DOI http://dx.doi.org/10.20947/S0102-3098a0306

ARTIGO ORIGINAL



# Desenvolvendo uma população brasileira sintética derivada do Censo Demográfico de 2010

Cleônidas Tavares de Souza Junior*
Desmond Campbell**
Srinivasa Vittal Katikireddi***
Paulo Victor Maciel da Costa****
Gervásio Ferreira dos Santos*****
Mauricio Lima Barreto******
Roberto Fernandes Silva Andrade*******

O Censo brasileiro de 2010 contém uma riqueza de informações que podem permitir pesquisas e subsidiar políticas em saúde, educação, economia e outros setores. O Censo fornece dados publicamente disponíveis em duas formas. Primeiro, tabelas de contingência no nível municipal, para estratos definidos por raça, gênero e educação. Segundo, microdados com informações pessoais. Para preservar o anonimato individual nos dados, o Censo reduziu algumas variáveis a categorias mais amplas e removeu dados com identificações pessoais. As estratégias de composição de dados das tabelas de contingência e dos microdados são diferentes e, ao comparar amostras de ambos os dados, descobrimos que a variável raça nos microdados ignora a presença de minorias em alguns municípios. Isso sugere que populações sintéticas baseadas no Censo de 2010 devem ser criadas usando tabelas de contingência. Nossa avaliação mostra que a população sintética assim criada mantém os valores e proporções das tabelas de contingência e apresenta totais próximos aos dos microdados.

**Palavras-chave:** População. Análise de coorte. Simulação de computador. Inferência estatística.

* Centro de Integração de Dados e Conhecimentos para Saúde, Fundação Oswaldo Cruz (Cidacs/Fiocruz), Salvador-BA, Brasil (cleonidas@gmail.com; https://orcid.org/0000-0002-0277-1154).

** Universidade de Glasgow, Glasgow, Reino Unido (desmond.campbell@glasgow.ac.uk; https://orcid.org/0000-0003-1085-714X).

*** Universidade de Glasgow, Glasgow, Reino Unido (vittal.katikireddi@glasgow.ac.uk; https://orcid.org/0000-0001-6593-9092).

**** Centro de Integração de Dados e Conhecimentos para Saúde, Fundação Oswaldo Cruz (Cidacs/Fiocruz), Salvador-BA, Brasil (paulo.vcosta@fiocruz.br; https://orcid.org/0000-0002-8326-2131).

***** Centro de Integração de Dados e Conhecimentos para Saúde, Fundação Oswaldo Cruz (Cidacs/Fiocruz), Salvador-BA, Brasil (gervasiofsantos@gmail.com; https://orcid.org/0000-0002-3545-3590).

****** Centro de Integração de Dados e Conhecimentos para Saúde, Fundação Oswaldo Cruz (Cidacs/Fiocruz), Salvador-BA, Brasil (mauricio.barreto@fiocruz.br; https://orcid.org/0000-0002-0215-4930).

******* Centro de Integração de Dados e Conhecimentos para Saúde, Fundação Oswaldo Cruz (Cidacs/Fiocruz), Salvador-BA, Brasil (randrade@ufba.br; https://orcid.org/0000-0002-9323-1400).

## Introdução

Um grande impedimento para ter dados de nível individual amplamente disponíveis é a exigência legal e ética de manter o anonimato em relação a informações confidenciais. Mesmo quando identificadores óbvios, como nome, foram removidos de um conjunto de dados, ainda pode ser possível reidentificar indivíduos a partir da combinação de informações para um conjunto de atributos. Ao fazer isso, qualquer outra informação no registro pode ser associada a esse indivíduo, inclusive informações confidenciais que o indivíduo nunca teve intenção de divulgar. Uma maneira de abordar isso é fornecer apenas estatísticas resumidas em algum nível agregado sobre variáveis. Outra possibilidade é criar um conjunto de dados sintéticos que contenham informações artificiais que imitem os dados do mundo real. Isso supera a restrição de confidencialidade: como nenhum dos dados pertence a pessoas reais e eles fornecem uma aproximação razoável da população real, dados sintéticos podem ser disponibilizados publicamente e aplicados a simulações de políticas em diferentes cenários, projeções populacionais e para produzir estimativas geográficas específicas a partir de variáveis armazenadas em diferentes bancos de dados. Em trabalhos de pesquisa em que não é possível usar uma população real, as populações sintéticas fornecem um valioso banco de ensaios que permitem avaliar antecipadamente o impacto de políticas públicas locais, como serviços de saúde, educação e segurança social (Prédhumeau; Manley, 2023; Li, Vidyattama; 2019).

Um conjunto de dados sintéticos é uma coleção de dados artificiais que imitam dados do mundo real (Bissett *et al.*, 2021). Em estudos de saúde, preservar a privacidade do paciente é uma exigência legal e ética. Nesses casos, portanto, as informações do paciente são agregadas de forma a manter os padrões observados nos dados originais e transformadas em conjuntos de dados sintéticos, criando pacientes falsos, porém, com características altamente realistas (Arora *et al.*, 2024). Uma maneira de dificultar a reidentificação de pessoas é criar dados sintéticos com modelos estocásticos projetados de tal forma que dados sintéticos gerados naturalmente não estabeleçam uma relação um-para-um entre indivíduos sintéticos e reais. Em situações em que é necessário preservar a identidade dos indivíduos e manter uma relação um-para-um entre os registros originais e sintéticos, utiliza-se a desidentificação por meio de métodos de anonimização (Gadotti *et al.*, 2024). Além disso, os dados populacionais sintéticos são relevantes para a avaliação de impactos heterogêneos na população (Tozluoğlu *et al.*, 2023).

Em estudos demográficos, as populações sintéticas foram usadas para projetar a futura distribuição da população australiana (Li; Vidyattama, 2019), extrair populações atuais e futuras da população do Canada (Prédhumeau; Manley, 2023) e investigar interações sociais (Chapuis; Taillandier; Drogoul, 2022). Na área da saúde, as populações sintéticas foram aplicadas em estudos sobre a transmissão de doenças infecciosas (Zhu *et al.*, 2024), a qualidade do ar (Etuman *et al.*, 2024), acidentes de trabalho (Nadal *et al.*, 2023) e em estudos sobre transplantes (Gunsalus *et al.*, 2024).

No Brasil, a geração de populações sintéticas aparece principalmente relacionada a estudos sobre transporte. Pianucci *et al.* (2019) criaram sinteticamente a população da cidade São Carlos (SP) para simular o transporte de pessoas. Sallard, Balać e Hörl (2020), com apoio dos fatores de expansão do Censo, produziram uma população sintética para a Região Metropolitana de São Paulo (RMSP) e Ajauskas e Strambi (2024) apresentam um gerador de populações sintéticas aplicado na RMSP e adaptado para criação de dados sintéticos para outras regiões do Brasil. Para aplicações diversas, Furtado (2020) criou um gerador de indivíduos e famílias artificiais para concentrações urbanas de aproximadamente cinquenta cidades brasileiras e com microdados. Ton *et al.* (2024) geraram populações sintéticas para diferentes países, incluindo o Brasil.

Embora existam diferentes aplicações para a população sintética, alguns obstáculos acabam por limitar a precisão da informação gerada. Uma população sintética similar à população real é dependente de dados precisos e abrangentes; dados imprecisos ou tendenciosos tendem a gerar uma população sintética que não se aproxima da realidade de uma população real (Tozluoğlu *et al.*, 2023).

O Instituto Brasileiro de Geografia e Estatística (IBGE) é um órgão público federal que, por meio do Censo, coleta, organiza e dissemina dados geográficos e sociais brasileiros. O IBGE divulgou dados do Censo de 2010 na forma de tabelas de contingência de estatísticas agregadas no nível municipal e microdados (amostras de dados em nível individual que fornecem valores para um conjunto maior de variáveis) (IBGE-Microdados, 2024). Os dados mais recentes da população brasileira (Censo de 2022) estão sendo disponibilizados em etapas. Por exemplo, até julho de 2025 os microdados não tinham sido divulgados. Dessa forma, o Censo de 2010 ainda continua sendo a fonte de informação com os conjuntos de dados mais completos sobre a população brasileira.

Os microdados do Censo de 2010 capturam características gerais da população, mas podem ser imprecisos para pequenos grupos populacionais por razões de manutenção da confidencialidade. Por exemplo, os municípios com códigos IBGE 5103205 e 2606200 possuem indígenas autodeclarados nas tabelas de contingência do Censo 2010, enquanto os microdados informam que não há indígenas nesses municípios.

No presente trabalho, o objetivo é apresentar um método para criação de populações sintéticas a partir da consolidação de subconjuntos de dados do Censo brasileiro de 2010. O estudo se concentra na criação de populações sintéticas agregando totais de todos os municípios brasileiros, gerando assim uma população sintética que reproduza com mais precisão detalhes das minorias raciais do que aqueles fornecidos pelos microdados divulgados.

## Métodos

*Fundamentação teórica*

A construção de uma população sintética pode ser apoiada por diferentes modelos teóricos e a escolha do modelo depende do tipo, quantidade e qualidade dos dados disponíveis. Os principais métodos para criação de populações sintéticas estão classificados em três modelos teóricos: reconstrução sintética (*synthetic reconstruction* – SR); otimização combinatória (*combinatorial optimization* – CO); e aprendizado estatístico (*statistical learning* – SL). Para embasar a tomada de decisão de qual modelo usar na criação de uma população sintética, Yaméogo *et al.* (2020) propõem, com base nas vantagens e limitações de cada modelo, um procedimento que infere um método de criação de populações sintéticas com base nas características dos dados da população (tipo, quantidade e qualidade) de cada estudo.

Para criar dados sintéticos, a reconstrução sintética (SR) usa basicamente dois procedimentos: um para ajuste e outro para alocação de dados. Partindo de uma amostra populacional, na fase de ajuste são atribuídos pesos aos indivíduos na amostra de modo que as somas dos pesos correspondam às somas marginais dos dados. Na alocação, indivíduos sintéticos são replicados a partir de conjuntos de indivíduos com as mesmas características e suas respectivas somas de pesos. Os métodos SR são dependentes de amostras consistentes; então, se a geração de indivíduos sintéticos for feita com amostras inconsistentes, somente indivíduos que aparecem nas amostras serão criados, deixando de fora outros conjuntos de indivíduos que aparecem na população real (Yaméogo *et al.*, 2020).

Os métodos mais conhecidos de SR são o *iterative proportional fitting* (IPF) e o *iterative proportional updating* (IPU). No IPF, uma tabela de contingência é ajustada iterativamente de modo que as somas marginais coincidam com valores conhecidos da população sem que, com isso, os valores das células da tabela sofram grandes mudanças (Beckman; Baggerly; Mckay, 1996). O IPU, uma extensão do IPF, foca em equilibrar os dados dos domicílios aos dados de nível individual. Em cada iteração do IPU, primeiro são ajustados os dados domiciliares com os valores marginais das tabelas e, em seguida, os dados dos indivíduos (Ye *et al.*, 2009).

Os métodos baseados na otimização combinatória (CO) minimizam as diferenças marginais com as tabelas de contingência, buscando, por meio de seleção e combinação de indivíduos, soluções otimizadas a partir de conjuntos finitos de soluções (Voas; Williamson, 2000). Um procedimento comum em CO é a replicação de agentes existentes nas amostras, no entanto, os métodos que são excessivamente dependentes da replicação podem gerar novos desafios conceituais e empíricos para geração e validação dos dados sintéticos (Zhu *et al.*, 2024; Etuman *et al.*, 2024). Uma desvantagem da CO é que, quando se trabalhar com populações muito grandes, as soluções ótimas aumentam a complexidade computacional, exigindo mais recursos para se chegar ao resultado (Yaméogo *et al.*, 2020).

No aprendizado estatístico (SL), as populações sintéticas são criadas a partir de probabilidades estimadas de amostras ou dados agregados de uma população real (Farooq *et al.*, 2013). Os métodos de SL usam dados parciais de uma população real para construir populações sintéticas que tenham distribuições empíricas similares às da população real. A partir de pequenas populações, métodos bayesianos, por exemplo, são usados para estimar taxas de mortalidade desagregadas por idade e sexo (Zhang *et al.*, 2019). Os métodos de SL são capazes de reproduzir indivíduos que não aparecem nas amostras, apresentam maneiras mais sistemáticas de imputar dados, mas falham ao relacionar as distribuições condicionais com as distribuições marginais das variáveis do estudo (Yaméogo *et al.*, 2020).

*Método para criação de populações sintéticas brasileiras*

Na primeira etapa deste trabalho, utilizamos os dados agregados para gerar a população sintética brasileira e, na segunda etapa, os microdados do Censo são expandidos para o nível individual, com o objetivo de permitir a validação da população sintética brasileira.

Na primeira etapa, inicialmente identificamos no Censo de 2010 onde aparecem as variáveis sociais código do município, domicílio, sexo, raça, situação escolar, estado civil, renda e idade. Essas informações podem ser obtidas de duas formas: pelo Sistema IBGE de Recuperação Automática (IBGE, 2024) e pelo *site* do IBGE-*Downloads* (IBGE-Downloads, 2024). Na Tabela 1 apresentamos as relações entre as variáveis de interesse e suas respectivas fontes de informação.

O objetivo da construção da Tabela 1 é indicar onde as variáveis sociais deste projeto aparecem dentro das tabelas de contingência do Censo. Essa identificação é necessária para verificar quais variáveis são comuns nas tabelas de contingência e assim estabelecer uma conexão entre elas. No modelo de agregação de dados do Censo de 2010 que propomos, primeiro conectamos as tabelas com variáveis que são comuns entre elas (por exemplo, municípios e estados) e, em seguida, estabelecemos novas conexões considerando proporções e ajustes como os apresentados na Tabela 2.

**TABELA 1**
**Relação entre tabelas de contingência e as variáveis sociais selecionadas – Censo de 2010**

| Tabelas de contingência | Estado | Município | Domicílio | Sexo | Raça | Idade | Situação escolar | Estado civil | Rendimentos |
|---|---|---|---|---|---|---|---|---|---|
| Tabelas 4.*.1.1 – População residente, por situação do domicílio e sexo | x | x | x | x | | | | | |
| Tabelas 4.* – População residente, por situação do domicílio e cor ou raça | x | x | x | | x | | | | |
| BR_TAB7 – População residente em domicílios particulares permanentes | x | x | x | | | x | | | |
| Tabelas 2.*.5.1 – População residente, por frequência a escola ou creche e rede de ensino que frequentavam | x | x | | | | | x | | |

(continua)

(continuaçnao)

| Tabelas de contingência | Estado | Município | Domicílio | Sexo | Raça | Idade | Situação escolar | Estado civil | Rendimentos |
|---|---|---|---|---|---|---|---|---|---|
| Tabelas 2.*.3.2 – Pessoas de 10 anos ou mais de idade, por estado civil | x | x | | | | | | x | |
| Tabelas 4.*.7.3 – Valor do rendimento nominal mediano mensal das pessoas de 10 anos ou mais de idade, total e com rendimento | x | x | x | | | | | | x |
| Tabelas 3.*.2.4 – Pessoas residentes em domicílios particulares, por cor ou raça, segundo o sexo e as classes de rendimento nominal mensal domiciliar *per capita* | x | | | x | x | | | | x |
| Tabelas 2.*.9.2 – Pessoas de 10 anos ou mais de idade, com rendimento, e valor do rendimento nominal médio e mediano mensal das pessoas de 10 anos ou mais de idade, com rendimento, por sexo | x | x | | x | | | | | x |
| Tabela 3.*.2.4 – Pessoas residentes em domicílios particulares, por cor ou raça, segundo o sexo e as classes de rendimento nominal mensal domiciliar *per capita* | x | | | x | x | | | | x |
| Tabela 4.*.7.3 – Valor do rendimento nominal mediano mensal das pessoas de 10 anos ou mais de idade, total e com rendimento, por situação do domicílio. | x | x | x | | | | | | x |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE.
Nota: O asterisco (*) entre a numeração das tabelas de contingência representa um conjunto com 27 tabelas, uma para cada unidade da federação brasileira.

**TABELA 2**
**Exemplo de ajuste proporcional de totais de indivíduos entre as variáveis raça e sexo no município de Acrelândia (Acre) – Censo de 2010**

*(a) Calculando proporções entre as variáveis raça e sexo*

| Informações sobre a raça | | | Informações sobre o sexo | |
|---|---|---|---|---|
| | | | Rural | |
| | | | 6.622 | |
| | | | Homens | Mulheres |
| Rural | Raça | Total | 3.646 | 2.976 |
| 6.622 | Branca | 1.600 | (1.600/6.622) * 3.646 | (1.600/6.622) * 2.976 |
| | Negra | 315 | (315/6622) * 3.646 | (315/6.622) * 2.976 |
| | Parda | 4.654 | (4.654/6.622) * 3.646 | (4.654/6.622) * 2.976 |
| | Amarela | 53 | (53/6.622) * 3.646 | (53/6.622) * 2.976 |
| | Indígena | 0 | (0/6.622) * 3.646 | (0/6.622) * 2.976 |
| | Não declarada | 0 | (0/6.622) * 3.646 | (0/6.622) * 2.976 |

*(b) Inferindo os totais de indivíduos por raça e sexo*

| Informações sobre a raça | | | Informações sobre o sexo | |
|---|---|---|---|---|
| | | | Rural | |
| | | | 6.622 | |
| | | | Homens | Mulheres |
| Rural | Raça | Total | 3.646 | 2.976 |
| 6.622 | Branca | 1.600 | 881 | 719 |
| | Negra | 315 | 173 | 142 |
| | Parda | 4.654 | 2.562 | 2.092 |
| | Amarela | 53 | 29 | 24 |
| | Indígena | 0 | 0 | 0 |
| | Não declarada | 0 | 0 | 0 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE.

Utilizando outro ponto de vista, ilustramos na Tabela 3 a aplicação da proporcionalidade. A Tabela 3a apresenta o total de habitantes de um município (Acrelândia) por *domicílio* e *sexo por domicílio*. A Tabela 3b informa o total de habitantes do mesmo município, por *domicílio* e *raça por domicílio*. A última coluna da Tabela 3b mostra a proporção de cada raça em relação ao domicílio e na Tabela 3c são registradas as proporções aferidas para as *raças* de pessoas do sexo masculino do município. Ao calcular a proporcionalidade entre as variáveis, não estamos assumindo nenhuma correlação entre elas, estamos usando a distribuição conhecida de uma variável para inferir a distribuição da outra e, dessa forma, manter coerentes os totais e subtotais das variáveis que estão sendo agregadas.

**TABELA 3**
**Amostra da população do município de Acrelândia (Acre) – Censo de 2010**

*(a) Tabela 4.2.1.1 – Informações sobre sexo e domicílio (amostra)*

| Código do município | População | Domicílio | População por tipo de domicílio | Sexo | Subtotais |
|---|---|---|---|---|---|
| 1200013 | 12.538 | Rural | 6.622 | Masculino | 3.646 |
| 1200013 | 12.538 | Rural | 6.622 | Feminino | 2.976 |
| 1200013 | 12.538 | Urbano | 5.916 | Masculino | 2.946 |
| 1200013 | 12.538 | Urbano | 5.916 | Feminino | 2.970 |

*(b) Tabela 4.2 – Informações sobre raça e domicílio (amostra)*

| Código do município | População | Domicílio | População por tipo de domicílio (a) | Raça | Subtotal por raça (b) | Proporção (b/a) * |
|---|---|---|---|---|---|---|
| 1200013 | 12.538 | Rural | 6.622 | Branca | 1.600 | 0,2416 |
| 1200013 | 12.538 | Rural | 6.622 | Negra | 315 | 0,0475 |
| 1200013 | 12.538 | Rural | 6.622 | Amarela | 53 | 0,0080 |
| 1200013 | 12.538 | Rural | 6.622 | Parda | 4.654 | 0,7028 |
| 1200013 | 12.538 | Rural | 6.622 | Indígena | 0 | 0,0000 |
| 1200013 | 12.538 | Rural | 6.622 | Sem declaração | 0 | 0,0000 |

*(c) Cálculo de agregação: Informações sobre domicílio, sexo e raça (Tabela 4.2 & Tabela 4.2.1.1)*

| Código do município | População | Domicílio | População por tipo de domicílio | Sexo | (a) Subtotal (sexo) | Raça por tipo de domicílio | Subtotal (raça) | (b) Proporção (raça) | (a*b) Subtotal (raça por sexo) |
|---|---|---|---|---|---|---|---|---|---|
| 1200013 | 12538 | Rural | 6.622 | Masculino | 3.646 | Branca | 1600 | 0,2416 | 881 |
| 1200013 | 12539 | Rural | 6.622 | Masculino | 3.646 | Negra | 315 | 0,0475 | 173 |
| 1200013 | 12539 | Rural | 6.622 | Masculino | 3.646 | Amarela | 53 | 0,0080 | 29 |
| 1200013 | 12539 | Rural | 6.622 | Masculino | 3.646 | Parda | 4.654 | 0,7028 | 2.562 |
| 1200013 | 12539 | Rural | 6.622 | Masculino | 3.646 | Indígena | 0 | 0,0000 | 0 |
| 1200013 | 12539 | Rural | 6.622 | Masculino | 3.646 | Sem declaração | 0 | 0,0000 | 0 |
| | | | | Total | 3.646 | | | Total | 3.645 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE.

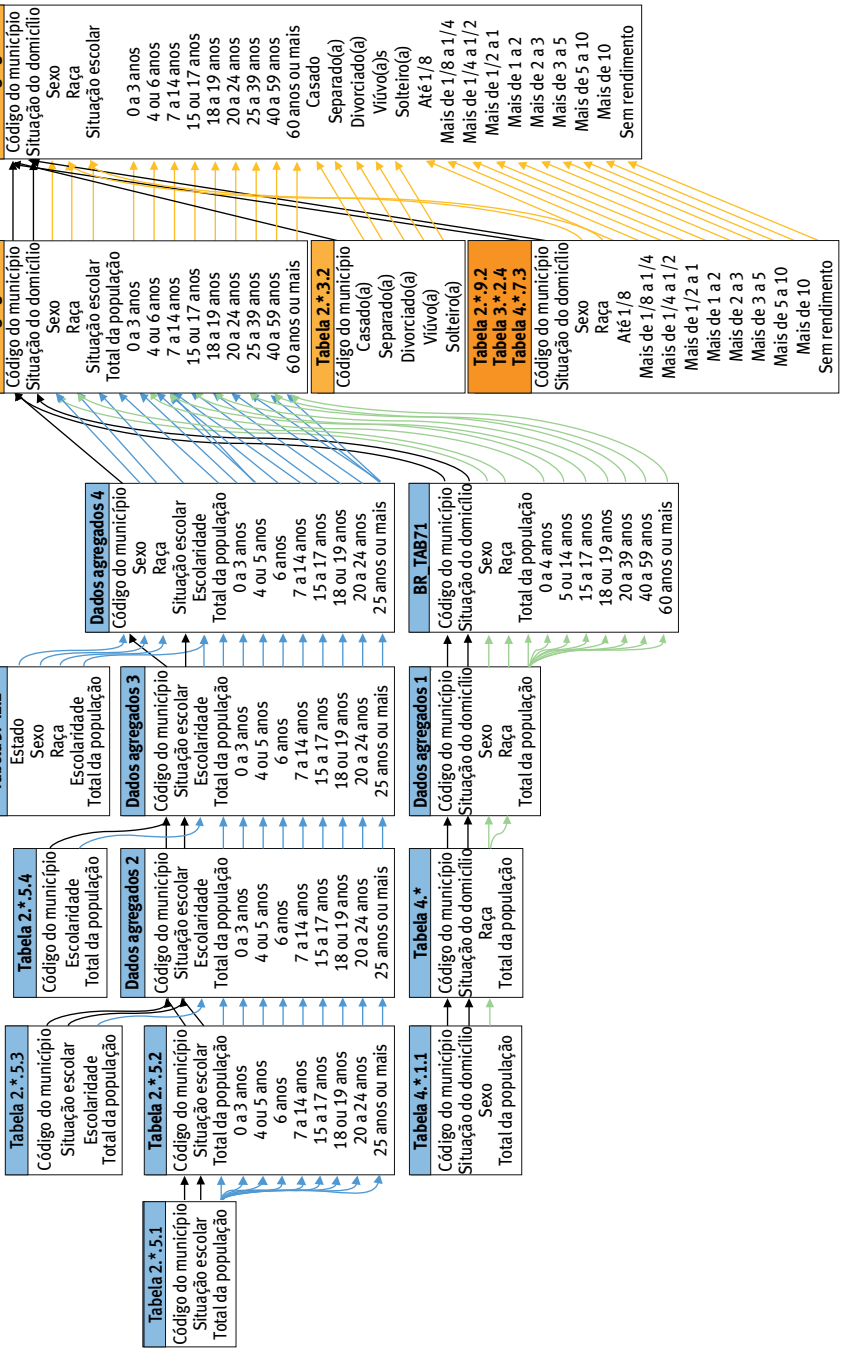
Devido a arredondamentos em algumas operações envolvendo dados entre diferentes colunas – por exemplo em *(a*b) subtotal (raça por sexo)* (Tabela 3c) –, é possível que algumas somas marginais não coincidam exatamente: na Tabela 3c, a soma das raças entre os indivíduos de sexo masculino (3.645) difere do total de pessoas do sexo masculino (3.646). Nesses casos, ajustamos essas diferenças usando as proporções encontradas (última coluna da Tabela 3b), selecionando aleatoriamente uma das raças e adicionando a diferença a ela.

O processo de criação de populações sintéticas que propomos é um modelo de reconstrução sintética baseado no método de *iterative proportional fitting* (IPF) que usa as somas marginais das matrizes para inferir valores para pares de variáveis. No entanto, em vez de manter entre as matrizes valores fracionários de indivíduos, como o IPF sugere, transformamos os conjuntos de indivíduos com valores fracionários em conjuntos de indivíduos com valores inteiros. Em seguida, por um sorteio baseado na distribuição de valores de uma das variáveis, adicionamos a um dos atributos do indivíduo as diferenças observadas entre os valores das tabelas de contingência e a dos dados sintéticos.

Depois de encontrar os valores da variável raça, um novo processo de conexão e aplicação de proporcionalidades é iniciado para a inclusão de uma nova variável (Figura 1). O propósito dessas operações (i.e., conexão entre variáveis, cálculo de proporções e ajustes) é estabelecer conexões entre as tabelas e garantir que os totais e subtotais não fiquem díspares entre as variáveis. Ignorar as diferenças causadas pelos arredondamentos pode gerar erros maiores nos totais populacionais das cidades, dos estados e do país. Com base nessas operações, agregamos as tabelas de contingência do Censo de 2010 conforme a Figura 1 (as setas pretas representam as variáveis-chave das conexões, as setas azuis referem-se a variáveis com valores extraídos de tabelas de contingência, as listas em azul são dados de escolaridade – e.g. número total de pessoas que frequentam, não frequentam mais e nunca frequentaram a escola –, as listas em verde são dados gerais da população e as listas em laranja consolidam dados individuais, escolaridade e gerais da população – e.g. raça, sexo, estado civil).

Embora o método de consolidação das tabelas seja simples (e.g., Tabela 3 e Figura 1), tivemos que fazer algumas inferências durante o processamento. Por exemplo, as *Tabelas 2.*.3.2 – Pessoas de 10 anos ou mais de idade, por estado civil*, na Tabela 1, restringem a população a indivíduos com mais de nove anos. Nessa situação inferimos que todos os menores de 10 anos são *solteiros*. Para as *Tabelas 2.*.9.2* e *Tabelas 4.*.7.3* (Tabela 1) inferimos que menores de 10 anos não têm renda. Essas inferências foram necessárias para manter os totais e subtotais coerentes entre as variáveis das populações municipais (e.g., a soma das *raças* dos indivíduos de *sexo masculino* na Tabela 3c tem que ser igual ao total de indivíduos do *sexo masculino*, assim como a soma dos indivíduos por *sexo* tem que ser o total da população do município). Depois de agregar os dados conforme nossa proposta, obtivemos uma tabela com 29 variáveis (Tabela 4) e 1,1 milhão de registros.

FIGURA 1
Diagrama de relacionamentos




Fonte: Elaboração dos autores.

A Tabela 4 está estruturalmente dividida em dois tipos de dados: os dados agregados, que registram os atributos dos indivíduos e o total de pessoas entre as faixas etárias; e as distribuições (probabilidades), que apresentam as probabilidades de os indivíduos pertencerem a um dos atributos das variáveis *estado civil* e *renda* associada. Essa divisão da Tabela 4 em totais de indivíduos e probabilidades dos indivíduos por *estado civil* e *renda* existe porque, durante o processo de conexão dos dados do Censo (Figura 1), nem sempre foi possível estabelecer uma relação um-para-um entre as variáveis das tabelas de contingência. Por isso, mantivemos os totais de indivíduos organizados na parte de dados agregados da Tabela 4 e as distribuições, inerentes aos estados civis e às rendas da população, na parte das distribuições. Por exemplo, na Figura 1, o estado civil (Tabela 2.*.3.2) não separa os indivíduos por *sexo*, então mantivemos as distribuições informadas pelo estado civil ao nível do município em uma parte específica da Tabela 4.

A população sintética brasileira foi criada a partir da Tabela 4. O método que desenvolvemos para a geração de populações sintéticas irá buscar, entre os intervalos de idades de cada linha da Tabela 4, o total de indivíduos a ser criado. O método irá atribuir a esses indivíduos os atributos registrados em cada linha e, para cada novo indivíduo sintético, sorteará uma *idade*, um *estado civil* e uma *renda* com bases nos respectivos intervalos de idades e distribuições de *estado civil* e *rendimento* da linha. Por exemplo, na primeira linha da Tabela 4, existem sete pessoas com idades entre zero e três anos, as quais se tornarão sete indivíduos sintéticos com idade entre zero e três anos que vão morar no município de Alta Floresta d'Oeste, na área rural, serão homens, brancos, estudantes, com 100% de chance de não terem renda e 100% de chance de serem solteiros. Para criar a população sintética brasileira, desenvolvemos o algoritmo SyntPopBr, cujo funcionamento está descrito no algoritmo 01.

**TABELA 4**
**Amostra dos dados agregados do município de Alta Floresta d'Oeste (Rondônia) – Censo de 2010**

| Tipo de dado | Características | Variáveis | Registro 01 | Registro 02 | Registro 03 |
|---|---|---|---|---|---|
| Dados agregados | Atributos dos indivíduos | Código do município | 1100015 | 1100015 | 1100015 |
| | | Domicílio | Rural | Rural | Rural |
| | | Sexo | Homem | Homem | Homem |
| | | Raça | Branco | Branco | Branco |
| | | Situação escolar | Estudante | Estudante | Estudante |
| | Total de indivíduos por idade (anos) | 00 a 03 | 7 | 1 | 0 |
| | | 04 a 06 | 0 | 58 | 25 |
| | | 07 a 14 | 0 | 0 | 420 |
| | | 15 a 17 | 0 | 0 | 0 |
| | | 18 a 19 | 0 | 0 | 0 |
| | | 20 a 24 | 0 | 0 | 0 |
| | | 25 a 39 | 0 | 0 | 0 |
| | | 40 a 59 | 0 | 0 | 0 |
| | | 60 ou mais | 0 | 0 | 0 |
| Distribuições (probabilidades) | Estado civil | Casado(a) | 0 | 0 | 0 |
| | | Separado(a) | 0 | 0 | 0 |
| | | Divorciado(a) | 0 | 0 | 0 |
| | | Viúvo(a) | 0 | 0 | 0 |
| | | Solteiro(a) | 1 | 1 | 1 |
| | Rendimentos (salários mínimos (x)) | $0 < x \leq 1/8$ | 0 | 0 | 0,03 |
| | | $1/8 < x \leq 1/4$ | 0 | 0 | 0,05 |
| | | $1/4 < x \leq 1/2$ | 0 | 0 | 0,17 |
| | | $1/2 < x \leq 1$ | 0 | 0 | 0,28 |
| | | $1 < x \leq 2$ | 0 | 0 | 0,23 |
| | | $2 < x \leq 3$ | 0 | 0 | 0,08 |
| | | $3 < x \leq 5$ | 0 | 0 | 0,06 |
| | | $5 < x \leq 10$ | 0 | 0 | 0,03 |
| | | $x \leq 10$ | 0 | 0 | 0,02 |
| | | $x = 0$ | 1 | 1 | 0,05 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE. Elaboração dos autores.

**Algoritmo 01: Criação da população sintética (SyntPopBr)**

```
1  Entrada: dados_agregados: dados agregados do Censo de 2010
2  Início:
3      criar vetores (identidade, código_município, domicílio, sexo, raça, situação escolar, estado civil, renda, idade)
4      Entrada: município: informar o código de um município
5      Entrada: seed: informar um valor para o gerador de números aleatórios (seed)
6      tab_tmp = dados agregados da variável município
7      identificação = 0
8      n = total de registro em tab_tmp
9      De 1 até n faça:
10         distribuição_do_estado_civil = obter a distribuição dos estados civis de tab_tmp.
11         m = total de intervalos de idades em tab_tmp
12         De 1 até m faça:
13             saldo = total de indivíduos no intervalo de idade
14             Enquanto (saldo > 0) faça:
15                 saldo = saldo – 1
16                 identificação = identificação + 1
17                 var_idade = sortear(uma idade no intervalo de idade atual)
18                 Se (var_idade < 10) então:
19                 var_estado_civil = 'solteiro'
20                 var_renda = 0
21                 Senão:
22                 var_estado_civil = sortear (um estado_civil com a distribuição_do_estado_civil)
23                 var_salario_minimo = média (rend_domicilio, rend_sexo)
24                 distribuição_das_rendas = obter a distribuição das rendas (de rend_1p8 a rend_sem de tab_tmp)
25                 var_renda = sortear (uma renda com a distribuição_das_rendas)
26                 var_renda = var_renda * var_salario_minimo
27                 Fim
28                 incluir (identificação, tab_tmp.município, tab_tmp.domicílio, tab_tmp.sexo, tab_tmp.raça, tab_tmp.situação_escolar, var_estado_civil, var_renda, var_idade) nos vetores (identidade, código_município, domicílio, sexo, raça, situação escolar, estado civil, renda, idade)
29             Fim
30         Fim
31     Fim
32     população_sintetica = criar tabela com (identidade, código_município, domicílio, sexo, raça, situação escolar, estado civil, renda, idade)
33     gravar população_sintetica como população_sintetica_censo2010.csv
34 Fim
```

O SyntPopBr recebe os dados agregados do Censo de 2010 (Tabela 4), o código de um município e o número de uma *semente* (linhas 1, 4 e 5). A *semente* controla o gerador de números pseudoaleatórios e permite reproduzir populações sintéticas com diferentes características (i.e., totais, subtotais e distribuições). Essas diferenças entre as populações geradas para um mesmo município poderão ser percebidas nas variáveis *idade*, *renda* e *estado civil*. Os laços de repetição, entre as linhas 9 e 31 do SyntPopBr, leem os dados agregados de um *município* e, a cada intervalo de idade (variáveis de *0 a 3 anos* até *60 anos ou mais* (Figura 1)), geram os valores da *renda* e do *estado civil* com base em distribuições conhecidas da população. A variável *idade* é atribuída aleatoriamente dentro de intervalo de idade atual e as demais variáveis são extraídas dos dados agregados do Censo.

*Expansão da população dos microdados*

Os microdados do Censo de 2010 referem-se a uma fração de aproximadamente 10% da população total brasileira. Para estimar os totais por domínios de interesse (e.g. sexo, domicílio, raça), o relatório metodológico do Censo Demográfico de 2010 (IBGE, 2013) esclarece que os pesos atribuídos entre as linhas nos microdados devem ser usados para estimar as variáveis de interesse da população. Sendo assim, conforme a equação (1) sugerida pelo relatório do Censo (IBGE, 2013, p. 641), estimativas de totais na população brasileira podem ser extraídas dos microdados por meio da variável *peso* (Tabela 5a) ou *fator de expansão* (i.e., frações da amostra ajustada de modo a representar uma parte da população ou parte de uma área geográfica). A estimativa para a contagem de $\hat{Y}$ na população brasileira é dada por:

$$\hat{Y} = \sum_{i=1}^{n} p_i y_i \quad (1)$$

Onde: $p_i$ é o peso associado à *i-ésima* unidade da amostra; $y_i$ corresponde ao valor associado à *i-ésima* unidade da amostra no domicílio; e $n$ é o número de unidades na amostra da busca em questão.

Com base na equação (1), a Tabela 5b apresenta os totais da população do município 1100015 por domicílio (1 – urbano; 2 – rural) e sexo (1 – masculino, 2 – feminino). Nesse caso em particular, estima-se que existam 6.970 homens residentes em domicílios urbanos no município 1100015 (e.g., Tabela 5b). Os pesos na equação (1) não são inteiros, eles foram calibrados com variáveis auxiliares baseadas na relação entre os domicílios pesquisados e os domicílios selecionados para a amostra, então, para encontrar o total de um extrato populacional, basta aplicar os pesos aos extratos de domicílios a serem pesquisados. Observamos que, quando as tabelas de contingência do Censo apresentam poucos indivíduos da mesma raça, o processo de estimação realizado pela equação (1) pode gerar totais de indivíduos menores que 1 para alguns municípios, o que provoca a ausência dessas raças nos municípios no processo de arredondamento do total para um número inteiro.

**TABELA 5**
**Microdados do município de Alta Floresta d'Oeste (Rondônia) – Censo de 2010**

*(a) Amostra dos microdados*

| Código do município | Domicílio | Sexo | Raça | Situação escolar | Estado civil | Renda | Idade | Peso |
|---|---|---|---|---|---|---|---|---|
| 1100015 | 1 | 1 | 4 | 3 | 5 | 0 | 5 | 8,705 |
| 1100015 | 1 | 2 | 4 | 1 | 5 | 523 | 16 | 8,705 |
| 1100015 | 1 | 1 | 1 | 3 | 5 | 1.532 | 90 | 9,818 |
| 1100015 | 1 | 2 | 1 | 3 | 5 | 3.541 | 72 | 9,495 |

*(b) Soma dos pesos para as variáveis domicílio, sexo e o código do município*

| Código do município | Domicílio | Sexo | Σ(peso) |
|---|---|---|---|
| 1100015 | 1 | 1 | 6,970 |
| 1100015 | 1 | 2 | 7,000 |
| 1100015 | 2 | 1 | 5,686 |
| 1100015 | 2 | 2 | 4,736 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE. Elaboração dos autores.

Selecionamos as variáveis *código do município, domicílio, sexo, raça, situação escolar, estado civil, renda e idade* por estarem comumente ligadas a estudos sobre desigualdades. A variável *código do município* corresponde ao *código IBGE* que identifica os municípios brasileiros, *domicílio* identifica a localidade onde os habitantes dos municípios moram (1 – áreas urbanas; 2 – áreas rurais), *sexo* separa homens e mulheres (1 – masculino; 2 – feminino), *raça* ou *cor* é uma variável autodeclarada (1 – branca; 2 – preta; 3 – amarela; 4 – parda; 5 – indígena; 9 – sem declaração), *situação escolar* identifica se a pessoa está estudando ou não (1 – frequenta a escola; 2 – já frequentou a escola; 3 – nunca frequentou a escola), *estado civil* classifica as pessoas em cinco categorias (1 – casado(a); 2 – desquitado(a) ou separado(a) judicialmente; 3 – divorciado(a); 4 – viúvo(a), solteiro(a); 5 – sem declaração), *renda* identifica o montante de dinheiro mensal que cada pessoa recebe como fruto de um emprego ou de um investimento. Nos microdados do Censo de 2010 há aproximadamente 20 milhões de indivíduos de todos os 5.565 municípios brasileiros. Reunimos em uma tabela (e.g., Tabela 5a) esses 20 milhões e as nove variáveis de nosso interesse (incluímos aqui a variável *peso*). Para expandir a população dos microdados ao nível do indivíduo usamos o algoritmo 02 (MicPopExpBr).

**Algoritmo 02: Expansão da população dos microdados (MicPopExpBr)**

```
1  Entrada: microdados: microdados da população brasileira do Censo de 2010
2  Início:
3      Entrada: município: informar o código de um município
4      tab_tmp = estimador (microdados, município)
5      criar vetores (identidade, código_município, domicílio, sexo, raça, situação escolar, estado civil, renda, idade)
6      identificação = 0
7      n = total de registro em tab_tmp
8      De 1 até n faça:
9          m = peso total do registro em tab_tmp
10         De 1 até m faça:
11             identificação = identificação + 1
12             identidade = identificação
13             incluir campos do registro atual de tab_tmp nos vetores (identidade, código_município, domicílio, sexo, raça, situação escolar, estado civil, renda, idade)
14         Fim
15     Fim
16     microdados_expandidos = criar tabela com (identidade, código_município, domicílio, sexo, raça, situação escolar, estado civil, renda, idade)
17     gravar microdados_expandidos_por_individuo_censo2010.csv
18 Fim
```

O algoritmo MicPopExpBr recebe os microdados do Censo de 2010, utiliza a equação (1) para calcular os totais de indivíduos entre as variáveis sociais deste estudo (ou seja, código da cidade, domicílio, sexo, raça, escolaridade, estado civil, renda e idade) e, para cada total identificado, cria (replica) seus respectivos indivíduos em uma nova tabela (cada linha desta nova tabela representa um indivíduo).

Como visto na seção de fundamentação teórica, a criação de uma população sintética pode ser apoiada por diferentes métodos e algoritmos. Por exemplo, o *Synthpop* (Nowok *et al.*, 2016) é um pacote do *software* estatístico R que protege a privacidade dos indivíduos de uma população causando distorções mínimas na versão sintética dos dados. No entanto, para evidenciar que existem diferenças entre as tabelas de contingência e os microdados do Censo de 2010, optamos por expandir os microdados de forma simples e sem distorções por meio do algoritmo MicPopExpBr.

## Avaliação

Em geral, o acesso a dados específicos e confidenciais de uma população está condicionado a normas de proteção de informações privadas e a autorizações específicas de comitês de ética e regras de confidencialidade. Nesse contexto, a validação de dados populacionais, ao nível do indivíduo, fica limitada a se ter acesso aos dados e às normas de descaracterização de dados privados (Leyk *et al.*, 2019). Por outro lado, a avaliação pode ser feita quando os dados do Censo estão disponíveis em diferentes estratos da

população, ou seja, conjuntos de dados em diferentes níveis do modelo são comparados a dados censitários em níveis mais específicos para determinar as semelhanças entre eles (Leyk *et al.*, 2019).

Ao nível municipal, começamos a análise registrando, na Tabela 6, os totais de indivíduos da população sintética brasileira e dos microdados expandidos do Censo conforme os atributos gerados pelo *código do município, domicílio*, *sexo*, *raça* e *situação escolar*. Os valores apresentados pelos microdados expandidos e pela população sintética variam entre eles por diferentes motivos, além dos relacionados pelo processo de descaracterização dos dados realizado pelo IBGE. Destacamos que nos microdados foram inseridos dados faltantes (i.e. NA) e que os totais de indivíduos são projetados com base em pesos que geram valores fracionários entre os totais de indivíduos. Nas tabelas de contingências, os atributos dos indivíduos foram separados em diferentes tabelas e, durante o nosso processo de conexão de dados entre as variáveis (Figura 1), algumas vezes não foi possível determinar com exatidão os totais de indivíduos entre duas variáveis (e.g., sexo e raça). Dessa forma, calculamos as proporções entre os indivíduos das duas variáveis e inferimos uma relação entre eles.

**TABELA 6**
**Amostra da comparação entre os atributos dos indivíduos na população sintética brasileira e nos microdados expandidos – Censo de 2010**

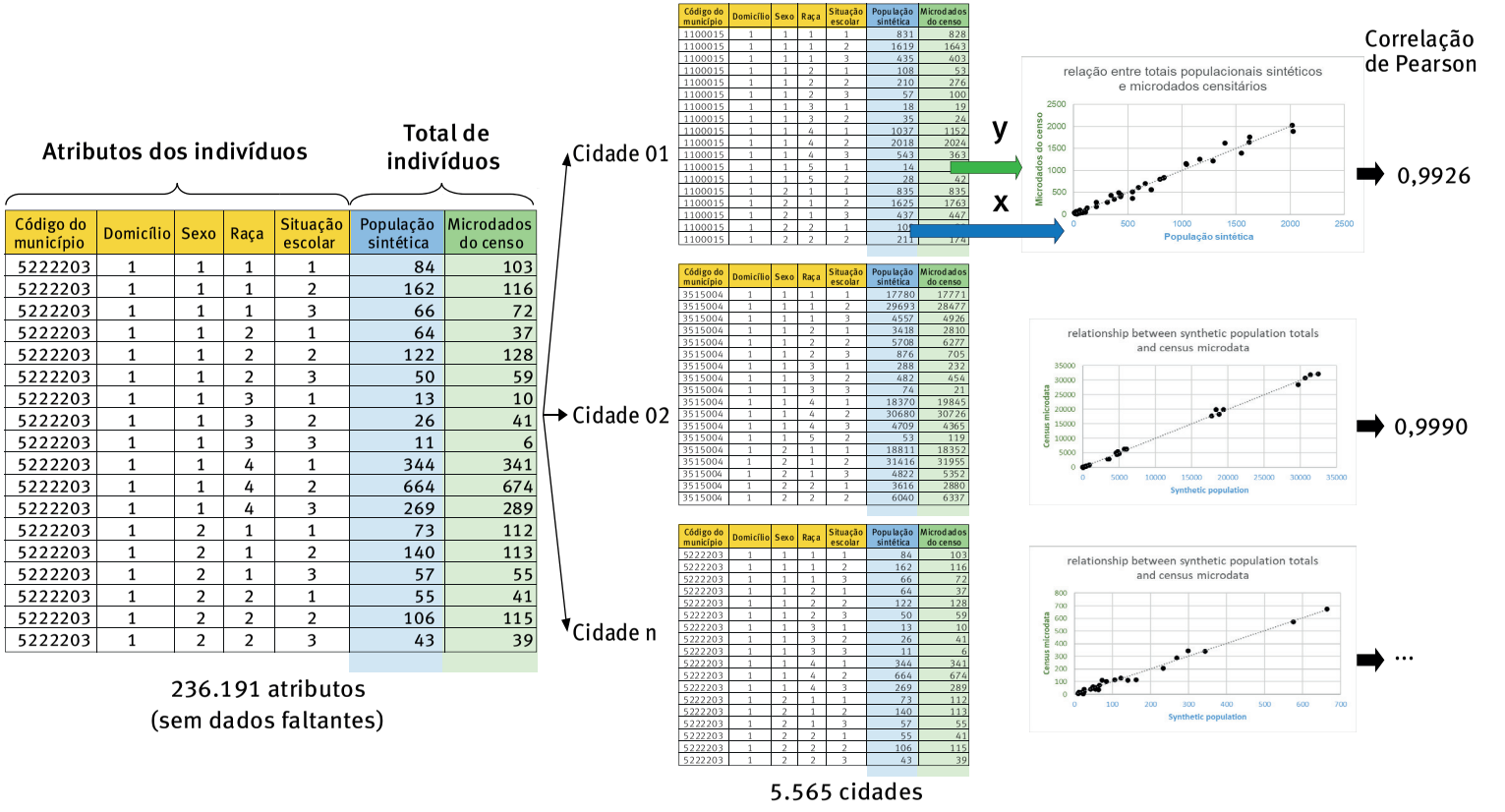
| Código do município | Atributos dos indivíduos | | | | Total de indivíduos | |
|---|---|---|---|---|---|---|
| | Domicílio | Sexo | Raça | Situação escolar | População sintética | Microdados expandidos |
| 1100015 | Urbano | Masculino | Amarela | Ex-estudante | 35 | 24 |
| 1100015 | Urbano | Masculino | Amarela | Nunca frequentou a escola | 9 | NA |
| 1100015 | Urbano | Masculino | Parda | Estudante | 1.037 | 1.152 |
| 1100015 | Urbano | Masculino | Indígena | Ex-estudante | 28 | 42 |
| 1100015 | Urbano | Masculino | Indígena | Nunca frequentou a escola | 8 | NA |
| 1100015 | Urbano | Feminino | Branco | Estudante | 835 | 835 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE. Elaboração dos autores.

Embora nenhum dos dois conjuntos de dados (total de indivíduos na população sintética e nos microdados expandidos) tenha valores iguais aos que deram origem a eles, ambos apresentam totais gerais por municípios similares entre si. Para demonstrar a semelhança entre eles, separamos os dados da Tabela 6 por município e, para cada atributo dos indivíduos, em um gráfico de coordenadas *xy*, projetamos no eixo *x* os totais observados na população sintética e, no eixo *y*, os totais dos microdados expandidos (Figura 2).

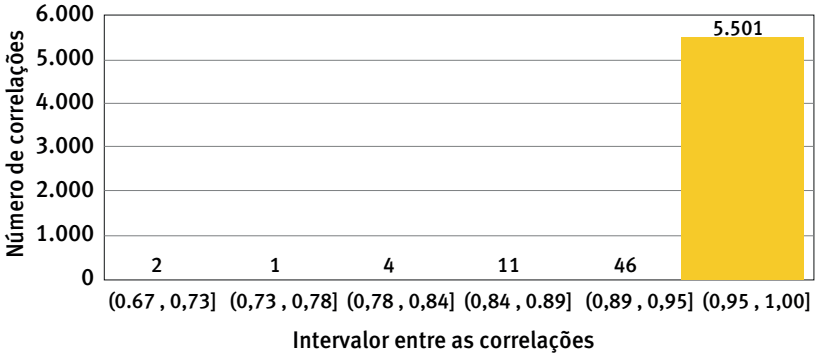
As correlações de Pearson para cada município foram calculadas com base nos totais apresentados pelos microdados e pela população sintética. Obtivemos dessa forma 5.565 correlações que dispusemos no histograma da Figura 3 para demonstrar, ao nível municipal, o quão semelhantes são os dados gerados pela população sintética e os microdados do Censo.

FIGURA 2
Correlação de Pearson entre os totais de indivíduos na população sintética e nos microdados expandidos




Fonte: Elaboração dos autores.

FIGURA 3
Histograma para a distribuição da correlação entre os totais nos microdados expandidos e os totais na população sintética




Fonte: Elaboração dos autores.

## Resultados

Ao nível municipal, os atributos dos indivíduos na população sintética brasileira e na população expandida dos microdados apresentam totais semelhantes de habitantes e correlações próximas de um (Figuras 2 e 3), demonstrando assim que, apesar dessas diferenças, os totais, ao nível municipal, mantêm semelhanças entre essas duas populações. Na Figura 2, os municípios com correlações abaixo de 0,8 apresentaram diferenças que variaram de 30% a 60% entre os totais das duas populações. Ao nível municipal, as principais diferenças encontradas entre essas populações surgem na representação de minorias raciais entre os municípios. Ao analisarmos os resultados da Tabela 6, notamos que a população dos microdados expandidos omitiu a presença de minorias raciais (grupos pequenos de indivíduos de mesma raça) em 1.885 municípios (um total de 15.918 indivíduos). Políticas públicas voltadas para a população indígena, por exemplo, podem falhar ao localizar pessoas autodeclaradas indígenas entre os municípios dos microdados expandidos (e.g. o município de Colniza descrito na seção de discussões). Já na população sintética brasileira que criamos esse problema não existe.

No nível federal, encontramos perfeita compatibilidade entre os dados sintéticos e os microdados expandidos do Censo em relação aos totais de indivíduos por sexo e, para as variáveis domicílio, raça e situação escolar, os totais gerados foram idênticos ou muito próximos (Tabela 7). No entanto, as variações nos números de indivíduos entre os atributos de raça demonstram que as distribuições de indivíduos nos microdados expandidos e na população sintética superestimam em meio milhão o total de pessoas pardas nos microdados expandidos. Os atributos *não declarada*, *branca* e *preta* merecem igual atenção.

**TABELA 7**
**Diferenças de números de indivíduos entre os microdados expandidos e a população sintética – Censo de 2010**

| Variáveis | Atributos | Microdados expandidos (a) | Pspulação sintética (b) | Diferenças (a-b) | Proporções (a/b) |
|---|---|---|---|---|---|
| População | Total | 190.755.799 | 190.755.799 | 0 | 1,00000 |
| Sexo | Feminino | 97.348.809 | 97.348.809 | 0 | 1,00000 |
| | Masculino | 93.406.990 | 93.406.990 | 0 | 1,00000 |
| População por tipo de domicílio | Urbana | 160.934.649 | 160.925.804 | 8.845 | 1,00005 |
| | Rural | 29.821.150 | 29.829.995 | -8.845 | 0,99970 |
| Raça | Branca | 90.621.281 | 91.051.639 | -430.358 | 0,99527 |
| | Parda | 82.820.452 | 82.277.337 | 543.115 | 1,00660 |
| | Preta | 14.351.162 | 14.517.956 | -166.794 | 0,98851 |
| | Amarela | 2.105.353 | 2.084.291 | 21.062 | 1,01010 |
| | Indígena | 821.501 | 817.968 | 3.533 | 1,00432 |
| | Não declarada | 36.051 | 6.608 | 29.443 | 5,45565 |
| Situação escolar | Estudantes | 59.565.188 | 59.564.697 | 491 | 1,00001 |
| | Ex-estudantes | 112.465.161 | 112.467.247 | -2.086 | 0,99998 |
| | Nunca frequentou a escola | 18.725.449 | 18.723.855 | 1.594 | 1,00009 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE. Elaboração dos autores.

Na Tabela 7, o valor mais discrepante está associado à variável raça com atributo *não declarada* (proporções (a/b) 5,45565). O nome da tabela de contingência do Censo 2010 que registra a raça da população brasileira por município é "*Tabelas 4.* – População residente, por situação de moradia e cor ou raça*" (ver Tabela 1). Nessa tabela, ao somarmos as pessoas que não declararam sua raça, obtemos um total de 6.608, o que corresponde ao valor encontrado em nossa população sintética. Os próprios dados do Censo de 2010 inferem valores diferentes para as pessoas que optaram por não declarar sua raça. O método usado para agregar os dados do Censo de 2010, os microdados e os algoritmos *MicPopExpBr* e *SyntPopBr* estão disponíveis no GitHub (Souza-Junior, 2024). Desenvolvemos os algoritmos usando o *software* estatístico R versão 4.4.2 (R Core Team, 2024) e o RStudio versão 2025.05.0+496 (Posit Team, 2025).

## Discussão

Em outros trabalhos, a criação de populações sintéticas brasileiras foi aplicada principalmente em estudos de transporte e de geração de dados (Quadro 1). Entre esses estudos, a variável raça, que é importante para a identificação de minorias entre os municípios e consequente análise de situações relacionadas à pobreza e saúde, é pouco explorada, aparecendo apenas no trabalho de Furtado (2020). Não encontramos populações sintéticas, ao nível do indivíduo, aplicadas a pesquisas sobre desigualdades e determinantes sociais na saúde.

**QUADRO 1**
**Populações sintéticas brasileiras em outros estudos**

| Autores | População do estudo | Aplicação | Variável raça |
|---|---|---|---|
| Pianucci *et al.* (2019) | São Carlos, SP | Transporte | Não |
| Sallard, Balać e Hörl (2020) | Região Metropolitana de São Paulo (RMSP) | Transporte | Não |
| Ajauskas e Strambi (2024) | Região Metropolitana de São Paulo (RMSP) | Transporte | Não |
| Furtado (2020) | 46 cidades | Geração de dados para pesquisas para outras pesquisas | Sim |
| Ton *et al.* (2024) | 90 países incluindo o Brasil | Geração de dados para pesquisas para outras pesquisas | Não |

Fonte: Elaboração dos autores.

Identificar com maior precisão a presença de minorias entre os municípios brasileiros é relevante para avaliar a segregação e a aplicação de políticas públicas mais igualitárias. A criação da população sintética brasileira que propomos agrega os dados das tabelas de contingência do Censo de modo a respeitar os totais e as proporções existentes entre as variáveis ao nível municipal, gerando dessa forma um conjunto de dados sintéticos que não ignoram a presença das minorias divulgadas pelo Censo de 2010.

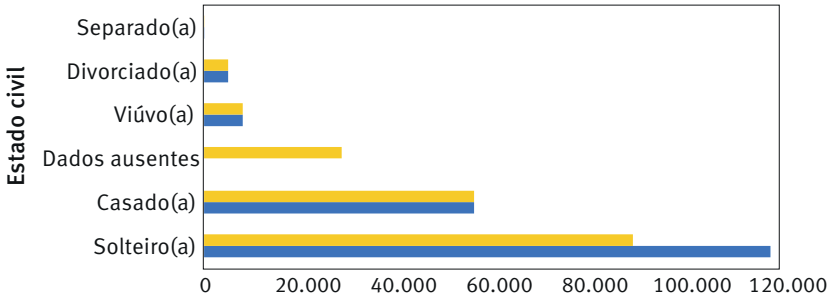
Entre as limitações de nossa proposta, destacamos quatro tópicos principais. O primeiro é que criamos indivíduos sintéticos a partir de somas marginais e ajustes pontuais. Para combinar as tabelas de contingência do Censo, assumimos independência condicional entre algumas variáveis, sendo que esta suposição pode não ser atendida na população real (e.g. Tabela 2). De fato, é provável que existam correlações dentro dessas variáveis que não podemos observar apenas a partir dos dados agregados disponíveis. Isso significa que os resultados que relacionam variáveis específicas entre si podem não ser confiáveis e, portanto, não devem ser usados para investigações específicas de relações entre variáveis independentes e dependentes. O segundo é que, devido à ausência de informações no Censo de 2010, inferimos que indivíduos sintéticos com menos de dez anos são solteiros e sem renda. O terceiro refere-se ao fato de que, apesar de mantermos as distribuições municipais de valores das variáveis estado civil e renda similar aos dados do Censo, atribuímos por sorteio esses valores aos indivíduos sintéticos. A diferença entre indivíduos nos microdados expandidos e os indivíduos da população sintética brasileira é que não existe entre eles uma relação de um-para-um (i.e., cada indivíduo que aparece na população sintética também aparece nos microdados); por outro lado, ao alterarmos as sementes estatísticas, criamos populações sintéticas com indivíduos sintéticos diferentes entre os municípios e sem, com isso, alterar a importância estatística das populações. No quarto tópico, destacamos que o esforço computacional (i.e., tempo de processamento e diversidade de dados a processar) para agregar os dados nas tabelas de contingência é maior que o esforço para agregar os

microdados. Para reduzir esse esforço, adequamos nosso método a práticas de programação que minimizam os tempos de processamento e reduzem os volumes processados (ver Krijkamp *et al.*, 2018; Belov; Tatarintsev; Nikulchev, 2021). Nosso método, à medida que inclui novas variáveis, também aumenta o esforço computacional. Por outro lado, os dados agregados conforme nossa proposta (Tabela 4) geram menor volume de dados do que os microdados. Indicamos a aplicação desse método a modelagens com poucas variáveis ou que estejam aparelhadas para um esforço computacional maior. Para análises inferenciais, recomendamos que os dados dos indivíduos (e.g. sexo, raça) sejam primeiro agregados por município, dado que a população sintética brasileira procura manter os totais das variáveis similares aos dados apresentados pelas tabelas de contingência do Censo de 2010. Ao nível do indivíduo, recomendamos o uso da população sintética para estudos que simulem modificações em uma variável específica (e.g. sexo, raça), visto que o total da variável, ao nível municipal, é similar aos apresentados pelas tabelas de contingência.

Para o Censo de 2010, o IBGE optou por registrar somente as nupcialidades e os rendimentos de pessoas com dez anos ou mais de idade (IBGE, 2013). A ausência destas informações na coleta de dados sobre indivíduos entre zero e nove anos foi interpretada como dados faltantes (i.e., NA) nos registros dos microdados. Na população sintética, inferimos que os indivíduos com idade inferior a dez anos têm "estado civil solteiro" e "rendimento zero". Esta discrepância pode ser observada na comparação do número de pessoas solteiras na Figura 4.

A construção de uma população sintética passa por diferentes etapas, entre elas a seleção das variáveis utilizadas para caracterizar a população. Nem sempre é possível retirar da fonte uma informação em sua plenitude, por isso, a imputação de valores ausentes pode ser necessária para completar dados de uma população. Conhecer o processo de construção da população sintética e suas variáveis evita, por exemplo, concluir que pessoas com menos de dez não têm renda. A população sintética é apenas um dos estágios iniciais de um processo de análise maior em estudos populacionais; como tal, pode ser usada como uma fonte de dados para previsão e microssimulação, sendo, por isso, importante alinhar as variáveis que caracterizam uma população sintética às variáveis que se pretende estudar.

**FIGURA 4**
**Comparação entre os estados civis na população sintética e nos microdados expandidos**




| | Solteiro(a) | Casado(a) | Dados ausentes | Viúvo(a) | Divorciado(a) | Separado(a) |
|---|---|---|---|---|---|---|
| Microdados expandidos | 89.653 | 56.435 | 28.775 | 8.063 | 5.019 | 2 |
| População sintética | 118.428 | 56.435 | 0 | 8.063 | 5.019 | 2 |

Fonte: Instituto Brasileiro de Geografia e Estatística – IBGE. Elaboração dos autores.

Nosso empenho, ao criarmos uma população sintética brasileira com os dados do Censo de 2010, foi garantir que não houvesse dados faltantes (i.e., NA) entre os atributos dos indivíduos e que os totais e subtotais das variáveis deste estudo estivessem alinhados às tabelas de contingência do Censo ao nível municipal. Nossa abordagem se diferencia das propostas baseadas na expansão dos microdados do Censo, ao apresentarmos uma população sintética sem dados faltantes e incluindo minorias que não aparecem nos microdados.

As tabelas de contingência e os microdados do Censo de 2010 foram projetados para atender a diferentes propósitos. As tabelas de contingência apresentam dados agregados por diferentes subconjuntos de variáveis e são úteis para a extração rápida (i.e., sem a necessidade de processamento de dados) de informações específicas da população (e.g. total de habitantes estratificados por sexo). Os microdados geram informações através de processamento de dados e permitem analisar suas informações a partir de diferentes combinações de variáveis (e.g. sexo e raça, renda e escolaridade). Ao agregarmos as tabelas de contingência e compararmos esses dados com os dados gerados pelos microdados, a depender do estrato analisado, diferenças sutis entre eles podem aparecer. Demostramos neste trabalho que algumas minorias estratificadas por raça, que aparecem nas tabelas de contingência, podem não aparecer nos microdados. Não investigamos o que pode ter motivado essas diferenças, mas conjecturamos que diferentes fatores podem ter contribuído para o caso, tais como: a amostra utilizada nos microdados deixou de contemplar algumas minorias; os critérios usados na descaracterização de dados privados dos microdados foram diferentes dos usados nas tabelas de contingências; etc.

É importante para o planejamento de políticas públicas de proteção e saúde a identificação de minorias em situação de vulnerabilidade entre os municípios brasileiros. Os microdados do Censo de 2010, por exemplo, não reportam a presença de indígenas na cidade de Colniza (código de município IBGE 5103205) no Mato Grosso. O município de

Colniza abriga as Terras Indígenas de Piripkura, Arara do Rio Branco e dos Kawahivas (um povo que vive em isolamento) e são reportados conflitos entre madeireiros e indígenas na região (PNUD; IPEA; FJP, 2013; MT tem três terras indígenas sob ameaça..., 2022). Embora os microdados não reportem a presença de indígenas no município, a tabela de contingência do Censo de 2010 (Tabela 4.25 – População residente, por situação do domicílio e cor ou raça, segundo os municípios – Mato Grosso – 2010) reporta, mas com dados agregados por tipo de domicílio e raça. Para identificar mais variáveis da população (e.g. sexo e idade) entre as tabelas de contingência do Censo, são necessárias outras consultas. O método de criação de populações sintéticas que propomos se alinha aos dados fornecidos pelas tabelas de contingência do Censo de 2010 e, ao integrar variáveis de diferentes tabelas, cria uma base de dados centralizada que possibilita que outros estudos (e.g. simulação e projeção) avaliem as populações municipais ao nível dos indivíduos e a partir de diferentes subconjuntos de variáveis.

Nossa contribuição para os estudos sobre a população brasileira foi apresentar uma população sintética para todos os municípios brasileiros, que pode ser criada a partir de uma fonte de informação com menor volume de dados e que inclui na variável raça minorias a nível municipal. A criação de uma população sintética é útil em diferentes contextos (e.g. preservação de dados privados e recuperação de informações populacionais pela agregação de dados). Embora os microdados do Censo de 2022 ainda não tenham sido publicados, nossa abordagem de recuperação de informações populacionais pela agregação de dados é uma solução que pode ser aplicada a qualquer Censo que tenha valores discrepantes entre as tabelas de contingência e os microdados.

## Considerações finais

O processo de construção de uma população sintética pode admitir diferentes escolhas e suposições. Como ela pode ser adaptada conforme as necessidades de cada estudo, compreender o processo de sua construção evita a criação de viés nos dados. Nossa proposta preserva os totais e proporções entre as variáveis e mantém os valores semelhantes aos dados populacionais do Censo. Além disso, também permite que as variáveis de um estudo sejam combinadas de diferentes maneiras e que novas variáveis sejam incluídas. Assim, entendemos que a população sintética brasileira aqui criada deve ser utilizada preferencialmente para evitar a sub-representação de alguns estratos populacionais.

Dados ao nível dos indivíduos podem ser inestimáveis para estudar desigualdades e determinantes da saúde da população. A criação de dados sintéticos ao nível do indivíduo permite reunir informações de diferentes setores (como informações demográficas, condições socioeconômicas, recebimento de políticas e resultados de saúde) e servir de fonte de informação para estudos sobre projeções populacionais e microssimulação.

A principal motivação desse trabalho resulta do fato de que os estudos sobre microssimulação são pouco explorados em investigações sobre a saúde da população brasileira.

As microssimulações em saúde têm sido intensamente desenvolvidas em países como Canadá, Estados Unidos e alguns países europeus, e são menos produzidas em países como Brasil e Argentina (Schofield *et al.*, 2018). Em relação à microssimulação, notamos uma carência de estudos sobre populações sintéticas brasileiras ao nível do indivíduo. Embora seja possível realizar microssimulações a partir de taxas e índices municipais (e.g. Rasella *et al.*, 2018), trabalhar com simulações a partir de indivíduos permite explorar mais profundamente as relações entre as características da população (e.g., raça, escolaridade, renda), transformando a população sintética brasileira em uma fonte de dados para análises de políticas públicas, predições populacionais e criação de cenários para estudos sobre a saúde.

Dados incompletos ou faltantes da população podem caracterizar um cenário populacional diferente da realidade e levar os gestores municipais a tomarem decisões equivocadas. A população sintética brasileira que apresentamos tem valores similares aos mostrados pelas tabelas de contingência do Censo de 2010 e mais completos (i.e. sem dados faltantes) que os apresentados pelos microdados do Censo de 2010. No entanto, apesar de a população sintética brasileira apresentar dados ao nível dos indivíduos, não desenvolvemos métodos para extrair informações ou análises a partir desses dados sintéticos.

Em aplicações futuras, pretendemos usar a população sintética brasileira para extrair, por meio de análises de grupamentos, informações sobre desigualdades sociais aos níveis municipais e estaduais. Além disso, pretendemos complementar a população sintética brasileira de 2010 com informações sobre mortalidade, natalidade e migração para projetar a população sintética brasileira para anos posteriores. Ao projetarmos a população sintética brasileira a partir de 2010, criaremos uma estrutura de dados que permitirão a realização de simulações a partir de diferentes extratos da população brasileira e a geração de informações que poderão subsidiar os processos de decisão dos gestores de políticas públicas.

## Referências

AJAUSKAS, R.; STRAMBI, O. Procedimento para geração de populações sintéticas com base em dados disponíveis no Brasil. **Transportes**, v. 32, n. 3, e2617, 2024.

ARORA, A.; WAGNER, S. K.; CARPENTER, R.; JENA, R.; KEANE, P. A. The urgent need to accelerate synthetic data privacy frameworks for medical research. **The Lancet Digit Health**, v. 7, n. 2, E157-E160, 2025.

BECKMAN, R. J.; BAGGERLY, K. A.; MCKAY, M. D. Creating synthetic baseline populations. **Transportation Research Part A: Policy and Practice**, v. 30, n. 6, p. 415-429, 1996.

BELOV, V.; TATARINTSEV, A.; NIKULCHEV, E. Choosing a Data storage format in the Apache Hadoop system based on experimental evaluation using Apache Spark. **Symmetry**, v. 13, Article 195, 2021.

BISSETT, K. R.; CADENA, J.; KHAN, M.; KUHLMAN, C. J. Agent based computational epidemiological modeling. **Journal of Indian Institute of Science,** v. 101, n. 3, p. 303-307, 2021.

CHAPUIS, K.; TAILLANDIER, P.; DROGOUL, A. Generation of synthetic populations in social simulations: a review of methods and practices. **Journal of Artificial Societies and Social Simulation**, v. 25, n. 2, Article 6, 2022.

DUARTE, L. T.; SILVA, D. B. D. N.; BRITO, J. A. D. M. Análise de paradados do Censo Demográfico 2010: uma investigação de fatores associados a erros não amostrais do levantamento de dados. **Revista Brasileira de Estudos de População**, v. 33, n. 3, p. 679-701, 2016.

ETUMAN, A. E.; BENOUSSAÏD, T.; CHARREIRE, H.; COLL, I. OLYMPUS-POPGEN: a synthetic population generation model to represent urban populations for assessing exposure to air quality. **PloS One**, v. 19, n. 3, Article e0299383, 2024.

FAROOQ, B.; BIERLAIRE, M.; HURTUBIA, R.; FLÖTTERÖD, G. Simulation based population synthesis. **Transportation Research Part B: Methodological**, v. 58, p. 243-263, 2013.

FELBERMAIR, S.; LAMMER, F.; TRAUSINGER-BINDER, E.; HEBENSTREIT, C. Generating synthetic population with activity chains as agent-based model input using statistical raster census data. **Procedia Computer Science**, v. 170, p. 273-280, 2020.

FURTADO, B. A. **Gerando famílias artificiais intraurbanas**: Censo 2010. Brasília: Ipea, 2020 (Nota Técnica, n. 78).

GADOTTI, A.; ROCHER, L.; HOUSSIAU, F.; CREŢU, A.; MONTJOYE, Y. Anonymization: the imperfect science of using data while preserving privacy. **Science Advances**, v. 10, n. 29, Article eadn7053, 2024.

GUNSALUS, P. R.; ROSE, J.; LEHR, C. J.; VALAPOUR, M.; DALTON, J. E. Creating synthetic populations in transplantation: a Bayesian approach enabling simulation without registry resampling. **PloS One**, 2024.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Metodologia do Censo Demográfico 2010**. Rio de Janeiro: IBGE, 2013.

IBGE – Instituto Brasileiro de Geografia e Estatística. Sistema IBGE de Recuperação Automática – Sidra, 2024. Disponível em: https://sidra.ibge.gov.br/home/pms/brasil. Acesso em: 01 set. 2024.

IBGE-DOWNLOADS. **IBGE – Downloads**, 2024. Disponível em: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html. Acesso em: 28 set. 2024.

IBGE-MICRODADOS. **IBGE – Censo demográfico**, 2024. Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?=&t=microdados. Acesso em: 28 set. 2024.

JIANG, N.; CROOKS, A. T.; KAVAK, H.; BURGER, A.; KENNEDY, W. G. A method to create a synthetic population with social networks for geographically-explicit agent-based models. **Computational Urban Science**, v. 2, n. 7, 2022.

KONDURI, K.; YOU, D.; GARIKAPATI, V.; PENDYALA, R. Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions. **Transportation Research Record**, v. 2563, n. 1, p. 40-50, 2016.

KRIJKAMP, E. M.; ALARID-ESCUDERO, F.; ENNS, E. A.; JALAL, H. J.; HUNINK, M. G. M.; PECHLIVANOGLOU, P. Microsimulation modeling for health decision sciences using R: a tutorial. **Medical Decision Making**, v. 38. n. 2, p. 400-422, 2018.

LEYK, S.; GAUGHAN, A. E.; ADAMO, S. B.; SHERBININ, A. de; BALK, D.; FREIRE, S.; ROSE, A.; STEVENS, F. R.; BLANKESPOOR, B.; FRYE, C.; COMENETZ, J.; SORICHETTA, A.; MACMANUS, K.; PISTOLESI, L.; LEVY, M.; TATEM, A. J.; PESARESI, M. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. **Earth System Science Data**, v. 11, p. 1385-1409, 2019.

LI, J.; VIDYATTAMA, Y. Projecting spatial population and labour force growth in Australian districts. **Journal of Population Research**, v. 36, p. 205-232, 2019.

MT tem três terras indígenas sob ameaça de madeireiros e grileiros vigiadas pela Força Nacional. **G1**. 10 de janeiro de 2022. Disponível em: https://g1.globo.com/mt/mato-grosso/noticia/2022/02/10/mt-tem-tres-terras-indigenas-sob-ameaca-de-madeireiros-e-grileiros-vigiadas-pela-forca-nacional.ghtml. Acesso em: 31 jul. 2025.

NADAL, J. O.; SANZ, G. C.; RIBAS, I. F.; MONDELO, P. R. Determining occupational accidents baseline ratios by considering a synthetic population: the case of Spain. **PloS One**, v. 18, n. 11, Article e0294707, 2023.

NICOLAIE, M. A.; FÜSSENICH, K.; AMELING, C.; BOSHUIZEN, H. C. Constructing synthetic populations in the age of big data. **Population Health Metrics**, n. 21, 2023.

NOWOK, B.; RAAB, G. M.; DIBBEN, C. synthpop: bespoke creation of synthetic data in R. **Journal of Statistical Software**, v. 74, n. 11, 2016.

PIANUCCI, M.; PITOMBO, C.; CUNHA, A.; LIMA, P. Previsão da demanda por viagens domiciliares através de método sequencial baseado em população sintética e redes neurais artificiais. **Transportes**, v. 27, 2019.

PNUD; IPEA; FJP. **Atlas do desenvolvimento humano no Brasil 2013**. PNUD Brasil, 2025. Disponível em: http://www.atlasbrasil.org.br/perfil/municipio/510325 #:~:text=De%20acordo%20com%20as%20estimativas,maioria%2C%20por%20homens%20e%20negros%20. Acesso em: 31 jul. 2025.

PRÉDHUMEAU, P.; MANLEY, E. A synthetic population for agent-based modelling in Canada. **Scientific Data**, v. 10, n. 148, 2023.

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna, Austria: R Foundation, 2024. Disponível em: https://www.r-project.org/. Acesso em: 02 ago. 2025.

RAGHUNATHAN, T. E. Synthetic data. **Annual Review Statistics and its Application**, v. 8, n. 129, p. 129-140, 2021.

RASELLA, D.; BASU, S.; HONE, T.; PAES-SOUSA, R.; OCKÉ-REIS, C. O.; MILLETT, C. Child morbidity and mortality associated with alternative policy responses to the economic crisis in Brazil: a nationwide microsimulation study. **Plos Medicine**, v.15, n. 5, Article e1002570, 2018.

POSIT TEAM. **RStudio: Integrated Development Environment for R**. Boston, MA: Posit Software, PBC, 2025. Disponível em: http://www.posit.co/. Acesso em: 02 ago. 2025.

SALLARD, A.; BALAĆ, M.; HÖRL, S. A synthetic population for the greater São Paulo metropolitan region. **Arbeitsberichte Verkehrs-und Raumplanung,** v. 1545, 2020.

SCHOFIELD, D.; ZEPPEL, M.; TAN, O.; LYMER, S.; CUNICH, M.; SHRESTHA, R. A brief, global history of microsimulation models in health: past applications, lessons learned and future directions. **International Journal of Microsimulation**, v. 11, n. 1, p. 97-142, 2018.

SOUZA-JUNIOR, C. T. D. Population. **GitHub repository**. Disponível em: https://github.com/Cleonidas-Tavares/Population. Acesso em: 16 nov. 2024. .

TEMPL, M.; KOWARIK, A.; MEINDL, B. Statistical disclosure control for micro-data using the R Package sdcMicro. **Journal of Statistical Software**, v. 67, n. 4, 2015.

TON, M. J.; INGELS, M. W.; DE BRUIJN, J. A.; DE MOEL, H.; REIMANN, L.; BOTZEN, W. J. W.; AERTS, J. C. J. H. A global dataset of 7 billion individuals with socio-economic characteristics. **Scientific Data**, v. 11, Article 1096, 2024.

TOZLUOĞLU, Ç.; DHAMAL, S.; YEH, S.; SPREI, F.; LIAO, Y.; MARATHE, M.; BARRETT, C. L.; DUBHASHI, D. A synthetic population of Sweden: datasets of agents, households, and activity-travel patterns. **Data in Brief**, v. 48, 2023.

VOAS, D.; WILLIAMSON, P. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. **International Journal of Population Geography**, v. 6, p. 349-366, 2000.

YAMÉOGO, B.; GASTINEAU, P.; HANKACH, P.; VANDANJON, P. Comparing methods for generating a two-layered synthetic population. **Transportation Research Record**, v. 2675, p. 136-147, 2020.

YE, X.; KONDURI, K.; PENDYALA, R.; SANA, B.; WADDELL. P. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. *In*: 88th ANNUAL MEETING OF THE TRANSPORTATION RESEARCH BOARD. **Proceedings** [...]. 2009. Washington, DC: [s.n.], 2009.

ZHANG, J. L.; BRYANT, J.; NISSEN, K. Bayesian small area demography. **Survey Methodology**, v. 45, n. 1, p. 13-29, 2019.

ZHU, K.; YIN, L.; LIU, K.; LIU, J.; SHI, Y.; LI, X.; ZOU, H.; DU, H. Generating synthetic population for simulating the spatiotemporal dynamics of epidemics. **Plos Computational Biology**, v. 20, Article e1011810, 2024.

## Sobre os autores

*Cleônidas Tavares de Souza Junior* é doutor em Modelagem Computacional de Sistemas Cognitivos pelo *Campus* Integrado de Manufatura e Tecnologia (Senai-Cimatec).

*Desmond Campbell* é doutor em Genética Estatística e mestre em Neurociência pelo King's College London, formado em Sistemas Eletrônicos pela Universidade de Ulster. Pesquisador associado na Escola de Saúde e Bem-Estar da Universidade de Glasgow.

*Srinivasa Vittal Katikireddi* é doutor e mestre em Saúde Pública pela Universidade de Glasgow e graduado em Medicina e Genética na Universidade de Edimburgo. Professor de Saúde Pública e Desigualdades em Saúde na Universidade de Edimburgo.

*Paulo Victor Maciel da Costa* é doutor em Demografia pelo Programa de Pós-Graduação em Demografia (PPGDem) da Universidade Federal do Rio Grande do Norte (UFRN).

*Gervásio Ferreira dos Santos* é doutor em Economia pela Universidade de São Paulo (USP).

*Mauricio Lima Barreto* é doutor em Epidemiologia pela London School of Hygiene & Tropical Medicine, mestre em Saúde Comunitária pela Universidade Federal da Bahia (UFBA) e médico pela UFBA.

*Roberto Fernandes Silva Andrade* é doutor em Física pela Universidade de Regensburg (Alemanha).

## Endereço para correspondência

*Cleônidas Tavares de Souza Junior*
Parque Tecnológico do Edifício Tecnocentro
Rua Mundo, 121, sala 315, Trobogy
41745-715 – Salvador-BA, Brasil

*Desmond Campbell*
Clarice Pears Building
90 Byres Road
G12 8TB – Glasgow, Reino Unido

*Srinivasa Vittal Katikireddi*
Clarice Pears Building
90 Byres Road
G12 8TB – Glasgow, Reino Unido

*Paulo Victor Maciel da Costa*
Parque Tecnológico do Edifício Tecnocentro
Rua Mundo, 121, sala 315, Trobogy
41745-715 – Salvador-BA, Brasil

*Gervásio Ferreira dos Santos*
Parque Tecnológico do Edifício Tecnocentro
Rua Mundo, 121, sala 315, Trobogy
41745-715 – Salvador-BA, Brasil

*Mauricio Lima Barreto*
Parque Tecnológico do Edifício Tecnocentro
Rua Mundo, 121, sala 315, Trobogy
41745-715 – Salvador-BA, Brasil

*Roberto Fernandes Silva Andrade*
Parque Tecnológico do Edifício Tecnocentro
Rua Mundo, 121, sala 315, Trobogy
41745-715 – Salvador-BA, Brasil

**CRediT**

Reconhecimentos: Não aplicável.
Financiamento: Esta pesquisa foi financiada pelo NIHR (NIHR134801) usando financiamento de desenvolvimento internacional do Reino Unido do governo do Reino Unido para apoiar a pesquisa em saúde global. As opiniões expressas nesta publicação são as dos autores e não necessariamente as do NIHR ou do governo do Reino Unido. O DDC e o SVK também reconhecem o financiamento do European Research Council (949582), Medical Research Council (MC_UU_12017_2) e Scottish Government Chief Scientist Office (SPHSU17).
Conflitos de interesse: Os autores certificam que não têm interesse pessoal, comercial, acadêmico, político ou financeiro que represente um conflito de interesses em relação ao manuscrito.
Aprovação ética: Os autores certificam que o trabalho não inclui seres humanos ou animais.
Disponibilidade de dados e material: os conteúdos já estão disponíveis.

Contribuições dos autores:

Cleônidas Tavares de Souza Junior: conceitualização; curadoria de dados; análise formal; investigação; *software*; validação; visualização; escrita – rascunho original; escrita – revisão e edição.

Desmond Campbell: conceitualização; análise formal; investigação; visualização; escrita – rascunho original; escrita – revisão e edição.

Srinivasa Vittal Katikireddi: conceitualização; análise formal; investigação; visualização; escrita – rascunho original; escrita – revisão e edição.

Paulo Victor Maciel da Costa: análise formal; investigação; validação; escrita – rascunho original.

Gervásio Ferreira dos Santos: análise formal; investigação; escrita – rascunho original.

Mauricio Lima Barreto: análise formal; investigação; escrita – rascunho original.

Roberto Fernandes Silva Andrade: conceitualização; análise formal; investigação; validação; visualização; escrita – rascunho original; escrita – revisão e edição.

Editores: Bernardo Lanza Queiroz, Júlia Almeida Calazans e Maria Carolina Tomás

## Abstract

*Developing a synthetic Brazilian population derived from the 2010 Census*

The 2010 Brazilian Census contains a wealth of information that could enable research and inform policies in health, education, the economy, and other sectors. The census provides publicly available information in two forms. Firstly, contingency tables are available at the municipal level, for strata defined by race, gender, and education. Secondly, microdata with personal information. To preserve individual anonymity in the data, the census collapsed some variables into broader categories and removed personally identifiable data. The data composition strategies of the contingency tables and the microdata are different and, when comparing samples of both data, we find that the race variable in the microdata ignores the presence of minorities in some municipalities. This suggests that synthetic populations based on the 2010 Census should be created using the contingency tables. Our evaluation shows that the so created synthetic population maintains the values and proportions of the contingency tables and presents totals close to those of the microdata.

**Keywords**: Population. Cohort analysis. Computer simulation. Statistical inference.

**Resumen**

*Desarrollo de una población brasileña sintética derivada del Censo de 2010*

El Censo Brasileño de 2010 contiene una gran cantidad de información que puede facilitar la investigación y apoyar el desarrollo de políticas en áreas como salud, educación, economía y otros sectores. El censo proporciona información disponible públicamente en dos formas. En primer lugar, existen tablas de contingencia disponibles a nivel municipal, para estratos definidos por raza, género y nivel educativo. En segundo lugar, se encuentran los microdatos, que contienen información personal. Para preservar el anonimato de los individuos, el censo agrupó algunas variables en categorías más amplias y eliminó los datos de identificación personal. Las estrategias de composición de datos en las tablas de contingencia y en los microdatos son diferentes y, al comparar muestras de ambos conjuntos, encontramos que la variable "raza" en los microdatos omite la presencia de minorías en algunos municipios. Esto sugiere que se deberían crear poblaciones sintéticas basadas en las tablas de contingencia del Censo de 2010. Nuestra evaluación muestra que la población sintética así creada conserva los valores y proporciones presentes en las tablas de contingencia, y presenta totales cercanos a los observados en los microdatos.

**Palabras clave**: Población. Análisis de cohorte. Simulación por computadora. Inferencia estadística.

Recebido para publicação em 20/05/2025
Aceito para publicação em 08/08/2025