

FORMULAÇÕES DE PROGRAMAÇÃO INTEIRA APLICADAS AO PROBLEMA DE ALOCAÇÃO ÓTIMA DE AMOSTRAS

José André de Moura Brito

ENCE/IBGE

Rua André Cavalcanti, 106, sala 503-A, Santa Teresa - Centro, Rio de Janeiro, RJ

jose.m.brito@ibge.gov.br

Leonardo Silva de Lima

UFPR

Av. Prefeito Lothario Meissner, 632, Jardim Botânico, 80210170 - Curitiba, PR

leonardo.delima@ufpr.br

Gustavo Silva Semaan

INFES/UFF

Estr. João Jasbick, s/n - Dezesete, Santo Antônio de Pádua - RJ

gustavosemaan@id.uff.br

RESUMO

Este artigo traz a proposta de duas formulações de programação inteira, que foram aplicadas à resolução do problema de alocação ótima em amostragem estratificada. Em uma das versões deste problema, busca-se minimizar uma soma ponderada de variâncias relativas associadas com m variáveis de pesquisa, fixado um tamanho de amostra n e que será alocado a L estratos populacionais. As duas formulações foram aplicadas em um conjunto de 21 populações, diversificadas quanto à origem, porte, valores de m e L , contemplando quatro diferentes cenários quanto ao valor do tamanho de amostra n . A partir dos resultados obtidos e de um conjunto de análises, foi possível avaliar a eficácia e viabilidade das duas formulações, frente a uma formulação da literatura.

PALAVRAS CHAVE. Alocação Ótima, Estratificação, Programação Inteira.

Tópicos (PM, OC)

ABSTRACT

This article proposes two integer programming formulations, which were applied to solve the optimal allocation problem in stratified sampling. In one of the versions of this problem, the aim is to minimize a weighted sum of relative variances associated with m research variables, fixing a sample size n and which will be allocated to L population strata. The two formulations were applied to a set of 21 populations, diverse in terms of origin, size, values of m and L , covering four different scenarios regarding the value of sample size n . Based on the results obtained and a set of analyses, it was possible to evaluate the effectiveness and viability of the two formulations, compared to a formulation from the literature.

KEYWORDS. Optimal Allocation. Stratification. Integer Programming.

Paper topics (MP, CO)

1. Introdução

Na era da informação, os órgãos oficiais de estatística como, por exemplo, o IBGE e o United States Census Bureau, vêm disponibilizando, sistematicamente, grandes quantidades de dados e estatísticas, sendo boa parte dessas informações obtidas e produzidas a partir da realização de suas pesquisas amostrais e levantamentos censitários. Destaca-se, ainda, que o acesso a essas informações é de suma importância para os governos, no que concerne ao planejamento e à implementação de suas políticas públicas, que abarcam um conjunto de ações desenvolvidas de forma a garantir diversos direitos à população em áreas como saúde, educação e segurança, promovendo a qualidade de vida e o bem-estar do povo.

Por sua vez, o planejamento e a consequente aplicação de tais pesquisas têm, por base, a utilização de amostragem probabilística, que constitui-se como um arcabouço metodológico indispensável aos órgãos de estatística oficial, abarcando um conjunto de técnicas que podem ser aplicadas separadamente ou combinadas na elaboração dos planos amostrais utilizados em suas pesquisas. O uso da amostragem probabilística [Lohr, 2021] permite a obtenção de estimativas de totais, médias e proporções relacionadas às variáveis da população alvo da pesquisa, por meio dos dados obtidos a partir uma amostra selecionada de um cadastro, que contém os registros de todos os elementos desta população. Como exemplo de pesquisa amostral temos a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD-Contínua), realizada pelo IBGE [IBGE, 2023]. Nesta pesquisa, uma amostra total de domicílios é coletada em um período de três meses onde, ao final desse ciclo, são produzidas as estimativas dos indicadores desejados.

A definição da técnica de amostragem a ser utilizada na pesquisa depende de diversos fatores, dentre os quais, destacam-se: características da população que será amostrada, recursos disponíveis para a pesquisa, recortes populacionais considerados (ex: unidades da federação, municípios etc) para a produção de estimativas, cadastro de seleção disponível, tempo disponível para a realização da pesquisa e nível de precisão desejado. Assim sendo, ao considerar tais fatores, podem ser adotadas diferentes técnicas de amostragem no planejamento da pesquisa, como, por exemplo, a amostragem por conglomerados, amostragem aleatória simples e amostragem estratificada [Lohr, 2021; Cochran, 1977].

Em particular, quando os responsáveis pelo planejamento amostral têm, como demanda, elaborar um plano que permita a produção estimativas (a partir da amostra) com um bom nível de precisão (menor variância possível), dado um orçamento limitado, uma alternativa frequentemente utilizada diz respeito à aplicação de uma amostragem estratificada. A estratificação permite explorar informações auxiliares prévias disponíveis para todas as N unidades populacionais, formando grupos de unidades homogêneas (estratos) e, em seguida, realizar uma amostragem independentemente dentro de cada um dos estratos [Lohr, 2021].

Ao aplicar a amostragem estratificada, em uma primeira fase devem ser determinados L estratos populacionais, mediante as informações disponíveis em um cadastro, sendo cada estrato constituído por N_h ($h = 1, \dots, L$) elementos da população e, na segunda fase, deve-se determinar os tamanhos de amostra n_h ($h = 1, \dots, L$) que serão alocados a cada um dos L estratos.

A determinação de tais tamanhos de amostra diz respeito à resolução do problema de alocação ótima de amostra. Neste problema, quando o interesse é restrito à determinação de estimativas de total (média, proporção etc) para uma única variável de pesquisa, pode-se utilizar, por exemplo, a alocação de Neyman [Lohr, 2021; Cochran, 1977] ou Potência [Bankier, 1988], de forma a determinar os valores de n_h , definindo-se, assim, o problema de alocação univariada.

A despeito de serem raras as pesquisas para as quais há uma única variável para fazer estimativas, a alocação de Neyman, baseada em determinada variável, pode ser utilizada considerando a sua correlação positiva com as demais variáveis da pesquisa. Todavia, quando devem ser produzi-

das estimativas com diferentes níveis de precisão para um conjunto de m variáveis e tais variáveis não estão fortemente correlacionadas, torna-se necessário o uso de um método que incorpore, no cálculo dos tamanhos de amostra, informações de todas as variáveis, o que implica resolver o problema de alocação ótima multivariada. Em particular, este problema pode ser abordado em duas versões, quais sejam: **(i)** Determinar os tamanhos de amostra n_h ($h = 1, \dots, L$) alocados aos L estratos populacionais, de forma a minimizar uma soma ponderada de variâncias relativas (nível de precisão), associadas às m variáveis de pesquisa, fixado previamente um tamanho total de amostra n ; **(ii)** Minimizar o tamanho total de amostra (somadas dos tamanhos de amostra n_h alocados aos estratos) sujeito a um nível de precisão pré-fixado para cada uma das m variáveis.

As duas versões do problema podem ser mapeadas em problemas de otimização não-linear inteira, havendo diversas abordagens da literatura que tratam do problema de forma direta, ou seja, considerando sua estrutura de não linearidade. Alternativamente a tais abordagens, propõe-se, neste artigo, duas formulações de programação linear inteira que tratam da primeira versão e incorporam, basicamente, uso de variáveis binárias e de linearização.

Além da introdução, este trabalho está dividido da seguinte forma: Na seção 2 são apresentados conceitos básicos de amostragem estratificada e a descrição do problema de alocação ótima. A seção 3 traz a apresentação das duas formulações propostas e na seção 4 são descritos os experimentos computacionais e análises realizadas a partir da comparação dos resultados produzidos pelas duas formulações e de uma formulação da literatura. Por fim, são apresentadas algumas conclusões e trabalhos futuros.

2. Amostragem Estratificada e Problema de Alocação Ótima Multivariada

Na amostragem estratificada [Lohr, 2021] uma população U formada por N unidades populacionais é dividida em L estratos representados por $U_1, \dots, U_h, \dots, U_L$ e compostos, respectivamente, por $N_1, \dots, N_h, \dots, N_L$ unidades populacionais. Os estratos formados cumprem as seguintes propriedades:

$$U_h \cap U_k = \emptyset, \quad h, k = 1, \dots, L \quad (h < k) \quad (1)$$

$$\cup_{h=1}^L U_h = U \quad (2)$$

$$N_1 + \dots + N_h + \dots + N_L = N \quad (3)$$

Uma população pode ser definida por N unidades correspondentes, por exemplo, a estabelecimentos comerciais, escolas e setores censitários¹. Sobre estas unidades é aplicada uma pesquisa por amostragem, ou seja, realizam-se entrevistas com uma parte da população, de forma a levantar informações relativas às variáveis de interesse associadas a tais unidades. A título de exemplificação, se a população é constituída por escolas, podem ser levantados quesitos como número de alunos e professores, número de salas de aula, número de laboratórios etc.

Em geral, tais estratos são definidos utilizando-se uma variável de estratificação X , cujos valores são conhecidos para todas as unidades de U , tal que X é representada por um vetor de N observações $X = (x_1, \dots, x_i, \dots, x_N)$. Uma vez definidos os estratos $U_1, \dots, U_L$, são obtidos os valores de N_h e S_{hx}^2 ($h = 1, \dots, L$) que correspondem, respectivamente, ao total de unidades e à variância populacional (em relação à variável X) em cada um dos L estratos, definida conforme a equação (4), em que $\bar{X}_h$ corresponde à média no h -ésimo estrato.

¹Segundo o IBGE, um setor censitário corresponde a uma unidade territorial estabelecida para fins de controle cadastral, formado por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios que permitam o levantamento por um recenseador - <https://censo2010.ibge.gov.br/materiais/guia-do-censo/operacao-censitaria.html>

$$S_{hx}^2 = \frac{1}{N_h - 1} \sum_{\forall i \in U_h} (x_i - \bar{X}_h)^2, \quad h = 1, \dots, L \quad (4)$$

O procedimento de definir os estratos populacionais utilizando a variável de estratificação X corresponde a resolver o problema de estratificação univariada [Brito et al., 2023; Hidiogrou e Kozak, 2018]. Em uma etapa posterior deste problema, após a obtenção dos valores de N_h e S_{hx}^2 , deve-se determinar os tamanhos de amostra n_h ($h = 1, \dots, L$) que serão alocados aos L estratos, ou seja, o quantitativo de elementos da população que serão selecionados para fazerem parte da amostra da pesquisa; o que corresponde a resolver o problema de alocação ótima. Para definir os valores de n_h é considerada a restrição definida na Equação (5), em que n_{min} e N_h correspondem, respectivamente, aos tamanhos mínimo e máximo de amostra no h -ésimo estrato².

$$n_{min} \leq n_h \leq N_h, \quad h = 1, \dots, L \quad (5)$$

Além disso, considerando que na pesquisa os dados são coletados para todas as unidades da amostra, sendo utilizados para produzir estimativas de total, média etc para um conjunto de m variáveis de pesquisa, denotadas por $Y_1, \dots, Y_j, \dots, Y_m$, a variância de cada variável Y_j é definida pela equação (6):

$$S_{hj}^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_{ij} - \bar{Y}_{hj})^2, \quad j = 1, \dots, m \quad (6)$$

sendo y_{ij} correspondente ao valor da variável Y_j para i -ésima unidade da população e $\bar{Y}_{hj}$ é a média das observações desta variável no h -ésimo estrato, tal que:

$$\bar{Y}_{hj} = \frac{1}{N_h} \sum_{\forall i \in U_h} y_{ij}, \quad h = 1, \dots, L, \quad j = 1, \dots, m \quad (7)$$

De acordo com [Cochran, 1977] e [Lohr, 2021], ao utilizar uma amostragem estratificada simples, define-se a variância do estimador de total t_{Y_j} para cada variável Y_j como:

$$V(t_{Y_j}) = \sum_{h=1}^L N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hj}^2 \quad (8)$$

Os valores de S_{hj}^2 também são calculados na etapa estratificação da variável X , uma vez que cada unidade da população do cadastro tem um valor associado à variável X e valores correspondentes para as variáveis $Y_1, \dots, Y_m$. Adicionalmente, para calcular os valores de $V(t_{Y_j})$, $j = 1, \dots, m$, deve-se determinar os tamanhos de amostra n_h alocados aos estratos.

Esta alocação é fundamental, pois permite à equipe de planejamento controlar a precisão das estimativas derivadas das m variáveis de pesquisa e considerar os seus custos de realização. Alternativamente, a partir da expressão da equação (8), pode-se adotar uma medida de precisão relativa, qual seja, a variância relativa do estimador de total, sendo $t_{Y_j} = \sum_{i=1}^N y_{ij}$ ($j = 1, \dots, m$) dada por:

$$\frac{V(t_{Y_j})}{t_{Y_j}^2}, \quad j = 1, \dots, m \quad (9)$$

²Neste artigo, seguindo o que está preconizado na literatura, o valor 2 foi adotado como o tamanho mínimo de amostra por estrato, mas tal valor pode ser alterado de forma a atender a demandas específicas de uma pesquisa.

A adoção desta medida é necessária, tendo em vista que, em muitas pesquisas baseadas em amostragem estratificada, algumas das variáveis de pesquisa apresentam diferentes ordens de grandeza, não fazendo sentido, desta forma, somar suas variâncias absolutas. Apresenta-se, a seguir, a formulação do problema.

2.1. Formulação do Problema de Alocação

O problema de alocação pode ser visto como um problema da mochila não linear inteiro (função objetivo não-linear e restrições lineares) conforme mencionado em Bretthauer e Shetty [2002]. Considere uma variação da formulação matemática do problema de alocação apresentada por Bretthauer e Shetty [2002], de acordo com as equações de (10) a (15). As variáveis de decisão são inteiras e dadas por n_h e representam os tamanhos das amostras para cada estrato populacional U_h ($h = 1, \dots, L$).

$$\text{Minimizar } \sum_{j=1}^m w_j \frac{1}{t_{Y_j}^2} \sum_{h=1}^L N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hj}^2 \quad (10)$$

Sujeito a:

$$n_1 + \dots + n_h + \dots + n_L = n \quad (11)$$

$$2 \leq n_h \leq N_h, h = 1, \dots, L \quad (12)$$

$$\sum_{j=1}^m w_j = 1 \quad (13)$$

$$0 < w_j < 1, j = 1, \dots, m \quad (14)$$

$$n_h \in \mathbb{Z}_+, h = 1, \dots, L \quad (15)$$

Os valores de w_j (pesos) são um dos parâmetros de entrada do modelo acima e são definidos previamente por especialistas no assunto ou pelos criadores da pesquisa, de forma cumprir as restrições (13) e (14). Se tais valores não forem especificados a priori, pesos iguais podem ser atribuídos a todas as m variáveis da pesquisa, ou seja, $w_j = 1/m$, $j = 1, \dots, m$. A função objetivo a ser minimizada em (10) corresponde à soma ponderada das variâncias relativas do estimador de total das m variáveis de pesquisa.

A restrição (11) garante que a soma dos tamanhos de amostra n_h alocados aos estratos é igual ao tamanho total de amostra n fixado a priori; a restrição (12) garante que o tamanho de amostra alocado a cada estrato é maior ou igual ao tamanho mínimo de amostra definido a priori e que tal tamanho de amostra não exceda o total populacional no estrato. Por fim, a restrição (15) garante que os tamanhos de amostra alocados aos estratos sejam inteiros.

Analisando esta formulação, pode-se observar que, no processo de alocação, há a necessidade de buscar-se o equilíbrio entre a precisão (qualidade das estimativas), função da variabilidade traduzida pelas variâncias das m variáveis e o custo da pesquisa, que está intrinsecamente associado aos tamanhos de amostra n_h alocados aos L estratos. Adicionalmente, a formulação apresenta não linearidade em sua função objetivo e têm variáveis inteiras (tamanhos de amostra).

O problema de alocação ótima multivariada foi primeiramente estudado por Neyman [1934] no desenvolvimento de sua teoria para resolver o problema de alocação univariada. A importância deste problema e a sua complexidade computacional, tendo em vista ser um problema de programação não linear inteiro, têm motivado, nas últimas décadas, o desenvolvimento

de vários trabalhos com abordagens baseadas em programação linear inteira, programação convexa, programação não linear, programação dinâmica, programação multiobjetivo, heurísticas e metaheurísticas. Dentre os diversos trabalhos encontrados na literatura que abordam a versão (i) considerada neste trabalho, destacam-se: [Kish, 1976; Khan e Ahsan, 2003; García e Cortez, 2006, 2008; Khan et al., 2011, 2012; Ali et al., 2013; Raghav et al., 2014; Varshney et al., 2015; Varshney, 2019; Friedrich et al., 2018; Brito et al., 2015, 2022].

Alguns destes trabalhos trazem métodos que resolvem o problema de forma direta no que concerne à questão de não linearidade, mais especificamente, utilizando programação não linear ou convexa [Bazaraa et al., 2006]. Todavia, a aplicação de tais métodos tende a produzir, em geral, soluções correspondentes a tamanhos de amostra n_h não inteiros, cujo arredondamento conduz à perda de otimalidade quanto à função objetivo definida na equação (10) ou de viabilidade quanto às restrições (11) ou (12) [Brito et al., 2015].

Em [García e Cortez, 2006] são propostas algumas abordagens para o problema, usando diferentes tipos de função objetivo associadas com as variâncias das m variáveis de pesquisa, e diferentes métodos de resolução, tais como lexicográfico, ϵ restrito e da distância. Em [Khan et al., 2011] os autores resolvem o problema de alocação utilizando uma formulação de programação convexa com uma função objetivo não linear, que é resolvida via técnica de aproximação de Chebyshev. Em [Khan e Ahsan, 2003] propõe-se uma formulação de programação não linear para o problema definido na Seção 2.1, sendo proposta uma solução baseada na aplicação de um algoritmo de programação dinâmica. Em [Varshney, 2019] propõe-se uma solução para o problema da Seção 2.1 baseado em uma função Gamma e usando um algoritmo de programação por metas lexicográficas.

Em [Raghav et al., 2014] os autores também resolvem o problema de alocação na presença de não-resposta, considerando a resolução de um problema multiobjetivo não linear inteiro. Em [Brito et al., 2015] os autores propuseram duas formulações de programação binária para resolver as duas versões do problema de alocação e cujo número de variáveis é proporcional ao tamanho (N) da população. Por fim, [Brito et al., 2022] propuseram uma heurística para a resolução das duas versões do problema de alocação multivariada. Tal heurística combina conceitos da metaheurística BRKGA com funções de penalidade.

3. Formulações Propostas

Nesta seção são apresentadas duas novas formulações de programação linear inteira para o problema definido de (10)-(15) e que tiveram, por base, os trabalhos de [Billionnet e Djebali, 2006; Bretthauer e Shetty, 2002] e fazem uso de variáveis binárias e de uma linearização. Tais formulações garantem o ótimo global correspondente à solução do problema de alocação.

3.1. Formulação - A (F_A)

Nesta formulação, representa-se cada variável n_h correspondente ao tamanho de amostra alocado ao h -ésimo estrato como uma soma de variáveis binárias, tal que:

$$n_h = \sum_{i=0}^{q_h} 2^i x_i^h, \quad 2^{q_h} \leq U_h < 2^{q_h+1}, \quad (16)$$

sendo $U_h = \min(n - 2(L - 1), N_h)^3$ e $x_i^h \in \{0, 1\}$, $h = 1, \dots, L$.

Cada variável n_h presente no denominador da função objetivo apresentada em (10) produz não linearidade no problema. Assim sendo, para efetuar a linearização da função, pode-se definir $z^h = \frac{1}{n_h}$ e combinar esta relação com a equação (16), produzindo-se as seguintes restrições:

³O uso de tal expressão garante o tamanho mínimo de amostra em cada estrato.

$$z^h \times \sum_{i=0}^{q_h} 2^i x_i^h = 1, \quad h = 1, \dots, L \quad (17)$$

Como as variáveis x_i^h são binárias e as variáveis z^h são contínuas, torna-se necessário aplicar uma linearização nos termos $x_i^h \times z^h$. Para isso, utilizou-se o que está proposto em [Billionnet e Djebali, 2006], definindo-se variáveis t_i^h tais que $t_i^h = x_i^h \times z^h$ e considerando as seguintes restrições:

$$t_i^h \geq z^h - z_{sup}^h(1 - x_i^h) - z_{inf}^h x_i^h, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h; \quad (18)$$

$$t_i^h \leq z^h - z_{inf}^h, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h; \quad (19)$$

$$t_i^h \leq (z_{sup}^h - z_{inf}^h)x_i^h, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h, \quad (20)$$

em que z_{sup} e z_{inf} correspondem, respectivamente, aos limitantes inferior e superior definidos a priori para cada variável z^h . Aqui, considerando que $z^h = \frac{1}{n_h}$, temos que z^h assume valor máximo quando $n_h = n_{min}$ e assume valor mínimo quando $n_h = U_h$, ou seja, $z_{inf} = \frac{1}{U_h}$ e $z_{sup} = \frac{1}{n_{min}}$.

Em seguida, efetuando-se a substituição de cada produto por este conjunto de restrições e utilizando as equações (16) e (17), define-se a seguinte formulação de programação linear inteira mista, em que a função objetivo em (21) correspondente à função de (10); a restrição (22) corresponde à restrição (11) e a restrição (23) corresponde à restrição (12).

$$\text{Minimizar} \sum_{j=1}^m w_j \frac{1}{t_{Y_j}^2} \sum_{h=1}^L N_h^2 z^h - N_h S_{hj}^2 \quad (21)$$

Sujeito a:

$$\sum_{h=1}^L \sum_{i=0}^{q_h} 2^i x_i^h = n \quad (22)$$

$$\sum_{i=1}^{q_h} x_i^h \geq 1, \quad h = 1, \dots, L \quad (23)$$

$$\sum_{i=0}^{q_h} 2^i t_i^h = 1, \quad h = 1, \dots, L \quad (24)$$

$$t_i^h \geq z^h - \frac{1}{n_{min}}(1 - x_i^h) - \frac{1}{U_h} x_i^h, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h \quad (25)$$

$$t_i^h \leq z^h - \frac{1}{U_h}, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h \quad (26)$$

$$t_i^h \leq \left(\frac{1}{n_{min}} - \frac{1}{U_h} \right) x_i^h, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h \quad (27)$$

$$x_i^h \in \{0, 1\}, \quad h = 1, \dots, L, \quad i = 0, \dots, q_h \quad (28)$$

$$z^h \in \mathbb{R}_+, \quad h = 1, \dots, L. \quad (29)$$

3.2. Formulação - B (F_B)

Esta formulação é baseada no trabalho de [Bretthauer e Shetty, 2002]. Inicialmente, considere o problema a seguir em que f_i e g_i são funções convexas, diferenciáveis e não lineares:

$$(P1) \text{ Minimizar } \sum_{i=1}^r f_i(s_i) \quad (30)$$

Sujeito a:

$$\sum_{i=1}^r g_i(s_i) \leq b \quad (31)$$

$$l_i \leq s_i \leq u_i, \quad i = 1, \dots, r \quad (32)$$

$$s_i \in \mathbb{Z}_+, \quad i = 1, \dots, r \quad (33)$$

Este problema pode ser convertido em um problema de programação binária. Mais especificamente, para $i = 1, \dots, r$ e $j = 1, \dots, u_i - l_i$ define-se $p_{ij} = f_i(l_i + j) - f_i(l_i + j - 1)$ e $q_{ij} = g_i(l_i + j) - g_i(l_i + j - 1)$. Estes termos são utilizados como coeficientes do seguinte problema de programação binária (P2), equivalente ao problema (P1) definido por (30)-(33).

Uma vez obtida a solução para o problema (P2), ou seja, os valores das variáveis x_{ij} , a solução de (P1) pode ser obtida utilizando a seguinte expressão: $s_i = l_i + \sum_{j=1}^{u_i-l_i} x_{ij}$ para $i = 1, \dots, r$.

$$(P2) \text{ Minimizar } \sum_{i=1}^r \sum_{j=1}^{u_i-l_i} p_{ij} x_{ij} + \sum_{i=1}^r f_i(l_i) \quad (34)$$

Sujeito a:

$$\sum_{i=1}^r \sum_{j=1}^{u_i-l_i} q_{ij} x_{ij} \leq b - \sum_{i=1}^r g_i(l_i) \quad (35)$$

$$x_{ij} \in \{0, 1\}, \quad i = 1, \dots, r, \quad j = 1, \dots, u_i - l_i \quad (36)$$

Analisando-se (P1), observa-se uma correspondência com a formulação definida por (10)-(15), uma vez que, a função objetivo em (10) corresponde a uma soma de funções objetivo convexas, não-lineares e com variáveis inteiras. Além disso, as restrições definidas em (11), (12) e (15) podem ser mapeadas, respectivamente, nas restrições definidas em (31), (32) e (33).

Em específico, define-se $g_h(n_h) = n_h$, $h = 1, \dots, L$, ou seja, cada função g_h corresponde a cada função $g_i(x_i)$, o termo b corresponde ao tamanho total de amostra n e os termos 2 e U_h correspondem, respectivamente, aos termos l_i e u_i . Por fim, definimos $f_h(n_h) = C_Y^h N_h^2 (\frac{1}{n_h} - \frac{1}{N_h})$, tal que $C_Y^h = \sum_{j=1}^m w_j \frac{1}{t_{Yj}^2} S_{hj}^2$, em que $f_h(n_h)$ tem correspondência com a função $f_i(s_i)$.

Considerando tal correspondência entre estes problemas e aplicando a transformação definida por (34)-(36), chega-se à seguinte formulação de programação inteira 0-1.

$$\text{Minimizar } \sum_{h=1}^L \sum_{j=1}^{U_h-2} [f_h(2+j) - f_h(2+j-1)] x_{hj} + \sum_{h=1}^L f_h(2) \quad (37)$$

$$\sum_{h=1}^L \sum_{j=1}^{U_h-2} x_{hj} = n - 2L \quad (38)$$

$$x_{hj} \in \{0, 1\}, \quad h = 1, \dots, L, \quad j = 1, \dots, U_h - 2 \quad (39)$$

A Tabela 1 a seguir traz o número de variáveis e restrições das formulações F_A e F_B e da formulação proposta por Brito et al. [2015] e denominada BSSM_FD.

Tabela 1: Número de Variáveis e Restrições das Formulações

Formulação	Variáveis	Restrições
A	$L + 2 \cdot \sum_{h=1}^L (q_h + 1)$	$1 + 2L + 3 \cdot \sum_{h=1}^L (q_h + 1) + L$
B	$\sum_{h=1}^L U_h - 2L$	1
BSSM.FD	$\sum_{h=1}^L N_h - 2L$	$L + 1$

4. Experimentos Computacionais

A presente seção traz os resultados computacionais obtidos a partir da aplicação das duas formulações apresentadas na Seção 3 e da formulação proposta por [Brito et al., 2015]. As três formulações foram implementadas em linguagem R⁴, utilizando o pacote Gurobi (disponível no R) com rotinas do solver GUROBI versão 11.0. Todos os experimentos foram realizados em um computador com 128GB de memória RAM e dotado de processador Intel i9-13900F de 5.60 GHz e com sistema operacional Windows 11 (64 bits).

Para avaliar as formulações foram utilizadas 21 populações, sendo a maioria também utilizada em [Brito et al., 2015; Friedrich et al., 2018; Brito et al., 2022] e cujas informações são apresentadas na Tabela 2. Nesta tabela, a coluna Pop traz o nome da população e as colunas rotuladas por N , L e m contêm, respectivamente, as informações relativas ao tamanho da população (porte), número de estratos populacionais e número de variáveis de pesquisa. Adicionalmente, a Figura 1 traz os percentuais relativos às correlações (total de 63) de Pearson [Triola, 2021] calculadas entre as variáveis utilizadas em cada uma das 21 populações⁵, indicando que boa parte das variáveis não têm correlação significativa. Tal fato ratifica a necessidade de usar uma abordagem de alocação multivariada, corroborando com o que está posto na introdução.

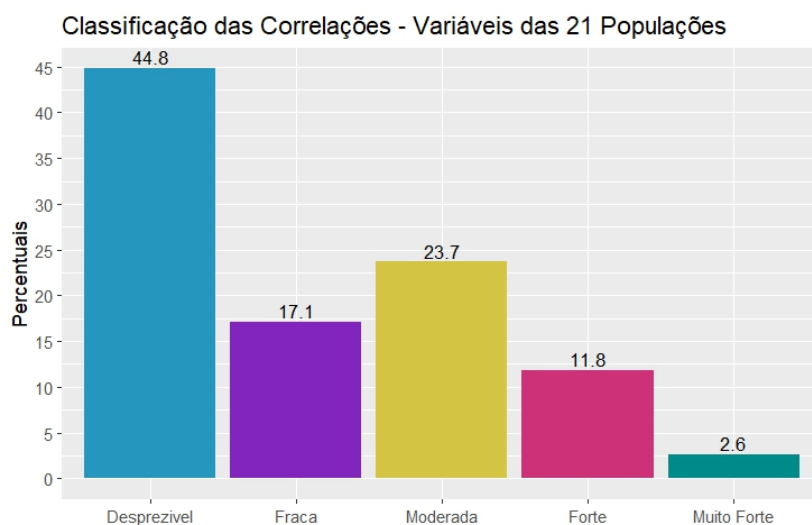
Tabela 2: Populações Utilizadas nos Experimentos

Pop	N	L	m	Pop	N	L	m
Amelia	1.631.293	103	3	Enem_mun_BA	31.279	397	3
Basesintetica	28.286	2	2	Enem_mun_MS	8.865	75	3
bethel	1.000	8	3	Enem_mun_PA	24.009	138	3
Cafe	20.472	3	3	Enem_mun_SC	22.176	251	3
CensoSP_APON	367.308	633	3	Inep_Nordeste	75.084	12	2
CensoSP_MUN	367.308	39	3	Inep_Norte	23.945	12	2
Census2021	35.672	6	3	Inep_Sudeste	59.778	12	2
Dados_educ_sup_2023_uf31	12.678	342	6	Inep_Sul	25.911	12	2
Dados_educ_sup_2023_ufs	178.246	27	6	MunicSu	2.896	7	4
ENAP_TUDO	194.695	12	2	Producao_Municipal	4.826	6	2
				Setores_MG_CENSO_2022	50.663	853	3

⁴Ambiente RStudio - versão 2023.12.0+369

⁵Em cada uma das populações foram consideradas todas as combinações das m variáveis tomadas duas a duas.

Figura 1: Gráfico com as Correlações



Em relação aos experimentos realizados com as 21 populações, foram considerados quatro cenários possíveis, correspondentes a diferentes tamanhos de amostra (n) fixados a priori quanto ao tamanho total da população (N), quais sejam: 5%, 10%, 20% e 30%. Adicionalmente, em cada um destes cenários, as três formulações foram aplicadas nas 21 populações considerando um tempo máximo de processamento de 60 minutos⁶. Os resultados obtidos encontram-se sumarizados na Tabela 3, que traz as estatísticas resumo (mínimo, quartis, média e máximo) dos tempos de processamento (em segundos).

Analisando esta tabela, observa-se que a formulação F_B foi a que teve a melhor performance, demandando, independente do cenário avaliado, tempos de processamento máximos da ordem de 8 segundos (relativo à população Amelia com mais de um milhão e meio de observações), seguida pela formulação BSSM_FD proposta por [Brito et al., 2015], com tempos inferiores a 12 segundos em 75% dos casos e com tempo máximo da ordem de 2,5 minutos relativo à alocação de um tamanho de amostra de 30% na população Amelia.

Tabela 3: Tempos de Processamento das Formulações

Amostra	n=5%N						n=10%N					
	Min	Q ₁	Med	Média	Q ₃	Max	Min	Q ₁	Med	Média	Q ₃	Max
Formulação	0,0	0,3	1,4	10,7	6,9	111,1	0,0	0,5	1,5	9,6	8,5	71,0
BSSM_FD	0,0	0,3	1,4	10,7	6,9	111,1	0,0	0,5	1,5	9,6	8,5	71,0
F_A	0,1	1,1	3605,1	2414,3	3612,1	3622,4	0,0	51,8	3605,1	2418,4	3609,9	3627,0
F_B	0,0	0,0	0,1	0,7	0,6	7,6	0,0	0,0	0,2	0,8	0,7	7,6
Amostra	n=20%N						n=30%N					
	Min	Q ₁	Med	Média	Q ₃	Max	Min	Q ₁	Med	Média	Q ₃	Max
Formulação	0,0	1,1	1,9	14,8	10,8	124,7	0,0	1,2	3,1	18,5	11,5	150,6
BSSM_FD	0,0	1,1	1,9	14,8	10,8	124,7	0,0	1,2	3,1	18,5	11,5	150,6
F_A	0,0	53,4	3605,1	2431,2	3610,6	3632,0	0,0	214,9	3605,4	2592,1	3608,2	3628,0
F_B	0,0	0,1	0,1	0,8	0,8	7,7	0,0	0,1	0,2	0,9	0,8	7,6

Adicionalmente, considerando o tempo máximo de processamento de 1 hora, observou-se que as formulações BSSM_FD e F_B produziram o ótimo global para todas as populações em todos

⁶Em alguns casos, o solver GUROBI terminou a execução e retornou a solução com tempos de processamento ligeiramente superiores a 3600 segundos.

os cenários, enquanto a formulação F_A produziu o ótimo global em cerca de 33% dos casos, em cada um dos cenários.

5. Conclusões e Trabalhos Futuros

O presente trabalho trouxe a proposta de uma formulação de programação linear inteira mista e programação binária para o problema de alocação ótima multivariada, considerando a minimização de uma soma ponderada de variâncias relativas de um conjunto de m variáveis de pesquisa. As duas formulações e uma formulação de programação binária proposta por [Brito et al., 2015] foram aplicadas em 21 populações diversificadas quanto ao porte (valor de N), à origem e ao número de estratos e variáveis de pesquisa, com destaque para as populações Amélia, com mais de 1,5 milhão de observações distribuídas em 103 estratos e CensoSP_APON com quase 368.000 observações e 633 estratos.

A partir dos resultados e análises apresentadas, considerando, concomitantemente, os tempos de processamento e as quantidades de ótimos globais produzidos pelas formulações, conclui-se que a formulação F_B é a mais recomendada para a resolução do problema de alocação multivariada, seguida pela formulação BSSM_FD. A formulação F_A , a despeito do número inferior de variáveis quando comparada a estas formulações, apresenta um número bem maior de restrições, além de demandar tempo de processamento bem superior às formulações F_B e BSSM_FD.

Em trabalhos futuros serão implementadas e avaliadas as formulações associadas à segunda versão do problema de alocação multivariada, qual seja, minimizar o tamanho de amostra, considerando um conjunto de m níveis de precisão previamente fixados e associados às m variáveis de pesquisa. Adicionalmente, objetivando reduzir o tempo de processamento da formulação F_A e, concomitantemente, produzir mais ótimos globais, será considerado o estudo do cálculo de novos valores para os limitantes z_{sup} e z_{inf} , mediante o desenvolvimento de uma heurística baseada no trabalho de [Brito et al., 2022].

Referências

- Ali, I., Raghav, Y., e Bari, A. (2013). Compromise allocation in multivariate stratified surveys with stochastic quadratic cost function. *Journal of Statistical Computation and Simulation*, 83(5): 962–976.
- Bankier, M. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42:174–177.
- Bazaraa, M., Sherali, H., e Shetty, C. (2006). *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York.
- Billionnet, A. e Djebali, K. (2006). Résolution d'un problème combinatoire fractionnaire par la programmation linéaire mixte. *RAIRO Operations Research*, 40:97–111.
- Brethauer, M. K. e Shetty, B. (2002). The nonlinear knapsack problem - algorithms and applications. *European Journal of Operational Research*, 138:459–472.
- Brito, J., Semaan, G., de Lima, L., e Fadel, A. (2023). Heuristic algorithm for univariate stratification problem. *RAIRO*, 57:3033–3060.
- Brito, J., Fadel, A., e Semaan, G. (2022). A genetic algorithm applied to optimal allocation in stratified sampling. *Communications in Statistics - Simulation and Computation*, 51(7):3714–3732.

- Brito, J., Silva, P., Semaan, G., e Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41(2):427–442.
- Cochran, W. (1977). *Sampling Techniques*. New York, John Wiley, 3rd edition.
- Friedrich, U., Münnich, R., e Rupp, M. (2018). Multivariate optimal allocation with box-constraints. *Austrian Journal of Statistics*, 47:33–52.
- García, J. A. D. e Cortez, L. U. (2006). Optimum allocation in multivariate stratified sampling: multi-objective programming. Technical report, Gobierno de México.
- García, J. e Cortez, L. (2008). Multi-objective optimisation for optimum allocation in multivariate stratified sampling. *Survey Methodology*, 34(2):215–222.
- Hidiroglou, M. e Kozak, M. (2018). Stratification of skewed populations: A comparison of optimisation-based versus approximate methods. *International Statistical Review*, 86(1):87–105.
- IBGE (2023). Pesquisa nacional por amostra de domicílios contínua notas técnicas versão 1.15. Notas técnicas 1.15, IBGE.
- Khan, M., Ali, I., e Ahmad, Q. (2011). Chebyshev approximate solution to allocation problem in multiple objective surveys with random costs. *American Journal of Computational Mathematics*, 1:247–251.
- Khan, M., Ali, I., Raghav, Y., e Bari, A. (2012). Allocation in multivariate stratified surveys with non-linear random cost function. *American Journal of Operations Research*, 2:100–105.
- Khan, M. e Ahsan, M. (2003). A note on optimum allocation in multivariate stratified sampling. *The South Pacific Journal of Natural Science*, 21:91–95.
- Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society, Series A*, 139(1):80–95.
- Lohr, S. (2021). *Sampling - Design and Analysis*. Chapman and Hall/CRC, New York.
- Neyman, J. (1934). On the two different aspects of the representative method: The method of representative sampling and the method of purposive sampling. *Journal of the Royal Statistical Society*, 97(4):558–625.
- Raghav, Y. S., Ali, I., e Bari, A. (2014). Multi-objective nonlinear programming problem approach in multivariate stratified sample surveys in the case of non-response. *Journal of Statistical Computation and Simulation*, 84(1):22–36.
- Triola, M. (2021). *Elementary Statistics*. Pearson, 14th edition.
- Varshney, R., Khan, M. G. M., Fatima, U., e Ahsan, M. J. (2015). Integer compromise allocation in multivariate stratified surveys. *Annals of Operations Research*, 226(1):659–668.
- Varshney, R. M. (2019). Optimum allocation in multivariate stratified sampling design in the presence of nonresponse with gamma cost function. *Journal of Statistical Computation and Simulation*, 89(13):2454–67.

Agradecimentos: Ao CNPq - Projeto Universal 405044/2021-6.