	26	**22**	**22**
354693	1509	81	24,5	24	19	**18**
354672	1232	82	12,5	25	**24**	**24**
354752	6454	82	28,3	**31**	32	**31**
354651	1096	84	31,0	21	21	**17**
354642	2552	86	36,0	24	21	**23**
354761	15446	87	97,5	29	25	**28**
354679	1787	90	10,2	26	**25**	**25**
354712	32089	96	10,5	**31**	27	**31**
354663	3564	103	19,1	32	33	**31**
354644	682	104	15,3	22	22	**19**
354754	14564	114	60,2	37	**35**	**35**
35466	2706	118	43,5	29	19	**28**
354649	5565	143	17,9	41	**39**	**39**
354744	47923	147	116	**35**	**35**	39
35472	63565	149	13,0	36	**33**	35
354781	61368	171	123,3	**40**	37	48

* Ótimo Global

sultado foi produzido pelo algoritmo VNDS, o que representa 83% das populações. Enquanto o algoritmo de [Kozak 2004] produziu a solução vencedora em 15 populações e o algoritmo de [Lavallée e Hidiroglou 1988] apenas em 7 populações.

Assim, o algoritmo proposto aqui, só não produziu a solução vencedora em 6 populações. Mesmo assim, em 4 dessas 6 populações o resultado encontrado foi muito próximo da solução vencedora. Acredita-se que o algoritmo apresenta um desempenho inferior, quando o tamanho populacional (N) é muito grande, isso se deve ao custo de processamento, pois nesses casos, o critério de parada é o tempo máximo de processamento, o que tende a abreviar a quantidade de iterações.

De forma a avaliar a qualidade dos resultados produzidos pelo algoritmo VNDS, para cada uma das três populações da Tabela 2, foi feita uma simulação com 100 execuções do algoritmo, e armazenado o tamanho amostral (n) resultante de cada execução. Assim, cada população

apresentou 100 resultados e as medidas provenientes dessa simulação estão apresentadas na tabela, na seguinte ordem: mínimo, 1º quartil, mediana (ou 2º quartil), 3º quartil, percentil 90 e máximo. Por exemplo, para a população 354636, se o algoritmo VNDS fosse executado 100 vezes, em pelo menos 90 dessas execuções, o resultado produzido seria $n = 17$. Mostrando, portanto, que houve pouca variabilidade nos resultados produzidos, o que indica a robustez do método.

Tabela 2: Medidas de posição para uma simulação de 100 réplicas do algoritmo VNDS

População	Mínimo	Q_1	Mediana (Q_2)	Q_3	P_{90}	Máximo
354631	19	20	20	20	20	23
354636	17	17	17	17	17	18
354652	20	20	20	20	21	22

Ressalte-se que esse algoritmo considera um processo de computação mais intensivo, o que permite explorar uma quantidade bem maior de soluções. 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 de [Lavallée e Hidiroglou 1988],[Kozak 2004] e da ordem de minutos ou até horas (com um máximo de seis horas) para o algoritmo proposto. Entretanto, a qualidade das soluções obtidas aqui foi significativamente melhor, o que pode ocasionar uma redução nos custos, devido à necessidade de uma amostra menor. A versão desse algoritmo ainda não é a definitiva. Assim, outros testes podem ainda ser necessários, tanto para melhorar performance, quanto para a melhor determinação dos parâmetros utilizados. No entanto, acredita-se que os resultados aqui apresentados mostram que essa nova abordagem heurística parece ser bastante promissora.

5. Conclusão e Trabalhos Futuros

O algoritmo proposto apresenta-se como uma alternativa aos métodos propostos na literatura para a resolução do problema de estratificação considerando o objetivo de minimizar o tamanho amostral. Para populações pequenas ($N < 1.000$) e médias ($1.000 < N < 10.000$), os resultados produzidos foram de boa qualidade e o tempo de processamento foi baixo. Entretanto, para populações grandes ($N > 10.000$) o algoritmo proposto apresentou soluções de boa qualidade, mas com um tempo computacional alto. Assim, o uso deste método mostrou-se mais adequado para problemas com pequenas e médias populações, pois a qualidade é muito superior aos demais.

O problema de estratificação é estudado desde [Dalenius 1951] 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, considerando a utilização da metaheurística VNDS [Hansen et al. 2001] combinada com a alocação ótima de [Brito et al. 2015]. Assim, foi possível criar um algoritmo como mais uma alternativa promissora para resolução do problema univariado.

Como trabalhos futuros, pretende-se generalizar o método, para assim ser possível resolver o problema de estratificação considerando: qualquer número para L; a possibilidade de incluir os custos por estratos associados à alocação da amostra; a possibilidade de haver ou não estrato certo. E, além disso, melhorar a eficiência do algoritmo, para que possa resolver o problema em um tempo menor.

Referências

Azevedo, R. V. (2004). Estudo comparativo de métodos de estratificação Ótima de populações assimétricas. Master's thesis, Escola Nacional de Ciências Estatísticas, Rio de Janeiro.

Bankier, M. D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42:174–177.

Bolfarine, H. e Bussab, W. O. (2005). *Elementos de amostragem*. Edgard Blücher, 1 edition.

Brito, J. A., Maculan, N., Brito, L. R. e Montenegro, F. M. T. (2011). Um algoritmo grasp aplicado ao problema de estratificação. In *XLIII Simpósio Brasileiro de Pesquisa Operacional (SBPO)*.

Brito, J. A. e Montenegro, F. M. T. (2007). Algoritmos heurísticos aplicados ao problema de estratificação ótima em populações assimétricas. In *XXXIX Simpósio Brasileiro de Pesquisa Operacional (SBPO)*.

Brito, J. A., Silva, P. L. d. N., Semaan, G. S. e Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41(2):427–442.

Cochran, W. G. (1977). *Sampling Techniques, 3rd Edition*. John Wiley.

Dalenius, T. (1951). The problem of optimum stratification. *Scandinavian Actuarial Journal*, p. 133–148.

Dalenius, T. e Hodges, J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 285(54):88–101.

Ekman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30(1):219–229.

Gunning, P. e Horgan, J. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Statistics Canada*, 2(30):159–166.

Hansen, P. e Mladenović, N. (2001). Variable neighborhood search: Principles and applications. *European Journal of Operational Research*, 130(3):449 – 467.

Hansen, P., Perez-Brito, D. e Mladenović, N. (2001). Variable neighborhood decomposition search. *Journal of Heuristics*, 7(4):335–350.

Hedlin, D. (2000). A procedure for stratification by an extended ekman rule. *Journal of Official Statistics*, 1(16):15–29.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of a large units in survey design. *The American Statistician*, 1(40):27–31.

IBGE, F. I. B. d. G. e. E. (2015). *Pesquisa Anual de Comércio 2013*, volume 25.

Keskintürk, T. e Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics and Data Analysis*, 52:53–67.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5):797–806.

Lavallée, P. e Hidiroglou, M. A. (1988). On the stratification of skewed populations. *Survey Methodology*, 14:33–43.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2 edition.

Mladenović, N. e Hansen, P. (1997). Variable neighborhood search. *Computer Ops Res*, 24(11):1097–1100.