UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM DE SISTEMAS COMPLEXOS

RAFAEL PEÇANHA WAISSMAN

Métodos de Alocação Fortuita: extensão para múltiplos grupos e sua avaliação empírica

São Paulo
2022

RAFAEL PEÇANHA WAISSMAN

Métodos de Alocação Fortuita: extensão para múltiplos grupos e sua avaliação empírica

Versão corrigida

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Modelagem de Sistemas Complexos.

Área de concentração: Sistemas Complexos

Orientador: Prof. Dr. Marcelo de Souza Lauretto

Coorientador: Prof. Dr. Julio Michael Stern

São Paulo

2022

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Peçanha Waissman, Rafael
Métodos de Alocação Fortuita: extensão para múltiplos grupos e sua avaliação empírica / Rafael Peçanha Waissman; orientador, Marcelo de Souza Lauretto; coorientador, Julio Michael Stern. -- São Paulo, 2022.
78 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Modelagem de Sistemas Complexos, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.
Versão corrigida

1. Amostragem. 2. Planejamento de experimentos. 3. Otimização. 4. Haphazard Intentional Sampling. I. Lauretto, Marcelo de Souza, orient. II. Stern, Julio Michael, coorient. III. Título.

Dissertação de autoria de Rafael Peçanha Waissman, sob o título **"Métodos de Alocação Fortuita: extensão para múltiplos grupos e sua avaliação empírica"**, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Modelagem de Sistemas Complexos, na área de concentração Sistemas Complexos, aprovada em 13 de Junho de 2022 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Marcelo de Souza Lauretto

Instituição: Universidade de São Paulo

Presidente

Prof. Dr. Alexandre da Silva Freire

Instituição: Universidade de São Paulo

Profa. Dra. Karina Valdivia Delgado

Instituição: Universidade de São Paulo

Prof. Dr. Rion Brattig Correia

Instituição: Instituto Gulbenkian de Ciência

Dedico este trabalho aos meus pais, Maria da Glória e Sérgio.

Agradecimentos

Ao Prof. Marcelo Lauretto, pela constante dedicação e disponibilidade durante a orientação, pelo entusiasmo com este trabalho e pelo imenso aprendizado proporcionado nesse período.

Aos Profs. Julio M. Stern e Alexandre S. Freire, pelos valiosos ensinamentos e pelas contribuições ao desenvolvimento deste trabalho.

Ao Miguel G. Ribeiro, pelo trabalho em equipe no desenvolvimento das publicações.

Aos professores da EACH por, em cada disciplina, me apresentarem fontes de conhecimentos fascinantes e me inspirarem a aplica-los em contextos profundamente gratificantes.

Aos amigos da USP e do TCM, pelo grande apoio, incentivo e motivação à concretização deste trabalho.

À Nic.

Aos meus pais e irmãs.

Muito Obrigado!

Resumo

WAISSMAN, Rafael Peçanha. **Métodos de Alocação Fortuita**: extensão para múltiplos grupos sua avaliação empírica. 2022. 78 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022.

O planejamento de um estudo ou experimento muitas vezes envolve a definição do método de seleção dos elementos que compõem uma amostra ou grupo de alocação, ponto que abrange discussões acerca do papel da aleatorização. O uso da aleatorização é amplamente defendido por duas razões principais: (i) os métodos frequentistas de inferência são baseados na premissa da amostragem aleatória; e (ii) a aleatorização evita qualquer possibilidade de interferência humana (intencional ou não), proporcionando, assim, maior isenção às pesquisas experimentais. Por outro lado, diversos autores questionam esse paradigma, especialmente porque os métodos baseados em aleatorização pura têm alta probabilidade de proverem amostras não representativas da população ou alocações nas quais os grupos experimentais formados tenham perfis significativamente diferentes entre si, o que pode potencialmente comprometer os próprios resultados e conclusões das pesquisas. Métodos já desenvolvidos conciliam uma maior eficiência no balanceamento das covariáveis com uma redução significativa na possibilidade de interferência humana. Uma abordagem conhecida como Alocação Intencional Fortuita busca alcançar esse objetivo por meio de modelos de otimização de distâncias entre as covariáveis de interesse, estendidos com a incorporação de covariáveis artificiais aleatorizadas. Essa abordagem se mostra promissora em proporcionar maior garantia de balanceamento, desacoplamento e poder de inferência. Contudo, os estudos já realizados com esta abordagem analisam apenas a alocação em dois grupos (controle / tratamento). O presente projeto propõe a generalização de métodos já desenvolvidos para os casos de alocações em mais de dois grupos. São apresentados estudos de caso, com discussões acerca do uso prático do método de Alocação Intencional Fortuita no planejamento de experimentos.

Palavras-chaves: Amostragem. Planejamento de experimentos. Otimização. Haphazard Intentional Sampling.

Abstract

WAISSMAN, Rafael Peçanha. **Haphazard Allocation Methods**: extension for multiple groups and empirical evaluation. 2022. 78 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2022.

The design of studies and experiments often involves defining a method for selecting the elements that constitutes a sample or allocation group, an issue that raises debates about the role of randomization. The use of randomization is widely advocated for two main reasons: (i) frequentist inference methods are based on the premise of random sampling; and (ii) randomization avoids any possibility of human interference (intentional or not), providing greater impartiality for experimental research. On the other hand, several authors have refuted this paradigm, particularly because methods based on pure randomization have a high probability of providing samples that are not representative of the population or allocations in which experimental groups profiles are significantly different from each other, which can potentially undermine a research's results and conclusions. Some already developed methods combines greater efficiency in balancing covariates with a significant reduction in the possibility of human interference. An approach known as Haphazard Intentional Allocation seeks to achieve this goal by optimizing distance models between the covariates of interest, expanded with the addition of random artificial covariates. This approach seems promising in providing greater assurance of balance, decoupling, and inference power. However, the studies conducted with this approach have only considered allocations in two groups ("control/treatment" designs). This project proposes the generalization of already developed methods to cases of allocations in more than two groups. Case studies are presented, with discussions on the practical use of the Haphazard Intentional Allocation method in experiments design.

Keywords: Sampling. Design of experiments. Optimization. Haphazard Intentional Sampling.

Lista de figuras

Figura 1 – Exemplo de convergência para o valor esperado de uma variável com distribuição normal padrão. 22

Figura 2 – Exemplo de distribuição correlacionada e após transformação no caso de duas variáveis 28

Figura 3 – Função de Perda de Mahalanobis e tempo de processamento para os métodos Alocação Fortuita (diferentes valores de λ^*), Rerandomização e Aleatorização pura. 34

Figura 4 – Valores de discrepâncias obtidos para média (vermelho) e variância (azul), considerando $n = 20$ e $g = 2$. 35

Figura 5 – Alocações aleatórias dos 54 elementos da base de dados das estações de ar (ver Seção 4.1) em 3 grupos de 18 para as Normas L_1, L_2, $(L_2)^2$ e L_∞. 1000 alocações para cada Norma. 40

Figura 6 – Distância de Mahalanobis e Kappa de Fleiss em relação à λ^* para o método de Alocação Fortuita. 51

Figura 7 – Diferenças entre as médias de cada um dos 3 Grupos (18, 18, 18) e a média global, considerando os métodos Alocaçã Fortuita, Rerandomização e Aleatorização simples. 52

Figura 8 – Diferenças entre as médias de cada um dos 4 Grupos (13, 13, 14, 14) e a média global, considerando os métodos Alocação Fortuita, Rerandomização e Aleatorização simples. 52

Figura 9 – Diferenças entre as médias de cada um dos 5 Grupos (10, 11, 11, 11, 11) e a média global, considerando os métodos Alocação Fortuita, Rerandomização e Aleatorização simples. 53

Figura 10 – Diferenças máximas entre as combinações de pares, para cada método - 3 Grupos (18, 18, 18). 54

Figura 11 – Diferenças máximas entre as combinações de pares, para cada método - 4 Grupos (13, 13, 14, 14). 55

Figura 12 – Diferenças máximas entre as combinações de pares, para cada método - 5 Grupos (10, 11, 11, 11, 11). 55

Figura 13 – Somatório das distâncias entre os pares de grupos e *Gap* relativo, para os tempos (30s, 60s, 120s, 180s, 360s e 1440s). 56
Figura 14 – Distribuição das 133 cidades consideradas. 57
Figura 15 – Tempo necessário para atingir $Gap = 0.01$. 58
Figura 16 – Distância de Mahalanobis e Kappa de Fleiss em relação à λ^* para o método de Alocação Fortuita. - Rorainópolis (34 setores) e Iguatu (176 setores) 59
Figura 17 – Soma das diferenças entre as médias dos dois grupos, para cada método - São Paulo (18.182 setores). 61
Figura 18 – Soma das diferenças entre as médias de cada grupo e a média global, para cada método - Serrana. 63
Figura 19 – Diferenças máximas entre as combinações de pares, para cada método - Serrana. 64
Figura 20 – Taxas de Infecção estimadas e reais após a aplicação das vacinas. . . . 65
Figura 21 – Balanceamento das covariáveis - Rio de Janeiro (10.158 setores). 70
Figura 22 – Balanceamento das covariáveis para cada método - Brasília (4.293 setores). 71
Figura 23 – Balanceamento das covariáveis para cada método - Campos dos Goytacazes (647 setores). 72
Figura 24 – Balanceamento das covariáveis para cada método - Marília (331 setores). 73
Figura 25 – Balanceamento das covariáveis para cada método - Iguatu (176 setores). 74
Figura 26 – Balanceamento das covariáveis para cada método - Cruzeiro do Sul (86 setores). 75
Figura 27 – Balanceamento das covariáveis para cada método - Oiapoque (39 setores). 76
Figura 28 – Balanceamento das covariáveis para cada método - Corrente (35 setores). 77
Figura 29 – Balanceamento das covariáveis para cada método - Rorainópolis (34 setores). 78

Lista de tabelas

Tabela 1 – Distâncias de Mahalanobis e Kappa de Fleiss 100 alocações de 300 segundos para cada método 50
Tabela 2 – Parâmetros λ e tempos de processamento para cada faixa de setores censitários. 60
Tabela 3 – Taxas de Eficácia das Vacinas 62
Tabela 4 – Taxas de Infecção estimadas (%) - Percentis 65

Lista de abreviaturas e siglas

Bin	Distribuição Binomial
IBGE	Instituto Brasileiro de Geografia e Estatística
LTS	Suporte de longo prazo
MDPI	Multidisciplinary Digital Publishing Institute
NO	Monóxido de nitrogênio
NO2	Dióxido de nitrogênio
NOx	Óxido de nitrogênio
O3	Ozônio
PM10	Material particulado de 10 micrômeros
SARS-CoV2	Síndrome Respiratória Aguda Grave Coronavírus 2
Temp	Temperatura do ar
UFPel	Universidade Federal de Pelotas
UR	Umidade relativa
WS	Velocidade do vento

Lista de símbolos

n	Quantidade de elementos
μ	Média populacional
σ	Desvio padrão populacional
$\overline{x}$	Média amostral
s	Desvio padrão amostral
E	Valor esperado
P	Probabilidade
$\mathcal{N}$	Distribuição normal
Bin	Distribuição binomial
$\|\|x\|\|_p$	Norma de x (não especificada)
$\|\|x\|\|_1$	Norma "City block"de x
$\|\|x\|\|_2$	Norma euclideana de x
$\|\|x\|\|_\infty$	Norma infinito de x
$\|\|x\|\|_{(\gamma 1,\gamma 2)}$	Norma Híbrida de x
$\boldsymbol{X}$	Matriz de covariáveis selecionadas
m	Quantidade de covariáveis
$\overline{\boldsymbol{X}}$	Vetor de médias das colunas da matriz X
$\boldsymbol{X}^*$	Matriz de covariáveis selecionadas após transformação
$cov(\boldsymbol{X})$	Matriz de covariância de $\boldsymbol{X}$
$D_M(\boldsymbol{x},\boldsymbol{y})$	Distância de Mahalanobis entre $\boldsymbol{x}$ e $\boldsymbol{y}$
$\mathbf{1}$	Vetor composto por 1 na dimensão apropriada
$\boldsymbol{w}$	Vetor de alocação no caso de dois grupos

M	Função de perda baseada na Distância de Mahalanobis
$\boldsymbol{Z}$	Matriz de covariáveis artificiais
$\boldsymbol{Z}^*$	Matriz de covariáveis artificiais após transformação
λ	Parâmetro de calibração da perturbação adicionada
$\boldsymbol{A}^*$	Matriz de covariáveis selecionadas e covariáveis artificiais após transformação
H	Função de perda baseada na norma híbrida
w_i	Elemento i de $\boldsymbol{w}$
g	Quantidade de grupos
$\boldsymbol{R}$	Matriz auxiliar do modelo para dois grupos
r_{ji}	Covariável j do indivíduo i em $\boldsymbol{R}$
$\boldsymbol{s}$	Vetor de valores ideais para o caso de dois grupos
s_j	Valor ideal para a covariável j para o caso de dois grupos
b_j	Variável de desvio ao valor ideal na covariável j a ser minimizada
$\boldsymbol{n}$	Conjunto da quantidade de elementos em cada grupo
$\boldsymbol{Q}^{(k)}$	Matriz auxiliar do modelo para o caso estendido
q_{jik}	Covariável j do indivíduo i do grupo k em $\boldsymbol{Q}$
$\boldsymbol{W}$	Matriz de alocação para o caso estendido
w_{ik}	Elemento i do grupo k de $\boldsymbol{W}$
$\boldsymbol{S}$	Matriz de valores ideais para o caso estendido
s_{jk}	Valor ideal para a covariável j no grupo k para o caso estendido
b_{jk}	Variável de desvio ao valor ideal na coovariável j do grupo k a ser minimizada
d_k	Variável de desvio ao valor ideal no grupo k a ser minimizada

Sumário

1 Introdução 16
1.1 Aleatorização e Intencionalidade 17
1.2 Objetivos 18
1.3 Motivação 19
1.4 Estrutura do documento 19

2 Revisão da literatura 20
2.1 Planejamento de experimentos 20
2.1.1 Introdução à convergência em variáveis aleatórias 20
2.1.2 Dimensionamento de amostras 24
2.1.3 Aspectos práticos do uso da aleatorização pura 25
2.2 Conceitos Básicos em Álgebra Linear 26
2.2.1 Normas Vetoriais 26
2.2.2 Fatoração de Cholesky 27
2.2.3 Distância de Mahalanobis 28
2.3 O Método de Alocação Intencional Fortuita 29
2.4 Trabalhos correlatos em métodos de amostragem/alocação . 32
2.4.1 Rerandomização 32
2.4.2 Balanceamento considerando médias e variâncias das variáveis 34

3 Extensões para o caso de múltiplos grupos 37
3.1 Considerações sobre a generalização dos modelos 37
3.2 Otimização no caso da norma L_1 40
3.2.1 Formulação para dois grupos (L_1) 40
3.2.2 Formulação proposta para mais de dois grupos (L_1) 42
3.3 Otimização no caso da norma L_∞ 43
3.3.1 Formulação do problema para dois grupos (L_∞) 43
3.3.2 Formulação proposta para mais de dois grupos (L_∞) . . . 44
3.4 Otimização no caso da norma híbrida (L_1, L_∞) 45
3.4.1 Formulação do problema para dois grupos (L_1, L_∞) 45

3.4.2 Formulação proposta para mais de dois grupos (L_1, L_∞) . 46
3.5 Avaliação do desacoplamento no caso de múltiplos grupos . 47

4 Estudos de caso 49
4.1 Estudo de Caso - Estações de monitoramento 49
4.1.1 Resultados - balanceamento x desacoplamento 50
4.1.2 Experimentos auxiliares 56
4.2 Estudo de caso - Epicovid19 57
4.2.1 Experimentos iniciais 58
4.2.2 Definição dos níveis de perturbação 59
4.2.3 Resultados - Balanceamento 60
4.3 Estudo de caso - Projeto S 62
4.3.1 Resultados - Balanceamento e estimação das eficácias . . . 62

5 Conclusão 66

Referências 68

Apêndice A – Balanceamento - Estudo de Caso Epicovid 19 70

1 Introdução

Técnicas de inferência baseadas em amostragem ou planejamento de experimentos constituem ferramentas de inquestionável utilidade e aplicabilidade interdisciplinar como ensaios clínicos, avaliações de políticas públicas, estudos de impacto econômico, procedimentos de auditoria e pesquisas em geral. No contexto da modelagem de sistemas complexos, as diferentes técnicas e ferramentas de modelagem muitas vezes ensejam a estimação de parâmetros para viabilizar uma reprodução computacional de fenômenos observados. Estudos ou experimentos podem ser necessários, por exemplo, na identificação de propriedades emergentes de uma população ou na inferência sobre as características dos agentes.

Pesquisas por amostragem baseiam-se em selecionar, dentro de uma população, um subconjunto no qual parâmetros populacionais sobre variáveis desconhecidas são estimados através de métodos de inferência. Nesses casos, é importante que a amostra possua perfil similar à população em relação a certas covariáveis (p.ex. sexo, idade, estrato socioeconômico, etc) potencialmente correlacionadas com a variável a ser estimada, evitando estimativas viesadas. Exemplo dessa categoria é um estudo recente conduzido para estimar a prevalência do SARS-CoV2 nos principais municípios brasileiros (UNIVERSIDADE FEDERAL DE PELOTAS, 2021). Uma vez que a prevalência do SARS-CoV2 está correlacionada com a idade dos indivíduos e com sua condição socioeconômica, é importante que as unidades amostrais obtidas tenham perfis similares aos da população em relação a essas covariáveis.

Estudos experimentais baseados em grupos de controle/tratamento objetivam avaliar o efeito de intervenções aplicadas a um subgrupo de uma amostra, em relação a outro subconjunto, no qual nenhuma intervenção é aplicada. Exemplos dessa abordagem são estudos duplos-cegos, que buscam avaliar os efeitos de um medicamento, fornecendo a substância real aos grupos de tratamento e o placebo ao grupo de controle. Nesses casos, covariáveis desbalanceadas podem ser inadequadas para a determinação de um efeito comparativo das intervenções, pondo em xeque as conclusões de um experimento. Por exemplo, o efeito do medicamento em estudo pode ser atenuado ou potencializado em um grupo que possui proporção elevada de indivíduos com determinada condição (ex. idade avançada, caso mais severo, consumo de álcool).

Como mencionado nos parágrafos acima, todo delineamento amostral busca a obtenção de amostras/alocações balanceadas, para proporcionar inferências robustas. Em planos amostrais baseados em aleatorização, este balanceamento é, em tese, alcançado através de amostras suficientemente grandes, com base na convergência assintótica dos erros amostrais para zero, garantida pelo Teorema do Limite Central. No entanto, são frequentes os casos práticos em que há restrições na disponibilidade de recursos, configurando-se possivelmente em limitações nos tamanhos das amostras. Em tais situações, métodos que envolvem algum grau de otimização podem proporcionar um nível de balanceamento que dificilmente seria obtido de modo puramente aleatório. Assim, embora o uso da aleatorização seja o padrão predominante em pesquisas experimentais, numerosas são as discussões acerca do uso desse método.

1.1 Aleatorização e Intencionalidade

Uma etapa do planejamento do estudo a ser realizado envolve a definição do método utilizado para a seleção dos elementos que comporão a amostra ou os subgrupos de alocação. O uso da aleatorização em testes foi popularizado por Fisher (FISHER, 1935), estabelecendo um novo paradigma para a validade estatística em pesquisa experimental. Sob discussões que permeiam as diversas filosofias de amostragem, diferentes abordagens são concebidas para justificar ou contrapor o uso da aleatorização, transcendendo o aspecto do balanceamento. Bonassi *et al.* (2009), por exemplo, utilizam uma perspectiva baseada em teoria dos jogos, concluindo que a aleatorização pode proporcionar soluções para problemas de intersubjetividade e reduzir efeitos de viesamentos desconhecidos.

Ensaios randomizados são a principal prática para testar a eficácia de tratamentos em estudos clínicos e também são amplamente utilizados em outras áreas, como auditoria governamental, sendo recomendada em diversos manuais orientativos (TRIBUNAL DE CONTAS DA UNIÃO, 2020; INTERNATIONAL ORGANIZATION OF SUPREME AUDIT INSTITUTIONS, 2019).

Por outro lado, publicações recentes vêm incentivando o uso de métodos baseados em amostragem/alocação intecional. Sob uma perspectiva pragmática de pesquisa médica com ensaios clínicos, Diniz *et al.* (2016) apresentam alguns contrapontos às principais justificativas para o uso da aleatorização em experimentos.

1. "Cada unidade da amostra deve ter a mesma probabilidade de ser selecionada."

Contraponto: A amostragem aleatória aduz que qualquer amostra, balanceada ou não, possui a mesma probabilidade de ser selecionada. Em termos de eficiência do estudo isso é questionável, pois, se é possível minimizar a chance de se obter uma amostra desbalanceada, pode não ser conveniente depender de uma análise posterior para corrigir um problema que poderia ter sido evitado inicialmente.

2. "A Aleatorização evita a interferência humana no procedimento de alocação." Contraponto: De fato, a utilização de métodos que não são completamente aleatórios aumenta a preocupação de alguém obter algum controle intencional sobre a alocação, de forma a interferir com um objetivo específico. No entanto, métodos que conciliam otimização com algum grau de aleatoriedade podem minimizar o determinismo amostral, de modo a tornar muito baixa a possibilidade de um indivíduo intervir, tampouco ter um pré-conhecimento dos elementos que serão selecionados.

Esses pontos são apenas parte de um extenso debate acerca do uso de métodos de otimização. O método de *Alocação Intencional Fortuita* apresentado em Lauretto *et al.* (2012, 2017, 2019) combina técnicas de otimização com perturbações aleatórias, apresentando ganhos relevantes em termos de balanceamento, poder de inferência e eliminação de interferências externas. Contudo, a formulação desenvolvida contempla apenas dois grupos de alocação, o que limita suas possibilidades de aplicação.

1.2 Objetivos

O objetivo geral deste trabalho é propor uma extensão do método de *Alocação Intencional Fortuita* apresentado em Lauretto *et al.* (2017, 2019), para casos genéricos onde há mais de dois grupos de alocação e avaliar metodologia proposta nos quesitos de balanceamento e desacoplamento. Os objetivos específicos são:

a) analisar aspectos da formulação original e estendê-la para múltiplos grupos;
b) construir uma biblioteca computacional com as extensões propostas;
c) avaliar empiricamente a utilização dessa abordagem em estudos de caso.

Os resultados deste trabalho foram apresentados por Miguel *et al.* (2021), no *International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (MaxEnt2021) e posteriormente publicados na revista *Entropy* (MIGUEL *et al.*, 2022).

1.3 Motivação

A possibilidade de aprimoramento das atuais práticas de pesquisa motiva esse trabalho em todos os contextos aplicáveis, podendo proporcionar o desenvolvimento de desenhos de experimentos e planos de amostragem mais eficientes.

A generalização proposta torna viável a experimentação concomitante em diversos grupos possibilitando, por exemplo, a avaliação de mais de um tipo de tratamento, ou opção de ação governamental. Os estudos de caso buscam avaliar, dentre outros critérios, o efeito do uso desse método no dimensionamento amostral, quando comparado a métodos de aleatorização, usualmente utilizados e previstos em orientações normativas.

1.4 Estrutura do documento

O Capítulo 2 apresenta uma revisão teórica, incluindo conceitos introdutórios de planejamento de experimentos e contextualiza o trabalho com a linha de pesquisa e publicações correlatas.

O Capítulo 3 apresenta a formulação proposta para a generalização do *Método de Alocação Fortuita* no caso de múltiplos grupos e as respectivas adaptações nos métodos de avaliação.

O Capítulo 4 apresenta os Estudos de Caso. O primeiro é um caso de seleção de múltiplos grupos de estações de monitoramento atmosférico, utilizando a formulação proposta e estendendo os experimentos realizados por Lauretto *et al.* (2019). O segundo apresenta uma avaliação empírica da formulação original em um caso de amostragem, baseado em um estudo de base populacional para estimação da prevalência da SARS-CoV2 (UNIVERSIDADE FEDERAL DE PELOTAS, 2021). O terceiro é inspirado no Projeto S (INSTITUTO BUTANTAN, 2020), um estudo epidemiológico conduzido na população do município de Serrana/SP, para testar a eficácia da vacina Coronavac; este experimento considera uma adaptação do estudo original, na qual quatro diferentes tipos de vacina são avaliadas. Por fim, o Capítulo 5 apresenta a conclusão do trabalho.

2 Revisão da literatura

Este capítulo apresenta uma introdução ao planejamento de experimentos, os conceitos matemáticos fundamentais para o desenvolvimento do trabalho, o método de *Alocação Fortuita*, contemplando a formulação para dois grupos e os métodos correlatos de alocação encontrados na literatura. Esses tópicos visam uma contextualização do presente trabalho no ambiente de pesquisa em que ele está inserido e proporcionar um embasamento teórico para melhor compreensão de métodos encontrados na literatura e da extensão proposta.

2.1 Planejamento de experimentos

Esta seção apresenta brevemente propriedades e conceitos de convergência de amostras aleatórias e o Teorema do Limite Central. Apresenta também aspectos práticos da utilização da Aleatorização pura em um contexto de limitação no tamanho da amostra e um breve histórico de desenhos de experimentos utilizados para alcançar o balanceamento.

2.1.1 Introdução à convergência em variáveis aleatórias

Uma definição de amostragem é o procedimento de obter um subconjunto qualquer de n unidades de uma população. Na prática, em estudos e experimentos, é usual que as variáveis de interesse dessa população possuam distribuições desconhecidas. O método mais comum de obter amostras é por meio da *aleatorização simples*, no qual cada unidade da população possui a mesma probabilidade de ser selecionada. A caracterização da amostra proporciona conhecimentos acerca da população, de modo que podemos obter informações sobre a distribuição original da qual a amostra foi extraída.

Considerando uma sequência de variáveis aleatórias, independentes e identicamente distribuídas, $x_1, x_2, \ldots, x_n$, extraída de uma população com média μ e variância σ^2, podemos definir a média e a variância amostrais como:

$$\overline{x} = \frac{1}{n}\sum_{1}^{n} x_i \qquad\qquad s^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i - \overline{x})^2 \tag{1}$$

Para qualquer sequência de variáveis aleatórias com valor esperado finito, temos que a soma dos valores esperados é igual ao valor esperado da soma e, no caso de uma função linear, $E(aX + b) = aE(X) + b$, então: (DeGroot e Schervish (2012, p. 218/219))

$$E(x_1 + \cdots + x_n) = E(x_i) + \cdots + E(x_n) \tag{2}$$

$$E(\overline{x}) = E(\frac{1}{n}(x_1 + \cdots + x_n)) = \frac{1}{n}E(x_1 + \cdots + x_n) \tag{3}$$

As Equações (2) e (3) implicam que o valor esperado para a média de uma amostra equivale à média populacional (DeGroot e Schervish (2012, p. 350)):

$$E(\overline{x}) = \frac{1}{n}\sum_{1}^{n} E(x_i) = \frac{1}{n}n\mu = \mu \tag{4}$$

A capacidade de uma amostra caracterizar adequadamente uma população está relacionada ao tamanho da amostra. Uma definição superficial da ideia de convergência é que, quanto maior a quantidade de elementos da sequência de variáveis, mais ela se assemelha à população da qual foi extraída. Essa aproximação é descrita pela Lei dos Grande Números.

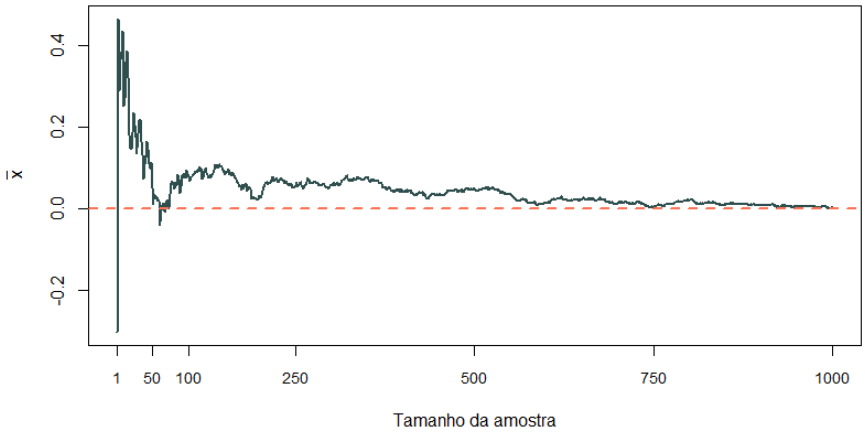
Lei (fraca) dos grandes números

Define-se (DeGroot e Schervish (2012, p. 353)) a Lei dos Grandes Números como a convergência em probabilidade $\overline{x} \xrightarrow{p} \mu$, geralmente apresentada como a Equação (5), que mostra que quando o tamanho da amostra tende ao infinito, a probabilidade da diferença entre a média amostral e o valor esperado ser menor do que qualquer $\epsilon > 0$ é 1:

$$\lim_{n\to\infty} P(|\overline{x} - \mu| < \epsilon) = 1 \tag{5}$$

A Figura 1 apresenta a média de amostras de tamanhos crescentes simuladas a partir da distribuição normal padrão ($\mathcal{N}(0, 1)$). A linha horizontal tracejada representa a verdadeira média da distribuição, ($\mu = 0$)

Figura 1 – Exemplo de convergência para o valor esperado de uma variável com distribuição normal padrão.



Fonte: Elaborado pelo autor.

No planejamento de um estudo, a Lei dos Grandes Números tem grande influência no dimensionamento amostral, pois à medida que o tamanho da amostra aumenta, maior a garantia de proximidade entre a média amostral e a média populacional. Juntamente com a ideia de convergência descrita pelo Teorema do Limite Central, esse conceito embasa grande parte dos métodos tradicionais de inferência e das formulações utilizadas no dimensionamento de amostras.

Teorema do Limite Central

No caso de amostras aleatórias simples extraídas de qualquer distribuição com média μ e variância σ^2, a distribuição amostral de $\overline{x}$ aproxima-se de $\mathcal{N}(\mu, \sigma^2/n)$, à medida que n aumenta. Ou seja, o Teorema do Limite Central estabelace que $\sqrt{n}(\overline{x} - \mu)/\sigma$ possui uma distribuição normal padrão limite:

$$\frac{\sqrt{n}(\overline{x} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \tag{6}$$

Decorre disso que a aleatorização pura garante, assintoticamente, que a variável será balanceada. No entanto, isso não garante que o balanceamento será proporcionado pela aleatorização pura no caso de uma amostra específica.

Além de um dos principais teoremas da estatística, o Teorema do Limite Central é o embasamento para as técnicas tradicionais de estimação e possibilita, por exemplo, a obtenção de intervalos de confiança aproximados para $\overline{x}$ em relação à média populacional,

baseados na distribuição normal. Conforme apresentado em Bolfarine e Bussab (2007), para um tamanho de amostra suficientemente grande:

$$P\big(\frac{|\overline{x}-\mu|}{\sqrt{\sigma^2/n}} \leq z_\alpha\big) \simeq 1-\alpha \tag{7}$$

Considerando que $(-z_\alpha, z_\alpha)$ são as ordenadas que delimitam uma área igual a $1-\alpha$ na densidade da $\mathcal{N}(0,1)$, um intervalo de confiança para μ, com coeficiente de confiança $(1-\alpha)$, pode ser:

$$P\big(\overline{x} - z_\alpha\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{x} + z_\alpha\sqrt{\frac{\sigma^2}{n}}\big) \simeq 1-\alpha, \tag{8}$$

$$\big(\overline{x} - z_\alpha\sqrt{\frac{\sigma^2}{n}}; \overline{x} + z_\alpha\sqrt{\frac{\sigma^2}{n}}\big) \tag{9}$$

Ou, no caso da variância populacional, σ^2, ser desconhecida:

$$\big(\overline{x} - z_\alpha\sqrt{\frac{s^2}{n}}; \overline{x} + z_\alpha\sqrt{\frac{s^2}{n}}\big) \tag{10}$$

Intervalos de confiança proporcionam mais informação à estimação de um parâmetro, pois podemos definir um intervalo que contenha esse parâmetro com um nível de confiança conhecido e a amplitude do intervalo nos fornece uma ideia da precisão da estimativa (DeGroot e Schervish (2012)).

No contexto de experimentação, distribuições obtidas a partir de amostras viabilizam a realização de testes de hipóteses, pois possibilitam indicar se um intervalo de confiança para um parâmetro contém o valor associado à hipótese em teste. Assim, para estabelecer um tamanho adequado da amostra, pode ser útil uma abordagem baseada na função de poder do teste, que quantifica a probabilidade de se cometer um erro do tipo I (rejeitar uma hipótese verdadeira) ou do tipo II (não rejeitar uma hipótese falsa).

Grande parte dos estudos realizados contempla alguma inferência envolvendo técnicas de estimação ou testes. Em ambos os casos, para que os objetivos do estudo sejam alcançados satisfatoriamente e a amostra viabilize a obtenção de uma conclusão aproveitável, é relevante a realização de um planejamento amostral.

2.1.2 Dimensionamento de amostras

Nos casos descritos na subseção anterior, o tamanho da amostra pode ser determinado a partir da definição do nível de confiança, da fixação de uma amplitude máxima admitida para o intervalo e da variabilidade da população, geralmente estimada (Bolfarine e Bussab (2007, p. 70)). Em estudos experimentais, a determinação do tamanho da amostra pode ser justificada por meio da função de poder, controlando a probabilidade de erro desejada (Kutner *et al.* (2005, p.653)). De modo geral, o dimensionamento amostral sempre envolve um trade-off entre a economia de recursos e o ganho na precisão da estimativa ou no poder de inferência. Uma solução adequada pode ser bem diferente em situações específicas, de acordo com o caso prático.

Situações onde há ampla disponibilidade de elementos com baixo custo do procedimento de análise apresentam a situação mais favorável para o planejamento amostral, uma vez que as condições de convergência apresentadas na Subseção 2.1.1 são facilmente atendidas. Pesquisas de mercado em massa realizadas por email e pesquisas eleitorais realizadas por ligação telefônica são exemplos desse caso.

No entanto, outras situações podem limitar o tamanho da amostra, seja pelo alto custo de se realizar um procedimento de coleta ou pela indisponibilidade de novos elementos. Esses cenários ocorrem, por exemplo, em pesquisas envolvendo condições clínicas incomuns ou raras, estudos que dependem de procedimentos de alto custo ou que envolvem um conjunto limitado de elementos, como instituições (escolas, hospitais, etc.) ou regiões geográficas (estados, subprefeituras, setores censitários, etc.).

Diante dessa limitação, é mais provável que os parâmetros amostrais apresentem diferenças substanciais em relação aos parâmetros populacionais, como ilustrado na figura 1. Em estudos experimentais, o mesmo ocorre em grupos de alocação com poucos elementos, uma vez que o processo de alocação é fundamentalmente o mesmo da amostragem. Além disso, modelos estatísticos baseados em amostras inadequadamente pequenas possuem maior risco de apresentar ajustes espúrios, levando a equívocos nas relações entre as variáveis e na comparação de efeitos entre os grupos.

Enquanto o princípio da convergência garante inúmeras propriedades desejáveis em amostras grandes, limitações no tamanho da amostra são comuns em casos práticos e um desafio para o planejamento amostral. A necessidade de utilizar uma amostra pequena

pode motivar a reconsideração do método de amostragem/alocação, uma vez que, nesses casos, a usual amostragem aleatória simples pode proporcionar riscos para o estudo.

2.1.3 Aspectos práticos do uso da aleatorização pura

O uso da aleatorização pura em desenho de experimentos foi popularizado por Ronald Fisher e é considerado um padrão de qualidade na prática clínica contemporânea, principalmente por evitar viés de seleção e justificar o uso de métodos frequêntistas de inferência. No entanto, como discorrido em Lock (2011), a possibilidade desse método proporcionar desbalanceamento, colocando o experimento em risco, foi relatada pelo próprio Fisher.

Um exemplo típico é o caso de definir as alocações de um conjunto de elementos, caracterizados por m covariáveis, em dois grupos. Conforme apresentado em Lock (2011), ao avaliar o balanceamento, a chance de ao menos uma covariável apresentar uma diferença significante entre os dois grupos ao nível α é $(1-\alpha)^m$, o que representa uma probabilidade de 40% no caso de 10 covariáveis e $\alpha = 5\%$. Esse cenário é ainda mais desfavorável quando envolve múltiplos grupos. Uma magnitude da ineficácia do uso da aleatorização pura para garantir o balanceamento nesses casos pode ser ilustrada com o exemplo a seguir.

Considere que k amostras de tamanho n são extraídas aleatoriamente de uma população infinita e que desejamos avaliar m covariáveis binárias, todas com proporção populacional igual a 50%. Definimos que a amostra está balanceada em relação à população se $n/2(1-a) \leq n\hat{p} \leq n/2(1+a)$, onde a é um percentual de tolerância arbitrário e $n\hat{p}$ denota a quantidade de uma das duas categorias em cada covariável. A probabilidade de ao menos uma das covariáveis apresentar algum desbalanceamento pode ser obtida por:

$$P(desbalanceamento) = 1 - \left[\left(\sum_{i=\frac{n}{2}(1-a)}^{\frac{n}{2}(1+a)} Bin(i, 0.5) \right)^m \right]^k$$

Onde $Bin(n, p)$ denota a distribuição binomial com probabilidade de sucesso p e n repetições.

Considerando, como exemplo, um caso em que 4 amostras de 40 elementos são extraídas de uma população infinita com 20 covariáveis de interesse e que uma covariável é considerada balanceada se apresentar desvio menor que 15% ($17 \leq n\hat{p} \leq 23$), a chance

de obter as quatro amostras balanceadas, utilizando a aleatorização simples, é de aproximadamente uma em 70 bilhões.

Experimentos conduzidos, tanto na literatura de referência em Morgan e Rubin (2015) e Lauretto *et al.* (2017, 2019), quanto neste trabalho também ilustram a magnitude do (des)balanceamento proporcionado com a utilização da aleatorização pura (ver Seções 4.1, 4.2). Soluções para superar esse problema em amostras pequenas surgem na literatura de modo quase concomitante à ideia de aleatorização e alguns dos métodos tradicionalmente utilizados são apresentados neste trabalho. Para melhor apresentar o Método de *Alocação Fortuita* e ilustrar certos conceitos matemáticos utilizados formulação, uma revisão é apresentada nas Seções seguintes.

2.2 Conceitos Básicos em Álgebra Linear

Esta seção apresenta alguns conceitos matemáticos utilizados nas medidas de balanceamento nas tranformações matriciais dos métodos de otimização.

2.2.1 Normas Vetoriais

Seja V um espaço vetorial, $||\cdot||$ é uma Norma se satisfaz as propriedades:

- $||a\boldsymbol{x}|| = |a|\,||\boldsymbol{x}||, \forall \boldsymbol{x} \in V, a \in \mathbb{R}$
- $||\boldsymbol{x}+\boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||, \forall \boldsymbol{x}, \boldsymbol{y} \in V$
- $||\boldsymbol{x}|| \geq 0, e \;\; ||\boldsymbol{x}|| = 0 \iff \boldsymbol{x} = 0$

Seja $\mathcal{X}$ um espaço vetorial com norma $||\cdot||_p$, é possível computar a distância $||\boldsymbol{x}-\boldsymbol{y}||_p$ entre dois vetores como:

$$||\boldsymbol{d}||_p = (|d_1|^p + |d_2|^p + |d_3|^p \ldots |d_n|^p)^{\frac{1}{p}} \quad d_i = |x_i - y_i| \tag{11}$$

$$||\boldsymbol{x}||_p = (\sum_{1}^{n} |x_i|)^{\frac{1}{p}} \quad d_i = |\boldsymbol{x}_i - \boldsymbol{y}_i| \tag{12}$$

Quando p é igual a $1, 2$ e ∞, a equação (12) corresponde, respectivamente, às distâncias de Manhattam ($||\boldsymbol{d}||_1 = |\boldsymbol{x}-\boldsymbol{y}|$), euclidiana ($||\boldsymbol{d}||_2 = (|\boldsymbol{x}-\boldsymbol{y}|^2)^{1/2}$) e Tchebycheff ($||\boldsymbol{d}||_\infty = \max|\boldsymbol{x}-\boldsymbol{y}|$).

A norma híbrida, apresentada por Ward e Wendell (1980), descreve uma aproximação da norma euclidiana por meio da combinação linear entre as normas $||.||_1$ e $||.||_\infty$ em $\mathbb{R}^2$.

$$||\boldsymbol{d}||_{(\gamma 1,\gamma 2)} = \gamma_1 ||\boldsymbol{d}||_1 + \gamma_2 \sqrt{2} ||\boldsymbol{d}||_\infty \tag{13}$$

Os coeficientes γ_1 e γ_2 representam os pesos atribuídos a cada componte. Neste trabalho consideraremos $\gamma_1 = \gamma_2 = 1$. Uma generalização dessa norma pode ser obtida pela substituição de $\sqrt{2}$ por $\sqrt{m}$, descrevendo a inclinação da componente $||.||_\infty$ em $\mathbb{R}^m$. Como apresentado na Seção 2.3, uma vantagem dessa norma híbrida é aproximar o problema de minimização da norma $||.||_2$ (que é um problema de programação quadrática), através da formulação de um problema de programação linear.

2.2.2 Fatoração de Cholesky

A fatoração de Cholesky é um caso particular da decomposição *LU* (lower-upper) que, de modo geral, exige menos operações aritiméticas e pode ser obtida por uma rotina simples (SOLOMON, 2020):

Seja $\boldsymbol{A} = a_{i,j}$ uma matriz simétrica, positiva definida. A Fatoração de Cholesky fornece a decomposição $\boldsymbol{C}\boldsymbol{C}^t$, $\boldsymbol{C} = c_{i,j}$, na qual:

- Os elementos da diagonal são obtidos por:

$$\begin{aligned} c_{1,1} &= a_{1,1}^{1/2} \\ c_{i,i} &= \left(a_{i,i} - \sum_{k=1}^{i-1} c_{i,k}^2\right)^{1/2} \qquad i = 2, 3 \ldots, n \end{aligned}$$

- E os demais elementos são obtidos por:

$$\begin{aligned} c_{i,1} &= \frac{a_{i,1}}{c_{1,1}}, \qquad i = 2, 3 \ldots, n \\ c_{i,j} &= \frac{\left(a_{i,j} - \sum_{k=1}^{j-1} c_{i,k} c_{j,k}\right)}{c_{j,j}} \qquad j = 2, 3, \ldots n-1 \end{aligned}$$

Neste trabalho, a fatoração de Cholesky é utilizada na inversa da matriz de covariância dos dados, permitindo uma transformação matricial de rotação e redimensionamento do conjunto de dados, gerando uma estrutura de covariância não correlacionada,

de modo que a Distância de Mahalanobis pode ser calculada diretamente pela distância euclidiana (ver Seção 2.3).

2.2.3 Distância de Mahalanobis

A distância de Mahalanobis é uma variação da distância euclidiana (L_2), que possui propriedades de invariância de escala e considera as correlações entre os dados. Essa distância pode ser calculada entre dois vetores $\boldsymbol{x}^{(1)}$ e $\boldsymbol{x}^{(2)} \in \mathbb{R}^m$, como:

$$D_M(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \left[(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})^t \, cov(\boldsymbol{X})^{-1} \, (\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})\right]^{1/2} \tag{14}$$

Onde $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ denota uma matriz de de observações independentes e igualmente distribuídas, proveniente da mesma população original de $\boldsymbol{x}^{(1)}$ e $\boldsymbol{x}^{(2)}$.

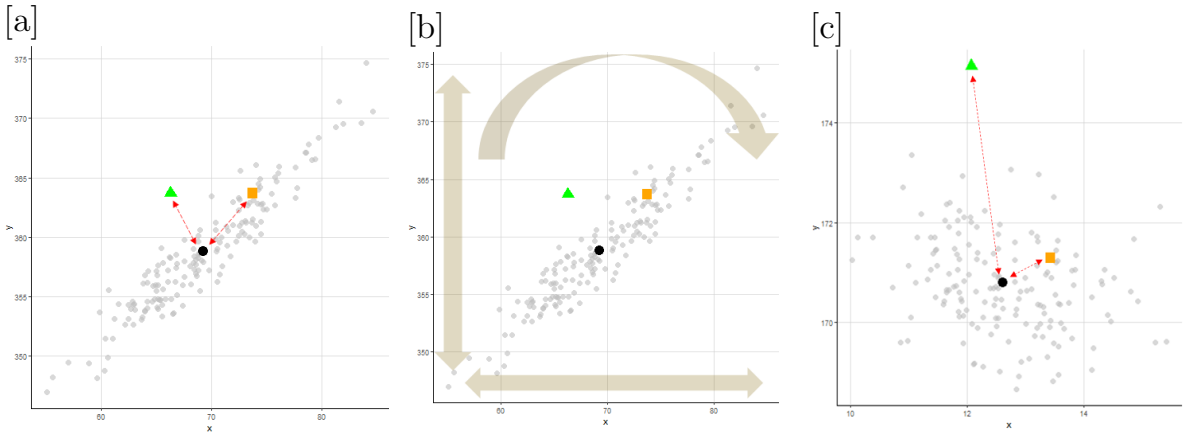
Denotando $\overline{\boldsymbol{X}}$ como vetor de média das colunas da matriz $\boldsymbol{X}$, a distância de um vetor qualquer em relação a $\overline{\boldsymbol{X}}$ pode ser calculada por:

$$D_M(\boldsymbol{x}^{(k)}, \overline{\boldsymbol{X}}) = \left[(\boldsymbol{x}^{(k)} - \overline{\boldsymbol{X}})^t \, cov(\boldsymbol{X})^{-1} \, (\boldsymbol{x}^{(k)} - \overline{\boldsymbol{X}})\right]^{1/2} \tag{15}$$

Essa distância é equivalente à distância euclidiana, para o caso em que $cov(\boldsymbol{X})$ é a matriz identidade. Assim, também pode ser caclculada por uma prévia transformação de $\boldsymbol{X}$ para um conjunto de dados não correlacionados, no qual as variáveis possuem variância 1 (Figura 2). Seja $\boldsymbol{L}$ a matriz triangular superior referente à decomposição de Cholesky de $cov(\boldsymbol{X})^{-1}$. Definimos a transformação:

$$\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{L} \tag{16}$$

Figura 2 – Exemplo de distribuição correlacionada e após transformação no caso de duas variáveis



Fonte: Elaborado pelo autor.

Após a normalização, a distância de Mahalanobis pode ser calculada pela distância euclicianda, de modo que as equações (14) e (15) são equivalentes a:

$$D_M(\boldsymbol{x}^{*(1)}, \boldsymbol{x}^{*(2)}) = \|\boldsymbol{x}^{*(1)} - \boldsymbol{x}^{*(2)}\|_2 \tag{17}$$

$$D_M(\boldsymbol{x}^{*(k)}, \overline{\boldsymbol{X}^*}) = \|\boldsymbol{x}^{*(k)} - \overline{\boldsymbol{X}^*}\|_2 \tag{18}$$

No exemplo bivariado apresentado na Figura 2, os pontos destacados (triângulo e quadrado) possuem a mesma distância euclideana ao centro da distribuição correlacionada, na matriz original (a). Após a transformação de rotação e reescalonamento dos eixos (b), as posições dos pontos consideram a estrutura de covariância e a distância de Mahalanobis pode ser obtida pela distância euclideana diretamente (c). No contexto deste trabalho, a distância de Mahalanobis é utilizada para verificar o balanceamento das variáveis entre subconjuntos de dados.

2.3 O Método de Alocação Intencional Fortuita

O método de *Alocação Intencional Fortuita* apresentado por Lauretto *et al.* (2012, 2017, 2019) caracteriza-se por conciliar a otimização do balancemento com algum grau de aleatoriedade no processo de amostragem/alocação. Esse método utiliza programação por metas para minimizar os desvios entre cada variável nos dois grupos. A especificação formal do problema de otimização está mais detalhadamente descrita na Seção 3 e é baseada nas formulações propostas por Romero (1991) para problemas de otimização nas normas L_p.

A medida de balanceamento considerada é a distância de Mahalanobis (ver Seção 2.2.3). A quantidade de elementos em cada grupo é denotada por n_1 e n_2, e o vetor linha $\boldsymbol{w}$, de dimensão $n = n_1 + n_2$, indica a alocação de cada elemento nos grupos 1 ($\boldsymbol{w}_i = 1$) ou 2 ($\boldsymbol{w}_i = 0$). Definindo a matriz $\boldsymbol{X} \in R^{n \times m_X}$ como a matriz de covariáveis observadas para as unidades amostrais candidatas, as médias de cada covariável sobre os elementos alocados, respectivamente aos grupos 1 e 2, podem ser expressas como:

$$\overline{\boldsymbol{X}^*}^{(1)} = \frac{\boldsymbol{w}}{n_1}\boldsymbol{X}^* \quad e \quad \overline{\boldsymbol{X}^*}^{(2)} = \frac{(\boldsymbol{1} - \boldsymbol{w})}{n_2}\boldsymbol{X}^* \tag{19}$$

De modo que a distância de Mahalanobis entre as médias $\overline{\boldsymbol{X}^*}^{(1)}$ e $\overline{\boldsymbol{X}^*}^{(2)}$ (as quais dependem de $\boldsymbol{X}$ e de $\boldsymbol{w}$) pode ser calculada como:

$$M(\boldsymbol{w}, \boldsymbol{X}) = m^{-1}\|\overline{\boldsymbol{X}^*}^{(1)} - \overline{\boldsymbol{X}^*}^{(2)}\|_2 \tag{20}$$

A otimização do balanceamento consiste em encontrar uma alocação que torne os grupos mais similares possíveis, ou seja, buscar uma alocação $\boldsymbol{w}$ que minimiza a função $M(\boldsymbol{w}, \boldsymbol{X})$, podendo ser formulada como o problema:

$$\begin{gathered} \min \quad M(\boldsymbol{w}, \boldsymbol{X}) \\ \text{sujeito a} \\ \mathbf{1}\boldsymbol{w}^t = n_1 \\ \mathbf{1}(\mathbf{1} - \boldsymbol{w})^t = n_2 \\ \boldsymbol{w} \in \{0, 1\}^n \end{gathered} \tag{21}$$

No entanto, esse problema descreve um método puramente intencional e determinístico de balanceamento. O método de *Alocação Fortuita* busca conciliar a minimização da equação (20) com algum grau de aleatorização ao processo de alocação. Para isso, é definida uma matriz covariáveis artificiais, $\boldsymbol{Z} \in \mathbb{R}^{n \times m_Z}$, composta por m_Z variáveis aleatórias com distribuição Normal padrão. O balanceamento por esse método consiste na resolução do problema (22):

$$\begin{gathered} \min \quad (1 - \lambda) M(\boldsymbol{w}, \boldsymbol{X}) + \lambda M(\boldsymbol{w}, \boldsymbol{Z}) \\ \text{sujeito a} \\ \mathbf{1}\boldsymbol{w}^t = n_1 \\ \mathbf{1}(\mathbf{1} - \boldsymbol{w})^t = n_2 \\ \boldsymbol{w} \in \{0, 1\}^n \end{gathered} \tag{22}$$

O parâmetro λ atua como regulador do grau de determinismo das alocações, controlando a magnitude das perturbatções acrescidas, proporcionando alocações que podem variar de totalmente determinísticas ($\lambda = 0$) a totalmente aletatórias ($\lambda = 1$). A quebra de determinismo proporcionada pela introdução da componente aleatória $\boldsymbol{Z}$ no processo de alocação atende a dois requisitos preconizados pelos defensores da aleatorização: (i) a eliminação de interferências humanas e (ii) a possibilidade de se realizar inferências sob o paradigma frequentista. Seguindo a nomenclatura adotada por Lauretto *et al.* (2019), usaremos o termo *desacoplamento* para nos referirmos a esta quebra no determinismo da alocação.

A obtenção de uma medida do desacoplamento consiste, inicialmente, em repetir o processo de alocação $R \approx 500$ vezes. Em seguida, utiliza-se o coeficiente de Yule (1912) para computar o balanço entre o número de repetições em que cada par de unidades foi

alocado no mesmo grupo e o número de repetições em que cada par foi alocado em grupos diferentes: formalmente, denotando por $z_{pq}(i,j)$ a quantidade de vezes que um par de unidades $(i,j) \in \{1,2,3...n\}^2, i < j$, é alocado respectivamente aos grupos $(p,q) \in \{0,1\}^2$, esse coeficiente pode ser calculdado por

$$\mathbf{Y}(i,j) = \frac{\sqrt{z_{00}(i,j)z_{11}(i,j)} - \sqrt{z_{01}(i,j)z_{10}(i,j)}}{\sqrt{z_{00}(i,j)z_{11}(i,j)} + \sqrt{z_{01}(i,j)z_{10}(i,j)}} \tag{23}$$

As formulações nas equações (21) e (22) são problemas de Programação Quadrática Inteira Mista. Lauretto *et al.* (2017), Lauretto *et al.* (2019) transformam essas formulações em um problema de Programação Linear Inteira Mista, através da aproximação da distância de Mahalanobis baseada na norma Híbrida (apresentada na subseção 2.2.1):

$$H(\boldsymbol{w}, \boldsymbol{A}) = m^{-1}\Big(\|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_1 + \sqrt{m}\|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_\infty\Big) \tag{24}$$

Onde a matriz $\boldsymbol{A}^*$ é composta pela matriz das covariáveis de interesse e artificiais:

$$\boldsymbol{X}^* = (1-\lambda)(\boldsymbol{X}\boldsymbol{L}) \qquad \text{e} \qquad \boldsymbol{Z}^* = \lambda(\boldsymbol{Z}\boldsymbol{L})$$

$$\boldsymbol{A}^* = \begin{bmatrix} \boldsymbol{X}^* & \vdots & \boldsymbol{Z}^* \end{bmatrix} \tag{25}$$

Uma transformação de λ em λ^* é aconselhável para equilibrar os pesos das covariáveis artificiais e de interesse, que possuem dimensões distintas:

$$\lambda = \lambda^* \,/\, [\lambda^*(1 - m_z/m_x) + m_z/m_x] \tag{26}$$

Assim, uma minimização da função 24 pode ser formulada como o sequinte problema de Programação Linear Inteira Mista:

$$\begin{aligned} \min \quad & (1-\lambda)H(\boldsymbol{w}, \boldsymbol{X}) + \lambda H(\boldsymbol{w}, \boldsymbol{Z}) \\ & \text{sujeito a} \\ & \mathbf{1}\boldsymbol{w}^t = n_1 \\ & \mathbf{1}(\mathbf{1} - \boldsymbol{w})^t = n_2 \\ & \boldsymbol{w} \in \{0,1\}^n \end{aligned} \tag{27}$$

Variações desse método podem ser formuladas para outras funções de distância. No Capítulo 3 são apresentadas formulações desse método para as normas L_1 e L_∞ e uma proposta de generalização para mais de dois grupos de alocação. O caso generalizado envolve uma adaptação das medidas de balanceamento (Seção 3.1) e desacoplamento consideradas.

2.4 Trabalhos correlatos em métodos de amostragem/alocação

2.4.1 Rerandomização

Morgan e Rubin (2012) discorrem sobre a recomendação de re-aleatorizar as alocações dos grupos antes do início do experimento, caso seja identificado que a configuração das alocações apresenta um desbalanceamento indesejável. O método de *Rerandomização* proposto visa mitigar o problema do desbalanceamento por meio de sucessivas alocações aleatórias e da definição de um critério de aceitação. De modo geral, a ideia desse método pode ser resumida nos seguintes passos (MORGAN; RUBIN, 2015):

1 - Coletar os dados.

2 - Definir um critério de aceitação para o balanceamento.

3 - Alocar os elementos nos grupos por meio de aleatorização.

4 - Checar o atendimento do critério definido no passo 2: Seguir para o passo 5 se o critério for atendido, senão, retornar ao passo 3.

5 - Conduzir o experimento utilizando a configuração final obtida no passo 4.

6 - Analisar os resultados utilizando testes de aleatorização, mantendo apenas as aleatorizações simuladas que satisfazem o critério de balanceamento definido em 2.

Define-se $\varphi(\boldsymbol{X}, \boldsymbol{w})$ como a função de aceitação/rejeição onde $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, é uma matriz composta por n elementos a serem alocados e m covariáveis de interesse e $\boldsymbol{w}$ o vetor n-dimensional que indica a alocação de cada elemento. Uma alocação é considerada "aceitável"se a distância entre as médias dos dois grupos, $D_M(\overline{\boldsymbol{X}}^{(1)}, \overline{\boldsymbol{X}}^{(2)})$, é inferior à um limite a, estabelecido:

$$\varphi(\boldsymbol{x}, \boldsymbol{w}) = \begin{cases} 1, & se \quad D_M(\overline{\boldsymbol{X}}^{(1)}, \overline{\boldsymbol{X}}^{(2)}) \leq a \\ 0, & se \quad D_M(\overline{\boldsymbol{X}}^{(1)}, \overline{\boldsymbol{X}}^{(2)}) > a \end{cases} \tag{28}$$

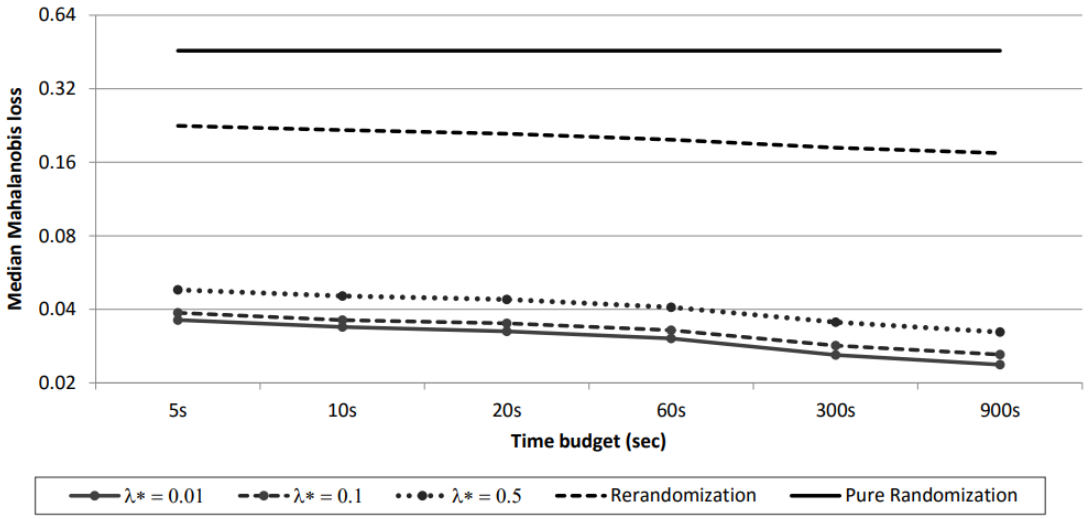
Os trabalhos de Morgan e Rubin (2012) e Morgan e Rubin (2015) consideram $D_M(\overline{\boldsymbol{X}}^{(1)}, \overline{\boldsymbol{X}}^{(2)})$ como a distância de Mahalanobis entre as médias dos grupos de alocação, $\overline{\boldsymbol{X}}^{(1)}$ e $\overline{\boldsymbol{X}}^{(2)}$, que pode ser apresentada como:

$$D_M(\overline{\boldsymbol{X}}^{(1)}, \overline{\boldsymbol{X}}^{(2)}) = \frac{n_1 n_2}{n} (\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)})^t cov(\boldsymbol{X})^{-1} (\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)}) \tag{29}$$

onde os vetores $\overline{\boldsymbol{X}}^{(1)}$ e $\overline{\boldsymbol{X}}^{(2)}$, que representam as médias das covariáveis dos grupos de tratamento e controle respectivamente, são obtidos pela equação (19). O critério de aceitação, a, pode ser definido de acordo com o poder computacional disponível, bem como de acordo com a importância relativa das covariáveis, definindo diferentes níveis de balanceamento para cada covariável (MORGAN; RUBIN, 2015). Inferências baseadas em testes de aleatorização podem ser realizadas considerando um subcojunto de amostras balanceadas $\{\boldsymbol{w}|\varphi(\boldsymbol{x},\boldsymbol{w}) = 1\}$.

Esse método é eficaz em proporcionar algum balanceamento da média mantendo um grau de desacoplamento próximo ao da aleatorização simples (LAURETTO *et al.*, 2017). No entanto, algumas situações práticas podem inviabilizar sua utilização. Um desses casos é a grande quantidade de covariáveis: Em uma amostra aleatória simples a probabilidade ao menos uma de m covariváeis ser desbalanceada a uma nível de significância α é proporcional a $1 - (1 - \alpha)^m$, ou seja, o número de alocações aleatórias necessarias cresce exponencialmente em relação à quantidade de covariváveis. A necessidade de um limite excessivamente baixo para o critério de tolerância, a, também pode inviabilizar esse método computacionalmente, além de possibilitar que o subconjunto de soluções aceitáveis se torne restrito a ponto de afetar o grau de desacomplamento proporcionado pelo método. Uma comparação do desempenho computacional desse método foi apresentada em Lauretto *et al.* (2017). A Figura 3 apresenta a mediana de $M(\boldsymbol{w},\boldsymbol{X})$ para 500 alocações obtidas também pelos métodos de *Alocação Fortuita* e de *Aleatorização Pura*.

Figura 3 – Função de Perda de Mahalanobis e tempo de processamento para os métodos Alocação Fortuita (diferentes valores de λ^*), Rerandomização e Aleatorização pura.



Fonte: Lauretto *et al.* (2017).

Para os experimentos realizados em Lauretto *et al.* (2017), Lauretto *et al.* (2019) e no presente trabalho (Seções 4.1 e 4.2) foi realizada uma adaptação do método de *Rerandomização*, de modo que a rotina busque a melhor alocação proporcionada em um determinado intervalo de tempo, dispensando a definição de um critério de aceitação.

2.4.2 Balanceamento considerando médias e variâncias das variáveis

Bertsimas, Johnson e Kallus (2015) apresentam um método baseado na minimização das diferenças entre as médias e das diferenças entre as variâncias para o problema de alocação. Esse médodo propõe alocar deterministicamente os elementos em g grupos e, após, aleatorizar os tratamentos atribuídos a cada grupo. Uma versão simplificada consiste em alocar uma covariável x_i, $i = 1, 2, ...n$ em grupos de $k = \frac{n}{g}$ elementos, que padronizada, é representada por $x'_i = (x_i - \overline{x})/\hat{\sigma}$, onde $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. A matriz $\boldsymbol{W}_{ip} \in \{0, 1\}$, indica a alocação do elemento i no grupo p, de modo que as componentes relativas à média e variância em cada grupo podem ser expressas por:

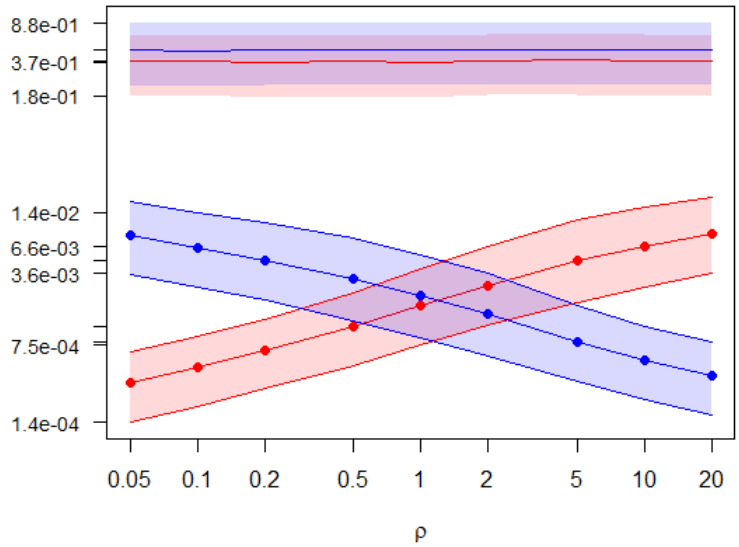
$$\mu_p(\boldsymbol{W}) = \frac{1}{k}\sum_{i=1}^{n} x'_i \boldsymbol{W}_{ip} \quad e \quad \sigma_p^2(\boldsymbol{W}) = \frac{1}{k}\sum_{i=1}^{n}(x'_i)^2 \boldsymbol{W}_{ip} \tag{30}$$

Sendo ρ um parâmetro de trade-off entre a otimização da média e da variância o problema é formulado como a função objetivo:

$$Z_g(\rho) = \min \quad \max_{p \neq q}(\, |\mu_p(\boldsymbol{W}) - \mu_q(\boldsymbol{W})| + \rho|\sigma_p^2(\boldsymbol{W}) - \sigma_q^2(\boldsymbol{W})|) \tag{31}$$

Após a alocação dos elementos, o método propõe aleatorizar os tratamentos atribuídos a cada grupo. Um efeito da variação do parâmetro ρ sobre o balanceamento é apresentado na Figura 4, que exibe as discrepâncias obtidas para a média (vermelho) e variância (azul) entre dois grupos (em unidades de desvio padrão), resultante de 3000 alocações. Para cada alocação, 20 elementos foram gerados aleatoriamente com distribuição normal padrão. As bandas representam as médias dos valores de discrepância acima e abaixo da média: $\mathbb{E}[X|X \geq \mathbb{E}[X]]$ e $\mathbb{E}[X|X \leq \mathbb{E}[X]]$. As bandas na parte superior correspondem às discrepâncias obtidas por meio de aleatorização simples.

Figura 4 – Valores de discrepâncias obtidos para média (vermelho) e variância (azul), considerando $n = 20$ e $g = 2$.



Fonte: Adaptado de Bertsimas, Johnson e Kallus (2015).

Bertsimas, Johnson e Kallus (2015) argumentam que a minimização da distancia, considerando apenas os dois primeiros momentos ($\rho = 0.5$), produz redução da discrepância em momentos de maior ordem, quando comparada à aleatorização. Também é mencionado que variáveis não controladas, correlacionadas com aquelas consideradas na otimização, também são balanceadas.

A formulação escolhida (eq: 31) minimiza a distância máxima entre as combinações de pares de grupos. A formulação proposta no presente trabalho

baseia-se na minimização das distâncias entre cada grupo e o conjunto total de dados.

3 Extensões para o caso de múltiplos grupos

Este capítulo apresenta as formulações dos problemas de otimização nas normas L_1, L_∞ e na norma híbrida. Os modelos são baseados em técnicas programação por metas, cuja a base para as formulações no contexto das normas L_p é apresentada em Romero (1991). Nas subseções seguintes, são apresentadas, para cada norma, as formulações dos problemas originais (para dois grupos) e as propostas de generalização (para mais grupos de alocação). O problema de encontrar uma configuração que torne os dois grupos similares, buscando a minimização de uma distância genérica em L_p, pode ser descrito como um problema do tipo:

$$\min \|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_p \tag{32}$$

$$\text{sujeito a}$$

$$\sum_{i=1}^{n} w_i = n_1$$

$$\boldsymbol{w} \in \{0,1\}^n$$

$$i = 1 \ldots, n$$

Onde $\boldsymbol{w}$ é o vetor das alocações e $\boldsymbol{A}^*$ denota uma matriz genérica de covariáveis reais e artificiais, transformada pela fatoração de Cholesky conforme apresentado na Seção 2.3, Equação 25.

3.1 Considerações sobre a generalização dos modelos

O Método de Alocação Fortuita busca a minimização da distância entre vetores compostos pelas médias de dois grupos. A generalização desse método para o caso de g grupos ($g \geq 2$) envolve buscar alocações que minimizem os $\binom{g}{2}$ pares de distâncias entre grupos, como sugerido inicialmente em Lock (2011). Uma abordagem possível consiste na minimização das distâncias entre a média de cada grupo e a média global. Essas duas abordagens podem ser interpretadas como formas de minimização da variância entre as médias dos grupos, quando considerada norma L_2^2, conforme apresentado a seguir.

A variância de um conjunto de observações $\boldsymbol{x} = x_1, x_2 \dots x_n$ pode ser expressa tanto em termos da diferença entre cada observação e a média, $\sum_{i=1}^{n} \frac{x_i}{n} = \overline{x}$, quanto em termos das diferenças entre os pares de observações sendo possível estabelecer a relação:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i}^{n}(x_i - \overline{x})^2 \\
&= \frac{1}{n-1}\Big(\sum_{i}^{n} x_i^2 + \sum_{i}^{n}\overline{x}^2 - \sum_{i}^{n} 2x_i\overline{x}\Big) = \frac{1}{n-1}\Big(\sum_{i}^{n} x_i^2 + n\overline{x}^2 - 2n\overline{x}^2\Big) = \frac{1}{n-1}\Big(\sum_{i}^{n} x_i^2 - n\overline{x}^2\Big) \\
&= \frac{1}{2(n-1)}\Big(\big(\sum_{i}^{n} x_i^2 - n\overline{x}^2\big) + \big(\sum_{j}^{n} x_j^2 - n\overline{x}^2\big)\Big) = \frac{1}{2n(n-1)}\big(n\sum_{i}^{n} x_i^2 - 2\sum_{i}^{n}\sum_{j}^{n} x_i x_j + n\sum_{j}^{n} x_j^2\big) \\
&= \frac{1}{2n(n-1)}\sum_{i}^{n}\sum_{j}^{n}\big(x_i - x_j\big)^2 = \frac{1}{n(n-1)}\sum_{1\le i<j\le n}^{n}(x_i - x_j)^2
\end{aligned}
\tag{33}
$$

Essa equação demonstra a relação linear entre as duas abordagens, de modo que ao minimizar os desvios quadráticos de cada grupo à média global do conjunto de dados, os desvios entre as médias de todos os grupos será necessariamente minimizada. Essa relação também é válida para o caso multivariado, conforme demonstrado a seguir.

Um conjunto de dados $\boldsymbol{A^*} \in \mathbb{R}^{n\times m}$ com média $\overline{\boldsymbol{A^*}}$, pode ser particionado em g grupos de dimensões $n_k = n_1, n_2 \dots n_g$, $k = 1, 2 \dots g$, formando os subconjuntos $\boldsymbol{A^*}^{(k)} \in \mathbb{R}^{n_k\times m}$.

$$
\begin{aligned}
||\overline{\boldsymbol{A^*}}^{(1)} - \overline{\boldsymbol{A^*}}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(1)} - \overline{\boldsymbol{A^*}}_1|^2 + |\overline{\boldsymbol{A^*}}_2^{(1)} - \overline{\boldsymbol{A^*}}_2|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(1)} - \overline{\boldsymbol{A^*}}_m|^2 \\
||\overline{\boldsymbol{A^*}}^{(2)} - \overline{\boldsymbol{A^*}}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(2)} - \overline{\boldsymbol{A^*}}_1|^2 + |\overline{\boldsymbol{A^*}}_2^{(2)} - \overline{\boldsymbol{A^*}}_2|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(2)} - \overline{\boldsymbol{A^*}}_m|^2 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
||\overline{\boldsymbol{A^*}}^{(g)} - \overline{\boldsymbol{A^*}}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(g)} - \overline{\boldsymbol{A^*}}_1|^2 + |\overline{\boldsymbol{A^*}}_2^{(g)} - \overline{\boldsymbol{A^*}}_2|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(g)} - \overline{\boldsymbol{A^*}}_m|^2
\end{aligned}
\tag{34}
$$

De modo que:

$$
\sum_{k=1}^{g} ||\overline{\boldsymbol{A^*}}^{(k)} - \overline{\boldsymbol{A^*}}||_2^2 = \sum_{k=1}^{g} |\overline{\boldsymbol{A^*}}_1^{(k)} - \overline{\boldsymbol{A^*}}_1|^2 + \sum_{k=1}^{g} |\overline{\boldsymbol{A^*}}_2^{(k)} - \overline{\boldsymbol{A^*}}_2|^2 \cdots + \sum_{k=1}^{g} |\overline{\boldsymbol{A^*}}_m^{(k)} - \overline{\boldsymbol{A^*}}_m|^2 \tag{35}
$$

Enquanto o somatório dos $\binom{g}{2}$ desvios quadrádicos, para todo $1 \le p < q \le g$:

$$
\begin{aligned}
||\overline{\boldsymbol{A^*}}^{(1)} - \overline{\boldsymbol{A^*}}^{(2)}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(1)} - \overline{\boldsymbol{A^*}}_1^{(2)}|^2 + |\overline{\boldsymbol{A^*}}_2^{(1)} - \overline{\boldsymbol{A^*}}_2^{(2)}|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(1)} - \overline{\boldsymbol{A^*}}_m^{(2)}|^2 \\
||\overline{\boldsymbol{A^*}}^{(1)} - \overline{\boldsymbol{A^*}}^{(3)}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(1)} - \overline{\boldsymbol{A^*}}_1^{(3)}|^2 + |\overline{\boldsymbol{A^*}}_2^{(1)} - \overline{\boldsymbol{A^*}}_2^{(3)}|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(1)} - \overline{\boldsymbol{A^*}}_m^{(3)}|^2 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
||\overline{\boldsymbol{A^*}}^{(2)} - \overline{\boldsymbol{A^*}}^{(3)}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(2)} - \overline{\boldsymbol{A^*}}_1^{(3)}|^2 + |\overline{\boldsymbol{A^*}}_2^{(2)} - \overline{\boldsymbol{A^*}}_2^{(3)}|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(2)} - \overline{\boldsymbol{A^*}}_m^{(3)}|^2 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
||\overline{\boldsymbol{A^*}}^{(p)} - \overline{\boldsymbol{A^*}}^{(q)}||_2^2 &= |\overline{\boldsymbol{A^*}}_1^{(p)} - \overline{\boldsymbol{A^*}}_1^{(q)}|^2 + |\overline{\boldsymbol{A^*}}_2^{(p)} - \overline{\boldsymbol{A^*}}_2^{(q)}|^2 \cdots + |\overline{\boldsymbol{A^*}}_m^{(p)} - \overline{\boldsymbol{A^*}}_m^{(q)}|^2
\end{aligned}
\tag{36}
$$

De modo que:

$$\sum_{1\leq p<q\leq g}^{g} ||\overline{\boldsymbol{A}^*}^{(p)} - \overline{\boldsymbol{A}^*}^{(q)}||_2^2 = \sum_{1\leq p<q\leq g}^{g} |\overline{\boldsymbol{A}^*}_1^{(p)} - \overline{\boldsymbol{A}^*}_1^{(q)}|^2 + \sum_{1\leq p<q\leq g}^{g} |\overline{\boldsymbol{A}^*}_2^{(p)} - \overline{\boldsymbol{A}^*}_2^{(q)}|^2 \cdots + \sum_{1\leq p<q\leq g}^{g} |\overline{\boldsymbol{A}^*}_m^{(p)} - \overline{\boldsymbol{A}^*}_m^{(q)}|^2 \tag{37}$$

Pela relação dada no caso de uma variável na equação (33) temos que:

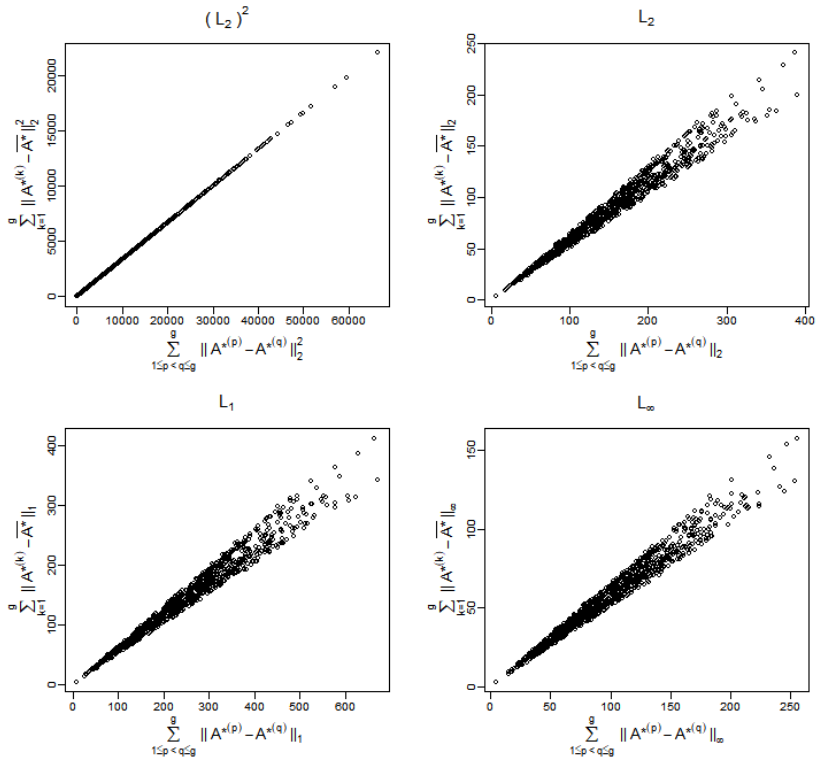
$$\begin{aligned}
\sum_{k=1}^{g} |\overline{\boldsymbol{A}^*}_1^{(k)} - \overline{\boldsymbol{A}^*}_1|^2 &= \frac{1}{g} \sum_{1\leq p<q\leq g}^{g} |\overline{\boldsymbol{A}^*}_1^{(p)} - \overline{\boldsymbol{A}^*}_1^{(q)}|^2 \\
\sum_{k=1}^{g} |\overline{\boldsymbol{A}^*}_2^{(k)} - \overline{\boldsymbol{A}^*}_2|^2 &= \frac{1}{g} \sum_{1\leq p<q\leq g}^{g} |\overline{\boldsymbol{A}^*}_2^{(p)} - \overline{\boldsymbol{A}^*}_2^{(q)}|^2 \\
&\vdots \qquad\qquad\qquad \vdots \\
\sum_{k=1}^{g} |\overline{\boldsymbol{A}^*}_m^{(k)} - \overline{\boldsymbol{A}^*}_m|^2 &= \frac{1}{g} \sum_{1\leq p<q\leq g}^{g} |\overline{\boldsymbol{A}^*}_m^{(p)} - \overline{\boldsymbol{A}^*}_m^{(q)}|^2
\end{aligned} \tag{38}$$

Das equações (33), (35) e (37), resulta que (HOPCROFT; KANNAN, 2014, p.264, Lemma 8.3):

$$\sum_{1\leq p<q\leq g}^{g} ||\overline{\boldsymbol{A}^*}^{(p)} - \overline{\boldsymbol{A}^*}^{(q)}||_2^2 = g \sum_{k=1}^{g} ||\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}||_2^2 \tag{39}$$

A Figura 5 apresenta resultados simulados para as duas abordagens. Apesar de a relação linear não se sustententar para as normas L_1 e L_∞, a forte correlação entre as duas abordagens torna viável a utilização prática da formulação baseada nas distâncias entre cada grupo e a média, e adequada para a finalidade do presente trabalho.

Figura 5 – Alocações aleatórias dos 54 elementos da base de dados das estações de ar (ver Seção 4.1) em 3 grupos de 18 para as Normas L_1, L_2, $(L_2)^2$ e L_∞. 1000 alocações para cada Norma.



Fonte: Elaborado pelo autor.

As subseções a seguir apresentam, para cada norma, as formulações originais do método de *Alocação Fortuita* (para dois grupos), seguidas pelas formulações que correspondem à extensão proposta neste trabalho.

3.2 Otimização no caso da norma L_1

3.2.1 Formulação para dois grupos (L_1)

Uma distância em um espaço normado L_1 pode ser calculada como $||\boldsymbol{x}-\boldsymbol{y}||_1 = \sum_{i=1}^{n} |x_i - y_i|$. A otimização do problema (32) para essa norma consiste em minimizar o somatório das diferenças entre as médias de cada covariável. Esse problema pode ser formulado como a minimização do valor de $\sum_{j=1}^{m} b_j$ para as restrições:

$$-b_j \leq \overline{\boldsymbol{A}}^{(1)}_{*j} - \overline{\boldsymbol{A}}^{(2)}_{*j} \leq b_j \tag{40}$$

onde $\overline{\boldsymbol{A}^*}^{(1)}$ e $\overline{\boldsymbol{A}^*}^{(2)}$ representam as médias de cada coluna de $\boldsymbol{A}^*$, para cada grupo de alocação:

$$\overline{\boldsymbol{A}^*}^{(1)} = \frac{\boldsymbol{A}^*\boldsymbol{w}}{n_1} \quad e \quad \overline{\boldsymbol{A}^*}^{(2)} = \frac{\boldsymbol{A}^*(\boldsymbol{1}-\boldsymbol{w})}{n_2} \tag{41}$$

Da equação 40, temos que $b_j \geq 0$ representa um limite para o módulo da diferença entre as médias da coluna $\boldsymbol{A^*}_j$ computadas sobre as unidades amostrais alocadas ao grupo 1 e 2. Logo, minimizar a soma de b_j significa minimizar a soma das diferenças entre essas médias. A diferença entre os dois grupos de alocação pode ser representada como:

$$\frac{\boldsymbol{A}^*\boldsymbol{w}}{n_1} - \frac{\boldsymbol{A}^*(\boldsymbol{1}-\boldsymbol{w}^t)}{n_2} = \frac{n_1+n_2}{n_1.n_2}\boldsymbol{A}^*\boldsymbol{w} - \boldsymbol{1}\frac{\boldsymbol{A}^*}{n_2} \tag{42}$$

Definimos que:

$$\boldsymbol{R} = \frac{n_1+n_2}{n_1 n_2}\boldsymbol{A}^{*t} \qquad e \qquad \boldsymbol{s} = \boldsymbol{1}\frac{\boldsymbol{A}^*}{n_2} \tag{43}$$

Definimos que $\boldsymbol{R} = r_{j,i}$ é uma matriz $m \times n$, e $\boldsymbol{s} = s_j$ um vetor de dimensão m, sendo possível definir as restrições da formulação L_1 como:

$$r_{j,\bullet}\boldsymbol{w} - s_j \leq b_j \tag{44}$$

$$r_{j,\bullet}\boldsymbol{w} - s_j \geq -b_j \tag{45}$$

Que também podem ser escritas como:

$$r_{j,\bullet}\boldsymbol{w} - b_j \leq s_j \tag{46}$$

$$-r_{j,\bullet}\boldsymbol{w} - b_j \leq -s_j \tag{47}$$

Assim, é possível formular um modelo para a otimização da distância nessa Norma como:

$$\min \sum_{j=1}^{m} b_j, \tag{48}$$

sujeito a

$$\sum_{i=1}^{n} r_{ji}w_i - b_j \leq s_j,$$

$$\sum_{i=1}^{n} -r_{ji}w_i - b_j \leq -s_j,$$

$$\sum_{i=1}^{n} w_i = n_1$$

$$w \in \{0,1\}^n$$

$$i = 1 \dots n, \qquad j = 1 \dots m, \qquad b \geq 0$$

3.2.2 Formulação proposta para mais de dois grupos (L_1)

Um modelo generalizado para mais de dois grupos pode ser formulado por meio da minimização da distância entre cada grupo e a média global. Na Norma L_1, o somatório das distâncias de g vetores em relação a um vetor de médias globais, $\overline{\boldsymbol{x}}$, é representado pelo somatório das diferenças: $\sum_{k=1}^{g} |\boldsymbol{x}^{(k)} - \overline{\boldsymbol{x}}| = \sum_{j=1}^{m}(|x_j^{(1)} - \overline{x}_j| + |x_j^{(2)} - \overline{x}_j| \dots + |x_j^{(g)} - \overline{x}_j|)$, de modo que uma generalização da função objetivo (32) pode ser formulada como:

$$\min \sum_{k=1}^{g} \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_1 \tag{49}$$

Para esse caso, a designação de cada elemento a um grupo é denotada pela matriz de alocação $\boldsymbol{W} = w_{i,k} \in \{0,1\}^{n \times g}$, onde $w_{i,k} = 1$ representa a alocação do elemento i ao grupo k. A quantidade de elementos em cada grupo é representada pelo vetor $\boldsymbol{n} = \{n_1, n_2 \dots n_g\}$, estabelecendo as restrições $\sum_{1}^{n} w_{\bullet,k}^t = n_k$. A minimização da distância entre as médias de cada grupo, $\overline{\boldsymbol{A}^*}^{(k)} = \frac{\boldsymbol{A}^* w_{\bullet,k}}{n_k}$, e o vetor das médias globais das covariáveis, $\overline{\boldsymbol{A}^*} = \boldsymbol{1}\frac{\boldsymbol{A}^*}{n}$, pode ser formulada com a minimização de $\sum_{j=1}^{m}\sum_{k=1}^{g} b_{j,k}$ em:

$$-b_{jk} \leq \overline{\boldsymbol{A}^*}_j^{(k)} - \overline{\boldsymbol{A}^*}_j \leq b_{jk} \tag{50}$$

Uma possível definição das restrições desse problema de Programação Linear Inteira Mista é apresentada a seguir: Considere as sequência de matrizes $\boldsymbol{Q}^{(k)} = q_{j,i}^{(k)}$:

$$\boldsymbol{Q}^{(k)} = \frac{\boldsymbol{A}^{*t}}{n_k} \tag{51}$$

De modo que é possível escrever (50) com as restrições:

$$q_{j,\bullet}^{(k)}\, w_{\bullet,k} - b_{j,k} \leq \overline{\boldsymbol{A}}_j^* \tag{52}$$

$$-q_{j,\bullet}^{(k)}\, w_{\bullet,k} - b_{j,k} \leq \overline{\boldsymbol{A}}_j^* \tag{53}$$

Definindo a matriz de valores ideais, $\boldsymbol{S} = s_{j,k}$, formada por g colunas que representam as médias globais, $\overline{\boldsymbol{A}^*}$, um modelo considerando essa generalização pode ser formulado como o problema:

$$\min \sum_{j=1}^{m} \sum_{k=1}^{g} b_{jk}, \tag{54}$$

$$\text{sujeito a}$$

$$\sum_{i=1}^{n} q_{jik} w_{ik} - b_{jk} \leq s_{jk},$$

$$\sum_{i=1}^{n} -q_{jik} w_{ik} - b_{jk} \leq -s_{jk},$$

$$\sum_{i=1}^{n} w_{ik} = n_k,$$

$$\sum_{k=1}^{g} w_{ik} = 1,$$

$$w \in \{0,1\}^{n \times g}$$

$$w_{ik} = 0 \quad \forall i < k$$

$$k = 1 \ldots g, \quad i = 1 \ldots n, \quad j = 1 \ldots m, \quad b \geq 0$$

3.3 Otimização no caso da norma L_∞

3.3.1 Formulação do problema para dois grupos (L_∞)

Uma distância entre dois vetores na norma L_∞ pode ser definida como $||(\boldsymbol{x} - \boldsymbol{y})||_\infty = \max |x_i - y_i|$. Esse problema pode ser formulado como a minimização do valor de b considerando a restrição:

$$-b \leq \frac{\boldsymbol{A}^* \boldsymbol{w}}{n_1} - \frac{\boldsymbol{A}^* (\boldsymbol{1} - \boldsymbol{w})}{n_2} \leq b \tag{55}$$

A diferença entre os grupos de alocação pode ser apresentada considerando as definições em (43),

$$r_{j,\bullet} \boldsymbol{w} - b \leq s_j \tag{56}$$

$$-r_{j,\bullet} \boldsymbol{w} - b \leq -s_j \tag{57}$$

Sendo possível formular um modelo para a otimização da distância nessa Norma como:

$$\min \ b, \tag{58}$$

$$\text{sujeito a}$$

$$\sum_{i=1}^{n} r_{ji} w_i - b \leq s_j$$

$$\sum_{i=1}^{n} -r_{ji} w_i - b \leq -s_j$$

$$\sum_{i=1}^{n} w_i = n_1$$

$$w \in \{0,1\}^n$$

$$j = 1 \ldots m, \quad i = 1 \ldots n, \quad b \geq 0$$

3.3.2 Formulação proposta para mais de dois grupos (L_∞)

Na Norma L_∞, as distâncias de g vetores em relação a um vetor de médias globais, $\overline{\boldsymbol{x}}$, podem ser consideradas como o máximo das diferenças entre as médias de cada covariável: $max|\boldsymbol{x}^{(k)} - \overline{\boldsymbol{x}}|$, de modo que uma generalização da função objetivo (32) pode ser formulada como:

$$\min \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_\infty \tag{59}$$

Que, diante das mesmas definições de (50), pode ser escrita como:

$$-b \leq \overline{\boldsymbol{A}^*}_j^{(k)} - \overline{\boldsymbol{A}^*}_j \leq b \tag{60}$$

De modo que um modelo de nessa norma pode ser formulado pela minimização de b com as restrições:

$$q_{j,\bullet}^{(k)} \, w_{\bullet,k} - b \leq \overline{\boldsymbol{A}}_j^* \tag{61}$$

$$-q_{j,\bullet}^{(k)} \, w_{\bullet,k} - b \leq \overline{\boldsymbol{A}}_j^* \tag{62}$$

Assim, o problema generalizado de minimização em L_∞ pode ser formulado como:

$$\min \quad b, \tag{63}$$

$$\text{sujeito a}$$

$$\sum_{i=1}^{n} q_{jik} w_{ik} - b \leq s_{jk},$$
$$\sum_{i=1}^{n} -q_{jik} w_{ik} - b \leq -s_{jk},$$
$$\sum_{i=1}^{n} w_{ik} = n_k,$$
$$\sum_{k=1}^{g} w_{ik} = 1,$$
$$w_{ik} = 0 \quad \forall i < k$$
$$w \in \{0,1\}^{n \times g}$$
$$k = 1 \ldots g, \quad i = 1 \ldots n, \quad j = 1 \ldots m, \quad b \geq 0$$

3.4 Otimização no caso da norma híbrida (L_1, L_∞)

3.4.1 Formulação do problema para dois grupos (L_1, L_∞)

A norma híbrida, baseada na publicação de Ward e Wendell (1980), pode ser considerada uma aproximação da norma L_2 (ver Seção 2.2.1). A otimização da função objetivo (32) nessa norma pode ser apresentada como:

$$\min \|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_{(\gamma 1, \gamma 2)} = \min \left(\gamma_1 \|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_1 + \gamma_2 \sqrt{m} \|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_\infty \right) \tag{64}$$

Esse problema pode ser formulado como a minimização do valor de $\gamma_1 \sum_{j=1}^{m} b_j + \gamma_2 \sqrt{m}\, d$ considerando as restrições:

$$-b_j \leq \overline{\boldsymbol{A}^*}_j^1 - \overline{\boldsymbol{A}^*}_j^2 \leq b_j \tag{65}$$

$$-b_j + \ d \geq 0 \tag{66}$$

A restrição (65) pode ser definida em termos de $\mathbf{R}$ e $\mathbf{s}$ (equação (43)), como nos demais casos. A adição da restrição (66) garante que $\max \|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_1 = \|\overline{\boldsymbol{A}^*}^{(1)} - \overline{\boldsymbol{A}^*}^{(2)}\|_\infty$.

$$r_{j,\bullet} \boldsymbol{w} - b_j \leq s_j \tag{67}$$

$$-r_{j,\bullet} \boldsymbol{w} - b_j \leq -s_j \tag{68}$$

$$b_j + \ d \geq 0 \tag{69}$$

Considerando $\gamma_1 = \gamma_2 = 1$, o problema pode ser formulado nessa norma como:

$$\min \sum_{j=1}^{m} b_j + \sqrt{m}\, d \tag{70}$$

$$\text{sujeito a}$$

$$\sum_{i=1}^{n} r_{ji} w_i - b_j \leq s_j,$$

$$\sum_{i=1}^{n} -r_{ji} w_i - b_j \leq -s_j,$$

$$-b_j + \ d \geq 0$$

$$\sum_{i=1}^{n} w_i = n_1$$

$$w \in \{0,1\}^n$$

$$j = 1 \ldots m, \quad i = 1 \ldots n, \quad b \geq 0, \quad d \geq 0$$

3.4.2 Formulação proposta para mais de dois grupos (L_1, L_∞)

Para o caso de múltiplos grupos, a otimização da função objetivo (32) nessa norma pode ser apresentada como:

$$\min \sum_{k=1}^{g} \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_{(\gamma 1, \gamma 2)} = \min \sum_{k=1}^{g} \left(\gamma_1 \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_1 + \gamma_2 \sqrt{m} \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_\infty \right) \tag{71}$$

Esse problema pode ser formulado como a minimização do valor de $\gamma_1 \sum_{j=1}^{m} \sum_{k=1}^{g} b_{j,k} + \gamma_2 \sqrt{m} \sum_{k=1}^{g} d$ considerando as restrições:

$$-b_{j,k} \leq \overline{\boldsymbol{A}^*}_j^{(k)} - \overline{\boldsymbol{A}^*}_j \leq b_{j,k} \tag{72}$$

$$-b_{j,k} + \ d_k \geq 0 \tag{73}$$

Essas restrições podem ser definidas considerando a matriz $\boldsymbol{Q}$ (equação (51)) como nos demais casos.

$$q_{j,\bullet}^{(k)} w_{\bullet,k} - b_{j,k} \leq \overline{\boldsymbol{A}}_j^* \tag{74}$$

$$-q_{j,\bullet}^{(k)} w_{\bullet,k} - b_{j,k} \leq -\overline{\boldsymbol{A}}_j^* \tag{75}$$

$$-b_{j,k} + \ d_k \geq 0 \tag{76}$$

Considerando $\gamma_1 = \gamma_2 = 1$. Assim como no caso de dois grupos, a restrição (76) garante que $\max \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_1 = \|\overline{\boldsymbol{A}^*}^{(k)} - \overline{\boldsymbol{A}^*}\|_\infty$. O problema pode ser formulado nessa norma como:

$$\min \ \sum_{j=1}^{m} \sum_{k=1}^{g} b_{jk} + \sqrt{m} \ d_k \tag{77}$$

sujeito a

$$\sum_{i=1}^{n} q_{jik} w_{ik} - b_{jk} \leq s_{jk},$$

$$\sum_{i=1}^{n} -q_{jik} w_{ik} - b_{jk} \leq -s_{jk},$$

$$-b_{jk} + \ d_k \geq 0$$

$$\sum_{i=1}^{n} w_{ik} = n_k$$

$$\sum_{k=1}^{g} w_{ik} = 1,$$

$$w_{ik} = 0 \quad \forall i < k$$

$$w \in \{0,1\}^{n \times g}$$

$$k = 1 \ldots g, \quad i = 1 \ldots n, \quad j = 1 \ldots m, \quad b \geq 0, \quad d \geq 0$$

3.5 Avaliação do desacoplamento no caso de múltiplos grupos

O coeficiente de associação apresentado por Yule (1912), considerado na formulação original (equação (23)), é utilizado para verificar o grau de concordância de uma variável binária, portanto, inadequado como medida de desacoplamento para mais de dois grupos. A métrica utilizada neste trabalho é o coeficiente Kappa de Fleiss (FLEISS, 1971), uma métrica inicialmente idealizada para mensurar no grau de concordância entre múltipos avaliadores. Essa métrica mostra-se adequada para medir o grau de concordância das repetições do processo de alocação. Como o método de *Alocação Intencional Fortuita* pode variar de totalmente determinístico a totalmente aleatório, espera-se um decréscimo no grau de concordância das diversas alocações com a variação do parâmetro λ, definido na seção 2.3.

No contexto dos experimentos conduzidos neste trabalho, esse coeficiente busca medir o grau de concordância de R repetições do processo de alocação de n elementos da base de dados em g grupos. Sendo $R_{i,k}$ a quantidade de vezes que o elemento i foi alocado no grupo k, temos que p_k é a proporção de todas as alocações ao grupo k.

$$p_k = \frac{1}{Rn} \sum_{i=1}^{n} R_{i,k} \tag{78}$$

O grau de concordância para as alocações do elemento i no total de repetições do processo de alocação pode ser concebido pela proporção de pares concordantes dentre as $R(R-1)$ permutações possíveis.

$$P_i = \frac{1}{R(R-1)} \sum_{k=1}^{g} R_{i,k}(R_{i,k} - 1) = \frac{1}{R(R-1)} (\sum_{k=1}^{g} R_{i,k}^2 - R) \tag{79}$$

Assim, a média do grau de concordância para todos os elementos pode ser obtida por:

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \frac{1}{nR(R-1)}(\sum_{i=1}^{n}\sum_{k=1}^{g} R_{i,k}^2 - nR) \tag{80}$$

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{g} \frac{R_{i,k}(R_{i,k}-1)}{R(R-1)} \tag{81}$$

Em uma situação onde os elementos são alocados nos grupos aleatoriamente, temos que o grau de concordância esperado é:

$$\overline{P}_e = \sum_{k=1}^{g} (\frac{1}{n}\sum_{i=1}^{n} \frac{R_{i,k}}{R})^2 \tag{82}$$

O coeficiente Kappa (κ) é a relação entre diferença do grau de concordancia obtido e o grau de concordancia esperado no caso aleatório, $\overline{P}_i - \overline{P}_e$ e a diferença entre um caso de concordância perfeita e o esperado no caso aleatório, $1 - \overline{P}_e$.

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{83}$$

Em geral, o coeficiente κ varia no intervalo [0,1]. Em casos específicos, onde o grau de discordância é superior ao esperado pelo caso aleatório, esse coeficiente pode ser negativo.

4 Estudos de caso

4.1 Estudo de Caso - Estações de monitoramento

Esta Seção apresenta uma extensão do caso discorrido por Lauretto *et al.* (2019), onde é analisado o problema de selecionar estações de monitoramento atmosférico para a instalação de um sensor que, devido ao alto custo, não pode contemplar todas as unidades. A base de dados (COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO, 2018) é composta por 54 estações de monitoramento e 18 covariáveis. Além das coordenadas geográficas da estação, foram consideradas as seguintes variáveis, que correspondem à observações de um ano (agosto de 2017 a julho de 2018): material particulado de 10 micrômetros (PM10), monóxido de nitrogênio (NO), dióxido de nitrogênio (NO2), Óxidos de nitrogênio (NOx), ozônio (O3), temperatura do ar (Temp), umidade relativa (UR) e velocidade do vento (WS). Para cada uma dessas 8 variáveis foram obtidas as medianas das observações nos períodos de chuva (outubro a março) e seca (abril a setembro).

No presente caso, é considerado um cenário hipotético em que diferentes tipos de sensores devem ser alocados nas estações, mas não é possível a instalação de mais de um tipo de sensor na mesma estação, de modo que é necessário selecionar mais de dois grupos.

Com o uso do método proposto no presente trabalho, foram analisados os casos de 3, 4 e 5 grupos. Algumas adaptações foram necessárias em relação ao caso estudado em Lauretto *et al.* (2019). A métrica de balanceamento considerada no caso de múltiplos grupos é a diferença da média de cada grupo em relação à média do conjunto de dados (ver Seção 3.1). O critério de desacoplamento considerado foi o coeficiente Kappa de Fleiss, apresentado na Seção 3.5. Assim como realizado em Lauretto *et al.* (2019), foi realizada uma avaliação do trade-off entre balanceamento e desacoplamento proporcionados pela proposta de extensão ao método *Alocação Intencional Fortuita* ($\lambda^* \in \{0.001, 0.05, 0.1, 0.2, 0.3, 0.4\}$) e a comparação com os métodos de *Rerandomização* e *Aleatorização pura*.

Para cada método, foram realizadas 100 repetições do processo de alocação, cada uma considerando o tempo de 300 segundos. Neste e nos demais estudos de caso deste trabalho os experimentos computacionais foram realizados em um computador com processador AMD RYZEN 1920X (3.5 ghz 12 núcleos, 24 threads), placa mãe ASROCK x399, 32GB RAM DDR4, Linux Ubuntu 18.04.5 LTS Server. As rotinas foram implementadas com o uso de software R. v. 3.6.1. (R Core Team, 2020) e os problemas de otimização linear foram reslvidos com a utilização do software Gurobi v. 9.0.1 (GUROBI, 2021).

4.1.1 Resultados - balanceamento x desacoplamento

A tabela 1 apresenta os valores obtidos com a variação de λ^* e os outros dois métodos. Destaca-se que o método de *Alocação Fortuita* proporcionou os melhores balanceamentos, mesmo para os maiores valores de λ^* considerados. Destaca-se também que o método de *Rerandomização* obteve praticamente os mesmos valores de desacoplamento que o método de *Aleatorização pura*.

Tabela 1 – Distâncias de Mahalanobis e Kappa de Fleiss 100 alocações de 300 segundos para cada método

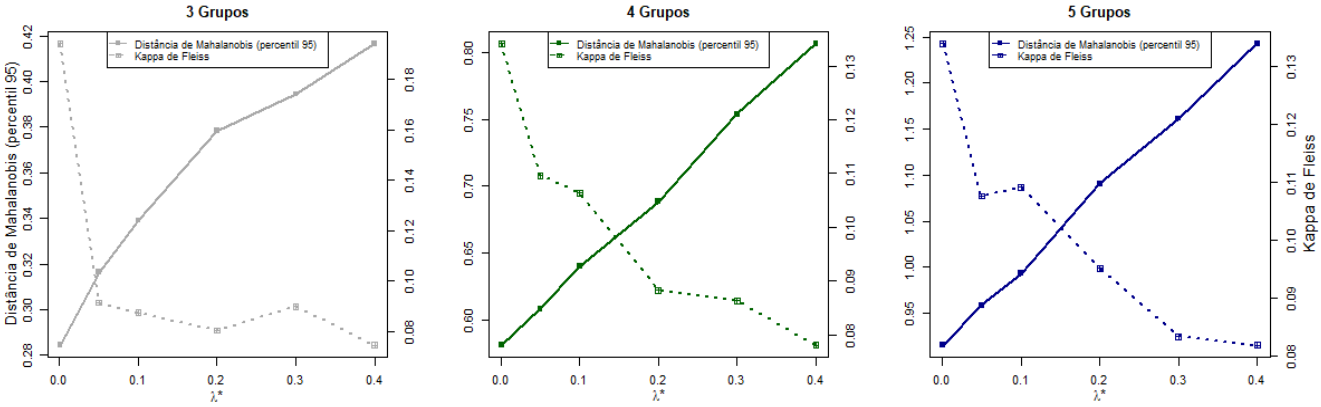
		Distância de Mahalanobis		
	Método	**Mediana**	**Percentil 95**	**Kappa**
3 Grupos (18, 18, 18)	Aloc. Fortuita ($\lambda^* = 0.01$)	0.27	0.28	0.194
	Aloc. Fortuita ($\lambda^* = 0.05$)	0.30	0.32	0.091
	Aloc. Fortuita ($\lambda^* = 0.1$)	0.32	0.34	0.087
	Aloc. Fortuita ($\lambda^* = 0.2$)	0.34	0.38	0.080
	Aloc. Fortuita ($\lambda^* = 0.3$)	0.36	0.39	0.090
	Aloc. Fortuita ($\lambda^* = 0.4$)	0.39	0.42	0.074
	Rerandomização	0.77	0.81	0.001
	Aleatorização pura	1.37	1.61	0.001
4 Grupos (13, 13, 14, 14)	Aloc. Fortuita ($\lambda^* = 0.01$)	0.55	0.58	0.134
	Aloc. Fortuita ($\lambda^* = 0.05$)	0.57	0.61	0.110
	Aloc. Fortuita ($\lambda^* = 0.1$)	0.60	0.64	0.106
	Aloc. Fortuita ($\lambda^* = 0.2$)	0.65	0.69	0.088
	Aloc. Fortuita ($\lambda^* = 0.3$)	0.69	0.75	0.086
	Aloc. Fortuita ($\lambda^* = 0.4$)	0.73	0.81	0.078
	Rerandomização	1.35	1.40	0.001
	Aleatorização pura	1.98	2.25	0.001
5 Grupos (10, 11, 11, 11, 11)	Aloc. Fortuita ($\lambda^* = 0.01$)	0.87	0.92	0.134
	Aloc. Fortuita ($\lambda^* = 0.05$)	0.90	0.96	0.108
	Aloc. Fortuita ($\lambda^* = 0.1$)	0.94	0.99	0.109
	Aloc. Fortuita ($\lambda^* = 0.2$)	1.01	1.09	0.095
	Aloc. Fortuita ($\lambda^* = 0.3$)	1.07	1.16	0.083
	Aloc. Fortuita ($\lambda^* = 0.4$)	1.15	1.24	0.082
	Rerandomização	1.89	1.94	0.000
	Aleatorização pura	2.57	2.78	0.001

Fonte: Autor, adaptado de Lauretto *et al.* (2019).

A Figura 6 exibe o comportamento dos aspectos analisados diante da variação de λ^*. Com $\lambda^* = 0.01$, a perturbação aleatória introduzida pelas variáveis artificiais no problema de otimização é pequena, portanto, espera-se um valor de Kappa próximo de 1. À medida em que λ^* aumenta, espera-se um decréscimo no valor de Kappa. Contudo, observa-se que, para $\lambda^* = 0.01$, o valor de Kappa já é relativamente baixo. Isso pode ser atribuído ao tempo limite

para a definição da alocação (300s), que não é suficiente para se obter a solução ótima do problema de otimização. Pela mesma razão, observa-se também que, embora haja uma tendência no decréscimo do coeficiente Kappa com o incremento de λ^*, este decréscimo não é perfeitamente monótono. Esse mesmo fator também pode justificar curvas mais suavisadas para os casos de 4 e 5 grupos, tendo em vista que esses casos demandam mais tempo para que sejam encontradas soluções próximas da ótima.

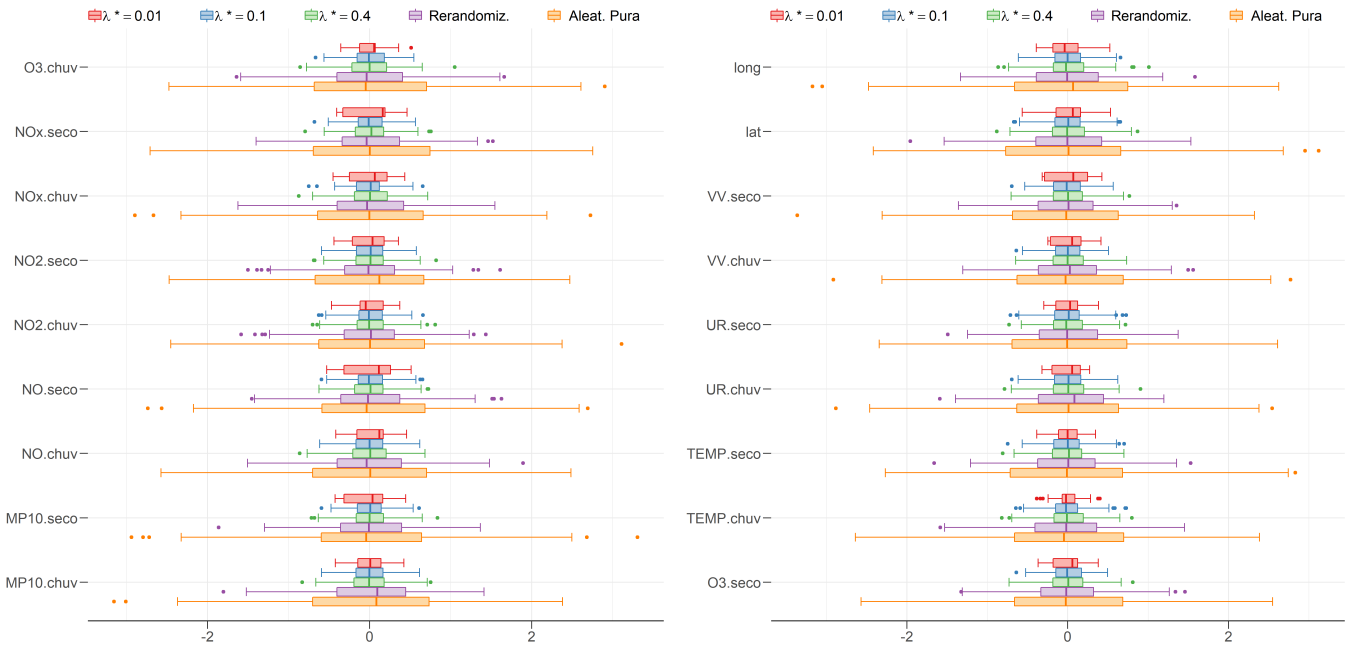
Figura 6 – Distância de Mahalanobis e Kappa de Fleiss em relação à λ^* para o método de Alocação Fortuita.



Fonte: Autor, adaptado de Lauretto *et al.* (2019).

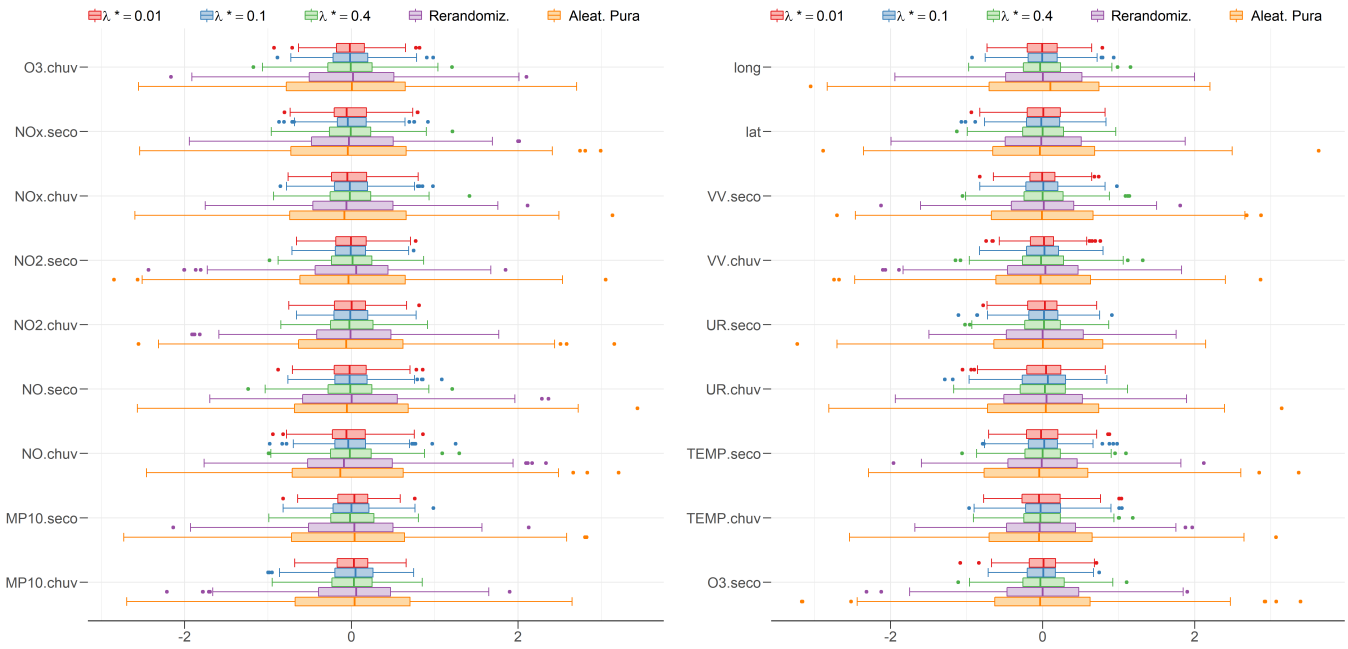
A avaliação do balanceamento em cada covariável pode ser verificada pelas diferenças entre as médias de cada grupo e da base de dados global, ou seja, $\overline{\boldsymbol{X}}_j^{(k)} - \overline{\boldsymbol{X}}_j$, onde $\overline{\boldsymbol{X}}$ é o vetor das médias das covariáveis da base de dados e $\overline{\boldsymbol{X}}^{(k)}$ é o vetor de médias dos elementos alocados no grupo k. As figuras 7, 8 e 9 apresentam as distribuições das diferenças em cada uma das 18 covariáveis . Os valores estão normalizados pelo desvio padrão das diferenças obtidas pelo método de aleatorização pura, ou seja, $(\overline{\boldsymbol{X}}_j^{(k)} - \overline{\boldsymbol{X}}_j)/s_j$, onde s_j é o desvio padrão das diferenças $(\overline{\boldsymbol{X}}_j^{(k)} - \overline{\boldsymbol{X}}_j)$ obtidas para as $100 \times g$ alocações utilizando métoto de aleatorização pura.

Figura 7 – Diferenças entre as médias de cada um dos 3 Grupos (18, 18, 18) e a média global, considerando os métodos Alocaçã Fortuita, Rerandomização e Aleatorização simples.



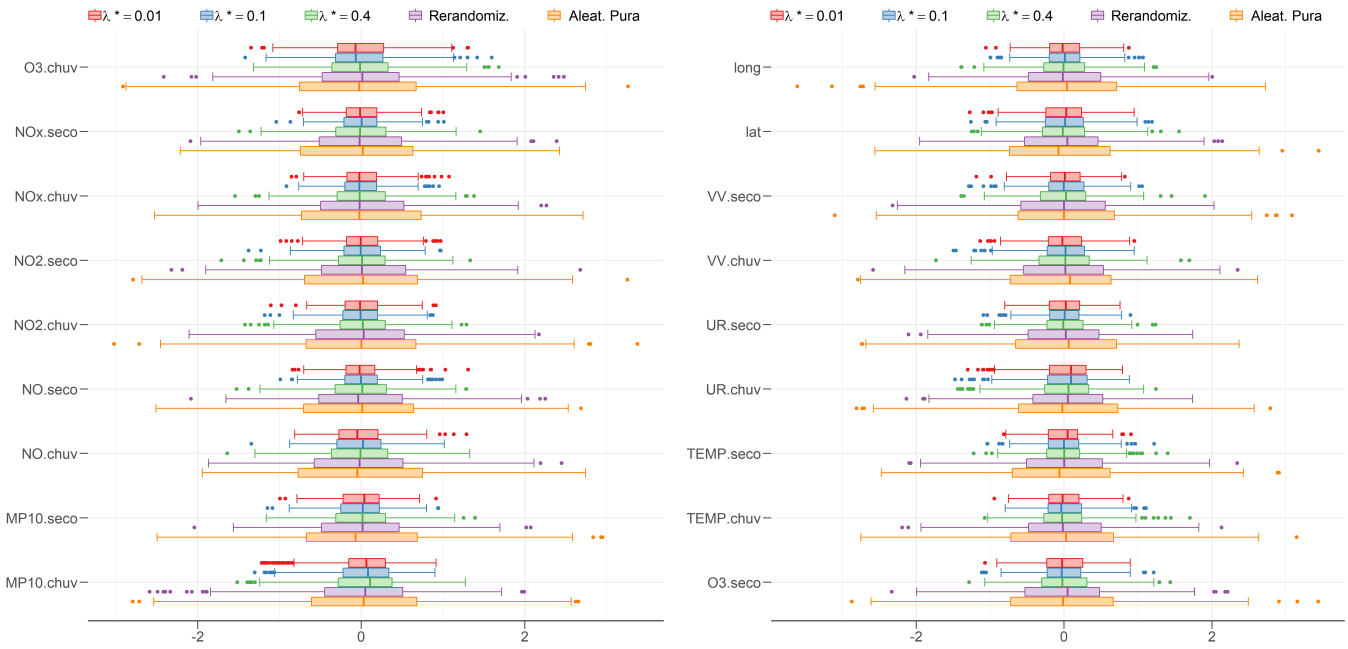
Fonte: Elaborado pelo autor.

Figura 8 – Diferenças entre as médias de cada um dos 4 Grupos (13, 13, 14, 14) e a média global, considerando os métodos Alocação Fortuita, Rerandomização e Aleatorização simples.



Fonte: Elaborado pelo autor.

Figura 9 – Diferenças entre as médias de cada um dos 5 Grupos (10, 11, 11, 11, 11) e a média global, considerando os métodos Alocação Fortuita, Rerandomização e Aleatorização simples.



Fonte: Elaborado pelo autor.

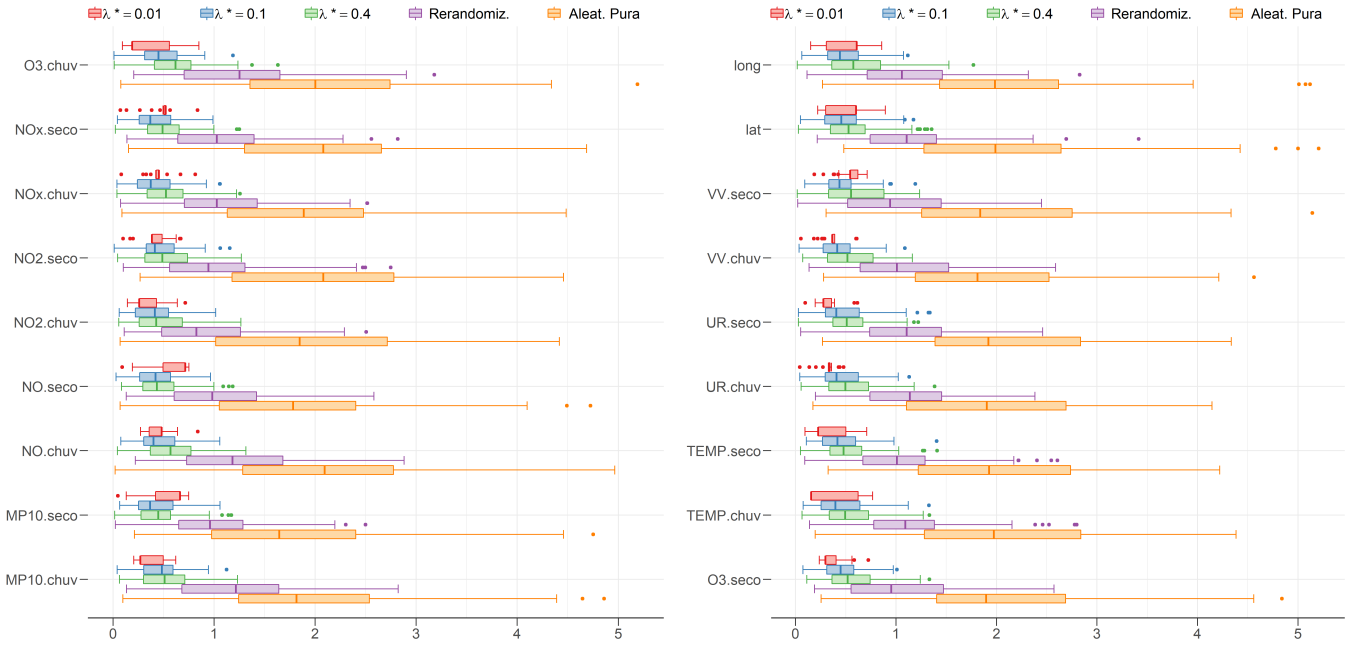
Com o aumento da quantidade de grupos é observado um aumento das diferenças nas covariáveis. Isso se justifica, em parte, pelo Teorema do Limite Central, uma vez que amostras menores tendem a apresentar maiores diferenças em relação à média populacional, ou seja, a média amostral de uma amostra de tamanho 11 (5 grupos) tende a apresentar maior diferença em relação à média global dos 54 elementos, quando comparada a uma amostra de tamanho 18 (3 grupos).

As alocações obtidas por meio de Aleatorização Pura apresentaram as maiores diferenças e extrapolaram 2 desvios-padrões nos três casos. As alocações obtidas pelo método de *Alocação Fortuita* apresentam as menores diferenças em todas as covariáveis, mesmo quando a perturbação adicionada é relativamente elevada ($\lambda^* = 0.4$). No caso de 3 grupos, essas diferenças não ultrapassam 1 desvio-padrão (calculado sobre as alocações obtidas por *Aleatorização Pura*).

O balanceamento no caso de múltiplos grupos de alocação também pode ser medido como as diferenças máximas entre os pares. Nesse caso, a medida de balanceamento para cada covariável foi obtida pela diferença máxima entre cada uma das $\binom{g}{2}$ combinações de pares (p, q) de grupos de alocação. As figuras 10, 11 e 12 apresentam as distribuições dos valores de $\max |\overline{\boldsymbol{X}}_j^{(p)} - \overline{\boldsymbol{X}}_j^{(q)}|/s_j$ obtidos nas 100 repetições, para cada método, onde s_j é o desvio padrão dos resultados de $\max |\overline{\boldsymbol{X}}_j^{(p)} - \overline{\boldsymbol{X}}_j^{(q)}|$ utilizando o método de *aleatorização pura*.

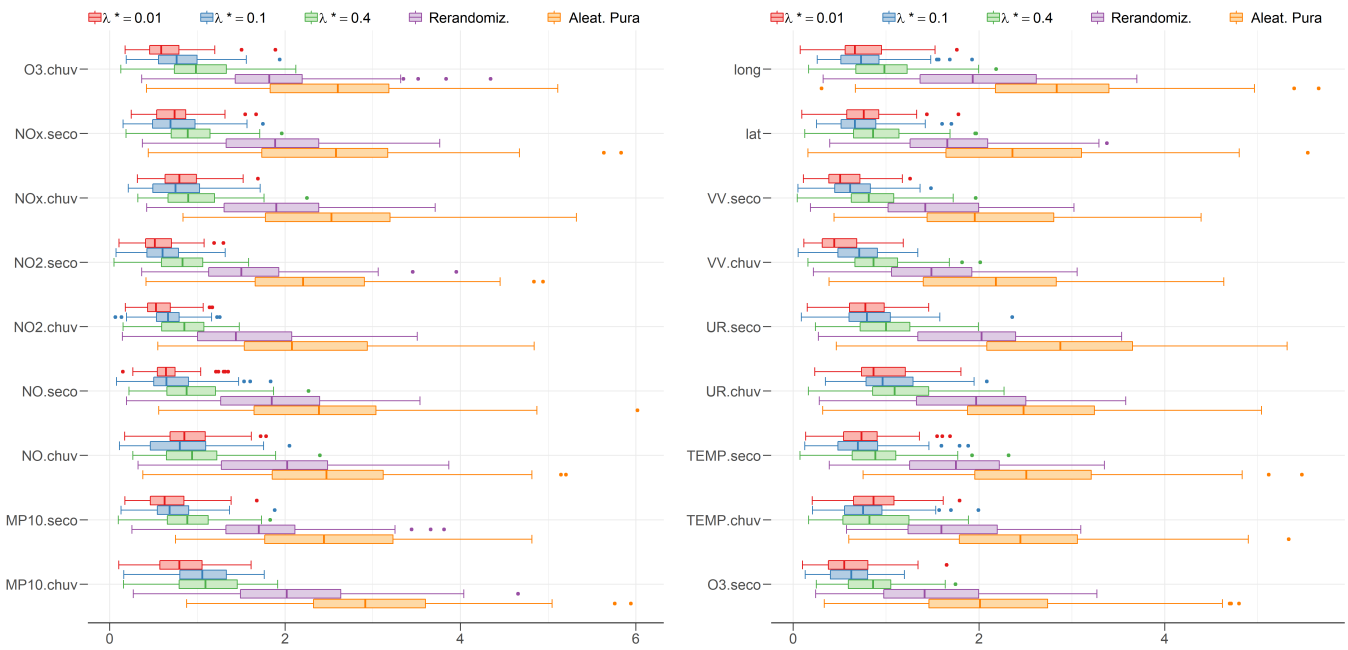
Destaca-se que, apesar de o modelo formulado ser baseado na minimização das distâncias entre os grupos e a média global, a generalização proposta foi capaz de reduzir substancialmente a diferença máxima pareada entre os grupos de alocação, bem como preservar a coerência dos resultados em relação à variação de λ^* e os demais métodos, quando comparado aos resultados das figuras 7 a 9, conforme discorrido na Seção 3.1.

Figura 10 – Diferenças máximas entre as combinações de pares, para cada método - 3 Grupos (18, 18, 18).



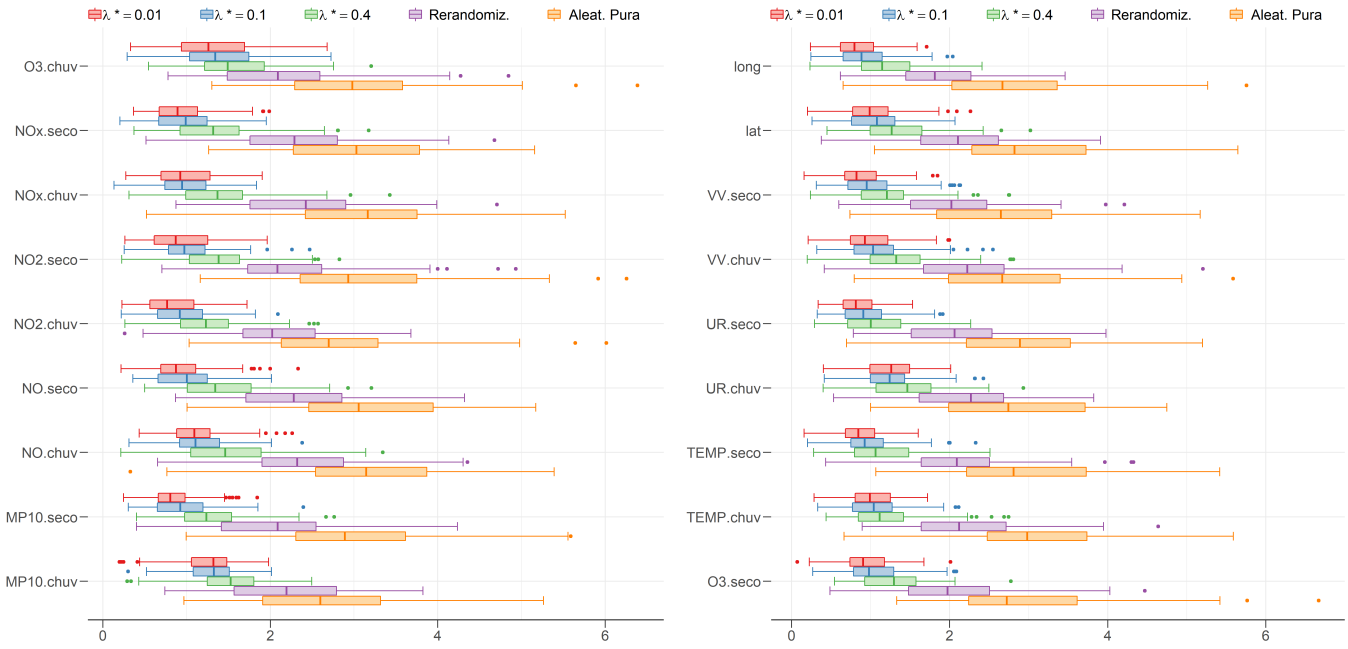
Fonte: Elaborado pelo autor.

Figura 11 – Diferenças máximas entre as combinações de pares, para cada método - 4 Grupos (13, 13, 14, 14).



Fonte: Elaborado pelo autor.

Figura 12 – Diferenças máximas entre as combinações de pares, para cada método - 5 Grupos (10, 11, 11, 11, 11).



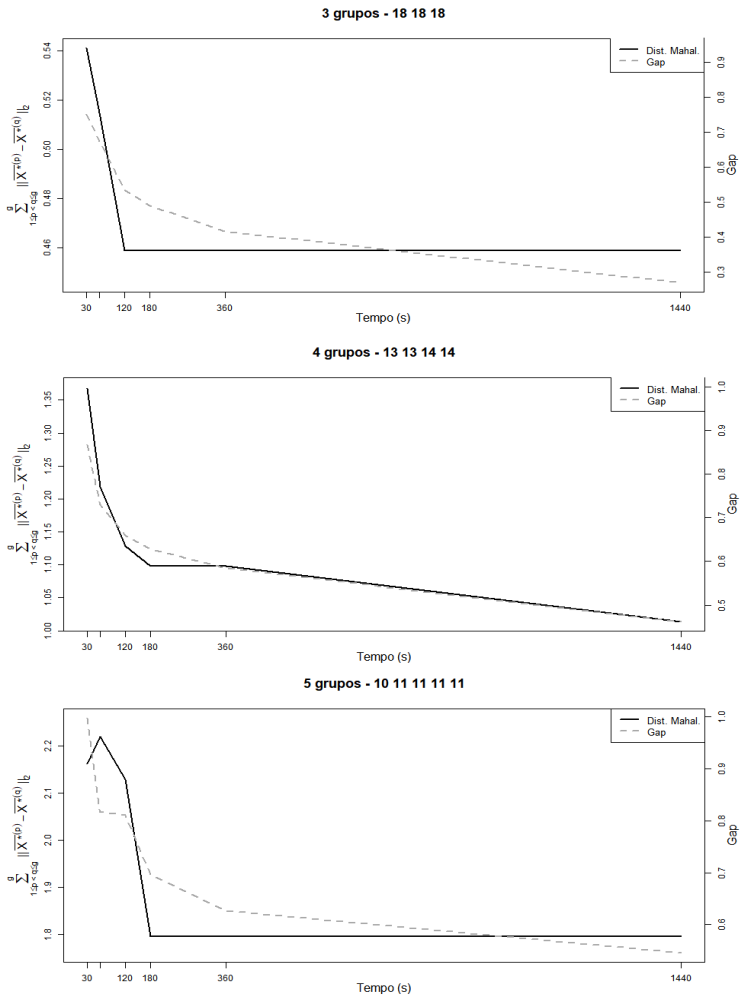
Fonte: Elaborado pelo autor.

4.1.2 Experimentos auxiliares

A definição de parâmetros utilizados no experimento ensejou uma avaliação empírica, de modo a verificar a capacidade do modelo minimizar o somatório da distância entre as combinações de pares de grupos e o *Gap* relativo ($\frac{Sol.Incumbente - LimitanteInferior}{Sol.Incumbente}$). Os resultados são apresentados na Figura 13 e consideram o caso determinístico ($\lambda = 0$). O tempo definido (300s) concilia uma estabilização dos resultados de otimização com a viabilidade de realizar um número apropriado de repetições dos procedimentos de alocação.

O comportamento observado na faixa de 30s à 60s, no caso de 5 grupos, deve-se ao fato de o modelo proposto minimizar as distâncias em relação ao centróide e ser baseado na norma híbrida, de modo que a otimização de $\sum_{i=1}^{k} ||\overline{\boldsymbol{X}^*}^{(k)} - \overline{\boldsymbol{X}^*}||_{(\gamma 1, \gamma 2)}$ pode apresentar uma relação não-monotônica em relação à $\sum_{1 \leq p < q \leq g}^{g} ||\overline{\boldsymbol{X}^*}^{(p)} - \overline{\boldsymbol{X}^*}^{(q)}||_2$.

Figura 13 – Somatório das distâncias entre os pares de grupos e *Gap* relativo, para os tempos (30s, 60s, 120s, 180s, 360s e 1440s).



Fonte: Elaborado pelo autor.

4.2 Estudo de caso - Epicovid19

O estudo de base populacional Epicovid19-BR, UNIVERSIDADE FEDERAL DE PELOTAS (2021), teve como objetivo *analisar a evolução de casos na população brasileira, por meio de uma amostragem de participantes em 133 "cidades sentinelas"* (Figura 14). A pesquisa baseou-se em inquéritos populacionais realizados por meio de visitas domiciliares. Em cada município, foram selecionados por meio de amostragem, 25 setores censitários.

Um setor censitário é uma unidade sócio-geográfica definida por uma área com população equivalente a cerca de 200 domicílios e é a menor porção territorial utilizada pelo IBGE para planejar e realizar levantamentos de dados do Censo e Pesquisas Estatísticas. Nesses setores foram selecionados aleatoriamente 10 domicílios, nos quais um indivíduo foi selecionado para a aplicação dos questionários e testes.

Este estudo de caso comparou a metodologia utilizada na seleção da amostra dos setores por meio dos métodos de *Aleatorização pura*, *Realeatorização* e *Alocação Fortuita*.

Figura 14 – Distribuição das 133 cidades consideradas.

Fonte: UNIVERSIDADE FEDERAL DE PELOTAS (2021).

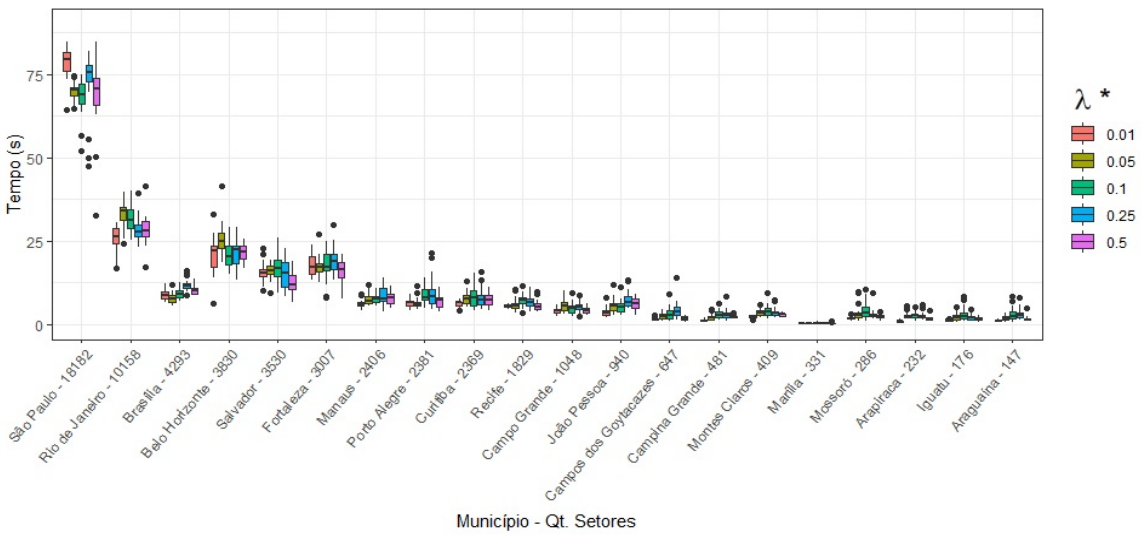
4.2.1 Experimentos iniciais

Os experimentos realizados baseiam-se em selecionar, em cada município, amostras de 25 setores censitários. Foram consideradas inicialmente as seguintes variáveis na implementação: Pessoas por Domicílios, Classe Social em 2019, percentual de Raça Branca e ln da Renda média em 2019.

A quantidade de setores censitários em cada município varia de 34 (Rorainópolis) a 18.182 (São Paulo). Essa grande variabilidade ensejou uma avaliação preliminar do tempo necessário para o problema atingir determinada qualidade de solução, no caso de cada município. Assim, para cada valor de λ, foram realizadas 30 repetições do processo de alocação, verificando o tempo necessário para encontrar uma solução incumbente com determinado Gap ($Gap_{abs} = Incumbente - Limitante\ inferior$). O Gap absoluto foi considerado por estabelecer uma distância máxima para a diferença entre as médias em todos os municípios.

Foram verificados os tempos necessários para atingir um valor de $Gap = 0.01$. Os resultados apresentados na figura 15 ilustram o comportamento do tempo utilizado em relação à quantidade de setores, no entanto, para a definição das faixas de tempo utilizadas nos experimentos, outros pré-experimentos empíricos foram realizados, de modo a definir uma calibração adequada do parâmetro λ e um tempo de CPU apropriado, que seja suficientemente grande para o *solver* encontrar soluções subótimas, mas também viável para a realização de um grande número de repetições.

Figura 15 – Tempo necessário para atingir $Gap = 0.01$.



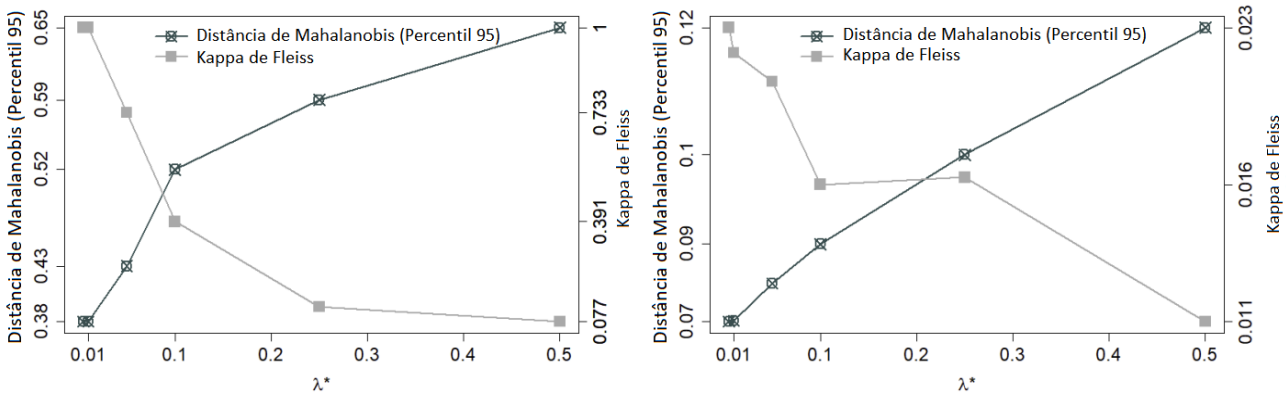
Fonte: Elaborado pelo autor.

4.2.2 Definição dos níveis de perturbação

Para que o *Método de Alocação Fortuita* não proporcione amostras previsíveis / determinísticas, torna-se necessária uma prévia avaliação do nível de perturbação adicionado ao procedimento de alocação, definida pelo parâmetro λ.

Uma calibração empírica foi realizada por meio da geração das curvas que comparam o nível de otimização (Distância de Mahalanobis) com o nível de desacoplamento (Kappa de Fleiss). A Figura 16 apresenta as curvas de Rorainópolis (34 setores) e Iguatu (176 setores). Os municípios foram caracterizados por 15 variáveis sócioeconomicas: Idade - Pessoas entre 70 e 75 (%), Idade - Pessoas com mais de 76 anos (%), Renda abaixo de meio salário mínimo (%), Renda entre 2 e 3 salários mínimos (%), Renda entre 3 e 5 salários mínimos (%), Renda entre 5 e 10 salários mínimos (%), Sem renda (%), População negra (%), Imóveis alugados (%), Condição da casa - outras condições (%), Casas com dois banheiros (%), Casas com três banheiros (%), Esgoto - fossa rudimentar (%), Esgoto - sistema de drenagem pluvial (%), Renda média - 2019.

Figura 16 – Distância de Mahalanobis e Kappa de Fleiss em relação à λ^* para o método de Alocação Fortuita. - Rorainópolis (34 setores) e Iguatu (176 setores)



Fonte: Autor. Adaptado de Miguel *et al.* (2021).

Para problemas com poucos setores, verificam-se alocações puramente determinísticas ($\kappa = 1$) quando λ^* é próximo de zero. No caso de problemas medianos, há a tendência de existir uma maior quantidade de soluções quase-ótimas, de modo que, mesmo para níveis baixos de perturbação o valor de κ é próximo de zero.

Essa análise baseou a definição das faixas de tempo e dos valores de λ^* utilizados nos experimentos para a avaliação do balencamento, apresentados na Tabela 2.

Tabela 2 – Parâmetros λ e tempos de processamento para cada faixa de setores censitários.

Setores	λ^*	**Tempo (s)**
< 50	0.1	5
50 - 4000	0.01	30
>4000	0.001	120

Fonte: Miguel *et al.* (2021).

4.2.3 Resultados - Balanceamento

Neste estudo de caso, o balanceamento das covariáveis tem como objetivo gerar amostras que são melhores estimadores do percentual de prevalência da SARS-CoV2 em cada município. Esses percentuais foram gerados por meio de um modelo de regressão auxiliar, apresentado em Miguel *et al.* (2022).

Dos 133 municípios da pesquisa, foram selecionados 10 para a realização dos experimentos de balanceamento. Essa seleção foi realizada de modo a abranger as diversas dimensões populacionais, traduzidas nas respectivas quantidades de setores. Considerando uma classificação ordenada dos 133 municípios, foram escolhidos os três maiores (1º São Paulo , 2º Rio de Janeiro e 3º Brasília), os três menores (131º Oiapoque, 132º Corrente e 133º Iguaçu) e quatro intermediários (30º Campos dos Goytacazes, 60º Marília, 90º Iguatu e 120º Cruzeiro do Sul). Os procedimentos de alocação foram repetidos 300 vezes para cada um dos 10 municípios, considerando as 15 variáveis sócio-economicas.

O balanceamento avaliado baseia-se nas diferenças entre as médias das covariáveis dos setores selecionados na amostra, $\overline{\boldsymbol{X}}^{(1)}$, e a médias das variáveis dos demais setores do município, $\overline{\boldsymbol{X}}^{(0)}$. A Figura 17 apresenta o resultado obtido para São Paulo. Os boxplots ilustram os valores de $(\overline{\boldsymbol{X}}_j^{(1)} - \overline{\boldsymbol{X}}_j^{(0)})/s_j$, onde s_j corresponde ao desvio padrão das diferenças de $(\overline{\boldsymbol{X}}_j^{(1)} - \overline{\boldsymbol{X}}_j^{(0)})$, obtidas para as 300 alocações utilizando método de aleatorização pura. Os resultados para os demais municípios estão apresentados no apêndice A.

Figura 17 – Soma das diferenças entre as médias dos dois grupos, para cada método - São Paulo (18.182 setores).

Fonte: Autor. Adaptado de Miguel *et al.* (2021).

4.3 Estudo de caso - Projeto S

O Projeto S foi uma iniciativa desenvolvida pelo Instituto Butantan que teve como objetivo a avaliação da eficácia da vacina Coronavac e envolveu a oferta da vacina para toda a população elegível do município de Serrana/São Paulo, que possui aproximadamente 45 mil habitantes (INSTITUTO BUTANTAN, 2020). O estudo de caso aqui apresentado, inspirado pelo Projeto S, considera uma situação hipotética na qual 4 opções de vacinas são avaliadas. Nesse caso ilustrativo, cada vacina é aplicada a um subconjunto da população residente de determinadas regiões do município, definidas por setores censitários.

O município de Serrana possui 45 setores censitários que, neste estudo, são caracterizados pelas mesmas 15 variáveis sócioeconomicas, consideradas para os procedimentos de alocação na seção 4.2. Para a calibração empírica dos parâmetros utilizados (tempo = 30s e $\lambda^* = 0.01$), o tempo de processamento e o valor de λ^* considerados foram os mesmos do estudo de caso anterior, apresentados na tabela 2. Foram geradas 300 alocações para cada método. A rotina dos experimentos consistiu em alocar os setores em 4 grupos (12, 11, 11 ,11) por cada método de alocação estudado e, para cada alocação, aplicar as taxas de eficácia relativas a cada vacina (Tabela 4), sobre o percentual de prevalência.

Tabela 3 – Taxas de Eficácia das Vacinas

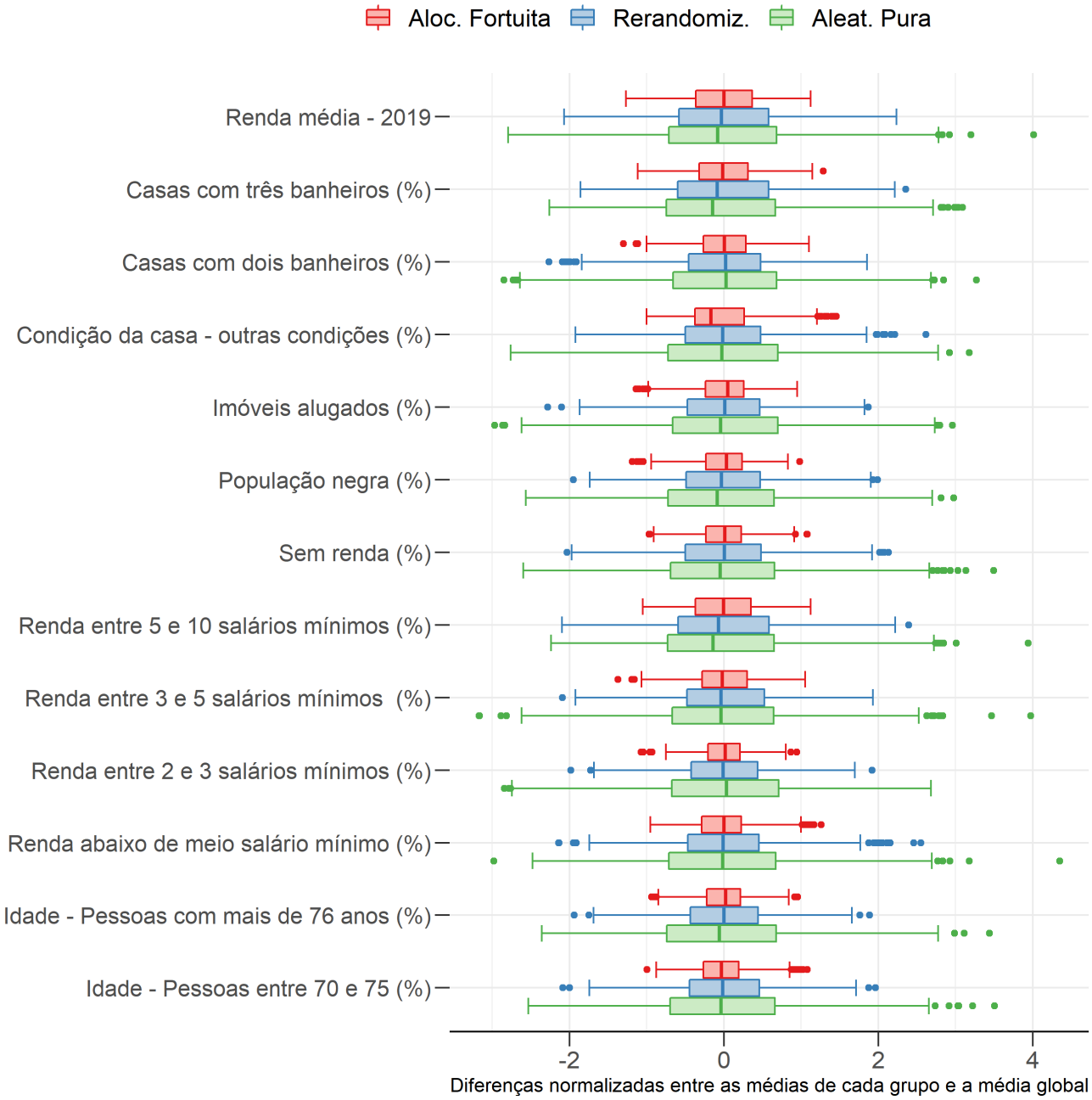
Vacina	**Eficácia (%)**
CORONAVAC/SINOVAC (controle)	50
AZTRAZENECA/OXFORD	70.4
MODERNA	94.5
PFIZER/BIONTEC	95

Fonte: Azevedo *et al.* (2021).

4.3.1 Resultados - Balanceamento e estimação das eficácias

A Figura 18 apresenta os valores de $(\overline{\boldsymbol{X}}_j^{(k)} - \overline{\boldsymbol{X}}_j)/s_j$, onde s_j corresponde ao desvio padrão das diferenças de $(\overline{\boldsymbol{X}}_j^{(k)} - \overline{\boldsymbol{X}}_j)$, obtidas para as $300 \times g$ alocações utilizando métoto de aleatorização pura. Assim como no estudo de caso anterior, o método de *Alocação Fortuita* apresentou as menores diferenças para essa medida de balanceamento.

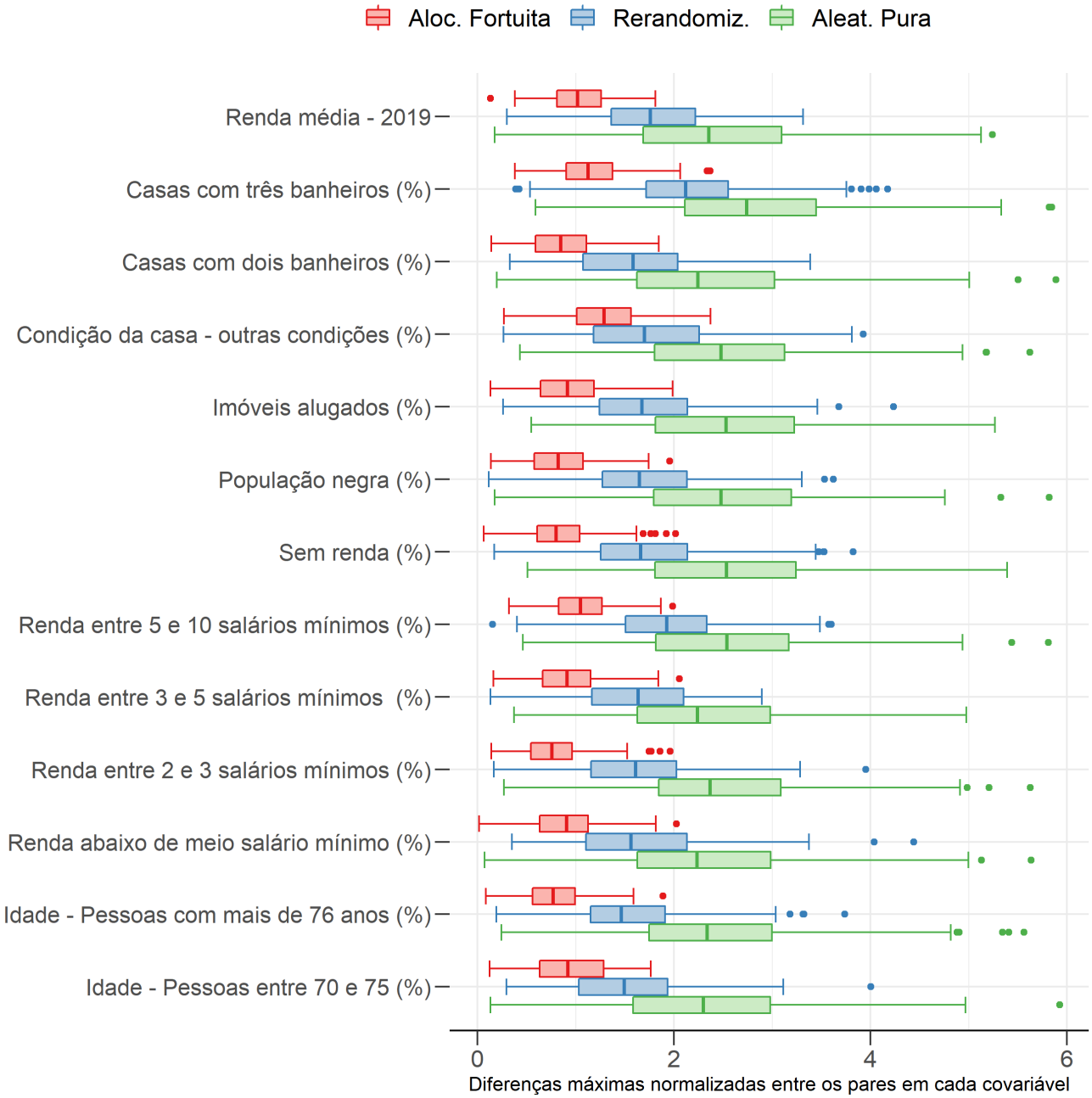
Figura 18 – Soma das diferenças entre as médias de cada grupo e a média global, para cada método - Serrana.



Fonte: Autor. Adaptado de Miguel *et al.* (2022).

No caso da avaliação da diferenças entre os pares, a Figura 19 apresenta as distribuições dos valores de $\max|\overline{\boldsymbol{X}}_j^{(p)} - \overline{\boldsymbol{X}}_j^{(q)}|/s_j$ obtidos nas 300 repetições, para cada método, onde s_j é o desvio padrão dos resultados de $\max|\overline{\boldsymbol{X}}_j^{(p)} - \overline{\boldsymbol{X}}_j^{(q)}|$ utilizando o método de aleatorização pura. Assim como no estudo de caso anterior, o método de *Alocação Fortuita* proporcionou os melhores valores de balanceamento, enquanto os métodos de *Rerandomização* e *Aleatorização Pura* apresentaram valores superiores a 4 e 5 desvios padrão, respectivamente.

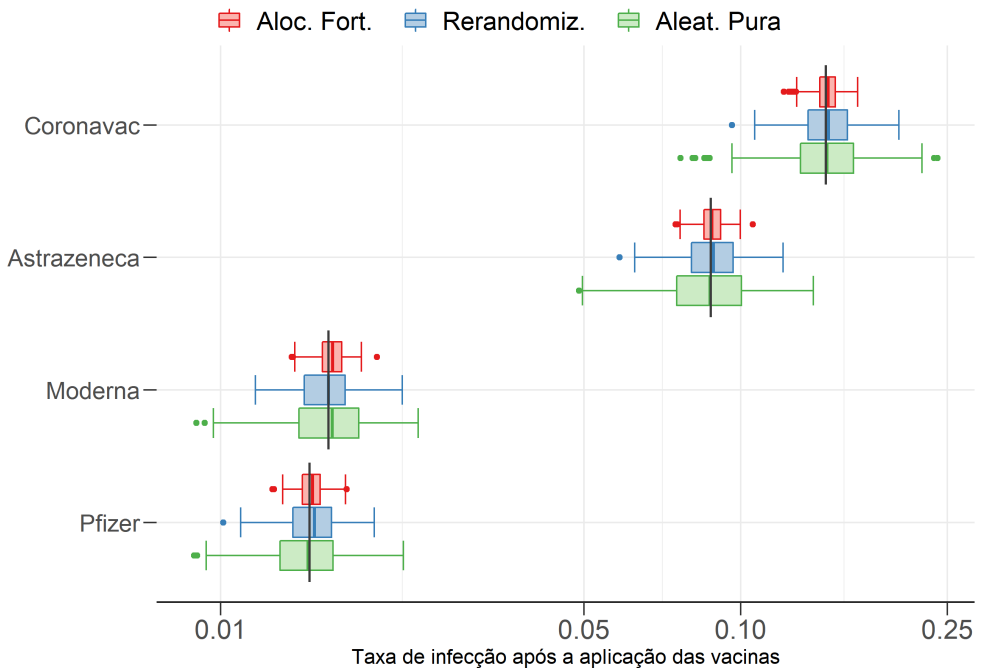
Figura 19 – Diferenças máximas entre as combinações de pares, para cada método - Serrana.



Fonte: Autor. Adaptado de Miguel *et al.* (2022).

A figura 20 apresenta as distribuições das taxas de infecção estimadas por cada método (boxplots) as taxas de infecção reais (barras verticais). A tabela 4 apresenta o 5º e 95º percentil desse resultado. Esse resultado mostra o desempenho desses métodos em proporcionar o balanceamento de uma variável não controlada no procedimento de otimização, no caso, o percentual de prevalência após a aplicação das vacinas. O método de *Alocação Fortuita* apresentou as menores variações em torno dos verdadeiros percentuais de prevalência, indicando que esse procedimento reduz a chance de erros substanciais na aferição das eficácias das vacinas.

Figura 20 – Taxas de Infecção estimadas e reais após a aplicação das vacinas.



Fonte: Autor. Adaptado de Miguel *et al.* (2022).

Tabela 4 – Taxas de Infecção estimadas (%) - Percentis

Vacina	Método	Perc. 05	Mediana	Perc. 95
CORONAVAC/SINOVAC (controle)	Aloc. Fortuita	13.11	14.74	16.09
	Rerandomização	11.62	14.76	17.90
	Aleat. Pura	10.22	14.68	19.88
AZTRAZENECA/OXFORD	Aloc. Fortuita	7.9	8.82	9.61
	Rerandomização	7.12	8.87	10.74
	Aleat. Pura	6.33	8.74	11.83
MODERNA	Aloc. Fortuita	1.46	1.65	1.78
	Rerandomização	1.31	1.61	1.94
	Aleat. Pura	1.15	1.64	2.17
PFIZER/BIONTEC	Aloc. Fortuita	1.36	1.50	1.64
	Rerandomização	1.18	1.51	1.78
	Aleat. Pura	1.06	1.47	1.92

Fonte: Autor. Baseado em Miguel *et al.* (2022).

5 Conclusão

Este trabalho apresentou as principais referências de um núcleo temático de publicações sobre o aprimoramento da eficiência de métodos de pesquisa, por meio de técnicas amostragem/alocação. Foi proposta uma modelagem baseada no método de *Alocação Fortuita* para o caso de múltiplos grupos, bem como a implementação da formulação proposta e de diversas rotinas para geração e avaliação das alocações para cada método estudado. Os resultados foram apresentados nos estudos de caso e possibilitaram as seguintes conclusões:

- A modelagem proposta para o uso do método de Alocação Fortuita mostrou-se a mais eficaz em proporcionar melhor balancemanto em todas as métricas avaliadas.

- Os resultados indicaram boas taxas de desacoplamento, mesmo para perturbações pequenas (valores de λ^* próximos de zero). No entando, cabe destacar que a taxa de desacoplamento pode variar substancialmente de acordo com características das bases de dados e o valor de λ^* deve ser calibrado no contexto de cada aplicação específica.

- O método de aleatorização pura resultou em desbalanceamentos substanciais, sobretudo nas comparações das diferenças máximas entre os pares de grupos, que chegaram a 6 desvios-padrão em dois estudos de casos.

- Esses valores ultrapassaram 4 desvios padrões para o método de *Rerandomização*. No entanto, cabe considerar que melhores resultados na utilização desse método podem ser obtidos por meio do aumento do tempo ou na capacidade de processamento.

Os resultados mostram o potencial do método de *Alocação Fortuita* selecionar subgrupos mais representativos e proporcionar estimativas com maior precisão, tornando pesquisas e experimentos mais eficientes. Em um contexto interdisciplinar, essa conclusão tem contribuições substanciais para a modelagem de sistemas complexos, uma vez que favorece a precisão e a atualização dos parâmetros utilizados no desenvolvimento modelos.

Os resultados deste trabalho geraram as seguintes publicações:

Miguel, M. G. R.; Waissman, R. P.; Lauretto, M. S.; Stern, J. M. Survey Optimization via the Haphazard Intentional Sampling Method. In: The 40th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. S.l.: MDPI, 2021. DOI:10.3390/psf2021003004

Miguel, M. G. R.; Waissman, R. P.; Lauretto, M. S.; Stern, J. M. Haphazard Intentional Sampling in Survey and Allocation Studies on COVID-19 Prevalence and Vaccine Efficacy. Entropy, MDPI AG, v. 24, n. 2, p. 225, jan. 2022. DOI:10.3390/e24020225

Como futuros objetos de pesquisa, podem-se mencionar as seguintes frentes:

a) A aplicação do método apresentado em vários campos como estudos clínicos, procedimentos de auditoria e aprendizado de máquina.

b) Novas modelagens contemplando formulações baseadas em outras funções de distância e extensões para outras estatísticas além da média, como a variância. Também podem ser consideradas medidas que comparam diretamente as distribuições. Como os métodos de solução dos *Solvers* comerciais estão em constante aprimoramento, o modelo baseado na norma quadrática, sem a adaptação para a Norma Híbrida, também pode ser considerado.

c) A investigação de propriedades assintóticas de convergência e um embasamento analítico, que podem fundamentar métodos de inferência aplicáveis especificamente ao contexto do método de *Alocação Fortuita*.

Os dados brutos utilizados neste trabalho estão disponíveis em:
⟨https://qualar.cetesb.sp.gov.br/qualar/home.do⟩ Acesso em: 5 jun. 2019.
⟨http://www.epicovid19brasil.org/?page_id=472⟩ Acesso em: 10 out. 2021.

Os códigos de implementação estão disponíveis em:
⟨https://github.com/marcelolauretto/Haphazard_MaxEnt2021⟩ Acesso em: 10 abril 2022.
⟨https://github.com/rafaelpw/Metodos_Aloc_Fortuita/⟩ Acesso em: 10 abril 2022.

Referências

AZEVEDO, T. C. P. de; FREITAS, P. V. de; CUNHA, P. H. P. da; MOREIRA, E. A. P.; ROCHA, T. J. M.; BARBOSA, F. T.; SOUSA-RODRIGUES, C. F. de; RAMOS, F. W. da S. Efficacy and landscape of covid-19 vaccines: a review article. *Revista da Associação Médica Brasileira*, FapUNIFESP (SciELO), v. 67, n. 3, p. 474–478, mar. 2021. Citado na página 62.

BERTSIMAS, D.; JOHNSON, M.; KALLUS, N. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, Institute for Operations Research and the Management Sciences (INFORMS), v. 63, n. 4, p. 868–876, ago. 2015. Citado 2 vezes nas páginas 34 e 35.

BOLFARINE, H.; BUSSAB, W. *Elementos de amostragem*. Sao Paulo: Edgard Blucher, 2007. ISBN 9788521203674. Citado 2 vezes nas páginas 23 e 24.

BONASSI, F. V.; NISHIMURA, R.; STERN, R. B.; GOGGANS, P. M.; CHAN, C.-Y. In defense of randomization: a subjectivist bayesian approach. In: . [S.l.]: AIP, 2009. Citado na página 17.

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO. *Sistema de Informações de Qualidade do Ar*. 2018. Acesso em: 5 jun. 2019. Disponível em: ⟨http://qualar.cetesb.sp.gov.br/qualar/home.do.html⟩. Citado na página 49.

DEGROOT, M.; SCHERVISH, M. *Probability and statistics*. Boston: Addison-Wesley, 2012. ISBN 978-0321500465. Citado 2 vezes nas páginas 21 e 23.

DINIZ, J.; FOSSALUZA, V.; PEREIRA, C. A. de B.; WECHSLER, S. Rain dance: the role of randomization in clinical trials. *Open Access Journal of Clinical Trials*, Informa UK Limited, Volume 8, p. 21–32, jul. 2016. Citado na página 17.

FISHER, R. A. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935. Citado na página 17.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, American Psychological Association (APA), v. 76, n. 5, p. 378–382, 1971. Citado na página 47.

GUROBI. *Gurobi Optimizer Reference Manual*. 2021. Disponível em: ⟨http://www.gurobi.com⟩. Citado na página 49.

HOPCROFT, J.; KANNAN, R. *Foundations of Data Science*. [S.l.: s.n.], 2014. Citado na página 39.

INSTITUTO BUTANTAN. *Projeto S*. 2020. ⟨https://projeto-s.butantan.gov.br/⟩. Acesso em 16 fev. 2022. Citado 2 vezes nas páginas 19 e 62.

INTERNATIONAL ORGANIZATION OF SUPREME AUDIT INSTITUTIONS. *GUID - 9020 - Evaluation of Public Policies*. Viena, 2019. Disponível em: ⟨issai.org/pronouncements/guid-9020-evaluation-of-public-policies/⟩. Citado na página 17.

KUTNER, M. H.; WASSERMAN, W.; NACHTSHEIM, C. J.; NETER, J. *Applied linear statistical models*. Boston: McGraw-Hill Irwin, 2005. ISBN 9780072386905. Citado na página 24.

LAURETTO, M. D. S.; NAKANO, F.; PEREIRA, C. A. de B.; STERN, J. M. Intentional sampling by goal optimization with decoupling by stochastic perturbation. In: . [S.l.]: AIP, 2012. Citado 2 vezes nas páginas 18 e 29.

LAURETTO, M. S.; STERN, R.; RIBEIRO, C.; STERN, J. Haphazard intentional sampling techniques in network design of monitoring stations. *Proceedings*, MDPI AG, v. 33, n. 1, p. 12, nov. 2019. Citado 10 vezes nas páginas 18, 19, 26, 29, 30, 31, 34, 49, 50 e 51.

LAURETTO, M. S.; STERN, R. B.; MORGAN, K. L.; CLARK, M.; STERN, J. M. Haphazard intentional allocation and rerandomization to improve covariate balance in experiments (maxent 2016). In: *Proceedings of the 36th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. [S.l.]: American Institute of Physics, 2017. Citado 6 vezes nas páginas 18, 26, 29, 31, 33 e 34.

LOCK, K. F. *Rerandomization to Improve Covariate Balance in Randomized Experiments*. Tese (Doutorado) — Harvard University, Cambrige, Massachusetts, 3 2011. Citado 2 vezes nas páginas 25 e 37.

MIGUEL, M. G. R.; WAISSMAN, R. P.; LAURETTO, M. S.; STERN, J. M. Survey Optimization via the Haphazard Intentional Sampling Method. In: *The 40th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. [S.l.]: MDPI, 2021. Citado 4 vezes nas páginas 18, 59, 60 e 61.

MIGUEL, M. G. R.; WAISSMAN, R. P.; LAURETTO, M. S.; STERN, J. M. Haphazard Intentional Sampling in Survey and Allocation Studies on COVID-19 Prevalence and Vaccine Efficacy. *Entropy*, MDPI AG, v. 24, n. 2, p. 225, jan. 2022. Citado 5 vezes nas páginas 18, 60, 63, 64 e 65.

MORGAN, K. L.; RUBIN, D. B. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 40, n. 2, p. 1263–1282, abr. 2012. Citado na página 32.

MORGAN, K. L.; RUBIN, D. B. Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, Informa UK Limited, v. 110, n. 512, p. 1412–1421, out. 2015. Citado 3 vezes nas páginas 26, 32 e 33.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: ⟨https://www.R-project.org/⟩. Citado na página 49.

ROMERO, C. *Handbook of Critical Issues in Goal Programming*. Oxford: Pergamon Press, 1991. ISBN 9780080406619. Citado 2 vezes nas páginas 29 e 37.

SOLOMON, J. *NUMERICAL ALGORITHMS: methods for computer vision,machine learning, and graphics*. S.l: CRC PRESS, 2020. ISBN 0367575639. Citado na página 27.

TRIBUNAL DE CONTAS DA UNIÃO. *Manual de Auditoria Operacional*. 4. ed. Brasilia, 2020. Disponível em: ⟨portal.tcu.gov.br/controle-externo/normas-e-orientacoes/normas-de-fiscalizacao/auditoria-operacional.htm⟩. Citado na página 17.

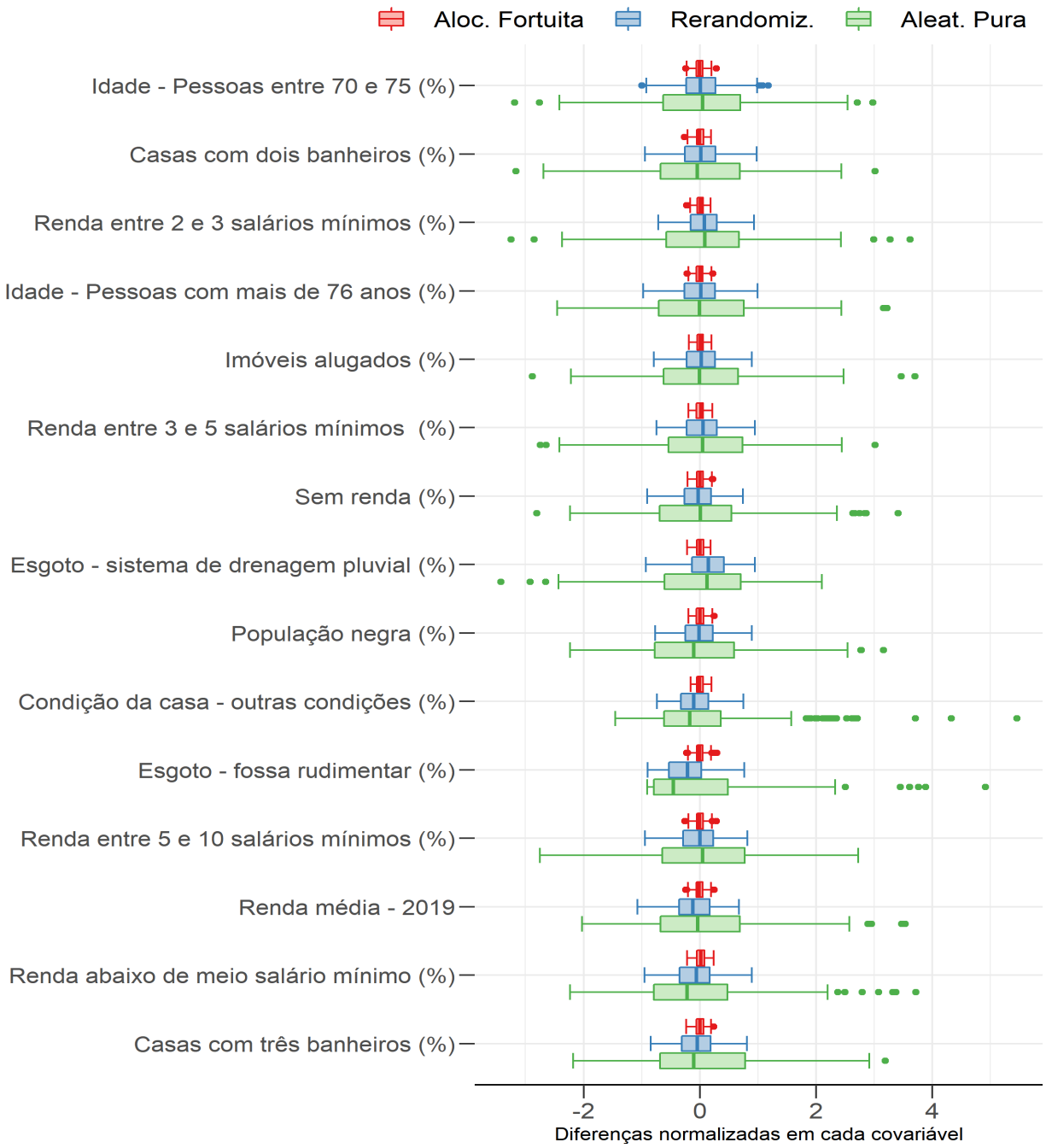
UNIVERSIDADE FEDERAL DE PELOTAS. *Evolução da Prevalência de Infecção por Covid-19 no Brasil: Estudo de base populacional*. 2021. Disponível em: ⟨www.epicovid19brasil.org/?page_id=472⟩. Citado 3 vezes nas páginas 16, 19 e 57.

WARD, J. E.; WENDELL, R. E. Technical note—a new norm for measuring distance which yields linear location problems. *Operations Research*, Institute for Operations Research and the Management Sciences, v. 28, n. 3-part-ii, p. 836–844, jun. 1980. Citado 2 vezes nas páginas 27 e 45.

YULE, G. U. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, JSTOR, v. 75, n. 6, p. 579, maio 1912. Citado 2 vezes nas páginas 30 e 47.

Apêndice A – Balanceamento - Estudo de Caso Epicovid 19

Figura 21 – Balanceamento das covariáveis - Rio de Janeiro (10.158 setores).

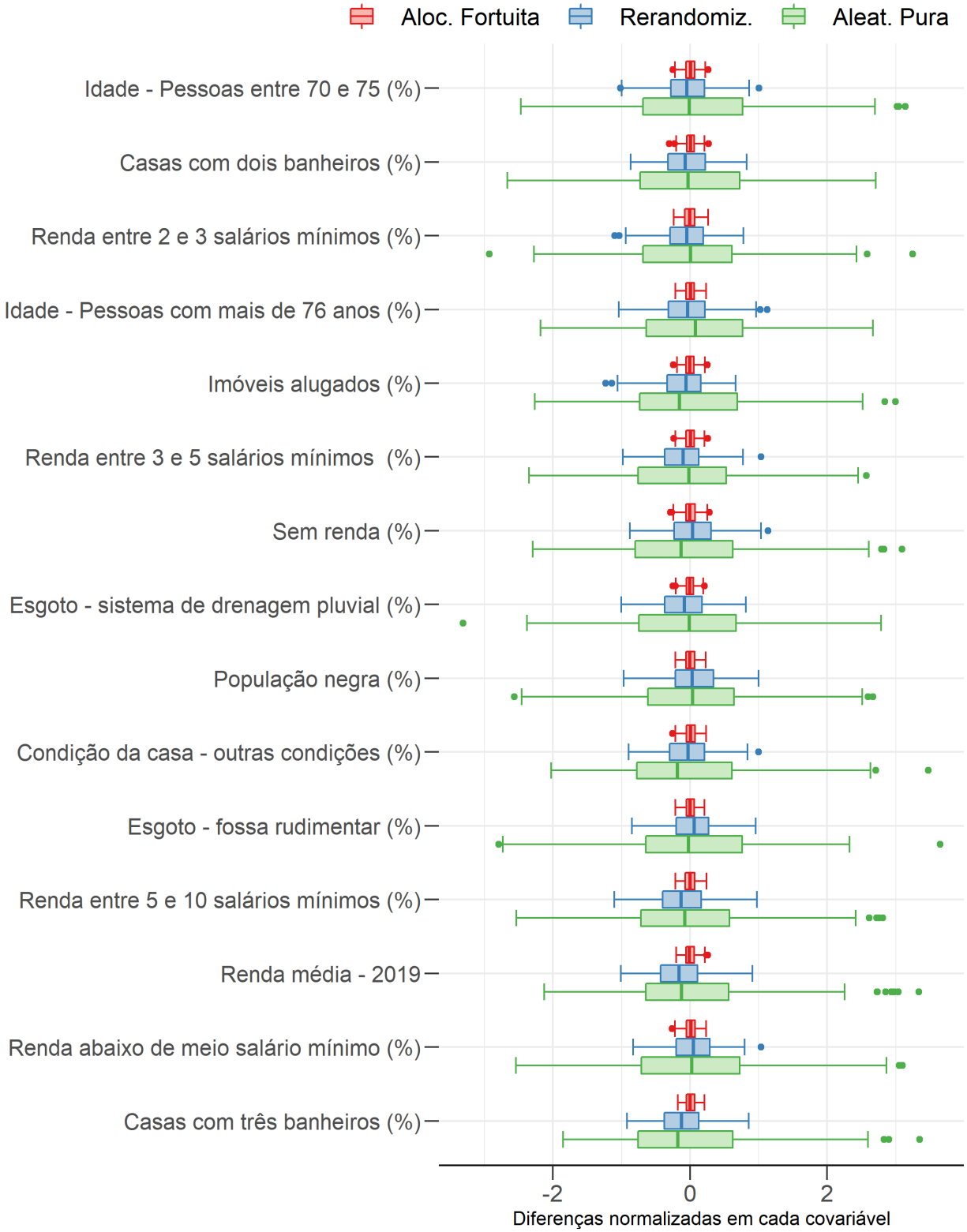


Fonte: Elaborado pelo autor.

Figura 22 – Balanceamento das covariáveis para cada método - Brasília (4.293 setores).

Fonte: Elaborado pelo autor.

Figura 23 – Balanceamento das covariáveis para cada método - Campos dos Goytacazes (647 setores).



Fonte: Elaborado pelo autor.

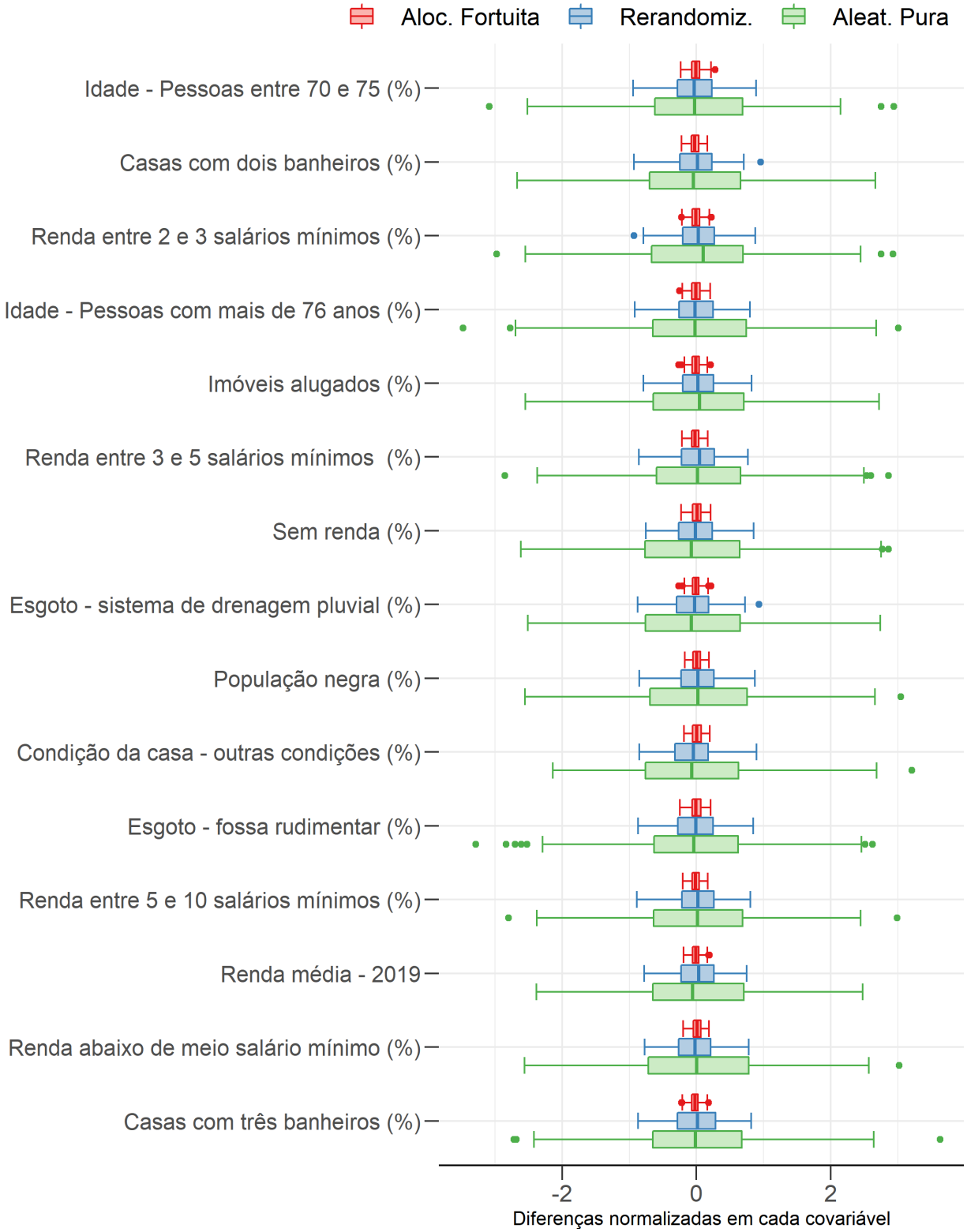
Figura 24 – Balanceamento das covariáveis para cada método - Marília (331 setores).

Fonte: Elaborado pelo autor.

Figura 25 – Balanceamento das covariáveis para cada método - Iguatu (176 setores).

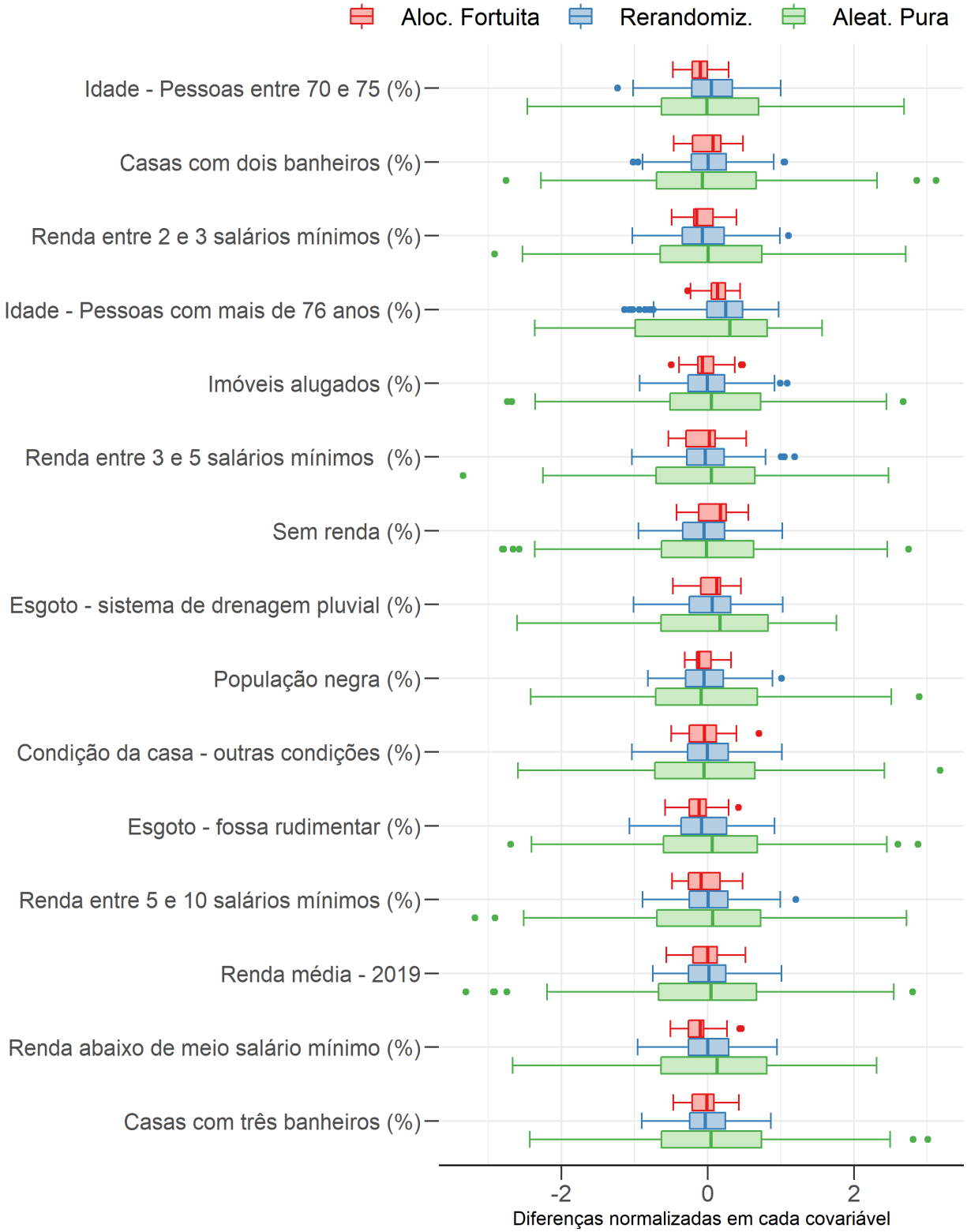
Fonte: Elaborado pelo autor.

Figura 26 – Balanceamento das covariáveis para cada método - Cruzeiro do Sul (86 setores).



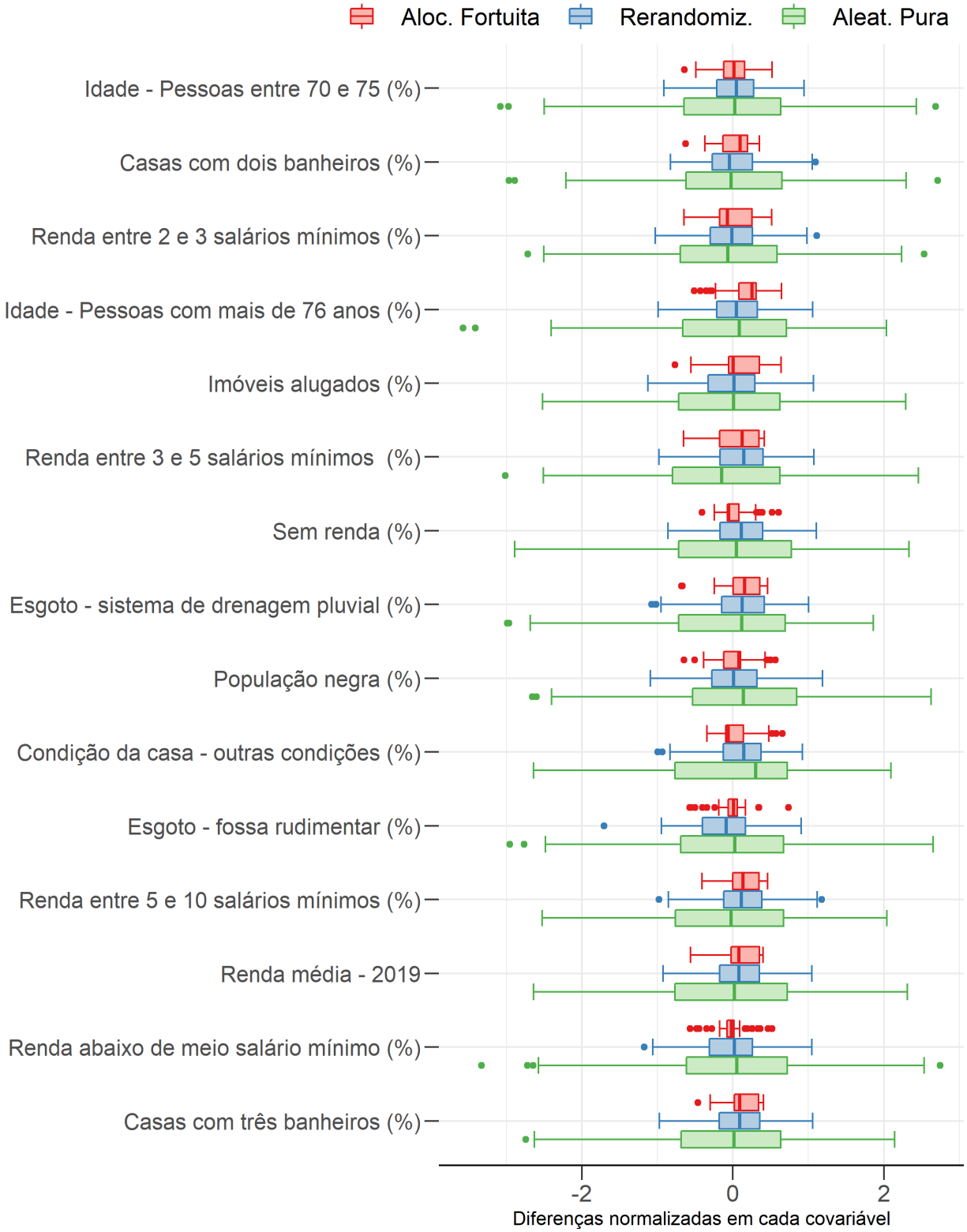
Fonte: Elaborado pelo autor.

Figura 27 – Balanceamento das covariáveis para cada método - Oiapoque (39 setores).



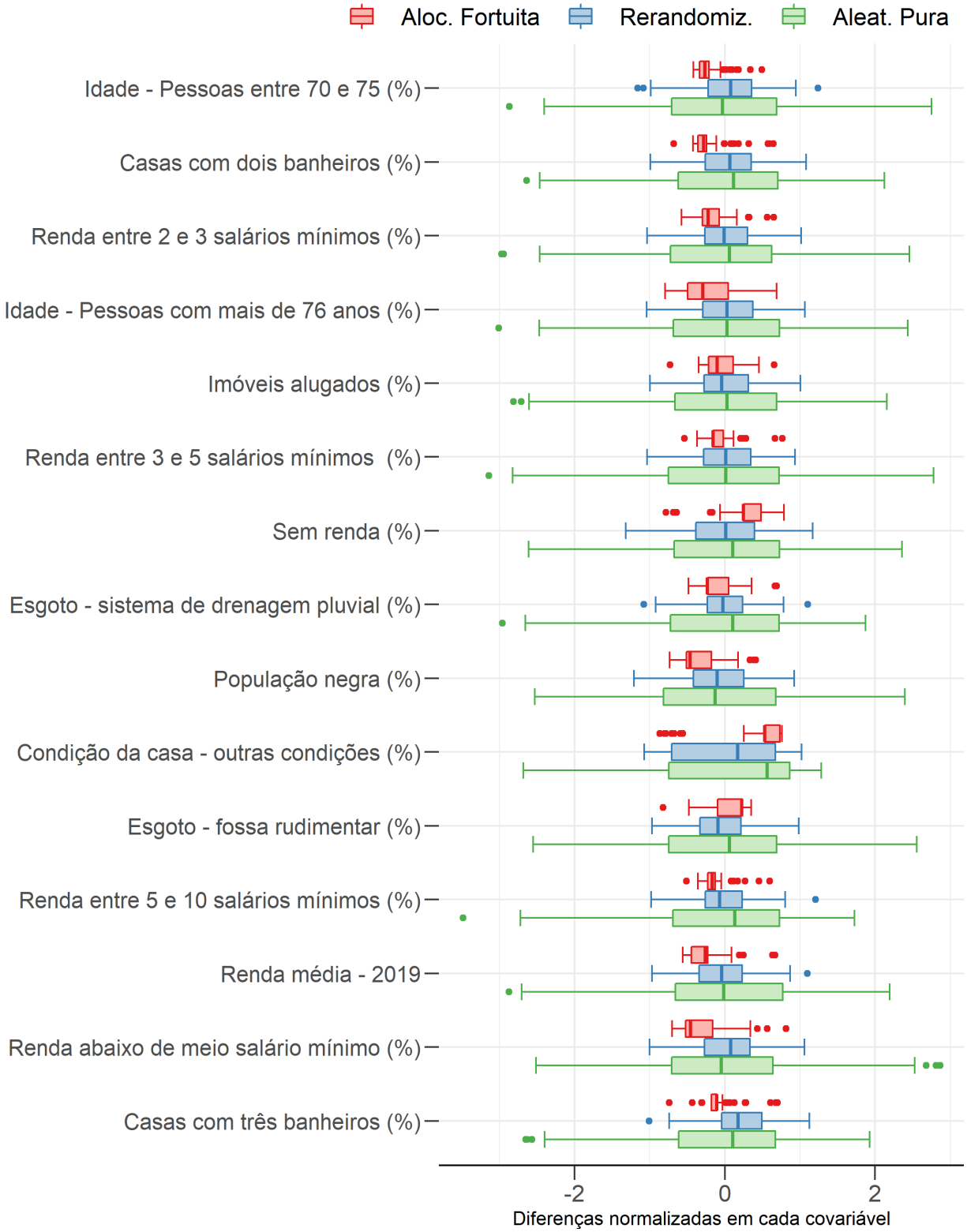
Fonte: Elaborado pelo autor.

Figura 28 – Balanceamento das covariáveis para cada método - Corrente (35 setores).



Fonte: Elaborado pelo autor.

Figura 29 – Balanceamento das covariáveis para cada método - Rorainópolis (34 setores).



Fonte: Elaborado pelo autor.