Nota Técnica

GERANDO FAMÍLIAS ARTIFICIAIS INTRAURBANAS: CENSO 2010

Bernardo Alves Furtado

Nº 78

Diset
Diretoria de Estudos e Políticas
Setoriais de Inovação e Infraestrutura

Novembro de 2020

ipea

Nota Técnica

GERANDO FAMÍLIAS ARTIFICIAIS INTRAURBANAS: CENSO 2010

Bernardo Alves Furtado

Nº 78

Diset
Diretoria de Estudos e Políticas
Setoriais de Inovação e Infraestrutura

ipea

Governo Federal

Ministério da Economia
Ministro Paulo Guedes

ipea **Instituto de Pesquisa Econômica Aplicada**

Fundação pública vinculada ao Ministério da Economia, o Ipea fornece suporte técnico e institucional às ações governamentais – possibilitando a formulação de inúmeras políticas públicas e programas de desenvolvimento brasileiros – e disponibiliza, para a sociedade, pesquisas e estudos realizados por seus técnicos.

Presidente
Carlos von Doellinger

Diretor de Desenvolvimento Institucional
Manoel Rodrigues Junior

Diretora de Estudos e Políticas do Estado, das Instituições e da Democracia
Flávia de Holanda Schmidt

Diretor de Estudos e Políticas Macroeconômicas
José Ronaldo de Castro Souza Júnior

Diretor de Estudos e Políticas Regionais, Urbanas e Ambientais
Nilo Luiz Saccaro Júnior

Diretor de Estudos e Políticas Setoriais de Inovação e Infraestrutura
André Tortato Rauen

Diretora de Estudos e Políticas Sociais
Lenita Maria Turchi

Diretor de Estudos e Relações Econômicas e Políticas Internacionais
Ivan Tiago Machado Oliveira

Assessor-chefe de Imprensa e Comunicação (substituto)
João Cláudio Garcia Rodrigues Lima

Ouvidoria: http://www.ipea.gov.br/ouvidoria
URL: http://www.ipea.gov.br

Nota Técnica

GERANDO FAMÍLIAS ARTIFICIAIS INTRAURBANAS: CENSO 2010

Bernardo Alves Furtado

Nº 78

Diset
Diretoria de Estudos e Políticas
Setoriais de Inovação e Infraestrutura

Novembro de 2020

ipea

EQUIPE TÉCNICA

Bernardo Alves Furtado
Técnico de planejamento e pesquisa na Diretoria de Estudos e Políticas Setoriais de Inovação e Infraestrutura (Diset) do Ipea; e bolsista produtividade do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

DOI: http://dx.doi.org/10.38116/ntdiset78

SUMÁRIO

1 INTRODUÇÃO

Esta *Nota Técnica* descreve o processo computacional de geração de indivíduos e famílias artificiais, a partir de dados do censo demográfico de 2010. As famílias geradas podem ser utilizadas em processos de simulação e modelagem mantendo a proporção e os detalhes da população de interesse, sem necessidade de manipulação direta dos dados do Instituto Brasileiro de Geografia e Estatística (IBGE). Adicionalmente, o processo computacional inclui ainda o *download* do arquivo de limites geoespaciais (em formato *shapefile*) de todos os setores censitários e sua junção em áreas de ponderação (APs). O *site* oficial que disponibiliza esses dados no IBGE não contém todos os municípios e regiões metropolitanas. Finalmente, informações de outras tabelas não presentes neste estudo podem utilizar os processos aqui descritos de forma similar.

Os resultados do processo computacional incluem indivíduos, suas características e a composição de quais indivíduos pertencem à mesma família. A respeito de cada membro da família é possível identificar: *i)* a AP de residência; *ii)* o gênero; *iii)* a idade; *iv)* os anos de instrução; *v)* a cor; e *vi)* o salário. O tamanho médio das famílias segue os padrões observados nos setores censitários.

O repositório – escrito em linguagem de programação Python – está disponível em: <https://github.com/BAFurtado/censo2010>. Adicionalmente, a geração de famílias foi utilizada em uma simulação de violência doméstica,[1] cujo código também está disponível de forma aberta em: <https://github.com/BAFurtado/home_violence>.

2 MÉTODO

O método utilizado segue duas etapas, detalhadas em seguida. Primeiro, localizam-se os dados na base oficial do IBGE. Em seguida, ocorre o processo gerador das famílias artificiais.

2.1 Dados

A extração de dados oficiais do IBGE segue a lista de tarefas, de forma automática.

1) *Download do site* ftp do IBGE dos arquivos zipados de cada Unidade da Federação (UF) referentes ao censo de 2010, com as informações tanto dos setores censitários quanto dos microdados dos resultados gerais da amostra.
2) Extração dos dados em pastas e subpastas.
3) Identificação e localização dos arquivos de interesse, nos setores e nos microdados da amostra, referentes às informações necessárias. Inclui a identificação de variáveis específicas.
4) Leitura dos arquivos de interesse e extração das seguintes informações:

Setores censitários

a) gênero e idade: por exemplo, 134 homens de 45 anos – nas planilhas Pessoa11 e Pessoa12, de cada UF;
b) cor: branca, preta, amarela, parda, indígena, distribuídas nas variáveis V002-V006, da planilha Pessoa03; e
c) salários e tamanho médio das famílias: de acordo com as informações das planilhas Basico, variáveis V003-V004 e V009-V010.

Amostra

a) anos de instrução: a partir das informações da variável V6400.

Cada informação é processada de modo a refletir proporcionalmente a presença de cada valor em cada área de ponderação. Com isso, o resultado intermediário da execução do arquivo read_amostra.py é a gravação de quatro planilhas, com as seguintes descrições:

- código da AP (inteiro), gênero (2=homens, 1=mulheres), idade em anos e o número de pessoas com aquelas características. Tabela: num_people_age_gender_AP.csv;
- código da AP, cor e o percentual de pessoas de cada cor. Tabela: etnia_AP.csv;

1. Ver: Madeira, L. M.; Furtado, B. A.; Dill, A. *Vida*: simulando violência doméstica em tempos de quarentena. 2020. Manuscrito.

- código da AP, anos de instrução (1: sem instrução e ensino fundamental incompleto; 2: fundamental completo e médio incompleto; 3: médio completo e superior incompleto; 4: superior completo; 5: não determinado) e o percentual de pessoas naquela faixa. Tabela: quali_AP.csv; e
- código da AP, número médio de pessoas na família, variância do número médio de pessoas na família, valor do rendimento nominal médio mensal das pessoas de 10 anos ou mais de idade e variância. Tabela: average_variance_family_wages.csv.

2.2 Geração

O processo gerador de famílias artificiais pode ser invocado para cada uma das 46 áreas de concentração da população (ACPs).[2,3] Como resultado, obtêm-se indivíduos com suas características e listagem de quais indivíduos pertencem a cada família.

Além da escolha da ACP de interesse, o usuário pode escolher também o número inicial de famílias. O tamanho da população final gerada dependerá do número médio de membros por família nas APs da ACP escolhida.

O processo se inicia com a seleção de quais municípios – e, portanto, quais APs – compõem a ACP escolhida. Na sequência o número de pessoas por idade e gênero é lido para cada AP.

Os anos de instrução são transformados proporcionalmente em anos de estudo de modo aleatório. Para cada indivíduo designam-se uma cor e um salário.

Todos esses processos obedecem às proporcionalidades presentes nos dados oficiais do IBGE. No caso da designação de cor, por exemplo, sabe-se que, em 2010, havia 0.1948% de pessoas que se autodenominaram de cor branca na AP 1302603005001. Portanto, há exatamente esse percentual de probabilidade de um indivíduo receber a cor branca na população gerada artificialmente.

Para o caso de salário e número médio de membros na família, foi utilizada uma distribuição normal simples, a partir da média e do desvio-padrão obtidos nas tabelas do IBGE, por APs.

A alocação de indivíduos em família é feita de modo que crianças não fiquem em famílias sem pelo menos um adulto. Adicionalmente, dado que a construção das famílias se deu para um estudo de violência doméstica, incluem-se, na medida da disponibilidade de homens e mulheres adultos, domicílios com a presença de ambos. Entretanto, visto que o número de adultos é variável, há famílias compostas apenas por um adulto, de ambos os gêneros.

O resultado da chamada do arquivo generator.py é, portanto, uma planilha com as características por indivíduos e uma listagem de quais indivíduos pertencem a quais famílias.

2.3 *Shapefiles* APs

O IBGE disponibiliza os *shapefiles* dos setores censitários por UF. Isso significa dizer que os polígonos dos limites, traçados de forma computacional, por meio das latitudes e longitudes, estão disponíveis para representar espacialmente cada setor. As APs também estão disponíveis. Todavia, a disponibilidade restringe-se apenas a algumas APs, não incluindo, por exemplo, aquelas situadas imediatamente nos limites externos do Distrito Federal, em regiões periféricas, que, no entanto, estão incluídas na ACP de Brasília. Cristalina, no interior de Goiás, é outro exemplo de município com mais de uma AP que não está disponibilizada no *site* ftp oficial do IBGE.

Desse modo, foi necessário construir as APs, e essa construção se dá a partir da listagem disponibilizada pelo IBGE, que informa quais setores censitários estão contidos em cada AP. Dados os *shapefiles* dos setores e a listagem, é possível construir os *shapefiles* das APs.

2. ACP é a denominação escolhida pelo IBGE para caracterizar as áreas urbanas densas nas quais há movimento pendular diário. *Grosso modo*, significa dizer que é a área urbana central das metrópoles brasileiras, excluídas as porções rurais, cujo movimento pendular é de baixa magnitude. Para mais informações, ver: IBGE – Instituto Brasileiro de Geografia e Estatística. *Arranjos populacionais e concentrações urbanas do Brasil*. Rio de Janeiro: IBGE, 2015.

3. As seguintes ACPs estão disponíveis para o processo gerador: "MANAUS", "BELEM", "MACAPA", "SAO LUIS", "TERESINA", "FORTALEZA", "CRAJUBAR", "NATAL", "JOAO PESSOA", "CAMPINA GRANDE", "RECIFE", "MACEIO", "ARACAJU", "SALVADOR", "FEIRA DE SANTANA", "ILHEUS - ITABUNA", "PETROLINA - JUAZEIRO", "BELO HORIZONTE", "JUIZ DE FORA", "IPATINGA", "UBERLANDIA", VITORIA", "VOLTA REDONDA - BARRA MANSA", "RIO DE JANEIRO", "CAMPOS DOS GOYTACAZES", "SAO PAULO", "CAMPINAS", "SOROCABA", "SAO JOSE DO RIO PRETO", "SANTOS", "JUNDIAI", "SAO JOSE DOS CAMPOS", "RIBEIRAO PRETO", "CURITIBA", "LONDRINA", "MARINGA", "JOINVILLE", "FLORIANOPOLIS", "PORTO ALEGRE", "NOVO HAMBURGO - SAO LEOPOLDO", "CAXIAS DO SUL", "PELOTAS - RIO GRANDE", "CAMPO GRANDE", "CUIABA","GOIANIA", "BRASILIA".

O mesmo processo (mesmas funções) foi utilizado, qual seja: *download* dos arquivos, extração dos arquivos, identificação dos arquivos de interesse, leitura e manipulação.

Para uso neste trabalho, as APs foram restritas àquelas pertencentes às 46 ACPs listadas na nota de rodapé 3 deste texto. Caso o usuário queira montar todas as APs do Brasil, é necessário comentar o código da linha 53 do arquivo sectors_into_APs.py. Os *shapefiles* das APs, por UFs, estão disponibilizados no diretório "data" do repositório público: <https://github.com/BAFurtado/censo2010/tree/master/data/areas>.

2.4 *Caveats*

A manipulação de grande conjunto de dados leva a algumas especificidades, que, por vezes, tornam os processos menos automáticos do que deveriam ser. Ao construir este trabalho, notamos algumas inconsistências no conjunto de dados que anotamos aqui.

1) A primeira questão é exatamente a ausência de todos os *shapefiles* das APs do país disponíveis para *download*. Não foi possível identificar qual a razão da ausência.
2) A posição dos arquivos extraídos não é sempre a mesma. Algumas UFs possuem organização de diretórios distinta, provavelmente em virtude do tamanho dos dados. Com isso, não é possível utilizar a posição dos arquivos extraídos. Foi necessário, portanto, referir-se expressamente ao nome do arquivo de interesse (Pessoa11, por exemplo) para localizar os dados.
3) Dentro dos arquivos, alguns nomes de colunas estão com todas as letras maiúsculas ('CD_APONDE'), outras vezes, apenas algumas encontram-se desse modo ('CD_APonde').
4) Um arquivo continha uma coluna extra.
5) A variância média dos setores, agregados em APs, é muito alta (magnitude da ordem de 10e5); com isso, a amostragem normalizada de salários apresenta valores muito próximos entre si. Uma estimativa mais fidedigna deve incluir um número maior de famílias ou a simulação (geração de dados) repetidamente.
6) Finalmente, o "nome" dos arquivos dos setores de São Paulo utiliza o código do Rio de Janeiro ('33'). Com isso, os arquivos estavam sendo sobrescritos e os dados do interior de São Paulo não apareciam. Neste caso, foi feito o *download* de forma não automática.

3 RESULTADOS

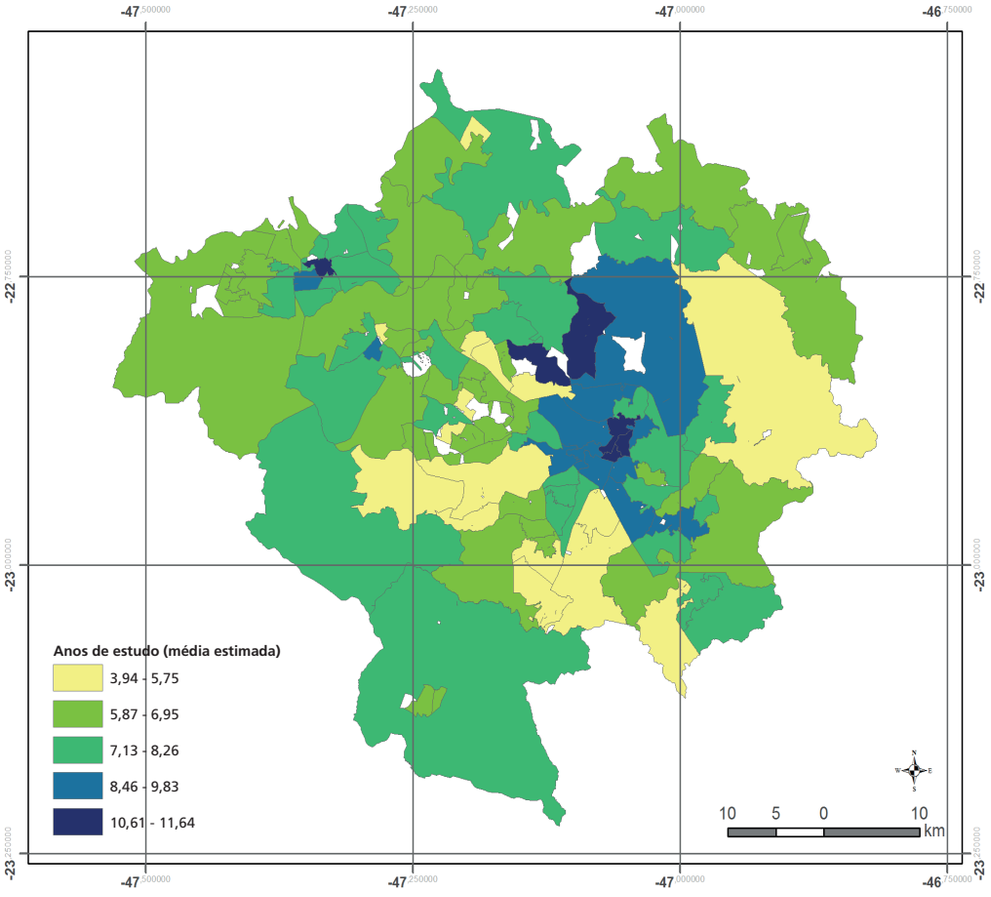
Os processos descritos nesta *Nota Técnica* geram como resultados: famílias artificiais e *shapefiles* das APs do IBGE. Ressalte-se que essa geração de famílias permite simular várias combinações de famílias possíveis. Assim, pode-se inserir variabilidade, respeitando-se as proporcionalidades dos dados, e gerar quantos conjuntos de famílias fidedignas se deseje para utilizar em outros processos de modelagem. A existência de inúmeras famílias – mantidas as características originais – permite, assim, modelar processos de Monte Carlo, pseudossignificância (por repetição), modelagem numérica e modelagem baseada em agentes, por exemplo.

A título de exemplo, ilustramos as características das famílias para a AP de Campinas.

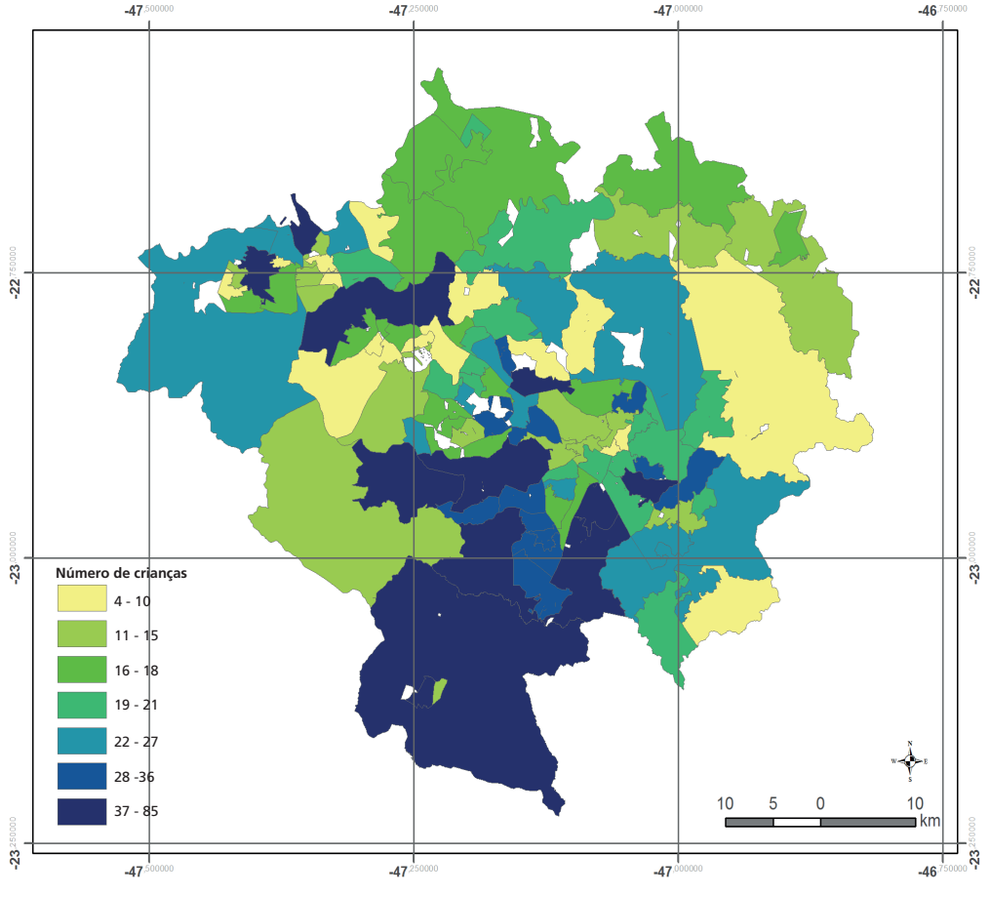
MAPA 1

AP de Campinas

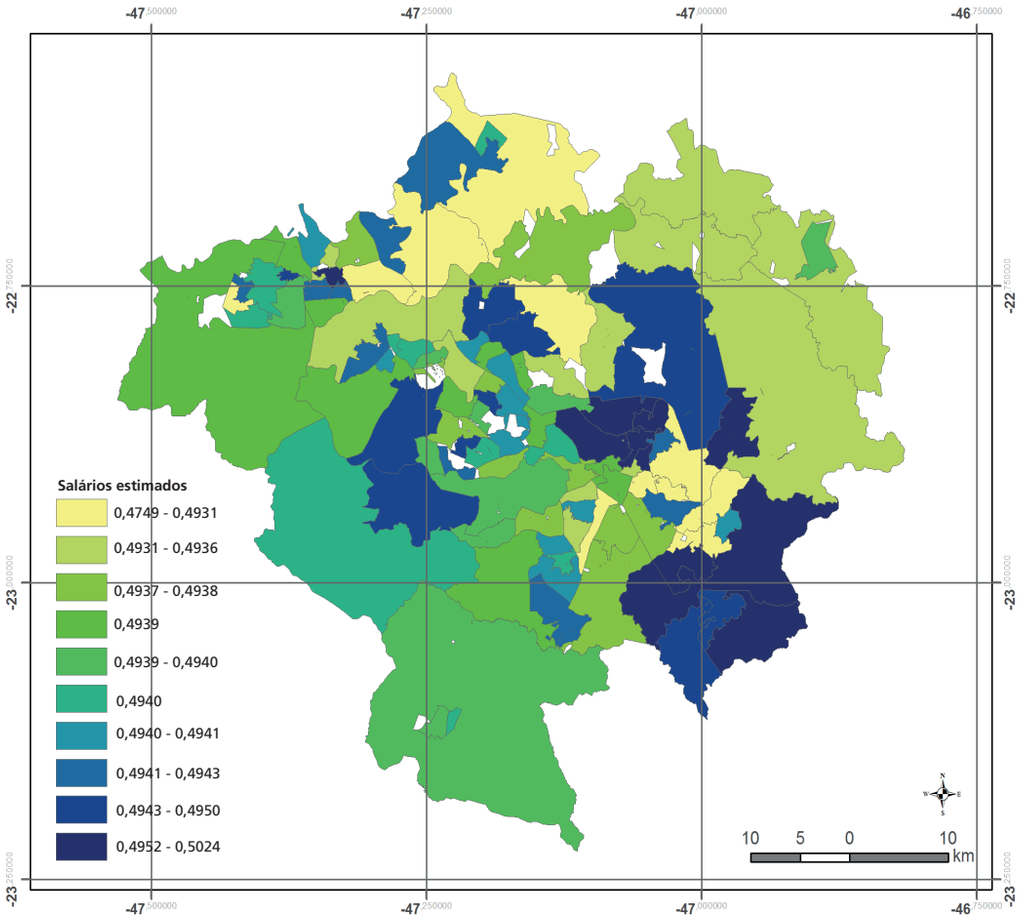
1A – Anos de estudo



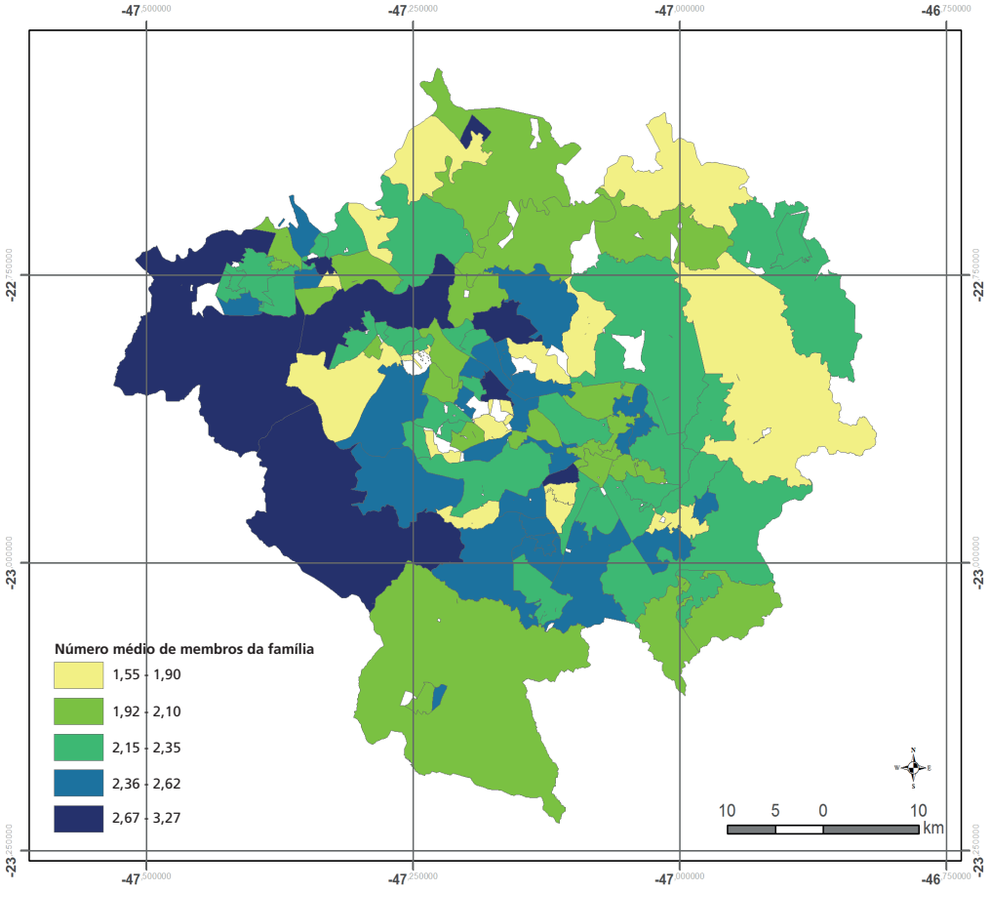
1B – Número de crianças



1C – Salários estimados



1D – Membros da família



Elaboração do autor.

Ipea – Instituto de Pesquisa Econômica Aplicada

Assessoria de Imprensa e Comunicação

EDITORIAL

Coordenação
Reginaldo da Silva Domingos

Supervisão
Carlos Henrique Santos Vianna

Revisão
Bruna Oliveira Ranquine da Rocha
Carlos Eduardo Gonçalves de Melo
Elaine Oliveira Couto
Lis Silva Hall
Mariana Silva de Lima
Marlon Magno Abreu de Carvalho
Vivian Barros Volotão Santos
Laysa Martins Barbosa Lima (estagiária)

Editoração
Aline Cristine Torres da Silva Martins
Mayana Mendes de Mattos

Capa
Danielle de Oliveira Ayres
Flaviane Dias de Sant'ana

The manuscripts in languages other than Portuguese published herein have not been proofread.

Livraria Ipea
SBS – Quadra 1 – Bloco J – Ed. BNDES, Térreo
70076-900 – Brasília – DF
Tel.: (61) 2026-5336
Correio eletrônico: livraria@ipea.gov.br

Composto em linux libertine 10/13 (texto)
Ubuntu bold (títulos, gráficos e tabelas)
Rio de Janeiro - RJ

Missão do Ipea
Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas.

ipea Instituto de Pesquisa Econômica Aplicada

MINISTÉRIO DA ECONOMIA

PÁTRIA AMADA BRASIL GOVERNO FEDERAL