CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

MÉTODO DE BUSCA DECOMPOSTA EM VIZINHANÇA VARIÁVEL COM RECONEXÃO POR CAMINHOS PARA O PROBLEMA DE ESTRATIFICAÇÃO ÓTIMA UNIVARIADO

Breno Tiago Novello Trotta de Oliveira

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do Título de Mestre em Engenharia de Produção e Sistemas.

Orientador

Leonardo Silva de Lima

Coorientador

José André de Moura Brito (ENCE/IBGE)

Rio de Janeiro

Março 2017

MÉTODO DE BUSCA DECOMPOSTA EM VIZINHANÇA VARIÁVEL COM RECONEXÃO POR CAMINHOS PARA O PROBLEMA DE ESTRATIFICAÇÃO ÓTIMA UNIVARIADO

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do Título de Mestre em Engenharia de Produção e Sistemas.

Breno Tiago Novello Trotta de Oliveira

Banca Examinadora:

Presidente, Prof. Dr. Leonardo Silva de Lima (CEFET/RJ) (Orientador)

Prof. Dr. José André de Moura Brito (ENCE/IBGE) (Coorientador)

Prof. Dr. Diego Moreira de Araujo Carvalho (CEFET/RJ)

Prof. Dr. Pedro Luis do Nascimento Silva (ENCE/IBGE)

Prof. Dr. Pedro Henrique González Silva (CEFET/RJ)

Rio de Janeiro

Março 2017

Ficha catalográfica elaborada pela Biblioteca Central do CEFET-RJ

O48 Oliveira, Breno Tiago Novello Trotta de
Método de busca decomposta em vizinhança variável com reconexão por caminhos para o problema de estratificação ótima univariado / Breno Tiago Novello Trotta de Oliveira.—2017.
85f. + apêndices : il. (algumas color.) , grafs. , tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2017.
Bibliografia : f. 82-85
Orientador : Leonardo Silva de Lima
Coorientador : José André de Moura Brito

1. Otimização combinatória. 2. Amostragem (Estatística). 3. Heurística. 4. Estratificação social. 5. Probabilidades. I. Lima, Leonardo Silva de (Orient.). II. Brito, José André de Moura (Coorient.). III. Título.

CDD 519.64

DEDICATÓRIA

Dedico essa dissertação
à minha esposa Ingrid
e à minha mãe Eloisa.

AGRADECIMENTOS

Ao Professores Leonardo Silva de Lima (D.Sc.) e José André de Moura Brito (D.Sc.), por terem aceitado me orientar nesse tema e por terem dedicado seus tempos para o desenvolvimento do trabalho.

À minha mãe, Eloisa, que me incentivou a dar continuidade nos estudos, apesar de toda dificuldade enfrentada.

À minha esposa, Ingrid, que demonstrou compreensão nos momentos que estive ausente e por seu companheirismo durante todo o curso.

Aos colegas do IBGE, em especial à minha chefe Deolinda, por incentivar e aceitar a flexibilidade dos horários do trabalho.

Aos colegas de turma conhecidos no CEFET-RJ.

EPÍGRAFE

”É melhor ter uma resposta aproximada para uma pergunta certa
do que uma resposta exata para uma pergunta errada.”
John Tukey

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, considerando um dos seguintes objetivos: *(i)* minimizar a variância de um estimador de total; *(ii)* minimizar o tamanho amostral. Em função de sua relevância prática e alta complexidade computacional, diversos métodos heurísticos têm sido propostos na literatura. Não obstante, boa parte desses métodos produz soluções de qualidade apenas razoável e com algumas limitações de aplicabilidade. Neste trabalho, foi proposto um algoritmo para resolução do problema de estratificação ótima univariado com o objetivo *(ii)* de minimizar o tamanho da amostra dado um nível de precisão fixado. Para problemas de menor complexidade o método de enumeração exaustiva é aplicado, enquanto em problemas de maior complexidade o algoritmo é baseado nas metaheurísticas *Variable Neighborhood Decomposition Search* com *Path Relinking*. O método proposto foi aplicado em cem populações (sendo 25 já utilizadas na literatura e 75 da Pesquisa Anual do Comércio), com quatro diferentes números de estrato e comparado com dois métodos bem conhecidos da literatura. O método foi capaz de produzir a melhor solução para 97% dos casos considerados, e além disso, em 38% ainda foi capaz de garantir que a solução corresponde a um mínimo global. Destarte, apresenta-se como uma alternativa promissora aos métodos existentes na literatura, embora leve a um maior tempo computacional.

Palavras-chave: Metaheurísticas. Otimização. Estratificação. Amostragem.

ABSTRACT

Optimum stratification problem is associated to the probability sampling area. This problem aims to delimit the population strata and define a sample allocation considering one of the following objectives: (i) minimize a variance of a total estimator; (ii) 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. Many of these methods produce only reasonable quality solutions and have some practical limitations. In this work, we propose an algorithm to solve the univariate optimal stratification problem in order to *(ii)* minimize the sample size given a fixed precision level. We used an exhaustive enumeration method to small problems, while in problems with greater complexity the algorithm is based on Variable Neighborhood Decomposition Search with Path Relinking metaheuristic. The proposed method was applied to a hundred populations (25 from the literature and 75 from Annual Trade Survey), with four different stratum numbers and our results were compared to two very well-known methods. Our method achieved the best results for 97% of the considered cases. In addition, it was still able to guarantee that the solution corresponds to a global minimum in 38% of the cases. Therefore, it turns out that our algorithm is a promising alternative to the existing methods in the literature although it leads to a longer computational time.

Keywords: Metaheuristics. Optimization. Stratification. Sampling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAS Amostragem Aleatória Simples

AE Amostragem Estratificada

AEC Amostragem Estratificada por Corte

CAGED Cadastro Geral de Empregados e Desempregados

CBS Cadastro Básico de Seleção

CEMPRE Cadastro Central de Empresas

CNAE Classificação Nacional de Atividades Econômicas

EVP Exato, VNDS, PR

IBGE Instituto Brasileiro de Geografia e Estatística

PAC Pesquisa Anual do Comércio

PR Reconexão por Caminhos (*Path-Relinking*)

RAIS Relação Anual de Informações Sociais

UF Unidade da Federação

VNDS Busca Decomposta em Vizinhança Variável (*Variable Neighborhood Decomposition Search*)

VNS Busca em Vizinhança Variável (*Variable Neighborhood Search*)

SUMÁRIO

1 - Introdução

O problema de estratificação está associado à área de amostragem probabilística. Por sua vez, a amostragem corresponde ao processo de selecionar um subconjunto de unidades de uma população, de modo que se possa inferir as características de interesse da população com certo grau de precisão, conforme Lohr (2010). Mais especificamente, na amostragem probabilística 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, amostragem estratificada, amostragem sistemática e amostragem por conglomerados.

A estratificação consiste em particionar a população em subpopulações (denominadas estratos), de tal forma que não há sobreposição e juntas, essas subpopulações, abrangem a totalidade da população. Depois de determinados os estratos, seleciona-se uma amostra em cada um deles independentemente. Algumas motivações para o uso da amostragem estratificada são: melhoria da precisão das estimativas; representar os diferentes grupos dentro de uma população; questões administrativas.

O problema de estratificação vem sendo estudado desde Dalenius (1950), sendo propostas novas abordagens até os dias atuais. Esse problema 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 métodos propostos na literatura foram desenvolvidos para atender ao primeiro objetivo, como Ekman (1959); Dalenius e Hodges (1959); Hedlin (2000); Gunning e Horgan (2004); Keskintürk e Er (2007); Brito et al. (2011) e muitos outros estudos, enquanto o segundo objetivo foi abordado em Hidiroglou (1986); Lavallée e Hidiroglou (1988); Kozak (2004); Brito e Montenegro (2007).

1.1 - Objetivo

O presente trabalho de dissertação traz a proposta de um novo método em alternativa aos métodos da literatura, com o intuito de atender ao objetivo *(ii)* do problema de estratificação univariado, para produzir novos pontos de corte para qualquer população, e ainda, capaz de atender a uma restrição adicional: um tamanho de amostra mínimo por estrato. Os pontos de corte (ou limites dos estratos) é que definem a solução do problema, pois a partir deles que se particiona (estratifica) a população e assim, pode-se fazer a alocação da amostra, para então, se calcular o tamanho amostral.

1.2 - Justificativa

Os métodos 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 do IBGE como, por exemplo, na Pesquisa Anual do Comércio (PAC), na Pesquisa Anual de Serviços (PAS), na Pesquisa Anual da Indústria da Construção (PAIC) etc.

Tal fato ficou evidenciado quando não foi possível aplicar diretamente os métodos existentes à estratificação da PAC, com o intuito de reduzir o tamanho amostral da pesquisa. Como essa pesquisa abrange todo o território nacional, a população de empresas é composta por pouco mais de 2 milhões de estabelecimentos comerciais. Devido a custos operacionais e outros fatores, o IBGE realiza a pesquisa através de um plano amostral que utiliza amostragem estratificada por corte, segundo o porte da empresa. Atualmente, a delimitação (ou corte) desses estratos obedece a um critério gerencial, que estabelece que todas as empresas com 20 ou mais pessoas ocupadas (empregados) são classificadas como pertencentes ao estrato certo e todas as unidades desse estrato são incluídas na amostra. As demais empresas são classificadas no estrato amostrado, que ainda se subdivide em três estratos, também de acordo com o porte da empresa, a saber: A_1, empresas de 0 a 4 pessoas ocupadas; A_2, empresas de 5 a 9 pessoas ocupadas; A_3, empresas de 10 a 19 pessoas ocupadas. Nesses estratos utiliza-se amostragem aleatória simples sem reposição das unidades elementares. Portanto, nos dias atuais, os pontos de corte utilizados na PAC são: 4, 9 e 19.

Mesmo considerando o uso de amostragem, em todo o país foram aplicados mais de 77 mil questionários em 2014. Em relação à amostra utilizada em 2007, houve um aumento

de 38% na quantidade de empresas para serem entrevistadas. Esse aumento implica maiores custos operacionais e uma maior carga de trabalho em todas as etapas da pesquisa. Por isso, pretende-se aplicar o método proposto, para definir novos pontos de corte, com o intuito de reduzir a amostra total dessa pesquisa.

1.3 - Aspectos Metodológicos

O método proposto determina os pontos de corte através de um algoritmo, que tem um procedimento de resolução exata e dois procedimentos de resolução baseados nas metaheurísticas Busca Decomposta em Vizinhança Variável (VNDS, tradução livre de *Variable Neighborhood Decomposition Search*) proposta por Hansen, Mladenović e Perez-Brito (2001) e Reconexão por Caminhos (PR, tradução livre de *Path-Relinking*) proposta originalmente por Fred Glover no livro de Barr, Helgason e Kennington (1996). Na etapa seguinte, a alocação ótima proposta por Brito et al. (2015) é utilizada.

Inicialmente desenvolvido para solucionar o problema de estratificação associado à PAC, o método proposto pode ser aplicado para qualquer tipo de população, inclusive para populações com valores negativos (outra limitação dos métodos existentes), podendo ainda, haver ou não a restrição de tamanho amostral mínimo por estrato. Para validar essa afirmação, foi feito um experimento empírico, em que foram utilizadas 100 populações com as mais variadas características – sendo 25 populações selecionadas da literatura (sem a presença da restrição mencionada) e 75 populações reais utilizadas na PAC (com a presença da restrição mencionada) – incluindo populações muito grandes, algo ainda não testado na literatura.

1.4 - Estrutura

Neste trabalho pretende-se solucionar um problema de amostragem probabilística, em que o método proposto consiste em um algoritmo baseado em metaheurísticas. Por isso, antes de apresentar a metodologia implementada, necessita-se apresentar os conceitos e o referencial teórico dessas duas áreas de estudo. Assim, essa dissertação está estruturada em mais cinco capítulos, além desse capítulo introdutório.

O capítulo dois traz os conceitos básicos sobre a amostragem e os principais conceitos sobre a amostragem estratificada, uma descrição detalhada sobre o problema de estratificação ótima e uma revisão bibliográfica dos métodos que serão utilizados para compa-

ração.

O terceiro capítulo traz uma descrição dos conceitos de metaheurísticas, e em particular, sobre os métodos Busca em Vizinhança Variável (VNS, tradução livre de *Variable Neighborhood Search*), VNDS e PR.

No capítulo quatro está a metodologia, em que se apresenta a base de dados das populações da literatura. Traz ainda o plano amostral da PAC e a delimitação dessa base de dados. Além disso, há uma descrição detalhada do algoritmo proposto apresentado com exemplos, o que facilita a compreensão do procedimento de estratificação, do procedimento de resolução exata e dos procedimentos de resolução baseada nas metaheurísticas.

O experimento empírico e os resultados computacionais estão no quinto capítulo. Nas cem populações mencionadas, e considerando quatro diferentes números de estratos, foram produzidos resultados para o método proposto e comparado com dois métodos bem conhecidos da literatura – Kozak (2004) e Lavallée e Hidiroglou (1988) – portanto, no total foram produzidos 1200 casos (100 populações x 3 métodos x 4 números de estratos). O método proposto foi capaz de produzir a melhor solução para 97% dos casos considerados, e além disso, em 38% dos casos foi também capaz de garantir que essas soluções correspondem a mínimos globais. Em especial, nas populações da PAC considerando cinco estratos, o método proposto apresentou o menor tamanho de amostra total entre todos os testes realizados, o que representaria uma redução amostral de 92% se comparado a metodologia atual da pesquisa.

Por fim, no capítulo seis estão as conclusões e possíveis tópicos de pesquisa para desenvolvimentos futuros.

2 - Amostragem

O problema de estratificação ótima está associado à área de amostragem probabilística. Por sua vez, a amostragem corresponde ao processo de selecionar um subconjunto de unidades de uma população, de modo que se possa inferir as características de interesse da população com certo grau de precisão, conforme Lohr (2010). Por isso, antes de apresentar o problema da estratificação, há a necessidade de introduzir alguns conceitos básicos de amostragem.

2.1 - Conceitos Básicos

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 (também chamada de censo), mas na prática, essa forma de realização da pesquisa dificilmente ocorre, seja por problemas geográficos, logísticos ou 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, para que se possa inferir sobre características de interesse da população. As principais vantagens ao utilizar amostragem, em vez da enumeração completa da população, segundo Cochran (1977) são: a redução dos custos, a coleta de dados de forma mais rápida e mais abrangente e uma maior acurácia na coleta das informações. Adicionalmente, Kish (1965) ainda considera como vantagem a viabilidade de uma amostra quando os elementos são destruídos[1], pois um censo seria inviável. 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; minimização do erro de medição. Abaixo há a definição de alguns conceitos importantes que serão úteis para um melhor acompanhamento do texto.

[1] Exemplos: teste de resistência de lâmpadas, exame de sangue etc.

Definição 2.1 (População ou Universo)**.** *É o conjunto de todas as unidades para o qual se quer obter informações ou fazer inferências, denotada pela letra* U*.*

Definição 2.2 (Amostra)**.** *É o subconjunto de unidades da população que é selecionada para medir ou observar, denotada pela letra* $\mathbb{S}$*.*

Definição 2.3 (Cadastro)**.** *É a lista de unidades que formam a população de onde a amostra é selecionada.*

Definição 2.4 (Unidade ou Elemento)**.** *É um único indivíduo ou objeto a ser medido ou observado na pesquisa.*

Como os parâmetros populacionais, em geral, são desconhecidos, utiliza-se os estimadores (funções dos dados amostrais) para realizar medidas das variáveis de interesse da população. Para facilitar o entendimento, deve ficar clara a diferença entre parâmetro e estimador, que podem ser definidos da seguinte forma:

Definição 2.5 (Parâmetro)**.** *É uma medida usada para descrever uma característica da população.*

Definição 2.6 (Estimador)**.** *É uma estatística adequada para estimar o valor de um parâmetro a partir dos dados amostrais, ou seja, é uma função dos dados amostrais que serve para estimar um parâmetro populacional.*

Assim, seja uma população finita $U = \{1, 2, \ldots, i, \ldots, N\}$ em que cada valor desse conjunto representa o rótulo da unidade na população de tamanho (N), denotadas por $i = 1, 2, \ldots, N$ ou $i \in U$. Denota-se o vetor $Y_U = \{y_1, y_2, \ldots, y_N\}$ como o vetor populacional, onde cada elemento y_i representa o valor da variável de interesse a ser estudada para $i \in U$. Dentre os parâmetros mais estudados com relação a uma determinada variável y, podem ser destacados os três seguintes: o total populacional

$$Y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i, \tag{2.1}$$

a média populacional

$$\overline{Y} = \frac{1}{N} \sum_{i \in U} y_i \tag{2.2}$$

e a variância populacional

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \overline{Y})^2. \tag{2.3}$$

Como os parâmetros são desconhecidos, em geral deve-se utilizar estimadores para conseguir medir as caraterísticas de interesse através da amostra, a qual conforme Definição 2.2, é um subconjunto da população, tal que $\mathbb{S} \subseteq U$ com cardinalidade n. E assim como na notação populacional, na notação amostral o índice i representa o rótulo da unidade na amostra, denotado por $i \in \mathbb{S}$. Os estimadores relevantes serão apresentados na próxima seção após cada plano amostral.

Com os conceitos básicos anteriores, e considerando que há disponível um cadastro com os requisitos desejáveis[2], na seção seguinte, é apresentada uma descrição breve de alguns tipos de planos amostrais de interesse para o desenvolvimento do presente estudo.

2.2 - Tipos de Planos Amostrais

A maioria das pesquisas realizadas pelo IBGE e pelos órgãos de estatística oficial pelo 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 planos amostrais são: amostragem aleatória simples, amostragem estratificada, amostragem sistemática e amostragem por conglomerados. Entretanto, os dois últimos fogem do escopo desse trabalho e não serão abordados. Para mais detalhes ver Cochran (1977).

O método mais básico de seleção é a **Amostragem Aleatória Simples (AAS)** que consiste em escolher n elementos dentre todas as N unidades da população de maneira aleatória e com igual probabilidade. A seguir, a mesma notação de Lohr (2010) é utilizada como base para a definição dos principais estimadores desse método de seleção. Sob amostragem aleatória simples, denote $\hat{Y}_{AAS}$ como o estimador não viciado do total populacional, dado por

$$\hat{Y}_{AAS} = \frac{N}{n} \sum_{i \in \mathbb{S}} y_i = N\overline{y} \tag{2.4}$$

e $V(\hat{Y}_{AAS})$ como a variância do estimador de total, dada por

$$V(\hat{Y}_{AAS}) = N^2(1 - \frac{n}{N})\frac{S_y^2}{n}, \tag{2.5}$$

[2]Conter informação suficiente que permita identificar e localizar cada unidade da população; não haver redundâncias; ter precisão e estar atualizado.

em que S_y^2 é dado pela equação 2.3.

Outro esquema de amostragem muito comum também é a **Amostragem Estratificada (AE)**. Conforme Cochran (1977) e Lohr (2010), este plano 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 e denotadas por $E_1, E_2, \ldots, E_L$) não se sobrepõem e, juntas, abrangem a totalidade da população, ou seja:

$$U = E_1 \cup E_2 \cup \ldots \cup E_L$$

$$E_j \cap E_k = \varnothing, \quad \forall j \neq k$$

$$N = \sum_{h=1}^{L} N_h$$

Uma vez determinados os estratos populacionais, seleciona-se uma amostra aleatória simples para cada estrato h, denotada por $\mathbb{S}_h$ para $h = 1, \ldots, L$, sendo as seleções feitas independentemente nos diferentes estratos, de modo que a amostra total é dada pela união das amostras de cada estrato, cuja notação é: $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \ldots \cup \mathbb{S}_L$. Cada amostra $\mathbb{S}_h$ tem um tamanho associado, denotado por n_h, de modo que o tamanho de amostra total é dado pela soma dos tamanhos amostrais de cada estrato, cuja notação é: $n = \sum_{h=1}^{L} n_h$.

Algumas motivações para o uso da estratificação são: melhoria da precisão das estimativas; representar os diferentes grupos dentro de uma população; garantir o espalhamento da amostra; questões administrativas. A forma de definir o tipo de estratificação a ser utilizada depende do objetivo da pesquisa, ou seja, se será utilizada a estratificação natural ou a estratificação estatística, conforme Freitas (2002). O primeiro caso ocorre quando há interesse em estimar parâmetros em cada um dos estratos formados, em geral, são consideradas divisões administrativas, tais como localização, lote de um produto, atividade econômica etc. O segundo caso ocorre quando se busca a formação de grupos homogêneos, segundo algum conjunto de características de interesse, visando aumentar a eficiência na estimação.

Para uma variável de interesse y, o estimador do total populacional sob amostragem estratificada é denotado por $\hat{Y}_{AE}$. Segundo Lohr (2010), este estimador é definido por

$$\hat{Y}_{AE} = \sum_{h=1}^{L} N_h \overline{y}_h, \tag{2.6}$$

sendo que $\overline{y}_h = \frac{1}{n_h}\sum_{i\in\mathbb{S}_h} y_i$ é a média amostral no h-ésimo estrato A variância do estimador de total $(\hat{Y}_{AE})$ é dada por

$$V(\hat{Y}_{AE}) = \sum_{h=1}^{L} N_h^2(1-\frac{n_h}{N_h})\frac{S_{hy}^2}{n_h}, \tag{2.7}$$

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

Outra etapa da AE é o procedimento que, segundo Bolfarine e Bussab (2005), 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

$$n_h = n \cdot a_h. \tag{2.8}$$

A escolha do procedimento de alocação é muito importante porque dele depende a precisão dos estimadores, conforme Bolfarine e Bussab (2005). Considerando essa questão, são utilizadas na literatura de amostragem as equações 2.9 a 2.13 para a_h, de acordo com o método de alocação escolhido:

$$a_h = \frac{1}{L} \quad (h = 1,\ldots,L) \tag{2.9}$$

$$a_h = \frac{N_h}{N} \quad (h = 1,\ldots,L) \tag{2.10}$$

$$a_h = \frac{N_h S_h/\sqrt{c_h}}{\sum_{j=1}^{L} N_j S_j/\sqrt{c_j}} \quad (h = 1,\ldots,L) \tag{2.11}$$

$$a_h = \frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j} \quad (h = 1,\ldots,L) \tag{2.12}$$

$$a_h = \frac{X_h^p}{\sum_{j=1}^{L} X_j^p} \quad (h = 1,\ldots,L) \tag{2.13}$$

A equação 2.9 refere-se à Alocação Uniforme, quando todas as amostras dos estratos (n_h) deverão ter o mesmo tamanho. A equação 2.10 refere-se à Alocação Proporcional,

em que o tamanho amostral é distribuído proporcionalmente ao tamanho populacional dos estratos (N_h). A equação 2.11 refere-se à Alocação Ótima, que é utilizada quando há grande variabilidade dos estratos e quando o custo para se obter os dados de cada estrato h, denotado por c_h, pode variar. A equação 2.12 refere-se à Alocação Ótima de Neyman, que é um caso particular da Alocação Ótima, que ocorre quando todos os custos por estrato são iguais, tal que $c_h = c, \forall h$. Essas equações, de 2.9 a 2.12, podem ser encontradas em Cochran (1977); Bolfarine e Bussab (2005).

Por fim, a equação 2.13 refere-se à Alocação Potência, proposta por Bankier (1988), que é um tipo de alocação proporcional, mas a distribuição é feita proporcionalmente a uma variável auxiliar x, que é correlacionada à variável de interesse y. Sendo $X_h^p = (\sum_{i \in E_h} x_i)^p$ o total populacional do h-ésimo estrato elevado a p, um parâmetro que pode assumir qualquer valor no intervalo $0 < p < \infty$. Entretanto, na prática, define-se o valor de p como a raiz quadrada ($p = 1/2$) ou como a raiz cúbica ($p = 1/3$) dessa variável auxiliar, conforme Lavallée e Hidiroglou (1988).

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

$$n = \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{V(\hat{Y}_{AE}) + \sum_{h=1}^{L} N_h S_h^2}. \tag{2.14}$$

Entre os tipos de alocação, a mais utilizada é a de Neyman (equação 2.12), 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, $n_h > N_h$.

Considerando o problema de alocação, Brito et al. (2015) desenvolveram uma formulação de programação inteira que garante o ótimo global para a alocação. Com o auxílio de variáveis binárias, foi possível formular o problema de alocação como um problema de programação inteira binária, de tal forma que a alocação sempre produzirá tamanhos amostrais com valores inteiros e menores que os tamanhos populacionais do estrato. Essa formulação é aplicável tanto ao caso univariado, como ao multivariado (quando há mais de uma variável

de estratificação).

Um outro plano amostral que pode ser utilizado é a **Amostragem Estratificada por Corte (AEC)**, que é uma variação da amostragem estratificada. Entretanto, difere da AE, apenas por um ponto: no último estrato todas as unidades da população compõem a amostra, tal que $n_L = N_L$. Nesse plano amostral, usualmente, a variável utilizada para a estratificação é a variável auxiliar x, em vez da variável de interesse y. Essa variável auxiliar x é, usualmente, tratada como uma medida de tamanho da população. A Amostragem Estratificada por Corte é utilizada quando a população de estudo apresenta uma alta assimetria [3] na variável de interesse e/ou na variável auxiliar. Como no último estrato (L) estão as maiores unidades 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 dificuldade do problema está em definir os pontos de corte, ou limites dos estratos, que serão utilizados para destacar o estrato certo e delimitar os estratos amostrados. Assim, 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.

Para compreender melhor a importância dos pontos de corte, suponha que se queira realizar um estudo com os mercados de uma cidade, de tal forma que os maiores mercados devem obrigatoriamente participar da amostra. Primeiramente, deve-se escolher uma variável auxiliar conhecida que seja útil para medir o tamanho do mercado. Pode-se utilizar o faturamento, a quantidade de funcionários, a quantidade de produtos vendidos etc. Suponha que foi escolhido o faturamento e, em um segundo momento, é preciso fazer um estudo para definir (ou já saber a priori) qual o valor de faturamento será utilizado para separar os grandes mercados dos demais. Suponha que foram definidos os seguintes valores para estratificar: R$ 5 milhões e R$ 20 milhões. Assim, mercados com faturamento menor que R$ 5 milhões serão considerados pequenos, mercados com faturamento entre R$ 5 milhões e R$ 20 milhões serão considerados médios e mercados com faturamento maior que R$ 20 milhões serão considerados grandes. Enquanto esse último grupo é o estrato certo, os dois primeiros correspondem aos estratos amostrados.

[3] A distribuição dos dados é assimétrica quando está mais concentrada de um lado da distriubição do que do outro. Segundo DeGroot e Schervish (2012), seja X uma variável aleatória com média μ e desvio-padrão σ, a assimetria de X é dada por $E[(X-\mu)^3]/\sigma^3$. Valores negativos indicam assimetria negativa (ou assimetria à esquerda) e valores positivos indicam assimetria positiva (ou assimetria à direita). Comumente, valores absolutos maiores que um são considerados altos, Doane e Seward (2011).

O procedimento de estratificação por corte do exemplo acima pode ser formalizado conforme Azevedo (2004), considerando $Y_U = \{y_1, y_2, ..., y_N\}$ o vetor populacional relacionado com a variável de interesse y e $X_U = \{x_1, x_2, ..., x_N\}$ o vetor populacional relacionado com a variável de estratificação x, e supondo, sem perda de generalidade, que $x_1 \leq x_2 \leq ... \leq x_N$. Essas observações são alocadas aos L estratos, segundo os pontos de corte $b_1 < b_2 < \ldots < b_{L-1}$. Os estratos são definidos segundo:

$$E_1 = \{i \in U | \, x_i \leq b_1\}, \tag{2.15}$$

$$E_h = \{i \in U | \, b_{h-1} < x_i \leq b_h\}, \quad h = 2, 3, \ldots, L-1 \tag{2.16}$$

$$E_L = \{i \in U | \, x_i > b_{L-1}\}. \tag{2.17}$$

Assim, para a construção de L estratos são necessários $(L-1)$ pontos de corte, como no exemplo dos supermercados, em que foram construídos três estratos a partir de dois pontos de corte. Portanto, 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 guiar a alocação da amostra, pois, em geral, as informa-ções da variável de interesse y são desconhecidas. Seja o total populacional associado à variável de estratificação x dado por $X = \sum_{i \in U} x_i = \sum_{h=1}^{L} \sum_{i \in E_h} x_i$, então o estimador do total populacional sob AEC é denotado por $\hat{X}_{AEC}$ e definido por

$$\hat{X}_{AEC} = X_L + \sum_{h=1}^{L-1} N_h \overline{x}_h, \tag{2.18}$$

sendo $X_L = \sum_{i \in E_L} x_i$ o total populacional do estrato certo e $\overline{x}_h = \dfrac{1}{n_h} \sum_{i \in \mathbb{S}_h} x_i$ a média amostral de x no h-ésimo estrato. A variância do estimador de total ($\hat{X}_{AEC}$) é dada por

$$V(\hat{X}_{AEC}) = \sum_{h=1}^{L-1} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_{hx}^2}{n_h} \tag{2.19}$$

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

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

$$CV(\hat{X}_{AEC}) = \frac{\sqrt{V(\hat{X}_{AEC})}}{X}. \tag{2.20}$$

2.3 - Definição do Problema de Estratificação

O problema de estratificação é mais usual para populações assimétricas, em que a Amostragem Estratificada por Corte 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:

(i) Determine o objetivo do problema;

(ii) Fixe o número de estratos (L);

(iii) Escolha o método de alocação;

(iv) Escolha o método de seleção;

(v) Escolha a variável de estratificação;

(vi) Calcule os $(L-1)$ pontos de corte $(b_1, ..., b_{L-1})$, para poder dividir a população em L estratos;

(vii) Calcule os tamanhos amostrais $n_1, n_2, ..., n_L$ de acordo com o método de alocação do item (iii);

(viii) Selecione as unidades em cada estrato h de acordo com o método de seleção do item (iv) e de acordo com o tamanho n_h do item (vii).

A etapa (i) consiste em definir o objetivo dentre os dois objetivos possíveis, a saber: *(1º)* minimizar a variância do estimador de total, ou seja, maximizar a precisão, considerando o tamanho de amostra fixo; *(2º)* minimizar o tamanho amostral, ou seja, minimizar o custo, considerando a precisão fixada previamente. A maioria dos métodos propostos na literatura foram concebidos para atender ao primeiro objetivo, como Ekman (1959), Dalenius e Hodges (1959), Glasser (1962), Hedlin (2000), Gunning e Horgan (2004), Keskintürk e Er (2007), Almeida (2007), Brito et al. (2011), Rao, Khan e Reddy (2014), Veiga (2015) e outros estudos, enquanto o segundo objetivo foi explorado na literatura apenas em Hidiroglou (1986),

Lavallée e Hidiroglou (1988), Kozak (2004), Brito e Montenegro (2007). Observe que 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 (ii) consiste em definir a quantidade de estratos que se pretende particionar a população. A etapa (iii) consiste em utilizar algum dos métodos de alocação, sendo o de Neyman o mais comum na literatura. A etapa (iv) consiste em escolher um dos métodos de amostragem probabilística citados na seção 2.2, em que o mais usual é a AAS, e por isso, foi o método de seleção adotado nesta dissertação. [4]

Na etapa (vi) utiliza-se a variável de estratificação escolhida no item (v) para encontrar os pontos de corte que possibilitam a minimização do objetivo. Além disso, essa etapa 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. Enquanto a etapa (vii) corresponde ao segundo nível do problema de estratificação, em que os tamanhos amostrais(n_h) são alocados por estratos de forma a somar o tamanho amostral total (n), considerando o objetivo especificado na etapa (i).

Como o segundo nível do problema já foi resolvido por Brito et al. (2015), o método proposto nesse trabalho utiliza a alocação apresentada nesse artigo. Portanto, o método proposto se concentra apenas no primeiro nível do problema de estratificação com o intuito de solucionar o segundo objetivo (minimizar o tamanho amostral). Assim, o problema de estratificação, aqui tratado, consistirá em determinar os pontos de corte ($b_1 < b_2 < \ldots < b_{L-1}$) visando minimizar o custo total (C_T) da pesquisa, dado por

$$C_T = \sum_{h=1}^{L} c_h n_h. \tag{2.21}$$

Porém, se for assumido que os custos por unidade dos estratos (c_h) são iguais ou desconhecidos para todos os estratos, o objetivo se resume a minimizar o tamanho amostral total $n = \sum_{h=1}^{L} n_h$.

Além disso, é preciso considerar algumas restrições habituais a todo processo de amostragem estratificada, como, por exemplo, a restrição que visa evitar a criação de estratos

[4] Essas três etapas estão presentes em qualquer procedimento de estratificação.

vazios, dada por

$$N_h \geq 1 \quad (h = 1, \ldots, L). \tag{2.22}$$

Deve-se, também, garantir que o tamanho amostral do estrato não será maior que o tamanho da população no estrato, tal que

$$n_h \leq N_h \quad (h = 1, \ldots, L), \tag{2.23}$$

e garantir que os tamanhos de amostra nos estratos são números inteiros positivos, de modo que

$$n_h \in \mathbb{Z}_+ \quad (h = 1, \ldots, L). \tag{2.24}$$

Como citado anteriormente, para minimizar o custo da pesquisa (ou o tamanho amostral), deve-se previamente fixar a precisão. Optou-se por utilizar o coeficiente de variação do estimador de total da variável de estratificação ($CV(\hat{X}_{AEC})$ — Equação 2.20), de modo que, se garanta que seja menor que um valor prefixado τ ($0 < \tau < 1$), isto é,

$$CV \leq \tau. \tag{2.25}$$

Como esse é um problema de estratificação por corte geral, note que não foi feita nenhuma restrição para que o último estrato seja o estrato certo. Vale ainda destacar, que esse é o caso univariado, ou seja, há apenas uma variável de estratificação. Ultimamente tem sido estudado o caso multivariado, quando há pelo menos duas variáveis de estratificação, como em Lednicki e Wieczorkowski (2003); Kozak, Verma e Zieliński (2007); Ballin e Barcaroli (2013); Brito, Semaan e Brito (2014).

2.4 - Métodos da Literatura para Minimização do Tamanho Amostral

Conforme comentado na seção anterior, a maioria dos métodos propostos na literatura foram concebidos para atender ao primeiro objetivo (minimizar a variância do estimador de total). Em relação ao segundo objetivo do problema de estratificação, há menos métodos do que para o primeiro. Portanto, a seguir são explicitados dois desses métodos para minimizar o custo ou o tamanho amostral.

O método desenvolvido por Lavallée e Hidiroglou (1988) é o mais conhecido e utilizado para atender ao segundo objetivo do problema de estratificação. Esse método traba-

lha com o número de estratos e o nível de precisão desejado (medido pelo coeficiente de variação) fixos, de tal modo que sejam produzidos limites de estratificação que minimizem o tamanho amostral total. No que diz respeito à alocação, o método utiliza a alocação potência (Equação 2.13).

A equação do tamanho amostral a ser minimizada é dada por

$$n = N_L + \frac{N \cdot \sum_{h=1}^{L-1} (W_h S_h)^2 (W_h \overline{X}_h)^{-p} \cdot \sum_{h=1}^{L-1} (W_h \overline{X}_h)^p}{CV^2 \cdot N \cdot \overline{X}^2 + \sum_{h=1}^{L-1} W_h S_h^2}, \tag{2.26}$$

sendo $W_h = N_h/N$ a proporção populacional do h-ésimo estrato. Para resolver essa equação, Lavallée e Hidiroglou (1988) derivaram em relação ao tamanho amostral n e igualaram a zero, obtendo a seguinte equação do segundo grau:

$$\alpha_h b_h^2 + \beta_h b_h + \gamma_h = 0 \tag{2.27}$$

sendo,

$$\begin{aligned} \alpha_h &= FT_h - FT_{h+1} \qquad h = 1, ..., L-2 \\ \alpha_h &= FT_h - AB \qquad h = L-1 \end{aligned} \tag{2.28}$$

$$\begin{aligned} \beta_h &= F(K_h - K_{h+1} - 2\overline{X}_h T_h + 2\overline{X}_{h+1} T_{h+1}) + 2AB(\overline{X}_h - \overline{X}_{h-1}) \quad h = 1, ..., L-2 \\ \beta_h &= F(K_h - 2\overline{X}_h T_h) + 2AB\overline{X}_h \quad h = L-1 \end{aligned} \tag{2.29}$$

$$\begin{aligned} \gamma_h &= FT_h(\overline{X}_h^2 + S_h^2) - FT_{h+1}(\overline{X}_{h+1}^2 + S_{h+1}^2) - AB(\overline{X}_h^2 - \overline{X}_{h+1}) \quad h = 1, ..., L-2 \\ \gamma_h &= FT_h(\overline{X}_h^2 + S_h^2) - AB\overline{X}_h^2 - F^2 \qquad h = L-1 \end{aligned} \tag{2.30}$$

As constantes A, B, F são dadas por:

$$A = \sum_{h=1}^{L-1} (W_h \overline{X}_h)^p \quad , B = \sum_{h=1}^{L-1} (W_h S_h)^2 (W_h \overline{X}_h)^{-p} \quad , F = CV^2 N \overline{X}^2 + \sum_{h=1}^{L-1} W_h S_h^2 \tag{2.31}$$

Os termos K_h e T_h são dados por:

$$K_h = Bp(W_h\overline{X}_h)^{p-1} - Ap(W_hS_h)^2(W_h\overline{X}_h)^{-p-1} \quad , \quad T_h = AW_h(W_h\overline{X}_h)^{-p} \tag{2.32}$$

As raízes da equação 2.27 são dadas pela fórmula de Bhaskara:

$$b_h = \frac{-\alpha_h \pm \sqrt{\beta_h^2 - 4\alpha_h\gamma_h}}{2\alpha_h} \tag{2.33}$$

Embora os parâmetros populacionais $W_h, \overline{X}_h, S_h^2$ sejam derivados de uma função de densidade, Lavallée e Hidiroglou (1988) consideraram que a população é finita e, por isso, puderam substituí-los por suas expressões analíticas usuais, como definido na seção 2.2. Como citado anteriormente, na prática, define-se o valor do parâmetro p igual à raiz quadrada ($p = 1/2$) ou igual à raiz cúbica ($p = 1/3$) dessa variável auxiliar.

O método de Lavallée e Hidiroglou (1988) pode ser resumido no Algoritmo 2.1. Esse algoritmo iterativo termina quando a diferença entre os limites atuais (b_h'') e os limites da iteração anterior (b_h') for menor que a tolerância desejada ε, tal que $\varepsilon \in (0, 1)$.

Algoritmo 2.1 Algoritmo de Lavallée e Hidiroglou (1988)

Entrada:

1 Passo 1: Ordenar a variável de estratificação de modo crescente, fazendo $b_0 = x_1$ e $b_L = x_N$;
2 Passo 2: Definir valores iniciais $b_1' < ... < b_{L-1}'$ para os limites, tais que $b_h' \in (b_0, b_L)$;
3 **repita**
4 Passo 3: Calcular W_h , $\overline{X}_h$ e S_h^2 com base nos limites b_h' atuais;
5 Passo 4: Calcular e substituir por novos pontos de corte $b_1'' < ... < b_{L-1}''$ mediante a equação 2.33;
6 **até** *MAX* $\frac{L-1}{h=1}|b_h'' - b_h'| < \varepsilon$;
7 **Saída:** Melhor Solução para $b_1, ..., b_{L-1}$

Em um estudo mais recente, Rivest (2002) propôs uma generalização do método de Lavallée e Hidiroglou (1988), utilizando um modelo de regressão para calcular os valores da variável de interesse, ao invés de substituí-lo pela variável de estratificação, além de possibilitar a escolha da forma de alocação entre Neyman e Potência. Adicionalmente, Lednicki e Wieczorkowski (2003) se basearam no trabalho de Rivest (2002) para propor apenas uma modificação na forma de resolver o problema de minimização, utilizando o algoritmo Simplex. Em contrapartida, Bramati (2012) argumentou que o modelo linear de Rivest (2002) é altamente sensível à presença de dados contaminados (presença de outliers na variável de interesse e/ou na variável de estratificação) e propôs um estimador de regressão robusto, que não é vulnerável. No entanto, quando os dados não estão contaminados, e são gerados

a partir de uma distribuição simétrica, esta abordagem é menos eficaz do que a anterior. Assim, o método original de Lavallée e Hidiroglou (1988) ainda se mantém como o principal, tendo sido desenvolvidas apenas novas variantes para ele.

Um outro método foi proposto por Kozak (2004), em que ele cria um algoritmo de busca aleatória para resolver o problema de minimização do tamanho de amostra total. Nesse estudo, o autor cita o trabalho de Lednicki e Wieczorkowski (2003), mas faz uma crítica quanto à utilização do algoritmo Simplex, pois considera que o método é pouco eficiente em relação ao tempo de processamento, especialmente quando há um grande número de variáveis ou de observações.

A equação utilizada é muito parecida com 2.26, alterando apenas a forma de alocação para Neyman em vez da alocação potência:

$$n = N_L + \frac{(\sum_{h=1}^{L-1} W_h S_h)^2}{CV^2 \cdot \overline{X}^2 + \frac{1}{N}\sum_{h=1}^{L-1} W_h S_h^2} \tag{2.34}$$

Essa abordagem considera as seguintes restrições:

$$N_h \geq 2, \qquad h = 1, ..., L \tag{2.35}$$

$$2 \leq n_h \leq N_h, \qquad h = 1, ..., L-1 \tag{2.36}$$

Note que Kozak, alterou a restrição 2.22 para 2.35, com o intuito de evitar a criação de estratos com apenas uma unidade populacional. Também o fez com a restrição 2.23, fixando um tamanho mínimo de duas unidades amostrais por estrato, criando assim a restrição 2.36.

Assim, o método de busca aleatória que visa minimizar a equação 2.34, sujeito às restrições 2.35 e 2.36 pode ser resumido no algoritmo iterativo 2.2. O parâmetro q é um número inteiro de acordo com o tamanho da população, não devendo ser maior que 5 e não podendo ser igual a 1, e ainda para populações menores (até algumas centenas de observações) deve ser igual a 2 ou 3, segundo Kozak (2004).

No passo 4 do algoritmo há dois critérios de parada possíveis: quando r atingir o número de R iterações e quando atingir m iterações sem melhoria da função objetivo. Esses dois parâmetros (R e m) já estão definidos no algoritmo disponibilizado pelo autor, sendo R fixado em 10.000 e m variando entre 50 e 500, de acordo com as características da população.

Portanto, quando um dos critérios for satisfeito na linha 14, o processo iterativo é interrompido.

Algoritmo 2.2 Algoritmo de Kozak (2004)

Entrada:

1 Passo 1: Ordenar a população segundo a variável de estratificação;
2 Passo 2:
3 Definir um vetor inicial **b** $= (b_1, ..., b_{L-1})$ com os limites dos estratos;
4 Calcular n segundo a equação 2.34;
5 Passo 3: **Para** $r = 0$ *até* R **Faça**
6 | Passo 3.a: Calcular **b'**, tal que **b'** $= b_i + j$, onde $j \in (-q, -1) \cup (1, q)$;
7 | Passo 3.b: Calcular n' segundo a equação 2.34;
8 | Passo 3.c:
9 | **Se** *as restrições 2.35 e 2.36 forem satisfeitas e* $n' \leq n$ **Então**
10 | | $\mathbf{b}_{r+1} =$ **b'**
11 | **Senão**
12 | | $\mathbf{b}_{r+1} =$ **b**
13 | **Fim Se**
14 | Passo 4: Até Satisfazer o Critério de Parada;
15 **Fim Para**
16 **Saída:** Melhor Solução **b**

Considerando os resultados obtidos nos experimentos empíricos de Kozak (2004), o método de Kozak conseguiu produzir os melhores resultados comparados aos demais métodos testados e também foi o mais rápido entre eles. Já em Kozak (2006), o autor adaptou seu método de busca aleatória para atender ao primeiro objetivo do problema de estratificação, ou seja, minimizar a precisão. No trabalho seguinte, Kozak e Verma (2006) avaliaram o algoritmo quanto aos dois objetivos, em que apresentou melhores resultados que o método de Gunning e Horgan (2004). Entretanto, para o segundo objetivo, o algoritmo de busca aleatória produziu resultados equivalentes aos obtidos pelo método de Lavallée e Hidiroglou (1988), sendo ambos melhores que os resultados obtidos pelo método de Gunning e Horgan (2004). Esse insucesso do método de Gunning e Horgan (2004) para o método de Lavallée e Hidiroglou (1988) era, de certa forma, inesperado, pois nos artigos de Gunning e Horgan (2004) e Horgan (2006), os resultados obtidos pelo método de Gunning e Horgan (2004) eram melhores que o de Lavallée e Hidiroglou (1988). Por isso, Kozak e Verma (2006) sugerem que há a necessidade de novos experimentos com populações assimétricas reais. A versão multivariada do algoritmo para os dois objetivos está em Kozak, Verma e Zieliński (2007).

3 - Metaheurísticas

Atualmente, as metaheurísticas têm sido aplicadas em diversos problemas de otimização combinatória. Segundo Blum e Roli (2003), isso se deve à importância dos problemas de otimização combinatória para o campo científico e o mundo industrial, e também, à complexidade desses problemas. Por isso, de forma a facilitar o entendimento da metaheurística que é proposta neste trabalho, para a resolução do problema de estratificação, se faz necessário apresentar os conceitos básicos e os métodos que serão mencionados e utilizados neste estudo.

3.1 - Conceitos Básicos

Em linhas gerais, a resolução de um problema de otimização está associada à escolha da melhor configuração de um conjunto de variáveis para alcançar os objetivos, sendo tais variáveis contínuas ou discretas. Conforme Papadimitriou e Steiglitz (1982), quando as variáveis de decisão são contínuas e o espaço de busca é finito, fica caracterizado o problema de otimização, que pode ser definido, como no exemplo abaixo:

$$\begin{aligned} \text{minimizar} \quad & f(x) \\ \text{sujeito a} \quad & g(x) \geq 0 \\ & x \geq 0 \end{aligned}$$

em que $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ e as funções $f(x)$ e $g(x)$ podem ser lineares ou não lineares.

Um conjunto de todas as soluções viáveis é dado por $\mathcal{S} = \{x \in \mathbb{R}^n | g(x) \geq 0\}$, em que x satisfaz todas as restrições e $\mathcal{S}$ é chamado de espaço de busca. Para resolver um problema de otimização combinatória de minimização deve-se encontrar uma solução $x^* \in \mathcal{S}$ com o valor mínimo da função objetivo, ou seja, $f(x^*) \leq f(x), \forall x \in \mathcal{S}$. Essa solução x^* é chamada de mínimo global.

Conforme a complexidade, um problema de otimização pode ser resolvido por méto-

dos exatos ou por métodos não exatos. Os métodos exatos produzem soluções ótimas, caracterizadas pelos mínimos (ou máximos) globais, provando sua otimalidade. Entretanto, para problemas com alto grau de complexidade, há um grande custo de tempo de processamento e sem a garantia, até mesmo, de produzir um mínimo (ou máximo) local. Entre os métodos dessa classe destacam-se: programação dinâmica, *branch and bound* e programação por restrições.

Por outro lado, dentre os métodos não exatos, destacam-se as heurísticas e as metaheurísticas. Conforme Resende e Sousa (2003), as heurísticas são métodos que produzem soluções viáveis e de boa qualidade. Já as metaheurísticas produzem, em geral, soluções de melhor qualidade do que as heurísticas, pois são métodos que coordenam procedimentos de busca com estratégias de mais alto nível, de modo a criar um processo capaz de escapar de mínimos locais de baixa qualidade e realizar uma busca robusta no espaço de soluções de um problema, Glover e Kochenberger (2003).

De forma a facilitar o entendimento são apresentados a seguir algumas definições importantes, segundo Blum e Roli (2003).

Definição 3.1 (Vizinhança de x)**.** *A estrutura de vizinhança é uma função* $\eta : \mathcal{S} \rightarrow 2^{|\mathcal{S}|}$ *que atribui a cada* $x \in \mathcal{S}$ *um conjunto de vizinhos* $\eta(x) \subseteq \mathcal{S}$*.*

Definição 3.2 (Minimo Local)**.** *Uma solução de mínimo local, em uma estrutura de vizinhança* η*, é uma solução* x' *tal que* $\forall x \in \eta(x') : f(x') \leq f(x)$ *.*

Um mínimo local difere do mínimo global pela vizinhança η, considerando um espaço de busca $\mathcal{S}$, pois no mínimo local, o ponto é ótimo para aquela vizinhança, enquanto no mínimo global, o ponto é ótimo para todo o espaço de busca. Assim, conforme Talbi (2009), para o mesmo problema de otimização, um ótimo local para uma vizinhança η_1 pode não ser um ótimo local para uma vizinhança diferente η_2.

A Figura 3.1 ajuda a exemplificar as últimas definições, em que há quatro soluções viáveis (x_1, x_2, x_3, x_4) do espaço de busca $\mathcal{S}$ e cada solução tem sua vizinhança representada pela região pontilhada ao redor de cada ponto e denotadas por $\eta(x_1), \eta(x_2), \eta(x_3), \eta(x_4)$, respectivamente. A solução x_1 não representa um mínimo local, pois existe uma solução melhor na vizinhança $\eta(x_1)$. Por outro lado, x_2, x_3, x_4 são mínimos locais, pois não existem soluções melhores na respectivas vizinhanças. E, além disso, a solução x_2 é um mínimo global, pois não existe nenhuma solução melhor para todo o espaço de busca $\mathcal{S}$, ou seja, é o valor mínimo da função objetivo e por isso, denota-se x_2 por x^*.

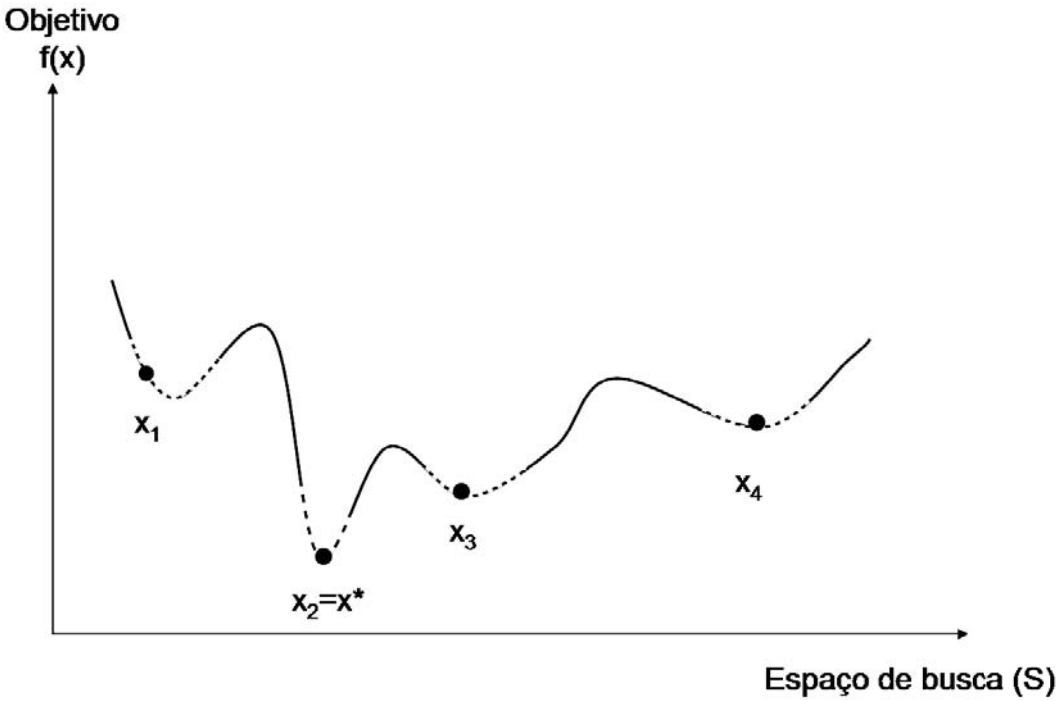

Figura 3.1: Exemplos de mínimos locais para um problema de minimização
Fonte: Elaboração Própria.

Embora, ainda hoje, seja difícil fornecer uma definição exata de uma metaheurística, Blum e Roli (2003) listaram nove características comuns a todos os algoritmos dessa classe, das quais, aqui foram destacadas quatro:

- Metaheurísticas são estratégias que orientam o processo de busca.
- O objetivo é explorar eficientemente o espaço de busca, a fim de produzir soluções de boa qualidade em tempo computacional factível.
- Os algoritmos são, geralmente, não-determinísticos.
- Pode incorporar mecanismos para evitar de ficar “preso” em regiões limitadas do espaço de busca.

3.2 - Variable Neighborhood Search (VNS)

A Busca em Vizinhança Variável (tradução livre de *Variable Neighborhood Search*) proposta por Mladenović e Hansen (1997) explora, incrementalmente, vizinhanças mais distantes da solução corrente, partindo da atual para a nova solução se, e somente se, uma melhora na função objetivo ocorrer. O método explora o fato de que o uso de várias vizinhanças

de busca local pode gerar diferentes ótimos locais e como o ótimo global coincide com o ótimo local de uma vizinhança, em algum momento, ele deve ser alcançado.

Algoritmo 3.1 Pseudo Código do VNS

Entrada:

1 Defina o conjunto de estruturas de vizinhanças η_k ($k = 1$, ..., k_{max}) que serão usadas na busca;
2 $x = x_0$; $< $ Encontre uma solução inicial $>$
3 **Enquanto** *Não atingir o Critério de Parada* **Faça**
4 $k = 1$;
5 **Enquanto** $k \leq k_{max}$ **Faça**
6 x_1 = Perturbação(x);
7 x_2 = Busca Local(x_1);
8 **Se** $f(x_2) < f(x)$ **Então**
9 $x = x_2$;
10 $k = 1$;
11 **Senão**
12 $k = k + 1$
13 **Fim Se**
14 **Fim Enquanto**
15 **Fim Enquanto**
16 **Saída:** Melhor Solução

O método é composto por cinco fases, conforme Algoritmo 3.1:

(i) Inicialização: consiste em escolher uma solução inicial viável para o problema. Geralmente, utiliza-se um método aleatório por ser mais rápido, mas podem ser geradas soluções de baixa qualidade o que fará o algoritmo demorar para convergir, Talbi (2009). Como alternativa, podem-se utilizar soluções conhecidas ou aplicar métodos construtivos gulosos[5] para gerar soluções iniciais.

(ii) Perturbação: Esse procedimento tem o objetivo de escapar das soluções de mínimo local de baixa qualidade, buscando soluções mais distantes, a fim de explorar completamente o espaço de busca $\mathcal{S}$. Entretanto, ela precisa guardar características da solução corrente. Na linha 6 do algoritmo, é gerada uma solução x_1 a partir da k-ésima vizinhança da solução x.

(iii) Busca local: A cada iteração procura-se uma solução melhor que a solução atual dentro da vizinhança. Existem três tipos básicos de busca local, conforme [Talbi (2009);

[5]Um algoritmo guloso produz uma solução viável, fazendo uma sequência de escolhas em pontos de decisão. A opção escolhida é a que parece melhor no momento no problema atual, mas não leva em consideração os resultados de subproblemas associados, para mais detalhes ver Cormen et al. (2002).

Souza (2011)] : primeira melhoria (tradução livre de *first improvement*), seguindo alguma ordem de exploração, assim que for encontrado um vizinho com solução melhor que a solução atual, a busca termina; melhor melhoria (tradução livre de *best improvement*), em que todos os vizinhos são testados exaustivamente até encontrar o melhor vizinho; seleção aleatória (tradução livre de *random selection*), em que um vizinho é escolhido aleatoriamente e é aceito se for melhor que a solução corrente. Caso contrário, testa-se outro vizinho aleatoriamente, sendo o procedimento interrompido após algumas iterações sem melhoria.

(iv) Substituição: Representada pelas linhas 8 até 13, significa que se a solução obtida x_2, for estritamente melhor que a solução x corrente, o valor atual é substituído e a busca reinicia na primeira vizinhança η_1. Caso contrário, faz-se a busca a partir da próxima vizinhança η_{k+1}, até que seja encontrada uma solução melhor e volte a busca para a vizinhança η_1. A ideia é explorar uma vizinhança ao máximo, enquanto resultados satisfatórios forem sendo obtidos, Santos (2014). A Figura 3.2 ilustra esse procedimento, em que se inicia a busca na vizinhança η_1, caso não encontre um vizinho com solução melhor, faz-se a busca na vizinhança η_2, depois na vizinhança η_3, até a vizinhança $\eta_{k_{max}}$. Se, em alguma dessas vizinhanças, for encontrado um vizinho com solução melhor, retorna-se para a vizinhança η_1 e reinicia-se o procedimento.

(v) Critério de parada: Pode ser um tempo máximo de uso da CPU, o número máximo de iterações, ou o número máximo de iterações entre duas melhorias.

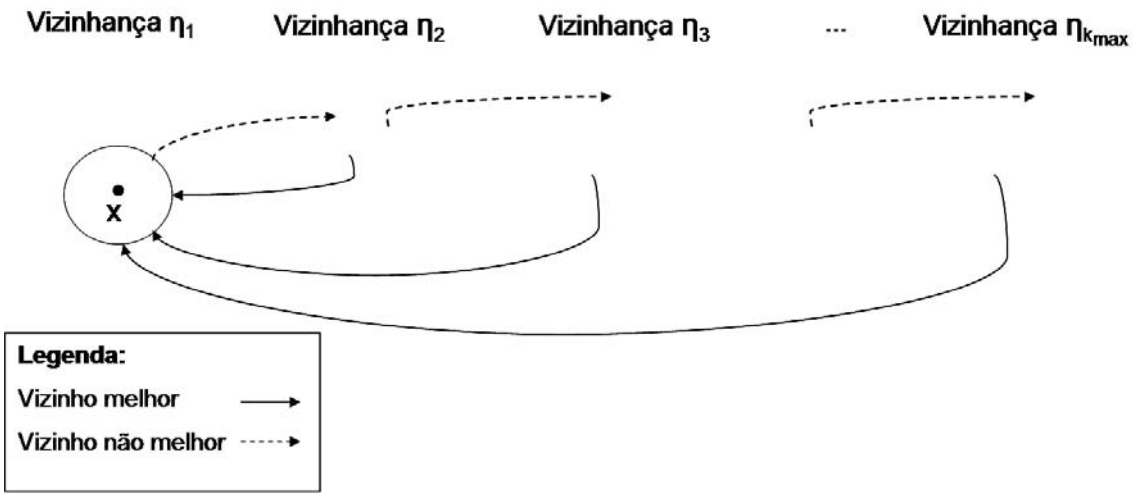

Figura 3.2: O princípio da vizinhança variável
Fonte: Adaptado de Talbi (2009).

Além disso, cabe ainda ao pesquisador definir: as estruturas de vizinhança η_k (linha 1), que podem ser arbitrárias, aninhadas ou ainda, uma sequência com aumento de cardinalidade, tal que $|\eta_1| < |\eta_2| < ... < |\eta_{k_{max}}|$, conforme Blum e Roli (2003); um critério de parada (linha 3) dentre as opções do item (v); e o parâmetro k_{max} (linha 5), que representa o número máximo de vizinhanças a serem visitadas. Como saída, o algoritmo fornece a melhor solução encontrada.

3.3 - Variable Neighborhood Decomposition Search (VNDS)

A Busca Decomposta em Vizinhança Variável (tradução livre de *Variable Neighborhood Decomposition Search*) proposta por Hansen, Mladenović e Perez-Brito (2001) consiste em uma variação do método VNS, 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 e tanto a busca local são definidas em subproblemas (problemas de tamanhos menores do que o inicial).

Algoritmo 3.2 Pseudo Código do VNDS

Entrada:
1 Defina o conjunto de estruturas de vizinhanças η_k ($k = 1$, ..., k_{max}) que serão usadas na busca;
2 $x = x_0$; $< $ Encontre uma solução inicial $>$
3 **Enquanto** *Não atingir o Critério de Parada* **Faça**
4 $k = 1$;
5 **Enquanto** $k \leq k_{max}$ **Faça**
6 x_1 = Perturbação(x); $< x_1$ difere de x por k atributos $>$
7 x_2 = Busca Local(x_1, k); $<$ Busca somente nos k atributos permitidos $>$
8 **Se** $f(x_2) < f(x)$ **Então**
9 $x = x_2$;
10 $k = 1$;
11 **Senão**
12 $k = k + 1$
13 **Fim Se**
14 **Fim Enquanto**
15 **Fim Enquanto**
16 **Saída:** Melhor Solução

Segundo Hansen, Mladenović e Perez-Brito (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 $x_1 \in \eta_k(x)$), o segundo resolve a cada iteração um subproblema em algum subespaço $V_k \subseteq \eta_k(x)$, com $x_1 \in V_k$. 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 reinicia da primeira vizinhança, conforme Algoritmo 3.2.

Em relação ao Algoritmo 3.1 a diferença está nas linhas 6 e 7. Como só é permitido mexer em k variáveis, a perturbação de x produzirá uma solução x_1 muito similar a x, 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, Mladenović e Perez-Brito (2001); Hansen e Mladenović (2001)], esse processo de decompor em subproblemas tende a ser mais eficiente (economizando tempo computacional) e para problemas muito grandes, esse método tende a produzir melhores resultados que o VNS.

3.4 - Path-Relinking (PR)

A Reconexão por Caminhos (tradução livre de *Path-Relinking*) proposta, originalmente, por Fred Glover no primeiro capítulo do livro de Barr, Helgason e Kennington (1996) é uma estratégia de intensificação com o objetivo de explorar trajetórias de soluções obtidas previamente por uma Busca Tabu ou por uma Busca Dispersa (tradução livre de *Scatter Search*). Segundo Glover e Kochenberger (2003) o método tem este nome por dois motivos: gerar um novo caminho entre soluções previamente ligadas por uma série de movimentos executados durante a busca; gerar um caminho entre soluções anteriormente ligadas a outras soluções, mas não entre si.

Conforme Talbi (2009), a ideia principal é gerar e explorar a trajetória no espaço de busca conectando uma solução inicial x_i e uma solução denominada de solução alvo x_t. No caminho entre essas duas soluções são criadas várias soluções intermediárias. Como resposta obtém-se a melhor dessas soluções intermediárias, conforme Algoritmo 3.3.

Algoritmo 3.3 Pseudo Código da Reconexão_Caminhos

Entrada:

1 Solução inicial x_i e Solução alvo x_t;
2 $x = x_i$;
3 **Enquanto** $Dist(x_i, x_t) \neq 0$ **Faça**
4 | Encontre o melhor movimento m que diminui $Dist(x \oplus m, t)$; $< m \in \Delta >$;
5 | $x = x \oplus m$; $<$ Aplique o movimento m na solução $x >$;
6 **Fim Enquanto**
7 **Saída:** Melhor Solução encontrada na trajetória entre x_i e x_t

Seja Δ o conjunto de componentes diferentes das soluções x_i e x_t e para cada componente $m \in \Delta$, define-se $x \oplus m$ como sendo a solução obtida de x pela complementação do

valor atual de x_m, [Talbi (2009); Gendreau e Potvin (2010)]. A função $Dist(x_i, x_t)$ pode, em tese, assumir qualquer medida de distância da literatura. Neste trabalho, são consideradas as medidas de distância mais alinhadas com o problema a ser resolvido. Uma delas, conhecida como distância de Hamming, é dada pelo número de entradas diferentes nos elementos das soluções x_i e x_t. Dessa forma, o Algoritmo 3.3 é interrompido quando a solução atual for igual a solução alvo.

Um exemplo para a Distância de Hamming está na tabela abaixo, em que as soluções estão em codificação binária. Nesse exemplo, a solução testada é a X1 em relação às demais. Assim, note que a distância entre X1 e X2 é 5, porque cinco bits diferem (os bits em negrito). O mesmo raciocínio é usado para as distâncias entre (X1 e X3), (X1 e X4) e (X1 e X5).

Tabela 3.1: Exemplo para a Distância de Hamming

Solução	Codificação Binária							Dist(X1,Xi)
X1	0	0	0	1	0	1	0	-
X2	**1**	**1**	**1**	1	**1**	1	**1**	5
X3	0	0	0	**0**	**1**	1	**1**	3
X4	0	0	**1**	1	0	**0**	**1**	3
X5	0	**1**	0	1	0	1	0	1

Fonte: Adaptado de Han, Kamber e Pei (2011)

Outra medida possível poderia ser a Distância de Minkowski de ordem h. Conforme Han, Kamber e Pei (2011), para $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ e $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$, a distância de ordem h entre esses dois pontos é dada por:

$$dist(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{ip} - x_{jp}|^h} \tag{3.1}$$

No caso particular em que $h = 2$ tem-se a Distância Euclidiana, e quando $h = 1$ tem-se a Distância de Manhattan ou distância do "quarteirão". Por exemplo, sejam os pontos $x_1 = (1, 2)$ e $x_2 = (3, 5)$, conforme Figura 3.3, a distância Euclidiana é representada pelo comprimento da reta ligando os dois pontos, medindo $\sqrt{|3-1|^2 + |5-2|^2} = \sqrt{2^2 + 3^2} = 3,6$ unidades. Enquanto a distância de Manhattan é representada pela soma dos comprimentos das retas pontilhadas, medindo $|3-1| + |5-2| = 2 + 3 = 5$ unidades.

Observe que se as soluções x_i e x_t são binárias, a distância de Hamming é um caso particular da distância de Minkowski. Caso isso não seja verdade, não se pode afirmar que os valores das distâncias sejam iguais. No entanto, como o que importa para o critério de

parada do algoritmo é que a distância entre os pontos seja nula, as duas distâncias terão valor nulo quando estiverem no mesmo ponto. Nos algoritmos desenvolvidos nessa dissertação, optou-se por utilizar a distância de Manhattan.

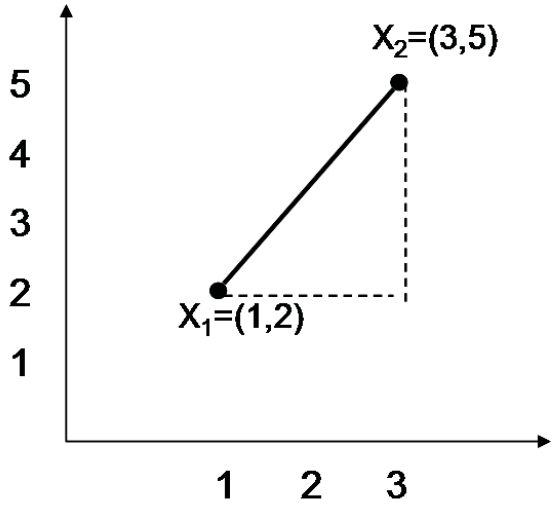

Figura 3.3: Exemplo de Distância Euclidiana e Distância de Manhattan
Fonte: Adaptado de Han, Kamber e Pei (2011).

Vários tipos de estratégia de reconexão podem ser considerados para definir a direção dos movimentos, Talbi (2009):

- Reconexão para Frente (tradução livre de *Forward Relinking*): partindo de x_i em direção a x_t;
- Reconexão para Trás (tradução livre de *Backward Relinking*): partindo de x_t em direção a x_i;
- Reconexão para Frente e para Trás (tradução livre de *Back and Forward Relinking*): as duas trajetórias anteriores são construídas em paralelo;
- Reconexão Mista (tradução livre de *Mixed Relinking*): assim como no item anterior, as duas trajetórias são construídas em paralelo, mas ambas vão em direção a uma solução guia x_g que está na mesma distância entre x_i e x_t.

Tais tipos de estratégias estão ilustradas na Figura 3.4, em que as soluções x_i, x_t e x_g são representadas pelos pontos pretos. Os pontos brancos representam as soluções intermediárias e os pontos cinzas, as melhores soluções produzidas por cada trajetória. Assim, a Reconexão para Frente é a linha pontilhada, a Reconexão para Trás é a linha tracejada, a Reconexão para Frente e para Trás são as linhas pontilhadas e tracejadas, conjuntamente.

Por fim, a Reconexão Mista é a linha sólida, em que as trajetórias vindas de direções opostas se encontram no ponto x_g.

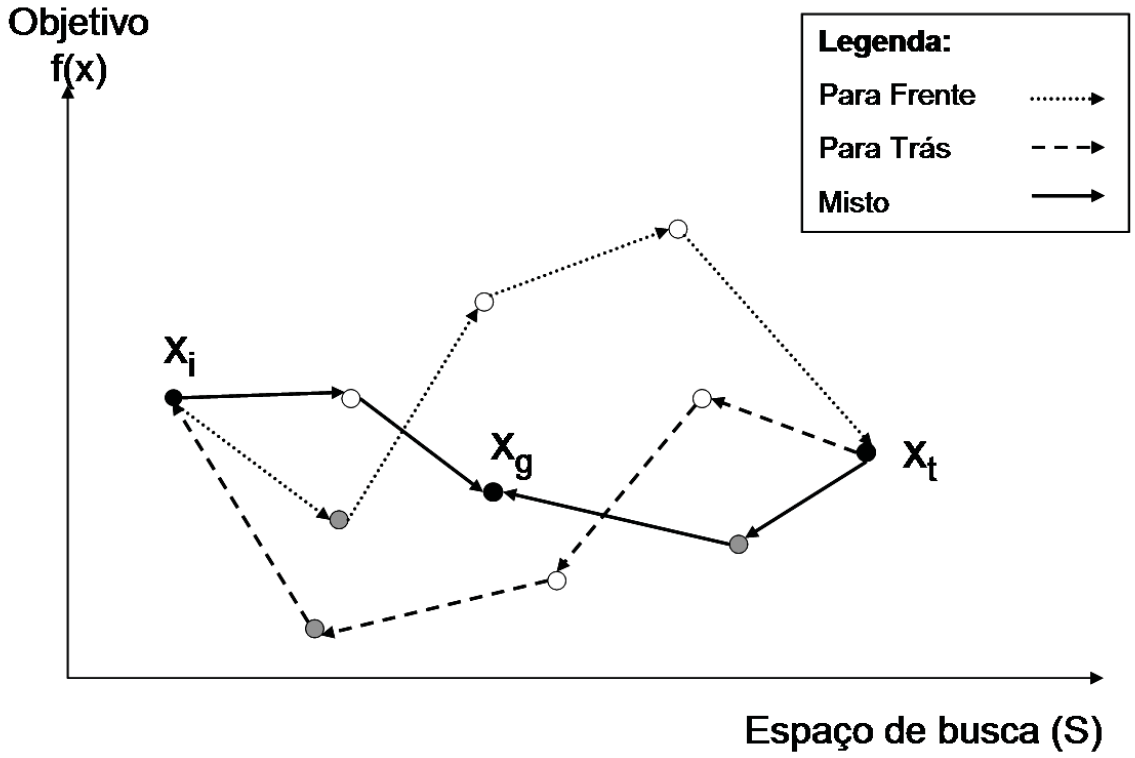

Figura 3.4: Diferentes Estratégias de Reconexão por Caminhos
Fonte: Adaptado de Talbi (2009).

4 - Metodologia

Uma vez que os conceitos de amostragem e de metaheurísticas já foram apresentados nos capítulos anteriores, pode-se compreender melhor o método que foi proposto para resolver o problema de estratificação. Inicialmente, são explicitadas as bases de dados que foram utilizadas para testar o método, e no fim do capítulo, os métodos existentes são comparados ao método proposto.

4.1 - Bases de Dados

A literatura existente para a resolução do problema de estratificação utiliza, usualmente, populações disponíveis em pacotes estatísticos ou pela geração de populações a partir de distribuições estatísticas. Dessa forma, os trabalhos da literatura existentes são comparáveis. Nessa dissertação serão apresentados os resultados dessas populações usuais, descritas na subseção a seguir.

Por outro lado, serão utilizadas, também, populações com duas características diferentes do usual: populações de tamanho muito grande e com uma restrição não muito habitual. Essas duas características estão presentes na Pesquisa Anual do Comércio, mais especificamente, essa pesquisa tem uma restrição particular que deve ser considerada e será apresentada na subseção 4.1.2.

Assim, considerando os dois tipos de bases de dados foram utilizadas 100 populações no experimento empírico realizado nessa dissertação.

4.1.1 - Populações da Literatura

A maioria dos métodos existentes na literatura para resolução do problema de estratificação, tanto para o primeiro objetivo, como para o segundo objetivo, utilizaram populações disponíveis em pacotes estatísticos ou optaram pela geração de populações a partir de distribuições estatísticas. Tais populações, por exemplo, estão descritas e foram utilizadas em Hedlin (2000); Gunning e Horgan (2004); Keskintürk e Er (2007); Brito et al. (2011);

Veiga (2015).

Dentre essas populações da literatura, optou-se por utilizar 25 delas, onde a maioria está disponibilizado gratuitamente em um software de estatística, chamado R (http://www.r-project.org) nos pacotes (http://cran.r-project.org/web/packages/available_packages_by_name.html): *stratification*, *GA4Stratification* e *sampling*. Há ainda algumas populações que foram fornecidas pelos autores. As descrições dessas populações estão apresentadas na Tabela A.1 do Apêndice, indicando o pacote do software R a que a população pertence ou a referência bibliográfica que a define.

Tabela 4.1: Informações básicas das 25 populações da literatura

ID	Tamanho Populacional (N)	Valores Distintos (w)	Mínimo	Máximo	Assimetria
U01	430	353	50	24.250	4,6
U02	1.000	1.000	358	986	-0,7
U03	1.000	1.000	0	13	2,7
U04	1.000	1.000	0,1	23,4	1,4
U05	3.369	1.129	40	28.000	6,4
U06	16.057	225	0	690.000	2,7
U07	487	487	63.582,9	10.446.591,8	10,1
U08	4.000	51	3	72	1,4
U09	4.000	2.837	243	28.578	2,7
U10	2.000	581	6	2.793	3,5
U11	10.000	5.453	62	74.398	4,2
U12	284	264	173	47.074	8,7
U13	2.000	2.000	141,2	486.366,5	8,6
U14	1.000	1.000	73,6	127,3	0,0
U15	284	68	4	671	8,5
U16	800	402	1	473.510	22,2
U17	284	277	347	59.877	7,9
U18	338	101	18	280	2,3
U19	2.896	881	0	3.634	2,7
U20	589	589	1.097	5.416.418,8	9,3
U21	357	200	70	977	2,1
U22	1.038	116	10	198	2,9
U23	677	576	200	9.623	2,5
U24	1.000	1.000	12,7	60,5	0,6
U25	16.057	6.405	-11.244	86.010	4,5

As características básicas das 25 populações da literatura estão sumarizadas na Tabela 4.1, em que a primeira coluna representa o código de identificação da população, a segunda coluna é o tamanho populacional (N), a terceira coluna é a quantidade de valores distintos (w), a quarta coluna é o valor mínimo, a quinta coluna é o valor máximo e a última

coluna é o coeficiente de assimetria. Vale destacar três pontos: só há três populações com tamanho superior a $N = 5.000$; a população $U25$ é a única que apresenta valores negativos; há uma população com assimetria negativa (U02), outra com assimetria nula (U14) e a maior assimetria positiva é de $22,2$ da população U16.

As 75 populações restantes utilizadas nesse trabalho foram obtidas a partir de dados reais da Pesquisa Anual de Comércio, as quais serão descritas na próxima subseção.

4.1.2 - Pesquisa Anual de Comércio

A Pesquisa Anual de Comércio (PAC), realizada pelo Instituto Brasileiro de Geografia e Estatística (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, salários, receita, entre outras variáveis, segundo domínios definidos por níveis geográficos em combinação com classes de atividades da Classificação Nacional de Atividades Econômicas (CNAE), conforme IBGE (2015).

A população alvo da pesquisa é constituída pelo conjunto de empresas comerciais pertencentes ao Cadastro Central de Empresas (CEMPRE) do IBGE situadas no Território Nacional, classificadas no setor de comércio de acordo com a CNAE. O Cadastro Básico de Seleção (CBS) é construído a partir do CEMPRE e atualizado anualmente pelos resultados da Relação Anual de Informações Sociais (RAIS) e do Cadastro Geral de Empregados e Desempregados (CAGED), ambos disponibilizados pelo Ministério do Trabalho. A unidade amostral é a empresa, que é, também, a unidade de investigação e unidade de análise da pesquisa.

4.1.2.1 - Plano Amostral

O plano amostral da pesquisa, conforme IBGE (2015), utiliza amostragem estratificada. Mais especificamente, são definidos dois níveis de estratos: natural e final. O estrato natural é formado a partir da combinação da Unidade da Federação (UF) e da classificação de atividade a três ou quatro dígitos da CNAE [6], ou seja, é definido por uma questão administrativa, assim como a definição de estratificação natural apresentada na seção 2.2. Por exemplo, as empresas da CNAE 4762 não estarão todas no mesmo estrato final, isto é, apenas estarão agrupadas se elas pertencerem a mesma UF. Por outro lado, os estratos finais

[6] A quatro dígitos para seis Unidades da Federação: RJ, SP, MG, PR, SC e RS. As demais Unidades da Federação formam estratos com CNAE a três dígitos.

são construídos pela subdivisão de cada estrato natural de acordo com o porte da empresa, ou seja, são definidos de modo a criar grupos ainda mais homogêneos, o que implica, portanto, uma estratificação estatística.

Os estratos finais são construídos de acordo com o porte da empresa, definidos por um critério gerencial, de modo que, 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 estratos, também de acordo com o porte da empresa: A_1, empresas com 0 até 4 pessoas ocupadas; A_2, empresas com 5 a 9 pessoas ocupadas; A_3, empresas com 10 a 19 pessoas ocupadas. Nesses estratos, utiliza-se amostragem aleatória simples sem reposição das unidades elementares que, por sua vez, correspondem às empresas.

O uso da amostragem estratificada por corte na PAC se justifica, pois há uma grande concentração de pequenas empresas e poucas empresas de maior porte. Esse comportamento é típico de uma distribuição assimétrica à direita, o que se torna evidente na Figura 4.1, em que a mediana está mais próxima do 1º quartil, enquanto a média é maior que o 3º quartil. Além disso, na Figura 4.2 pode-se ver que mais de 1 milhão de empresas possuem até duas pessoas ocupadas.

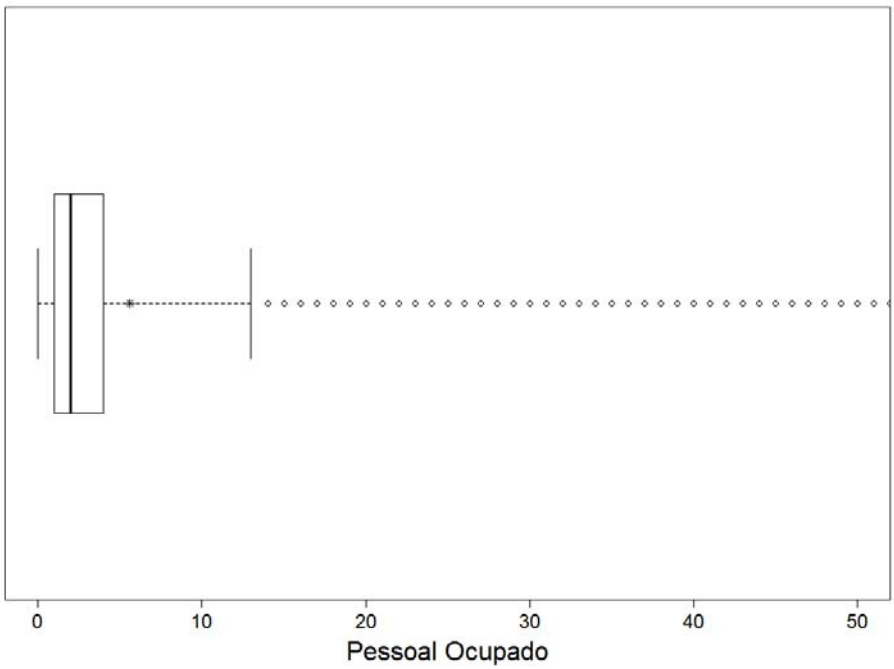

Figura 4.1: Boxplot da população de empresas da PAC 2014, segundo o número de pessoas ocupadas

Fonte: CBS PAC 2014.

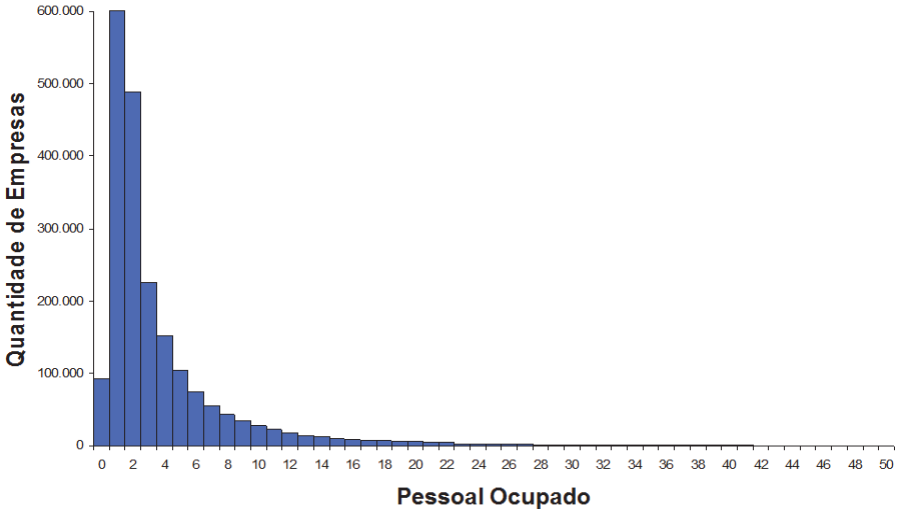

Figura 4.2: Frequência absoluta da população de empresas da PAC 2014, segundo o número de pessoas ocupadas

Fonte: CBS PAC 2014.

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 10%. A amostra de empresas é obtida por amostragem aleatória simples sem reposição 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 $n_C = N_C$. 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 (2.20) e utiliza-se a alocação de Neyman – Equação (2.12) – para substituí-las na Equação (2.14). Após algumas operações algébricas, tem-se:

$$n_A = \frac{N_A^2 (\sum_{h=1}^{3} W_h S_h)^2}{CV(\hat{X}_{AEC})^2 \cdot X^2 + N_A \sum_{h=1}^{3} W_h S_h^2}, \tag{4.1}$$

sendo que $N_A = N - N_C$ é o tamanho populacional do estrato amostrado (A), $W_h = N_h / N_A$ 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 dado pela Equação (2.20).

Além disso, arbitrou-se um número mínimo de cinco empresas, quando houver, em cada estrato final amostrado, ou seja, $n_{A_h} \geq 5$ $(h = 1, 2, 3)$. Essa restrição é para evi-

tar 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 precaução é necessária, pois, em alguns estratos específicos, a taxa de não-resposta é alta. A não-resposta ocorre quando um questionário foi a campo, mas não foi respondido pela empresa, devido a inúmeros motivos. Por isso, há a necessidade de alterar a restrição 2.23, pois é preciso definir um limite mínimo para o tamanho amostral por estrato final, fixado em 5. Isto é, o tamanho amostral do estrato (n_h) deve pertencer ao intervalo discreto $[5, N_h]$, só sendo permitido ser menor que 5, quando o tamanho populacional do estrato (N_h) for menor que 5, o que em notação matemática é dado por

$$min\{5, N_h\} \leq n_h \leq N_h \qquad h = 1, ..., L-1. \tag{4.2}$$

Essa restrição não é atendida pelos métodos existentes na literatura. Mesmo havendo uma restrição similar no método de Kozak (2004), conforme Equação 2.36, o valor mínimo não é um parâmetro ajustável, é um parâmetro fixado em dois.[7]

4.1.2.2 - Delimitação

A Pesquisa Anual do Comércio abrange todo o território nacional e a população da pesquisa contempla pouco mais de 2 milhões de estabelecimentos comerciais. Por ser tão abrangente, a aplicação de um censo é inviável e, por isso, utiliza-se amostragem para produzir estatísticas oficiais confiáveis desse ramo da economia. Mesmo considerando apenas a amostra, em todo o país foram aplicados mais de 77 mil questionários em 2014, conforme Tabela 4.2. Além disso, é possível ver que de 2007 para 2014, houve um aumento de 54% de empresas selecionadas no estrato certo, enquanto os tamanhos dos demais estratos se mantiveram praticamente inalterados ou até diminuíram seu tamanho amostral. Em geral, houve um aumento de 38% a mais de empresas para serem entrevistadas. Esse aumento implica maiores custos operacionais e uma maior carga de trabalho em todas as etapas da pesquisa. Por esse motivo, há o interesse em reduzir do tamanho da amostra.

Nessa dissertação, optou-se por limitar os experimentos relacionados à PAC 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. Nessa Unidade da Federação, conforme descrito na seção 4.1.2.1, a população de empresas é sub-

[7] Embora o algoritmo disponibilizado pelo autor permita $n_h = 1$.

Tabela 4.2: Comparação do tamanho amostral da PAC em 2007 e 2014, segundo o tipo do estrato final

Estrato Final	Tamanho Amostral (2007)	Tamanho Amostral (2014)	Variação (em %)
A_1	6.504	5.493	-16
A_2	4.146	4.195	1
A_3	3.974	4.016	1
C	41.859	64.285	54
TOTAL	56.483	77.989	38

Fonte: PAC 2007 e PAC 2014

divida de acordo com classificação de atividade a quatro dígitos, o que resulta em 75 estratos naturais. Portanto, cada estrato natural será tratado como uma população independente, totalizando assim, 75 populações para aplicação do método e, assim, minimizar o tamanho amostral total e calcular os limites dos pontos de corte de cada uma delas.

As características básicas das 75 populações estão sumarizadas na Tabela A.2 do Apêndice, em que a primeira coluna representa o 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 auxiliar (pessoal ocupado) e a última coluna é o coeficiente de assimetria. Assim, há os mais variados tipos de populações, desde muito pequenas com apenas $N = 54$ empresas, até muito grandes com 65.431 empresas, também com valores para w variando de 6 a 384 e todas populações com assimetria positiva variando de $1,2$ a $195,8$.

4.2 - Aplicação do Método: o Algoritmo Proposto

Para cada uma das 100 populações citadas na seção 4.1 (25 populações da literatura e 75 populações associadas à base de dados originada no CBS da PAC 2014), foram calculados novos limites para os pontos de corte, visando a minimização do tamanho amostral total. Com esse objetivo, foram aplicados os métodos de Lavallée e Hidiroglou (1988), Kozak (2004) (descritos na seção 2.4) e também o método proposto nessa dissertação. Entretanto, os dois primeiros apresentam uma limitação, pois não foram concebidos para receber a restrição adicional 4.2.

Na PAC, como não é considerado o custo unitário por estrato, o objetivo resume-se a minimizar o tamanho amostral total (n). E como há a definição do estrato certo, deve-se fazer $n_L = N_L$.

Todas as restrições apresentadas (de 2.22 a 2.25 e 4.2) são duras (tradução livre de *hard constraints*), que conforme Linden (2012), são aquelas restrições que obrigatoriamente devem ser obedecidas, pois, caso contrário, são produzidas soluções inviáveis. Como os métodos de Lavallée e Hidiroglou (1988) e de Kozak (2004) não foram concebidos para atender à restrição 4.2, algumas adaptações são necessárias.

Como alternativa a esses métodos, propõe-se aqui um novo método capaz de atender a todas as restrições, utilizando a alocação proposta por Brito et al. (2015) que garante o ótimo global para o problema de alocação (segundo nível) dentro de um tempo computacional factível.

Todavia, até o momento ainda não existe um método para encontrar o ótimo global desse problema de minimização no primeiro nível, a não ser um método exaustivo, que considera todas as soluções possíveis. Entretanto, calcular todas as combinações possíveis é inviável para populações de tamanho grande. Assim sendo, para problemas de maior porte, deve-se procurar utilizar algoritmos baseados em metaheurísticas, o que justifica a proposta desse trabalho de dissertação.

A ideia central do algoritmo proposto é baseada no processo de discretização apresentado por Brito et al. (2006). Nesse processo, considere o vetor populacional $X_U = \{x_1, x_2, ..., x_N\}$ 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 (vi) do procedimento de estratificação (seção 2.3), são necessários $(L-1)$ pontos de corte para definir os L 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 ($\mathbf{b}^* = \{b_1^*, b_2^*, b_3^*\}$) para populações com pequenos valores para w, pois é possível testar todas as combinações, uma vez que o problema se resume a uma combinação de w em $L-1$, ou seja, $\binom{w}{L-1}$. Por exemplo, para a população 354542 da PAC, em que $w = 6$ e $L - 1 = 3$, há 20 combinações possíveis de soluções para serem testadas.

A Tabela 4.3 mostra as soluções ótimas globais para a população citada no parágrafo anterior. Só é possível garantir que essas soluções correspondem a ótimos globais, porque

a enumeração exaustiva e a alocação de Brito et al. (2015) foram utilizadas. Portanto, existem três soluções ótimas ($\mathbf{b}^* = \{0, 1, 4\}$; $\mathbf{b}^* = \{1, 2, 4\}$ e $\mathbf{b}^* = \{1, 3, 4\}$) associadas ao tamanho amostral mínimo ($n^* = 17$), e todas as outras dezessete combinações de soluções possíveis produzirão ou $n > 17$ ou soluções inviáveis, soluções estas que não atendem a alguma restrição. O procedimento geral do que foi feito nesse exemplo, está representado no Algoritmo EXATO 4.3 que será apresentado duas páginas a frente.

Tabela 4.3: Mínimo Global da população 354542 da PAC

Código do Estrato Natural	b_1^*	b_2^*	b_3^*	Tamanho amostral*
354542	0	1	4	17
	1	2	4	
	1	3	4	

A partir da ideia de enumeração exaustiva exemplificada 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 Reconexão por Caminhos (PR). Em alusão a esses três procedimentos citados (Exato, VNDS, PR) foi dado o nome de **Algoritmo EVP**, de acordo com a primeira letra de cada palavra. A estrutura resumida está apresentada no Algoritmo 4.1: caso o problema seja considerado pequeno, obtém-se a solução ótima global a partir do procedimento de enumeração exaustiva; caso contrário, o algoritmo tentará produzir a melhor solução baseando-se nas metaheurísticas VNDS e PR, até atingir um critério de parada.

Algoritmo 4.1 Estrutura Resumida do Algoritmo EVP

```
Se Problema Pequeno Então
    EXATO;
Senão
    Enquanto Não Satisfizer um Critério de Parada Faça
        VNDS;
        PR;
    Fim Enquanto
Fim Se
```

Vale ressaltar que o tamanho do problema está relacionado ao tamanho populacional (N) e ao número de combinações de w em $L-1$, $\binom{w}{L-1}$, lembrando que w é a quantidade de elementos únicos do conjunto Q. Portanto, a partir de experimentos prévios, definiu-se que um problema é considerado pequeno quando atende, simultaneamente, a dois critérios: $N \leq 4.000$ e o valor de $\binom{w}{L-1} \leq 1.000.000$.

Tendo como referência a estrutura citada, apresenta-se o algoritmo EVP completo

(4.2), disponível no Apêndice B e implementado em linguagem R, por ser um software de ampla utilização entre os estatísticos. Como dados de entrada utiliza-se o vetor populacional (X_U) e os parâmetros iniciais de estratificação, a saber: o número de estratos (L); o coeficiente de variação fixado (τ), ver restrição 2.25; o tamanho mínimo amostral por estrato (n_{min}); e o tamanho mínimo populacional por estrato (N_{min}). Além desses, há os parâmetros adicionais necessários para a execução do algoritmo, a saber: número de vizinhos máximo (t_{max}); a amplitude da perturbação (r_1); a amplitude da busca local (r_2); tamanho máximo de soluções elite (p_{max}); e os critérios de parada – o número máximo de iterações (i_{max}), número de iterações sem redução do valor da função objetivo ($notBest$) e tempo máximo de processamento ($cpuTime$).

Algoritmo 4.2 Algoritmo EVP Completo

Entrada: X_U; Parâmetros Iniciais: L, τ, n_{min}, N_{min}; Parâmetros adicionais: t_{max},r_1, r_2, p_{max}, $notBest$, i_{max}, $cpuTime$

```
Q=deduplica(X);
Se Problema Pequeno Então
    b* = Algoritmo EXATO (4.3);
Senão
    Defina o conjunto de estruturas de vizinhanças η_k (k = 1, ..., k_max) que serão usadas na busca
    b = b_0;                    < Encontre uma solução inicial >
    i = 1;
    Enquanto Não Satisfizer um Critério de Parada Faça
        k = 1;
        Enquanto k ≤ k_max Faça
            d = binário(k, k_max);   <Vetor binário aleatório de tamanho k_max, com k elementos iguais a 1>
            b_1 = PERTURBAÇÃO(b, k, d);   <b_1 difere de b por k atributos, de acordo com d >
            b_2 = BUSCA_LOCAL(b_1, k, d);   <Busca somente nos k atributos, de acordo com d >
            Se f(b_2) ≤ f(b) Então
                b = b_2;
                k = 1;
            Senão
                k = k + 1
            Fim Se
        Fim Enquanto
        Atualiza pool com b
        Se i ∈ {5, 25, 45, 65, ..., i_max} Então
            b_3=PR_ADAPTADO (pool)
            Se f(b_3) ≤ f(b) Então
                b = b_3
            Fim Se
        Fim Se
        i = i + 1
    Fim Enquanto
Fim Se
Saída: Melhor Solução b
```

Na linha 1 do algoritmo, define-se o conjunto Q a partir da retirada das duplicações de X_U. E já na linha 2 é feito o teste acerca do tamanho do problema. Portanto, se o problema for considerado pequeno, obtém-se a solução ótima global a partir da aplicação do Algoritmo EXATO 4.3. Caso contrário, utiliza-se o algoritmo proposto que é executado a partir da linha 5 na tentativa de produzir a melhor solução baseada nas metaheurísticas.

O Algoritmo 4.3 inicia enumerando todas as soluções possíveis (C) a partir do conjunto Q. Na linha 2 aplica-se, em cada uma das soluções, o método de alocação ótima de Brito et al. (2015), produzindo assim, todos os tamanhos amostrais (n) associados à respectiva solução. Todos esses resultados são representados por A. Destarte, pode-se afirmar que, pelo menos, um desses resultados é um mínimo global. Na linha 3, a função MÍNIMO possibilita verificar qual dos tamanhos amostrais (n) é o menor dentre todos os resultados. Em caso de empate, mantém-se o primeiro resultado. Como saída obtém-se o vetor solução ótimo $\mathbf{b}^* = \{b_1^*, b_2^*, \ldots, b_{L-1}^*\}$, associado ao menor tamanho amostral produzido.

Algoritmo 4.3 Pseudo Código do Algoritmo EXATO

Entrada: X_U e Parâmetros: L, τ, n_{min}, N_{min}

1 $C = \binom{w}{L-1}$;
2 $A =$ ALOCAÇÃO_OTIMA(C);
3 $\mathbf{b}^*$ = Solução associada ao MÍNIMO$\{A\}$
4 **Saída:** $\mathbf{b}^*$

A partir da linha 5, o Algoritmo proposto 4.2 produzirá a melhor solução viável tendo como base as metaheurísticas VNDS (linhas 5 a 20) e PR (linhas 21 a 27). A partir desse ponto, as metaheurísticas utilizam os outros parâmetros adicionais citados anteriormente.

A metaheurística VNDS, assim como o VNS, apresenta cinco fases, conforme descrito na seção 3.2, a inicialização, a perturbação, a busca local, a substituição e o critério de parada. Na fase de inicialização (linha 6), define-se um vetor $\mathbf{b}_0$ viável, selecionando aleatoriamente $(L-1)$ valores do conjunto Q para serem os elementos $b_1, b_2, \ldots, b_{L-1}$. E 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 de vizinhança 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 $j = 1, 2, \ldots, w$.

No VNDS a decomposição consiste em fixar todos os atributos, exceto por k atributos. Por isso, deve-se especificar a definição desse parâmetro. Assim, no método proposto, k corresponde à quantidade de elementos do vetor **b** que não serão fixos. Esses elementos livres serão modificados nas fases da perturbação e da busca local. Por exemplo, se $k = 1$

então tem-se $(L-2)$ elementos do vetor fixos e apenas um elemento livre. Portanto, o valor de $k_{max} = L-1$ (é o próprio tamanho do vetor **b**), assim, quando $k = L-1$ todos os elementos são livres.

Para ilustrar o método, 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 $Q = \{1,2,3,4,5,7,8,10,15,31\}$. Supondo L = 4 e conforme o processo de inicialização descrito acima, gera-se um vetor aleatório inicial $\mathbf{b} = \mathbf{b}_0 = \{3,7,10\}$. Com isso, baseando-se no procedimento de estratificação por cortes representado pelas equações 2.15, 2.16, 2.17, 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 $N_4 = 2$.

Na linha 11 do Algoritmo 4.2, define-se um vetor binário **d** de tamanho k_{max} com k elementos iguais a 1 e os demais elementos com valor 0. Esse vetor **d** serve para indicar quais são os elementos livres do vetor **b** que serão modificados nas fases de perturbação e de busca local.

Na fase de perturbação (linha 12) gera-se um novo vetor solução $\mathbf{b}_1$ que difere de **b** por k elementos. Continuando o exemplo acima, em que $\mathbf{b} = \{3,7,10\}$ ou $\mathbf{b} = \{q_3, q_6, q_8\}$ e para $k = 1$, apenas um elemento deve ser modificado. Então, sorteia-se aleatoriamente qual dos três elementos será o escolhido. Por exemplo, para $\mathbf{d} = \{0,1,0\}$ será o elemento b_2, tal que o novo vetor $\mathbf{b}_1$ será $\{3, b_2', 10\}$. Em que b_2' segue a 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 4.3, o novo valor de b_2' será um elemento qualquer do conjunto $\{4,5,8,10\}$. Sendo produzido, por exemplo, um novo vetor solução $\mathbf{b}_1 = \{3,4,10\}$ ou $\mathbf{b}_1 = \{q_3, q_4, q_8\}$.

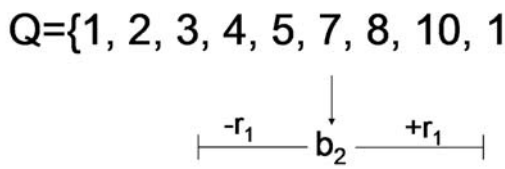

Figura 4.3: Exemplo de Estrutura de Vizinhança para o Algoritmo EVP

A função PERTURBAÇÃO (Algoritmo 4.4) define a forma geral do que foi apresentado nesse exemplo. Considerando o exemplo anterior, em que $r_1 = 2$, o conjunto R é dado por $R = \{-2,-1,1,2\}$, então é necessário escolher um elemento de R para aplicar a perturbação e produzir uma nova solução, o que ocorre na linha 3 representado pelo

valor R_1. Por exemplo, o valor poderia ser $R_1 = -1$ indicando que a perturbação nos elementos livres será na mesma direção. Sendo o vetor acima $\mathbf{b} = \{q_3, q_6, q_8\} = \{3, 7, 10\}$ e para $k = 2$, com elementos livres b_2 e b_3 ($\mathbf{d} = \{0, 1, 1\}$), o novo vetor solução seria $\mathbf{b}_1 = \{q_3, q_{6+R_1}, q_{8+R_1}\} = \{q_3, q_{6-1}, q_{8-1}\} = \{q_3, q_5, q_7\} = \{3, 5, 8\}$.

Algoritmo 4.4 Pseudo Código da função PERTURBAÇÃO

Entrada: Q, **b** e Parâmetros: k, r_1, **d**

1 q_j = Índices(**b** $\in Q$, **d**); $<$ Obtém os índices de **b** no conjunto Q, para os elementos livres ($d_j = 1$) $>$

2 $R = \{-r_1, \ldots, -1, 1, \ldots, +r_1\}$; $<$ Conjunto de valores variando de $-r_1$ até $+r_1$ (sem o zero)$>$

3 R_1=Sorteia(R); $<$Sorteia aleatoriamente um elemento de R$>$

4 $q_j = q_j + R_1$; $<$ Os índices q_j são atualizados com o valor de R_1 $>$

5 $\mathbf{b}_1 = \mathbf{b} + q_j$ $<$ A solução $\mathbf{b}_1$ é criada com os índices de q_j nas k posições de $d_j = 1$ e nas outras posições com o vetor **b** $>$

6 **Saída:** $\mathbf{b}_1$

Na fase de busca local (linha 13 do Algoritmo 4.2), a partir do vetor $\mathbf{b}_1$, produz-se a melhor solução dentre os t_{max} vizinhos da vizinhança, solução essa chamada de $\mathbf{b}_2$. Assim como na perturbação, a estrutura de vizinhança é a mesma, porém utiliza r_2 como parâmetro para a amplitude da busca local, em que r_2 é independente de r_1, podendo ser igual ou diferente deste. Continuando o exemplo anterior, tinha-se para $k = 1$ e como elemento livre b_2, o vetor solução era $\mathbf{b}_1 = \{3, 4, 10\}$. Então, para $t_{max} = 2$ e $r_2 = 3$, serão testados dois vizinhos para b_2', lembrando que agora, b_2' deverá pertencer ao conjunto $\{1, 2, 3, 5, 7, 8\}$. Gerando assim duas possíveis soluções para $\mathbf{b}_2$, tais como, $\mathbf{b}_{21} = \{3, 5, 10\}$ e $\mathbf{b}_{22} = \{3, 8, 10\}$. Então, calcula-se $f(\mathbf{b}_{21})$ e $f(\mathbf{b}_{22})$, aplicando a alocação ótima de Brito et al. (2015). O menor valor de n encontrado entre eles determinará o vetor $\mathbf{b}_2$ que será testado na fase de substituição.

A função BUSCA_LOCAL (Algoritmo 4.5) define a forma geral do que foi apresentado nesse exemplo, só diferindo do algoritmo 4.4 por considerar o parâmetro r_2, por repetir a busca em t_{max} vizinhos (mantendo apenas a melhor solução) e por permitir que a busca nos elementos livres não seja na mesma direção. Considerando o exemplo anterior, em que $r_2 = 3$, o conjunto R é formado por $R = \{-3, -2, -1, 1, 2, 3\}$. Para $k = 2$, precisa-se escolher um subconjunto de R de tamanho $k = 2$, para produzir uma nova solução, o que ocorre na linha 4 representado pelo conjunto $R_k \subseteq R$. Por exemplo, o conjunto R_k poderia ser $\{-1, 2\}$ indicando que o movimento dos elementos livres não são na mesma direção. Considerando o vetor anterior $\mathbf{b}_1 = \{q_3, q_5, q_7\} = \{3, 5, 8\}$ e com elementos livres b_2 e b_3 ($\mathbf{d} = \{0, 1, 1\}$), conforme linhas 5 e 6, um novo vetor solução seria $\mathbf{b}_{2t} = \{q_3, q_{5+R_{k1}}, q_{7+R_{k2}}\} = \{q_3, q_{5-1}, q_{7+2}\} = \{q_3, q_4, q_9\} = \{3, 4, 15\}$. Esse procedimento é repetido t_{max} vezes e depois calcula-se o tamanho amostral para todas as soluções ($f(\mathbf{b}_{2t})$), sendo a saída do

algoritmo, a solução associada ao menor tamanho amostral produzido.

Algoritmo 4.5 Pseudo Código da função BUSCA_LOCAL

Entrada: Q, $\mathbf{b}_1$ e Parâmetros: k, r_2, t_{max}, **d**

1 **Para** $t = 1$ *até* t_{max} **Faça**

2 q_j = Índices($\mathbf{b}_1 \in Q$, **d**); $<$ Obtém os índices de $\mathbf{b}_1$ no conjunto Q, para os elementos livres $(d_j = 1) >$

3 $R = \{-r_2, \ldots, -1, 1, \ldots, +r_2\}$; $<$ Conjunto de valores variando de $-r_2$ até $+r_2$ (sem o zero)$>$

4 R_k=Sorteia(k, R); $<$Sorteia k elementos de R$>$

5 $q_j = q_j + R_k$; $<$ Os índices q_j são atualizados com os valores $R_k >$

6 $\mathbf{b}_{2t} = \mathbf{b}_1 + q_j <$ A solução $\mathbf{b}_{2t}$ é criada com os índices de q_j nas k posições de $d_j = 1$ e nas outras posições com o vetor $\mathbf{b}_1 >$

7 **Fim Para**

8 $\mathbf{b}_2$ = Solução associada ao MÍNIMO$\{f(\mathbf{b}_{2t})\}$;

9 **Saída:** $\mathbf{b}_2$

A fase da substituição representada pelas linhas 14 até 19 do algoritmo 4.2 significa que se a solução produzida $\mathbf{b}_2$ determina um tamanho de amostra menor ou igual ao tamanho amostral associado à solução **b** corrente, substitui-se o vetor **b** e faz-se $k = 1$ novamente. Caso contrário, faz-se $k = k + 1$, até atingir k_{max}. Por permitir a substituição da solução corrente por outra solução equivalente, isso dá ao algoritmo a possibilidade de explorar outras vizinhanças, pois a solução inicial na próxima iteração será diferente.

A partir da linha 21 o algoritmo EVP segue o conceito da metaheurística de Reconexão por Caminhos. Especificamente, na linha 21, cria-se um conjunto elite das melhores soluções obtidas, também chamado de *pool*, que começa vazio e é limitado de tamanho, conforme ideia de Glover em Gendreau e Potvin (2010). Portanto, o pool começa vazio e vai sendo preenchido com a solução corrente **b**, até atingir seu tamanho máximo (p_{max}). Após o conjunto elite atingir seu tamanho máximo, a solução corrente **b** somente será adicionada se for melhor que alguma das soluções já existentes no pool, sendo a pior solução descartada do conjunto, aquela que produz o maior tamanho amostral.

Na linha 23 do Algoritmo 4.2 aplica-se a função PR_ADAPTADO (Algoritmo 4.6) para produzir uma solução $\mathbf{b}_3$. A função considera todas as combinações das p_{max} soluções do pool tomadas duas a duas, ou seja, são feitas ao todo $\binom{p_{max}}{2}$ combinações, em que trata-se uma solução como a inicial(x_i) e a outra como a final(x_t). Então aplica-se o Algoritmo 3.3 da Reconexão por Caminhos, considerando a estratégia de Reconexão para Frente e a Distância de Manhattan, em que a melhor solução intermediária é mantida como resposta do algoritmo. A função PR_ADAPTADO garante uma busca quase exaustiva entre as soluções elite permitindo, assim, explorar de forma intensiva a vizinhança dessas soluções, o que re-

presenta um custo computacional alto. Por isso, para evitar que o algoritmo EVP consuma um tempo de processamento muito elevado, decidiu-se aplicá-la apenas em algumas iterações do algoritmo 4.2, conforme linha 22.

Algoritmo 4.6 Pseudo Código da função PR_ADAPTADO

Entrada: pool, p_{max}

1 $x_i \in$ pool e $x_t \in$ pool, tal que: $x_i \neq x_t$; $\quad i = 1, \ldots, (p_{max} - 1)$; $\quad t = 2, \ldots, p_{max}$.
2 **Enquanto** $Dist(x_i, x_t) \neq 0$ **Faça**
3 Encontre o melhor movimento m que diminui $Dist(x \oplus m, t)$;
4 $x = x \oplus m$; $\quad <$ Aplique o movimento m na solução $x >$;
5 $x_f =$ Melhor Solução encontrada na trajetória entre x_i e x_t
6 **Fim Enquanto**
7 **Saída:** x_f

Para ilustrar e tendo como referência o procedimento de Reconexão por Caminhos apresentado em Brito et al. (2010), suponha como exemplo, um pool com $p_{max} = 3$ e o problema de estratificação para $L = 4$. Portanto, pode-se ter as seguintes soluções $\mathbf{b}_1 = \{1, 4, 8\}$, $\mathbf{b}_2 = \{2, 7, 10\}$ e $\mathbf{b}_3 = \{3, 8, 10\}$. Nesse caso, o algoritmo irá testar as três combinações possíveis: ($x_i = \mathbf{b}_1, x_t = \mathbf{b}_2$), ($x_i = \mathbf{b}_1, x_t = \mathbf{b}_3$) e ($x_i = \mathbf{b}_2, x_t = \mathbf{b}_3$). Tomando como exemplo apenas a primeira combinação ($x_i = \mathbf{b}_1, x_t = \mathbf{b}_2$), os movimentos intermediários estão apresentados na Tabela 4.4, essas soluções intermediárias (x_f) são então avaliadas mediante o cálculo da função objetivo através da aplicação da alocação de Brito et al. (2015), sendo mantida a solução de melhor qualidade. Observe nessa tabela que a trajetória da Reconexão por Caminhos é realizada de forma que a distância de Manhattan entre a solução inicial(x_i) e a solução intermediária(x_f) é reduzida em uma unidade até que a solução alvo(x_t) seja atingida. De acordo com o Algoritmo 4.6, ainda seriam calculadas as soluções intermediárias para a segunda combinação ($x_i = \mathbf{b}_1, x_t = \mathbf{b}_3$) e para a terceira combinação ($x_i = \mathbf{b}_2, x_t = \mathbf{b}_3$). A melhor de todas as soluções intermediárias, será considerada o vetor $\mathbf{b}_3$.

Tabela 4.4: Exemplo dos movimentos executados pelo procedimento de Reconexão por Caminhos

x_i	1	4	8
x_f	2	4	8
x_f	2	5	8
x_f	2	6	8
x_f	2	7	8
x_f	2	7	9
x_t	2	7	10

Retornando ao algoritmo EVP (4.2), as linhas 24 até 26 representam que se a solução $\mathbf{b}_3$ determina um tamanho de amostra menor ou igual ao tamanho amostral associado à solução **b** corrente, substitui-se o vetor **b**. Enquanto o critério de parada não é satisfeito, novas iterações ($i = i + 1$) são executadas. O algoritmo 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 ($cpuTime$).

4.3 - Métodos da Literatura Utilizados para Comparação

De forma a avaliar os resultados produzidos pelo algoritmo EVP proposto, foram utilizados para comparação os dois métodos que têm apresentado os melhores resultados na literatura, descritos na seção 2.4: Kozak (2004) e Lavallée e Hidiroglou (1988). Contudo, eles apresentam uma limitação, pois não foram concebidos para contemplar restrições adicionais, como a restrição 4.2 e, por isso, para as 75 populações da PAC, algumas adaptações foram necessárias.

O método de Kozak (2004) está implementado em linguagem R na função *strata.LH* disponível no pacote *stratification*. Por não ter o seu código fonte disponível, optou-se nas populações da PAC por ignorar a restrição 4.2 para, assim, utilizar o algoritmo disponível apenas para gerar o vetor **b**. E a partir desses resultados, calcular novos tamanhos amostrais que atendam a essa restrição, o que pode ocasionar CV maior que 10%. Vale lembrar que Kozak, alterou a restrição 2.22 ($N_h \geq 1$) para 2.35 ($N_h \geq 2$), com o intuito de evitar a criação de estratos com apenas uma unidade populacional. Para possibilitar a comparação entre os métodos, essa restrição (2.35) foi mantida para os demais métodos.

Já o método de Lavallée e Hidiroglou (1988), no caso das populações da PAC, não precisou de ajuste, pois foi implementado em código aberto em linguagem SAS em Azevedo (2004), o que permitiu a inclusão da restrição 4.2, sendo utilizado para o parâmetro p a raiz quadrada $(1/2)$. Para as populações da literatura, utilizou-se a função *strata.LH* disponível no pacote *stratification*, utilizando o algoritmo de Sethi (1963) e o mesmo parâmetro ($p = 1/2$) para a alocação potência.

Em Kozak e Verma (2006) foi apresentada uma medida de eficiência relativa para comparar métodos que minimizam o tamanho amostral dado que a precisão é fixa. Nesse trabalho de dissertação, essa medida é dada por

$$EFF_{i,EVP} = \frac{n_i}{n_{EVP}} \cdot 100, \tag{4.3}$$

em que i representa os métodos de Kozak (2004) ou de Lavallée e Hidiroglou (1988), n_i representa o tamanho amostral produzido por esses métodos e n_{EVP} representa o tamanho amostral produzido pelo método proposto. Portanto, valores superiores a 100 indicam que o método proposto apresentou uma solução melhor, enquanto valores inferiores a 100 indicam que o método proposto apresentou uma solução pior e por fim, valores iguais a 100 indicam que os métodos se equivalem.

5 - Resultados Computacionais

5.1 - Ambiente Computacional

Os três algoritmos (citados nas seções 4.2 e 4.3) foram executados em um computador com 6GB de memória RAM, com processador i7 de 2.2GHz com 4 núcleos e sistema operacional Windows de 64 bits. As abreviações LH88, KOZAK04, EVP referem-se, respectivamente, aos algoritmos de Lavallée e Hidiroglou (1988), Kozak (2004) e ao novo algoritmo EVP proposto.

Para cada uma das 100 populações citadas na seção 4.1 (25 populações da literatura e 75 populações associadas à base de dados originada no CBS da PAC 2014), foram calculados novos limites para os pontos de corte, visando a minimização do tamanho amostral total, a partir da aplicação dos três métodos supracitados. Foram testados, ainda, os resultados para o número de estratos (L) variando de 3 a 6, conforme feito em outros trabalhos da literatura, tais como Gunning e Horgan (2004); Er (2011); Brito, Semaan e Brito (2014), e também por recomendação de Cochran (1977) que diz que um número maior de estratos traria pouco ganho, em relação aos custos adicionais que viriam junto. Portanto, no total serão produzidos 1200 resultados (100 populações x 3 métodos x 4 números de estratos).

Os parâmetros do problema de estratificação das populações da literatura, para os três métodos, foram: CV $\leq 10\%$, $N_h \geq 2$, $n_L = N_L$ e $n_h \geq 1$. Para as populações da PAC, foram utilizados os mesmos parâmetros, exceto pela substituição da última restrição por 4.2, ou mais especificamente, por $n_h \geq 5$.

Para o algoritmo EVP proposto nesse trabalho de dissertação, além dos parâmetros de estratificação acima, outros parâmetros adicionais também são necessários. Após testes preliminares com algumas combinações de valores para esses parâmetros, optou-se por utilizar os seguinte 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 $cpuTime = 5000$ segundos (aproximadamente 1 hora e 23 minutos). Além disso, o algoritmo exato 4.3 foi executado para as populações que atendam ao seguinte critério: tamanho da população $N \leq 4.000$ e combinação de soluções possíveis $\binom{w}{L-1} \leq 1.000.000$, conforme explicitado na seção 4.2. O método de alocação proposto por Brito et al. (2015) está implementado em linguagem R disponível no pacote *MultAlloc* (http://cran.r-project.org/web/packages/MultAlloc/index.html).

5.2 - Populações da Literatura

Considerando inicialmente $L = 3$, os resultados obtidos para os tamanhos amostrais totais ($n = n_1 + n_2 + n_3$), para o coeficiente de variação estimado (CV) medido em percentual e para o tempo de processamento medido em segundos, estão apresentados na Tabela 5.1. Note que na população U25, os dois algoritmos da literatura não puderam ser aplicados, pois essa população apresenta valores negativos, e portanto, apenas o algoritmo EVP foi capaz de produzir resultados.

Doravante, denomina-se por **solução vencedora**, a solução associada ao menor tamanho amostral de cada população produzida por um dos três algoritmos, que atenda a todas as restrições. Essas soluções estão nas células sombreadas em cinza. Por exemplo, na população U25 a solução vencedora veio do algoritmo EVP proposto. Adicionalmente, note que o empate é permitido, por exemplo, na população U01, todos os três algoritmos produziram a solução vencedora.

Considerando as soluções vencedoras, observa-se que o algoritmo EVP teve um desempenho levemente superior ao de Kozak (2004). Das 25 populações consideradas no estudo, o algoritmo EVP produziu a solução vencedora em todas. Enquanto o algoritmo de Kozak (2004) produziu a solução vencedora em 24 populações e o algoritmo de Lavallée e Hidiroglou (1988) apenas em 9 populações. Contudo, nos casos de empate, o algoritmo de Kozak (2004) tende a produzir coeficientes de variação estimados inferiores aos produzidos pelo algoritmo EVP.

Entre as soluções vencedoras produzidas pelo algoritmo EVP, ainda há aquelas que foram produzidas pelo algoritmo exato (4.3), marcadas com um asterisco na coluna do tamanho amostral, portanto, são soluções ótimas globais. Assim, a partir de agora, a solução produzida pelo Algoritmo 4.3 será denominada apenas por **solução ótima**. Portanto, conforme o esperado, considerando a definição do critério de problema pequeno, há 20 soluções ótimas na Tabela 5.1, o que representa 80% dos resultados para o caso em que $L = 3$.

Tabela 5.1: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento Produzidos por Algoritmo para as 25 Populações da Literatura ($L = 3$)

ID	LH88			KOZAK04			EVP		
	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
U01	29	9,7	≈ 1	29	9,7	≈ 1	29*	10,0	157
U02	4	6,0	≈ 1	4	6,0	≈ 1	4*	10,0	857
U03	38	9,8	≈ 1	37	9,9	≈ 1	37*	10,0	970
U04	16	9,5	≈ 1	15	9,8	≈ 1	15*	10,0	911
U05	61	9,8	≈ 1	60	9,9	≈ 1	60*	10,0	3.626
U06	24	9,6	≈ 1	23	9,9	≈ 1	23	9,9	28
U07	24	9,6	≈ 1	24	9,5	≈ 1	24*	10,0	139
U08	8	9,8	≈ 1	8	9,8	≈ 1	8*	9,8	8
U09	21	9,8	≈ 1	21	9,8	≈ 1	21	10,0	35
U10	32	9,7	≈ 1	31	9,9	≈ 1	31*	10,0	589
U11	43	9,8	≈ 1	42	9,9	≈ 1	42	10,0	57
U12	18	9,4	≈ 1	17	9,7	≈ 1	17*	10,0	30
U13	33	9,7	≈ 1	32	9,9	≈ 1	32	10,0	22
U14	4	3,9	≈ 1	4	3,9	≈ 1	4*	9,3	851
U15	18	9,6	≈ 1	18	9,6	≈ 1	18*	10,0	3
U16	23	9,5	≈ 1	22	9,8	≈ 1	22*	10,0	129
U17	17	9,6	≈ 1	16	10,0	≈ 1	16*	10,0	34
U18	11	9,5	≈ 1	11	9,5	≈ 1	11*	10,0	5
U19	32	9,7	≈ 1	31	9,9	≈ 1	31*	10,0	1.884
U20	26	9,7	≈ 1	26	9,7	≈ 1	26*	10,0	233
U21	16	9,4	≈ 1	15	9,8	≈ 1	15*	10,0	19
U22	19	9,6	≈ 1	18	9,9	≈ 1	18*	10,0	14
U23	25	9,6	≈ 1	24	9,8	≈ 1	24*	10,0	248
U24	6	7,8	≈ 1	5	9,2	≈ 1	5*	10,0	874
U25	N/A	N/A	N/A	N/A	N/A	N/A	51	10,0	86

N/A: Não se Aplica

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

Em relação ao tempo de processamento, a diferença é relevante, enquanto os algoritmos LH88 e KOZAK04 levaram menos de um segundo para cada população, os tempos do algoritmo EVP variaram de 3 segundos a 3626 segundos (aproximadamente 1 hora). Entretanto, tal comportamento já era esperado, uma vez que o algoritmo EVP resulta de um processo de computação mais intensivo, que permite explorar uma quantidade bem maior de soluções, ou ainda produzir soluções ótimas a partir da enumeração exaustiva.

Para os demais números de estratos ($L = 4, 5, 6$), os resultados estão nas Tabelas 5.2, 5.3 e 5.4, respectivamente – essas tabelas têm a mesma estrutura da Tabela 5.1. Observe que nessas tabelas há resultados em negrito sublinhado para o método de Lavallée e Hidiroglou (1988), pois para essas populações o método não convergiu e com isso a restrição

$N_h \geq 2$ não foi atendida, sendo portanto, soluções inviáveis. E como era de se esperar, em função do critério para aplicação da resolução exata, o número de soluções ótimas produzidas pelo algoritmo EVP foi caindo conforme o valor de L foi aumentando: quatro soluções ótimas para $L = 4$, duas soluções ótimas para $L = 5$ e nenhuma solução ótima para $L = 6$.

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 nas Tabelas 5.2 e 5.3, e também, a população U01 da Tabela 5.4.

Tabela 5.2: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento Produzidos por Algoritmo para as 25 Populações da Literatura ($L = 4$)

ID	LH88			KOZAK04			EVP		
	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
U01	21	9,3	≈ 1	20	9,6	≈ 1	20	10,0	3
U02	5	3,5	≈ 1	5	3,4	≈ 1	5	8,0	2
U03	21	9,4	≈ 1	19	9,9	≈ 1	19	10,0	26
U04	10	9,2	≈ 1	9	9,7	≈ 1	10	9,8	10
U05	33	9,8	≈ 1	32	9,9	≈ 1	33	10,0	49
U06	14	9,1	≈ 1	13	9,5	≈ 1	12	10,0	229
U07	14	9,5	≈ 1	14	9,5	≈ 1	14	9,9	5
U08	7	7,7	≈ 1	5	9,7	≈ 1	5*	9,9	122
U09	12	9,3	≈ 1	11	9,8	≈ 1	13	9,7	52
U10	17	9,4	≈ 1	16	9,7	≈ 1	16	10,0	17
U11	21	9,7	≈ 1	20	9,9	≈ 1	22	10,0	172
U12	10	8,9	≈ 1	9	9,6	≈ 1	9	9,9	3
U13	18	9,5	≈ 1	17	9,8	≈ 1	17	9,9	30
U14	5	2,2	≈ 1	5	2,2	≈ 1	5	3,3	5
U15	11	8,9	≈ 1	10	9,3	≈ 1	10*	10,0	45
U16	**31**	9,1	≈ 1	15	9,8	≈ 1	15	9,9	6
U17	11	8,7	≈ 1	10	9,4	≈ 1	10	9,7	2
U18	8	7,9	≈ 1	7	8,8	≈ 1	7*	10,0	171
U19	17	9,5	≈ 1	16	9,8	≈ 1	16	9,8	34
U20	17	9,0	≈ 1	15	9,7	≈ 1	15	9,9	5
U21	10	8,8	≈ 1	9	9,4	≈ 1	9	9,9	2
U22	12	9,0	≈ 1	11	9,5	≈ 1	11*	10,0	551
U23	14	8,9	≈ 1	12	9,7	≈ 1	12	9,9	7
U24	5	6,3	≈ 1	5	6,3	≈ 1	5	6,9	3
U25	N/A	N/A	N/A	N/A	N/A	N/A	33	10,0	129

N/A: Não se Aplica

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

Tabela 5.3: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento Produzidos por Algoritmo para as 25 Populações da Literatura ($L = 5$)

ID	LH88			KOZAK04			EVP		
	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
U01	**26**	9,7	≈ 1	13	9,5	≈ 1	14	10,0	12
U02	6	2,2	≈ 1	6	2,2	≈ 1	6	4,3	4
U03	12	9,6	≈ 1	12	9,5	≈ 1	12	9,9	21
U04	8	8,3	≈ 1	6	9,8	≈ 1	6	9,9	8
U05	**71**	9,8	≈ 1	20	9,7	≈ 1	20	10,0	73
U06	10	8,6	≈ 1	9	9,1	≈ 1	8	10,0	351
U07	10	9,1	≈ 1	10	9,2	≈ 1	10	10,0	6
U08	6	7,0	≈ 1	6	6,5	≈ 1	6*	9,8	1.264
U09	9	8,6	≈ 1	8	9,3	≈ 1	10	10,0	33
U10	12	8,6	≈ 1	10	9,6	≈ 1	10	9,9	13
U11	14	9,2	≈ 1	13	9,6	≈ 1	13	9,7	165
U12	7	9,2	≈ 1	7	9,0	≈ 1	7	9,3	4
U13	12	9,2	≈ 1	11	9,6	≈ 1	12	9,9	25
U14	6	1,5	≈ 1	6	1,5	≈ 1	6	3,8	3
U15	9	8,2	≈ 1	7	9,5	≈ 1	7*	9,9	845
U16	**29**	9,5	≈ 1	10	9,2	≈ 1	10	10,0	4
U17	8	8,7	≈ 1	7	9,3	≈ 1	7	9,9	4
U18	6	7,6	≈ 1	6	7,7	≈ 1	6	9,0	2
U19	12	8,8	≈ 1	11	9,3	≈ 1	11	9,9	25
U20	11	8,7	≈ 1	10	9,1	≈ 1	10	9,9	11
U21	6	9,2	≈ 1	6	9,2	≈ 1	6	9,6	4
U22	**13**	9,3	≈ 1	7	9,5	≈ 1	7	9,9	4
U23	10	8,7	≈ 1	8	9,7	≈ 1	8	9,9	4
U24	6	4,2	≈ 1	6	4,2	≈ 1	6	8,4	7
U25	N/A	N/A	N/A	N/A	N/A	N/A	25	10,0	115

N/A: Não se Aplica

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

Tabela 5.4: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento Produzidos por Algoritmo para as 25 Populações da Literatura ($L = 6$)

ID	LH88			KOZAK04			EVP		
	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
U01	**25**	9,0	≈ 1	10	8,8	≈ 1	9	9,8	21
U02	7	1,6	≈ 1	7	1,6	≈ 1	7	2,7	3
U03	10	8,7	≈ 1	9	9,0	≈ 1	10	9,9	15
U04	7	7,1	≈ 1	7	7,1	≈ 1	7	8,1	11
U05	**65**	9,9	≈ 1	13	9,7	≈ 1	15	10,0	32
U06	9	7,7	≈ 1	7	8,7	≈ 1	7	9,7	63
U07	**20**	8,2	≈ 1	7	9,2	≈ 1	8	9,4	15
U08	7	5,1	≈ 1	7	4,7	≈ 1	7	9,2	9
U09	7	8,2	≈ 1	7	8,1	≈ 1	8	9,9	87
U10	9	8,6	≈ 1	7	9,8	≈ 1	7	10,0	16
U11	**18**	9,4	≈ 1	9	9,9	≈ 1	12	10,0	85
U12	8	6,5	≈ 1	8	6,5	≈ 1	8	8,9	4
U13	10	8,2	≈ 1	8	9,4	≈ 1	8	9,9	16
U14	7	1,1	≈ 1	7	1,1	≈ 1	7	1,8	2
U15	8	6,7	≈ 1	7	8,7	≈ 1	7	9,8	3
U16	**27**	9,2	≈ 1	8	7,9	≈ 1	8	10,0	10
U17	7	7,7	≈ 1	7	7,6	≈ 1	7	8,6	3
U18	7	5,4	≈ 1	7	5,3	≈ 1	7	10,0	2
U19	10	8,2	≈ 1	8	9,2	≈ 1	8	9,5	18
U20	9	8,5	≈ 1	8	9,0	≈ 1	8	9,9	7
U21	**9**	8,0	≈ 1	7	6,8	≈ 1	7	8,3	3
U22	**13**	9,0	≈ 1	7	7,0	≈ 1	7	9,3	4
U23	8	7,9	≈ 1	7	8,8	≈ 1	7	9,3	6
U24	7	3,0	≈ 1	7	3,0	≈ 1	7	4,5	4
U25	N/A	N/A	N/A	N/A	N/A	N/A	23	9,9	51

N/A: Não se Aplica

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

A Tabela 5.5 traz a eficiência relativa dada pela Equação 4.3, lembrando que valores superiores a 100 indicam que o método EVP apresentou uma solução melhor que o método comparado. Na população U25 os algoritmos LH88 e KOZAK04 não são aplicáveis, pois não foram concebidos para solucionar o problema para populações com valores negativos. Em outros casos, o algoritmo LH88 não conseguiu produzir soluções viáveis. Em ambas situações, a célula da tabela é preenchida com “N/A”, o que conta como um resultado favorável para o método proposto. A eficiência relativa para o método de Lavallée e Hidiroglou (1988) variou de 88 a 140, já a eficiência relativa para o método de Kozak (2004) variou de 75 a 113. Especificamente, na comparação entre o método proposto e o método de Lavallée e Hidiroglou (1988), ocorreram 62 resultados favoráveis ao algoritmo EVP, 34 empates

e apenas 4 resultados favoráveis para LH88. Na comparação entre o método proposto e o método de Kozak (2004), ocorreram 7 resultados favoráveis ao algoritmo EVP, 81 empates e 12 resultados favoráveis para KOZAK04, totalizando 100 resultados em cada comparação.

Tabela 5.5: Eficiência Relativa entre os Métodos por Número de Estratos para as 25 Populações da Literatura

ID	$EFF_{LH88/EVP}$				$EFF_{KOZAK04/EVP}$			
	L=3	L=4	L=5	L=6	L=3	L=4	L=5	L=6
U01	100	105	N/A	N/A	100	100	93	111
U02	100	100	100	100	100	100	100	100
U03	103	111	100	100	100	100	100	90
U04	107	100	133	100	100	90	100	100
U05	102	100	N/A	N/A	100	97	100	87
U06	104	117	125	129	100	108	113	100
U07	100	100	100	N/A	100	100	100	88
U08	100	140	100	100	100	100	100	100
U09	100	92	90	88	100	85	80	88
U10	103	106	120	129	100	100	100	100
U11	98	95	108	N/A	100	91	100	75
U12	106	111	100	100	100	100	100	100
U13	100	106	100	125	100	100	92	100
U14	100	100	100	100	100	100	100	100
U15	100	110	129	114	100	100	100	100
U16	105	N/A	N/A	N/A	100	100	100	100
U17	106	110	114	100	100	100	100	100
U18	100	114	100	100	100	100	100	100
U19	103	106	109	125	100	100	100	100
U20	100	113	110	113	100	100	100	100
U21	107	111	100	N/A	100	100	100	100
U22	106	109	N/A	N/A	100	100	100	100
U23	104	117	125	114	100	100	100	100
U24	120	100	100	100	100	100	100	100
U25	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

N/A: Não se Aplica

A Figura 5.1 resume as soluções vencedoras para todos os números de estratos considerados. Dos 100 resultados produzidos (25 populações x 4 números de estratos), o método proposto apresentou a solução vencedora em 88, o método de Kozak (2004) em 93 e o método de Lavallée e Hidiroglou (1988) apenas em 31 resultados. Portanto, percebe-se que o algoritmo EVP venceu quando o números de estratos é três, uma vez que é o único método que consegue produzir uma solução para a população U25, que apresenta valores negativos, e nas demais populações o algoritmo EVP obteve os mesmos resultados do algoritmo de Kozak (2004). Nos demais estratos, outra vez, esses dois algoritmos produziram resultados

quase equivalentes, com leve vantagem para KOZAK04. Embora, mesmo nos casos em que venceu do algoritmo EVP, foi por uma diferença de apenas uma unidade amostral, com raríssimos casos em que a diferença superou esse valor. Por fim, o algoritmo LH88 foi o que apresentou o pior resultado entre os três métodos.

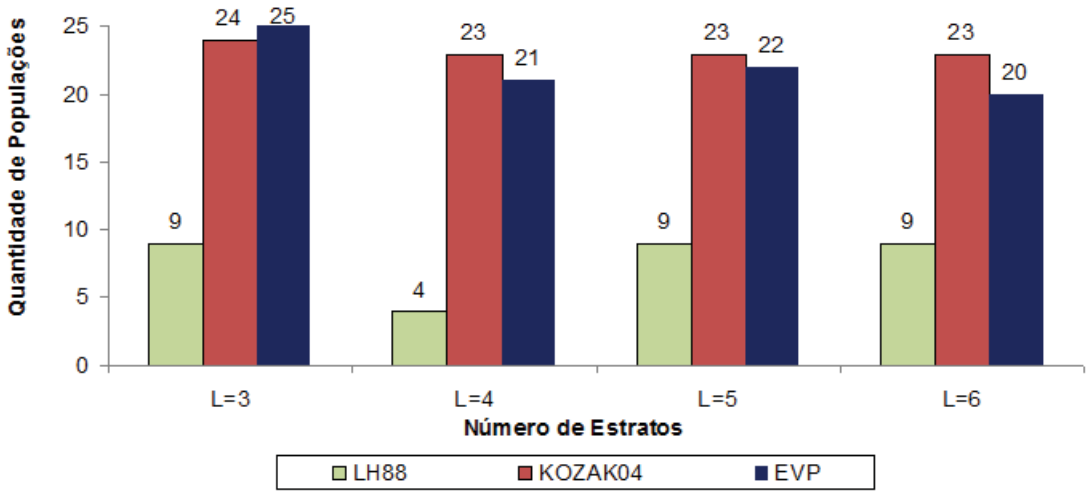

Figura 5.1: Quantidade de Soluções Vencedoras Produzidas por Algoritmo e por Número de Estratos, para as 25 Populações da Literatura

5.3 - Populações da PAC

Considerando inicialmente o caso em que $L = 4$, pois é o número de estratos definido na PAC, serão apresentados, também, os resultados seguindo a metodologia atual da pesquisa descrita na seção 4.1.2.1, denominada de $ATUAL$, apenas para evidenciar o ganho de qualidade ao se utilizar um dos métodos citados. Os resultados obtidos para os tamanhos amostrais totais ($n = n_1 + n_2 + n_3 + n_C$), para o coeficiente de variação estimado medido em percentual e para o tempo de processamento medido em segundos, estão apresentados na Tabela 5.6.

Como citado anteriormente, havia a possibilidade do algoritmo de Kozak (2004) ajustado gerar soluções inviáveis, ou seja, que ultrapassariam o limite de CV de 10%, fato que ocorreu em nove populações. E ainda, no método de Lavallée e Hidiroglou (1988) em quatro populações a restrição $N_h \geq 2$ não foi atendida. Assim, ambas situações representam soluções inviáveis (grifadas de negrito sublinhado) e não podem ser consideradas para encontrar a solução vencedora (células sombreadas em cinza).

Tabela 5.6: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento produzidos por Algoritmo para as 75 populações da PAC ($L = 4$)

Código do	ATUAL		LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	n	CV(%)	t(s)	n	CV(%)	t(s)	n	CV(%)	t(s)
35461	82	9,7	25	8,5	≈ 1	20	9,7	≈ 1	19	9,9	1.150
35462	150	4,9	28	9,7	≈ 1	29	9,4	≈ 1	27*	10,0	842
35466	309	2,9	29	9,5	≈ 1	19	**12,6**	≈ 1	28*	10,0	4.596
35472	1.995	8,6	36	9,8	1,8	33	10,0	5	33	10,0	2.822
35473	1.002	4,7	22	8,9	≈ 1	18	9,9	≈ 1	18	9,9	599
354511	584	4,2	42	9,8	1,7	40	9,9	≈ 1	40	10,0	191
354512	33	9,7	19	6,7	≈ 1	19	7	≈ 1	17*	9,8	4
354530	865	8,3	44	9,7	1,8	43	9,8	2,2	42	9,9	479
354541	160	8,2	31	9,6	≈ 1	30	9,8	≈ 1	30	10,0	27
354542	27	9,9	**15**	5,9	≈ 1	17	6,6	≈ 1	17*	8,3	1
354631	53	4,6	22	8,1	1	22	8,1	≈ 1	19*	10,0	54
354632	60	5,2	22	9,2	≈ 1	21	9,2	≈ 1	20*	9,9	79
354633	305	4,2	30	9,8	≈ 1	30	9,7	≈ 1	29*	10,0	3.356
354634	112	3,8	24	9,6	1,1	23	9,8	≈ 1	23*	10,0	237
354635	129	2,4	22	9,2	≈ 1	19	**10,1**	≈ 1	19*	10,0	973
354636	24	6,0	23	5,1	≈ 1	18	7,3	≈ 1	17*	9,7	2
354637	148	4,4	29	9,6	1	30	9,4	≈ 1	28*	10,0	1.295
354639	253	2,6	30	9,0	≈ 1	28	9,6	≈ 1	27*	10,0	3.125
354641	122	4,9	26	9,5	≈ 1	25	9,7	≈ 1	24*	10,0	305
354642	141	4,5	24	9,7	1,1	21	**10,6**	≈ 1	23*	10,0	1.935
354643	50	4,4	21	6,6	≈ 1	21	6,6	≈ 1	18*	10,0	75
354644	145	1,0	22	8,4	1,3	22	8,4	≈ 1	19*	10,0	1.020
354645	233	3,0	30	9,6	≈ 1	29	9,7	≈ 1	28*	10,0	1.899
354646	178	1,9	25	7,2	≈ 1	20	9,2	≈ 1	19*	10,0	3.220
354647	120	2,4	23	8,2	1	22	8,6	≈ 1	21*	10,0	889
354649	449	4,5	41	9,7	1,1	39	9,9	≈ 1	39	10,0	35
354651	135	1,8	21	7,4	≈ 1	21	7,4	≈ 1	17*	9,9	729
354652	88	2,4	28	6,9	≈ 1	21	8,5	≈ 1	20*	10,0	186
354661	101	2,1	21	9,3	1,1	21	9,3	≈ 1	19*	9,9	266
354663	323	4,5	32	9,6	≈ 1	33	9,4	≈ 1	31*	10,0	3.607
354665	63	4,1	29	6,7	≈ 1	20	9,5	≈ 1	19*	9,9	76
354671	86	4,7	23	9	≈ 1	23	8,8	≈ 1	21*	9,8	108
354672	158	2,9	25	9,7	1,1	24	9,8	≈ 1	24*	10,0	539
354673	103	2,6	23	9,4	1,3	21	9,9	≈ 1	21*	10,0	244
354674	33	1,4	17	5,4	≈ 1	17	5,2	≈ 1	14*	9,9	3
354679	229	3,6	26	9,6	≈ 1	25	9,9	≈ 1	25*	10,0	947

Tabela 5.6: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 4$) (continuação)

Código do	ATUAL		LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	n	CV(%)	t(s)	n	CV(%)	t(s)	n	CV(%)	t(s)
354681	96	2,3	24	8,7	≈ 1	20	9,8	≈ 1	20*	10,0	152
354682	21	0,2	16	,6	≈ 1	13	1	≈ 1	11*	3,4	1
354683	77	2,4	24	7,2	≈ 1	20	8,5	≈ 1	19*	10,0	86
354684	135	2,2	26	8,5	≈ 1	22	9,5	≈ 1	21*	9,9	446
354685	181	1,8	24	9,1	≈ 1	23	9,2	≈ 1	21*	10,0	641
354686	131	5,5	26	9,5	≈ 1	23	**10,1**	≈ 1	24*	10,0	333
354687	259	5,2	30	9,7	≈ 1	29	9,9	≈ 1	29*	10,0	2.281
354689	187	6,4	**26**	9,8	≈ 1	25	10,0	≈ 1	25*	10,0	1.004
354691	90	0,4	17	4,3	≈ 1	17	4,2	≈ 1	14*	9,8	208
354692	50	1,5	20	3,7	≈ 1	20	3,7	≈ 1	14*	9,8	26
354693	110	1,6	24	5,1	1	19	7,5	≈ 1	18*	9,9	1.086
354711	1.769	0,4	30	9,7	1,2	25	**11,2**	≈ 1	29	9,9	165
354712	527	8,6	**31**	9,8	1,2	27	**10,1**	1,8	28	10,0	306
354713	115	4,7	19	7,7	1,1	19	7,9	≈ 1	17	9,0	436
354741	125	7,0	32	9,6	≈ 1	29	9,7	≈ 1	28*	10,0	692
354742	193	8,4	27	9,7	≈ 1	25	10,0	≈ 1	25	10,0	72
354743	114	8,1	**21**	9,7	1,2	21	9,8	≈ 1	21	9,9	14
354744	1.546	7,0	35	9,9	1,8	35	9,8	2,4	35	9,9	812
354751	182	9,2	35	9,8	≈ 1	26	9,9	≈ 1	26	10,0	126
354752	201	8,2	31	9,9	1	31	9,8	≈ 1	30	10,0	52
354753	169	6,0	22	7,2	≈ 1	22	7,2	≈ 1	20	9,9	107
354754	400	7,0	37	9,7	1,2	35	10,0	≈ 1	35	10,0	107
354755	152	8,8	31	9,5	1,2	30	9,4	≈ 1	28	10,0	190
354756	32	9,1	22	9,6	≈ 1	18	9,4	≈ 1	17*	9,8	14
354757	86	8,9	22	9,7	≈ 1	21	9,8	≈ 1	21*	9,9	239
354759	205	8,4	28	9,7	1,2	27	9,9	≈ 1	27	10,0	60
354761	317	8,7	29	9,5	1,3	24	**10,7**	≈ 1	27	10,0	77
354762	33	9,3	19	8,5	≈ 1	21	8,7	≈ 1	17*	9,8	9
354763	152	8,1	25	8,3	1	25	8,1	≈ 1	21	10,0	457
354771	547	4,2	26	9,6	1,3	22	**11,1**	≈ 1	25	10,0	211
354772	280	8,4	34	9,9	≈ 1	33	10,0	≈ 1	33	10,0	149
354773	63	8,6	30	9,7	≈ 1	23	9,7	≈ 1	22*	10,0	106
354774	125	8,4	23	9,8	≈ 1	21	9,8	≈ 1	21	10,0	73
354781	1.151	7,0	40	9,6	3,1	37	**10,2**	2,8	38	10,0	2.726
354782	583	5,5	39	9,9	1,1	40	9,7	≈ 1	39	10,0	178
354783	81	8,3	25	9,5	≈ 1	21	10,0	≈ 1	21*	10,0	227
354784	73	8,6	23	9,8	≈ 1	17	9,9	≈ 1	17	9,9	95

Tabela 5.6: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 4$) (continuação)

Código do	ATUAL		LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	n	CV(%)	t(s)	n	CV(%)	t(s)	n	CV(%)	t(s)
354785	32	9,2	19	9,0	≈ 1	18	9,4	≈ 1	17*	9,9	13
354789	837	8,6	31	9,8	2	29	9,7	2,7	28	10,0	5.000
TOTAL	20.475		1.993			1.837			1.784		

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

A partir de uma análise das soluções vencedoras da Tabela 5.6, é possível notar que o algoritmo EVP teve um desempenho muito superior aos demais. Das 75 populações consideradas no estudo, em todas o melhor resultado foi produzido pelo algoritmo EVP. Enquanto o algoritmo de Kozak (2004) produziu a solução vencedora em 24 populações e o algoritmo de Lavallée e Hidiroglou (1988) apenas em 2 populações. Além disso, conforme o esperado, considerando as definições do critério de problema pequeno e de solução ótima descrita na seção anterior, conclui-se que há 47 populações (marcadas com um asterisco na coluna do tamanho amostral) que atenderam ao critério para o caso em que $L = 4$, o que representa 63% dos resultados da tabela.

Vale ainda destacar que, mesmo sem a execução do algoritmo exato, o algoritmo EVP produziu soluções melhores do que as soluções dos métodos comparados. Por exemplo, como nas populações 35461, 354530, 354713, entre outras.

Analisando a linha de total dessa tabela, observa-se que há um ganho considerável na redução do tamanho amostral total ao se utilizar qualquer um dos três algoritmos, em relação ao método atual. Se o método proposto fosse utilizado, haveria uma redução de 91% do tamanho amostral para o Estado de São Paulo, saindo das atuais 20.475 unidades amostrais para apenas 1.784. Assim, se poderia sugerir que a pesquisa adotasse o método proposto em que são calculados os pontos de corte, de forma a otimizar os resultados, de cada população individualmente. Entretanto, além de outros aspectos econômicos, a metodologia atual é mais simples de compreender, pois estratifica qualquer população considerando os pontos de corte 4, 9, 19 pessoas ocupadas na empresa.

Ainda considerando os tamanhos amostrais produzidos por cada algoritmo da Tabela 5.6, pode-se calcular para cada população um *gap* em relação ao método ATUAL, dado por $GAP_i = \frac{n_{ATUAL} - n_i}{n_i}$, em que i representa os algoritmos LH88, KOZAK04 e EVP, conforme

Brito et al. (2011). Os resultados estão na Tabela 5.7 na seguinte ordem: mínimo, 1º quartil, mediana (ou 2º quartil), média, 3º quartil e máximo. Portanto, a média e a mediana do algoritmo EVP são maiores do que dos outros dois algoritmos, o que indica um ganho maior desse algoritmo em relação aos outros dois.

Tabela 5.7: Medidas de posição dos *gaps*, segundo os algoritmos

Algoritmo	Mínimo	Q1	Mediana (Q2)	Média	Q3	Máximo
LH88	0,04	2,49	4,36	8,20	7,33	57,97
KOZAK04	0,33	2,85	4,44	7,45	6,95	59,45
EVP	0,41	3,19	5,42	9,00	8,04	60,00

Para os demais números de estratos ($L = 3, 5, 6$), os resultados estão nas Tabelas A.3, A.4 e A.5 do Apêndice, respectivamente – essas tabelas têm a mesma estrutura da Tabela 5.6, exceto por não apresentar os resultados para o método $ATUAL$, que conforme apresentado na seção 4.1.2.1 está definido apenas para $L = 4$. Em relação às soluções ótimas produzidas pelo algoritmo EVP, ocorreu o esperado de acordo com o critério utilizado. A quantidade de soluções ótimas foi caindo, conforme o valor de L foi aumentando, vide Tabela 5.8.

Tabela 5.8: Quantidade de Soluções Ótimas por Número de Estratos das 75 Populações da PAC

Número de estratos (L)	3	4	5	6
Quantidade de soluções ótimas	47	47	23	8

Considerando o tempo de processamento da Tabela 5.6, a diferença é significativa, enquanto os algoritmos LH88 e KOZAK04 levaram no máximo 5 segundos, os tempos do algoritmo EVP variaram de 1 segundo a 5.000 segundos. Entretanto, conforme mencionado anteriormente, o algoritmo EVP resulta de um processo de computação mais intensiva, o que acarreta um consumo de tempo de processamento alto.

Adicionalmente, na Tabela A.5 do Apêndice aparece um valor de mais de 18.000 segundos (aproximadamente 5 horas), isso ocorre porque não há um limite de tempo quando o algoritmo exato (4.3) é executado. E além disso, mesmo para as soluções não exatas, existem alguns resultados com tempo superior a 5.000 nas Tabelas A.4 e A.5 do Apêndice. Isso pode ocorrer, pois o critério de parada só é testado ao final do algoritmo, entre uma iteração e outra. Assim, em um caso excepcional, uma iteração pode ter passado pelo teste de parada, quando estava perto de satisfazer o critério, e então, na próxima iteração poderá

executar todos os comandos do VNDS de $k = 1$ até k_{max}, e pode ainda, ter que aplicar o PR, que tem um custo computacional alto.

A população 354542, no caso de $L = 6$, não apresenta soluções viáveis, todos os algoritmos falharam. Tal fato é explicado pela existência de apenas seis valores distintos, o que cria exatamente seis estratos, fazendo com que a variância dentro dos estratos seja nula sempre, o que gera erro nos algoritmos. Por isso, esse resultado foi desconsiderado, e portanto, foram produzidos 299 resultados (75 populações x 4 números de estratos - 1) em cada método, que servirão para análise.

Foi montada a tabela da eficiência relativa para as populações da PAC, com a mesma estrutura da Tabela 5.5 (por questões de espaço e organização, ela está apresentada no Apêndice A.6). Lembrando que, a célula da tabela preenchida com "N/A" conta como um resultado favorável para o método proposto, porque o método comparado não foi capaz de produzir uma solução viável. A eficiência relativa do método de Lavallée e Hidiroglou (1988) variou de 100 a 183, já a eficiência relativa do método de Kozak (2004) variou de 96 a 150. Especificamente, na comparação entre o método proposto e o método de Lavallée e Hidiroglou (1988), ocorreram 271 resultados favoráveis ao algoritmo EVP, 28 empates e nenhum resultado favorável para LH88. Na comparação entre o método proposto e o método de Kozak (2004), ocorreram 246 resultados favoráveis ao algoritmo EVP, 52 empates e apenas um resultado favorável para KOZAK04, totalizando 299 resultados em cada comparação.

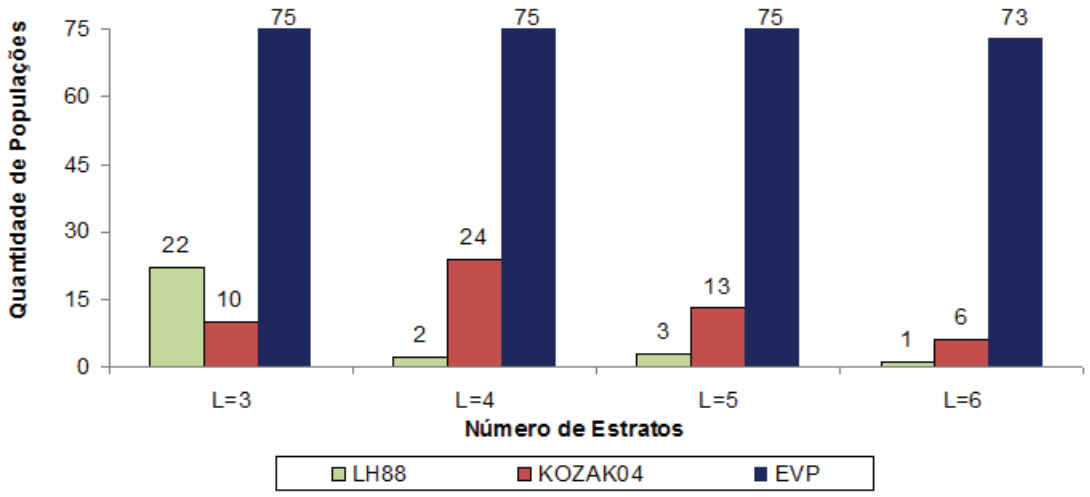

Figura 5.2: Quantidade de Soluções Vencedoras Produzidas por Algoritmo e por Número de Estratos, para as 75 Populações da PAC

A Figura 5.2 resume as soluções vencedoras para todos os números de estratos considerados. Assim, percebe-se que o algoritmo EVP proposto venceu de forma incontestável para todos os números de estratos, com o algoritmo de Kozak (2004), tendo um desempenho regular e o algoritmo de Lavallée e Hidiroglou (1988), tendo o pior desempenho. O algoritmo EVP só não venceu em um dos 299 resultados produzidos, mostrando que é o método mais adequado para a resolução do problema de estratificação quando há a restrição $n_h \geq 5$, pois apresenta resultados de qualidade superior do que os resultados obtidos pelos métodos concorrentes.

O método proposto apresenta-se como uma alternativa aos métodos existentes, considerando a relevância dos resultados obtidos aqui. Em especial, a execução do algoritmo EVP para cinco estratos ($L = 5$) apresentou o menor tamanho de amostra total (n) entre todos os testes realizados com as populações de São Paulo da PAC, com apenas 1.594 empresas, o que geraria uma redução amostral de 92% se comparado ao método atual da pesquisa. Sendo assim, poderia ser sugerido ao IBGE que considere a possibilidade de avaliar a aplicabilidade do método proposto e de outro parâmetro para L. Entretanto, é necessário levar em conta também os aspectos da teoria econômica.

5.4 - Avaliação Geral do Algoritmo Proposto

Considerando as 100 populações (25 da literatura + 75 da PAC), calculou-se o percentual de soluções vencedoras para cada número de estrato (L) testado, apresentado na Figura 5.3. Considerando essa análise, o algoritmo EVP proposto foi o que apresentou os melhores resultados, variando de 94% a 100% na capacidade de produzir o melhor resultado para a população.

Ainda na mesma figura, tem-se o percentual total com base nos 400 resultados (100 populações x 4 números de estratos) que cada algoritmo produziu. Assim, de modo geral, nesse estudo empírico, o algoritmo EVP foi capaz de produzir a solução vencedora em 97% dos casos, enquanto o algoritmo de Kozak (2004) produziu a solução vencedora em 37% dos casos e o de Lavallée e Hidiroglou (1988) apenas em 15% dos casos.

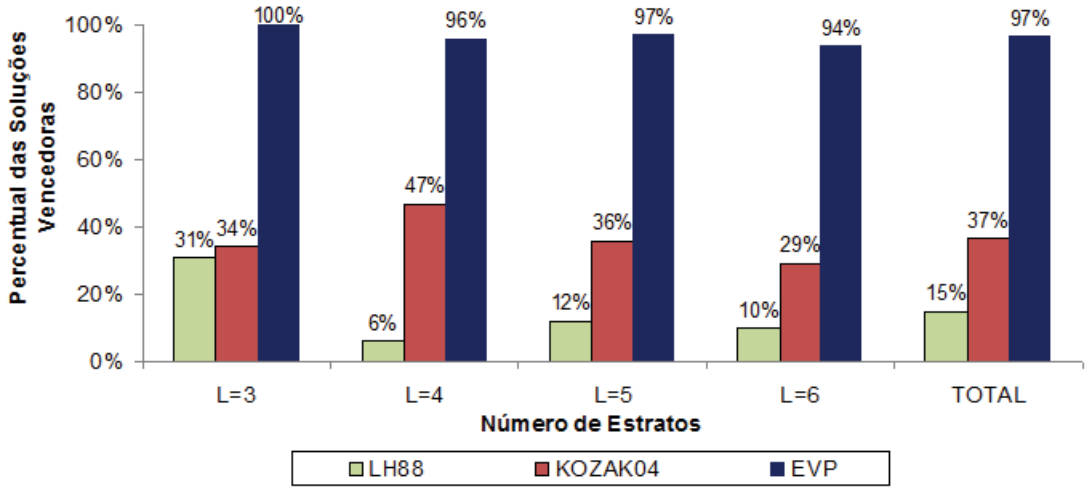

Figura 5.3: Percentual de Soluções Vencedoras Produzidas por Algoritmo e por Número de Estratos, para as 100 Populações Consideradas

A Tabela 5.9 traz as quantidades de soluções ótimas produzidas pelo algoritmo EVP, a partir da execução do algoritmo exato (4.3). Portanto, como já citado anteriormente, o número de soluções ótimas é inversamente proporcional ao número de estratos (L), sendo assim, o método é capaz de produzir a solução ótima em 67% das populações para $L = 3$, mas apenas em 8% das populações para $L = 6$. De modo geral, conseguiu-se produzir 151 soluções ótimas entre os 400 resultados produzidos pelo algoritmo EVP, o que dá um aproveitamento de aproximadamente 38%.

Tabela 5.9: Quantidade de Soluções Ótimas por Número de Estratos das 100 Populações Consideradas

Número de estratos (L)	3	4	5	6	Total
Quantidade de soluções ótimas	67	51	25	8	151

Quando o método proposto produz a solução com base no algoritmo exato (4.3), ele sempre produzirá a mesma resposta em qualquer execução, pois é a solução ótima. Por outro lado, isso nem sempre acontece quando o método proposto produz a solução com base nas metaheurísticas VNDS e PR, denominada a partir de agora, simplesmente por **solução metaheurística**. Essa é uma desvantagem do uso de metaheurísticas, pois uma execução pode apresentar um resultado diferente de outra execução.

Portanto, precisa-se avaliar a estabilidade e qualidade das soluções metaheurísticas produzidas pelo algoritmo EVP. Assim, fez-se uma simulação com 100 execuções do algoritmo, a partir de dez populações selecionadas, armazenando o tamanho amostral (n) resul-

tante de cada execução. Essas dez populações (cinco da literatura e cinco da PAC) foram selecionadas aleatoriamente, dentre aquelas que apresentaram solução metaheurística para todos os números de estratos. Os resultados da simulação para $L = 3$ estão na Tabela 5.10, portanto, cada população apresentou 100 resultados e a partir deles, foram obtidas as medidas de posição, que estão na seguinte ordem: moda (o resultado mais frequente), mínimo, 1º quartil, mediana (ou 2º quartil), 3º quartil e máximo.

Tabela 5.10: Medidas de Posição para uma Simulação de 100 Réplicas do Algoritmo EVP ($L = 3$)

População	Moda	Mínimo	Q1	Mediana (Q2)	Q3	Máximo
U06	23	23	23	23	23	23
U09	21	21	22	24	26	45
U11	42	42	43	45	52	68
U13	32	32	32	34	39	45
U25	50	50	50	53	58	73
354541	59	59	59	59	59	59
354711	55	55	55	55	55	55
354742	54	54	54	54	54	54
354743	45	45	45	45	45	45
354774	41	41	41	41	41	42

De acordo com essa tabela, em todas as dez populações selecionadas o valor mais frequente (a moda) foi igual ao valor mínimo, indicando que ao se fazer uma execução qualquer do algoritmo, essa execução tenderá a apresentar o menor valor possível. Em cinco populações (U06, 354541, 354711, 354742, 354743) a resposta do algoritmo foi a mesma para todas as 100 execuções, mas para as demais populações houve variação de resultados.

Por exemplo, para a população U09 e considerando o valor do 3º quartil (Q3) da tabela, pode-se afirmar que em pelo menos 75 das 100 execuções, o resultado produzido será um tamanho amostral menor ou igual a 26 ($n \leq 26$). Note que para quatro populações da literatura (U09, U11, U13, U25) há pelo menos uma execução que produziu resultados ruins, apresentados na última coluna da tabela. Isso se explica, pelo fato do algoritmo seguir um processo de busca aleatória (não direcionada), fazendo com que, em alguns casos raros, a resposta final do algoritmo não seja tão boa quanto às demais execuções, tendo produzido um ótimo local que está distante do ótimo global.

Tomando como base a moda, se pode afirmar que na maioria das vezes a execução do algoritmo irá produzir a melhor solução possível. E se baseando na mediana, afirma-se que em pelo menos 50% das execuções o algoritmo irá produzir uma boa solução, o que

indica a robustez do método.

Os resultados da simulação para os demais números de estratos ($L = 4, 5, 6$) estão, respectivamente, nas Tabelas A.7, A.8 e A.9 do Apêndice, com a mesma estrutura da Tabela 5.10. Em que o comentário acima sobre a moda nem sempre é válido, enquanto o comentário a respeito da mediana continua válido. Vale notar que para $L = 4$, os valores mínimos produzidos para as populações U09, U11 e U25 apresentadas na Tabela A.7 são menores que os tamanhos amostrais apresentados na Tabela 5.2. O mesmo comentário é válido para $L = 5$, nas populações U09, U13 e U25 da Tabela A.8 em comparação com a Tabela 5.3. E ainda, para $L = 6$ na Tabela A.9, as populações U09, U11 e U25 em comparação com a Tabela 5.4 e as populações 354711, 354743 e 354774 em comparação com a Tabela A.5.

Em relação ao tempo de processamento, após várias execuções do algoritmo EVP até se chegar a versão final do código, há indícios para acreditar que os principais fatores que influenciam o tempo são o número de estratos(L) e o tamanho populacional(N). Portanto, o algoritmo EVP levará mais tempo para aquelas populações muito grandes ($N > 10.000$) e para estratificar em 5 ou 6 ou mais estratos.

Em linhas gerais, o método proposto apresenta-se como uma alternativa aos métodos existentes na literatura para a resolução do problema de estratificação univariado, considerando o objetivo de minimizar o tamanho amostral, principalmente, quando há a restrição para o tamanho amostral mínimo por estrato, em que o método mostrou-se o mais adequado. Quando não há essa restrição, o método apresentou resultados superiores ao de Lavallée e Hidiroglou (1988) e quase equivalentes ao de Kozak (2004). Portanto, considera-se que a utilização do método proposto para qualquer problema de estratificação constitui uma boa alternativa.

Antes de aplicar o algoritmo EVP proposto, caso se queira saber previamente se a solução produzida será uma solução ótima, recomenda-se analisar a população em relação ao tamanho populacional (N), a quantidade de valores distintos(w) e o número de estratos (L) que será utilizado. Quando esses três fatores forem números pequenos, aumenta-se a chance de obter a solução ótima – por números pequenos entenda-se: $N < 1.000$, $w < 200$ e $L \leq 4$. Vale lembrar que, no caso do procedimento de resolução exato ser aplicado, não há limite de tempo, podendo exceder o parâmetro $cpuTime$ especificado.

Ressalte-se que esse algoritmo resulta de um processo de computação mais intensiva, pois 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 para o algoritmo EVP proposto neste trabalho. 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. Assim, o uso deste método pode vir a ser mais vantajoso, apesar do custo computacional.

Conforme o explicitado acima, conclui-se que o método proposto representa um *trade-off* entre tempo e qualidade dos resultados. Se a preferência for pelo último, deve-se aplicar o método sem hesitação, entretanto se a preferência for pelo tempo reduzido, deve-se aplicar os outros métodos da literatura.

6 - Conclusões e Extensões

As primeiras abordagens para o problema de estratificação remetem à década de 50 e, até hoje, é um dos problemas estatísticos que persiste sem solução definitiva para a delimitação dos cortes de estratificação. Esse problema pode ser formulado considerando dois objetivos possíveis: minimizar a variância de um estimador ou minimizar o tamanho amostral. Na literatura, os métodos que mais obtiveram resultados positivos para o segundo objetivo foram Lavallée e Hidiroglou (1988) e Kozak (2004). Todavia, tais métodos não permitem a inclusão da restrição de tamanho amostral mínimo por estrato e também não permitem que haja valores negativos na variável de estratificação da população. A primeira restrição citada é muito importante para algumas pesquisas amostrais realizadas no âmbito do IBGE. Para contornar essa limitação, esse trabalho teve como objetivo desenvolver um novo método capaz de atender a essa restrição.

O objetivo do trabalho foi alcançado, pois foi possível criar um método como mais uma alternativa promissora à resolução do problema de estratificação univariado, levando em conta a restrição citada ou não. Seja o problema de estratificação apresentado na seção 2.3, o método proposto (denominado de algoritmo EVP) resolve a etapa (vii) (busca dos pontos de corte) através da enumeração exaustiva ou através das metaheurísticas VNDS proposta por Hansen, Mladenović e Perez-Brito (2001) e PR proposta originalmente por Fred Glover no livro de Barr, Helgason e Kennington (1996).

Na etapa (viii) utilizou-se a alocação ótima proposta por Brito et al. (2015). Nos casos em que a solução é obtida através da enumeração exaustiva, pode-se garantir que essa solução corresponde a um mínimo global, pois a alocação de Brito et al. (2015) também é um método exato.

Para verificar a qualidade das soluções foi realizado um experimento contendo 100 populações citadas na seção 4.1 (25 populações da literatura e 75 populações do Estado de São Paulo associadas à base de dados originada no CBS da PAC 2014) visando a minimização do tamanho amostral. Ainda foram testados os resultados para o número de

estratos (L) variando de 3 a 6. Além disso, foram também obtidos os resultados através dos métodos de Lavallée e Hidiroglou (1988) e Kozak (2004), com o intuito de poder compará-los. Portanto, no total foram produzidos 1200 resultados (100 populações x 3 métodos x 4 números de estratos).

O método proposto foi capaz de produzir a solução vencedora em 97% dos resultados possíveis, sendo que em 38% ainda foi capaz de garantir que a solução correspondia a um mínimo global. Além disso, no caso específico de $L = 3$ conseguiu produzir o mínimo global em 67% dos casos. Quando há a restrição para o tamanho amostral mínimo por estrato, o algoritmo EVP proposto apresentou um desempenho muito superior aos demais métodos, e quando não há essa restrição, o método apresentou resultados superiores ao de Lavallée e Hidiroglou (1988) e quase equivalentes ao de Kozak (2004).

Especificamente nos testes realizados com as populações da PAC, o algoritmo EVP produziu os melhores resultados para todos os números de estratos considerados nesse estudo, sendo capaz de reduzir o tamanho amostral total em 91%, se comparado com a metodologia atual da pesquisa. Portanto, o método proposto apresenta-se como uma alternativa que pode ser considerada pelo IBGE. Entretanto, mudanças metodológicas precisam ser avaliadas pelas áreas responsáveis do Instituto, levando em conta a viabilidade do método e também os aspectos da teoria econômica.

Para produzir soluções de boa qualidade, o algoritmo EVP utiliza um procedimento que demanda um custo computacional mais intensivo que explora uma quantidade bem maior de soluções. Dessa forma, esse processo intensivo resulta em tempos computacionais maiores do que de seus concorrentes. Evidenciando, assim, o trade-off entre tempo e qualidade dos resultados, sendo o último a preferência do método proposto.

A contribuição do trabalho está no algoritmo EVP proposto, capaz de resolver, dentro de um tempo razoável, qualquer problema de estratificação univariado de minimização do tamanho amostral, com presença ou não da restrição de tamanho amostral mínimo por estrato. E também por ser aplicável a todo tipo de população, mesmo aquelas com valores negativos ou aquelas com tamanho populacional muito grande, sendo ainda capaz de produzir o mínimo global em determinados casos. Portanto, se apresenta como uma alternativa aos métodos existentes na literatura à resolução do problema de estratificação, com a disponibilização do código fonte no Apêndice B.

Além disso, até o presente momento, esse estudo já resultou em um artigo publicado no XLVIII Simpósio Brasileiro de Pesquisa Operacional, intitulado de “Metaheurística VNDS

Aplicada ao Problema de Estratificação Ótima", quando foram apresentados os resultados parciais, vide Oliveira, Lima e Brito (2016). Adicionalmente, há mais um artigo em fase final de produção que será submetido à *5th International Conference on Variable Neighborhood Search* em parceria com o periódico *Electronic Notes in Discrete Mathematics*.

Convém salientar ainda que, com o avanço tecnológico dos processadores dos computadores nos próximos anos, o algoritmo EVP deverá apresentar uma queda significativa no tempo de processamento. Além disso, é possível diminuir esse tempo através da otimização do código fonte e da utilização de técnicas de paralelização.

Como possíveis extensões têm-se: a generalização do método para o problema de estratificação multivariado, quando há mais de uma variável de estratificação; desenvolver o método para atender também ao primeiro objetivo do problema de estratificação, minimizar a variância de um estimador (considerando o tamanho de amostra fixo).

Em trabalhos futuros, ainda se pode testar outras Unidades da Federação, outras pesquisas ou outras populações. Há ainda a possibilidade de se testar outras metaheurísticas para auxiliar a busca.

Referências Bibliográficas

Almeida, A. G. R. (2007). “Métodos para delimitação de estratos em populações assimétricas: uma comparação”. Diss. de mestrado. Rio de Janeiro: Escola Nacional de Ciências Estatísticas.

Azevedo, R. V. (2004). “Estudo Comparativo de Métodos de Estratificação Ótima de Populações Assimétricas”. Diss. de mestrado. Rio de Janeiro: Escola Nacional de Ciências Estatísticas.

Baillargeon, S.; Rivest, L.-P. (2011). “The construction of stratified designs in R with the package stratification”. Em: *Survey Methodology* 37.1, pp. 53–65.

Ballin, M.; Barcaroli, G. (2013). “Joint determination of optimal stratification and sample allocation using genetic algorithm”. Em: *Survey Methodology* 39.2, pp. 369–393.

Bankier, M. D. (1988). “Power Allocations: Determining Sample Sizes for Subnational Areas”. Em: *The American Statistician* 42, pp. 174–177.

Barr, R.; Helgason, R.; Kennington, J., eds. (1996). *Interfaces in Computer Science and Operations Research*. Boston: Kluwer Academic Publishers.

Blum, C.; Roli, A. (2003). “Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison”. Em: *ACM Computing Surveys* 35.3, pp. 268–308.

Bolfarine, H.; Bussab, W. O. (2005). *Elementos de amostragem*. 1ª ed. Edgard Blücher.

Bramati, M. (2012). “Robust Lavallée-Hidiroglou stratified sampling strategy”. Em: *Survey Research Methods* 6.3, pp. 137–143.

Brito, J. A. M.; Maculan, N.; Brito, L. R.; Montenegro, F. M. T. (2011). “Um algoritmo GRASP aplicado ao problema de estratificação”. Em: *XLIII Simpósio Brasileiro de Pesquisa Operacional (SBPO)*. Ubatuba,SP.

Brito, J. A. M.; Montenegro, F. M. T. (2007). “Algoritmos heurísticos aplicados ao problema de estratificação ótima em populações assimétricas”. Em: *XXXIX Simpósio Brasileiro de Pesquisa Operacional (SBPO)*. Fortaleza,CE.

Brito, J. A. M.; Montenegro, F. M. T.; Maculan, N.; Azevedo, R. V. (2006). “Propostas para o problema de estratificação em amostras considerando alocação proporcional”. Em: *XXXIX Simpósio Brasileiro de Pesquisa Operacional (SBPO)*. Goiânia,GO.

Brito, J. A. M.; Ochi, L.; Montenegro, F. M. T.; Maculan, N. (2010). “An iterative local search approach applied to the optimal stratification problem”. Em: *International Transactions in Operational Research* 17, pp. 753–764.

Brito, J. A. M.; Semaan, G. S.; Brito, L. R. (2014). “Algoritmos heurísticos aplicados ao problema de estratificação ótima”. Em: *XLVI Simpósio Brasileiro de Pesquisa Operacional (SBPO)*. Salvador,BA.

Brito, J. A. M.; Silva, P. L. N.; Semaan, G. S.; Maculan, N. (2015). “Integer programming formulations applied to optimal allocation in stratified sampling”. Em: *Survey Methodology* 41.2, pp. 427–442.

Chambers, R.; Dunstan, R. (1986). “Estimating distribution functions from survey data”. Em: *Biometrika* 73.3, pp. 597–604.

Cochran, W. G. (1977). *Sampling Techniques*. 3ª ed. New York: John Wiley & Sons.

Cormen, T.; Leiserson, C.; Rivest, R.; Stein, C. (2002). *Algoritmos: Teoria e Prática*. 2ª ed. Editora Campus Elsevier.

Dalenius, T. (1950). “The problem of optimum stratification”. Em: *Skandinavisk Aktuarietidskrift*, pp. 203–213.

Dalenius, T.; Hodges, J. (1959). “Minimum variance stratification”. Em: *Journal of the American Statistical Association* 285.54, pp. 88–101.

DeGroot, M. H.; Schervish, M. J. (2012). *Probability and Statistics*. Fourth Edition. Boston: Pearson Education, Inc.

Doane, D. P.; Seward, L. E. (2011). “Measuring Skewness: A Forgotten Statistic?” Em: *Journal of Statistics Education* 19.2.

Ekman, G. (1959). “An approximation useful in univariate stratification”. Em: *The Annals of Mathematical Statistics* 30.1, pp. 219–229.

Er, S. (2011). “Comparison of the Efficiency of the Various Algorithms in Stratified Sampling when the Initial Solutions are Determined with Geometric Method”. Em: *International Journal of Statistics and Applications* 1, pp. 1–10.

Freitas, M. P. S. (2002). “Estratificação para a Amostra de uma Pesquisa Domiciliar sobre Mercado de Trabalho”. Diss. de mestrado. Rio de Janeiro: Escola Nacional de Ciências Estatísticas.

Gendreau, M.; Potvin, J.-Y., eds. (2010). *Handbook of Metaheuristics*. 2ª ed. Vol. 146. Springer.

Glasser, G. (1962). “On the complete coverage of large units in a statistical study”. Em: *Review of the International Statistical Institute* 30.1, pp. 28–32.

Glover, F.; Kochenberger, G. A., eds. (2003). *Handbook of Metaheuristics*. Springer.

Gunning, P.; Horgan, J. M. (2004). “A new algorithm for the construction of stratum boundaries in skewed populations”. Em: *Survey Methodology* 30.2, pp. 159–166.

Han, J.; Kamber, M.; Pei, J. (2011). *Data Mining: Concepts and Techniques*. 3ª ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hansen, P.; Mladenović, N. (2001). “Variable neighborhood search: Principles and applications”. Em: *European Journal of Operational Research* 130.3, pp. 449–467.

Hansen, P.; Mladenović, N.; Perez-Brito, D. (2001). “Variable Neighborhood Decomposition Search”. Em: *Journal of Heuristics* 7.4, pp. 335–350.

Hedlin, D. (2000). “A procedure for stratification by an extended ekman rule”. Em: *Journal of Official Statistics* 16.1, pp. 15–29.

Hidiroglou, M. A. (1986). “The construction of a self-representing stratum of large units in survey design”. Em: *The American Statistician* 40.1, pp. 27–31.

Horgan, J. M. (2006). “Stratification of skewed populations: A review”. Em: *International Statistical Review* 74.1, pp. 67–76.

IBGE (2015). *Pesquisa Anual de Comércio 2013*. Vol. 25, pp. 1–110.

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

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, p. 643.

Kozak, M. (2004). “Optimal stratification using random search method in agricultural surveys”. Em: *Statistics in Transition* 6.5, pp. 797–806.

Kozak, M. (2006). “Multivariate Sample Allocation: Application of a Random Search Method”. Em: *Statistics in Transition* 7.4, pp. 889–900.

Kozak, M.; Verma, M. R. (2006). “Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency”. Em: *Survey Methodology* 32.2, pp. 157–163.

Kozak, M.; Verma, M. R.; Zieliński, A. (2007). “Modern Approach to Optimum Stratification: Review and Perspectives”. Em: *Statistics in Transition* 8.2, pp. 223–250.

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

Lednicki, B.; Wieczorkowski, R. (2003). “Optimal stratification and sample allocation between subpopulations and strata”. Em: *Statistics in Transition* 6.2, pp. 287–305.

Linden, R. (2012). *Algoritmos Genéticos*. 3ª ed. Editora Ciência Moderna, p. 475.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2ª ed., p. 608.

Mladenović, N.; Hansen, P. (1997). “Variable neighborhood search”. Em: *Computer Ops Res* 24.11, pp. 1097–1100.

Oliveira, B. T. N. T.; Lima, L. S.; Brito, J. A. M. (2016). “Metaheurística VNDS Aplicada ao Problema de Estratificação Ótima”. Em: *XLVIII Simpósio Brasileiro de Pesquisa Operacional (SBPO)*. Vitória - ES. URL: http://www.sbpo2016.iltc.br/pdf/155837.pdf.

Papadimitriou, C. H.; Steiglitz, K. (1982). *Combinatorial Optimization - Algorithms and Complexity*. New York: Dover Publications, Inc.

Rao, D.; Khan, M.; Reddy, K. (2014). “Optimum stratification of a skewed population”. Em: *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering* 8.3, pp. 492–495.

Resende, M. G. C.; Sousa, J. P., eds. (2003). *Metaheuristics: Computer Decision-Making*. New York: Kluwer Academic Publishers.

Rivest, L.-P. (2002). “A Generalization of the Lavallé and Hidiroglou Algorithm for Stratification in Business Surveys”. Em: *Survey Methodology* 28.2, pp. 191–198.

Santos, N. M. (2014). “Um estudo de problemas de clusterização com restrições de capacidade e de conexidade”. Tese de doutorado. Niterói: Universidade Federal Fluminense.

Sethi, V. (1963). “A Note on the optimum Stratification of Populations for Estimating the Population Means”. Em: *Australian Journal of Statistics* 5, pp. 20–33.

Souza, M. J. F. (2011). *Inteligência Computacional para Otimização*. Ouro Preto, MG. URL: http://www.decom.ufop.br/prof/marcone/Disciplinas/InteligenciaComputacional.

Talbi, E.-G. (2009). *Metaheuristics: from design to implementation*. New Jersey: John Wiley & Sons, Inc.

Veiga, T. M. (2015). “Um estudo comparativo de métodos para delimitação de estratos em populações assimétricas”. Diss. de mestrado. Rio de Janeiro: Escola Nacional de Ciências Estatísticas.

A - Tabelas Adicionais

Tabela A.1: Descrições das 25 populações da literatura

ID	Nome	Pacote no R	Referência	Descrição
U01	BeeFarms	-	Chambers e Dunstan (1986)	Fazendas australianas de gado estratificadas em sete estratos definidos por regiões industriais
U02	beta103	GA4Stratification	-	População gerada aleatoriamente da distribuição Beta com os parâmetros a=10 e b=3
U03	chi1	GA4Stratification	-	População gerada aleatoriamente da distribuição Qui-Quadrado com 1 grau de liberdade
U04	chi5	GA4Stratification	-	População gerada aleatoriamente da distribuição Qui-Quadrado com 5 graus de liberdade
U05	debtors	stratification	-	População de devedores em uma firma irlandesa
U06	HHINCTOT	stratification	-	Rendimento familiar antes dos impostos em 2001 no Canadá.
U07	iso2004	GA4Stratification	-	Vendas líquidas de empresas industriais turcas em 2004. População dividida por 1000.
U08	Kozak1	-	Kozak e Verma (2006)	Populações dadas no artigo de Kozak e Verma
U09	Kozak2	-	Kozak e Verma (2006)	Populações dadas no artigo de Kozak e Verma
U10	Kozak3	-	Kozak e Verma (2006)	Populações dadas no artigo de Kozak e Verma
U11	Kozak4	-	Kozak e Verma (2006)	Populações dadas no artigo de Kozak e Verma
U12	me84	Sampling	-	Quantidade de funcionários municipais de 284 municípios da Suécia em 1984
U13	mrts	stratification	-	População simulada a partir da Pesquisa Mensal de Vendas e Comércio da Statistics Canada
U14	p100e10	GA4Stratification	-	População gerada aleatoriamente da distribuição Normal sendo $\mu = 100$ e $\sigma = 10$
U15	p75	GA4Stratification	-	População em milhares de 284 municípios na Suécia em 1975
U16	pop800	-	Hedlin (2000)	Gerada aleatoriamente da distribuição Log-Normal($X = e^Z$), onde Z segue uma N($\mu = 4; \sigma^2 = 2,7$)
U17	rev84	Sampling	-	Valores dos imóveis em milhões de Coroas suecas de 284 municípios em 1984.
U18	SugarCaneFarms	-	Chambers e Dunstan (1986)	População de fazendas de cana de açúcar da Austrália
U19	Swiss	Sampling	-	Informações sobre os municípios suíços (2003)
U20	TaxableIncome	Sampling	-	Rendimento tributável de municípios da Bélgica em 2001(em euros). População dividida por 1000.
U21	Usbanks	stratification	-	Recursos em milhões de dólares de grandes bancos norte-americanos comerciais.
U22	Uscities	stratification	-	População em milhares de cidades norte-americanas em 1940.
U23	Uscolleges	stratification	-	Quantidade de estudantes em faculdades dos EUA de quatro anos em 1952-1953.
U24	Rchisq2-30	-	Baillargeon e Rivest (2011)	População gerada aleatoriamente da distribuição Qui-Quadrado com 30 graus de liberdade
U25	M101	stratification	-	Despesas familiares com recreação em 2001 no Canadá

Tabela A.2: Informações básicas das 75 populações da PAC

Código do Estrato Natural	Tamanho Populacional (N)	Valores Distintos (w)	Assimetria
35461	32.645	54	37,4
35462	1.941	79	12,3
35466	2.706	118	43,5
35472	63.565	149	13,0
35473	9.168	88	10,6
354511	9.938	233	27,7
354512	540	14	8,0
354530	41.377	157	78,3
354541	4.598	77	10,4
354542	112	6	1,2
354631	463	51	6,8
354632	773	50	11,3
354633	3.608	106	23,7
354634	912	73	4,5
354635	1.097	91	28,1
354636	149	24	4,0
354637	1.864	88	12,2
354639	2.278	119	24,2
354641	1.425	67	8,3
354642	2.552	86	36,0
354643	751	44	22,3
354644	682	104	15,3
354645	1.776	106	12,3
354646	1.486	103	21,9
354647	1.207	81	22,7
354649	5.565	143	17,9
354651	1.096	84	31,0
354652	609	68	9,9
354661	605	75	7,9
354663	3.564	103	19,1

Tabela A.2: Informações básicas das populações da PAC (continuação)

Código do Estrato Natural	Tamanho Populacional (N)	Valores Distintos (w)	Assimetria
354665	705	50	16,2
354671	716	56	5,7
354672	1.232	82	12,5
354673	655	73	7,5
354674	65	31	4,3
354679	1.787	90	10,2
354681	451	70	4,1
354682	54	21	5,1
354683	421	57	9,1
354684	708	80	9,2
354685	777	91	8,7
354686	1.561	66	5,8
354687	3.745	101	7,7
354689	3.603	76	16,9
354691	527	81	22,2
354692	255	47	10,7
354693	1.509	81	24,5
354711	5.493	384	45,7
354712	32.089	96	10,5
354713	6.423	71	78,4
354741	3.260	68	11,9
354742	6.020	70	10,8
354743	4.167	56	19,5
354744	47.923	147	116
354751	15.059	66	21,0
354752	6.454	82	28,3
354753	7.405	69	65,1
354754	14.564	114	60,2
354755	11.033	74	46,6
354756	1.084	29	15,3
354757	3.811	47	18,3

Tabela A.2: Informações básicas das populações da PAC (continuação)

Código do Estrato Natural	Tamanho Populacional (N)	Valores Distintos (w)	Assimetria
354759	9.214	68	22,8
354761	15.446	87	97,5
354762	1.347	23	23,6
354763	10.287	64	78,3
354771	14.438	130	91,7
354772	11.987	92	33,1
354773	2.327	41	11,2
354774	6.566	55	59,7
354781	61.368	171	123,3
354782	11.634	135	37,5
354783	3.667	49	11,3
354784	5.686	48	70,0
354785	1.439	25	6,5
354789	65.431	107	195,8

Fonte: CBS 2014

Tabela A.3: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 3$)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
35461	72	9,9	≈ 1	42	9,9	≈ 1	42	10,0	100
35462	47	9,8	≈ 1	51	9,2	≈ 1	47*	10,0	96
35466	50	9,8	≈ 1	40	**11,8**	≈ 1	49*	10,0	78
35472	76	10,0	1	76	10,0	3	76	10,0	194
35473	34	9,9	≈ 1	33	10,0	≈ 1	33	10,0	16
354511	76	9,9	≈ 1	81	9,4	≈ 1	75	10,0	29
354512	24	9,7	≈ 1	24	9,6	≈ 1	23*	10,0	1
354530	86	10,0	1	89	9,7	1	86	10,0	103
354541	60	9,8	≈ 1	60	9,8	≈ 1	59	9,9	8
354542	13	9,2	≈ 1	13	9,2	≈ 1	12*	9,9	1
354631	29	9,6	≈ 1	30	9,3	≈ 1	28*	9,9	3
354632	34	9,9	≈ 1	34	9,8	≈ 1	34*	10,0	4
354633	58	9,8	≈ 1	61	9,5	≈ 1	57*	10,0	58
354634	45	9,8	≈ 1	46	9,6	≈ 1	44*	10,0	9
354635	37	9,7	≈ 1	30	**12,5**	≈ 1	36*	10,0	22
354636	19	7,2	≈ 1	19	7,2	≈ 1	17*	9,6	1
354637	48	9,7	≈ 1	50	9,4	≈ 1	47*	10,0	30
354639	47	9,9	≈ 1	44	**10,4**	≈ 1	46*	10,0	56
354641	46	9,8	≈ 1	48	9,5	≈ 1	45*	10,0	12
354642	45	9,8	≈ 1	35	**12,1**	≈ 1	44*	10,0	46
354643	26	9,6	≈ 1	18	**13,1**	≈ 1	24*	10,0	4
354644	28	9,4	≈ 1	26	**10,4**	≈ 1	27*	10,0	22
354645	51	9,8	≈ 1	54	9,4	≈ 1	50*	10,0	39
354646	29	9,8	≈ 1	22	12,8	≈ 1	29*	10,0	55
354647	32	9,6	≈ 1	28	**11,2**	≈ 1	31*	10,0	18
354649	71	9,9	≈ 1	75	9,5	≈ 1	70	10,0	19
354651	27	9,8	≈ 1	17	**13,9**	≈ 1	26*	10,0	25
354652	27	10,0	≈ 1	27	10,0	≈ 1	27*	10,0	6

Tabela A.3: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 3$) (continuação)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
354661	33	9,7	≈ 1	36	8,6	≈ 1	33*	10,0	10
354663	56	9,9	≈ 1	60	9,5	≈ 1	56*	10,0	69
354665	29	9,7	≈ 1	27	**10,5**	≈ 1	29*	10,0	4
354671	35	9,8	≈ 1	36	9,6	≈ 1	35*	9,9	4
354672	45	9,6	≈ 1	48	8,9	≈ 1	44*	10,0	15
354673	35	9,6	≈ 1	36	9,4	≈ 1	34*	10,0	7
354674	16	8,5	≈ 1	15	9,1	≈ 1	14*	9,9	1
354679	47	10,0	≈ 1	50	9,5	≈ 1	47*	10,0	23
354681	33	9,7	≈ 1	35	9,2	≈ 1	33*	10,0	5
354682	14	0,9	≈ 1	11	1,5	≈ 1	9*	3,5	1
354683	26	9,5	≈ 1	23	**10,6**	≈ 1	25*	9,9	4
354684	32	9,9	≈ 1	34	9,3	≈ 1	32*	10,0	12
354685	39	10,0	≈ 1	41	9,6	≈ 1	39*	10,0	15
354686	47	9,8	≈ 1	47	9,7	≈ 1	46*	10,0	11
354687	62	9,9	≈ 1	63	9,8	≈ 1	62*	10,0	49
354689	56	10,0	≈ 1	58	9,7	≈ 1	56*	10,0	28
354691	14	8,0	≈ 1	14	8,0	≈ 1	12*	9,9	7
354692	15	7,6	≈ 1	15	7,4	≈ 1	14*	9,8	2
354693	22	7,7	≈ 1	22	7,8	≈ 1	20*	9,8	36
354711	56	9,7	≈ 1	44	**12,3**	≈ 1	55	9,9	10
354712	63	9,9	1	62	10,0	1	62	10,0	67
354713	28	9,5	≈ 1	18	**14,1**	≈ 1	27	10,0	19
354741	52	9,9	≈ 1	55	9,5	≈ 1	52*	10,0	17
354742	55	9,8	≈ 1	54	9,9	≈ 1	54	10,0	11
354743	45	9,9	≈ 1	46	9,7	≈ 1	45	9,9	7
354744	70	9,9	1	71	9,8	1	69	10,0	222
354751	51	9,9	≈ 1	50	9,9	≈ 1	50	10,0	25
354752	59	9,9	≈ 1	62	9,5	≈ 1	58	10,0	12

Tabela A.3: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 3$) (continuação)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
354753	26	9,6	≈ 1	18	**12,8**	≈ 1	25	9,9	27
354754	64	9,8	≈ 1	66	9,6	≈ 1	63	10,0	26
354755	51	9,9	≈ 1	52	9,6	≈ 1	50	9,9	21
354756	29	9,9	≈ 1	30	9,6	≈ 1	29*	9,9	2
354757	42	9,8	≈ 1	41	9,9	≈ 1	41*	10,0	9
354759	53	9,8	≈ 1	53	9,8	≈ 1	52	9,9	17
354761	55	9,9	≈ 1	48	**10,7**	≈ 1	54	10,0	33
354762	27	9,7	≈ 1	27	9,7	≈ 1	26*	10,0	2
354763	33	9,6	≈ 1	24	**12,7**	≈ 1	32	9,9	27
354771	46	9,8	≈ 1	32	**13,7**	≈ 1	46	10,0	35
354772	65	9,9	≈ 1	68	9,6	≈ 1	64	10,0	20
354773	39	9,7	≈ 1	38	9,8	≈ 1	37*	10,0	6
354774	43	9,8	≈ 1	42	9,8	≈ 1	41	10,0	12
354781	66	9,9	1	62	**10,5**	2	66	10,0	230
354782	77	9,9	≈ 1	83	9,4	≈ 1	77	10,0	25
354783	42	9,8	≈ 1	42	9,8	≈ 1	41*	10,0	10
354784	32	9,7	≈ 1	19	**13,6**	≈ 1	30	10,0	20
354785	30	9,9	≈ 1	28	9,8	≈ 1	28*	10,0	2
354789	55	9,9	1	55	9,8	1	54	10,0	170

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

Tabela A.4: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 5$)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
35461	**21**	7,0	1	22	6,7	1	20	9,8	6018
35462	31	8,5	≈ 1	25	8,0	≈ 1	23	10,0	93
35466	31	9,3	1	23	8,3	≈ 1	22	9,0	160
35472	**27**	9,5	1	22	9,4	7	22	9,8	7.428
35473	**21**	7,4	1	22	6,7	≈ 1	19	10,0	1.545
354511	27	9,5	1	25	9,7	≈ 1	25	10,0	439
354512	23	5,4	≈ 1	22	5,8	≈ 1	19*	7,0	56
354530	36	9,7	1	27	9,7	3	26	10,0	1.536
354541	27	9,2	1	24	8,9	≈ 1	22	9,6	142
354542	**20**	5,4	≈ 1	22	5,3	≈ 1	22*	1,0	1
354631	27	6,6	≈ 1	23	6,7	≈ 1	19*	10,0	870
354632	27	7,1	≈ 1	22	7,2	≈ 1	20*	9,9	1.340
354633	26	8,6	≈ 1	26	8,5	≈ 1	23	10,0	155
354634	23	7,9	≈ 1	22	8,3	≈ 1	22	8,9	16
354635	23	7,2	1	23	6,5	≈ 1	22	7,6	16
354636	28	3,0	≈ 1	24	4,5	≈ 1	17*	9,6	10
354637	28	9,7	1	26	7,6	≈ 1	23	10,0	63
354639	25	8,0	2	25	7,9	≈ 1	23	9,6	74
354641	26	9,4	≈ 1	22	8,1	≈ 1	22*	9,9	9.642
354642	26	8,8	1	25	7,3	≈ 1	22	9,4	54
354643	26	5,4	≈ 1	24	5,5	≈ 1	19*	9,8	893
354644	24	6,0	1	24	5,7	≈ 1	22	8,3	12
354645	26	8,4	1	23	8,5	≈ 1	22	9,3	24
354646	24	5,8	1	24	5,8	≈ 1	21	9,5	70
354647	27	7,6	≈ 1	26	5,7	≈ 1	21	9,1	34
354649	27	9,3	1	25	9,7	≈ 1	24	9,9	117
354651	22	6,0	1	22	6,1	≈ 1	19	9,3	35
354652	33	6,3	≈ 1	25	5,7	≈ 1	20*	10,0	3.494

Tabela A.4: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 5$) (continuação)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
354661	24	6,8	1	23	7	≈ 1	22	9,3	7
354663	25	9,0	1	23	9,3	≈ 1	22	9,9	101
354665	33	5,6	≈ 1	22	7,1	≈ 1	19*	9,9	1.127
354671	28	7,1	≈ 1	24	7,2	≈ 1	21*	9,8	1.976
354672	24	7,2	1	24	7,1	≈ 1	22	9,6	39
354673	24	6,9	2	24	6,8	≈ 1	21	10,0	21
354674	22	3,4	≈ 1	22	3	≈ 1	16*	9,3	20
354679	25	7,8	≈ 1	24	8,2	≈ 1	23	9,0	41
354681	24	6,7	1	24	6,6	≈ 1	20*	10,0	2.900
354682	**17**	0,4	≈ 1	16	0,5	≈ 1	13*	1,4	4
354683	25	4,9	1	25	5,1	≈ 1	19*	10,0	1.295
354684	27	6,5	1	23	7,5	≈ 1	22	9,3	17
354685	27	6,0	1	23	7,6	≈ 1	22	9,6	7
354686	24	9,3	1	22	8,0	≈ 1	22*	10,0	8.351
354687	22	9,0	1	22	8,9	≈ 1	22	9,6	145
354689	**25**	9,4	1	22	8,0	≈ 1	22	9,5	91
354691	22	2,7	1	22	2,7	≈ 1	16	8,8	10
354692	25	3,0	≈ 1	20	3,7	≈ 1	16*	9,5	287
354693	29	4,4	1	23	4,5	≈ 1	19	8,9	68
354711	29	7,3	2	28	7,5	≈ 1	24	10,0	1.185
354712	**22**	9,3	≈ 1	22	8,5	2	22	9,8	3.657
354713	24	5,6	≈ 1	23	5,3	≈ 1	19	9,6	749
354741	25	8,4	1	25	7,9	≈ 1	22*	10,0	15.020
354742	24	9,2	≈ 1	22	8,7	≈ 1	22	9,7	275
354743	**21**	8,8	≈ 1	22	7,2	≈ 1	22	9,5	130
354744	32	9,5	1	26	9,1	3	24	10,0	5.967
354751	22	8,6	1	23	8,7	≈ 1	22	9,7	498
354752	28	9,4	≈ 1	25	8,7	≈ 1	22	10,0	111

Tabela A.4: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 5$) (continuação)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
354753	27	5,4	≈ 1	24	6,4	≈ 1	19	10,0	690
354754	32	9,5	1	25	9,0	≈ 1	24	10,0	584
354755	31	7,5	≈ 1	25	8,8	≈ 1	24	9,1	399
354756	22	6,8	≈ 1	22	7,1	≈ 1	19*	9,2	136
354757	24	7,2	≈ 1	23	7,4	≈ 1	21*	9,9	3.833
354759	25	9,2	≈ 1	23	8,4	≈ 1	22	9,8	269
354761	26	8,9	1	23	8,9	≈ 1	22	9,6	547
354762	24	5,2	≈ 1	22	6,1	≈ 1	19*	9,1	87
354763	29	6,0	1	24	7,4	≈ 1	21	9,9	634
354771	27	7,3	1	27	7,5	≈ 1	22	9,9	4.255
354772	29	9,4	≈ 1	24	9,7	≈ 1	23	10,0	5.649
354773	25	7,6	≈ 1	22	7,9	≈ 1	22*	9,6	1.423
354774	22	8,4	≈ 1	22	7,2	≈ 1	21	9,9	356
354781	34	8,3	3	28	9,2	4	25	9,9	5.750
354782	27	9,1	1	26	9,2	≈ 1	25	9,9	489
354783	23	8,4	≈ 1	23	7,5	≈ 1	21*	10,0	4229
354784	22	6,6	≈ 1	22	6,4	≈ 1	22	8,5	146
354785	23	6,0	≈ 1	22	5,8	≈ 1	19*	9	115
354789	27	8,6	2	25	8,4	4	24	10,0	4.617

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

Tabela A.5: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 6$)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
35461	26	5,6	1	27	5,5	1	24	7,3	3.687
35462	30	6,5	1	30	5,9	≈ 1	25	9,4	216
35466	28	7,3	1	28	6,3	≈ 1	24	8,6	328
35472	26	7,3	2	27	6,7	5	28	9,2	5124
35473	26	5,4	1	27	4,9	1	25	9,5	758
354511	28	8,8	1	27	7,6	≈ 1	27	8,6	1.178
354512	27	5,4	≈ 1	24	5,3	≈ 1	21*	6,5	5
354530	33	8,1	1	27	8,5	4	27	8,9	2.519
354541	28	7,2	1	27	6,8	≈ 1	24	9,9	226
354542	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
354631	32	4,4	≈ 1	28	4,7	≈ 1	22	9,6	15
354632	31	4,9	≈ 1	27	5,3	≈ 1	21	9,7	109
354633	31	7,2	1	29	6,4	≈ 1	24	9,9	448
354634	26	7,5	1	27	5,6	≈ 1	24	8,8	30
354635	28	5,6	1	28	4,5	≈ 1	21	10	88
354636	33	2,4	≈ 1	27	3,9	≈ 1	18*	8,1	35
354637	31	6,4	1	30	5,7	≈ 1	24	9,9	112
354639	30	5,7	1	31	5,4	≈ 1	24	9,6	128
354641	27	7,1	1	27	6,2	≈ 1	24	9,4	47
354642	29	6,5	1	28	5,7	≈ 1	24	9,7	227
354643	31	3,5	≈ 1	26	4,6	≈ 1	21	9,1	17
354644	29	4,1	1	30	3,8	≈ 1	21	9,8	31
354645	26	8,1	1	27	6,3	≈ 1	24	10,0	99
354646	29	5,2	1	29	4,2	≈ 1	25	9,9	34
354647	31	4,9	1	31	3,9	≈ 1	23	9,6	88
354649	27	9,0	2	27	7,5	≈ 1	27	8,6	498
354651	27	5,5	1	27	4,3	≈ 1	22	9,5	48
354652	30	5,0	1	30	3,9	≈ 1	21	9,0	43

Tabela A.5: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 6$) (continuação)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
354661	29	6,0	1	28	5,3	≈ 1	24	6,7	18
354663	29	8,3	1	27	6,9	≈ 1	25	9,9	271
354665	30	5,1	1	27	5,2	≈ 1	21	9,5	23
354671	33	5,4	≈ 1	27	6,1	≈ 1	22	9,4	25
354672	29	6,2	1	28	5,7	≈ 1	25	9,8	42
354673	29	6,1	1	29	5,1	≈ 1	21	9,9	22
354674	27	2,8	≈ 1	27	2,2	≈ 1	18*	9,9	105
354679	29	7,0	1	28	5,8	≈ 1	24	9,0	77
354681	29	6,2	1	27	5,3	≈ 1	24	9,3	10
354682	22	0,3	≈ 1	21	0,2	≈ 1	15*	0,9	10
354683	30	4,3	1	29	4,2	≈ 1	22	9,7	9
354684	30	6,0	1	28	5,4	≈ 1	23	10,0	40
354685	32	5,4	1	27	5,5	≈ 1	22	9,9	31
354686	27	6,6	1	27	6,0	≈ 1	24	9,1	77
354687	27	7,8	1	27	6,1	≈ 1	24	9,5	384
354689	26	6,7	1	27	6,2	≈ 1	25	9,0	157
354691	27	2,4	1	27	1,8	≈ 1	19	9,5	22
354692	30	1,9	≈ 1	25	2,4	≈ 1	18	6,8	11
354693	28	4,0	1	28	3,3	≈ 1	20	9,3	97
354711	33	5,5	2	32	6,1	≈ 1	27	10,0	412
354712	26	6,2	1	27	6,1	3	24	9,7	6.184
354713	29	3,7	1	26	4,4	≈ 1	22	7,9	747
354741	30	6,8	1	28	6,4	≈ 1	24	9,5	196
354742	27	6,7	1	27	6,1	≈ 1	24	9,4	914
354743	26	5,7	1	27	5,3	≈ 1	25	8,0	1.755
354744	31	7,4	2	29	7,2	4	27	9,3	6.888
354751	27	6,1	1	27	8,2	≈ 1	24	9,9	2.455
354752	30	7,0	1	29	7,3	≈ 1	27	9,8	501

Tabela A.5: Tamanho Amostral, Coeficiente de Variação e Tempo de Processamento obtidos por Algoritmo para as 75 populações da PAC ($L = 6$) (continuação)

Código do	LH88			KOZAK04			EVP		
Estrato Natural	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)	n	CV(%)	Tempo(s)
354753	31	3,9	1	29	4,3	≈ 1	22	9,6	538
354754	30	7,3	1	29	6,6	1	24	9,9	2.160
354755	29	6,2	1	29	6,9	≈ 1	25	9,4	738
354756	27	4,5	1	27	4,3	≈ 1	21*	9,0	760
354757	28	5,6	≈ 1	28	5,1	≈ 1	23	9,8	298
354759	28	6,4	1	28	6,4	≈ 1	24	9,7	602
354761	28	6,3	1	28	6,4	≈ 1	24	9,7	1.732
354762	28	4,5	≈ 1	27	4,2	≈ 1	21*	8,2	309
354763	29	5,1	2	29	5,0	≈ 1	22	9,9	2.001
354771	32	6,0	1	28	6,1	1	27	9,4	6.993
354772	31	6,8	1	27	7,5	≈ 1	27	9,6	2.868
354773	28	6,2	1	27	6,1	≈ 1	23*	10,0	18.548
354774	27	6,5	1	27	5,5	≈ 1	24	8,9	381
354781	39	6,3	3	29	7,3	5	28	10,0	5.976
354782	31	7,6	1	32	6,3	≈ 1	26	9,7	688
354783	28	6,8	≈ 1	28	5,1	≈ 1	24	10,0	159
354784	27	4,9	1	27	4,3	≈ 1	24	9,0	253
354785	27	4,1	≈ 1	27	4,5	≈ 1	21*	9,6	416
354789	31	5,8	2	27	7,5	3	27	9,7	5.320

N/A: Não se Aplica

* Mínimo Amostral Ótimo

≈ 1: Aproximadamente 1 segundo

Tabela A.6: Eficiência relativa entre os métodos por número de estratos para as 75 populações da PAC

Código do	$EFF_{LH88/EVP}$				$EFF_{KOZAK04/EVP}$			
Estrato Natural	L=3	L=4	L=5	L=6	L=3	L=4	L=5	L=6
35461	171	132	N/A	N/A	100	105	110	113
35462	100	104	135	120	109	107	109	120
35466	102	104	141	117	N/A	N/A	105	117
35472	100	109	N/A	N/A	100	100	100	96
35473	103	122	N/A	N/A	100	100	116	108
354511	101	105	108	104	108	100	100	100
354512	104	112	121	129	104	112	116	114
354530	100	105	138	122	103	102	104	100
354541	102	103	123	117	102	100	109	113
354542	108	N/A	N/A	N/A	108	100	100	N/A
354631	104	116	142	145	107	116	121	127
354632	100	110	135	148	100	105	110	129
354633	102	103	113	129	107	103	113	121
354634	102	104	105	N/A	105	100	100	113
354635	103	116	105	133	N/A	N/A	105	133
354636	112	135	165	183	112	106	141	150
354637	102	104	122	129	106	107	113	125
354639	102	111	109	125	N/A	104	109	129
354641	102	108	118	113	107	104	100	113
354642	102	104	118	121	N/A	N/A	114	117
354643	108	117	137	148	N/A	117	126	124
354644	104	116	109	138	N/A	116	109	143
354645	102	107	118	N/A	108	104	105	113
354646	100	132	114	116	N/A	105	114	116
354647	103	110	129	135	N/A	105	124	135
354649	101	105	113	100	107	100	104	100
354651	104	124	116	123	N/A	124	116	123
354652	100	140	165	143	100	105	125	143

Tabela A.6: Eficiência relativa entre os métodos por número de estratos para as 75 populações da PAC (continuação)

Código do	$EFF_{LH88/EVP}$				$EFF_{KOZAK04/EVP}$			
Estrato Natural	L=3	L=4	L=5	L=6	L=3	L=4	L=5	L=6
354661	100	111	109	121	109	111	105	117
354663	100	103	114	116	107	106	105	108
354665	100	153	174	143	N/A	105	116	129
354671	100	110	133	150	103	110	114	123
354672	102	104	109	116	109	100	109	112
354673	103	110	114	138	106	100	114	138
354674	114	121	138	150	107	121	138	150
354679	100	104	109	121	106	100	104	117
354681	100	120	120	121	106	100	120	113
354682	156	145	N/A	N/A	122	118	123	140
354683	104	126	132	136	N/A	105	132	132
354684	100	124	123	130	106	105	105	122
354685	100	114	123	145	105	110	105	123
354686	102	108	109	113	102	N/A	100	113
354687	100	103	100	113	102	100	100	113
354689	100	N/A	N/A	N/A	104	100	100	108
354691	117	121	138	142	117	121	138	142
354692	107	143	156	167	107	143	125	139
354693	110	133	153	140	110	106	121	140
354711	102	103	121	122	N/A	N/A	117	119
354712	102	N/A	N/A	N/A	100	N/A	100	113
354713	104	112	126	132	N/A	112	121	118
354741	100	114	114	125	106	104	114	117
354742	102	108	109	113	100	100	100	113
354743	100	N/A	N/A	N/A	102	100	100	108
354744	101	100	133	115	103	100	108	107
354751	102	135	100	113	100	100	105	113
354752	102	103	127	111	107	103	114	107
354753	104	110	142	141	N/A	110	126	132

Tabela A.6: Eficiência relativa entre os métodos por número de estratos para as 75 populações da PAC (continuação)

Código do	$EFF_{LH88/EVP}$				$EFF_{KOZAK04/EVP}$			
Estrato Natural	L=3	L=4	L=5	L=6	L=3	L=4	L=5	L=6
354754	102	106	133	125	105	100	104	121
354755	102	111	129	116	104	107	104	116
354756	100	129	116	129	103	106	116	129
354757	102	105	114	122	100	100	110	122
354759	102	104	114	117	102	100	105	117
354761	102	107	118	117	N/A	N/A	105	117
354762	104	112	126	133	104	124	116	129
354763	103	119	138	132	N/A	119	114	132
354771	100	104	123	119	N/A	N/A	123	104
354772	102	103	126	115	106	100	104	100
354773	105	136	114	122	103	105	100	117
354774	105	110	105	113	102	100	105	113
354781	100	105	136	139	N/A	N/A	112	104
354782	100	100	108	119	108	103	104	123
354783	102	119	110	117	102	100	110	117
354784	107	135	100	113	N/A	100	100	113
354785	107	112	121	129	100	106	116	129
354789	102	111	113	115	102	104	104	100

N/A: Não se Aplica

Tabela A.7: Medidas de Posição para uma Simulação de 100 Réplicas do Algoritmo EVP ($L = 4$)

População	Moda	Mínimo	Q1	Mediana (Q2)	Q3	Máximo
U06	12	12	12	12	12	13
U09	14	11	13	15	18	27
U11	30	20	26	30	34	50
U13	18	17	18	20	25	39
U25	26	25	28	32	37	84
354541	30	30	30	30	30	30
354711	29	29	29	29	29	30
354742	25	25	25	25	25	26
354743	21	21	21	21	21	21
354774	21	21	21	21	21	21

Tabela A.8: Medidas de Posição para uma Simulação de 100 Réplicas do Algoritmo EVP ($L = 5$)

População	Moda	Mínimo	Q1	Mediana (Q2)	Q3	Máximo
U06	9	8	8	9	9	24
U09	10	8	10	12	14	28
U11	19	13	17	21	28	54
U13	13	11	13	16	19	55
U25	28	16	23	27	31	57
354541	22	22	22	22	22	23
354711	25	24	24	25	26	29
354742	22	22	22	22	22	22
354743	22	22	22	22	22	22
354774	22	21	21	22	22	22

Tabela A.9: Medidas de Posição para uma Simulação de 100 Réplicas do Algoritmo EVP ($L = 6$)

População	Moda	Mínimo	Q1	Mediana (Q2)	Q3	Máximo
U06	7	7	7	7	9	24
U09	8	7	9	10	13	22
U11	14	10	14	18	23	50
U13	12	8	12	14	21	51
U25	17	12	17	20	26	61
354541	24	24	24	24	25	27
354711	28	24	27	28	30	36
354742	24	24	24	24	25	27
354743	24	23	24	24	24	27
354774	24	22	23	24	24	25

B - Código em R do Algoritmo Proposto

```
# cat(
# ”
# EVP(X,L,cv,nhmin,Nmin,imax,tmax,pmax,notbest,range_s,range_b)
# onde
# X: vetor com os valores populacionais
# L: nº de estratos
# cv: coeficiente de variação fixado
# nhmin: tamanho mínimo amostral por estrato
# Nmin: tamanho mínimo populacional por estrato
# imax: nº maximo de iterações
# tmax: nº de vizinhos maximo
# pmax: nº maximo de soluções no POOL
# notbest: nº de iterações sem melhoria
# range_s: intervalo de shaking (distancia máxima do vizinho)
# range_b: intervalo de busca (distancia máxima do vizinho)
# cpu_time: tempo maximo de CPU (medido em segundos)
# ”
# )
evp=function(X,L,cv,nhmin,Nmin,imax,tmax,pmax,notbest,range_s,range_b,cpu_time)
{
library(sampling)
library(MultAlloc)
```