

Universidade de Brasília
Departamento de Estatística

Comparação entre Planos Amostrais Complexos Equiprovãveis e Não Equiprováveis

André Vale Araujo

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

André Vale Araujo

Comparação entre Planos Amostrais Complexos Equiprováveis e Não Equiprovãveis

Orientador: Prof. Dr. Alan Ricardo da Silva

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Agradecimentos

Agradeço primeiramente à minha família, aos meus pais e aos meus irmãos, pelo apoio constante e por estarem sempre ao meu lado. Sem o incentivo que me ofereceram ao longo desta jornada, este trabalho não teria sido possível.

Expresso também minha gratidão ao Professor Dr. Alan, cuja orientação, disponibilidade e, principalmente, paciência foram fundamentais para o desenvolvimento deste trabalho. Seus ensinamentos contribuíram de maneira significativa para o meu crescimento acadêmico e profissional.

Agradeço, ainda, aos meus amigos, tanto aqueles que fiz durante a graduação quanto os que já me acompanhavam antes, pelo companheirismo, pelas conversas, pelo apoio nos momentos difíceis e pela alegria compartilhada ao longo desta trajetória.

Resumo

Este trabalho tem como objetivo comparar planos amostrais complexos equiprováveis e não equiprováveis, analisando suas características, vantagens e limitações. Abordam-se aspectos fundamentais da amostragem complexa, com ênfase em problemas operacionais como superdimensionamento e subdimensionamento, que podem comprometer a seleção das unidades e a correta expansão dos pesos amostrais. Para tratar essas questões, são desenvolvidas macros em SAS capazes de identificar e corrigir automaticamente ambos os problemas, garantindo a viabilidade estatística dos planos. A metodologia é aplicada a quatro estados brasileiros selecionados de forma aleatória (Alagoas, Rio de Janeiro, Paraná e Acre), permitindo comparar custo, variância e distribuição dos erros entre os desenhos. Os resultados evidenciam que, embora o plano equiprovável simplifique a inferência e promova pesos constantes, ele é mais suscetível a desafios operacionais, enquanto o plano não equiprovável oferece maior flexibilidade e aderência às diferenças populacionais, ainda que demande ponderação adequada. O estudo contribui com uma análise prática e teórica dos delineamentos amostrais e fornece ferramentas replicáveis para o planejamento de levantamentos domiciliares em larga escala.

Palavras-chave: Amostragem Complexa, EPSEM, Probabilidade Proporcional ao Tamanho, Superdimensionamento, Subdimensionamentos, Pesos Amostrais.

Lista de Tabelas

2.1 Exemplo de 10 setores com os domicílios 17

2.2 Exemplo não equiprovável com cálculo do m_i e peso amostral 18

2.3 Exemplo com ajuste de pesos após detecção de superdimensionamento . . . 23

2.4 Exemplo equiprovável com subdimensionamento 23

2.5 Ajuste para correção do subdimensionamento (fusão dos setores 7 e 8) . . . 24

2.6 Exemplo equiprovável com subdimensionamento corrigido 24

2.7 Progressão dos estágios e problemas de dimensionamento até o 2º estágio . 28

2.8 Progressão dos estágios e problemas de dimensionamento até o 3º estágio . 28

2.9 Progressão dos estágios e problemas de dimensionamento até o 4º estágio . 28

3.1 Variáveis utilizadas do banco de setores censitários do IBGE 30

4.1 Interações e agrupamentos por ordem de vizinhos de Alagoas 38

4.2 Comparação entre os planos amostrais de Alagoas 40

4.3 Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (AL) 41

4.4 Interações e agrupamentos por ordem de vizinhos de Rio de Janeiro 43

4.5 Comparação entre os planos amostrais do Rio de Janeiro 45

4.6 Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (RJ) 45

4.7 Interações e agrupamentos por ordem de vizinhos do Paraná 47

4.8 Comparação entre os planos amostrais do Paraná 49

4.9 Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (PR) 49

4.10 Interações e agrupamentos por ordem de vizinhos do Acre 52

4.11 Comparação entre os planos amostrais do Acre 54

4.12 Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (AC) 54

4.13 Soma dos pesos amostrais por Unidade da Federação e plano amostral . . . 55

Lista de Figuras

4.1 Mapa de municípios do estado de Alagoas após resolução de subdimensionamento 37

4.2 Mapa de municípios do estado de Alagoas selecionados para a amostra EPSEM 37

4.3 Mapa de setores censitários do estado de Alagoas após resolução de subdimensionamento 38

4.4 Mapa de municípios do estado do Alagoas selecionados para a amostra Não EPSEM 39

4.5 Mapa de municípios do estado do Rio de Janeiro após resolução de subdimensionamento 42

4.6 Mapa de municípios do estado do Rio de Janeiro selecionados para a amostra EPSEM 43

4.7 Mapa de setores censitários do estado do Rio de Janeiro após resolução de subdimensionamento 43

4.8 Mapa de municípios do estado do Rio de janeiro selecionados para a amostra Não EPSEM 44

4.9 Mapa de municípios do estado do Paraná após resolução de subdimensionamento 46

4.10 Mapa de municípios do estado do Paraná selecionados para a amostra EPSEM 47

4.11 Mapa de setores censitários estado do Paraná após resolução de subdimensionamento 47

4.12 Mapa de municípios do estado do Paraná selecionados para a amostra Não EPSEM 48

4.13 Mapa de municípios do estado do Acre após resolução de subdimensionamento 51

4.14 Mapa de municípios do estado do Acre selecionados para a amostra EPSEM 51

4.15 Mapa de setores censitários estado do Acre após resolução de subdimensionamento 52

4.16 Mapa de municípios do estado do Acre selecionados para a amostra Não EPSEM 53

Sumário

1 Introdução **8**

1.1 Objetivos 9

2 Amostragem Complexa **10**

2.1 Introdução 10

2.2 Amostragem por Conglomerados em Dois Estãgios com Tamanhos Iguais 10

2.3 Amostragem por Conglomerados em dois estágios com tamanhos desiguais 13

2.3.1 Seleção com Probabilidade Proporcional ao Tamanho 13

2.3.2 Custo e tamanho de amostra 14

2.3.3 Proporções 19

2.4 Amostra Equiprovãvel 19

2.5 Superdimensionamento 20

2.6 Subdimensionamento 22

2.7 Amostragem por Conglomerados em Três Estãgios 24

3 Materiais e Métodos **29**

3.1 Introdução 29

3.2 Materiais 29

3.3 Métodos 30

3.3.1 Algoritmo de solução para Superdimensionamento 30

3.3.2 Algoritmo de solução para Subdimensionamento 31

3.3.3 Planos Amostrais 33

4 Análise dos resultados **35**

4.1 Introdução 35

4.2 Estado de Alagoas 36

4.2.1 Amostra Equipróvavel 36

4.2.2 Amostra Não Equipróvavel 39

4.3 Estado do Rio de Janeiro 41

4.3.1 Amostra Equipróvavel 42

4.3.2 Amostra Não Equipróvavel 44

4.4 Estado do Paraná 45

4.4.1 Amostra Equipróvavel 46

4.4.2 Amostra Não Equipróvavel 48

4.5 Estado do Acre 50

4.5.1 Amostra Equiprovável 50

4.5.2 Amostra Não Equiprovável 53

4.6 Resumo Geral 55

5 Conclusão **57**

Referências 58

Apêndice 60

A Macro SAS para subdimensionamento **60**

B Macro SAS para superdimensionamento **69**

Capítulo 1

Introdução

A amostragem é uma ferramenta importante da estatística, permitindo inferências sobre a população a partir de um subconjunto representativo, com foco na obtenção de estimativas precisas ao menor custo possível (Cochran, 1977). Em pesquisas de grande escala, como levantamentos socioeconômicos, pesquisas eleitorais e estudos de saúde pública, a utilização de planos amostrais complexos torna-se mais apropriada para melhor representar a heterogeneidade da população (Lohr, 2010). Embora a Amostragem Aleatória Simples (AAS) e outras técnicas, como a amostragem estratificada e a por conglomerados, sejam igualmente importantes, elas individualmente não seriam as mais indicadas para lidar com populações heterogêneas.

Por outro lado, os planos amostrais complexos, que necessariamente envolvem estratificação e conglomeração, além da seleção de unidades com probabilidades desiguais (Skinner e Wakefield, 2017), também podem utilizar estimadores do tipo razão e regressão para ajustar por outras variáveis. No contexto de grandes pesquisas nacionais, como a Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua, planos amostrais complexos têm sido amplamente utilizados, pois possibilitam a coleta de dados representativos de diversas regiões e grupos sociais de maneira eficiente (IBGE, 2014).

Em planos amostrais complexos de múltiplos estágios, a desigualdade no tamanho das unidades exige o uso de seleção de unidade considerando uma Probabilidade Proporcional ao Tamanho (PPT) (ou do inglês *Probability Proportional to Size* (PPS)). No entanto, essa abordagem pode gerar desafios na seleção das unidades, como a ocorrência de unidades amostradas superdimensionadas e subdimensionadas (ou, do inglês, *oversized* e *undersized*, respectivamente) (Kish, 1965). Unidades superdimensionadas são aquelas que têm probabilidade maior do que 1 de estarem na amostra, enquanto as unidades subdimensionadas são aquelas que possuem menos elementos do que a quantidade de unidades

a serem selecionadas (Kish, 1965).

Para exemplificar, considere um plano amostral com dois estágios, sendo os setores censitários as Unidades Primárias de Amostragem (UPA), e os domicílios as Unidades Secundárias de Amostragem (USA). O problema de unidades superdimensionadas pode ocorrer no primeiro estágio, quando existirem setores censitários com uma quantidade muito grande de domicílios. Uma forma de resolver isso seria alocar aqueles setores censitários com probabilidades maiores do que 1 em um estrato, e os restantes em outro estrato.

Já o problema de unidades subdimensionadas pode ocorrer no segundo estágio, quando a quantidade de domicílios a serem selecionados para a amostra for maior do que a quantidade de domicílios existentes naquele setor censitário. Uma forma de resolver isso é agrupar os domicílios de setores censitários contíguos até que a quantidade de domicílios seja maior ou igual à quantidade de domicílios a serem selecionados para a amostra. Note que esses ajustes (ou preparação da base de dados) devem ser feitos antes de qualquer seleção. Em planos amostrais com múltiplos estágios, as unidades intermediárias (entre o primeiro e o último estágio) podem apresentar ambos os problemas simultaneamente.

Quando se trabalha com unidades de tamanho desigual, uma forma de evitar pesos também desiguais é utilizar um método de seleção equiprovável (ou do inglês *Equal Probability of Sampling Method* (EPSEM)), conforme enunciado por Kish (1965). A Amostragem Complexa Equiprovável garante que todas as unidades da população tenham a mesma probabilidade de serem selecionadas, o que simplifica a inferência estatística, reduz a variância das estimativas e aumenta a robustez dos resultados (Kish, 1965). No entanto, os problemas de superdimensionamento e subdimensionamento podem ocorrer com mais frequência.

1.1 Objetivos

Esse trabalho tem como objetivo geral comparar os planos amostrais complexos equiprováveis e não equiprováveis, analisando suas características, vantagens e limitações em diferentes contextos.

Os objetivos específicos são:

- Comparar o custo e a variância entre os dois métodos;
- Desenvolver uma macro em SAS para ajustar subdimensionamento e superdimensionamento;

Capítulo 2

Amostragem Complexa

2.1 Introdução

A amostragem complexa é uma abordagem metodológica que combina técnicas de seleção, como estratificação em diversos estágios de conglomeração, frequentemente com probabilidades distintas de seleção. Essa estratégia é adequada quando a população de interesse apresenta heterogeneidade ou está organizada em subgrupos definidos, como regiões geográficas, unidades administrativas ou agrupamentos institucionais (Kish, 1965; Lohr, 2010).

Neste trabalho, não serão abordados os métodos de Amostragem Aleatória Simples (AAS), Amostragem Estratificada (AE) ou Amostragem por Conglomerados em Um Estágio (AC1). O foco será direcionado imediatamente para a Amostragem por Conglomerados em Dois Estágios (AC2) e Três Estágios (AC3).

A seguir, serão detalhadas as teorias necessárias para um desenho amostral complexo, com enfoque nos métodos equiprováveis e não equiprováveis, em que cada elemento da população pode ser selecionado com probabilidades iguais ou distintas. Serão apresentados ainda os procedimentos de verificação e ajuste para unidades superdimensionadas e subdimensionadas, aspectos característicos desse tipo de delineamento amostral.

2.2 Amostragem por Conglomerados em Dois Estágios com Tamanhos Iguais

A amostragem em dois estágios, também denominada subamostragem, consiste na seleção inicial de UPAs, seguida da seleção de USAs dentro das UPAs selecionadas

(Cochran, 1977). Essa técnica apresenta vantagens em contextos nos quais o custo de obtenção de uma lista completa da população é elevado, mas é viável realizar listagens dentro de conglomerados previamente selecionados (Kish, 1965).

Neste tipo de plano, todas as UPAs possuem o mesmo número de USAs, sendo a seleção realizada por AAS tanto no primeiro quanto no segundo estágio (Cochran, 1977). Denotando por a o número de UPAs selecionadas, de um total A disponíveis na população, e por m o número de USAs selecionadas, de um total M existentes dentro de cada conglomerado, o tamanho total da amostra é dado pela expressão:

$$n = am \tag{2.1}$$

Para conglomerados, a média amostral por subelemento $\bar{\bar{y}}$ é dada por (Cochran, 1977):

$$\bar{\bar{y}} = \sum_{i=1}^{a} \sum_{j=1}^{m} \frac{y_{ij}}{am} \tag{2.2}$$

onde y_{ij} é a unidade amostrada da i-ésima UPA da j-ésima USA.

A variância, em comparação com a Amostragem Aleatória Simples, apresenta uma forma mais complexa, pois envolve o cálculo das variâncias nos dois estágios de seleção (Cochran, 1977), sendo dada pela equação:

$$\text{Var}(\bar{\bar{y}}) = (1 - f_1) \frac{S_1^2}{a} + (1 - f_2) \frac{S_2^2}{am} \tag{2.3}$$

onde:
S_1^2 é a variância entre as médias das UPAs, dada por $\frac{1}{A-1} \sum_{i=1}^{A} (\bar{Y}_i - \bar{\bar{Y}})^2$;
S_2^2 é a Variância dentro das UPAs, dada por $\frac{1}{A(M-1)} \sum_{i=1}^{A} \sum_{j=1}^{M} (Y_{ij} - \bar{Y}_i)^2$;
f_1 é a fração de amostragem no primeiro estágio, $f_1 = a/A$;
f_2 é a fração de amostragem no segundo estágio, $f_2 = m/M$.

É possível perceber que quando $m = M$, o segundo termo da variância desaparece $(1 - \frac{m}{M} = 0)$, e a equação se reduz a $\text{Var}(\bar{\bar{y}}) = \left(1 - \frac{a}{A}\right) \frac{S_1^2}{a}$, que é exatamente a variância no caso de apenas um estágio.

O custo total do levantamento apresenta uma estrutura linear que inclui componentes fixos e variáveis, sendo dado pela equação:

$$C = c_0 + c_1 a + c_2 am \tag{2.4}$$

em que C é o custo total da amostragem; c_0 representa o custo fixo, que independe do

tamanho da amostra; c_1 é o custo por UPA; e c_2 é o custo por USA.

Para a otimização do desenho amostral, o primeiro passo é determinar a quantidade ótima m de USAs por UPA. Quando as variâncias populacionais e os custos são conhecidos, o valor ótimo de m é dado por (Cochran, 1977):

$$m_{\text{ótimo}} = \frac{S_2}{\sqrt{S_1^2 - \frac{S_2^2}{M}}} \sqrt{\frac{c_1}{c_2}} \tag{2.5}$$

É possível perceber que, quando consideramos o caso de apenas um estágio, o valor ótimo naturalmente seria $m = M$, pois todas as unidades dentro do conglomerado devem ser incluídas. É importante notar que esta equação não se aplica diretamente ao caso de um estágio, pois, neste caso, a decisão de incluir todas as USAs ($m = M$) já está pré-estabelecida.

Na prática, como as variâncias populacionais raramente são conhecidas, utiliza-se uma versão simplificada que depende do coeficiente de correlação intraclasse (ρ), que mede a homogeneidade dentro dos conglomerados (Kish, 1965), sendo dada pela equação:

$$m_{\text{ótimo}} = \sqrt{\frac{1-\rho}{\rho} \frac{c_1}{c_2}} \tag{2.6}$$

O coeficiente ρ pode ser estimado de duas formas. A primeira baseia-se nos produtos cruzados dos desvios dada pela equação:

$$\rho = \frac{2 \sum_{i=1}^{A} \sum_{i<j}^{M} \sum_{j<l}^{M} (Y_{ij} - \bar{\bar{Y}})(Y_{il} - \bar{\bar{Y}})}{(M-1)(AM-1)S^2} \tag{2.7}$$

onde $S^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})}{N-1}$ é a variância da população.

A segunda forma, mais operacional, utiliza o Efeito do Planejamento Amostral (EPA) (ou do inglês *Design Effect (Deff)*), conceito central no planejamento amostral (Kish, 1965), em que o EPA expressa o aumento da variância decorrente do uso de conglomerados em vez da amostragem aleatória simples, sendo dada por:

$$\rho = \frac{\text{EPA} - 1}{M - 1} \tag{2.8}$$

A determinação do valor de ρ depende da escolha de um valor de EPA, a qual deve ser fundamentada em estudos prévios ou em julgamentos técnicos especializados, devido à particularidade de cada população (Kish, 1965).

Para um orçamento fixo C, o número de UPAs a é dado pela equação:

$$a = \frac{C - c_0}{c_1 + c_2 m_{\text{ótimo}}} \tag{2.9}$$

Quando o objetivo é atingir uma precisão específica, expressa por um erro relativo máximo r, então a é determinado pela equação (Kish, 1965):

$$a = \frac{z^2_{\gamma/2} CV^2 \left(\rho + \frac{1-\rho}{m_{\text{ótimo}}}\right)}{r^2} \tag{2.10}$$

onde $z_{\gamma/2}$ é o quantil da distribuição normal para um nível de confiança γ, CV é o coeficiente de variação, dado por $S/\bar{Y}$, e r é o erro relativo máximo aceitável.

Na ausência de informações precisas sobre ρ e CV, recomenda-se a realização de um estudo piloto, conforme destacado por Lohr (2010). Na prática, conglomerados de tamanhos iguais são raros, particularmente em pesquisas domiciliares. Diante dessa realidade, a próxima seção abordará os métodos para tratamento de conglomerados com tamanhos desiguais.

2.3 Amostragem por Conglomerados em dois estágios com tamanhos desiguais

Quando as unidades primárias de amostragem possuem tamanhos muito diferentes, como setores censitários com números variados de domicílios, utiliza-se a abordagem com tamanhos desiguais. Este trabalho abordará a seleção das UPAs por Probabilidade Proporcional ao Tamanho (PPT), método mais adequado para essa situação.

2.3.1 Seleção com Probabilidade Proporcional ao Tamanho

A seleção PPT é especialmente adequada para situações em que os conglomerados possuem tamanhos desiguais. Este método refere-se à seleção com reposição, onde a probabilidade de seleção do i-ésimo conglomerado é dada por $\pi_i = M_i/M_0$, sendo M_i o tamanho do conglomerado e $M_0 = \sum_{i=1}^{A} M_i$ o total de unidades na população. Na prática, como a obtenção precisa de M_i é frequentemente difícil, utiliza-se uma medida de sua grandeza M_i^* altamente correlacionada com seu tamanho, resultando em $z_i = M_i^*/M_0^*$. O estimador PPT para o total populacional é dado pela equação:

$$\hat{Y}_{PPT} = \frac{1}{a}\sum_{i=1}^{a}\frac{y_i}{z_i} \tag{2.11}$$

Para a média temos:

$$\bar{Y}_{PPT} = \frac{1}{aM_0^*}\sum_{i=1}^{a}\frac{y_i}{z_i} \tag{2.12}$$

em que a representa o número de conglomerados selecionados; z_i denota a probabilidade de seleção do i-ésimo conglomerado; e y_i corresponde ao valor da variável de interesse no i-ésimo conglomerado, também dado por $y_i = M_i\bar{y}_i$.

De acordo com Cochran (1977), este estimador é não viesado quando o número de unidades secundárias m_i é fixo em cada conglomerado. Sua eficiência é maximizada quando os totais dos conglomerados Y_i são proporcionais aos seus respectivos tamanhos M_i. Além disso, conforme destacado por Kish (1965), a utilização do método PPT facilita o controle das variações de tamanho entre os conglomerados, proporcionando uma seleção mais equilibrada e representativa.

A variância estimada do estimador PPT é dada pela equação:

$$\widehat{\mathrm{Var}}(\hat{Y}_{PPT}) = \sum_{i=1}^{a}\frac{\left(\frac{y_i}{z_i} - \hat{Y}_{PPT}\right)^2}{a(a-1)} \tag{2.13}$$

em que $\widehat{\mathrm{Var}}(\hat{Y}_{PPT})$ é a variância estimada do estimador PPT do total; $\hat{Y}_i$ é o total estimado no i-ésimo conglomerado que também pode ser obtido como $\hat{Y}_i = M_i\bar{y}_i = y_i$; e $\hat{Y}_{PPT}$ é o estimador PPT do total populacional, conforme a Equação (2.11).

A precisão deste estimador depende da homogeneidade das médias $\bar{y}_i$ entre os conglomerados. Quanto mais próximas forem essas médias do valor global $\hat{Y}_{PPS}/M_0$, menor será a variância da estimativa. Ainda segundo Kish (1965), a definição adequada do tamanho amostral em cada conglomerado é fundamental para otimizar a eficiência do levantamento.

2.3.2 Custo e tamanho de amostra

Em amostragem por conglomerados de tamanho desigual, o planejamento amostral requer considerações especiais sobre custos e tamanhos ótimos. Seguindo Cochran (1977) e Kish (1965), adota-se uma abordagem que considera um custo médio de amos-

tragem das Unidades Primárias de Amostragem (UPAs).

A função de custo de amostragem em dois estágios de tamanhos desiguais é dada pela equação:

$$C = c_0 + c_n a + c_2 \sum_{i=1}^{a} m_i + c_l \sum_{i=1}^{a} m_i \tag{2.14}$$

onde c_0 é o custo fixo, c_n é o custo fixo por UPA, c_2 é o custo por USA e c_l o custo de listar as subunidades das unidades selecionadas. O termo c_l é incluído porque o amostrista precisa listar os elementos das unidades selecionadas antes de fazer o sorteio (Cochran, 1977).

Essa equação não é usada dessa forma, dado que o custo depende de uma série particular de unidades que são selecionadas. Considere o custo médio das UPAs:

$$E[C] = c_0 + c_n a + c_2 a\bar{m} + c_l a\bar{m}$$

$$E[C] = c_0 + (c_n + c_l\bar{m})a + c_2 a\bar{m}$$

$$C = c_0 + c_1 a + c_2 a\bar{m} \tag{2.15}$$

onde: C é o custo total da amostragem; c_0 é o custo fixo da pesquisa; c_1 é o custo fixo por por UPA mais o custo médio de listagem das unidades; c_2 é o custo por unidade secundária (dentro das UPAs); a é o número de UPAs selecionadas; $\bar{m}$ é o tamanho médio dos conglomerados na amostra, $\bar{m} = \frac{\sum_{i=1}^{a} m_i}{a}$.

Para conglomerados de tamanhos desiguais, a variância da média amostral por subelemento $\bar{\bar{y}}$ é dada pela equação (Cochran, 1977):

$$Var(\bar{\bar{y}}) = (1 - f_1)\frac{S_b^2}{a} + (1 - f_2)\frac{S_2^2}{a\bar{m}} \tag{2.16}$$

onde:
f_1 é a fração de amostragem no primeiro estágio, $f_1 = \frac{a}{A}$;
f_2 é a fração de amostragem no segundo estágio, $f_2 = \frac{\bar{m}}{\bar{M}}$;
S_b^2 é a variância entre conglomerados;
S_2^2 é a variância dentro dos conglomerados;

A fração de amostragem no segundo estágio pode ser expressa de forma constante como (Cochran, 1977):

$$f_{2i} = \frac{m_i}{M_i} = f_2 = \frac{\bar{m}}{\bar{M}} \tag{2.17}$$

onde $\bar{M}$ é o tamanho médio populacional dos conglomerados calculado como $\bar{M} = \frac{\sum_{i=1}^{A} M_i}{A}$.

As componentes de variância são calculadas como:

$$S_b^2 = \frac{M_i^2(\bar{Y}_i - \bar{\bar{Y}})^2}{\bar{M}(A-1)} \tag{2.18}$$

$$S_2^2 = \sum_{i=1}^{A} \frac{M_i}{M_0} S_{2i}^2 \tag{2.19}$$

O tamanho ótimo médio dos conglomerados na amostra é calculado utilizado a desigualdade de Cauchy-Schwarz para chegar à seguinte equação (Cochran, 1977):

$$\bar{m}_{\text{ótimo}} = \frac{S_2}{\sqrt{S_b^2 - \frac{s_2^2}{\bar{M}}}} \sqrt{\frac{c_1}{c_2}} \tag{2.20}$$

A Equação (2.20) mantém a mesma forma básica da Equação (2.5) para conglomerados iguais, porém adaptada para considerar o tamanho médio. Como alternativa equivalente, pode-se calcular $\bar{m}_{\text{ótimo}}$ utilizando diretamente o coeficiente ρ conforme mostrado na Equação (2.6), seguindo a mesma abordagem já estabelecida.

Assim como no caso de conglomerados de tamanhos iguais, o número de UPAs (a) pode ser determinado por duas abordagens equivalentes: utilizando a Equação de custo (2.9) após o cálculo de $\bar{m}_{\text{ótimo}}$, ou aplicando diretamente a Equação (2.10) quando se busca atender a um erro relativo máximo especificado. Ambas as abordagens mantêm a mesma estrutura lógica do caso de tamanhos iguais, adaptando-se apenas o $m_{\text{ótimo}}$ para o tamanho médio $\bar{m}_{\text{ótimo}}$.

Caso seja desejado que os tamanhos da amostra no segundo estágio sejam diferentes, pode-se fazer a seguinte distribuição (Silva, 2025):

$$m_i = \frac{a\bar{m}M_i}{\sum_{i=1}^{a} M_i} \tag{2.21}$$

Note que primeiro é necessário selecionar a amostra de UPAs para depois calcular os m_i.

Em planos amostrais com conglomerados de tamanhos desiguais e em múltiplos estágios, as diferentes probabilidades de seleção criam um desequilíbrio na representação das unidades. Como destacado por Kish (1965), conglomerados maiores têm maior probabilidade de serem selecionados no primeiro estágio, enquanto suas unidades internas apresentam menor chance de inclusão no segundo estágio.

Esse desbalanceamento é corrigido através da aplicação de pesos amostrais, que

ajustam a contribuição de cada observação para refletir corretamente sua proporção na população-alvo (Kish, 1965). A omissão ou má especificação desses pesos pode levar a distorções nas estimativas, afetando tanto a precisão quanto a validade inferencial dos resultados (Cochran, 1977). O peso amostral pode ser calculado como:

$$W_{ij} = W_i \times W_{j|i} = W_1 \times W_2 \tag{2.22}$$

onde:

W_i é o peso do conglomerado i no primeiro estágio, dado por:

$$W_i = \frac{\sum_{i=1}^{A} M_i}{aM_i} = \frac{M_0}{aM_i} \tag{2.23}$$

sendo M_0 o total populacional, M_i o tamanho populacional do conglomerado i e a o número de conglomerados selecionados na amostra do primeiro estágio.

$W_{j|i}$ é o peso da unidade j dentro do conglomerado i, dado pela equação:

$$W_{j|i} = \frac{M_i}{m_i} \tag{2.24}$$

sendo M_i o tamanho populacional no conglomerado i e m_i o número de unidades selecionadas no conglomerado i.

Como exemplo, a Tabela 2.1 apresenta 10 setores censitários hipotéticos considerados como UPAs, totalizando 1.919 domicílios na população. Deseja-se selecionar uma amostra de $\bar{m} = 36$ domicílios, com $a = 5$ setores sorteados no primeiro estágio utilizando seleção com probabilidade proporcional ao tamanho (PPT).

Tabela 2.1: Exemplo de 10 setores com os domicílios

Setor	**Domicílios**	**Setor**	**Domicílios**
1	443	6	291
2	162	7	64
3	127	8	25
4	354	9	232
5	119	10	102
		Total	**1.919**

Suponha que foram selecionados na amostra os setores 1, 2, 7, 8 e 10. Para esses setores sorteados, calcula-se a quantidade de domicílios a serem selecionados no segundo estágio (m_i) de forma proporcional ao tamanho de cada setor, respeitando o total de $\bar{m} = 36$ domicílios por setor, em média.

A Tabela 2.2 apresenta os valores de M_i, os respectivos tamanhos de amostra m_i selecionados dentro de cada setor, os pesos do primeiro e segundo estágio (W_1 e W_2) e o peso final atribuído a cada unidade amostral (W_{ij}). A última coluna mostra o produto $m_i \times \text{Peso}_{ij}$, que representa a expansão de cada conglomerado para a população total.

Tabela 2.2: Exemplo não equiprovável com cálculo do m_i e peso amostral

Setor	M_i	m_i	W_1	W_2	$Peso_{ij}$	$m_i \times Peso_{ij}$
1	443	100	0,866	4,430	**3,838**	**383,800**
2	162	37	2,369	4,378	**10,373**	**383,801**
7	64	14	5,997	4,571	**27,414**	**383,796**
8	25	6	15,352	4,167	**63,967**	**383,802**
10	102	23	3,763	4,435	**16,687**	**383,801**
Total	**796**	**180**	–	–	–	**1.919**

Este exemplo mostra como o cálculo adequado dos pesos amostrais garante que cada unidade represente corretamente sua parcela da população. A última coluna evidencia que, embora os tamanhos m_i e os pesos W_{ij} variem entre os conglomerados, o produto $m_i \times W_{ij}$ permanece aproximadamente constante, totalizando exatamente os 1.919 domicílios da população. Isso confirma que o plano amostral não equiprovável foi devidamente calibrado, permitindo a expansão fiel dos resultados amostrais para a população-alvo.

Em contraste, nos planos equiprováveis - como em amostragens aleatórias simples ou em conglomerados de mesmo tamanho - todas as unidades têm a mesma probabilidade de seleção em cada estágio, tornando os pesos amostrais desnecessários para a estimação da média, conforme discutido por Cochran (1977). Nesses casos, dois requisitos fundamentais devem ser atendidos para garantir a equiprobabilidade: a homogeneidade no tamanho dos conglomerados e a uniformidade na fração amostral entre os estágios, conforme argumentado por Kish (1965).

Como ressalta Lohr (2010), quando essas condições são cumpridas, é possível obter estimativas pontuais para a média diretamente, sem a necessidade de ponderação. Isso simplifica tanto o desenho da amostra quanto o processo de análise estatística.

2.3.3 Proporções

Para proporções em amostragem por conglomerados em dois estágios (AC2), o estimador $\bar{p}$ e sua variância estimada são dados por:

$$\bar{p} = \frac{1}{a}\sum_{i=1}^{a} p_i = \frac{1}{am}\sum_{i=1}^{a}\sum_{j=1}^{m} p_{ij}, \tag{2.25}$$

onde: p_{ij} é indicador binário 0 ou 1 da característica de interesse na unidade j do conglomerado i e $p_i = \frac{1}{m}\sum_{j=1}^{m} p_{ij}$ é a proporção amostral no i-ésimo conglomerado.

A variância estimada de $\bar{p}$ é:

$$\widehat{\text{Var}}(\bar{p}) = \frac{(1-f_1)}{a(a-1)}\sum_{i=1}^{a}(p_i-\bar{p})^2 + \frac{f_1(1-f_2)}{a^2m(m-1)}\sum_{i=1}^{a} p_i(1-p_i) \tag{2.26}$$

2.4 Amostra Equiprovável

Uma amostra equiprovável (ou do inglês *Equal Probability of Selection Method* (EPSEM)) é aquela em que todos os elementos da população têm igual probabilidade de inclusão na amostra, análogo ao que ocorre em uma amostragem aleatória simples (Cochran, 1977). No entanto, em desenhos amostrais complexos, como na amostragem por conglomerados com tamanhos desiguais, com estratificação e múltiplos estágios, esse princípio requer um cuidadoso balanceamento das probabilidades nos múltiplos estágios (Kish, 1965).

Para um desenho amostral em dois estágios, onde os conglomerados possuem tamanhos desiguais (M_i), a probabilidade final de inclusão de uma unidade elementar, denotada por f, deve satisfazer a equação:

$$f = \frac{n}{N} = \underbrace{\frac{aM_i}{\sum_{i=1}^{A} M_i}}_{f_1} \times \underbrace{\frac{m_i}{M_i}}_{f_2} = \frac{am_i}{\sum_{i=1}^{A} M_i} \tag{2.27}$$

Como observado por Lohr (2010), o mecanismo compensatório que garante a equiprobabilidade funciona da seguinte maneira:

- Conglomerados maiores (M_i grande) têm maior probabilidade de seleção no primeiro estágio (f_1), mas menor fração de seleção interna (f_2);

- Conglomerados menores (M_i pequeno) têm menor f_1, mas maior f_2.

Essa compensação assegura que o produto $f_1 \times f_2$ seja constante para todas as unidades elementares da população, garantindo uma amostragem do tipo EPSEM. Para que essa condição seja satisfeita, é necessário que:

$$m_i = \frac{n}{a} \quad \text{(tamanho constante da subamostra)} \tag{2.28}$$

ou seja, o número de unidades selecionadas em cada conglomerado deve ser constante, independentemente de seu tamanho.

Planos equiprováveis (EPSEM) se destacam por sua simplicidade: estimadores como médias e proporções permanecem não viesados e sua variância pode ser estimada por equações tradicionais (Cochran, 1977). Em contraste, como discutido por Skinner e Wakefield (2017), planos não equiprováveis exigem o uso de pesos e técnicas especializadas, como a linearização por série de Taylor dentre outros métodos computacionais, para garantir inferência válida.

A Equação (2.29) resume a estrutura de compensação que garante a equiprobabilidade no plano amostral complexo:

$$f = f_1 \times f_2 = \left(\frac{aM_i}{\sum_{i=1}^{A} M_i} \right) \times \left(\frac{m}{M_i} \right) = \frac{am}{\sum_{i=1}^{A} M_i} = \frac{n}{N} \tag{2.29}$$

Embora o método EPSEM preserve a equiprobabilidade mesmo com conglomerados heterogêneos, sua aplicação prática pode enfrentar dificuldades operacionais. Em particular, a grande variação no tamanho dos conglomerados pode causar subdimensionamento e superdimensionamento.

2.5 Superdimensionamento

No contexto de amostragem complexa, a utilização de uma seleção PPT é uma estratégia essencial para otimizar a eficiência do plano amostral, seja em desenhos equiprováveis ou não equiprováveis. A probabilidade de seleção de um conglomerado i é dada por $f_{1i} = \frac{aM_i}{\sum_{i=1}^{A} M_i}$, onde M_i representa o tamanho do i-ésimo conglomerado, a o número de conglomerados amostrados, e A o total de conglomerados na população.

Entretanto, quando alguns conglomerados possuem tamanho M_i excessivamente grande, pode ocorrer o fenômeno conhecido como *superdimensionamento*, em que $f_{1i} > 1$ ou $\frac{\sum_{i=1}^{A} M_i}{a} < M_i$. Isso acaba por gerar uma amostra menor de elementos, visto que

o mesmo conglomerado pode ser selecionado mais de uma vez, além de se ter pesos amostrais menores do que 1 (o que é contra-intuitivo).

Esse problema é discutido por autores como Kish (1965) e Lohr (2010), que destacam a necessidade de adaptação do plano para manter sua viabilidade estatística e operacional. Segundo Lohr (2010), conglomerados com $f_{1i} > 1$ devem ser tratados de maneira diferenciada para evitar a seleção do mesmo conglomerado mais de uma vez.

Uma solução amplamente recomendada para o problema de superdimensionamento é a estratificação das unidades superdimensionadas, conforme proposto por Kish (1965). Nesse procedimento, os conglomerados com probabilidade de seleção superior a 1 (isto é, $f_{1i} > 1$) são alocados em um estrato separado, onde são selecionados com certeza (probabilidade igual a 1). Essa abordagem garante a correção da violação da condição $f_{1i} \leq 1$ e a adequação dos pesos amostrais.

Essa estratégia é particularmente relevante em pesquisas domiciliares ou populacionais que envolvem grandes capitais ou unidades primárias de grande porte. É importante destacar que esse problema pode ocorrer tanto no estágio inicial de seleção quanto em estágios intermediários do plano amostral.

Considere o exemplo Não equiprovável da Tabela 2.1, em que os setores 1, 2, 7, 8 e 10 foram selecionados com probabilidade proporcional ao tamanho. No segundo estágio, definiu-se a seleção de $\bar{m} = 36$ domicílios por setor. Observa-se que o Setor 1 é **superdimensionado**:

$$f_1 = \frac{aM_1}{\sum_{i=1}^{A} M_i} = \frac{5 \times 443}{1.919} = 1,152$$

Como $f_1 > 1$, o setor poderia ser sorteado mais de uma vez, configurando o problema de superdimensionamento.

Na prática, essa situação seria identificada automaticamente, já que o procedimento de sorteio indicaria múltiplas seleções para o mesmo setor. Para contornar o problema, aplica-se uma estratificação: o Setor 1 é isolado em um **estrato certo (Estrato I)**, incluído com peso 1, garantindo sua presença na amostra. Os demais setores são agrupados em um **segundo estrato (Estrato II)**, no qual os pesos são recalculados de forma a preservar a coerência da expansão em relação ao total populacional.

O peso do primeiro estágio, W_{1hi}, é definido da seguinte forma:

$$W_{1hi} = \begin{cases} 1, & \text{se } h = 1 \\ \frac{M_{02}}{a_2 \times M_i}, & \text{se } h = 2 \end{cases}$$

onde a_2 é o número de conglomerados alocados no estrato 2, podendo ser definido como $a_2 = a - a_1$, e M_{02} é o total populacional desse estrato, podendo ser expresso como $M_{02} = M_0 - M_{01}$.

Aplicando ao exemplo, o Setor 1 integra o Estrato I e recebe $W_{111} = 1$. Já o Setor 2, pertencente ao Estrato II, tem peso ajustado:

$$W_{122} = \frac{M_{02}}{a_2 \times M_2} = \frac{M_0 - M_{01}}{(a - a_1) \times M_2} = \frac{1.919 - 443}{(5-1) \times 162} = \frac{1.476}{4 \times 162} = 2,278$$

Assim, assegura-se que o setor superdimensionado participará da amostra apenas uma vez, enquanto os demais mantêm pesos proporcionais, garantindo a representatividade do total populacional.

No segundo estágio, os pesos são calculados independentemente da estratificação adotada no primeiro estágio, seguindo a fórmula usual:

$$W_{2j|hi} = \frac{M_i}{m_i}$$

Assim, tem-se:

- Para o Setor 1 (Estrato I):

$$W_{2j|11} = \frac{M_1}{m_1} = \frac{443}{100} = 4,430$$

- Para o Setor 2 (Estrato II):

$$W_{2j|22} = \frac{M_2}{m_2} = \frac{162}{37} \approx 4,378$$

O resultado final do ajuste encontra-se na Tabela 2.3.

2.6 Subdimensionamento

O subdimensionamento ocorre quando a dimensão populacional M_i de um conglomerado é insuficiente para atender à demanda amostral m_i pré-estabelecida ($M_i < m_i$),

Tabela 2.3: Exemplo com ajuste de pesos após detecção de superdimensionamento

Setor	Estrato	M_i	m_i	W_1	W_2	$Peso_{ij}$	$m_i \times Peso_{ij}$
1	(Estrato I)	443	100	1,000	4,430	**4,430**	**443,000**
2	(Estrato II)	162	37	2,278	4,378	**9,973**	**369,000**
7	(Estrato II)	64	14	5,766	4,571	**26,357**	**369,000**
8	(Estrato II)	25	6	14,760	4,167	**61,500**	**369,000**
10	(Estrato II)	102	23	3,618	4,435	**16,043**	**369,000**
Total	–	**796**	**180**	–	–	–	**1.919**

inviabilizando o sorteio. Esse cenário é mais comum em planos amostrais com tamanho fixo de conglomerados ($m_i = \bar{m}$) do que em desenhos com cálculo do m_i de diferentes tamanhos a partir do $\bar{m}$. Notavelmente, esse problema pode surgir tanto em estágios intermediários quanto no estágio final do plano amostral.

A literatura, como discutido em Kish (1965) e Cochran (1977), recomenda que, nesses casos, os conglomerados subdimensionados sejam combinados com unidades mais próximas, formando uma nova unidade composta com tamanho M_i' suficientemente grande para suportar a alocação definida ($M_i' \geq m_i$).

Contudo, essa solução pode induzir outro problema: o conglomerado resultante pode tornar-se excessivamente grande, levando a um caso de superdimensionamento em que a nova unidade composta tem probabilidade de seleção $f_{1i} > 1$. Assim, o ajuste requer um balanceamento cuidadoso entre o controle do subdimensionamento e a prevenção do superdimensionamento.

Considere o exemplo apresentado na Tabela 2.1, suponha que os setores sorteados foram 1, 2, 7, 8 e 10 de forma equiprovável com $\bar{m} = 36$ domicílios por setor, em média. A Tabela 2.4 apresenta o problema da distribuição, onde o conglomerado 8 (destacado em vermelho) possui apenas 25 domicílios, sendo insuficiente para atender à cota de 36 domicílios requerida, caracterizando um caso de subdimensionamento.

Tabela 2.4: Exemplo equiprovável com subdimensionamento

Setor	M_i	m_i	W_1	W_2	$Peso_{ij}$	$m_i \times Peso_{ij}$
1	443	36	0,866	12,306	**10,661**	**383,8**
2	162	36	2,369	4,500	**10,661**	**383,8**
7	64	36	5,997	1,778	**10,661**	**383,8**
8	25	36	–	–	–	–
10	102	36	3,763	2,833	**10,661**	**383,8**
Total	**796**	**180**	–	–	–	–

Para corrigir o subdimensionamento, propõe-se agrupar o conglomerado 8 (com 25 domicílios) com o conglomerado 7 (com 64 domicílios), resultando em um novo con-

glomerado com 89 domicílios, como mostrado na Tabela 2.5.

Tabela 2.5: Ajuste para correção do subdimensionamento (fusão dos setores 7 e 8)

Setor	**Domicílios**	**Setor**	**Domicílios**
1	443	6	291
2	162	7+8	64+25
3	127	9	232
4	354	10	102
5	119		
		Total	**1.919**

Diante da necessidade de agrupamento do setor 8 com o setor 7, procedeu-se à seleção de um novo setor para compor a amostra. O setor 9 foi escolhido, garantindo assim que todos os conglomerados atendessem ao tamanho mínimo requerido de 36 domicílios. A configuração final corrigida é apresentada na Tabela 2.6, onde se observa a adequação de todos os elementos amostrais aos critérios estabelecidos, e que os pesos são iguais em todas as unidades.

Tabela 2.6: Exemplo equiprovável com subdimensionamento corrigido

Setor	M_i	m_i	W_1	W_2	$Peso_{ij}$	$m_i \times Peso_{ij}$
1	443	36	0,866	12,306	**10,661**	**383,8**
2	162	36	2,369	4,500	**10,661**	**383,8**
7+8	89	36	4,312	2,472	**10,661**	**383,8**
9	232	36	1,654	6,444	**10,661**	**383,8**
10	102	36	3,763	2,833	**10,661**	**383,8**
Total	**1.028**	**180**	–	–	–	**1.919**

Na prática, o Instituto Brasileiro de Geografia e Estatística (IBGE) implementa uma solução eficiente para esse dilema por meio da definição de *áreas de ponderação*, como descrito em suas notas metodológicas (IBGE, 2014). Essas áreas são unidades geográficas artificiais construídas a partir da junção de setores censitários com características semelhantes e tamanhos adequados para garantir a viabilidade estatística e operacional da amostra. Cada área de ponderação é construída de modo a evitar conglomerados muito pequenos que não contenham pelo menos 400 domicílios.

2.7 Amostragem por Conglomerados em Três Estágios

Esta Seção amplia os conhecimentos da Amostragem por Conglomerados em dois estágios (AC2) para contemplar o caso de Amostragem por Conglomerados em três estágios (AC3). Seja A o número total de UPAs na população, M_i o número de USAs na i-ésima UPA e L_{ij} o número de Unidades Terciárias de Amostragem (UTA) na j-ésima

USA da i-ésima UPA. Na amostra, denota-se por a o número de UPAs selecionadas, m_i o número de USAs selecionadas na i-ésima UPA e l_{ij} o número de UTAs selecionadas na USA (i, j).

A função de custo esperado segue a expansão da Equação (2.15), e é dada por:

$$C = c_0 + c_1 a + c_2 a\bar{m} + c_3 a\bar{m}\bar{l} \tag{2.30}$$

onde:

c_1 é o custo fixo por UPA e eventuais custos médios de listagem;

c_2 é o custo por USA;

c_3 é o custo por UTA;

$\bar{m}$ é o número médio de USAs selecionadas por UPA: $\bar{m} = \frac{1}{a}\sum_{i=1}^{a} m_i$;

$\bar{l}$ é o número médio de UTAs selecionadas por USA: $\bar{l} = \frac{1}{a\bar{m}}\sum_{i=1}^{a}\sum_{j=1}^{m_i} l_{ij}$.

A equação da média amostral por subsubelemento é expressa por:

$$\bar{\bar{y}} = \frac{1}{a\bar{m}\bar{l}}\sum_{i=1}^{a}\sum_{j=1}^{\bar{m}}\sum_{k=1}^{\bar{l}} y_{ijk} \tag{2.31}$$

A variância da média amostral por subsubelemento $\bar{\bar{y}}$ é dada pela equação:

$$Var(\bar{\bar{y}}) = (1 - f_1)\frac{S_1^2}{a} + (1 - f_2)\frac{S_2^2}{a\bar{m}} + (1 - f_3)\frac{S_3^2}{a\bar{m}\bar{l}}, \tag{2.32}$$

onde:

S_1^2 é a variância entre UPAs (Unidades Primárias de Amostragem);

S_2^2 é a variância entre USAs (Unidades Secundárias de Amostragem) dentro de UPA, calculada como média ponderada pelos tamanhos das UPAs;

S_3^2 é a variância entre UTAs (Unidades Terciárias de Amostragem) dentro das USAs;

$f_1 = \frac{a}{A}$ é a fração amostral do primeiro estágio, com a UPAs selecionadas de um total de A UPAs na população;

$f_2 = \frac{\bar{m}}{\bar{M}}$ é a fração amostral do segundo estágio, representando a proporção média de USAs selecionadas por UPA;

$f_3 = \frac{\bar{l}}{\bar{L}}$ é a fração amostral do terceiro estágio, representando a proporção média de UTAs selecionadas por USA;

$\bar{M} = \frac{1}{A}\sum_{i=1}^{A} M_i$ é o número médio de USAs por UPA na população, onde M_i é o número de USAs na i-ésima UPA;

$\bar{L} = \frac{\sum_{i,j} L_{ij}}{\sum_i M_i}$ é o número médio de UTAs por USA na população.

As componentes de variância são calculadas como:

$$S_1^2 = \frac{1}{A-1}\sum_{i=1}^{A}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 \tag{2.33}$$

$$S_2^2 = \frac{1}{A(\bar{M}-1)}\sum_{i=1}^{A}\sum_{j=1}^{\bar{M}}\left(\bar{Y}_{ij} - \bar{Y}_i\right)^2 \tag{2.34}$$

$$S_3^2 = \frac{1}{A\bar{M}(\bar{L}-1)}\sum_{i=1}^{A}\sum_{j=1}^{\bar{M}}\sum_{k=1}^{\bar{L}}\left(Y_{ijk} - \bar{Y}_{ij}\right)^2 \tag{2.35}$$

onde:
$\bar{\bar{Y}} = \frac{1}{A\bar{M}\bar{L}}\sum_{i=1}^{A}\sum_{j=1}^{\bar{M}}\sum_{k=1}^{\bar{L}} Y_{ijk}$ é a média populacional verdadeira por UTA;
$\bar{Y}_i = \frac{1}{\bar{M}\bar{L}}\sum_{j=1}^{\bar{M}}\sum_{k=1}^{\bar{L}} Y_{ijk}$ é a média verdadeira da i-ésima UPA;
$\bar{Y}_{ij} = \frac{1}{\bar{L}}\sum_{k=1}^{\bar{L}} Y_{ijk}$ é a média verdadeira da j-ésima USA dentro da i-ésima UPA;
Y_{ijk} é o valor da variável de interesse para a k-ésima UTA dentro da j-ésima USA e da i-ésima UPA.

De acordo com Silva (2025), uma estratégia para determinação dos tamanhos amostrais em AC3 é utilizar a estrutura do ρ, podendo ser utilizada tanto para uma estimação de média quanto para proporções, determinando $\bar{l}_{\text{ótimo}}$.

Para determinar o tamanho ótimo das UTAs $\bar{l}_{\text{ótimo}}$ utiliza-se a equação:

$$\bar{l}_{\text{ótimo}} = \sqrt{\frac{1-\rho}{\rho}\frac{c_2}{c_3}} \tag{2.36}$$

Em seguida, o número médio de USAs por UPA, $\bar{m}_{ótimo}$, é calculado com base na precisão desejada, considerando um erro relativo máximo r, o coeficiente de variação populacional CV e a correlação intraclasse ρ entre as USAs. A equação é dada por:

$$\bar{m}_{ótimo} = \frac{z_{\gamma/2}^2 CV^2\left(\rho + \frac{1-\rho}{\bar{l}_{ótimo}}\right)}{r^2} \tag{2.37}$$

Por fim, o número de UPAs a pode ser calculado a partir da equação de custo total, assumindo como fixados os valores de $\bar{m}_{\text{ótimo}}$, $\bar{l}_{\text{ótimo}}$ e um custo disponível para a pesquisa, sendo dada por:

$$a = \frac{C - c_0}{c_1 + c_2\bar{m}_{\text{ótimo}} + c_3\bar{m}_{\text{ótimo}}\bar{l}_{\text{ótimo}}} \tag{2.38}$$

Assim, a abrangência da pesquisa depende do orçamento disponível, enquanto

a definição da amostra dentro de cada unidade do primeiro estágio é determinada pela margem de erro e pelo nível de precisão desejado (Silva, 2025).

Os pesos amostrais para três estágios são dados pela seguinte equação:

$$W_{ijk} = W_i \times W_{j|i} \times W_{k|ij} = W_1 \times W_2 \times W_3 \tag{2.39}$$

onde:

W_i é o peso do conglomerado i no primeiro estágio, dado por:

$$W_i = \frac{\sum_{i=1}^{A} M_i}{aM_i} = \frac{M_0}{aM_i} \tag{2.40}$$

sendo M_0 o total populacional, M_i o tamanho populacional na i-ésima UPA e a o número de UPAs selecionadas na amostra do primeiro estágio.

$W_{j|i}$ é o peso da unidade j dentro do conglomerado i, dado, no caso padrão, por:

$$W_{j|i} = \frac{M_i}{\bar{m}L_{ij}} \tag{2.41}$$

onde L_{ij} é o tamanho populacional da j-ésima USA da i-ésima UPA e $\bar{m}$ é o número médio de USAs selecionadas na amostra do segundo estágio. Para o caso não equiprovável, é dado por:

$$W_{j|i} = \frac{M_i}{m_i L_{ij}} \tag{2.42}$$

sendo m_i o número de USAs selecionadas na amostra do segundo estágio.

$W_{k|ij}$ é o peso da unidade k dentro da j-ésima USA do conglomerado i, dado, no caso padrão, por:

$$W_{k|ij} = \frac{L_{ij}}{\bar{l}} \tag{2.43}$$

onde $\bar{l}$ é o número médio de UTAs selecionadas na amostra do terceiro estágio. Para o caso não equiprovável, é dado por:

$$W_{k|ij} = \frac{L_{ij}}{l_{ij}} \tag{2.44}$$

sendo l_{ij} o número de UTAs selecionadas na amostra do terceiro estágio.

A ocorrência dos problemas de dimensionamento segue o padrão apresentado nas Tabelas 2.7, 2.8 e 2.9. Observa-se que o superdimensionamento sempre ocorre na UPA, enquanto o subdimensionamento sempre ocorre na última unidade amostral. Nas unidades intermediárias, como a USA e a UTA nos 3º e 4º estágios, podem ocorrer ambos os tipos

de problema. As Tabelas 2.7, 2.8 e 2.9 também indicam em qual estágio cada problema é resolvido e qual unidade é afetada. Assim, as setas ilustram que o subdimensionamento é solucionado no primeiro estágio, por meio do agrupamento de UPAs, o que, por sua vez, pode aumentar a ocorrência de superdimensionamento nas UPAs.

Tabela 2.7: Progressão dos estágios e problemas de dimensionamento até o 2º estágio

Parâmetro	**1º Estágio**	**2º Estágio**
UPA	Superdimensionada	—
USA	Subdimensionada	→ Subdimensionada

Tabela 2.8: Progressão dos estágios e problemas de dimensionamento até o 3º estágio

Parâmetro	**1º Estágio**	**2º Estágio**	**3º Estágio**
UPA	Superdimensionada	—	—
USA	Subdimensionada	→ Superdimensionada	—
UTA	—	Subdimensionada	→ Subdimensionada

Tabela 2.9: Progressão dos estágios e problemas de dimensionamento até o 4º estágio

Parâmetro	**1º Estágio**	**2º Estágio**	**3º Estágio**	**4º Estágio**
UPA	Superdimensionada	—	—	—
USA	Subdimensionada	→ Superdimensionada	—	—
UTA	—	Subdimensionada	→ Superdimensionada	—
UQA	—	—	Subdimensionada	→ Subdimensionada

Capítulo 3

Materiais e Métodos

3.1 Introdução

Este Capítulo descreve os materiais e métodos empregados no desenho de planos amostrais complexos para levantamentos domiciliares, abordando especificamente os desafios de subdimensionamento e superdimensionamento de unidades em seleções multietápicas. Foram implementados dois planos amostrais contrastantes – um equiprovável e outro não equiprovável – seguindo desenho em três estágios (municípios, setores censitários urbanos/rurais e domicílios), com restrições orçamentárias pré-definidas e utilizando algoritmos computacionais desenvolvidos para tratamento dos problemas de dimensionamento.

Os materiais incluem bases do IBGE para o processo de seleção da amostra. O método compreende desde o cálculo do tamanho amostral até a seleção final das unidades, com comparação sistemática entre os planos baseada em critérios de variância, custo e eficiência amostral, garantindo a viabilidade técnica dentro do orçamento estabelecido.

3.2 Materiais

Foram utilizados dois bancos de dados principais como suporte ao delineamento amostral:

Banco de setores censitários do Instituto Brasileiro de Geografia e Estatística (IBGE), contendo informações agregadas sobre domicílios por setor censitário, proveniente do Censo Demográfico de 2022. Para este estudo, foram utilizadas as seguintes variáveis:

Tabela 3.1: Variáveis utilizadas do banco de setores censitários do IBGE

Variável	**Descrição**
CD_UF	Código da Unidade da Federação
NM_UF	Nome da Unidade da Federação
CD_MUN	Código do município
NM_MUN	Nome do município
CD_SETOR	Geocódigo do setor censitário
SITUACAO	Classificação urbana/rural do setor
V0001	Total de pessoas no setor censitário
V0002	Total de domicílios no setor censitário

Banco do Cadastro Nacional de Endereços para Fins Estatísticos (CNEFE), que serviu como base para a seleção do terceiro estágio dos planos amostrais; Malha municipal digital dos estados, disponibilizado pelo IBGE; e malha de setores censitários dos estados, disponibilizado pelo IBGE.

3.3 Métodos

Nesta Seção, são apresentados os procedimentos metodológicos adotados para a realização do estudo, incluindo as ferramentas computacionais utilizadas e os planos amostrais implementados.

3.3.1 Algoritmo de solução para Superdimensionamento

O Algoritmo 1, utilizado para solucionar o problema de superdimensionamento, recebe como parâmetros de entrada: M_0 (tamanho total da população), a (tamanho da amostra desejado), um conjunto de **dados** contendo a variável M_i (tamanho populacional de cada unidade), além de outras variáveis de identificação. Como resultado, o algoritmo retorna o conjunto de dados original acrescido de duas colunas: `estrato_k`, que detalha a configuração dos estratos identificados, e `pesos`, contendo o peso calculado para cada unidade.

Algoritmo 1: Algoritmo para Correção de Superdimensionamento

para *cada unidade* i **faça**
  Calcule o peso inicial: $peso_i = \frac{M_0}{a \times M_i}$
fim
$k = 1$
Passo 2: Processamento Iterativo
enquanto *algum* $peso_i < 1$ **faça**
  para *cada unidade* i **faça**
    se $peso_i < 1$ **então**
      Marque $\texttt{estrato}_k = 1$
    caso contrário
      Marque $\texttt{estrato}_k = 0$
    fim
  fim
  Calcule:
  M_{02}: tamanho total da população no $\texttt{estrato}_k = 0$
  a_2: tamanho da amostra no $\texttt{estrato}_k = 0$
  para *cada unidade* i **faça**
    se $estrato_k = 1$ **então**
      $peso_i = 1$
    caso contrário
      $peso_i = \frac{M_{02}}{a_2 \times M_i}$
    fim
  fim
  $k = k + 1$
fim
Passo 3: Resultado Final
Gera tabela com a coluna $\texttt{estrato}_k$ detalhando as unidades no estrato os estratos

3.3.2 Algoritmo de solução para Subdimensionamento

O Algoritmo 2, utilizado para solucionar o problema de subdimensionamento, recebe como parâmetros de entrada: `dados` (conjunto de dados contendo as unidades), `id` (identificador único da unidade), `tamanho_min` (tamanho mínimo requerido), `tamanho_tot` (tamanho populacional total de cada unidade), `pop` (variável populacional utilizada nos cálculos), `vizinhos` (tabela de vizinhança com relações de adjacência) e `mapa` (arquivo shapefile opcional para visualização geográfica). Como resultado, o algoritmo retorna o conjunto de dados original acrescido de colunas: `id_final` (identificador final do agrupamento), `id_hist` (histórico completo de junções) e o número de junções realizadas por unidade.

Algoritmo 2: Algoritmo de Agregação para Unidades Subdimensionadas

Passo 1: Inicialização
Ordenar unidades por `tamanho_tot` (Crescente)
Inicializar `id_final` e $k = 1$
Verificar unidades subdimensionadas (`tamanho_tot` < `tamanho_min`) então `subdim` = 1
Passo 2: Primeira Agregação
para *cada unidade com* `subdim` = *1* **faça**
    se *unidade não foi agregada* ***E*** *Vizinho não foi agregado* **então**
        Agregar unidade ao vizinho e registrar junção no histórico com `juncao` = 1
    fim
fim
Atualizar `id_final` das novas unidades (agrupadas)
Recalcular `tamanho_tot`
reclassificar unidades subdimensionadas (`tamanho_tot` < `tamanho_min`) então `subdim` = 1
Passo 3: Loop Principal
enquanto *existem unidades* `subdim` = 1 **faça**
    Ordenar agrupamentos por `tamanho_tot` (crescente)
    para *cada unidade* `subdim` **faça**
        se *unidade não foi agregada* ***E*** *Vizinho não foi agregado* **então**
            Agregar unidade ao vizinho e registrar junção no histórico
        fim
    fim
    Expansão para Vizinhos de Segunda Ordem (Cres cente)
    se *nenhuma agregação realizada* **então**
        para *cada agrupamento* `subdim` **faça**
            Buscar vizinhos distantes
            Ordenar vizinhos por `tamanho_tot` (Crescente)
            se *unidade não foi agregada* ***E*** *Vizinho não foi agregado* **então**
                Agregar unidade ao vizinho e registrar junção no histórico
            fim
        fim
    fim
    Expansão para Vizinhos de Segunda Ordem Decrescente
    se *nenhuma agregação realizada* **então**
        Buscar vizinhos distantes
        Ordenar vizinhos por `tamanho_tot` (decrescente)
        para *cada agrupamento* `subdim` **faça**
            Buscar vizinhos mais distantes (terceira ordem)
            se *unidade não foi agregada* ***E*** *Vizinho não foi agregado* **então**
                Agregar unidade ao vizinho e registrar junção no histórico
            fim
        fim
    fim
    Atualizar `id_final` das novas unidades (agrupadas)
    $k \leftarrow k + 1$
    Recalcular `tamanho_tot` e `pop`
    Reclassificar unidades subdimensionadas (`tamanho_tot` < `tamanho_min`) então `subdim` = 1
fim
Passo 4: Finalização
Construir histórico completo de junções (`id_hist`)
Calcular número total de junções por unidade
Gerar tabela final com `id_final` e população consolidada dos agrupamentos
Produzir visualização geográfica

3.3.3 Planos Amostrais

Os planos amostrais equiprovável e não equiprovável adotam desenho em conglomerados de ttês estágios: municípios (UPAs), setores censitários estratificados em rurais e urbanos (USAs) e domicílios (UTAs). Os parâmetros de dimensionamento incluem: nível de confiança de 95%, custo total de R$140.000,00 (30% fixo), custos unitários de R$600,00 (município), R$12,00 (setor) e R$10,00 (domicílio), Deff = 2,5, proporção P de municípios com até 50 mil habitantes para cálculo do CV, e erro relativo máximo de 2,5% (2% para o Acre).

1. **Definição do tamanho da amostra:**

   - Em ambos os planos, calcula-se inicialmente um tamanho médio para as UTAs, em seguida o tamanho médio das USAs, e, com base nesses tamanhos e com o orçamento disponível, se calcula o tamanho das UPAs.
   - No **plano equiprovável**, este tamanho médio é utilizado de forma uniforme em todas as seleções, isto é, todas as unidades têm a mesma probabilidade de compor a amostra.
   - No **plano não equiprovável**, o cálculo é ponderado de acordo com o tamanho da unidade. Assim, unidades maiores possuem maior probabilidade de serem selecionadas, resultando em probabilidades diferenciadas.

2. **Problemas de super e subdimensionamentos:**

   - Durante a etapa de definição dos tamanhos amostrais, podem ocorrer situações de **superdimensionamento** (quando uma unidade possui tamanho superior ao desejado) ou de **subdimensionamento** (quando uma unidade possui tamanho inferior ao requerido).
   - No caso de **subdimensionamento**, as unidades menores são agrupadas até que atinjam o tamanho mínimo necessário, garantindo que possam compor adequadamente a amostra.
   - Já no **superdimensionamento**, a unidade é alocada em um estrato específico de forma a assegurar que seja selecionada apenas uma vez. Nessa etapa, os pesos são recalculados de acordo com o estrato, preservando a consistência da amostra.

3. **Seleção dos municípios (primeiro estágio):** em ambos os planos, os municípios são sorteados com probabilidade proporcional ao tamanho (PPT), utilizando o banco de dados de domicílios agregados por setor.

4. **Estratificação e seleção dos setores censitários (segundo estágio):** em ambos os planos, os setores são estratificados em *rurais* e *urbanos* com alocação proporcional e sorteados com PPT.

5. **Seleção dos domicílios (terceiro estágio):** em ambos os planos, os domicílios são sorteados por Amostragem Aleatória Simples (AAS) a partir do banco de dados do CNEFE.

6. **Organização da amostra final:**

   - Em cada etapa de seleção, é verificado se o produto entre o peso amostral daquela fase e o total de unidades selecionadas corresponde ao valor populacional total. Ao final, considera-se o produto acumulado de todos os pesos pelas respectivas unidades selecionadas, que deve corresponder ao total populacional, assegurando a calibração do desenho amostral em cada etapa.
   - No **plano equiprovável**, cada unidade da amostra recebe o mesmo peso amostral.
   - No **plano não equiprovável**, cada unidade recebe um peso específico.

7. **Análise:** em ambos os planos, procede-se à análise das informações nos domicílios sorteados. No entanto:

   - No **plano equiprovável**, as estimativas podem ser calculadas sem os pesos amostrais.
   - No **plano não equiprovável**, as estimativas devem incorporar os pesos amostrais para garantir resultados representativos da população.

8. **Variável Simulada:** A variável foi construída a partir de uma distribuição Bernoulli(p), onde a probabilidade p de sucesso corresponde à proporção efetiva de municípios com população inferior a 50 mil habitantes em cada Unidade da Federação.

A comparação entre os dois planos foi conduzida com base nos seguintes critérios:

- Tamanho amostral
- Cálculo do Erro Padrão
- Custos de implementação

Capítulo 4

Análise dos resultados

4.1 Introdução

O presente Capítulo tem como objetivo apresentar e discutir os resultados obtidos a partir da aplicação dos planos amostrais complexos equiprovável e não equiprovável nos estados selecionados para o estudo: Alagoas, Rio de Janeiro, Paraná e Acre. A análise busca avaliar as diferenças entre os dois delineamentos em termos de tamanho amostral, tratamento de super e subdimensionamento, custos e variâncias, conforme os objetivos do trabalho descrito na Seção 1.1.

Cada Seção apresenta, de forma detalhada, os resultados da execução dos algoritmos implementados para o dimensionamento e ajuste dos planos amostrais, conforme os critérios metodológicos definidos anteriormente. São apresentados os agrupamentos de unidades amostrais por ordem de vizinhança; as ocorrências de superdimensionamento e subdimensionamento; e as respectivas soluções aplicadas.

A estrutura analítica adotada segue a mesma lógica para todos os estados, permitindo comparação direta entre os dois métodos. Para cada Unidade da Federação (UF) são apresentados mapas de setores censitários e municípios antes e após as correções; tabelas de interações por ordem de vizinhança; e a comparação entre os planos amostrais equiprovável e não equiprovável quanto a tamanho de amostra esperado, custo estimado e variância da variável simulada.

As simulações e seleções amostrais foram implementadas no *software* SAS 9.4, utilizando programas desenvolvidos especificamente para operacionalizar o processo iterativo de agrupamento e o cálculo dos pesos amostrais. Esse ambiente permitiu o controle da aleatoriedade das seleções e a verificação da consistência entre os estágios.

4.2 Estado de Alagoas

O estado de Alagoas, selecionado aleatoriamente como estudo de caso para a aplicação dos planos amostrais, representa a unidade federativa com a menor extensão territorial entre as analisadas e a 19ª maior população do país.

A população total do estado é de **3.127.683** habitantes, distribuídos em **1.320.151** domicílios, organizados em **102** municípios (com proporção de municípios com até 50 mil habitantes de **9,8%**) e **6.267** setores censitários (desconsiderando setores com população 0) que compõem a base amostral disponível.

Conforme o método descrito no Capítulo 3, o dimensionamento amostral para Alagoas resultou na seleção de **14** UPAs, **45** USAs e **13** UTAs, totalizando **8.190** entrevistas previstas. O custo total esperado, calculado pela Equação (2.30), é de aproximadamente **R$ 139.860,00** para este desenho amostral.

O custo total da amostra, considerando a quantidade selecionada em cada estágio de seleção, pode ser obtido a partir da manipulação da Equação (2.30), conforme apresentado a seguir:

$$\begin{aligned} C &= c_0 + c_1 a + c_2 a\bar{m} + c_3 a\bar{m}\bar{l} \\ &= c_0 + c_1 a + c_2 \sum_{i=1}^{a} m_i + c_3 \sum_{j=1}^{m_i} l_{ij} \end{aligned} \tag{4.1}$$

onde m_i é quantidade de elementos selecionados no segundo estágio (setores censitários) e l_{ij} é a quantidade de elementos selecionados no terceiro estágio (domicílios).

4.2.1 Amostra Equiprόvavel

Para a realização dessa seleção, foi necessário solucionar os casos de subdimensionamento das unidades no primeiro estágio. Identificaram-se 65 municípios com esse problema, os quais foram tratados por meio de um processo iterativo de agrupamento. Na primeira iteração, foram realizados 27 agrupamentos; na segunda, 17; e na terceira, 5, resultando em um total de **53** municípios. Dessa forma, a configuração final das unidades do primeiro estágio ficou estabelecida conforme ilustrado na Figura 4.1.

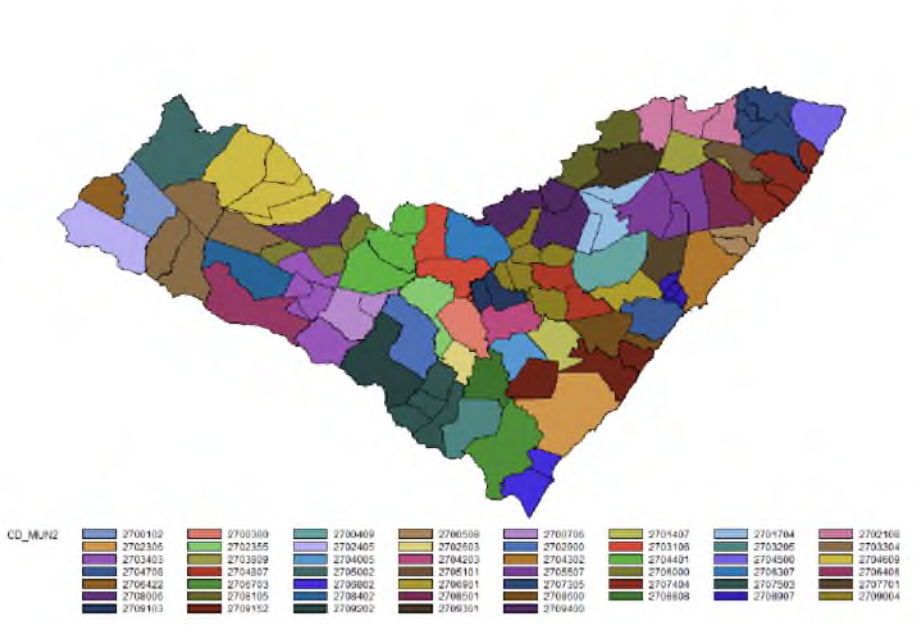


Figura 4.1: Mapa de municípios do estado de Alagoas após resolução de subdimensionamento

No segundo estágio, o problema foi solucionado apenas para os municípios previamente selecionados, conforme apresentado na Figura 4.2. Observa-se que esses municípios estão geograficamente bem distribuídos ao longo do território estadual, o que reflete a diversidade espacial contemplada pela amostra.

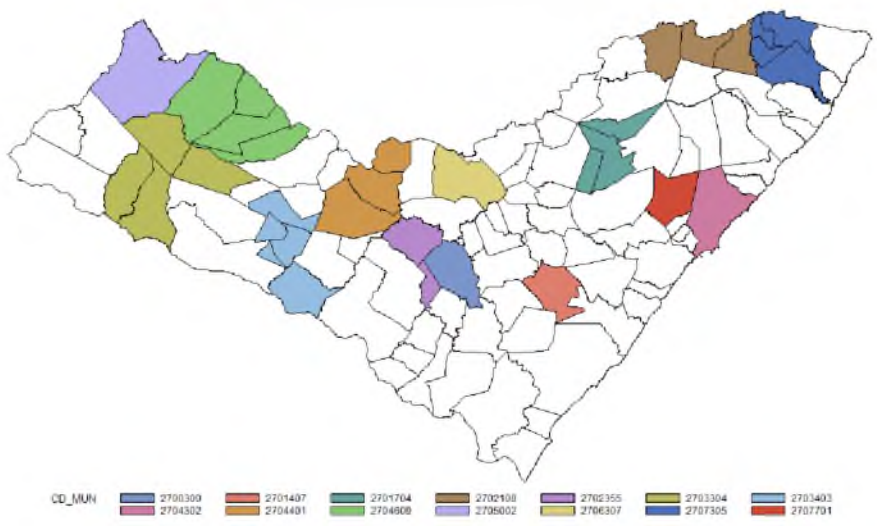


Figura 4.2: Mapa de municípios do estado de Alagoas selecionados para a amostra EPSEM

Identificaram-se 88 setores censitários subdimensionados dentre os 3.209 existentes nos municípios selecionados, os quais foram tratados mediante um processo iterativo de agrupamento. Do processo resultaram 82 agrupamentos, o que está alinhado com a

lógica do Algoritmo 2, que organiza os setores subdimensionados em ordem decrescente de prioridade, agrupando inicialmente os menores entre si com o objetivo de evitar o superdimensionamento, conforme detalhado na Tabela 4.1.

Tabela 4.1: Interações e agrupamentos por ordem de vizinhos de Alagoas

Interação	**Vizinhos de ordem**	**nº de agrupamentos**
1º	1º	72
2º	1º	8
3º	1º	1
4º	2º	1

Como resultado desse tratamento, obteve-se uma configuração final composta por 3.054 setores censitários. Embora a visualização completa dos limites e agrupamentos seja dificultada pelo fato de muitos setores apresentarem dimensões muito reduzidas na imagem, especialmente nas áreas urbanas da capital, a distribuição final das unidades pode ser observada na Figura 4.3.

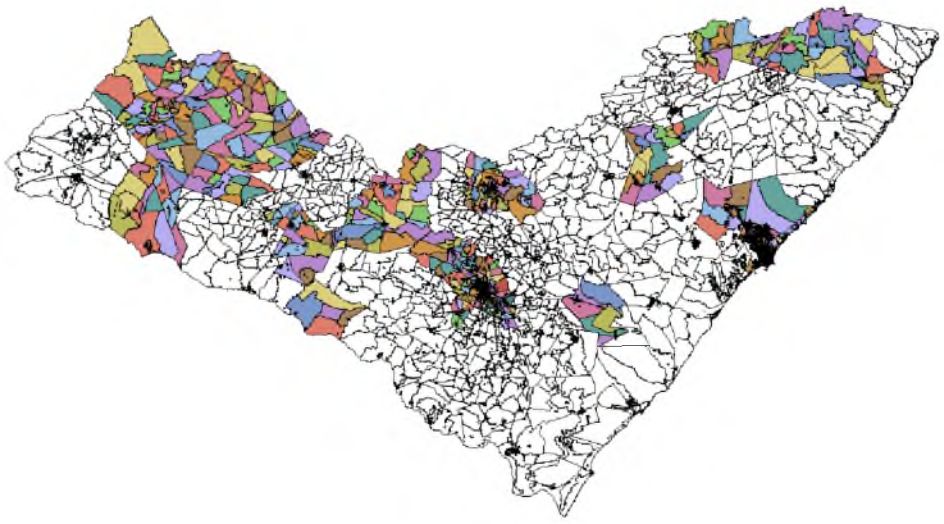


Figura 4.3: Mapa de setores censitários do estado de Alagoas após resolução de subdimensionamento

Já no caso de superdimensionamento, no primeiro estágio, duas unidades municipais foram agrupadas em um estrato certo, de modo a serem selecionadas apenas uma vez. No segundo estágio, 64 setores censitários apresentaram o mesmo problema e foram devidamente corrigidos.

No primeiro estágio foram selecionados 14 municípios, e no segundo, 630 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural. No terceiro e último estágio, foram sorteados 8.190 domicílios, alcançando exatamente os

valores estipulados de tamanho e custo. Em todas as etapas, verificou-se que o produto dos pesos amostrais pela população selecionada correspondia ao total populacional, confirmando que o planejamento amostral foi executado corretamente, conforme a Tabela 4.13.

4.2.2 Amostra Não Equiprόvavel

Como esperado para esse tipo de desenho, a ocorrência de problemas foi reduzida: não houve subdimensionamento, e o superdimensionamento ocorreu apenas no primeiro estágio, afetando duas unidades municipais, que foram agrupadas em um mesmo estrato para serem selecionadas apenas uma vez. No segundo estágio, não foram identificadas unidades superdimensionadas.

No primeiro estágio, foram selecionados 14 municípios, conforme ilustrado na Figura 4.4. Embora essa quantidade se mantenha igual em ambos os planos amostrais, verifica-se que o plano equiprovável inclui 53 municípios em sua composição, um acréscimo de 19 unidades. Este incremento é resultado dos agrupamentos realizados para correção dos casos de subdimensionamento.

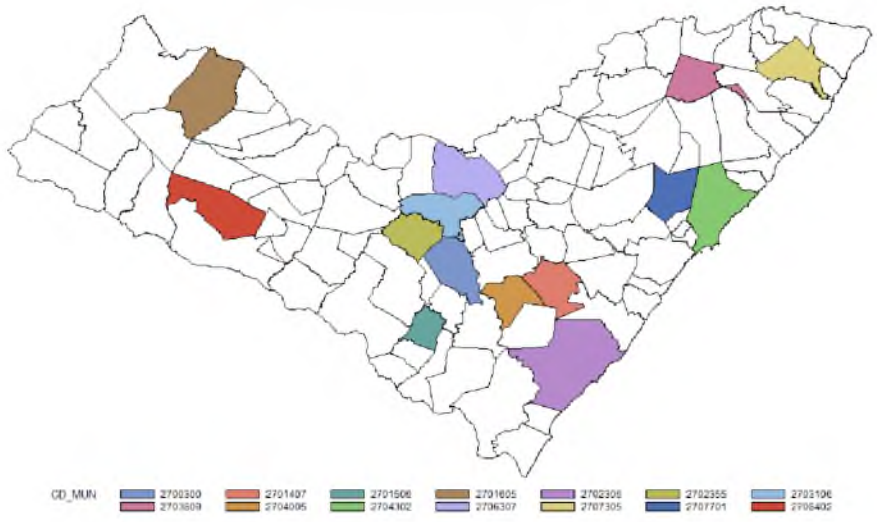


Figura 4.4: Mapa de municípios do estado do Alagoas selecionados para a amostra Não EPSEM

Enquanto no segundo estágio foram selecionados 632 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural, no terceiro e último estágio foram sorteados 8.212 domicílios. Em cada etapa, verificou-se que o produto dos pesos amostrais pela população selecionada correspondia ao total populacional, confirmando

que o planejamento amostral foi seguido corretamente, conforme a Tabela 4.13. Contudo, o total de unidades selecionadas ultrapassou o limite estipulado, resultando em um custo total de **R$140.104,00**, calculado conforme a Equação 4.1, o que excede o orçamento previsto.

A variável de interesse foi simulada conforme descrito no Capítulo 3, seguindo a proporção de municípios definida anteriormente. A Tabela 4.2 apresenta o resumo descritivo dos resultados obtidos para os dois planos amostrais avaliados: o equiprovável (EPSEM) e o não equiprovável (Não EPSEM).

Observa-se que o plano EPSEM manteve coerência com o desenho planejado, resultando em um erro padrão inferior ao obtido pelo plano Não EPSEM, mesmo apresentando um tamanho amostral ligeiramente menor. Essa redução no erro padrão indica maior precisão das estimativas sob o plano equiprovável, reforçando a eficiência do procedimento quando as probabilidades de inclusão são iguais (Kish, 1992).

Tabela 4.2: Comparação entre os planos amostrais de Alagoas

Plano	**Tamanho total da amostra**	**Custo (R$)**	**Erro Padrão**
Plano EPSEM	8.190	R$139.860,00	0,0030
Plano Não EPSEM	8.212	R$140.104,00	0,0051

Como discutido na Seção 2.7, a margem de erro foi estabelecida a partir do cálculo de $\bar{m}$, sendo esse valor fixado individualmente para cada UPA. A Tabela 4.3 apresenta os resultados obtidos por município, indicando o erro amostral associado a cada plano amostral. Embora nem todos os municípios selecionados sejam os mesmos, a comparação permanece válida, pois a variável foi simulada após a seleção da amostra. Caso tivesse sido gerada antes, seria necessário que as mesmas unidades fossem sorteadas para garantir uma comparação adequada.

Verifica-se que o plano EPSEM apresenta maior estabilidade entre os erros amostrais, com valores concentrados em torno do erro amostral desenhado de 0,025, definido para o cálculo do tamanho da amostra. Em contraste, o plano Não EPSEM evidencia maior dispersão, com erros amostrais variando amplamente entre os municípios, o que indica uma perda de controle sobre a precisão local das estimativas.

De modo geral, os resultados indicam que o plano EPSEM, além de apresentar menor custo total, proporcionou maior homogeneidade dos erros amostrais, característica desejável para garantir comparabilidade e robustez das estimativas. Por outro lado, o plano Não EPSEM, embora apresente custo e tamanho amostral ligeiramente superiores, resultou em erros amostrais mais elevados e heterogêneos.

Tabela 4.3: Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (AL)

Código	Erro EPSEM	Erro Não EPSEM
2700300	0.01751	0.01962
2701407	0.02365	0.05755
2701506	–	0.04842
2701605	–	0.03903
2701704	0.02242	–
2702108	0.03047	–
2702306	–	0.03386
2702355	0.02181	0.04398
2703106	–	0.04600
2703304	0.02455	–
2703403	0.02253	–
2703809	–	0.03724
2704005	–	0.05627
2704302	0.02615	0.00996
2704401	0.02165	–
2704609	0.02738	–
2705002	0.02301	–
2706307	0.02578	0.02170
2707305	0.02351	0.04521
2707701	0.02303	0.03059
2708402	–	0.03318

4.3 Estado do Rio de Janeiro

Selecionado aleatoriamente como estudo de caso, o estado do Rio de Janeiro trata-se da unidade federativa com a terceira maior população do país, concentrada em uma área territorial relativamente reduzida, resultando na maior densidade demográfica entre todos os estados analisados. Além disso, seu território inclui diversas ilhas, característica que enriquece a análise por acrescentar particularidades geográficas relevantes ao planejamento amostral.

A população total do estado é de **16.055.174** habitantes, distribuídos em **7.705.494** domicílios, organizados em **92** municípios (com proporção de municípios com até 50 mil habitantes de **40,22%**) e **40.519** (desconsiderando setores com população 0) setores censitários que compõem a base amostral disponível.

O dimensionamento amostral para o Rio de Janeiro resultou na seleção de **5** UPAs, **134** USAs e **12** UTAs, totalizando **8.040** entrevistas previstas. O custo total estimado, calculado conforme a Equação 2.30, foi de aproximadamente **R$133.440,00**.

4.3.1 Amostra Equiprόvavel

Da mesma forma que para o estado de Alagoas, foi necessário solucionar os casos de subdimensionamento das unidades no primeiro estágio. Identificaram-se 45 municípios com esse problema, os quais foram tratados por meio de um processo iterativo de agrupamento. Na primeira iteração, foram realizados 23 agrupamentos; na segunda, 11; e na terceira, 2, resultando em um total de **56** municípios. Dessa forma, a configuração final das unidades do primeiro estágio ficou estabelecida conforme ilustrado na Figura 4.5.

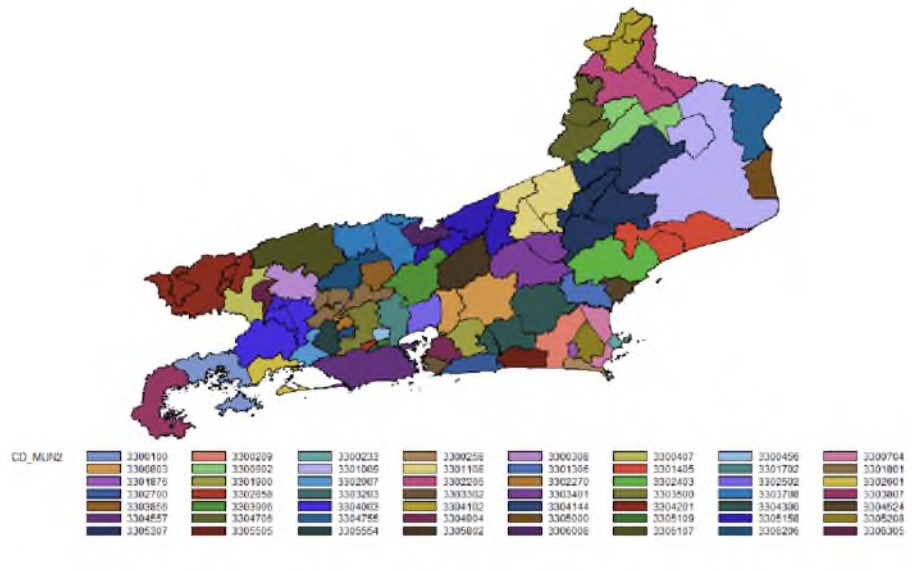


Figura 4.5: Mapa de municípios do estado do Rio de Janeiro após resolução de subdimensionamento

No segundo estágio, a solução foi aplicada apenas aos municípios previamente selecionados, conforme ilustrado na Figura 4.6. A amostra resultante não apresentou uma distribuição adequada ao longo do território, o que pode ser atribuído, em parte, ao reduzido tamanho amostral de apenas cinco municípios.

Dos 17.999 setores censitários pertencentes aos municípios selecionados, 177 foram identificados como subdimensionados. Ao todo, foram realizados 166 agrupamentos, resultando em 17.833 setores na configuração final, conforme detalhado na Tabela 4.4 e ilustrado na Figura 4.7, cuja visualização se apresenta ainda mais comprometida do que no caso de Alagoas.

No superdimensionamento, o primeiro estágio teve uma unidade municipal agrupada em um estrato certo, de modo a ser selecionada apenas uma vez. No segundo estágio, 10 setores censitários apresentaram o mesmo problema e foram devidamente corrigidos.

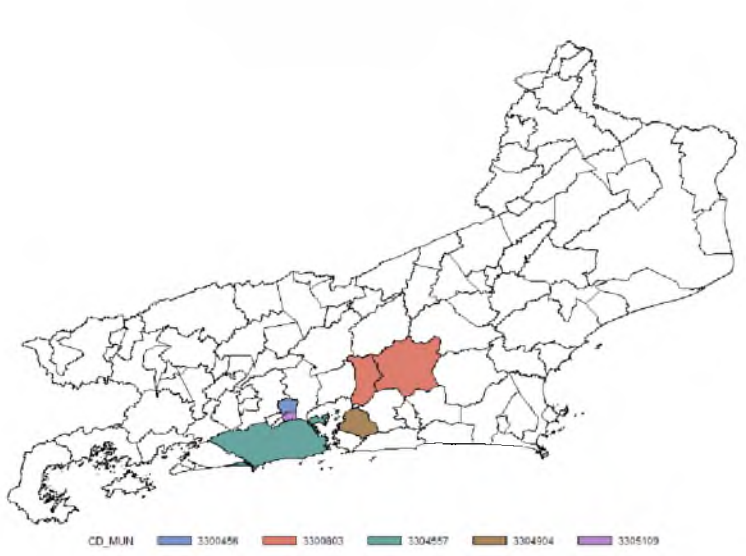


Figura 4.6: Mapa de municípios do estado do Rio de Janeiro selecionados para a amostra EPSEM

Tabela 4.4: Interações e agrupamentos por ordem de vizinhos de Rio de Janeiro

Interação	**Vizinhos de ordem**	**nº de agrupamentos**
1º	1º	140
2º	1º	19
3º	1º	5
4º	1º	2

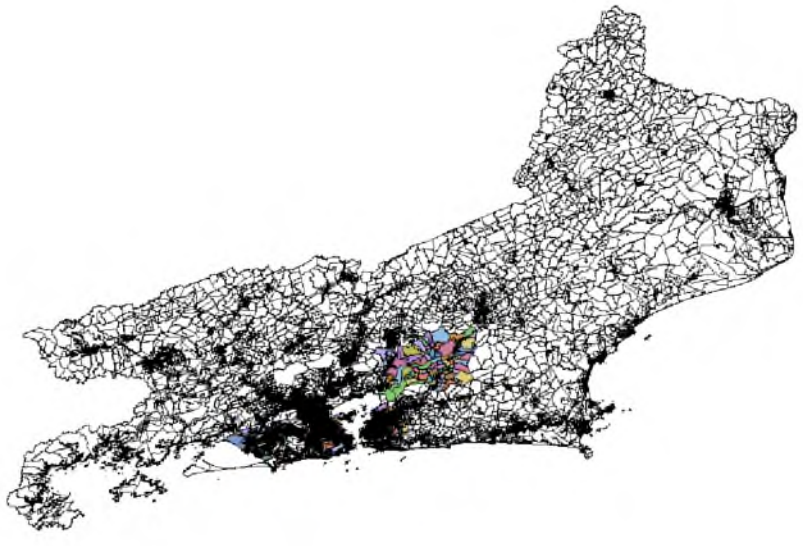


Figura 4.7: Mapa de setores censitários do estado do Rio de Janeiro após resolução de subdimensionamento

No primeiro estágio foram selecionados 5 municípios, e no segundo, 670 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural. No terceiro e último estágio, foram sorteados 8.040 domicílios, alcançando exatamente os

valores estipulados de tamanho e custo. Em cada etapa, confirmou-se que os pesos amostrais reproduziam o total populacional, indicando o correto seguimento do plano amostral, conforme a Tabela 4.13.

4.3.2 Amostra Não Equiprόvavel

Não houve subdimensionamento, e o superdimensionamento ocorreu apenas no primeiro estágio, afetando uma unidade municipal, que foi agrupada em um mesmo estrato para ser selecionada apenas uma vez. No segundo estágio, não foram identificadas unidades superdimensionadas.

No primeiro estágio, foram selecionados 5 municípios, conforme ilustrado na Figura 4.8. Em comparação, no delineamento amostral equiprovável, observa-se a inclusão de um município adicional na amostra em razão do subdimensionamento.

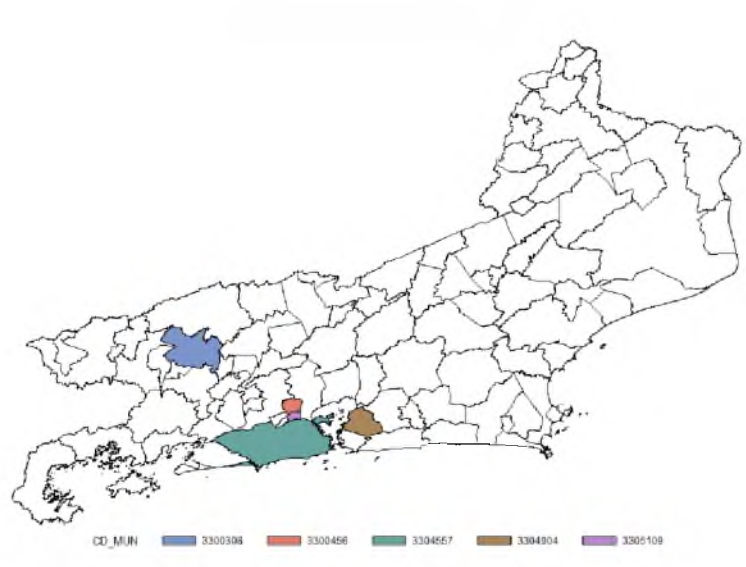


Figura 4.8: Mapa de municípios do estado do Rio de janeiro selecionados para a amostra Não EPSEM

Enquanto no segundo estágio foram selecionados 673 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural, no terceiro e último estágio foram sorteados 8.079 domicílios. Em cada etapa, confirmou-se que os pesos amostrais reproduziam o total populacional, indicando a correta estrutura do plano amostral, conforme a Tabela 4.13. O total de unidades selecionadas ficou acima do limite estipulado, resultando em um custo total de **R$133.866,00**, calculado conforme a Equação 2.30, mas ainda dentro do orçamento previsto.

A variável de interesse foi simulada conforme descrito no Capítulo 3. A Tabela

4.5 apresenta o resumo dos resultados obtidos para os dois planos avaliados. No Rio de Janeiro, os planos apresentaram tamanhos amostrais e custos bastante próximos, mas com diferenças relevantes quanto à precisão dos estimadores.

Tabela 4.5: Comparação entre os planos amostrais do Rio de Janeiro

Plano	**Tamanho total da amostra**	**Custo (R$)**	**Erro Padrão**
Plano EPSEM	8.040	R$133.440,00	0,0095
Plano Não EPSEM	8.079	R$133.866,00	0,0054

A Tabela 4.6 apresenta os erros amostrais por município. Observa-se que o plano EPSEM mantém erros próximos ao valor estabelecido, com pequena variação entre as unidades. Já o plano Não EPSEM apresenta dispersão mais elevada, incluindo municípios com erros substancialmente maiores, o que indica menor controle da precisão local.

Tabela 4.6: Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (RJ)

Código	**Erro EPSEM**	**Erro Não EPSEM**
3300456	0,02574	0,05118
3300803	0,02354	–
3304557	0,02463	0,01359
3304904	0,02683	0,03592
3305109	0,02474	0,05675
3300308	–	0,10066

De forma geral, apesar de o plano Não EPSEM apresentar erro padrão global menor, sua variabilidade municipal é sensivelmente maior. O plano EPSEM, por sua vez, apresenta maior regularidade dos erros e melhor aderência ao erro-alvo, configurando desempenho mais consistente para o estado.

4.4 Estado do Paraná

Selecionado aleatoriamente como estudo de caso para a aplicação dos planos amostrais, o estado do Paraná destaca-se por sua expressiva população e grande extensão territorial, características que o tornam um cenário relevante para testar a adaptabilidade dos desenhos amostrais em contextos de alta dispersão geográfica e densidade populacional variável.

A população total do estado é de **11.444.380** habitantes, distribuídos em **5.027.079** domicílios, organizados em **399** municípios (com proporção de municípios com até 50 mil

habitantes de **8,77%**) e **23.435** (desconsiderando setores com população 0) setores censitários que compõem a base amostral disponível.

O tamanho amostral para o Paraná resultou na seleção de **15** UPAs, **41** USAs e **13** UTAs, totalizando **7.995** entrevistas previstas. O custo total estimado, calculado conforme a Equação 2.30, foi de aproximadamente **R$138.330,00**.

4.4.1 Amostra Equiprόvavel

Foi necessário solucionar os casos de subdimensionamento das unidades no primeiro estágio. Identificaram-se 281 municípios com esse problema, os quais foram tratados por meio de um processo iterativo de agrupamento. Na primeira iteração, foram realizados 125 agrupamentos; na segunda, 57; na terceira, 21; na quarta, 2 e na quinta 1, resultando em um total de **193** municípios. Dessa forma, a configuração final das unidades do primeiro estágio ficou estabelecida conforme ilustrado na Figura 4.9.

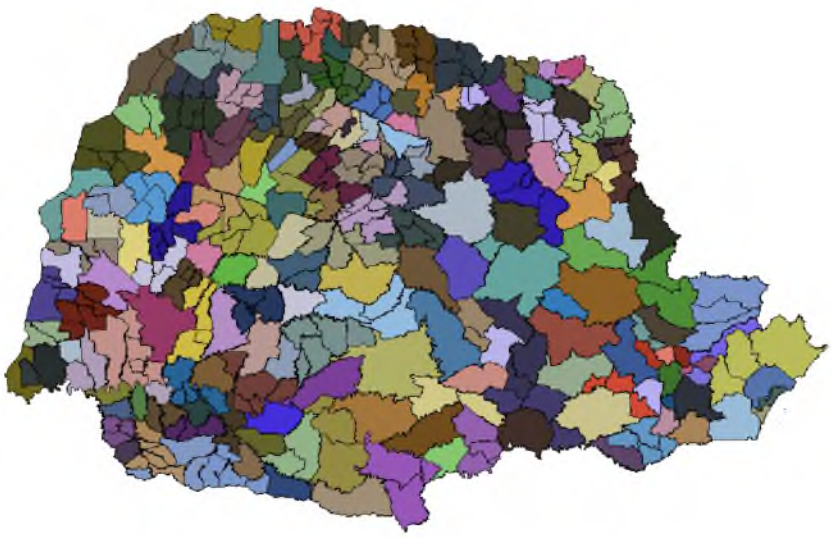


Figura 4.9: Mapa de municípios do estado do Paraná após resolução de subdimensionamento

No segundo estágio, a solução foi aplicada apenas aos municípios previamente selecionados, conforme ilustrado na Figura 4.10. A distribuição espacial dos municípios na amostra demonstra uma cobertura adequada, considerando a seleção de apenas 15 unidades entre 193 municípios após o processo de agrupamento.

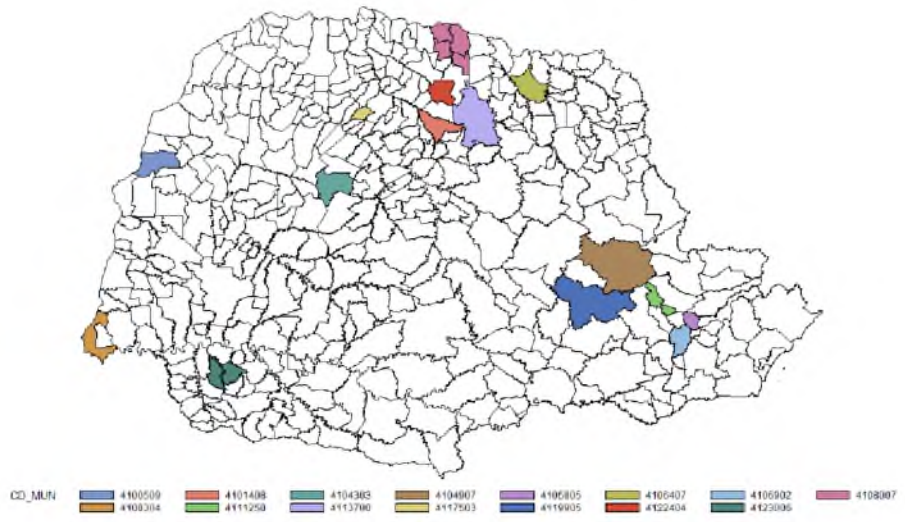


Figura 4.10: Mapa de municípios do estado do Paraná selecionados para a amostra EPSEM

Dos 6.815 setores censitários presentes nos municípios selecionados, foram identificadas 88 unidades subdimensionadas, tratadas por meio de um processo de agrupamento iterativo que resultou em 85 agrupamentos e em uma configuração final de 6.719 setores, conforme detalhado na Tabela 4.7 e na Figura 4.11.

Tabela 4.7: Interações e agrupamentos por ordem de vizinhos do Paraná

Interação	**Vizinhos de ordem**	**nº de agrupamentos**
1º	1º	75
2º	1º	5

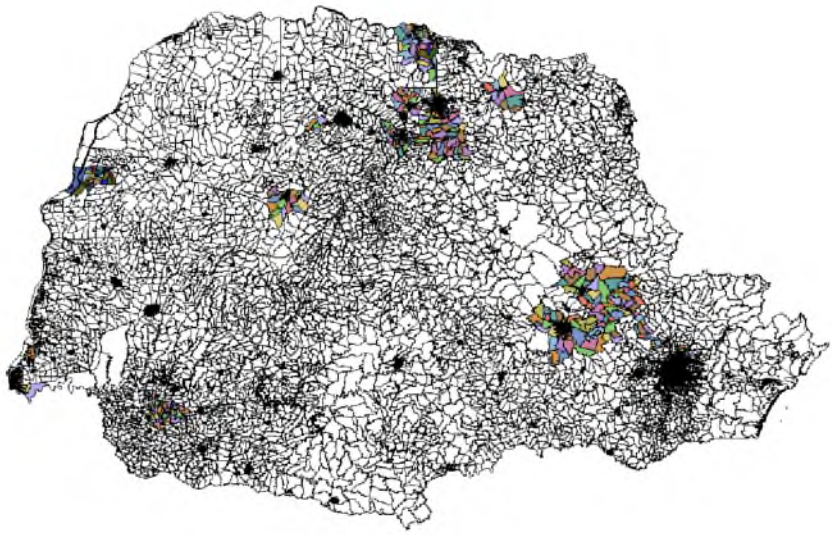


Figura 4.11: Mapa de setores censitários estado do Paraná após resolução de subdimensionamento

No estado do Paraná, houve apenas um caso de superdimensionamento. Assim, no primeiro estágio foram selecionados 15 municípios, e no segundo, 615 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural. No terceiro e último estágio, foram sorteados 7.995 domicílios, alcançando exatamente os valores estipulados de tamanho e custo. Em cada etapa, confirmou-se que os pesos amostrais reproduziam o total populacional, indicando o correto seguimento do plano amostral, de acordo com a Tabela 4.13.

4.4.2 Amostra Não Equipróvavel

A ocorrência de problemas foi reduzida: não houve subdimensionamento, e o superdimensionamento ocorreu apenas no primeiro estágio, afetando uma unidade municipal, que foi agrupada em um mesmo estrato para ser selecionada apenas uma vez. No segundo estágio, não foram identificadas unidades superdimensionadas.

No primeiro estágio, foram selecionados 15 municípios, conforme ilustrado na Figura 4.12. Já no delineamento amostral equiprovável, observa-se a inclusão de 5 municípios adicionais na amostra, quando comparado ao não equiprovável, em razão dos agrupamentos realizados.

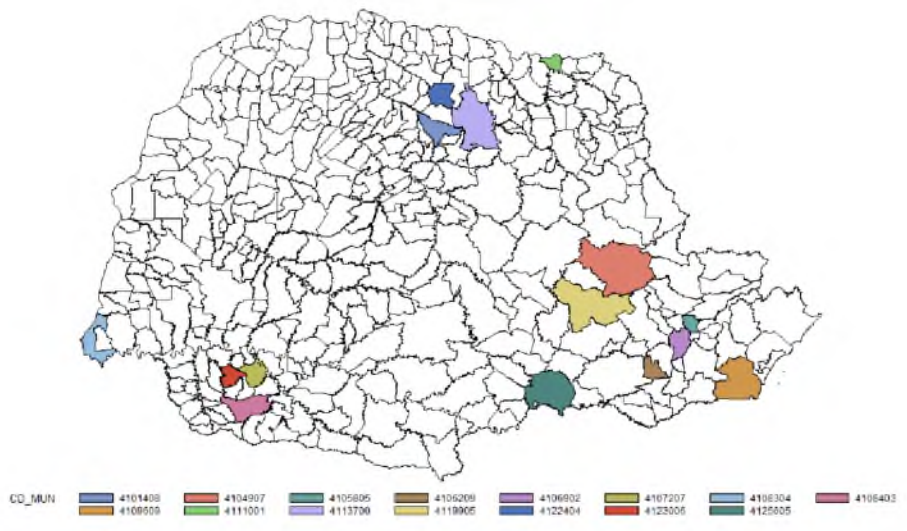


Figura 4.12: Mapa de municípios do estado do Paraná selecionados para a amostra Não EPSEM

Enquanto no segundo estágio foram selecionados 617 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural, no terceiro e último estágio foram sorteados 8.023 domicílios. Em cada etapa, confirmou-se que os pesos amostrais

reproduziam o total populacional, indicando o correto seguimento do plano amostral, conforme a Tabela 4.13. O total de unidades selecionadas ficou acima do limite estipulado, resultando em um custo total de **R$138.634,00**, calculado conforme a Equação 2.30, mas ainda dentro do orçamento previsto.

A variável de interesse foi simulada conforme descrito no Capítulo 3. A Tabela 4.8 apresenta o resumo dos resultados para os dois planos avaliados. No Paraná, os tamanhos amostrais e os custos novamente se mantêm próximos, com o plano EPSEM apresentando erro padrão inferior ao do Não EPSEM, mesmo com ligeira redução no tamanho da amostra.

Tabela 4.8: Comparação entre os planos amostrais do Paraná

Plano	**Tamanho total da amostra**	**Custo (R$)**	**Erro Padrão**
Plano EPSEM	7.995	R$138.330,00	0,0029
Plano Não EPSEM	8.023	R$138.634,00	0,0067

Tabela 4.9: Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (PR)

Código	**Erro EPSEM**	**Erro Não EPSEM**
4100509	0,02207	–
4101408	0,02612	0,03211
4104303	0,02435	–
4104907	0,02620	0,03918
4105805	0,02052	0,02967
4106407	0,02330	–
4106902	0,02288	0,00975
4108007	0,02488	–
4108304	0,02330	0,02521
4108403	–	0,03522
4109609	–	0,05390
4111001	–	0,08216
4111258	0,02460	–
4113700	0,02581	0,02074
4117503	0,03209	–
4119905	0,02423	0,01797
4122404	0,02397	0,05951
4123006	0,02969	0,12008
4125605	–	0,03890

A Tabela 4.9 apresenta os erros amostrais por município. Observa-se que o plano EPSEM mantêm valores próximos ao erro amostral especificado, com variação moderada entre as unidades. Já o plano Não EPSEM apresenta erros amostrais mais diferentes,

incluindo municípios com valores significativamente elevados, o que compromete a regularidade da precisão local.

De forma geral, o plano EPSEM demonstra desempenho mais consistente, apresentando menor erro padrão global e maior uniformidade entre os municípios. Já o plano Não EPSEM, embora tenha tamanho amostral ligeiramente superior, resulta em erros mais elevados e heterogêneos, o que reduz sua eficiência estatística no estado.

4.5 Estado do Acre

Selecionado aleatoriamente como estudo de caso, o estado do Acre caracteriza-se por possuir a 9ª maior área do Brasil em contraste com sua 25ª posição em população, uma combinação que gera elevada heterogeneidade populacional entre municípios e setores censitários e demanda ajustes específicos no planejamento amostral.

Diferentemente dos casos anteriores, para o Acre foi adotado um erro amostral de 2%, com o objetivo de aumentar o número de setores e reduzir o número de municípios selecionados, em virtude da quantidade reduzida de municípios existentes no estado. Adicionalmente, o desenho equiprovável considerou todos os setores censitários, inclusive aqueles com população igual a zero, enquanto o desenho não equiprovável excluiu esses setores da base amostral, pois estavam ocorrendo problemas de estimação.

A população total do estado é de **830.018** habitantes, distribuídos em **319.292** domicílios, organizados em **22** municípios (com proporção de municípios com até 50 mil habitantes de **9,09%**) e **2.215** setores censitários, ou **2.127** quando desconsiderados os setores com população igual a zero, que constituem a base amostral utilizada.

Para o tamanho amostral para o estado do Acre, no desenho equiprovável, resultou na seleção de **9** UPAs, **80** USAs e **11** UTAs, totalizando **7.920** entrevistas esperadas, com custo total estimado de **R$ 135.240,00**.

Por sua vez, o desenho não equiprovável resultou na seleção de **10** UPAs, **79** USAs e **11** UTAs, totalizando **8.690** entrevistas esperadas, com custo total estimado de **R$ 144.380,00**.

4.5.1 Amostra Equiprovável

Para corrigir o subdimensionamento das unidades do primeiro estágio, identificaram-se 7 municípios problemáticos, tratados por um processo iterativo de agrupamento: 8 agrupamentos na primeira iteração e 2 na segunda, totalizando **12** municípios. A confi-

guração final dessas unidades está apresentada na Figura 4.13.

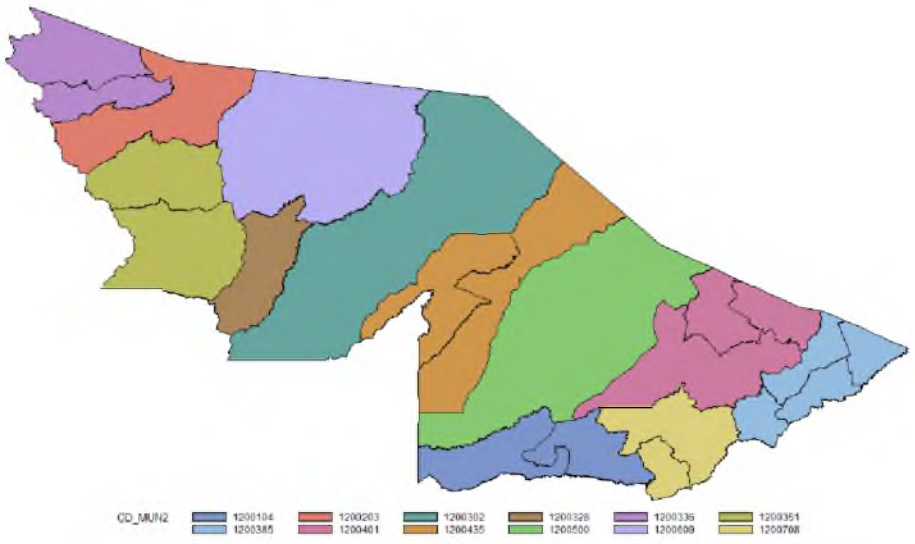


Figura 4.13: Mapa de municípios do estado do Acre após resolução de subdimensionamento

No segundo estágio, a solução foi aplicada apenas aos municípios previamente selecionados, conforme ilustrado na Figura 4.14. A distribuição espacial resultante demonstra uma cobertura adequada, considerando que 9 dos 12 municípios restantes após o agrupamento foram incluídos na amostra. Este resultado exemplifica a razão pela qual a alteração do erro amostral, mencionada anteriormente, foi necessária para selecionar um número menor de UPAs.

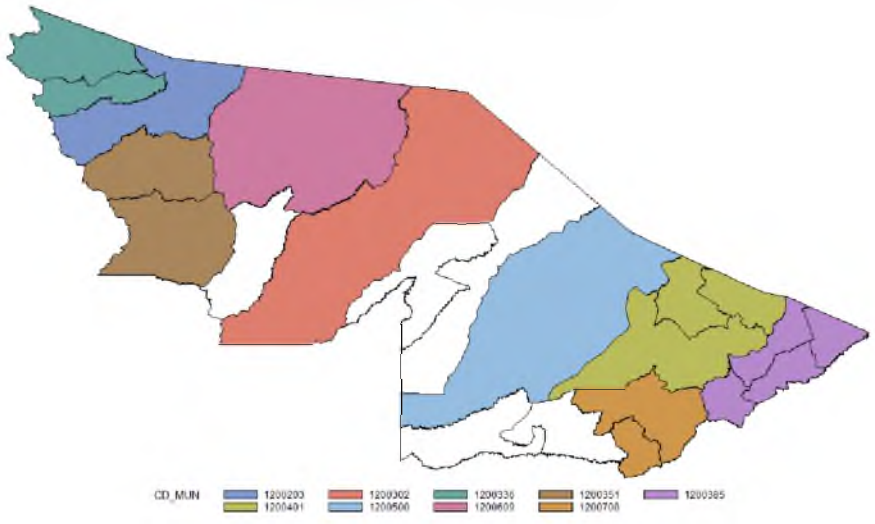


Figura 4.14: Mapa de municípios do estado do Acre selecionados para a amostra EPSEM

Identificaram-se 232 setores censitários subdimensionados dentre os 1.899 existentes, os quais foram tratados por meio de um processo iterativo de agrupamento, conforme detalhado na Tabela 4.10 e na Figura 4.15. Após os agrupamentos realizados, obteve-se um total de 1.684 setores censitários, como ilustrado na figura.

Tabela 4.10: Interações e agrupamentos por ordem de vizinhos do Acre

Interação	**Vizinhos de ordem**	**nº de agrupamentos**
1	1º	143
2	1º	37
3	1º	11
4	1º	3
5	2º	6
6	2º	5
7	2º	3
8	2º	3
9	2º	1
10	2º	2
11	2º	1

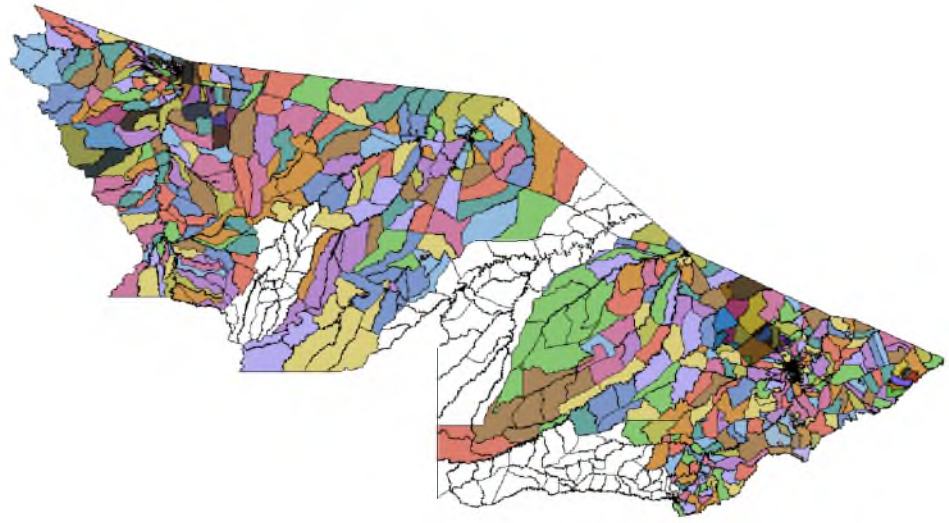


Figura 4.15: Mapa de setores censitários estado do Acre após resolução de subdimensionamento

Já no caso de superdimensionamento, no primeiro estágio, 3 unidades municipais foram agrupadas em um estrato certo, de modo a serem selecionadas apenas uma vez. No segundo estágio, 312 setores censitários apresentaram o mesmo problema e foram devidamente corrigidos.

No primeiro estágio foram selecionados 9 municípios, e no segundo, 720 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural. No

terceiro e último estágio, foram sorteados 7.920 domicílios, alcançando exatamente os valores estipulados de tamanho e custo. Em cada etapa, confirmou-se que os pesos amostrais reproduziam o total populacional, indicando o correto seguimento do plano amostral, conforme a Tabela 4.13.

4.5.2 Amostra Não Equiprovável

A ocorrência de problemas foi mínima: não houve subdimensionamento e o superdimensionamento limitou-se ao primeiro estágio, afetando apenas 2 unidades municipais, agrupadas em um mesmo estrato para seleção única. No segundo estágio, 137 setores censitários foram corretamente agrupados em seus respectivos estratos.

No primeiro estágio, foram selecionados 10 municípios, conforme apresentado na Figura 4.16. No delineamento amostral equiprovável, observa-se a inclusão de 7 municípios adicionais, em razão dos agrupamentos realizados, quando comparado ao não equiprovável, mesmo tendo selecionado originalmente 9 municípios.

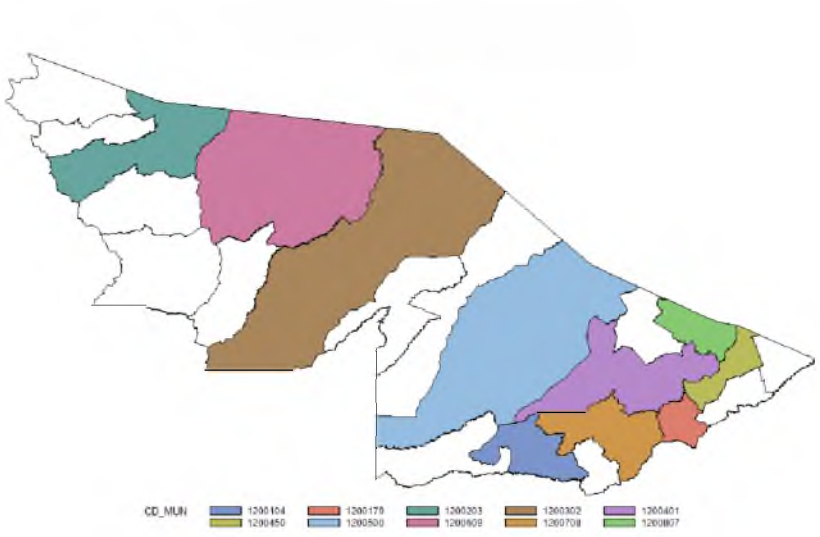


Figura 4.16: Mapa de municípios do estado do Acre selecionados para a amostra Não EPSEM

Enquanto no segundo estágio foram selecionados 790 setores censitários, estratificados com alocação proporcional entre as áreas urbana e rural, no terceiro e último estágio foram sorteados 8.697 domicílios. Em cada etapa, confirmou-se que os pesos amostrais reproduziam o total populacional, indicando a correta estrutura do plano amostral, conforme a Tabela 4.13. O total de unidades selecionadas ultrapassou o limite estipulado, resultando em um custo total de **R$144.450,00**, calculado conforme a Equação 2.30, o que excede o orçamento previsto.

A variável de interesse foi simulada conforme descrito no Capítulo 3. A Tabela 4.11 apresenta o resumo dos resultados obtidos para o Acre. Neste estado, o plano EPSEM apresentou tamanho amostral e custo menores, além de erro padrão substancialmente inferior ao do plano Não EPSEM, evidenciando melhor desempenho global.

Tabela 4.11: Comparação entre os planos amostrais do Acre

Plano	**Tamanho total da amostra**	**Custo (R$)**	**Erro Padrão**
Plano EPSEM	7.920	R$135.240,00	0,0015
Plano Não EPSEM	8.697	R$144.450,00	0,0046

A Tabela 4.12 apresenta os erros amostrais por município. Nota-se que o plano EPSEM mantém erros amostrais mais concentrados e próximos ao valor especificado, indicando maior estabilidade. Em contraste, o plano Não EPSEM apresenta variação mais ampla, com municípios exibindo erros elevados, o que reduz o controle da precisão local.

Tabela 4.12: Comparação dos erros amostrais por município entre os métodos EPSEM e Não EPSEM (AC)

Código	**Erro EPSEM**	**Erro Não EPSEM**
1200203	0,02129	0,02074
1200302	0,02059	0,01896
1200336	0,02072	–
1200351	0,02935	–
1200385	0,01970	–
1200401	0,01457	0,00978
1200500	0,02315	0,02539
1200609	0,02448	0,02406
1200708	0,02331	0,02511
1200104	–	0,03167
1200179	–	0,03321
1200450	–	0,03627
1200807	–	0,04024

De forma geral, o plano EPSEM apresentou desempenho superior, com menor custo, menor erro padrão e maior uniformidade dos erros amostrais entre os municípios. Já o plano Não EPSEM, apesar de incluir mais unidades, resultou em custos mais altos e maior heterogeneidade dos erros amostrais, reduzindo sua eficiência estatística e econômica.

4.6 Resumo Geral

A Tabela 4.13 apresenta a verificação da soma do produto entre o peso e a população em cada estágio de seleção, considerando todos os planos amostrais e as unidades analisadas. Esse procedimento assegura que os pesos foram corretamente especificados e, consequentemente, que o desenho amostral foi executado de forma adequada. É essencial que essa verificação seja realizada estágio a estágio, pois o peso final (terceiro estágio) resulta do produto dos pesos obtidos nas etapas anteriores, com estratificação no segundo estágio.

Tabela 4.13: Soma dos pesos amostrais por Unidade da Federação e plano amostral

UF	**Plano**	**1º estágio**	**2º estágio**	**3º estágio**
AL	Plano EPSEM	3.127.683 (14)	3.127.683 (630)	3.127.683 (8.190)
	Plano Não EPSEM	3.127.683 (14)	3.127.683 (632)	3.127.683 (8.212)
RJ	Plano EPSEM	16.055.174 (5)	16.055.174 (670)	16.055.174 (8.040)
	Plano Não EPSEM	16.055.174 (5)	16.055.174 (673)	16.055.174 (8.079)
PR	Plano EPSEM	11.444.380 (15)	11.444.380 (615)	11.444.380 (7.995)
	Plano Não EPSEM	11.444.380 (15)	11.444.380 (617)	11.444.380 (8.023)
AC	Plano EPSEM	830.018 (9)	830.018 (720)	830.018 (7.920)
	Plano Não EPSEM	830.018 (10)	830.018 (790)	830.018 (8.697)

Nota: Os valores entre parênteses se referem à quantidade de elementos amostrados.

No caso de Alagoas, no primeiro estágio da amostragem, foram selecionados 14 municípios representativos dos demais 88, com base em seus respectivos pesos. Ao multiplicar o total de pessoas (*V0001*) de cada município selecionado pelo seu peso e somar os resultados, obtém-se exatamente 3.127.683 pessoas, valor correspondente à população total do estado. No segundo estágio, adotando um desenho equiprovável, foram selecionados 45 setores censitários em cada um desses municípios, totalizando 630 setores. O peso dessa seleção é determinado pelo total de pessoas dos municípios escolhidos no primeiro estágio e, para manter a representatividade, é multiplicado pelo peso do primeiro estágio, de modo que a população de cada setor seja ponderada pelo peso combinado dos dois estágios. A soma desses produtos também resulta em 3.127.683 pessoas, o que confirma a consistência do processo. O mesmo princípio é aplicado no terceiro estágio, quando são considerados os domicílios, assegurando que a representatividade seja preservada em todas as etapas da seleção.

No caso não equiprovável, as mesmas verificações são realizadas. Entretanto, como os tamanhos das USAs e UTAs não são constantes, as probabilidades de seleção variam entre as unidades, resultando em pesos distintos dentro dos mesmos estágios,

calculados como o inverso das respectivas probabilidades de seleção, conforme detalhado na Seção 2.4.

Outro ponto que merece atenção é que, durante a estratificação, foram necessárias algumas correções manuais. Como a alocação é proporcional entre os setores rurais e urbanos, em alguns casos os setores rurais são muito pequenos, fazendo com que a proporção de setores censitários urbanos ultrapasse 0,99. Nesses casos, devido aos arredondamentos, o número total de setores a selecionar ($m_i = 45$) poderia ser excedido: o valor de m_{ih} para o estrato urbano seria arredondado para 45, enquanto o do rural, mesmo muito pequeno, seria arredondado para 1. Assim, para o mesmo município, seriam selecionados 46 setores em vez dos 45 previstos originalmente. Sendo assim, os resultados foram verificados manualmente e ajustados quando necessário, de modo a manter o número total de setores selecionados conforme o município.

Capítulo 5

Conclusão

Este trabalho tinha o objetivo de comparar os planos amostrais complexos equiprováveis e não equiprováveis, analisando suas características em diferentes contextos estaduais. A análise conduzida nos estados de Alagoas, Rio de Janeiro, Paraná e Acre permitiu obter conclusões sobre o desempenho dos dois métodos.

Os resultados mostraram que o plano EPSEM apresentou melhor controle do orçamento, mantendo-se próximo dos custos previstos, enquanto o plano não equiprovável superou o orçamento estabelecido, com diferença mais significativa no Acre, onde excedeu em R$ 9.210,00. É importante destacar que o plano não EPSEM apresentou tamanho amostral e custos ligeiramente maiores devido a questões de arredondamento no cálculo dos tamanhos amostrais (m_i e l_{ij}). Essa situação assemelha-se ao ocorrido na alocação proporcional durante a estratificação, onde arredondamentos levaram à seleção de unidades adicionais. Contudo, diferentemente do processo de estratificação, onde ajustes manuais foram viáveis para um número limitado de casos, a verificação e correção manual mostraram-se inviáveis para as mais de 500 USAs, o que contribuiu para o excesso unidades amostradas.

Em termos de desempenho estatístico, o plano EPSEM apresentou erros padrão menores em três dos quatro estados analisados e maior homogeneidade dos erros amostrais municipais, com valores mais próximos ao erro amostrais definidos.

Na implementação, as macros SAS desenvolvidas para tratamento de subdimensionamento e superdimensionamento mostraram-se funcionais. A macro de agrupamento iterativo por ordem de vizinhança resolveu os casos de subdimensionamento, enquanto a macro de estratificação para unidades superdimensionadas permitiu o agrupamento em estratos certos. O plano não equiprovável demandou menos correções, porém com menor controle sobre o tamanho amostral final.

Os resultados obtidos indicam que o plano EPSEM apresentou desempenho superior em termos de precisão das estimativas. Entretanto, conforme discutido ao longo da análise, esse delineamento mostrou-se mais suscetível a ocorrências de problemas operacionais, exigindo o desenvolvimento de macros específicas para contorná-los. Por outro lado, o plano Não EPSEM apresentou menor frequência de inconsistências, revelando-se mais simples de implementar, ainda que resulte em estimativas menos precisas.

Como limitações do estudo, destacam-se os problemas de arredondamento identificados tanto no processo de estratificação quanto no cálculo dos tamanhos amostrais m_i e l_{ij}, que impactaram principalmente o plano não EPSEM. Adicionalmente, a abordagem inconsistente em relação aos setores com população zero, incluídos no desenho equiprovável e excluídos no não equiprovável no caso do Acre, representa uma limitação metodológica que pode ter influenciado os resultados comparativos. Para trabalhos futuros, recomenda-se desenvolver técnicas específicas para a correção sistemática dos arredondamentos que afetaram o dimensionamento amostral e estabelecer um protocolo unificado para o tratamento de setores com população zero, de modo a garantir comparabilidade entre os diferentes planos amostrais.

Referências Bibliográficas

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons.

IBGE (2014). Instituto brasileiro de geografia e estatística. notas metodológicas da pesquisa nacional por amostra de domicílios contínua. Disponível em: `https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf`. Acesso em: 30 mar. 2025.

Kish, L. (1965). *Survey Sampling*. Wiley.

Kish, L. (1992). Weighting for unequal P_i. *Journal of Official Statistics*, 8(2):183–200.

Lohr, S. L. (2010). *Sampling: design and analysis*, (second ed.). Cengage Learning.

Silva, A. R. (2025). *Amostragem: Teoria e Aplicações*. Clube de Autores.

Skinner, C. & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, 32(2):165–175.

Apêndice A

Macro SAS para subdimensionamento

```
%macro undersized(data=,ID=,size=,tot_size=,undersized=,pop=,neighbor=,shape=,nneigh=3,order=)
proc sort data=&data;by &tot_size;run;
proc iml;
use &data;
read all var{&id &undersized};
use &neighbor;
read all var{&id &id.1} into NEIGH;
n=nrow(&undersized);
&id.2=&id||j(n,1,'0');
check3=0;
print "Iteration 1";
do i=1 to n;
if &undersized[i]=1 then do;
j=1;
viz=NEIGH[loc(NEIGH[,1]=&id[i]),2];
viz=viz[j,];
*print viz;
check=&id.2[loc(&id=viz),2];
if i>1 then check3=ncol(loc(&id.2[1:i-1,1]=&id[i]));
if check='0' & check3=0 then do;&id.2[i,1]=viz;&id.2[i,2]='1';
print (&id[i]) "joined with" viz[label=''];
end;
*check=ncol(loc(CD_MUN2=viz));
*if check>0 then CD_MUN2[i]=viz;
end;
end;
```

```
create _under_ from &id.2[colname={"&id.2" "join"}];
append from &id.2;
quit;
proc sql noprint;select mean(input(join,1.)) into:mjoin from _under_;quit;
%put &mjoin;

data &data.j;merge &data _under_;
if &id ne &id.2 then &id.c=trim(left(&id))||","||trim(left(&id.2));
else &id.c=&id;
&id.f=&id.c;
run;
/*proc freq data=&data.j noprint;
tables &id.2 /out=&data.2(rename=(count=&tot_size) drop=percent);
weight &tot_size;
run;*/
proc means data=&data.j nway noprint;
class &id.2;
var &tot_size;
output out=&data.2(drop=_type_ _freq_) sum=&tot_size;
run;
/*proc freq data=&data.j noprint;
tables &id.2 /out=&data.2p(rename=(count=&pop) drop=percent);
weight &pop;
run;*/
proc means data=&data.j nway noprint;
class &id.2;
var &pop;
output out=&data.2p(drop=_type_ _freq_) sum=&pop;
run;
data &data.2;merge &data.2 &data.2p;
_size_=&size;
if &tot_size<_size_ then &undersized=1;else &undersized=0;
run;
proc sql noprint;
select count(*) into:n_under from &data.2
where &undersized=1;
quit;
%put &n_under;

%let k=1;
%do %while(&n_under>0 and &k<=30);
```

```
proc sort data=&data.2;by &order &tot_size;run;
proc iml;
use &data.2;
read all var{&id.2 &undersized};
use &neighbor;
read all var{&id &id.1} into NEIGH;
n=nrow(&undersized);
&id.2=&id.2||j(n,1,'0');
check3=0;
print "Iteration %eval(&k+1)";
do i=1 to n;
if &undersized[i]=1 then do;
j=1;
g=10;
check=0;
do while(check=0 & j<=g);
viz=NEIGH[loc(NEIGH[,1]=&id.2[i,1]),2];
g=nrow(viz);
viz=viz[j,];
*print i viz;
check=ncol(loc(&id.2[,1]=viz));
if check>0 then do;
check2=&id.2[loc(&id.2[,1]=viz),2];
if i>1 then check3=ncol(loc(&id.2[1:i-1,1]=&id.2[i,1]));
if check2='0' & check3=0 then do;value_old=&id.2[i,1];&id.2[i,1]=viz;&id.2[i,2]='1';
print value_old[label=''] "joined with" viz[label=''];
end;
end;
j=j+1;
end;
end;
end;
create _under_ from &id.2[colname={"&id.2" "join"}];
append from &id.2;
quit;
proc sql noprint;select mean(input(join,1.)) into:mjoin from _under_;quit;
%put &mjoin;

%if &mjoin=0 and &n_under>&nneigh %then %do;
proc print data=&data.2 noobs;
var &id.2 &tot_size _size_ &undersized;
```

```
where &undersized=1;
run;

%put WARNING: **** Using Neighbors of Second Order ****;
proc sort data=&data.2;by &order &tot_size;run;
proc iml;
use &data.2;
read all var{&id.2 &undersized};
use &neighbor;
read all var{&id &id.1} into NEIGH;
n=nrow(&undersized);
&id.2=&id.2||j(n,1,'0');
check3=0;
print "Using Neighbors of Second Order";
do i=1 to n;
if &undersized[i]=1 then do;
j=1;
g=10;
check=0;
do while(check=0 & j<=g);
viz=NEIGH[loc(NEIGH[,1]=&id.2[i,1]),2];
g=nrow(viz);
viz=viz[j,];
l=1;
h=10;
check2='0';
do while(check=0 & l<=h);
viz2=NEIGH[loc(NEIGH[,1]=viz),2];
h=nrow(viz2);
viz2=viz2[l,];
*print i l h;
check=ncol(loc(&id.2[,1]=viz2));
if check>0 then do;
check2=&id.2[loc(&id.2[,1]=viz2),2];
if i>1 then check3=ncol(loc(&id.2[1:i-1,1]=&id.2[i,1]));
if check2='0' & check3=0 & &id.2[i,1]^=viz2 then do;value_old=&id.2[i,1];&id.2[i,1]=viz2;
print value_old[label=''] "joined with" viz2[label=''];
end;
end;
l=l+1;
end;
```

```
j=j+1;
end;
end;
end;
create _under_ from &id.2[colname={"&id.2" "join"}];
append from &id.2;
quit;
proc sql noprint;select mean(input(join,1.)) into:mjoin from _under_;quit;
%put &mjoin;

%end;

%if &mjoin=0 and &n_under<=&nneigh %then %do;
proc print data=&data.2 noobs;
var &id.2 &tot_size _size_ &undersized;
where &undersized=1;
run;
%put WARNING: **** Using Neighbors of Third Order ****;

proc sort data=&neighbor out=&neighbor.2;by &id descending &tot_size;run;
proc iml;
use &data.2;
read all var{&id.2 &undersized};
use &neighbor.2;
read all var{&id &id.1} into NEIGH;
n=nrow(&undersized);
&id.2=&id.2||j(n,1,'0');
check3=0;
print "Using Neighbors of Second Order";
do i=1 to n;
if &undersized[i]=1 then do;
j=1;
g=10;
check=0;
do while(check=0 & j<=g);
viz=NEIGH[loc(NEIGH[,1]=&id.2[i,1]),2];
g=nrow(viz);
viz=viz[j,];
l=1;
h=10;
check2='0';
```

```
do while(check=0 & l<=h);
viz2=NEIGH[loc(NEIGH[,1]=viz),2];
h=nrow(viz2);
viz2=viz2[l,];
*print i l h;
check=ncol(loc(&id.2[,1]=viz2));
if check>0 then do;
check2=&id.2[loc(&id.2[,1]=viz2),2];
if i>1 then check3=ncol(loc(&id.2[1:i-1,1]=&id.2[i,1]));
if check2='0' & check3=0 & &id.2[i,1]^=viz2 then do;value_old=&id.2[i,1];&id.2[i,1]=viz2;
print value_old[label=''] "joined with" viz2[label=''];
end;
end;
l=l+1;
end;
j=j+1;
end;
end;
end;
create _under_ from &id.2[colname={"&id.2" "join"}];
append from &id.2;
quit;
proc sql noprint;select mean(input(join,1.)) into:mjoin from _under_;quit;
%put &mjoin;

%end;

data &data.2;merge &data.2(keep=&id.2 &tot_size &pop rename=&id.2=&id) _under_;
if &id ne &id.2 then &id.c%eval(1+&k)=trim(left(&id))||","||trim(left(&id.2));
else &id.c%eval(1+&k)=&id;
run;
proc sort data=&data.2;by &id;run;
proc sort data=&data.j;by &id.%eval(1+&k);run;
data &data.j;merge &data.j &data.2(keep=&id &id.2 &id.c%eval(1+&k)
rename=(&id.2=&id.%eval(2+&k) &id=&id.%eval(1+&k)));
by &id.%eval(1+&k);
&id.f=CATX(",",&id.f,&id.c%eval(1+&k));
run;
/*proc freq data=&data.2 noprint;
tables &id.2 /out=&data.2p(rename=(count=&pop) drop=percent);
weight &pop;
```

```
run;*/
proc means data=&data.2 nway noprint;
class &id.2;
var &pop;
output out=&data.2p(drop=_type_ _freq_) sum=&pop;
run;
/*proc freq data=&data.2 noprint;
tables &id.2 /out=&data.2(rename=(count=&tot_size) drop=percent);
weight &tot_size;
run;*/
proc means data=&data.2 nway noprint;
class &id.2;
var &tot_size;
output out=&data.2(drop=_type_ _freq_) sum=&tot_size;
run;
data &data.2;merge &data.2 &data.2p;
_size_=&size;
if &tot_size<_size_ then &undersized=1;else &undersized=0;
run;
proc sql noprint;
select count(*) into:n_under from &data.2
where &undersized=1;
quit;
%put &n_under;
%let k=%eval(&k+1);
%end;

proc sql;select sum(&pop) as M0 from &data.2;quit;
%put &M0;
data &data.j;set &data.j;
*&id.f=CATX(",", of &id.c:);
joins=length(&id.f)-length(compress(&id.f,','))+1;
run;
%macro j;
proc sql noprint;select max(joins) into:max from &data.j;quit;
data &data.j;set &data.j;
%do i=1 %to &max;
cod&i=scan(&id.f,&i,',');
%end;
*if joins<&max then &id%left(%eval(&max-1))=scan(&id.f,joins,',');
run;
```

```
%mend j;
%j;
proc sort data=&data.j;by &id;run;
proc transpose data=&data.j out=&data.jt(drop=_name_);
by &id;var cod:;
run;
data &data.jt;set &data.jt;if col1 ne '';run;
data &data.jt;merge &data.jt &data.j(keep=&id &id%eval(1+&k));
by &id;
if &id%eval(1+&k) ne '';
run;
proc sort data=&data.jt nodupkey;by &id%eval(1+&k) col1;run;

proc freq data=&data.jt noprint;
tables &id%eval(1+&k) /out=_freq_mult_(drop=percent rename=count=freq);
run;
data &data.jt2;merge &data.jt _freq_mult_;by &id%eval(1+&k);
if freq>1;
run;
proc sql noprint;
select count(DISTINCT &id%eval(1+&k)) into :cnt_cod from &data.jt2;
select DISTINCT &id%eval(1+&k) into :cod1 - :cod%left(&cnt_cod)
from &data.jt2;
quit;
%put &cnt_cod &cod1;

%if &shape ne %then %do;
%macro jt;
data &shape.2;set &shape.;run;
proc sql;drop table _xj_;run;
%do i=1 %to &cnt_cod;
proc sql noprint;
select col1 into:group separated by "' '" from &data.jt
where &id%eval(1+&k)="&&cod&i";
quit;
/*%put &group;*/
data _x_;length var $5000.;var="'"||"&group"||"'";value="&&cod&i";
call symput("group&i",var);run;
proc append base=_xj_ data=_x_ force;run;
%put &&cod&i &&group&i;
%end;
```

```
proc sql noprint;select max(length(var)) into:lv from _xj_;quit;
%put &lv;
data _xj_;length var $&lv..;set _xj_;run;
data &shape.2;set &shape.2;
%do i=1 %to &cnt_cod;
if &id in (&&group&i) then &id.2="&&cod&i";
%end;
if &id.2='' then &id.2=&id;
run;
%mend jt;
%jt;

goptions reset=all reset=global;
proc gmap data=&data.2 map=&shape.2 all;
id &id.2;
choro &id.2;
run;
quit;
proc gmap data=&data.2 map=&shape.2 all;
id &id.2;
choro &id.2 /nolegend;
run;
quit;
%end;

%mend undersized;
```

Apêndice B

Macro SAS para superdimensionamento

```
%macro oversized(data=,size=,tot_size=,oversized=,strata=,M0=);
data &data;set &data;
M0=&M0;
a=&size;
if &tot_size>M0/a then &oversized.1=1;else &oversized.1=2;
weight_PSU=M0/(&size*&tot_size);
run;

%let k=2;
%let prob1=1;
%do %while(&prob1>0);
proc means data=&data noprint nway;
var &tot_size;
class &strata &oversized%eval(&k-1);
output out=_sum_pop_str_(drop=_type_ rename=_freq_=_freq_str) sum=_pop_str_;
run;

%if &strata= %then %do;
proc sql noprint;
select _pop_str_ into:M02 from _sum_pop_str_
where &oversized%eval(&k-1)=2;
select _freq_str into:Str1 from _sum_pop_str_
where &oversized%eval(&k-1)=1;
quit;
%put &M02 &Str1;
```

```
data &data;set &data;
if &oversized%eval(&k-1)=1 then weight_PSU=1;
if &oversized%eval(&k-1)=2 then weight_PSU=&M02/((&size-&Str1)*&tot_size);
if &tot_size>(&M02/(&size-&Str1)) then &oversized.&k=1;else &oversized.&k=2;
if &oversized%eval(&k-1)=1 then &oversized.&k=1;
if &oversized.&k=1 then _ah_=&Str1;
if &oversized.&k=2 then _ah_=&size-&Str1;
run;
%end;

%else %do;
data _sum_pop_str1_;set _sum_pop_str_(where=(&oversized%eval(&k-1)=1)
rename=_freq_str=_freq_str1 drop=_pop_str_);
run;
%if &k=2 %then %do;
proc sort data=&data;by &strata &oversized%eval(&k-1);run;
%end;
%else %do;
proc sort data=&data(drop=_freq_str1 _ah_ _pop_str_);by &strata &oversized%eval(&k-1);run;
%end;
data &data;merge &data _sum_pop_str1_;
by &strata;
if _freq_str1=. then _freq_str1=0;
run;
data &data;merge &data _sum_pop_str_;
by &strata &oversized%eval(&k-1);
if &oversized%eval(&k-1)=1 then weight_PSU=1;
if &oversized%eval(&k-1)=2 then weight_PSU=_pop_str_/((&size-_freq_str1)*&tot_size);
if &tot_size>(_pop_str_/(&size-_freq_str1)) then &oversized.&k=1;else &oversized.&k=2;
if &oversized%eval(&k-1)=1 then &oversized.&k=1;
if &oversized.&k=1 then _ah_=_freq_str1;
if &oversized.&k=2 then _ah_=&size-_freq_str1;
run;
%end;


proc sql noprint;
select count(*) into:prob1 from &data
where weight_PSU<1;
quit;
%put &prob1;
```

```
%let k=%eval(&k+1);
%end;
data &data;set &data;
&oversized=&oversized%eval(&k-1);
run;

%mend oversized;
```