RODRIGO AJAUSKAS

**PROCEDIMENTO PARA GERAÇÃO DE POPULAÇÕES SINTÉTICAS COM BASE
EM DADOS DISPONÍVEIS NO BRASIL**

SÃO PAULO

2021

RODRIGO AJAUSKAS

**PROCEDIMENTO PARA GERAÇÃO DE POPULAÇÕES SINTÉTICAS COM BASE
EM DADOS DISPONÍVEIS NO BRASIL**

Versão Corrigida

Dissertação apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Mestre em Ciências.

Área de concentração: Eng. de Transportes

Orientador: Prof. Dr. Orlando Strambi

SÃO PAULO

2021

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

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, ____ de _____ de 20__

Assinatura do autor: _____

Assinatura do orientador: _____

Catálogo-na-publicação

Ajauskas, Rodrigo

Procedimento para geração de populações sintéticas com base em dados disponíveis no Brasil / R. Ajauskas – versão corr. -- São Paulo, 2021.

144 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Transportes.

1.Planejamento de Transportes 2.Transportes Urbano 3.População
I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Transportes II.t.

RESUMO

A presente pesquisa teve como objetivos principais a disponibilização de um gerador de populações sintéticas adaptado para o Brasil e a produção de uma população sintética de domicílios e indivíduos para a Região Metropolitana de São Paulo. Modelos mais modernos de previsão da demanda por transportes exigem dados de entrada desagregados sobre a população que, apesar de serem coletados por institutos nacionais de geografia e estatística, são disponibilizados apenas de forma agregada por questões de privacidade. Assim, é necessário recriar esta população com suas respectivas características, em um processo conhecido na literatura por geração de populações sintéticas. O processo se baseia principalmente em dados disponibilizados pelo Censo, nas formas de: (i) totais marginais ou de controle para cada atributo de interesse, agregados por zonas (setores censitários); e (ii) amostras da população com informações completas, porém com menor precisão espacial, no nível de região (áreas de ponderação), a serem utilizadas como matrizes semente para a geração de uma população. Considerando as diferentes abordagens teóricas e a disponibilidade de códigos livres, foi selecionado o PopulationSim como gerador de populações sintéticas para adaptação nesta dissertação. O PopulationSim pertence à categoria de métodos de reconstrução sintética e se baseia no conceito de maximização de entropia, apresentando características de flexibilidade que o tornam vantajoso em relação a outros geradores, incluindo a possibilidade de se utilizar totais de controle inconsistentes, controles em múltiplas escalas geográficas e controlar tanto atributos de domicílios como de indivíduos. Na aplicação para a Região Metropolitana de São Paulo (RMSP), foram utilizados dados do Censo 2010 do IBGE e da Pesquisa OD 2017 de São Paulo. Foram selecionados atributos no nível dos domicílios e dos indivíduos para a geração da população. Um conjunto de códigos foi desenvolvido de forma complementar para permitir a leitura das bases de dados nacionais e a preparação dos arquivos utilizados pelo PopulationSim. Foram também desenvolvidos códigos adicionais para o procedimento de validação dos resultados. Os resultados obtidos apresentam indicadores de qualidade de ajuste satisfatórios; a validação dos geradores, porém, padece da ausência de uma base real de dados desagregados da população para comparação. Os arquivos contendo a população sintética de domicílios e indivíduos para a RMSP foram disponibilizados em um repositório onde constam também um tutorial de aplicação e os códigos desenvolvidos especificamente para aplicação no Brasil.

Palavras-chave: populações sintéticas, geradores de populações, sintetizadores de populações, modelos baseados em atividades

ABSTRACT

This research has two main objectives: (i) to adapt a population synthesizer and make it available to applications in Brazil and (ii) to provide results in the form of a synthetic population of households and individuals for the Metropolitan Area of São Paulo (RMSP). Modern travel demand forecasting models require disaggregate data about the population and its characteristics. While this data is usually collected in national censuses, they are only released as aggregate values, to preserve privacy. Therefore, the population needs to be recreated in a process known as population synthesis, based on data provided by the census in the form of (i) marginal or control totals, for each attribute of interest, and (ii) population samples, with complete information about each household and individual, although at a more aggregate spatial level, to be used as seed matrices in the synthesis process. Considering the theoretical approaches and the availability of open-source codes, the synthesizer PopulationSim was selected to be adapted to Brazil. PopulationSim is a synthetic reconstruction method that is based on the concept of entropy maximization and allows flexibility for the use of conflicting control totals, multiple geographic scales and controls both for household and persons attributes. The application of the synthesizer for generating RMSP's synthetic population used data from the 2010 national Census and from the Origin-Destination survey conducted in the region in 2017. Attributes of interest at the household and individual level were selected for the application. A package of codes was specifically developed for reading the Census and OD survey data files and preparing them in the format required by PopulationSim. Additional codes were developed to extend the validation procedures. The results present satisfactory values for the validation metrics; population synthesizers, though, lack the existence of a ground truth for the assessment of the results. The files containing the synthetic population of households and individuals for the Metropolitan Area of São Paulo are available for download at a data repository, which also includes a tutorial and the codes developed for the application of PopulationSim to any area in Brazil.

Keywords: synthetic population generation, population synthesizers, activity-based models

AGRADECIMENTOS

A maior parte do texto desta dissertação foi desenvolvida em 2020, ano marcado pelo isolamento social decorrente da pandemia da Covid-19. Foi, para mim, um ano de recolhimento e reflexão. Um ano monótono, porém de aprendizados, em que eu passei a dar um valor ainda maior às pequenas coisas da vida.

Agradeço aos meus pais e família, por todo apoio e carinho que sempre me deram.

Agradeço aos meus amigos da vida, que sempre estiveram por perto, presencialmente ou virtualmente (especialmente neste último ano de isolamento devido à pandemia). Agradeço, em especial, aos meus colegas de casa, que estiveram mais presentes do que nunca na minha vida neste ano e me propiciaram muitos momentos de alegria, apesar do isolamento.

Agradeço a Alexandre Seixas, pela compreensão da importância do desenvolvimento de uma pesquisa acadêmica e pela flexibilidade dada ao longo destes anos, o que possibilitou a realização deste trabalho.

Agradeço, em especial, ao professor Orlando Strambi. Além das conversas sobre diversos assuntos (incluindo transportes) que tivemos ao longo dos meus anos no programa de mestrado e em que eu pude aprender muito, a orientação e colaboração de Orlando foi vital para o amadurecimento e melhoria da qualidade do presente trabalho.

Agradeço aos desenvolvedores do projeto ActivitySim, sob o qual o gerador de populações sintéticas utilizado neste trabalho se encontra – e também aos inúmeros desconhecidos do fórum Stack Exchange, que ao longo da última década têm respondido a diversas questões sobre programação enfrentadas por outros usuários.

Por fim, gostaria de destacar a importância da filosofia de códigos abertos. Esta dissertação se apoiou em um projeto desenvolvido em código aberto em ambiente colaborativo. Apesar de considerar que estamos em tempos nebulosos quanto à manipulação da população através de informações falsas, acredito que a existência e acessibilidade a conteúdos de qualidade através da *internet* possa ser a chave para o desenvolvimento sustentável e redução das desigualdades nas próximas décadas. O futuro dirá.

LISTA DE ABREVIATURAS

ABM – *Activity-Based Models*

APOND – Área de Ponderação

CO – *Combinatorial Optimization*

EMR – Erro Médio Relativo

IPF – *Iterative Proportional Fitting*

IPU – *Iterative Proportional Updating*

MCMC – *Markov Chain Monte Carlo*

MMC – Método Monte Carlo

ODSP – Pesquisa OD 2017 de São Paulo

RAIS – Relação Anual de Informações Sociais

RMSE – *Root Mean Square Error*

RMSP – Região Metropolitana de São Paulo

SC – Setor Censitário

SL – *Statistical Learning*

SR – *Synthetic Reconstruction*

VAE – *Variational Autoencoder*

SUMÁRIO

1.	INTRODUÇÃO.....	1
1.1	Caracterização do problema	1
1.2	População sintética	1
1.3	Justificativa e objetivos	3
1.4	Estrutura da dissertação	5
2.	REVISÃO DA LITERATURA.....	6
2.1	Visão geral.....	6
2.2	Principais termos e definições utilizados.....	8
2.3	Procedimentos	10
2.3.1	Reconstrução Sintética	10
2.3.2	Otimização Combinatória.....	13
2.3.3	Comparações entre procedimentos na literatura.....	13
2.3.4	As Evoluções do IPF	14
2.3.5	Maximização de entropia.....	21
2.3.6	Aprendizagem estatística	23
2.3.7	Trabalhos em português.....	25
2.3.8	Fora da academia	26
2.4	Adaptações e aplicações	27
2.5	Validação	28
2.5.1	<i>Ground truth</i> e validação interna.....	28
2.5.2	Procedimentos e pós-processamento	29
2.5.3	Métricas de qualidade de ajuste.....	30
2.6	Esquema genérico de reconstrução sintética	31
2.6.1	IPF e IPU	31
2.6.2	Maximização de entropia.....	32
3.	CÓDIGOS DISPONÍVEIS E SELEÇÃO	33
3.1	Considerações iniciais	33

3.2	Códigos disponíveis.....	33
3.3	A escolha do PopulationSim.....	35
4.	FONTES DE DADOS, ATRIBUTOS E VALIDAÇÃO	37
4.1	Introdução	37
4.2	Fontes de dados e atributos comumente utilizados	37
4.3	Fontes de dados disponíveis no Brasil.....	38
4.3.1	Censo Demográfico do IBGE.....	38
4.3.2	Demais fontes	39
4.4	Atributos a utilizar na aplicação	43
4.5	Validação	44
4.5.1	Procedimentos realizados	44
4.5.2	Métricas de qualidade de ajuste.....	44
5.	POPULATIONSIM.....	47
5.1	Sobre.....	47
5.2	Documentação	48
5.3	Metodologia.....	48
5.3.1	Visão geral.....	48
5.3.2	Componentes principais	49
5.3.3	Exemplo de aplicação.....	57
5.3.4	Passos do procedimento computacional.....	61
5.4	Procedimento de validação.....	66
6.	EXTENSÃO DO POPULATIONSIM PARA O BRASIL	68
6.1	Introdução.....	68
6.2	Visão geral dos códigos desenvolvidos	68
6.3	Considerações sobre os dados de entrada.....	71
6.3.1	Censo Demográfico do IBGE.....	71
6.3.2	Pesquisa OD de São Paulo 2017	72
6.4	Relação de atributos escolhidos.....	74

6.5	Códigos para tratamento de dados de entrada	76
6.5.1	Área a sintetizar (base: Censo 2010)	77
6.5.2	Totais de controle (base: Censo 2010)	78
6.5.3	Matriz semente (base: Censo 2010).....	82
6.5.4	Totais de controle meta 1 (base: ODSP 2017)	83
6.6	Códigos para validação.....	87
6.6.1	Código original modificado.....	87
6.6.2	Código adicional.....	87
7.	APLICAÇÃO NA RMSP.....	88
7.1	Visão geral da RMSP	88
7.2	Dados de entrada	90
7.2.1	Zonas e população	90
7.2.2	Configurações (settings.yaml).....	91
7.2.3	Variáveis e importâncias (controls.csv).....	91
7.2.4	Totais de controle	93
7.2.5	Microdados expandidos x controles meta	94
7.3	Processamento	95
7.4	Demais considerações sobre a aplicação	96
7.4.1	Data-base e projeções	96
7.4.2	Apenas para domicílios particulares.....	96
8.	VALIDAÇÃO	98
8.1	Comparação de totais agregados	99
8.1.1	Totais de controle x sintetizado.....	99
8.1.2	Variáveis combinadas.....	100
8.2	Fatores de expansão.....	101
8.3	Avaliação por variável (categoria dos atributos).....	103
8.3.1	Para setores censitários	104
8.3.2	Para áreas de ponderação.....	106

8.4	Avaliação por zona (setor censitário e área de ponderação)	108
8.4.1	Para setores censitários	108
8.4.2	Para áreas de ponderação	110
8.5	Considerações sobre a validação	115
9.	CONCLUSÃO.....	117
9.1	Síntese do trabalho e conclusões	117
9.2	Limitações do procedimento	119
9.3	Trabalhos futuros	120
9.3.1	Melhorias do <i>software</i>	120
9.3.2	Aplicação da população gerada em um modelo “final”	120
9.3.3	Análise de sensibilidade de parâmetros do modelo	121
9.3.4	Uso de variáveis cruzadas como controles	121
9.3.5	Avaliação da qualidade de ajuste para variáveis não controladas	121
	REFERÊNCIAS	122
	APÊNDICE A - CONTEÚDO DIGITAL	128
	APÊNDICE B - CÓDIGOS DE GERAÇÃO DE POPULAÇÕES SINTÉTICAS	131
	APÊNDICE C - ATRIBUTOS UTILIZADOS NA APLICAÇÃO NA RMSP	132

1. INTRODUÇÃO

1.1 CARACTERIZAÇÃO DO PROBLEMA

Modelos baseados em atividades (ABM, do inglês *Activity-Based Models*) vêm ao longo das últimas décadas se popularizando como ferramenta para a estimação de demanda por transportes, como alternativa aos modelos convencionais baseados em viagens, como o Modelo de Quatro Etapas (McNally e Rindt, 2007). Ao tratar de maneira desagregada atividades e processos de tomada de decisão, ABMs são capazes de avaliar – ao menos a princípio – um leque maior de intervenções e suas respectivas respostas comportamentais por parte da população (Ortúzar e Willumsen, 2011).

ABMs operam no nível do indivíduo, cujo comportamento de viagens é inferido a partir de suas características demográficas e socioeconômicas – como idade, sexo, renda, educação e emprego – que influenciam os seus processos de tomada de decisão e padrões de atividades. Também são utilizados atributos no nível do domicílio, como o número de residentes, posse de veículos e número de trabalhadores, visto que decisões pessoais não dependem apenas das características do indivíduo, mas também da situação familiar. Apesar de estas informações serem coletadas pelo Censo Demográfico (IBGE, 2010) para toda a população, elas não são disponibilizadas de forma desagregada por questões de privacidade (ou são disponibilizadas como amostras para baixa resolução espacial).

Sendo assim, é necessário recriar esta população com suas características e alocá-la espacialmente, em um processo conhecido na literatura como geração de populações sintéticas (em inglês, *Synthetic Population Generation*).

1.2 POPULAÇÃO SINTÉTICA

Ory (2017) define população sintética de maneira informal, como “comunidades virtuais com descrições detalhadas de domicílios e das pessoas que vivem neles”. A ideia básica é que a população gerada esteja no grupo das melhores estimativas da população real, respeitando as informações agregadas (Ryan *et al.*, 2009).

O processo para geração da população sintética normalmente se baseia em dados disponibilizados por institutos nacionais de estatísticas, nas formas de:

- Amostras da população (da ordem de 1 a 5% do total¹) com informações completas dos domicílios e indivíduo, porém sem precisão espacial (na escala das chamadas áreas de ponderação²); e
- Totais agregados por setor censitário (contagem dos domicílios ou indivíduos que apresentam uma determinada característica – por exemplo, domicílios com renda per capita entre 2 e 3 salários mínimos, ou pessoas com idade entre 21 e 30 anos).

A Figura 1 esquematiza estas bases de dados, indicando a relação entre elas e o tamanho típico destas unidades:

Figura 1 – Exemplo dos dados de interesse disponíveis no Censo Brasileiro, incluindo informações e tamanho típicos

A população sintética, que é o produto a ser gerado, é uma base de dados em que cada linha corresponde a um domicílio (e/ou indivíduo), obtido a partir dos microdados disponíveis para a área de ponderação e alocado em um dado setor censitário pertencente a esta área, com suas respectivas características. Ao contar o número de domicílios da população sintética alocados em um dado setor censitário e que possuem uma determinada característica (por exemplo, renda entre 2 e 3 salários mínimos), este número deve ser o mais próximo possível daquele fornecido como referência para o procedimento. Geradores de populações sintéticas mais atuais permitem também considerar os totais informados em áreas maiores do que aquelas em que são fornecidos os microdados da amostra (como por exemplo, municípios).

¹ Proporções comuns em outros países. No Brasil, o IBGE fornece amostras que variam entre 5% e 50% da população do município, a depender do número de habitantes deste.

² Utilizando terminologias do Censo Brasileiro para facilitar o entendimento

A maioria dos procedimentos existentes gera a população sintética replicando (clonando) os domicílios disponíveis na amostra, alterando os seus pesos amostrais (presentes na base da amostra), e os alocando nos setores censitários.

1.3 JUSTIFICATIVA E OBJETIVOS

Diversos métodos de geração de populações sintéticas foram desenvolvidos desde o trabalho seminal de Beckman *et al.* (1996), conforme descrito no Capítulo 2. O procedimento de Beckman *et al.* (1996) apresentava limitações que diversos autores buscaram superar, seja seguindo a mesma linha de abordagem (reconstrução sintética) ou outras abordagens (como otimização combinatória e aprendizagem estatística).

Nas últimas duas décadas houve uma significativa evolução das abordagens de reconstrução sintética – destacando-se as linhas de (i) *Iterative Proportional Updating* de Ye *et al.* (2009) e Konduri *et al.* (2016), e (ii) de maximização de entropia, de Vovsha *et al.* (2015) e Paul *et al.* (2018a). Em contraste, observou-se que abordagens baseadas em otimização combinatória foram pouco aplicadas na área de planejamento de transportes desde o trabalho de Williamson *et al.* (1998).

Também são notáveis os desenvolvimentos mais recentes utilizando técnicas de aprendizagem estatística, em que se destacam Farooq *et al.* (2013) e, mais recentemente, Sun *et al.* (2018). Estes procedimentos, porém, encontram-se ainda restritos à própria geração da população sintética, não tendo sido observadas aplicações posteriores em modelagem de transportes. Em contraste, os procedimentos de reconstrução sintética são frequentemente vinculados a projetos de ABMs.

Sendo assim, as abordagens baseadas em reconstrução sintética mostram-se como um caminho seguro e consolidado para a área, após o desenvolvimento de diversas técnicas para superar as suas principais limitações.

Observou-se, neste panorama de desenvolvimento de diversos procedimentos para geração de populações sintéticas, que não há um procedimento padrão e “pronto para uso” para ser aplicado para dados disponíveis no Brasil, independentemente da região. Tampouco há bases de dados já consolidadas, sintéticas ou não, da população da Região Metropolitana de São Paulo (RMSP).

Visto que a geração da população sintética é um passo chave para a aplicação de ABMs (Ortúzar e Willumsen, 2011) – que vêm se mostrando como uma alternativa promissora ao Modelo de

Quatro Etapas para a estimação de demanda por transportes – considerou-se que o desenvolvimento de um procedimento amigável e adaptado ao contexto nacional (isto é, práticas de modelagem e com os dados disponíveis no Brasil) seja pertinente para contribuir com o desenvolvimento da área no país.

Dessa forma, a presente pesquisa tem como objetivos gerais:

1. Disponibilizar uma base de dados completa e desagregada da população³ da RMSP para facilitar, dentre outras aplicações, o uso de modelos baseados em atividades (ABMs)⁴;
2. Disponibilizar um procedimento para a geração de populações sintéticas que possa ser aplicado em qualquer área do Brasil.

E como objetivos específicos, destacam-se:

1. Avaliar e selecionar um gerador de população sintética;
2. Elaborar códigos de preparação de dados de entrada disponíveis no Brasil;
3. Aplicar o gerador de população sintética para uma região no Brasil;
4. Identificar os principais procedimentos de avaliação da qualidade dos resultados obtidos e aplicá-los para uma região no Brasil.

³ Na forma de dois arquivos principais: (i) um arquivo contendo informações sobre os indivíduos (aprox. 20 milhões de registros) e (ii) outro arquivo contendo informações sobre os domicílios (aprox. 6 milhões).

⁴ E outras aplicações que utilizam como dado de entrada bases desagregadas contendo (i) as características sociodemográficas e econômicas de domicílios e indivíduos e (ii) a respectiva distribuição espacial.

1.4 ESTRUTURA DA DISSERTAÇÃO

Após o presente capítulo de introdução, a dissertação é estruturada da seguinte forma:

- Capítulo 2: Revisão da Literatura
 - Apresenta uma revisão da literatura sobre geradores de populações sintéticas;
- Capítulo 3: Códigos Disponíveis e Seleção
 - Identifica os algoritmos de geração de populações sintética disponíveis (isto é, em código aberto) e elege o código a ser utilizado para a aplicação no Brasil;
- Capítulo 4: Fontes de Dados, Atributos e Validação
 - Apresenta as fontes de dados e atributos utilizados por outras aplicações de geração de populações sintéticas – e indica os atributos e os procedimentos de validação a serem utilizados na aplicação do presente trabalho;
- Capítulo 5: PopulationSim
 - Apresenta o gerador de populações sintéticas escolhido para ser utilizado;
- Capítulo 6: Extensão do PopulationSim para o Brasil
 - Descreve os códigos desenvolvidos para possibilitar a aplicação do PopulationSim em qualquer área do Brasil, para o tratamento de dados de entrada e para a validação dos resultados obtidos;
- Capítulo 7: Aplicação na RMSP
 - Relata aspectos relacionados à aplicação do PopulationSim na Região Metropolitana de São Paulo, incluindo dados de entrada e configurações;
- Capítulo 8: Validação
 - Avalia a população sintética gerada para a RMSP com base nos procedimentos e métricas de qualidade de ajuste descritos no Capítulo 4;
- Capítulo 9: Conclusão e Trabalhos Futuros
 - Resume as principais conclusões do trabalho, aponta suas limitações e indica possibilidades de trabalhos futuros.

Por fim, o documento apresenta três apêndices, contendo os *links* para o repositório dos códigos e sua documentação (Apêndice A) e informações complementares sobre o conteúdo.

2. REVISÃO DA LITERATURA

O presente capítulo apresenta uma revisão da literatura sobre geração de populações sintéticas, especialmente dos trabalhos considerados mais relevantes para a área de planejamento de transportes. Inicialmente são indicados os principais autores e suas linhas de desenvolvimento. Na sequência, os principais procedimentos desenvolvidos desde o procedimento clássico de Beckman *et al.* (1996) são apresentados, sendo identificadas as suas motivações e principais contribuições. O capítulo encerra com um esquema indicando o procedimento genérico seguido pelos geradores de populações sintéticas – após uma discussão acerca de meios levantados na literatura para validação dos resultados.

2.1 VISÃO GERAL

Ao longo das últimas décadas, diversos procedimentos para a geração de populações sintéticas foram desenvolvidos. Eles podem ser agrupados em três classes principais: (i) reconstrução sintética, (ii) otimização combinatória e (iii) aprendizagem estatística. De forma a introduzir o assunto, apresentam-se brevemente a seguir os artigos de destaque em cada uma destas linhas:

1. Beckman *et al.* (1996): utilizaram o método IPF (*Iterative Proportional Fitting*) para a geração da população sintética do simulador TRAMSIMS (Smith *et al.*, 1995). A abordagem pode ser enquadrada na categoria de algoritmos de reconstrução sintética (*Synthetic Reconstruction*, SR). Diversos outros procedimentos utilizaram o IPF e o trabalho seminal de Beckman *et al.* (1996) como base, incluindo os métodos baseados em maximização de entropia, como Vovsha *et al.* (2015);
2. Williamson *et al.* (1998): utilizaram a técnica *Simulated Annealing* de otimização combinatória (*Combinatorial Optimization*, CO), que realiza a troca de domicílios entre as múltiplas zonas sob uma dada área de ponderação caso as métricas de qualidade de ajuste utilizadas indiquem melhoria;
3. Farooq *et al.* (2013): aplicaram técnicas de aprendizagem estatística (*Statistical Learning*, SL), buscando resolver os principais problemas identificados nos métodos de SR e CO. A abordagem de Farooq *et al.* (2013), baseada em simulações, utiliza as distribuições conjuntas dos atributos da população real (dadas ou geradas a partir de premissas ou modelos paramétricos), através do uso de condicionais. Esta linha também é referida na literatura por “métodos baseados em modelos/simulações”.

A Figura 2 esquematiza as relações entre estes três artigos seminais e outros artigos relevantes identificados na área de pesquisa de geração de populações sintéticas que serão citados ao longo do presente capítulo.

Figura 2 – Visão geral dos principais artigos identificados e suas relações com três artigos de destaque⁵

⁵ São utilizados, nas terceiras linhas das caixas, termos que remetem ao trabalho, sejam eles a técnica proposta ou utilizada, a contribuição principal do trabalho ou o nome do gerador de populações sintéticas. Mais informações sobre estes trabalhos e termos podem ser encontradas ao longo do presente capítulo.

Alguns destes artigos foram agrupados, de forma a indicar algumas evoluções ou linhas similares aos originais: as evoluções de IPF (controles em indivíduos e domicílios e IPU), as abordagens que utilizam maximização de entropia e as abordagens que utilizam técnicas de inteligência artificial.

Observa-se na literatura uma predominância na utilização de métodos de reconstrução sintética, especialmente nas aplicações em modelagem de transportes. Métodos baseados em otimização combinatória são mais observados em aplicações na área de geografia, e em modelos relacionados à saúde, saneamento e avaliação de políticas de impostos (Harland *et al.*, 2012). A linha de aprendizagem estatística vem se destacando nos últimos anos, em um momento de desenvolvimento das teorias e aplicações de técnicas de inteligência artificial em diversos campos do conhecimento (Sun *et al.*, 2018). Entretanto, suas aplicações em ABMs ainda são restritas, apesar do bom desempenho observado ao se avaliar a sua aderência com a população real. Dessa forma, no presente trabalho, optou-se por focar os esforços nas abordagens de reconstrução sintética – incluindo maximização de entropia –, que vêm sendo aprimoradas nas duas últimas décadas por diversos autores.

2.2 PRINCIPAIS TERMOS E DEFINIÇÕES UTILIZADOS

- *Microdados* são bases de dados sobre *amostras* (parcelas do total) de domicílios ou pessoas (que são dois arquivos separados), indicando seus *atributos*. Estes registros são individualizados, anônimos e sem precisão espacial – com informações sobre apenas 5% dos domicílios de uma dada *área de ponderação*. A base de *microdados* é utilizada para construir a chamada *matriz semente* dos geradores de populações sintéticas;
- A *matriz semente* é uma seleção dos microdados, filtrados em função da abrangência desejada e dos atributos de interesse. Ela é transformada pela maioria dos procedimentos em *tabelas de contingência* (isto é, é agrupada através de contagens);
- *Totais marginais* (chamados de *totais de controle* quando são utilizados como parâmetro nos procedimentos) são contagens de ocorrências em uma dada zona, apresentadas para *categorias de atributos*, seja de domicílio ou de indivíduo (por exemplo, total de pessoas com idade entre 21 e 30 anos). Estas categorias funcionam como *variáveis* nos procedimentos de geração e validação de populações sintéticas;

- *Totais marginais/de controle*, no Censo brasileiro, estão disponíveis para *setores censitários*, que é a menor escala em que o Censo divulga informações. A base do Censo que contém esta informação é chamada de “Agregados por Setores Censitários”. *Áreas de ponderação* são formadas por múltiplos *setores censitários* (normalmente dezenas);
- O termo *zonas* é utilizado no presente trabalho de maneira genérica, para se referir a unidades geográficas – sejam elas *setores censitários*, *áreas de ponderação*, *distritos* ou *municípios*, a depender do contexto e do método. Diversas *zonas* formam as chamadas *regiões*;
- A escala *meta* se refere à maior escala geográfica utilizada por um gerador de população sintética, maior inclusive que as *áreas de ponderação*;
- *IPF (Iterative Proportional Fitting)* é um método de *ajustamento* (uma das etapas de métodos de *reconstrução sintética*). Entretanto, o acrônimo *IPF* é por vezes utilizado para se referir a todo o procedimento de geração de população sintética dos geradores que o utilizam;
- Apesar de não utilizados neste documento, são indicados a seguir os nomes de bases e escalas geográficas utilizadas nos EUA e Reino Unido, o que pode ser relevante para o leitor que consultar diretamente a literatura internacional:
 - *Setores censitários*, no censo brasileiro, correspondem a:
 - *Census Tract*, nos EUA⁶, e
 - *ED (Enumeration Districts)* e suas respectivas *SAS (Small Area Statistics)*, no Reino Unido;
 - E *área de ponderação* e *microdados*, respectivamente, a:
 - *PUMA* e *PUMS (Public Use Microdata Area e Sample)*, nos EUA, e
 - *SAR area* e *SAR (Samples of Anonymised Records)*, no Reino Unido.

⁶ Entretanto, diversas aplicações observadas adotavam a *TAZ (Traffic Analysis Zone)* como unidade para controle ao invés de *Census Tract*. As *TAZ* são fornecidas diretamente pelo censo dos EUA, através do *Census Transportation Planning Package*. Outras unidades menores do censo estadunidense são: *block* e *block group* (menores que o *tract*)

2.3 PROCEDIMENTOS

2.3.1 Reconstrução Sintética

O artigo de Beckman *et al.* (1996) é um dos mais reconhecidos para a geração de populações sintéticas, ao ser o primeiro a aplicar o procedimento conhecido por *Iterative Proportional Fitting* (IPF) para resolver o problema de geração de populações sintéticas. Conforme apontado por Miranda (2019), o trabalho de Beckman *et al.* (1996) é a publicação mais citada na área de geradores de populações sintéticas.

Utilizando o procedimento IPF, proposto por Deming e Stephan (1940), uma amostra de uma população, estruturada na forma de uma tabela de contingência, é rebalanceada de forma a satisfazer restrições de totais marginais dos seus atributos. Este é o primeiro de dois passos principais do método de Beckman *et al.* (1996), chamado na literatura de *fitting stage* (etapa de “ajustamento”).

Com a tabela de contingência ajustada, cada agente da amostra é clonado e passa a fazer parte da população de uma dada zona em função de sua probabilidade de ser incluído, calculada na etapa anterior. A seleção de agentes normalmente é realizada através de simulações de Monte Carlo. Este segundo passo corresponde à etapa de “geração” da população sintética.

A tese de doutorado de Müller (2017) pode ser consultada para uma definição verbal, formal e um exemplo numérico do método IPF, assim como o próprio trabalho original de Beckman *et al.* (1996). O exemplo numérico para duas variáveis, adaptado de Müller (2017), é apresentado na Figura 3. No exemplo, indivíduos são classificados segundo dois atributos: (i) status de trabalho, com três categorias: não trabalha (Ñ), trabalha em tempo parcial (P) e trabalha em tempo integral (I); e (ii) idade, segregada em quatro categorias. As variáveis n representam as somas das linhas e colunas da tabela de contingência (com matriz semente preparada a partir dos microdados). Os totais de controle de linha (ℓ_i) e coluna (c_j) são representados à direita e abaixo, respectivamente. Os fatores de ajuste $\ell_i/n_i^{*(k)}$ e $c_j/n_j^{*(k)}$, utilizados para atualizar os valores da tabela de contingência alternadamente ao longo das iterações k , são indicados na última linha e coluna, respectivamente.

(a) Tabela de contingência inicializada para o IPF

k = 0		j (idade)				$n_{i*}^{(k)}$	ℓ_i	$\ell_i / n_{i*}^{(k)}$
		0-14	15-34	35-64	65+			
i (trabalho)	Ñ	73	23	35	74	205	124	0,60
	P	0	42	17	15	74	83	1,12
	I	0	60	65	2	127	227	1,79
$n_{*j}^{(k)}$		73	125	117	91	406	-	$n_{**}^{(k)}$
c_j		88	132	115	99	-	434	n
$c_j / n_{*j}^{(k)}$		1,21	1,06	0,98	1,09			

(b) Tabela de contingência após ajustamento para controles de linhas

k = 1		j (idade)				$n_{i*}^{(k)}$	ℓ_i	$\ell_i / n_{i*}^{(k)}$
		0-14	15-34	35-64	65+			
i (trabalho)	Ñ	44,2	13,9	21,2	44,8	124	124	1,00
	P	0,0	47,1	19,1	16,8	83	83	1,00
	I	0,0	107,2	116,2	3,6	227	227	1,00
$n_{*j}^{(k)}$		44,2	168,3	156,4	65,2	434	-	$n_{**}^{(k)}$
c_j		88	132	115	99	-	434	n
$c_j / n_{*j}^{(k)}$		1,99	0,78	0,74	1,52			

Figura 3 – Exemplo de aplicação do método IPF de Beckman *et al.* (1996), parte 1/2
 Fonte: Müller (2017). Adaptado pelo autor.

(c) Tabela de contingência após ajustamento para controles de colunas

$k = 2$		j (idade)				$n_{i*}^{(k)}$	ℓ_i	$\ell_i / n_{i*}^{(k)}$
		0-14	15-34	35-64	65+			
i (trabalho)	$\tilde{N}$	88,0	10,9	15,6	68,0	182,5	124	0,68
	P	0,0	37,0	14,0	25,6	76,5	83	1,08
	I	0,0	84,1	85,4	5,4	175,0	227	1,30
$n_{*j}^{(k)}$		88	132	115	99	434	-	$n_{**}^{(k)}$
c_j		88	132	115	99	-	434	n
$c_j / n_{*j}^{(k)}$		1,00	1,00	1,00	1,00			

(d) Tabela de contingência após convergência

$k \rightarrow \infty$		j (idade)				$n_{i*}^{(k)}$	ℓ_i	$\ell_i / n_{i*}^{(k)}$
		0-14	15-34	35-64	65+			
i (trabalho)	$\tilde{N}$	88,0	1,8	2,5	31,7	124	124	1,00
	P	0,0	25,2	9,3	48,5	83	83	1,00
	I	0,0	105,0	103,2	18,8	227	227	1,00
$n_{*j}^{(k)}$		88	132	115	99	434	-	$n_{**}^{(k)}$
c_j		88	132	115	99	-	434	n
$c_j / n_{*j}^{(k)}$		1,00	1,00	1,00	1,00			

Figura 4 – Exemplo de aplicação do método IPF de Beckman *et al.* (1996), parte 2/2
 Fonte: Müller (2017). Adaptado pelo autor.

O procedimento indicado no exemplo das figuras anteriores (Figura 3 e Figura 4) é, na prática, realizado de maneira análoga simultaneamente para as “n” variáveis a serem utilizadas como controles.

Este procedimento é o primeiro de dois passos principais da geração de populações sintéticas pelos métodos de reconstrução sintética baseados em IPF, chamado na literatura internacional de *fitting stage* (etapa de ajustamento). O segundo passo corresponde à geração da população. A Figura 5 esquematiza o núcleo do procedimento, que serve como base para outros métodos.

Figura 5 – Os dois principais passos dos procedimentos baseados em IPF para geração de populações sintéticas

Moreno e Moeckel (2018) destacam que, enquanto existe uma série de procedimentos alternativos empregados para a fase de ajustamento (que serão discutidos ao longo deste capítulo), a etapa de geração normalmente se baseia em simulações de Monte Carlo, em que agentes são obtidos aleatoriamente a partir das distribuições conjuntas ajustadas a partir de uma matriz semente na primeira etapa (Ye *et al.*, 2009). Fogem desta regra os métodos baseados em otimização combinatória e alguns dos métodos de maximização de entropia e aprendizagem estatística, conforme apresentado no levantamento realizado por Moreno e Moeckel (2018).

2.3.2 Otimização Combinatória

Williamson *et al.* (1998) utilizaram otimização combinatória para o ajuste dos pesos da matriz semente. O processo iterativo é iniciado com uma seleção aleatória de domicílios da amostra para a composição de uma zona. Na sequência, um destes domicílios é substituído por outro. Caso a qualidade de ajuste melhore, a troca é concretizada (Harland *et al.*, 2012). A técnica empregada por Williamson *et al.* (1998) de *simulated annealing* é utilizada de para evitar seleções de agentes que não sejam ótimas (Templ, 2014).

Em suma, conforme sintetizado por Vovsha *et al.* (2015), estes métodos são baseados na realização da troca de domicílios entre as múltiplas zonas sob uma dada área de ponderação⁷ caso as métricas utilizadas de qualidade de ajuste melhorem.

2.3.3 Comparações entre procedimentos na literatura

Após o trabalho seminal de Beckman *et al.* (1996), outros geradores de populações sintéticas foram desenvolvidos. Bowman (2004) realizou uma comparação entre oito geradores de populações sintéticas disponíveis na época, todos baseados no procedimento IPF, incluindo o TRANSIMS de Beckman *et al.* (1996). O artigo indicou melhorias a serem incorporadas em futuros geradores de populações sintéticas baseados em IPF, destacando a importância da validação dos atributos adotados e da criação de um *software* que permitisse o ajustamento de atributos básicos sem a necessidade de programação adicional.

Mais recentemente, Müller e Axhausen (2010) elaboraram um levantamento do estado da arte dos geradores de populações sintéticas baseados em IPF. Seis geradores de população sintéticas publicados entre 2007 e a data do artigo foram avaliados. Os autores destacam que o desempenho destes geradores não foi avaliado comparativamente, uma vez que os tipos de

⁷ Termo do Censo brasileiro para se referir à área para a qual são fornecidas as informações da amostra

dados de entrada variam, assim como as métricas de qualidade de ajuste utilizadas nos processos de validação. Os autores encerram o artigo constatando que cada gerador tem sua vantagem e que um que fosse superior e que incorporasse as melhores abordagens ainda não havia sido desenvolvido. Também destacam a dificuldade em se produzir um gerador de populações sintéticas universal, no que se refere a diferentes dados de entrada considerando geografias e tipos de agentes diferentes, e recomendam o desenvolvimento de um programa de código aberto que possibilitasse a execução de rotinas para tarefas frequentemente enfrentadas na geração de populações sintéticas.

Ryan *et al.* (2009) realizaram uma comparação entre métodos CO e IPF utilizando uma população completa (de empresas e empregados, no caso) para avaliar seus desempenhos, e concluíram que CO produzia resultados com menor variância. Apesar disso, conforme destacado por Farooq *et al.* (2013), o tempo de convergência de métodos baseados em CO é muito alto, especialmente com uso de *simulated annealing* como ferramenta de otimização. Este fato pode ter contribuído para que métodos baseados em IPF tenham sido mais utilizados e desenvolvidos do que CO nas últimas décadas.

2.3.4 As Evoluções do IPF

O procedimento elaborado por Beckman *et al.* (1996) apresenta algumas limitações, que foram tratadas por outros autores desde então. Dentre elas, destacam-se: (i) o problema de valores nulos na amostra, (ii) os requisitos computacionais, (iii) a impossibilidade de controlar as variáveis tanto no nível de indivíduos como de domicílios e (iv) o uso de variáveis disponíveis para diferentes escalas geográficas. Estas limitações são explicadas nos próximos subitens e são apresentados os principais procedimentos desenvolvidos a partir do IPF de Beckman *et al.* (1996) que buscaram superá-las.

2.3.4.1 Valores nulos da amostra

As matrizes semente, obtidas a partir de microdados, contêm uma população relativamente pequena de indivíduos/domicílios. Com isso, é comum que uma determinada combinação de atributos não esteja representada nesta amostra, e que assim seu valor seja zero na tabela de contingência, apesar de na realidade existir em uma determinada zona. Como pode se observar no artigo de Beckman *et al.* (1996), das 11.760 células obtidas pela multiplicação do número de categorias de cada atributo (variáveis), apenas 609 apresentavam valores diferentes de zero.

No procedimento IPF, as células inicializadas com valor zero jamais mudam de valor, por se tratar de um procedimento multiplicativo. Para superar este problema, Beckman *et al.* (1996) inserem valores pequenos arbitrários no lugar destes zeros – entretanto isto pode introduzir uma distorção nas estruturas de correlações entre as variáveis. Guo e Bhat (2007) apresentam algumas alternativas para tratar este problema, como a redução do número de categorias dos atributos. Com faixas de valores maiores dentro de cada categoria de uma variável quantitativa (ou agregando diversas categorias para variáveis qualitativas), a chance de se obter valores nulos na amostra tende a ser menor.

A abordagem de Ye *et al.* (2009) para este problema se baseia em utilizar estimativas de escalas geográficas maiores (região) para substituir os “falsos-zeros” em escalas geográficas menores (zona). Um exemplo numérico dessa abordagem, para atributos de tamanho e renda de domicílios, é apresentado na Figura 6.

No exemplo, uma determinada zona não apresenta nenhuma combinação de renda baixa e tamanho 1 para domicílios. Entretanto, no nível da região que esta zona se encontra, são observados 2 dos 33 domicílios com estas características. Esta parcela, que corresponde a aproximadamente 6% do total, é aplicada no nível de zona, e na sequência a matriz de probabilidades é ajustada de forma a somar 1.

(a) Amostra da zona

Zona n		Renda	
		Alta	Baixa
Tamanho	1	3	0
	2	2	4
	>3	2	1
		Total	12

(b) Amostra da região

Região (contém zona n)		Renda	
		Alta	Baixa
Tamanho	1	7	2
	2	8	10
	>3	3	3
		Total	33

(c) Probabilidades da zona, valor zero destacado

Zona n		Renda	
		Alta	Baixa
Tamanho	1	0,25	0,00
	2	0,17	0,33
	>3	0,17	0,08
		Total	1,00

(d) Probabilidades da região

Região (contém zona n)		Renda	
		Alta	Baixa
Tamanho	1	0,21	0,06
	2	0,24	0,30
	>3	0,09	0,09
		Total	1,00

(e) Imputação de uma probabilidade

Zona n		Renda	
		Alta	Baixa
Tamanho	1	0,25	0,06
	2	0,17	0,33
	>3	0,17	0,08
		Total	1,06

(f) Amostra da zona após ajuste (divisão por 1,06)

Região (contém zona n)		Renda	
		Alta	Baixa
Tamanho	1	0,24	0,06
	2	0,16	0,31
	>3	0,16	0,08
		Total	1,00

Figura 6 – Abordagem de Ye *et al.* (2009b) para o problema de valores nulos na amostra
 Fonte: Ye *et al.* (2009b). Adaptado pelo autor.

2.3.4.2 Controles nos níveis de indivíduos e domicílios

Beckman *et al.* (1996) realizaram a sintetização de domicílios sem considerar as restrições dos totais de controle no nível do indivíduo. Por isso, são observadas inconsistências entre a população sintetizada e a real. Para resolver esse problema, alguns autores utilizaram métodos distintos, como a incorporação de controles no nível do indivíduo na etapa de geração, após uma etapa de ajustamento comum.

Guo e Bhat (2007), além de tratarem o problema de valores nulos na amostra, também buscaram satisfazer os controles no nível do indivíduo em seu artigo. Em sua abordagem, os autores utilizaram simulações de Monte Carlo, porém com pesos que são reduzidos uma vez que um

dado domicílio é selecionado, de forma a favorecer grupos com agentes que ainda se encontram sub-representados na população até o momento. Auld e Mohammadian (2010) consideram a influência dos atributos no nível do indivíduo para a seleção de um domicílio, sendo alterada a sua probabilidade de seleção. Müller (2017) classifica ambos os métodos como “métodos de ajustamento em nível único com seleção enviesada” em contraste a outros considerados “algoritmos multinível”.

O gerador de populações sintéticas do ABM Albatross, desenvolvido por Arentze *et al.* (2007), utiliza a técnica de matriz de relações para superar o problema em questão. O modelo, desenvolvido para os Países Baixos, exigia os dados no nível do domicílio – entretanto, apenas dados no nível do indivíduo estavam disponíveis. A matriz de relações designa posições em domicílios para os indivíduos – podendo ser (i) chefe de família ou cônjuge do chefe, (ii) independente (apenas uma pessoa) ou (iii) morador. Ao montar os domicílios, distribuições no nível de indivíduos são convertidas em distribuições de domicílios.

Uma linha de evolução do IPF que se destaca é a IPU: *Iterative Proportional Updating*. O procedimento, desenvolvido por Ye *et al.* (2009), é utilizado pelo sintetizador PopGen (MARG, 2016) e permite o uso de controles em mais de um nível hierárquico (como indivíduos e domicílios) ainda na etapa de ajustamento.

O algoritmo IPU parte da estrutura condensada de lista proposta em Pritchard e Miller (2009), em que é representada em cada linha a frequência de um determinado tipo de domicílio (em função das categorias dos atributos adotados). Na versão de Ye *et al.* (2009) as combinações consideram atributos no nível do indivíduo e domicílio e a cada iteração é analisada uma determinada categoria de um atributo, sendo modificados apenas os pesos das linhas que apresentam valor não-nulo para esta categoria, seja ela no nível de domicílio ou de indivíduo.

A Figura 7, adaptada de Müller (2017), ilustra o procedimento do IPU. No exemplo, *y* corresponde aos grupos de domicílios agregados de acordo com os atributos considerados (por exemplo, os domicílios de código entre 1 e 22 são considerados iguais neste procedimento); na sequência o atributo de posse (A1) ou não (A0) de automóvel é indicado na forma de duas *dummies*. Os domicílios também são categorizados em função do número de indivíduos que não trabalha (Ñ) e que trabalha (T). Na parte da direita da figura são indicadas as *k* iterações (e sua respectiva categoria A), que são realizadas utilizando o fator de ajuste calculado a partir da

relação entre o total de controle da categoria⁸ (c_A) e a frequência calculada daquela iteração, $F_A^{(k)}$. Nas iterações, o fator de ajuste multiplica os pesos dos grupos de domicílios (linhas) que apresentam valor não nulo para a categoria da respectiva iteração.

y	dummies		número. dom.	frequência		número. ind.	Iteração (k)					
	A0	A1		Ñ	T		Atributo (A)	Ñ	T	A0	A1	Ñ
1-22	0	1	22	2	1	66	$f_y^{(k)}$	1,00	0,74	1,66	1,66	1,64
23-43	0	1	21	1	1	42		1,00	0,74	1,66	1,66	1,64
44-64	0	1	21	3	0	63		1,00	0,74	0,74	0,74	0,74
65-80	1	0	16	2	0	32		1,00	0,74	0,74	0,39	0,39
81-96	1	0	16	2	1	48		1,00	0,74	1,66	0,86	0,86
97-108	1	0	12	1	0	12		1,00	0,74	0,74	0,39	0,39
109-119	0	1	11	2	0	22		1,00	0,74	0,74	0,74	0,74
120-128	0	1	9	1	0	9		1,00	0,74	0,74	0,74	0,74
129-136	1	0	8	1	2	24		1,00	0,74	1,66	0,86	0,86
137-144	0	1	8	1	2	24		1,00	0,74	1,66	1,66	1,64
145-151	1	0	7	1	1	14		1,00	0,74	1,66	0,86	0,86
152-158	1	0	7	3	0	21		1,00	0,74	0,74	0,39	0,39
159-164	0	1	6	0	1	6		1,00	1,00	2,23	2,23	2,21
165-170	0	1	6	0	2	12		1,00	1,00	2,23	2,23	2,21
171-173	1	0	3	0	1	3		1,00	1,00	2,23	1,16	1,16
174-175	0	1	2	0	3	6		1,00	1,00	2,23	2,23	2,21
176	1	0	1	0	2	2		1,00	1,00	2,23	1,16	1,16
Σ	70	106	176	279	127	406	$F_A^{(k)}$	279,0	101,7	86,2	146,1	255,0
							c_A	207	227	45	145	207
							$c_A / F_A^{(k)}$	0,74	2,23	0,52	0,99	0,81

Figura 7 – Exemplo de aplicação do método IPU de Ye *et al.* (2009)

Fonte: Müller (2017). Adaptado pelo autor.

Antes da primeira iteração (iteração 0), considerando os pesos iniciais dos domicílios como 1, seriam obtidos 279 indivíduos na população sintética que não trabalham. Entretanto, segundo o total de controle desta zona, 207 indivíduos da zona não trabalham. Então, é calculado um fator de correção dado pela razão entre estes números e obtem-se o resultado de 0,74. Este fator multiplica os pesos dos domicílios que possuem algum indivíduo para esta característica – enquanto os demais se mantem inalterados. Na sequência, o mesmo é feito para outro atributo (no caso, indivíduos que trabalham) com os novos pesos. Novamente, o fator de correção (novo)

⁸ Dado de entrada

é aplicado somente nos domicílios que apresentam ao menos um indivíduo nesta condição. O procedimento é repetido até um critério de convergência (em função da diferença entre pesos obtidos em cada iteração) ser alcançado.

A etapa de geração em Ye *et al.* (2009) também difere da abordagem clássica de seleção aleatória de domicílios: a probabilidade de um domicílio ser selecionado é igual ao seu peso dividido pela soma dos pesos dos domicílios pertencentes a um dado grupo de domicílios. Isto é feito porque domicílios que se enquadram em uma mesma combinação de categorias para atributos de domicílios podem ser compostos, naturalmente, por indivíduos distintos em relação a outras variáveis que não são controladas. Dessa forma, distribuições tanto no nível de domicílio quanto no nível de indivíduo são satisfeitas.

Por fim, o método de Ye *et al.* (2009) complementa a população sintética gerada com os domicílios necessários para que esta alcance o número total de cada zona. Isto é necessário porque há um arredondamento dos pesos para os números inteiros mais próximos e, como muitas células apresentam valor próximo de zero, a tendência é que estas se tornem zero e que o total fique abaixo do desejado. Para igualar os totais, os autores comparam as distribuições obtida e desejada (baseada nos totais de controle), e os grupos de domicílios cujos totais apresentam as maiores diferenças tem seus valores aumentados em 1.

Anos depois, Ye *et al.* (2020) analisaram o IPU e provaram, sob a perspectiva teórica, que o procedimento proposto em 2009 é incapaz de convergir para uma solução ótima de distribuição da população satisfazendo tanto os totais de controle no nível de indivíduo como domicílio. Ye *et al.* (2020) então propuseram uma extensão ao procedimento original que utiliza otimização binível.

Müller e Axhausen (2011) propuseram o algoritmo HIPF (*Hierarchical Iterative Proportional Fitting*), que tem como ideia principal a alternância entre os domínios de totais de controle de atributos de indivíduos e domicílios para a realização do IPF, convertendo os fatores de expansão do nível de domicílio para o nível de indivíduo e vice-versa. Para alternar entre estes domínios, é utilizado um passo de otimização de entropia, utilizando o princípio de inserir o mínimo de informação possível na distribuição. Em contraste aos métodos IPU (Ye *et al.*, 2009) e de entropia (Bar-Gera *et al.*, 2009; Lee e Fu, 2011), que consideram os controles no nível do indivíduo apenas na etapa de geração (segunda etapa), o HIPF já os considera na etapa de ajustamento (primeira etapa). Os autores, após a validação frente aos algoritmos IPU e de

entropia, concluem que o desempenho do HIPF é superior em termos de qualidade de ajuste, enquanto os seus tempos de execução e convergência são similares aos dos demais métodos.

Müller (2017), ao comparar os métodos, indica que os algoritmos IPU e baseados em entropia são altamente similares, tendo como única diferença o procedimento de ajuste de pesos: enquanto o primeiro altera os pesos igualmente, independentemente do número de pessoas de um domicílio de uma determinada categoria (conforme exemplificado na Figura 7), o segundo diferencia o “impacto” do rebalanceamento em função do número de pessoas do domicílio que se encontrem nesta categoria. Entretanto, ambos operam apenas alterando os pesos no nível do domicílio, como foi exemplificado para o IPU na Figura 7 – em contraste ao HIPF.

2.3.4.3 Controles em diferentes escalas geográficas

Conforme observado no item anterior, diversos pesquisadores desenvolveram métodos capazes de utilizar controles nos níveis de indivíduo e domicílio simultaneamente para uma determinada escala geográfica. Entretanto, é comum que alguns dos totais de controle estejam disponíveis apenas para escalas maiores.

Konduri *et al.* (2016), partindo do algoritmo desenvolvido por Ye *et al.* (2009), incorporaram ao algoritmo IPU a possibilidade de acomodar totais de controle em mais de uma escala geográfica. No procedimento, os pesos são inicialmente ajustados de forma a satisfazer os controles no nível de região, isto é, a maior escala geográfica tratada. Em um passo seguinte, os pesos são ajustados para cada unidade geográfica dentro da região em questão. Estes passos são repetidos e o procedimento iterativo chega ao fim uma vez que o critério de convergência é alcançado. No exemplo apresentado pelos autores, mil iterações foram suficientes para alcançar resultados satisfatórios.

Moreno e Moeckel (2018) expandiram o IPU de Konduri *et al.* (2016) para três escalas geográficas e aplicaram o modelo na região metropolitana de Munique, Alemanha. Os atributos estavam divididos em três níveis hierárquicos: indivíduos, domicílios (referentes à composição familiar) e habitações (referentes aos atributos dos imóveis). Os autores também indicam como balancear a quantidade e ordem dos atributos para buscar resultados de maior qualidade.

2.3.4.4 Requisitos computacionais

Em termos computacionais, Pritchard e Miller (2009) melhoraram significativamente o tempo de processamento e requerimento de memória dos algoritmos utilizando uma representação das matrizes por listas ao invés de tabelas de contingência.

Conforme destacado por Müller e Axhausen (2010), cada atributo (ou nova categoria de atributo) adicionado ao modelo como controle implica no aumento do número de dimensões da tabela de contingência – que cresce de maneira exponencial e passa a consumir mais memória para seu processamento. Na *sparse list representation* de Pritchard e Miller (2009), combinações cujos valores sejam nulos não são exibidas e as demais são registradas a partir de suas frequências. Com isso, um número maior de variáveis (atributos e suas respectivas categorias) pode ser utilizado. Para exemplificar, os dados do exemplo da Figura 3 são representados na forma de lista na Tabela 1 a seguir.

Tabela 1 – Representação por lista para realização do IPF

Domicílios	Trabalho (i)	Idade (j)	Número de indivíduos
1-22	Ñ	0-14	73
74-96	Ñ	15-34	23
97-131	Ñ	35-64	35
132-205	Ñ	65+	74
206-247	P	15-34	42
248-264	P	35-64	17
265-279	P	65+	15
280-339	I	15-34	60
340-404	I	35-64	65
405-406	I	65+	2

Fonte: Müller (2017). Adaptado pelo autor.

2.3.5 **Maximização de entropia**

Outro caminho explorado para resolver o problema do uso de controles em mais de um nível hierárquico foi o de maximização da entropia, adotado por Bar-Gera *et al.* (2009) e Lee e Fu (2011), através da resolução de um problema de otimização sujeito a restrições. Apesar de não utilizarem IPF, as técnicas baseadas em maximização de entropia também partem do princípio de rebalancear a matriz semente e, sendo assim, também podem ser enquadradas na categoria de reconstrução sintética. Bar-Gera *et al.* (2009) também introduzem a ideia de permitir a

flexibilização dos totais, acomodando algum nível de desvio entre os totais de controle e os totais obtidos.

Barthelemy e Toint (2013) também utilizaram maximização de entropia em um método que, diferente dos demais, não utiliza amostras como dado de entrada. O método consiste em gerar os indivíduos em um primeiro passo e, após estimar as distribuições conjuntas dos domicílios, juntar estes indivíduos para realizar a composição dos domicílios.

Vovsha *et al.* (2015) propuseram um procedimento baseado em maximização da entropia utilizando programação linear tanto para a realização do chamado balanceamento (etapa análoga ao ajustamento do IPF, mas que também é utilizada para a alocação dos domicílios nas zonas), como para a “integralização” dos pesos. Para a resolução dos problemas de programação linear, os autores propõem o uso do método iterativo de Newton-Raphson.

Em relação à integralização, Vovsha *et al.* (2015) destacam que os totais de controle geralmente não podem ser facilmente desagregados. Para o procedimento, são utilizados dois passos: (i) um “meta-balanceamento”, considerando que os totais de controles no nível de região “controlam” os totais no nível de zona; e (ii) a desagregação dos pesos por zonas.

Os autores destacaram em seu artigo três novas funcionalidades de seu gerador de populações sintéticas:

- (i) O uso de totais de controle imperfeitos, possibilitando o uso de graus de confiabilidade distintos para cada total de controle. A funcionalidade se justifica pelo fato de que dados de diferentes fontes podem ser conflitantes;
- (ii) A discretização otimizada dos fatores de expansão fracionais (a chamada “integralização”), que é utilizada como alternativa à geração por métodos como Monte Carlo;
- (iii) O uso de controles em múltiplas escalas geográficas, possibilitando que sejam utilizados totais de controle tanto em escala maior que a associada à semente (chamada de “meta”) como quantas escalas menores se desejar.

Paul *et al.* (2018a) propuseram melhorias em relação à metodologia proposta por Vovsha *et al.* (2015). A melhoria, segundo os autores, se encontra na correção do erro da última zona sintetizada (*last zone error*, em inglês), ao realizar tanto o balanceamento como a “integralização” de forma simultânea para todas as zonas. O artigo desenvolvido por Paul *et al.*

(2018a) apresenta o PopulationSim, implementação em código aberto da metodologia proposta pelos autores.

O procedimento desenvolvido por Vovsha *et al.* (2015) e posteriormente aprimorado por Paul *et al.* (2018a) é mais explorado no item 3.3 (que aborda a seleção do código a adaptar e justifica a sua escolha) e no Capítulo 5 (em que o PopulationSim é apresentado).

2.3.6 Aprendizagem estatística

Em 2013, Farooq *et al.* propuseram um novo método para a geração de populações sintéticas, utilizando técnicas de aprendizagem estatística. Os autores apresentam uma série de problemas identificados com as abordagens baseadas em IPF ou CO, dentre eles: (i) o excesso de confiança nos dados disponíveis para determinação dos pesos, com o uso de uma única tabela de contingência para ser ajustada a partir de uma amostra; (ii) o fato de clonarem os agentes, perdendo a heterogeneidade que pode não ter sido capturada nas amostras; e (iii) os problemas de escalabilidade para uso de um número elevado de atributos. Os autores destacam que os procedimentos baseados em definições de peso e amostragem focam no ajustamento a partir dos marginais, sem ênfase na estrutura de correlações.

A abordagem de Farooq *et al.* (2013), baseada em simulações, considera as distribuições conjuntas observadas na população real, através do uso de condicionais. A técnica de *Gibbs sampling*, um método de *Markov Chain Monte Carlo* – ou MCMC –, é utilizada de forma a simular uma sequência dependente de amostragens aleatórias para formar a população sintética a partir de distribuições conjuntas desconhecidas.

Entretanto, o procedimento precisa das distribuições condicionais entre os atributos a serem utilizados. Para isso, é realizada uma etapa de preparação destes condicionais, que normalmente não estão todos disponíveis para utilização. Para estes condicionais incompletos, os autores citam três alternativas: uso de (i) premissas, (ii) conhecimento do domínio e (iii) modelos paramétricos. Para o primeiro, os autores exemplificam com uma regra de que, caso não se saiba a distribuição condicional da idade, dado sexo e renda ($\text{idade} \mid \text{sexo}, \text{renda}$) pode se assumir que sexo é independente de idade e assim $(\text{idade} \mid \text{sexo}, \text{renda}) = (\text{idade} \mid \text{renda})$. Já para o segundo, conhecimento do domínio, um exemplo é assumir que crianças não têm renda. Por fim, para o uso de modelos paramétricos, é apresentado um modelo logit de escolha discreta para tipo de habitação a partir de renda, sexo e idade. Para a aplicação realizada por Farooq *et*

al. (2013), em Bruxelas (Bélgica), a renda é estimada a partir da renda média da zona, posse de automóveis, nível educacional e número de trabalhadores no domicílio.

De acordo com Farooq *et al.* (2013), diferentemente das duas outras abordagens, o método desenvolvido não se baseia na clonagem de registros a partir da definição de pesos de uma tabela de contingência, mas sim considerando a distribuição conjunta entre os atributos e obtendo agentes a partir dela, possibilitando maior grau de heterogeneidade. A validação, realizada frente a dados completos do censo da Suíça (*ground truth*), indicou que o método superou a abordagem clássica de IPF, realizada pelos autores para as mesmas condições. O método também apresenta como vantagens possibilitar o uso de atributos tanto discreto quanto contínuos.

Entretanto, conforme destacado por Casati *et al.* (2015), o método é restrito a dados não-hierárquicos – isto é, não permite combinações de atributos a níveis de domicílios e indivíduos – e não permite a imposição direta de totais de controle. Casati *et al.* (2015) então desenvolveram uma extensão para o método superando estes dois problemas, chamada de hMCMC (*Hierarchical Markov Chain Monte Carlo*). O método proposto cria os domicílios definindo as posições dos indivíduos, podendo estes serem donos (maior renda e maior idade, em caso de conflito), esposo ou filho (segundos e terceiros agentes dos domicílios); ou outro (demais agentes do domicílio).

Para satisfazer perfeitamente os totais de controle, um passo seguinte (de pós-processamento) é proposto, utilizando a técnica de *raking* generalizado (GR, *generalized raking*). A técnica, utilizada para a calibração de pesos amostrais de pesquisas e muito popular na área de metodologias de pesquisas estatísticas, não havia sido muito aplicada na área de planejamento de transportes segundo levantamento de Müller (2017).

O trabalho de Farooq *et al.* (2013), em um momento de desenvolvimento das teorias e aplicações de técnicas de inteligência artificial em diversos campos do conhecimento (Sun *et al.*, 2018), abriu o caminho para outras abordagens para a geração de populações sintéticas.

Sun e Erath (2015) propuseram uma abordagem baseada em redes bayesianas, considerando o problema de geração de populações sintéticas como a inferência de uma distribuição de probabilidade conjunta. Os autores utilizam rede bayesianas para modelar as interdependências entre atributos, tanto no nível de domicílio como de indivíduo.

Sun *et al.* (2018) propuseram uma estrutura para geração de populações sintéticas composta por três modelos principais: (i) análise de classes latentes integrada com (ii) fatoração tensorial, seguido por (iii) pós-processamento baseado em amostragem por rejeição. Os autores afirmam que o modelo proposto tem grande potencial para reproduzir associações tanto entre variáveis no mesmo nível como em níveis diferentes (domicílios e indivíduos). O modelo proposto naturalmente considera as associações entre variáveis nos níveis de domicílios e indivíduos.

Borysov *et al.* (2018) propõem um método baseado em VAE (*Variational Autoencoder*), um método de aprendizado não-supervisionado que busca aprender a distribuição conjunta dos dados, com o objetivo de superar as limitações dos métodos baseados em MCMC no que se refere ao número de atributos e suas respectivas classes.

2.3.7 Trabalhos em português

Foram identificadas poucas aplicações dos procedimentos de geração de populações sintéticas no Brasil ou em língua portuguesa. As aplicações identificadas, todas no âmbito acadêmico, utilizaram procedimentos distintos das principais abordagens encontradas em planejamento de transportes que foram apresentadas ao longo do presente capítulo.

Pianucci *et al.* (2019) desenvolveram um método para a modelagem da geração de viagens combinando o uso de redes neurais artificiais e população sintética. Para a geração da população sintética, os autores utilizaram o Método Monte Carlo (MMC), introduzido por Birkin e Clarke (1988). A modelagem foi elaborada em Visual Basic (VBA) e utilizou dados agregados por setores censitários e microdados do Censo Demográfico de 2010 (IBGE, 2010).

O Método de Monte Carlo (MMC), popular no campo da geografia (Harland *et al.*, 2012) e utilizado por Pianucci *et al.* (2019), também é conhecido por “probabilidades condicionais”, e tem sua população construída atributo a atributo. Harland *et al.* (2012) realizaram uma comparação entre três metodologias: probabilidades condicionais, *deterministic reweighting* e *simulated annealing*. Os autores constataram que a abordagem de probabilidades condicionais apresentou desempenho inferior ao método de *simulated annealing* de Voas e Williamson (2000).

Ribeiro (2011) buscou criar um modelo para estimação da geração e distribuição de viagens interurbanas. A geração de populações sintéticas foi um dos passos adotados para isso, sendo utilizados apenas dois atributos para indivíduos (idade e posição na estrutura domiciliar) além

do tamanho dos domicílios. Ribeiro (2011) utilizou o procedimento proposto por Miyamoto *et al.* (2010), que também é um Método de Monte Carlo. O método adotado se baseia na composição de domicílios em função dos atributos dos agentes selecionados.

2.3.8 Fora da academia

Em 2017, Ory (2017) publicou a matéria “A first step toward creating a digital planning laboratory is populating it”⁹ (“O primeiro passo para a criação de um laboratório digital de planejamento é alimentá-lo com uma população”, em uma tradução livre), veiculada no blog da Sidewalk Labs – uma empresa do grupo Alphabet, holding que possui a Google. Em uma linguagem acessível, porém técnica, a matéria destaca a importância de se criar uma população virtual que reflita as características reais da população.

O modelo em desenvolvimento, intitulado *Doppelgänger*¹⁰, foi produzido em código aberto no GitHub, utilizando técnicas de redes bayesianas e otimização convexa – creditando o trabalho de Vovsha *et al.* (2015), citado anteriormente. Entretanto, o projeto, ao menos em sua versão de código aberto, aparenta ter sido abandonado, sendo realizada sua última contribuição em março de 2018.

A Sidewalk Labs recentemente anunciou o Replica, “uma ferramenta que permite que planejadores explorem como, quando, onde e porque as pessoas se locomovem em uma região” (Sidewalk Labs, 2018). A ferramenta utiliza um modelo de comportamento relativo à demanda (*travel behaviour model*) calibrado a partir de dados de *smartphones* e uma população sintética.

Em 2019 estavam sendo desenvolvidos projetos pilotos para a ferramenta Replica em Portland¹¹, Chicago e Kansas City. Conforme destacado em artigo da The Intercept¹², as autoridades de transporte, entretanto, não estão seguras das fontes das informações a serem utilizadas. Foram buscados artigos científicos relacionados ao Replica e ao seu gerador de populações sintéticas, porém não se encontrou nenhuma referência.

⁹ Disponível em: <<https://medium.com/sidewalk-talk/a-first-step-toward-creating-a-digital-planning-laboratory-is-populating-it-beeb87d485f1>>. Acesso em: 9 abr. 2019.

¹⁰ *Doppelgänger* é uma lenda germânica. A palavra é uma junção das palavras alemãs *doppel* (que significa duplo, réplica ou duplicata) e *gänger* (andante, ambulante ou aquele que vaga).

¹¹ Termos do acordo disponíveis em: <http://opb-imgserve-production.s3-website-us-west-2.amazonaws.com/original/replica_portland_term_sheet_1544652344627.pdf>. Acesso em: 9 abr. 2019.

¹² Artigo disponível em: <<https://theintercept.com/2019/01/28/google-alphabet-sidewalk-labs-replica-cellphone-data/>>. Acesso em: 9 abr. 2019.

A empresa canadense INRO, que desenvolve o software de modelagem de demanda Emme, anunciou para a nova versão do software¹³ a inclusão de um gerador de populações sintéticas, podendo ser utilizado como “primeiro passo de um modelo baseado em agentes”. A Figura 8 indica recortes das imagens presentes na notícia. É importante destacar que o Emme é um software proprietário (pago) e um dos mais utilizados no mundo para a estimação de demanda através de modelos macroscópicos.

Figura 8 – Recortes da matéria “What’s New in Emme 4.5”
Fonte: INRO (2020)

2.4 ADAPTAÇÕES E APLICAÇÕES

Lim (2019) realizou uma adaptação do PopGen para a Austrália, incorporando procedimentos iniciais para o tratamento dos dados (de forma a adaptá-los para o formato exigido pelo gerador estadunidense) e de pós-processamento (para completar a população devido à falta de domicílios/pessoas na população sintética gerada pelo procedimento IPU). A autora destacou as dificuldades enfrentadas pelo fato de o PopGen ter sido desenvolvido para ser utilizado nos EUA.

Sallard *et al.* (2020) também apresentaram uma metodologia para criação de uma população sintética para a RMSP. Apesar de o nome do artigo a princípio sugerir que se trata apenas da

¹³ “News: What's New in Emme 4.5”. Publicado em: 14 dez. 2020. Acesso em: 15 dez. 2020.

criação da população sintética¹⁴, os autores apresentam uma metodologia completa para a produção de um “cenário de mobilidade final”, incluindo a combinação da população sintetizada com os diários de viagem da Pesquisa OD 2017 e a associação de locais de estudo e trabalho, obtidos através do Open Street Map, a estes indivíduos.

Sallard et al. (2020) utilizaram a plataforma “eqasim”, detalhada em Hörl e Balac (2021). Em sua etapa de geração da população sintética, os agentes são expandidos diretamente através dos pesos amostrais fornecidos nos microdados e posteriormente arredondados por *stochastic rounding*. Segundo Hörl e Balac (2021), a adoção de “possivelmente um dos métodos mais simples” é justificável devido ao fato da amostragem dos Censos ser relativamente grande quando comparada a outras fontes. Entende-se que isto é também possível, neste caso, devido ao fato de que a população gerada é posteriormente alocada a domicílios pelo “eqasim”, ganhando assim a resolução espacial necessária para o modelo.

2.5 VALIDAÇÃO

2.5.1 *Ground truth* e validação interna

Em 2011, Müller e Axhausen aplicaram o procedimento de geração de populações sintéticas HIPF na Suíça, país em que o registro completo da população é disponibilizado para pesquisadores, permitindo assim que o método fosse verificado frente à *ground truth* (isto é, uma base de dados real desagregada). A mesma base de dados foi utilizada para a elaboração dos testes e validação por Farooq *et al.* (2013). Com os dados reais, estes autores puderam avaliar não apenas a aderência da população simulada em relação aos totais de controle, mas também as suas distribuições conjuntas – possibilitando assim a realização de uma validação mais robusta do que a feita nos demais estudos que não dispunham desta informação.

Entretanto, diferente da Suíça e da Escandinávia (Voas e Williamson, 2000), o registro completo da população raramente é divulgado nos outros países (Farooq *et al.*, 2013) – e é exatamente isso o que justifica o tema de pesquisa de geração de populações sintéticas. Dessa forma, na ausência de uma base real desagregada, a validação dos resultados torna-se um desafio. Lim (2020) afirma que esta limitação para a validação é destacada em diversos artigos da área.

¹⁴ Entretanto, segundo Lovelace *et al.* (2015), os termos *population synthesis* e *spatial microsimulation* são por vezes utilizados como sinônimos.

Sun *et al.* (2015), para avaliar os resultados frente a uma população real, consideraram a amostra disponível de Singapura (cerca de 9.600 domicílios entrevistados em uma pesquisa de mobilidade) como o universo e utilizaram uma parcela dessa amostra como microdados (no caso, 96 observações para uma amostra de 1%).

O que é feito, na maioria das demais aplicações pesquisadas, é a comparação entre os totais agregados da população sintetizada e os totais de controle, para uma mesma escala geográfica, através de diversas métricas de qualidade de ajuste – ou seja, uma “validação interna”.

2.5.2 Procedimentos e pós-processamento

Além das métricas, Ye *et al.* (2009) recomendam que, para métodos que utilizam simulações de Monte Carlo (que são estocásticas), sejam geradas diversas bases de populações sintéticas e selecionada aquela com maior ajuste aos atributos no nível do indivíduo.

Outros autores, como Lim (2020) e Casati *et al.* (2015), incorporaram passos de “pós-processamento” após verificarem, na etapa de validação, que as bases geradas não foram satisfatórias.

Na ausência de uma base real desagregada, Voas e Williamson (2000) apresentam possibilidades de avaliação de desempenho além da avaliação das contagens das incidências para as variáveis controladas na população gerada. Os autores apresentam três “testes”: (i) o comum, comparando variáveis controladas; (ii) a avaliação da combinação de variáveis (idade por sexo, por exemplo) em relação a totais de controle de variáveis combinadas, quando disponível no Censo; e (iii) a avaliação de variáveis não-controladas porém para as quais há totais de controle disponíveis. É importante destacar que variáveis cruzadas podem também ser utilizadas como dado de entrada no modelo (ao usar como controle combinações de categorias de dois atributos).

Conforme destacado por Voas e Williamson (2000), uma outra utilidade da população sintética é a análise de variáveis combinadas não disponíveis no Censo. Ou seja, é possível avaliar, com base na população sintética gerada, por exemplo, a parcela de domicílios composta exclusivamente por pessoas entre 21 e 40 anos e cuja ocupação seja de “profissionais das ciências e das artes” em determinada zona. O segundo teste de Voas e Williamson (2000), supracitado, se mostra especialmente importante quando há o intuito de utilizar a população sintética gerada para estes fins.

2.5.3 Métricas de qualidade de ajuste

Voas e Williamson (2001) apresentaram um artigo específico para a comparação de métricas de qualidade de ajuste para populações sintéticas, avaliando três “famílias” de métricas: qui-quadrado, *z-score* e métricas derivadas do campo de teoria da informação. Os autores destacam que a escolha da métrica não é simples, citando complexidades relacionadas à: (i) absoluto x relativo, considerando opções de ponderação para aquelas relativas; (ii) predominância de “lado” dos erros; e (iii) variabilidade do erro.

Em relação ao qui-quadrado, Voas e Williamson (2001) destacam que o cálculo apresenta problemas para lidar com valores esperados (que estão no denominador da equação) menores que 5 – e especialmente iguais a 0. Para o segundo caso, a equação de Neyman-Pearson pode ser uma alternativa – ao substituir o denominador pela média simples entre os valores observados e esperados. No grupo do qui-quadrado, os autores citam também a estatística G^2 (teste de razão de verossimilhança) – que, assim como o qui-quadrado, apresenta problemas quando os valores das células são pequenos.

Como conclusão, Voas e Williamson (2001) afirmam que não há um conceito único de ajuste para o qual todos os modelos podem ser julgados e aconselham que se leve em consideração o *fitness for purpose* – isto é, a adequação à aplicação em questão – para eleger as métricas a utilizar.

Moreno e Moeckel (2018) indicam que são comumente utilizadas como medidas de qualidade de ajuste na área de geração de populações sintéticas: diferença média relativa absoluta (ou AARD, do inglês *Average Absolute Relative Difference*), raiz do erro quadrático médio (ou RMSE, do inglês *Root Mean Square Error*), diferenças de contagens e erros relativos. Os autores apresentam os resultados na forma de tabelas (para cada atributo, acumulando as zonas) e mapas (para cada zona e um único atributo por mapa).

Ye *et al.* (2009) utilizam o teste do qui-quadrado para comparar estatisticamente duas distribuições, e afirmam que a diferença média relativa absoluta (AARD) não é adequada para avaliações envolvendo pequenas áreas por não expressar as diferenças de magnitude. Os autores utilizam esta métrica descartando os valores nulos (desconsiderando estas células do cálculo do qui-quadrado e dos graus de liberdade para obtenção do valor-p).

2.6 ESQUEMA GENÉRICO DE RECONSTRUÇÃO SINTÉTICA

2.6.1 IPF e IPU

São esquematizados na Figura 9 os passos principais previstos para a geração da população sintética por métodos baseados em IPF, sendo destacadas as duas etapas principais: ajustamento e geração. Este procedimento foi esquematizado por ser o mais utilizado na literatura e na prática.

Apesar de não indicado na figura, pode haver também um procedimento de pós-processamento após a geração, de forma a melhorar a qualidade de ajuste dos resultados em relação aos dados de referência.

Figura 9 – Passos principais previstos para a geração da população sintética por métodos baseados em IPF

Após a obtenção dos microdados (para a matriz semente a ser ajustada) e dos agregados (totais por zona, a serem utilizados como controle), deve-se definir os atributos e respectivas categorias a utilizar. Então, pode-se filtrar a matriz semente e selecionar os totais de controle das zonas a sintetizar para os atributos e categorias de interesse.

Na sequência, ocorre o procedimento de geração da população sintética, composto por duas etapas. Na etapa de ajustamento, os pesos da matriz semente são alterados para respeitar os totais de controle de cada zona. Após o ajuste, há um processo de arredondamento dos valores resultantes, para gerar um número inteiro de domicílios ou indivíduos e possibilitar, na etapa seguinte de geração, a amostragem por simulações de Monte Carlo. Devido à natureza estocástica das simulações de Monte Carlo (o resultado deve ser diferente em cada simulação realizada), o procedimento é normalmente realizado múltiplas vezes, gerando várias bases de populações sintéticas.

Por fim, deve-se realizar a validação dos resultados. A partir da qualidade de ajuste, pode-se (i) selecionar a base gerada com os melhores resultados e/ou (ii) testar o uso de diferentes atributos e eventualmente reduzir ou alterar as categorias.

2.6.2 Maximização de entropia

O procedimento baseado em maximização de entropia de Vovsha *et al.* (2015) – e sua evolução de Paul *et al.* (2018) – pode ser entendido de maneira similar ao descrito no item anterior. A diferença principal se encontra no “núcleo” do procedimento: enquanto os procedimentos IPF e IPU se baseiam nos processos de ajustamento e geração, os supracitados utilizam como processos principais o balanceamento (análogo ao ajustamento) e a “integralização” no lugar da geração, sendo desnecessária a etapa de geração. O detalhamento deste procedimento, por ter sido o selecionado para aplicação no presente estudo, pode ser encontrado no Capítulo 5.

3. CÓDIGOS DISPONÍVEIS E SELEÇÃO

Após a revisão de diferentes métodos empregados para a geração de populações sintéticas realizada no Capítulo 2, é apresentado neste capítulo o levantamento dos códigos de programação disponíveis publicamente, seguido pela discussão acerca da escolha do software a utilizar para geração de populações sintéticas no Brasil: o PopulationSim, descrito em Paul *et al.* (2018a).

3.1 CONSIDERAÇÕES INICIAIS

Lim (2019), após identificar mais de uma dezena de geradores de populações sintéticas, afirma que muitos destes procedimentos são exageradamente complexos, carecem de detalhes sobre suas implementações e de transparência na validação, dificultando assim sua utilização por terceiros.

Ao se referirem a modelos baseados em atividade, Sallard *et al.* (2020) afirmam que muitos não são de “código aberto” ou carecem de documentações – e destacam que uma exceção é o ActivitySim, do qual o PopulationSim desenvolvido por Paul *et al.* (2018a) faz parte.

3.2 CÓDIGOS DISPONÍVEIS

Müller (2017) realizou a implementação de diversos procedimentos consagrados para a etapa de ajustamento, utilizando a linguagem de programação R: IPF hierárquico (HIPF), IPU, maximização de entropia e *generalized raking*. Seus códigos estão disponíveis em repositório aberto do GitHub¹⁵.

O código do projeto SILO (*Simple Integrated Land Use Orchestrator*), que inclui o gerador de populações sintéticas aperfeiçoado por Konduri *et al.* (2016), está disponível sob a licença *GNU General Public License*¹⁶ – um tipo de licença *copyleft*, que exige que projetos derivados sejam distribuídos pela mesma licença. O código foi elaborado na linguagem Java.

O PopGen (MARG, 2016), desenvolvido inicialmente a partir do algoritmo IPU de Ye *et al.* (2009), foi criado para aplicação no Arizona, EUA. O gerador PopGen possui duas versões, PopGen 1.1 e PopGen 2.0, ambas em código aberto. A primeira versão é mais amigável (possui

¹⁵ Müller, K. (2016) MultiLevelIPF: Implementation of algorithms that extend IPF to nested structures. R package version 0.3-2. Disponível em: <<http://krlmlr.github.io/MultiLevelIPF>>. Acesso em: 9 abr. 2019.

¹⁶ Os termos para a licença GNU GPL encontram-se disponíveis em: <<http://www.gnu.org/licenses/gpl.html>>. Acesso em: 10 abr. 2019.

uma interface gráfica para o usuário e baixa e processa dados do censo dos EUA automaticamente), além de possibilitar ajustes para superar inconsistência entre totais de controle nos níveis de indivíduo e domicílio. Entretanto, esta versão permite variáveis em apenas uma escala geográfica, algo superado pelo PopGen 2.0 que incorporou os avanços de Konduri *et al.* (2016). Além desta funcionalidade, o PopGen 2.0 é mais eficiente em termos de tempo de processamento, incorporando novos padrões de programação. Assim como o SILO, o PopGen é distribuído sob a licença GNU GPL e hospeda seus códigos fontes em páginas do GitHub¹⁷.

O SynthPop é outra implementação do IPU de Ye *et al.* (2009). O gerador utiliza Python e foca em alto desempenho computacional e “reusabilidade” do código.

Sun *et al.* (2018), que utilizaram uma estrutura hierárquica contendo três modelos baseados em técnicas de aprendizagem estatística, também disponibilizaram seus códigos, escritos para MATLAB, em repositório do GitHub.

Por fim, outros dois códigos estão disponíveis: (i) o PopSyn III, elaborado a partir do algoritmo de Vovsha *et al.* (2015) e (ii) o PopulationSim, apresentado em Paul *et al.* (2018a), elaborado a partir da abordagem de maximização de entropia de Vovsha *et al.* (2015), porém com alguns avanços em relação a este.

A Tabela 2 sumariza os códigos citados que estão abertos, sendo indicados método(s) e linguagem de programação principais utilizados por cada um deles. Os endereços eletrônicos dos *sites* destes projetos e/ou dos repositórios de códigos são disponibilizados no Apêndice B.

Tabela 2 – Relação de projetos com códigos disponíveis e seus respectivos métodos e linguagem de programação principal

Id	Projeto ou Autor(es)	Método(s)*	Linguagem de programação
1	Müller (2017)	ML IPF	R
2	SILO	IPU	Java
3	PopGen 2.0 (Maricopa)	IPU / Ent	Python
4	SynthPop	IPU	Python
5	Sun <i>et al.</i> (2018)	HMM	MATLAB
6	PopSyn III	Ent	R
7	PopulationSim	Ent	Python

* ML IPF: Multi-level Iterative Proportional Fitting / IPU: Iterative Proportional Updating / HMM: Hierarchical Mixture Model / Ent: Maximização de Entropia

¹⁷ Versão 2.0 disponível em <<https://github.com/foss-transportationmodeling/popgen>>. Acesso em: 10 abr. 2019.

3.3 A ESCOLHA DO POPULATIONSIM

Após o levantamento de diversas metodologias no Capítulo 2 e códigos disponíveis neste capítulo, foi escolhido o PopulationSim como o código principal a ser utilizado no presente trabalho. A escolha deste software se baseou em alguns critérios, como a qualidade do algoritmo, a disponibilidade de documentação, a plataforma de compartilhamento de código, a linguagem de programação, a qualidade do código.

A qualidade do algoritmo está vinculada ao fato deste ser uma evolução do artigo seminal de Vovsha *et al.* (2015). O procedimento proposto, baseado em maximização da entropia, incorpora a possibilidade de definir a importância de cada atributo e utilizar totais de controle em múltiplos níveis geográficos e imperfeitos – isto é, cujos totais em diferenças escalas podem divergir. O procedimento foi citado como inspiração pelos desenvolvedores da ferramenta de código fechado Replica (Ory, 2017) – que destacam a possibilidade de definir “pesos subjetivos” para a importância de diferentes atributos como um diferencial de procedimentos de maximização de entropia frente a métodos baseados em IPF ou IPU. Assim como os procedimentos mais recentes baseados em IPF, o procedimento de Paul *et al.* (2018a) possibilita o uso de controles tanto para domicílios como para indivíduos e em múltiplas escalas geográficas – porém evita os erros decorrentes da utilização de simulações de Monte Carlo ao utilizar o procedimento de “integralização”.

Em relação à documentação, o PopulationSim dispõe de um artigo científico publicado no TRB (Paul *et al.*, 2018a), um memorando técnico (RSG, 2017) e um site de documentação (PopulationSim, 2020). Além disso, o código é hospedado na plataforma GitHub, destinada à realização de projetos colaborativos, onde também podem ser vistas discussões sobre possíveis problemas e melhorias para o software, além do histórico de atualização de versões.

O algoritmo do PopulationSim foi escrito em Python, uma linguagem de fácil leitura e alto poder de processamento, que se popularizou ao longo da última década, especialmente para aplicações relacionadas a análise de dados. O código, conforme apontado no memorando técnico do projeto (RSG, 2017), segue uma série de boas práticas de programação e possui uma alta densidade de anotações para facilitar o entendimento, conforme exemplificado na Figura 10.

O algoritmo utiliza arquivos “.csv” para dados de entrada – de fácil leitura e manipulação através de blocos de notas e softwares de planilhas – e arquivos de extensão “.yaml” para

configurações. Arquivos “.yaml” podem ser lidos com o bloco de notas do Windows, como um “.txt”, e apresentam uma estrutura simplificada para alteração de configurações e dados de entrada, sem exigir que o usuário acesse os códigos principais em Python para a alteração de parâmetros do algoritmo.

```

34 class SimultaneousListBalancer(object):
35     """
36     Dual-zone simultaneous list balancer using Newton-Raphson method with control relaxation.
37
38     Simultaneously balances the household weights across multiple subzones of a parent zone,
39     ensuring that the total weight of each household across sub-zones sums to the parent hh weight.
40
41     The resulting weights are float weights, so need to be integerized to integer household weights
42     """
43
44     def __init__(self,
45                 incidence_table,
46                 parent_weights,
47                 controls,
48                 sub_control_zones,
49                 total_hh_control_col):
50         """
51
52         Parameters
53         -----
54         incidence_table : pandas DataFrame
55             incidence table with only columns for sub controls
56         parent_weights : pandas: Series
57             parent zone balanced (possibly integerized) aggregate target weights

```

Figura 10 – Recortes de código em Python do PopulationSim: alta densidade de anotações
Fonte: PopulationSim (2020)

Também contribuiu para a escolha o fato de que o PopulationSim possui desenvolvedores e uma comunidade que o utiliza, havendo assim potencial para melhorias. Além disso, o PopulationSim opera como um módulo do projeto do modelo baseado em atividades ActivitySim (ActivitySim, 2020) – contratado por um consórcio de agências de transportes de diversos estados dos EUA.

Maiores detalhes sobre o PopulationSim podem ser consultados nos documentos supracitados, assim como no site de documentação do projeto – além do descrito nos capítulos seguintes deste trabalho, em especial no Capítulo 5.

4. FONTES DE DADOS, ATRIBUTOS E VALIDAÇÃO

O presente capítulo apresenta os atributos utilizados em outras aplicações de geração de populações sintéticas encontradas na literatura e indica os atributos que serão adotados no presente trabalho, após identificar algumas fontes de dados disponíveis no Brasil. Adicionalmente, no final do capítulo, apresenta as métricas de validação a serem utilizadas.

4.1 INTRODUÇÃO

Um dos passos necessários para a geração de populações sintéticas, conforme indicado no item 2.6, é a definição dos atributos e de suas respectivas categorias a utilizar como controle.

Isto é necessário por dois motivos: (i) o “carregamento” de todos os atributos dos microdados para a população sintética geraria arquivos muito grandes para serem processados por computadores comuns – visto que há centenas de variáveis na base original; e (ii) o uso de muitos atributos controlados (isto é, com totais de controle a serem respeitados) aumenta a complexidade do problema a ser resolvido pelos algoritmos e consequentemente o tempo de processamento¹⁸.

4.2 FONTES DE DADOS E ATRIBUTOS COMUMENTE UTILIZADOS

Segundo Farooq *et al.* (2013), tradicionalmente são utilizados dados de duas fontes para a geração de populações sintéticas: censos e pesquisas de mobilidade. Estas informações estão disponíveis em duas formas: amostras e tabelas de contingências.

Moreno e Moeckel (2018) realizaram um levantamento das variáveis utilizadas em mais de vinte aplicações de geradores de populações sintéticas, observando que todos os trabalhos para transportes incluíam ao menos três variáveis: idade, sexo e número de moradores do domicílio. Outras variáveis comumente utilizadas são: status de emprego, renda, etnia (atributos individuais), número de veículos, número de crianças e tipo de domicílio (atributos dos domicílios). Os mesmos autores citam a importância de reduzir o número de atributos e categorias de forma a diminuir o erro resultante, caso isso não prejudique na aplicação final da população sintética.

¹⁸ Conforme descrito no Capítulo 7, o tempo de processamento do PopulationSim pode ser de horas ou até mesmo dias para um computador comum gerar uma população de milhões de pessoas e domicílios. Aplicações de outros geradores de populações sintéticas também levam várias horas para serem processados.

Ye *et al.* (2009) utilizaram três variáveis no nível do domicílio (tipo de composição familiar, tamanho, renda) e três no nível do indivíduo (sexo, idade e etnia).

Lim (2020) gerou populações sintéticas para quatro cidades australianas, utilizando como atributos: tipo de residência (isolada, geminada, edifício etc.), famílias por domicílio, número de pessoas e número de automóveis como atributos de domicílio; e gênero, idade e status de trabalho como atributos dos indivíduos.

A aplicação do PopulationSim na região sul do estado de Oregon, EUA (SOABM, 2020a), utilizou os seguintes atributos, com controles em três escalas (região, zonas de tráfego e microzonas): idade, categoria de ocupação no trabalho, renda domiciliar, número de pessoas por domicílio, trabalhadores por domicílio, presença de criança e tipo de residência.

4.3 FONTES DE DADOS DISPONÍVEIS NO BRASIL

4.3.1 Censo Demográfico do IBGE

O Censo Demográfico é uma pesquisa realizada a cada dez anos pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2010). Por realizar a pesquisa com virtualmente toda a população, o Censo é a principal base de dados utilizada para informações sociodemográficas no país, sendo a fonte principal de dados do presente trabalho.

Os resultados mais recentes do Censo foram coletados na pesquisa de 2010. O Censo de 2020, que deveria realizar a coleta de dados no ano de redação desta dissertação, foi adiado para 2021 devido à epidemia da Covid-19¹⁹. Sendo assim, a fonte de dados primária para o presente trabalho deve ser o Censo 2010 (IBGE, 2010).

A base de dados intitulada “Agregados por Setores Censitários” do Censo 2010 contém valores levantados junto à totalidade da população e é comumente utilizada para análises sociodemográficas. O setor censitário é a menor divisão geográfica utilizada pelo Censo, normalmente apresentando centenas ou alguns milhares de habitantes em suas áreas. A presente pesquisa deve utilizar esta unidade geográfica como referência, visto que os totais para diversos atributos são disponibilizados neste nível geográfico.

¹⁹ IBGE (2020). Censo 2020 adiado para 2021. Comunicados. Disponível em: <<https://www.ibge.gov.br/novo-portal-destaques/27161-censo-2020-adiado-para-2021.html>>. Acessado em: 15 dez. 2020.

Além da pesquisa do universo da população, faz parte do Censo um levantamento por amostragem, que é aplicado em apenas uma parcela na população, definida em função no número de habitantes de cada município, conforme indicado na Tabela 3.

Tabela 3 – Fração amostral dos domicílios e número de municípios, segundo as faixas de tamanho da população dos municípios

Classes de tamanho da população dos municípios (hab.)	Fração amostral de domicílios	Número de municípios
Até 2.500	50%	260
Mais de 2.500 até 8.000	33%	1.912
Mais de 8.000 até 20.000	20%	1.749
Mais de 20.000 até 500.000	10%	1.604
Mais de 500.000	5%	40
<i>Média/Total</i>	<i>11% (média no Brasil)</i>	<i>5.565</i>

Fonte: IBGE (2010). Adaptado pelo autor.

Nota-se que para municípios com mais de 500.000 habitantes a amostra é de apenas 5%. Visto que a maior parte da população da RMSP reside em municípios nesta faixa (o município de São Paulo corresponde a mais de 50% da população da RMSP), espera-se que a parcela da amostra disponível para a aplicação do presente trabalho seja de pouco mais de 5%.

Ao se comparar os números da Tabela 3 com as amostras adotadas por institutos de estatísticas de outros países, pode-se dizer que as amostras do censo brasileiro são relativamente grandes. Segundo Bierlaire *et al.* (2013), as amostras são normalmente de 5% ou menos da população total.

Os resultados da amostra são disponibilizados para as chamadas “áreas de ponderação”. Segundo o IBGE, “uma área de ponderação é um conjunto de setores censitários que constituem a menor área geográfica para a qual as estimativas provenientes da amostra são avaliadas em termos de precisão”.

4.3.2 Demais fontes

No Brasil, além das bases de dados do Censo, que incluem a informação da semente a expandir (microdados) e totais por setor censitário (agregados), foram identificadas outras duas fontes de dados que poderiam ser utilizadas como totais de controle: pesquisas “origem e destino” e a RAIS.

O uso de totais de controle de uma fonte distinta daquela em que foram coletados os microdados pode ser útil para incorporar informações que podem ser (i) mais atualizadas, (ii) mais confiáveis e/ou (iii) indisponíveis nos totais de controle do próprio Censo.

O IBGE utiliza um questionário mais extenso para entrevistar a parcela da população que integra os microdados²⁰. Por este motivo, parte das informações coletadas na amostra e disponibilizadas nos microdados não possui um total de controle disponível no próprio Censo. É o caso, por exemplo, dos atributos de (i) presença de criança no domicílio, (ii) posse de automóvel e (iii) tipo de ocupação quanto a emprego.

Para estas variáveis há duas opções: mantê-las na matriz semente sem controles ou as controlá-las com base em outras fontes de dados. Este é o caso da variável de presença de criança no domicílio, que foi inferida a partir dos registros de domicílios dos microdados (identificando se havia pessoas com menos de 18 anos no domicílio) e que não era disponibilizada na base de “agregados por setores censitários” (isto é, não é apresentado o número de domicílios com criança em cada setor censitário). Para as outras duas variáveis, são apresentadas a seguir as suas possíveis fontes para totais de controle.

Mais informações sobre as fontes de dados e variáveis utilizadas são apresentadas no Capítulo 6, em que o procedimento de extração dos dados é detalhado.

4.3.2.1 Pesquisas Origem e Destino

Realizadas em algumas cidades e regiões metropolitanas no Brasil, as chamadas Pesquisas de Origem e Destino (comumente chamadas de “pesquisas OD”) constituem uma importante fonte de dados para o entendimento dos deslocamentos intraurbanos, sendo utilizadas como base para o desenvolvimento de modelos de transportes, além de possibilitar a elaboração de matrizes origem-destino de viagens.

Em São Paulo, o Metrô²¹ realizou a sua primeira pesquisa origem e destino em 1967. Desde então e até a presente data, foram realizadas seis edições da pesquisa com intervalos de dez anos, conforme indicado na Figura 11, sendo a mais recente realizada no ano de 2017. Além disso, foram também realizadas outras duas pesquisas entre as três últimas pesquisas OD, chamadas de “pesquisas de mobilidade”, realizadas com uma amostra menor e questionário

²⁰ A título de comparação, o “Questionário Básico” (utilizado para toda a população) tem quatro páginas, enquanto o “Questionário da Amostra” é composto por dez páginas.

²¹ Companhia do Metropolitano de São Paulo

simplificado. O banco de dados da última pesquisa OD realizada em São Paulo foi divulgado no ano de 2019.

Figura 11 – Linha do tempo da pesquisa OD: seis edições, além de duas pesquisas de mobilidade

Fonte: Metrô (2020)

Diferentemente do Censo Demográfico do IBGE, a pesquisa OD não é um censo – isto é, não realiza a pesquisa com a totalidade da população. A pesquisa OD aplica o questionário em amostras de domicílios que são dimensionadas e definidas de forma a possibilitar um nível de confiança aceitável para a pesquisa e mantê-la viável do ponto de vista de execução e financeiro. Para o caso da RMSP, a amostra da pesquisa domiciliar da OD 2017 é de 32 mil domicílios, o que corresponde a aproximadamente 0,5% do total de domicílios da Região Metropolitana.

Para a obtenção de um panorama socioeconômico e de mobilidade da cidade – e para utilizar no desenvolvimento dos modelos tradicionais de demanda por transportes –, esta amostra é expandida em função da representatividade de cada domicílio, função da fração amostral correspondente.

Dentre as variáveis presentes tanto nos microdados do Censo e como na Pesquisa OD 2017 de São Paulo (ODSP), porém indisponíveis nos agregados por setor censitário do Censo, estão a (i) posse ou não de automóvel nos domicílios e (ii) o grau de instrução dos indivíduos. De forma a incorporar uma informação mais recente no modelo e utilizar a funcionalidade de "meta controles" do PopulationSim, os totais estimados para estas duas variáveis da Pesquisa OD de 2017²² foram utilizados como totais de controle na geração de população sintética do presente trabalho.

4.3.2.2 RAIS

A RAIS (Relação Anual de Informações Sociais) foi instituída por decreto federal em 1975 e se trata de um documento a ser entregue pelas empresas, "contendo elementos destinados a

²² A Pesquisa Origem e Destino 2017 de São Paulo é referida por vezes ao longo do presente documento apenas por "ODSP"

suprir as necessidades de controle, estatística e informações das entidades governamentais da área social" (Brasil, 1975). A base de dados formada pela compilação das informações fornecidas nas RAIS é a mais completa base sobre empregos formais do Brasil.

O Censo não apresenta informações sobre emprego na base de “agregados por setores censitários”, visto que não são realizadas perguntas deste tipo no questionário básico. Dessa forma, a base de dados da RAIS seria uma potencial fonte para a obtenção de totais de controle para variáveis de emprego.

Os microdados do Censo apresentam o tipo de ocupação quanto a emprego dos entrevistados, segundo a CBO (Classificação Brasileira de Ocupações) de 2002, mesma classificação utilizada na RAIS. Entretanto, a RAIS apresenta as informações para o município de trabalho do indivíduo – enquanto o Censo é uma pesquisa de base domiciliar. Por este motivo, o uso da base da RAIS para a geração de populações sintéticas fica limitado a casos em que se considere insignificante a parcela de trabalhadores de outros municípios – ou caso se agregue a informação para uma escala em que esta hipótese (número insignificante de trabalhadores residindo fora da região – e vice-versa) seja razoável.

Para a aplicação do presente trabalho, considerou-se que o uso destas informações de forma espacialmente desagregada poderia causar erros significativos devido ao elevado grau de integração entre os municípios que compõem a região metropolitana, em que é comum que um morador resida em um município e trabalhe em outro (algo que, inclusive, faz parte do conceito de região metropolitana). Adicionalmente, entende-se que a agregação desta informação para toda a região metropolitana não melhoraria o modelo – e não seria possível ao se utilizar dados da ODSP para uma escala menor. Isso porque a presente versão do PopulationSim não permite o uso de totais de controles em mais de uma escala meta – o que, neste caso, exigiria que os totais de controle meta da ODSP fossem também agregados para toda a região metropolitana e assim perdessem o seu componente espacial. Por fim, o fato de se tratar de uma base de empregos exclusivamente formais exigiria a realização de algum tipo de tratamento para sua compatibilização com o atributo correspondente do Censo.

Sendo assim, optou-se por não se utilizar a RAIS na presente pesquisa.

4.4 ATRIBUTOS A UTILIZAR NA APLICAÇÃO

Considerando (i) os atributos utilizados em outros geradores de população sintética, descritos anteriormente e (ii) os atributos comumente utilizados como parâmetros de modelos de demanda no Brasil, considerados relevantes do ponto de vista de comportamento de viagens no contexto nacional, foram selecionados os atributos para a aplicação do presente trabalho. Estes atributos são indicados na Tabela 4, acompanhados das escalas em que são realizados seus controles.

Tabela 4 – Atributos escolhidos para serem utilizados na aplicação e escala de seus controles

Nível	Descrição do atributo	Escala para controle
Domicílio	Número de moradores	Setor censitário
	Idade do responsável	Setor censitário
	Renda per capita em salários mínimos	Setor censitário
	Posse ou não de automóvel	Distrito/Município
	Presença ou não de criança	- (não disponível)
Indivíduo	Sexo do indivíduo	Setor censitário
	Idade do indivíduo	Setor censitário
	Grau de instrução	Distrito/Município
	Categoria de ocupação no trabalho	- (não utilizado ¹)

¹ Hipóteses para utilizar a RAIS (descritas no item anterior) não foram consideradas aceitáveis para a RMSP

Para a aplicação, os controles na escala de setor censitário são obtidos na base “Agregados por Setores Censitários” do Censo 2010. Já os controles na escala de distrito ou município são extraídos das tabelas de totais da Pesquisa OD 2017 de São Paulo.

Naturalmente, ressalta-se que estes atributos estão todos disponíveis para todos os domicílios ou indivíduos (a depender da variável) da base de microdados do IBGE.

A origem destes atributos, assim como a correspondência entre variáveis originais de cada fonte utilizada, é discutida no item 6.4, específico para a aplicação do gerador de populações sintéticas realizada.

4.5 VALIDAÇÃO

Baseado no levantamento feito no Capítulo 2 quanto a procedimentos e métricas para validação, são indicadas neste item as verificações a serem feitas para avaliar a qualidade dos resultados gerados.

4.5.1 Procedimentos realizados

A avaliação de qualidade do resultado principal, realizada na maioria dos artigos indicados no Capítulo 2, é a verificação da qualidade de ajuste (*goodness of fit*) entre os totais de controle e o sintetizado agregado, para cada zona e para cada variável, utilizando as métricas indicadas no item 4.5.2.

Utilizando estas mesmas métricas, é feita também a verificação do ajuste entre combinações de atributos e totais, não utilizados como controles, de variáveis combinadas (como por exemplo as combinações entre idade e sexo, disponível na base de agregados por setor censitário), procedimento sugerido por Voas e Williamson (2000).

Adicionalmente, é avaliado como os totais de controle da escala meta (maior que a escala da semente) se comparam com os microdados diretamente expandidos (através do peso amostral), conforme realizado por Vovsha *et al.* (2015).

As métricas descritas no item seguinte são calculadas no Capítulo 8 (i) na escala de área de ponderação, para todas as variáveis controladas; e (ii) na escala de setor censitário, apenas para as variáveis controladas nesta escala.

4.5.2 Métricas de qualidade de ajuste

A semelhança entre a tabela de controles (dado de entrada) e a sintetizada (agregados por zonas) é avaliada no Capítulo 8 através de quatro métricas: (i) o erro médio relativo, EMR; (ii) o desvio padrão do erro médio relativo, “DesvPad”; (iii) a raiz do erro quadrático médio, RMSE; e (iv) o qui-quadrado, “Qui-quad”.

No presente item são indicadas as equações utilizadas, apresentadas com suas variáveis já adaptadas para a avaliação de geradores de populações sintéticas.

4.5.2.1 Erro médio relativo (EMR)

O erro médio relativo (EMR) é dado pela razão entre (i) a diferença de sintetizado e controle e (ii) seu valor de referência (o controle).

$$EMR_{ij} = \frac{S_{ij} - C_{ij}}{C_{ij}}$$

onde i e j são as dimensões das tabelas (zona i e variável j),
S é o sintetizado e C é o controle.

A soma do erro médio relativo nas zonas, para um dado atributo, permite aferir se há um erro sistemático subestimando (valores negativos) ou superestimando (valores positivos) o total de controle.

4.5.2.2 Desvio padrão do erro médio relativo (σ)

A segunda métrica, desvio padrão do erro médio relativo, permite avaliar a dispersão destes valores, algo que pode não ser observado pela métrica anterior caso haja um equilíbrio entre erros positivos e negativos.

O desvio padrão do EMR (σ_j) para um dado atributo é calculado por:

$$\sigma_j = \sqrt{\frac{\sum_i |EMR_{ij} - \overline{EMR}_j|^2}{N}}$$

onde i e j são as dimensões das tabelas (zona i e atributo j), N é o número de zonas e $\overline{EMR}_j$ é a média entre as diferentes zonas do EMR para um dado atributo.

4.5.2.3 Raiz do erro quadrático médio (RMSE)

Para verificar a magnitude das variações em módulo – e destacando aquelas em que há erros maiores – é calculada a raiz do erro quadrático médio (ou RMSE, do inglês *Root Mean Square Error*), dada pela seguinte equação:

$$RMSE_j = \sqrt{\frac{\sum_i (S_{ij} - C_{ij})^2}{N_j}}$$

onde i e j são as dimensões das tabelas (zona i e atributo j),
C é o controle, S é o sintetizado e N é o número de zonas.

4.5.2.4 Qui-quadrado (χ^2)

É também calculado o valor da estatística de teste qui-quadrado, para a realização de um teste de aderência dos atributos controlados. O valor de cada célula (das tabelas de zona versus variável) é calculado através da seguinte equação:

$$\chi^2 = \frac{(C_{ij} - S_{ij})^2}{S_{ij}}$$

onde i e j são as dimensões da tabela (zona i e variável j),
C é o controle e S é o sintetizado.

O qui-quadrado é calculado célula a célula e então (i) agregado por variável, para verificar o ajuste das categorias de cada atributo, ou (ii) por zona, para verificar a qualidade da estimação para a zona como um todo. As equações a seguir indicam estes dois casos, respectivamente.

Qualidade de ajuste para cada variável j:

$$\chi_j^2 = \sum_i \frac{(C_{ij} - S_{ij})^2}{S_{ij}}$$

Qualidade de ajuste para cada zona i:

$$\chi_i^2 = \sum_j \frac{(C_{ij} - S_{ij})^2}{S_{ij}}$$

Após este cálculo, o teste de aderência entre as distribuições (por exemplo, resultados das zonas para uma determinada variável) pode ser realizado, através da comparação entre a estatística qui-quadrado calculada e a distribuição de referência. Um valor-p²³ de 95% (isto é, a probabilidade de que uma distribuição seja independente em relação à outra) é comumente utilizado (Voas e Williamson, 2000) como valor de referência para a avaliação da hipótese nula de dependência entre as distribuições – e verificar se há evidência ou não de que elas são similares.

²³ O valor-p também é chamado de probabilidade de significância

5. POPULATIONSIM

Neste capítulo é apresentado o gerador de populações sintéticas selecionado para a aplicação do presente trabalho, PopulationSim. Na sequência, é discutida a abordagem teórica utilizada pelo *software*, baseada no trabalho de Vovsha et al. (2015). Por fim, o procedimento de validação disponível junto ao gerador de populações sintéticas é brevemente descrito.

5.1 SOBRE

O PopulationSim é uma plataforma aberta para a geração de populações sintéticas (PopulationSim, 2020). Conforme indicado no capítulo anterior, o PopulationSim faz parte de um projeto maior de um software de modelagem baseada em atividades: o ABM ActivitySim, um projeto patrocinado por um consórcio de agências de transportes de diversos estados dos EUA, atualmente contando com 9 organizações²⁴ indicadas na Figura 12 (ActivitySim, 2020).

Figura 12 – Logotipo e agências de transporte patrocinadoras do projeto do ActivitySim
Fonte: ActivitySim (2020)

Foi desenvolvido pela consultoria RSG e nasceu do desejo do *Department of Transportation* de Oregon em construir uma plataforma aberta e compartilhada e que pudesse ser facilmente adaptada para as necessidades de planejamento de nível estadual, regional e urbano.

O PopulationSim exige, para sua aplicação, a instalação de Python e *libraries* utilizadas pelo procedimento. Os desenvolvedores sugerem a instalação do Anaconda para seu uso, um pacote de gestão e instalação de complementos de programação. Em sua atual versão, o PopulationSim não possui interface gráfica, sendo necessário que o usuário edite os arquivos de configuração através de leitores de texto e execute o algoritmo através do *prompt* de comando.

²⁴ *Metropolitan Planning Organizations* (MPOs) e *Departments of Transportation* (DOTs) dos estados da Califórnia, Geórgia, Oregon e Michigan e Washington – além da *Association of Metropolitan Planning Organizations*, a AMPO.

5.2 DOCUMENTAÇÃO

O PopulationSim tem seus códigos hospedados no repositório GitHub, assim como sua documentação. Além do código e documentação, o projeto utiliza ativamente até o presente momento a funcionalidade de *issues* do GitHub, em que os usuários e desenvolvedores identificam potenciais problemas com o software. O PopulationSim teve sua última modificação em outubro de 2020, dois meses antes da redação deste texto.

Além dos recursos *online*, há um relatório preliminar de uma versão anterior do algoritmo (RSG, 2017) e um artigo científico (Paul *et al.*, 2018a). O trabalho de Vovsha *et al.* (2015) contém o procedimento utilizado como base para Paul *et al.* (2018a) e também pode ser consultado.

5.3 METODOLOGIA

5.3.1 Visão geral

O algoritmo do PopulationSim é uma evolução do procedimento proposto por Vovsha *et al.* (2015) e, dessa forma, está fundamentado no conceito de maximização da entropia. Assim como a maioria dos geradores de populações sintéticas mais atuais, permite o uso de controles tanto para variáveis referentes a domicílios como a indivíduos – além de possibilitar a utilização de controles em múltiplas escalas geográficas.

A Figura 13 apresenta o funcionamento do algoritmo para uma situação simplificada em que são utilizados totais de controle em apenas uma escala, inferior àquela da matriz semente.

Figura 13 – Processos do PopulationSim para duas escalas geográficas
 Fonte: Paul *et al.* (2018a). Adaptado pelo autor

O procedimento, para este cenário de dados de entrada em apenas duas escalas, inicia com a agregação dos totais de controle (disponíveis na escala de zonas de tráfego no exemplo do software – ou setor censitário do Censo para a aplicação no Brasil) para a escala em que são disponibilizados os microdados (as chamadas PUMAs²⁵ do Censo estadunidense – ou áreas de ponderação do Censo brasileiro).

A seguir são aplicados, em sequência, dois dos principais algoritmos do PopulationSim: o balanceamento (redefinição dos pesos fornecidos pelo censo para cada domicílio, na amostra) e a integralização (transformação destes pesos em números inteiros). Ambos os processos são formulados como problemas de programação linear com restrições, buscando os valores ótimos das variáveis de decisão com base na função objetivo e restrições definidas (apresentadas no item seguinte).

Com os novos pesos de cada domicílio da base de microdados, devidamente “integralizados”, os algoritmos de balanceamento e integralização são novamente aplicados um após o outro – porém desta vez utilizando os totais de controle “puros” (isto é, sem agregação) como restrições do problema de programação linear. Como resultado deste procedimento, têm-se os pesos finais dos domicílios para cada setor censitário.

A evolução em relação ao procedimento original de Vovsha *et al.* (2015) se refere à correção do erro da última zona sintetizada. A melhoria é obtida através da implementação, tanto para o processo de balanceamento como para o de integralização, do cálculo simultâneo para todas as zonas que compõem cada área de ponderação – evitando assim que os erros residuais sejam absorvidos pela última zona.

5.3.2 Componentes principais

Os processos de balanceamento e integralização (conversão de números decimais para inteiros), citados anteriormente, podem ser considerados os dois algoritmos principais do procedimento do PopulationSim e são descritos com maiores detalhes nos subitens a seguir.

Quando se deseja utilizar controles em uma escala geográfica maior que a semente (a chamada escala meta), um terceiro algoritmo é utilizado pelo procedimento: a fatoração dos controles da escala meta. Este algoritmo também é descrito a seguir.

²⁵ *Public Use Microdata Areas*

A Figura 14 esquematiza os nomes traduzidos dos três componentes principais do PopulationSim supracitados.

Figura 14 – Componentes principais do PopulationSim

5.3.2.1 Balanceamento

O algoritmo de balanceamento, na primeira vez em que é aplicado no PopulationSim, processa a lista de domicílios dos microdados com seus respectivos pesos amostrais²⁶. Estes pesos amostrais são números decimais, calculados pelo próprio Censo, que quando multiplicados pelos atributos de cada domicílio, devem atender os totais de controle correspondentes das áreas de ponderação.

Para a aplicação do balanceamento, tabelas de incidência são geradas para os domicílios de cada área de ponderação, sendo indicadas as contribuições de cada domicílio (n) para cada variável (i) – sendo que as variáveis, como definidas aqui, correspondem às categorias dos atributos, sejam elas referentes a domicílio (como o número de moradores) ou indivíduo (como a idade). A Tabela 5 indica um exemplo de uma tabela de incidência.

²⁶ O algoritmo de balanceamento é aplicado, no mínimo, duas vezes pelo PopulationSim, conforme Figura 13. É aplicado três vezes caso se utilize controles na escala meta (conforme Figura 15) e quatro vezes, conforme exemplo do pôster de Paul *et al.* (2018b), caso se utilize, além da escala meta, mais de uma escala abaixo da área de ponderação. Em todas as aplicações exceto na primeira, deve-se entender “peso amostral” como “peso inicial”, sendo este a saída do passo anterior que deve ser atualizado.

Tabela 5 – Exemplo de tabela de incidência gerada a partir de microdados

Domicílio	Número de moradores (NP)				Idade das pessoas (AGEP)				Peso entrada	Peso saída
	1	2	3	4+	0-15	16-35	36-64	65+		
	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8		
n=1	1							1	20	X ₁
n=2		1			1	1			20	X ₂
n=3			1			1	2		20	X ₃
n=4				1		2	2		20	X ₄
n=5				1	1	3	2		20	X ₅
Controle (A_i)	250	250	250	300	400	1250	1100	250		

Há múltiplos conjuntos de pesos (x_n) que poderiam satisfazer os totais de controle das categorias dos atributos – trata-se de um sistema indeterminado. Entretanto, deseja-se que estes pesos sejam o mais parecido o possível com os pesos amostrais, de forma a minimizar as mudanças na distribuição das variáveis não controladas. Para isso, o procedimento de balanceamento de lista é formulado como um problema de maximização de entropia convexo, com a seguinte função objetivo:

$$\min \sum_n x_n \ln \frac{x_n}{w_n}$$

onde x_n é o peso final e w_n o peso inicial de cada domicílio n .

Sujeita às seguintes restrições:

$$\sum_n a_{in} * x_n = A_i, (\alpha_i),$$

$$x_n \geq 0$$

onde a_i são as incidências e α representa as variáveis duais que dão origem aos fatores de balanceamento.

Com esta formulação, quanto maior for a diferença entre o peso amostral inicial (w_n) e o final (x_n), mais distante de 1 é o valor da razão x_n/w_n – e consequentemente maior é o valor de seu logaritmo natural, que deve ser minimizado.

O problema de otimização descrito pode ser resolvido apenas se as restrições permitirem a existência de uma solução. Entretanto, na prática, os controles nem sempre são perfeitamente consistentes e, nestes casos, não haveria uma solução. Por este motivo, fatores de relaxamento foram introduzidos na formulação para permitir uma solução mesmo quando os controles não forem perfeitos.

Por outro lado, nem sempre é desejável flexibilizar os controles na mesma proporção. Alguns atributos podem ser mais importantes para uma determinada aplicação do que outras. Dessa forma, cada uma das variáveis (categorias dos atributos) tem a elas vinculado um parâmetro chamado de "importância", que multiplica o fator de relaxamento na função objetivo. Quanto maior for o valor da importância de uma variável em relação às demais, menores devem ser os erros obtidos para ela.

Com a introdução de fatores de relaxamento (z) e valores de importância (μ) na formulação, chega-se à seguinte função objetivo:

$$\min_{\{x_n, z_i\}} \sum_n x_n \ln \frac{x_n}{w_n} + \sum_i \mu_i z_i \ln z_i$$

onde z_i é o fator de relaxamento e μ_i o valor da importância de cada variável i

Sujeita às seguintes restrições:

$$\begin{aligned} \sum_n a_{in} * x_n &= A_i * z_i, (\alpha_i), \\ x_n &\geq 0, z_i \geq 0 \end{aligned}$$

Quanto maior for o valor da importância, mais o fator de relaxamento é penalizado ao se distanciar de 1.

De forma a reduzir a perda de heterogeneidade na população sintética, o PopulationSim permite a especificação de um valor máximo para restringir os fatores de expansão. Fatores de expansão são calculados pela razão entre o peso final (gerado pelo PopulationSim) e o peso amostral (dados pelo IBGE na base de microdados). O termo “fator de expansão” não deve ser confundido com o peso final do domicílio, que é utilizado para a expansão dos microdados e geração da população sintética.

$$E = \frac{x_n}{w_n}$$

onde E é o máximo fator de expansão permitido,
 x_n é o peso final e w_n o peso amostral de cada domicílio n

O fator de expansão é utilizado então para limitar o peso final dos domicílios – que também não podem assumir valores negativos.

$$0 \leq x_n \leq E * w_n$$

O problema descrito de otimização com restrições lineares pode ser resolvido iterativamente através do método de Newton-Raphson (Vovsha *et al.*, 2015).

Conforme citado anteriormente, o principal avanço do PopulationSim em relação ao procedimento de Vovsha *et al.* (2015) é a correção do erro da última zona, através da implementação dos procedimentos de balanceamento e integralização para todas as zonas de uma dada área de ponderação simultaneamente²⁷ – e não sequencialmente. A Tabela 6 mostra a tabela de incidência de uma área de ponderação formada por três setores censitários, cada um deles com seus respectivos totais de controle (A_{zi}) e pesos associados a cada domicílio (x_{zn}), a serem calculados.

Também são apresentados os pesos de cada domicílio calculados anteriormente para esta área de ponderação (x_n). Estes pesos, que são o resultado da etapa anterior, funcionam agora como controle da somatória dos pesos de cada domicílio para o conjunto de zonas.

Tabela 6 – Exemplo de tabela de incidência com totais de controle e pesos para uma única área de ponderação composta por três setores censitários

		Pessoas no domicílio (NP)				Idade da pessoa (AGE)				Peso	Peso	Peso	Peso
		1	2	3	4+	0-15	16-35	36-64	65+	agregado	Sector 1	Sector 2	Sector 3
		i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	x_n	x_{1n}	x_{2n}	x_{3n}
Domicílio	n=1	1							1	250	x_{11}	x_{21}	x_{31}
	n=2		1			1	1			250	x_{12}	x_{22}	x_{32}
	n=3			1			1	2		250	x_{13}	x_{23}	x_{33}
	n=4				1		2	2		150	x_{14}	x_{24}	x_{34}
	n=5				1	1	3	2		150	x_{15}	x_{25}	x_{35}
Controles	A_i	250	250	250	300	400	1250	1100	250	-			
	A_{1i}	50	50	50	60	80	250	220	50				
	A_{2i}	75	75	75	90	120	375	330	75				
	A_{3i}	125	125	125	150	200	625	550	125				

Fonte: Paul *et al.* (2018a). Adaptado pelo autor.

Ao realizar o balanceamento simultaneamente para todos os setores contidos dentro de uma área de ponderação, é necessário restringir os pesos de cada setor (x_{zn}) para que estes não ultrapassem o total da área de ponderação (x_n). Isso é feito através de uma redução proporcional dos pesos.

²⁷ Porém separadamente para cada área de ponderação

A formulação detalhada do balanceamento simultâneo pode ser encontrada em Paul *et al.* (2018a) e RSG (2017).

5.3.2.2 Integralização

O resultado do balanceamento é um vetor de novos pesos para domicílios. Entretanto, visto que a unidade domiciliar não deve ser fracionada na população sintética, os pesos dos domicílios devem ser, necessariamente, números inteiros. O simples arredondamento dos pesos implicaria na violação de totais de controle. Por essa razão, um novo problema de otimização é formulado para a parte decimal dos pesos, de maneira análoga ao algoritmo de balanceamento.

Para uma determinada zona, tem-se a seguinte função objetivo:

$$\min \sum_n y_n * \begin{pmatrix} \ln \frac{y_n}{x_n}, & \text{se } y_n = 1 \\ 0, & \text{se } y_n = 0 \end{pmatrix}$$

onde y_n é a parcela decimal integralizada e x_n a parcela decimal do peso original de cada domicílio n

E sujeito às seguintes restrições:

$$\begin{aligned} \sum_n a_{ni} * y_n &= \underline{A}_i \\ y_n &= \{0,1\} \end{aligned}$$

onde $\underline{A}_i$ é a soma dos resíduos excluía a parte inteira dos pesos, dada por:

$$\underline{A}_i = A_i - \sum_n a_{ni} * \text{int}(x_n)$$

A formulação detalhada do procedimento simultâneo para a integralização pode ser encontrada em Paul *et al.* (2018a) e RSG (2017).

5.3.2.3 Fatoração dos controles meta

Os processos indicados anteriormente, na Figura 13, se referem a uma situação em que são utilizados controles em apenas uma escala, inferior àquela da matriz semente. Para os casos em que se deseja utilizar também controles em uma escala superior à escala da semente, um processo adicional é necessário: a fatoração dos controles meta.

A Figura 15 indica os processos necessários para quando são utilizados controles em uma escala superior. Ao compará-la com a Figura 13, pode-se observar este processo adicional e a suas implicações: uma aplicação adicional do procedimento de balanceamento anteriormente.

A fatoração dos controles no nível meta tem como objetivo considerar a influência dos controles de uma escala maior que a área de ponderação na geração da população sintética. Para isso, são estimados controles “equivalentes” na escala da matriz semente, explicado através de um exemplo numérico no item seguinte

O processo de transformar os totais de uma escala menor (setores censitários) para uma escala maior (área de ponderação) é trivial: eles são simplesmente agregados. Porém, para totais disponíveis em uma escala maior que a semente (distrito/região), um procedimento mais elaborado é necessário.

Figura 15 – Processos do PopulationSim para três escalas geográficas (superior, semente e inferior) utilizando terminologias do Censo brasileiro
 Fonte: Paul *et al.* (2018a). Adaptado pelo autor

5.3.3 Exemplo de aplicação

5.3.3.1 Introdução

A presente seção apresenta um exemplo numérico, focando no processo de fatoração dos controles no nível meta – e incluindo um passo adicional não presente na atual versão do PopulationSim.

Considere uma região (escala meta), composta por duas áreas de ponderação (escala da matriz semente) com os seguintes domicílios, conforme indicado na Tabela 7.

Tabela 7 – Exemplo de matriz semente de domicílios para duas áreas de ponderação

id	APOND	PesAm	NP	HHAUT
1	1	10	2	1
2	1	4	3	0
3	1	6	2	1
4	2	15	1	0
5	2	9	2	1
6	2	6	3	0

onde APOND: área de ponderação, PesAm: peso amostral,
NP: número de pessoas, HHAUT: variável para posse (1) ou não (0) de automóvel

A aplicação do PopulationSim (deste exemplo) nessa região utilizará controles para os dois atributos para os quais há informações sobre os domicílios: o número de moradores (NP) e a posse ou não de automóvel (HHAUT).

Considere para o exemplo três escalas geográficas, da menor para a maior: setor censitário, área de ponderação e região. O atributo de número de moradores (NP) é controlado na menor escala, a de setor censitário. Já o de posse ou não de automóvel (HHAUT) é controlado na escala de região.

5.3.3.2 Balanceamento inicial

Ao se agregar os pesos amostrais (PesAm) dos domicílios por área de ponderação, tem-se para o atributo de número de pessoas (NP) indicado na Tabela 8.

Tabela 8 – Pesos amostrais agregados por NP e por APOND

NP	PesAm		Σ
	APOND 1	APOND 2	
1	0	15	15
2	16	9	25
3	4	6	10
Σ	20	30	50

São comparados na Tabela 9 (i) os microdados agregados da área de ponderação 1²⁸ com (ii) os totais de controle agregados para a mesma área, dados como exemplo²⁹. Observa-se que há uma diferença, a ser “corrigida” através do cálculo de novos pesos.

Tabela 9 – Comparação entre pesos dos microdados e totais de controle para a APOND 1

NP	Microdados	Agreg. SC
	Σ PesAm	Totais
1	0	0
2	16	8
3	4	12
Σ	20	20

onde “Microdados” corresponde à base de microdados do Censo (amostra) e “Agreg. SC” à base de agregados por setor censitário, também do Censo

Neste momento, o PopulationSim resolve o problema de otimização de balanceamento de lista, considerando não apenas esta variável do exemplo, mas todas aquelas que são controladas nos níveis inferiores à semente. Como resultado desta operação, é incluída uma nova coluna na tabela dos microdados, contendo os pesos preliminares calculados (PesPre), conforme indicado na Tabela 10.

Tabela 10 – Exemplo de matriz semente de domicílios com coluna de pesos preliminares

id	APOND	PesAm	NP	HHAUT	PesPre
1	1	10	2	1	5
2	1	4	3	0	12
3	1	6	2	1	3
4	2	15	1	0	16
5	2	9	2	1	8
6	2	6	3	0	6
Σ	-	50	-	-	50

²⁸ Apresentados na Tabela 8, na coluna PesAm/APOND 1

²⁹ Disponíveis originalmente por setor censitário e agregados para a área de ponderação 1

5.3.3.3 Fatoração dos controles no nível meta

Estes pesos preliminares são agregados, na Tabela 11, para as duas áreas de ponderação que compõem a região tratada no exemplo, em função das categorias do atributo HHAUT, a ser controlado no nível de região. As somatórias por categorias do atributo (última coluna), por conter as únicas duas áreas de ponderação da região, correspondem aos totais da região.

Tabela 11 – Pesos preliminares agregados por HHAUT e APOND

HHAUT	Σ PesPre		Σ
	APOND 1	APOND 2	
0	12	22	34
1	8	8	16
Σ	20	30	50

Observa-se na Tabela 11 que, de acordo com os totais de controle de níveis inferiores, 34 dos 50 domicílios não teriam automóvel (soma de HHAUT = 0). Observa-se também que a parcela de domicílios sem veículos (HHAUT = 0) na área de ponderação 2 é maior que na área de ponderação 1 (22/30 contra 12/20).

São então incluídos, como dados de entrada, os totais de controle no nível meta. Suponha que, segundo uma pesquisa OD realizada em data mais recente, 70% dos domicílios da região têm auto (Σ HHAUT1 = 35) e os outros 30%, não (Σ HHAUT0 = 15). Para considerar esta nova informação, é calculado um fator de ajuste do meta controle (abreviado para FMC), dado pela razão entre os totais no nível meta e no nível inferior, para cada categoria do atributo HHAUT, conforme indicado na Tabela 12.

Tabela 12 – Cálculo do FMC a partir da razão entre totais de controle meta e pesos agregados dos microdados

HHAUT	Microdados	OD	FMC
	Σ PesPre	ToContMeta	
0	34	15	0,44
1	16	35	2,19
Σ	50	50	-

Estes fatores são, na sequência, aplicados na tabela no nível de semente, conforme indicado na Tabela 13.

Tabela 13 – Pesos preliminares agregados por HHAUT e APOND após aplicação do FMC

HHAUT	$\Sigma \text{ PesPre} \cdot \text{FMC}$		Σ
	APOND 1	APOND 2	
0	5,3	9,7	15,0
1	17,5	17,5	35,0
Σ	22,8	27,2	-

5.3.3.4 Complemento do procedimento de fatoração

Observa-se que após a aplicação do FMC, o número de domicílios de cada zona segundo estes controles muda: a área de ponderação 1, por exemplo, passa a conter 22,8 domicílios (Tabela 13) ao invés de 20 (conforme total da Tabela 11, que, por ser calculado a partir dos pesos amostrais, corresponde ao total de controle do Censo). A depender da diferença entre os controles agregados e meta, este erro pode ser ainda mais significativo.

O algoritmo original do PopulationSim, até a data, parava na etapa referente à Tabela 13. Entretanto, entende-se que a expansão do processo de fatoração seria pertinente, através do cálculo e a aplicação de um segundo fator de ajuste (chamado de FMC') para corrigir estes totais, conforme Tabela 14.

Tabela 14 – Cálculo do FMC' a partir da comparação entre pesos do microdados e totais de controle meta após a aplicação do FMC para a região

APOND	Microdados	OD	FMC'
	$\Sigma \text{ PesAm}$	ToContMETA	
1	20	22.8	0.88
2	30	27.2	1.10
Σ	50	50	-

Aplicando o FMC' calculado na Tabela 14, obtêm-se os pesos atualizados da Tabela 15.

Tabela 15 – Pesos preliminares agregados por HHAUT e APOND após aplicação do FMC'

HHAUT	$\Sigma \text{ PesPre} \cdot \text{FMC} \cdot \text{FMC}'$		Σ
	APOND 1	APOND 2	
0	4.6	10.7	15.3
1	15.4	19.3	34.7
Σ	20.0	30.0	-

Observa-se que, após a aplicação do segundo fator, chega-se a (i) uma correção para o total de domicílios por área de ponderação (20 e 30, para as APONDS 1 e 2, respectivamente) e (ii) a um resultado próximo daquele utilizado como dado de entrada para o controle meta, para este exemplo (diferença de aproximadamente 2% entre 15 e 15,3 ou 35 e 34,7). Esta é uma proposição de melhoria do método, que dessa forma manteria o controle pleno sobre o total de domicílios em cada zona.

Esta melhoria não foi implementada no código do PopulationSim, até a presente versão, e, portanto, não prevalece na aplicação realizada.

5.3.3.5 Balanceamento e Integralização

Após a geração dos controles meta fatorizados, os procedimentos de balanceamento e integralização (modelados como problemas de otimização, descritos anteriormente) são aplicados – primeiramente para o nível de semente e, na sequência, para as demais escalas menores (na aplicação do presente trabalho, apenas uma: a escala de setores censitários). Como resultado, têm-se os pesos finais inteiros de cada domicílio em cada zona no nível mais desagregado possível (no caso, setor censitário). Estes pesos, ao serem utilizados para a replicação dos domicílios em cada setor censitário, dão origem às duas bases de dados da população sintética: domicílios e indivíduos.

5.3.4 Passos do procedimento computacional

São brevemente descritos a seguir os onze passos em que o PopulationSim é segmentado, desde a leitura dos dados de entrada à produção das bases de dados de saída (população sintética e arquivos auxiliares), conforme a Figura 16. São mencionados conceitos apresentados ao longo dos demais subitens do presente item.


```
# Steps for base mode
run_list:
  steps:
    - input_pre_processor
    - setup_data_structures
    - initial_seed_balancing
    - meta_control_factoring
    - final_seed_balancing
    - integerize_final_seed_weights
    - sub_balancing.geography=TAZ
    - expand_households
    - summarize
    - write_tables
    - write_synthetic_population
```

Figura 16 – Recorte do arquivo “settings.yaml”: os onze passos do PopulationSim para a geração de populações sintéticas
Fonte: PopulationSim (2020)

Para maiores informações sobre a geração de tabelas “de apoio” preparadas pelo software para auxiliar na geração das populações sintéticas, sugere-se a consulta à documentação online do PopulationSim (PopulationSim, 2020).

1. input_pre_processor

Lê os arquivos de entrada e os transforma em um formato adequado para a manipulação pelo PopulationSim nos passos seguintes (*pipeline tables*). Por padrão, lê os arquivos indicados na seção “input_table_list” do arquivo “settings.yaml”.

2. setup_data_structures

Estrutura os arquivos lidos, já em formato adequado, com as informações necessárias para os outros módulos. Neste passo, a tabela de incidência é construída, arquivos são filtrados para realizar a eliminação de informações desnecessárias e o *switch* de agrupamento de domicílios iguais segundo a tabela de contingência é lido.

Por padrão, a tabela de incidência é criada pelo PopulationSim agregando domicílios considerados iguais segundo as categorias de atributos adotadas como controles³⁰. Caso haja, por exemplo, dois domicílios com um único morador com mais de 65 anos, um com peso 100 e outro com peso 150 – e que estes sejam os dois únicos atributos (número de moradores e idade), os domicílios podem ser agregados e dar origem ao grupo de domicílios 1, com peso 250. Este exemplo é indicado nas tabelas a seguir (Tabela 16 e Tabela 17).

³⁰ Esta opção pode ser ligada ou desligada no arquivo “settings.yaml”

Tabela 16 – Exemplos de microdados de domicílios (esquerda) e pessoas (direita)

id_d	WGTP	NP	id_d	PWGTP ³¹	id_p	AGE
1	100	1	1	100	1	67
2	150	1	2	150	1	80

onde WGTP: peso amostral; d: domicílios; PWGTP: peso amostral dos domicílios e p: pessoas

Tabela 17 – Exemplo de tabela de incidência com agrupamento de domicílios

Grupo do domicílio	Pessoas no domicílio (NP)				Idade das pessoas (AGEP)				WGTP
	1	2	3	4+	0-15	16-35	36-64	65+	
	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	
n=1	1							1	250

As variáveis utilizadas na tabela de incidência (categorias dos atributos) provêm da definição da relação entre as variáveis da semente (idade do indivíduo, por exemplo) e totais de controle disponíveis (pessoas entre 36 e 64 anos que residem nos domicílios de um dado setor censitário). Esta relação é descrita pelo usuário no arquivo “controls.csv”, que é lido pelo PopulationSim no formato indicado na Figura 17.

```
target,geography,seed_table,importance,control_field,expression
p_age_03,SC,persons,500,AGEP3,(persons.AGEP >= 21) & (persons.AGEP <= 30)
```

Figura 17 – Recorte do arquivo “controls.csv”: correspondência entre semente e controle

Pode-se observar no fragmento do arquivo indicado anteriormente: a definição do novo nome para a variável (p_age_03), a escala na qual ela é controlada (SC – ou setor censitário), a que tipo de variável se refere (“persons” – ou seja, indivíduos e não domicílios), sua importância relativa (valor 500), seu nome na tabela de totais de controle (AGEP3) e, por último, a equação utilizada para classificar os indivíduos nesta categoria (registros na matriz semente de indivíduos cujo valor da variável AGEP é maior ou igual a 21 e menor ou igual a 30).

3. initial_seed_balancing

Primeira aplicação do procedimento de balanceamento, descrito anteriormente, em que são atualizados os pesos amostrais (entrada) dos domicílios através da criação dos chamados pesos preliminares (saída), utilizando os totais de controle de escalas inferiores agregados ao nível da semente.

³¹ O peso dos indivíduos é, por definição, o mesmo dos domicílios aos quais eles pertencem

4. meta_control_factoring

Aplicação do procedimento de fatoração, também descrito anteriormente. Neste passo, os totais de controle no nível meta são considerados para influenciar a geração da população sintética.

5. final_seed_balancing

Aplicação do procedimento de balanceamento utilizando os controles fatorados a partir do nível meta (passo 4) e os pesos preliminares, no nível de semente (passo 3), tendo como resultado os pesos “quase” finais (decimais) neste nível.

6. integerize_final_seed_weights

Integralização dos pesos do passo anterior, gerando os pesos finais por área de ponderação, a serem posteriormente desagregados entre as zonas da(s) escala(s) inferior(es).

7. sub_balancing.geography=TAZ

Passo para a aplicação dos procedimentos de balanceamento e integralização no(s) nível(is) especificado(s) – no caso, TAZ (equivalente ao setor censitário). Neste processo, são definidos os pesos dos domicílios em cada TAZ/setor censitário.

8. expand_households

Geração de uma lista de domicílios completa e expandida, sendo indicado, para cada registro (domicílio), o código do setor censitário (ou menor escala geográfica) no qual ela se localiza.

Em outras palavras, trata-se do “esqueleto” da população de domicílios sintéticos: uma tabela cujo número de linhas corresponde ao número final de domicílios já alocados em um determinado setor censitário – porém ainda sem estar ligada com os atributos dos domicílios, a serem obtidos nos microdados.

9. summarize

Geração dos arquivos de síntese da população gerada, sendo realizadas as contagens de domicílios e indivíduos por categoria em cada zona das escalas geográficas inferiores (setor censitário, no caso) e da semente (área de ponderação) – e comparando estes números com os totais de controle. A tabela de pesos de cada domicílio da semente, em cada etapa do PopulationSim, é também gerada neste passo.

10. write_tables

Criação dos arquivos especificados na lista “output_tables” do arquivo “settings.yaml”.

O arquivo “settings.yaml” da extensão para o Brasil do PopulationSim, elaborada neste trabalho, gera por padrão todas as tabelas possíveis, podendo o usuário eliminá-las da lista a gerar. Destacam-se, dentre as tabelas de saída (além da população sintética, produzida no passo seguinte):

- “Summary_TAZ”: população sintética agregada por setor censitário para cada atributo controlado nesta escala (sufixo “_result”) e comparada com os controles (sufixo “_control”), dando origem às variáveis de diferença (sufixo “_diff”)³²;
- “Summary_TAZ_PUMA”: idem ao anterior, porém agregados na escala de área de ponderação e incluindo os atributos controlados no nível meta;
- “Final_summary_hh_weights”: apresenta os pesos dos domicílios em cada passo do procedimento, incluindo o peso amostral (dado na matriz semente), o peso preliminar (calculado para a escala de área de ponderação) e o peso final, convertido para número inteiro;
- “Final_TAZ_weights”: indica o número de replicações de cada domicílio da semente em cada setor censitário. Estes pesos, quando somados para um dado domicílio, correspondem ao valor do peso final (coluna “TAZ_integer_weight”) da tabela “Final_summary_hh_weights” supracitada. Um exemplo desta tabela é indicado na Figura 18.

	A	B	C	D	E	F
1	hh_id	TAZ	balanced_weight	integer_weight	PUMA	REGION
2	3	355030801000001	0.2	0	3550308005039	355030801
3	365	355030801000001	2.3	2	3550308005039	355030801
4	372	355030801000001	1.9	2	3550308005039	355030801
5	377	355030801000001	1.4	2	3550308005039	355030801
6	505	355030801000001	0.9	1	3550308005039	355030801
7	789	355030801000001	0.0	0	3550308005039	355030801

Figura 18 – Recorte do arquivo “TAZ_integer_weight” do PopulationSim: pesos de cada domicílio (hh_id) em cada setor censitário (TAZ)

³² A geração deste arquivo é necessária para a aplicação do procedimento de validação adicional desenvolvido no presente trabalho (item 6.6.2)

11. write_synthetic_population

Cria os arquivos “synthetic_persons.csv” (indivíduos) e “synthetic_households.csv” (domicílios): as duas bases de dados que definem a população sintética, contendo os atributos definidos na seção “Synthetic Population Output Specification” do arquivo “settings.csv”.

5.4 PROCEDIMENTO DE VALIDAÇÃO

O PopulationSim também fornece uma rotina para o cálculo de métricas a serem observadas para a validação. Diferente dos demais códigos com extensão “.py” (isto é, “códigos puros” de Python), esta rotina foi elaborada para Jupyter Notebook, com a extensão “.ipynb” e executada via web. Um documento neste formato é composto por uma lista ordenada de células de entrada e saída que podem conter código, texto, gráficos e figuras – conforme exemplo da Figura 19.

Figura 19 – Recortes de um Jupyter Notebook

Além da diferença absoluta entre células, são calculadas três métricas para a avaliação dos resultados gerados: (i) o erro médio relativo, (ii) o desvio padrão do erro médio relativo e (iii) a raiz do erro quadrático médio (RMSE). O cálculo destas três métricas, que também foram incorporadas no procedimento de validação próprio desenvolvido no presente trabalho, foi indicado no item 4.5.2.

O *notebook* fornecido pelo PopulationSim, além de calcular as métricas, apresenta os resultados através de gráficos, conforme ilustrado na Figura 20.

Figura 20 – Recortes de exemplos de gráficos gerados para validação do PopulationSim
Fonte: PopulationSim (2020)

É importante lembrar que o *expansion factor* do PopulationSim deve ser entendido como o fator de expansão dado pela razão entre o peso final obtido e o peso amostral (dado nos microdados) – e não como simplesmente o peso final calculado (número de vezes que um dado domicílio será replicado na população sintética).

$$\text{Fator de expansão} = \frac{\text{peso final integralizado}}{\text{peso amostral (microdados)}}$$

6. EXTENSÃO DO POPULATIONSIM PARA O BRASIL

O presente capítulo apresenta considerações acerca dos códigos elaborados para auxiliar o usuário no processo de geração de população sintética utilizando o PopulationSim com base em dados disponíveis para o Brasil.

6.1 INTRODUÇÃO

Independentemente do procedimento utilizado, entende-se que os códigos necessários para a geração de populações sintéticas possam ser divididos em três grupos, conforme indicado na Figura 21:

Figura 21 – Classificação dos códigos para a geração de populações sintéticas

Os capítulos anteriores discutiam principalmente códigos do segundo grupo, “códigos de geração da população sintética” – que pode ser entendido como o núcleo desta atividade. Entretanto, por vezes é também necessário processar os dados de entrada (tratamento) e verificar a qualidade dos resultados gerados (validação).

6.2 VISÃO GERAL DOS CÓDIGOS DESENVOLVIDOS

O PopulationSim, desenvolvido nos EUA, foi elaborado para processar dados do Censo estadunidense. Visto que os dados coletados pelo Censo brasileiro são disponibilizados em formatos e agregações diferentes, foram inicialmente elaborados algoritmos para a extração das variáveis (i) de múltiplas tabelas (para o caso dos totais de controle, obtidos a partir da base “Agregados por Setores Censitários” do Censo 2010) e (ii) a partir de arquivos de comprimento fixo (para o caso da matriz semente, obtida a partir dos “Microdados” do Censo 2010). Os nomes dados para estes dois códigos são apresentados na Figura 22.

Figura 22 – Nomes dos códigos para produção do arquivo com totais de controle (esquerda) e da matriz semente (direita)

De forma a facilitar a especificação dos municípios de interesse para a sintetização da população, foi criado um terceiro código para filtrar a base de dados de setores censitários. Este código recebe como dados de entrada, além da base de correspondência geográfica do Censo, um arquivo de texto em que o usuário lista os municípios para os quais se deseja gerar a população sintética. Os nomes dados para estes dois arquivos são apresentados na Figura 23.

Figura 23 – Nomes do arquivo de especificação de municípios a sintetizar (esquerda) e do código para filtrar arquivo com correspondências entre escalas (direita)

Além disso, utilizando a capacidade do PopulationSim de utilizar controles também na escala meta (isto é, em uma escala maior do que aquela em que é disponibilizada a matriz semente), foram criados códigos para a extração e transformação de dados da Pesquisa OD de São Paulo de 2017. O nome do código é indicado na Figura 24.

Figura 24 – Nome do código para produção de controles meta a partir da Pesquisa OD de São Paulo de 2017

Por fim, foi elaborado um código para a validação, cujo nome é indicado na Figura 25. O código realiza o cálculo de métricas de interesse para a verificação da aderência entre os totais de controle e a contagem da população sintetizada, para cada categoria dos atributos nas escalas de setor censitário e áreas de ponderação), gerando planilhas com múltiplas abas para facilitar a verificação dos resultados por parte do usuário.

Figura 25 – Nome do código utilizado para validação

A Figura 26 indica como estes códigos se relacionam entre si e com o procedimento do PopulationSim (item “2. python run_populationsim”).

Figura 26 – Fluxo de ações para a aplicação do PopulationSim no Brasil

6.3 CONSIDERAÇÕES SOBRE OS DADOS DE ENTRADA

As fontes de dados a utilizar na presente aplicação foram introduzidas no Capítulo 4. No presente item são apresentadas discussões acerca da utilização dos dados destas fontes, incluindo a compatibilidade entre as bases geográficas utilizadas e a origem dos dados.

6.3.1 Censo Demográfico do IBGE

No Censo, áreas de ponderação são compostas por conjuntos de setores censitários (normalmente da ordem de grandeza de dezenas de setores por área de ponderação). Visto que a matriz semente e o total de controle na menor escala provêm da mesma fonte (Censo 2010), não há problemas quanto à adequação das variáveis ou das áreas utilizadas.

As figuras a seguir (Figura 27 e Figura 28) apresentam recortes das documentações do Censo, para a amostra (microdados) e para os totais de controle (agregados por setor censitário), respectivamente.

Variáveis do Registro de Domicílios	
VAR	NOME
V6532	RENDIMENTO DOMICILIAR PER CAPITA, EM N° DE SALÁRIOS MÍNIMOS, EM JULHO DE 2010

Figura 27 – Extrato da planilha “Layout Microdados Amostra”, que contém a descrição dos atributos apresentados para domicílios e pessoas
Fonte: IBGE (2010)

6.19 Arquivo Renda dos Domicílios (planilha DomicílioRenda_UF.xls o DomicílioRenda _UF.csv)

NOME DA VARIÁVEL	DESCRIÇÃO DA VARIÁVEL
V005	Domicílios particulares com rendimento nominal mensal domiciliar per capita de até 1/8 salário mínimo
V006	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1/8 a 1/4 salário mínimo
V007	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1/4 a 1/2 salário mínimo
V008	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1/2 a 1 salário mínimo

Figura 28 – Extrato do documento “Resultados do Universo por setor censitário”, que contém a descrição das variáveis utilizadas como totais de controle
Fonte: IBGE (2010)

A indicação de conexão entre as variáveis de renda indicadas na Figura 27 e na Figura 28 (V6532 dos microdados com V005 a V008 dos totais de controle), assim como as dos demais atributos, pode ser verificada no Apêndice C.

6.3.2 Pesquisa OD de São Paulo 2017

A menor unidade geográfica para os quais os dados expandidos da ODSP estão disponíveis é a chamada "zona OD". Na última edição da pesquisa, a RMSP foi dividida em 517 zonas, das quais 342 estão contidas no município de São Paulo e 175 se encontram nos demais municípios que compõem a região metropolitana.

Algumas zonas OD podem estar divididas em mais de uma área de ponderação, sendo cruzadas pelos limites destas. A Figura 29 apresenta um recorte da zona oeste do município de São Paulo em que isso pode ser observado.

Figura 29 – Incompatibilidade (linhas vermelhas pontilhadas) entre zonas OD (cores variadas) e áreas de ponderação (linhas pretas)

Fonte: Metrô (2020) e IBGE (2020). Elaboração própria.

Sendo assim, as próprias zonas OD não podem ser utilizadas como escala meta. Dessa forma, analisou-se na sequência a área geográfica de distritos. A ODSP disponibiliza³³ um arquivo do tipo *crosswalk* de correspondência geográfica, que indica em qual distrito cada zona está contida – o que confirma a possibilidade de se utilizar esta área geográfica, visto que áreas de ponderação também concordam geograficamente com distritos.

³³ Arquivo "Corresp2007_2017.xlsx"

Na Figura 30 é possível observar a concordância entre as zonas OD (cores variadas) e os limites dos distritos (linhas azuis) para o mesmo recorte da Figura 29.

Figura 30 – Compatibilidade entre zonas OD (cores variadas) e distritos no município de São Paulo (linhas azuis)

Fonte: Metrô (2020) e IBGE (2020). Elaboração própria.

Entretanto, para os demais municípios da RMSP, observou-se que algumas zonas OD estão divididas entre dois distritos. A Figura 31 exemplifica dois casos em que isso ocorre: um no município de Mairiporã (esquerda) e outro em Guarulhos (direita).

Figura 31 – Incompatibilidade entre limites dos distritos (azul) e zonas OD (cores variadas) fora do município de São Paulo.

Fonte: Metrô (2020) e IBGE (2020). Elaboração própria.

Sendo assim, para os 38 municípios da RMSP além de São Paulo, a escala meta adotada passa a ser o próprio município – e não o distrito.

6.4 RELAÇÃO DE ATRIBUTOS ESCOLHIDOS

São apresentados nas tabelas a seguir (Tabela 18 e Tabela 19) os atributos presentes nas matrizes sementes tratadas, tanto de domicílio como de indivíduos. Os atributos podem ter duas origens: (i) a base de microdados do Censo ou (ii) podem ser criados no código de tratamento de dados de entrada (como os atributos de “id”, que são códigos únicos de identificação).

Tabela 18 – Atributos de domicílio selecionados para a aplicação na RMSP

Atributos da semente	Descrição do atributo	Nível de controle¹
REGION	Id da região (distrito/município, escala meta)	- (id)
PUMA	Id da área de ponderação	- (id)
hhnum2	Id original (Censo) do domicílio	- (id)
hhnum	Id criado (sequencial e a partir de 1)	- (id)
WGTP	Peso amostral	- (peso)
TIPODOM	Tipo de domicílio (1 = domicílio particular)	-
NP	Número de moradores	SC
HHINCADJ	Renda per capita do domicílio em salários mínimos	SC
AGEHOH	Idade do responsável pelo domicílio	SC
HHAUT	Posse ou não de automóvel	Dist./Mun.
HHMOT	Posse ou não de motocicleta	-
HHCRIAN	Presença ou não de criança no domicílio	-

¹ SC = Setor Censitário, fonte: Censo 2010 | Dist./Mun. = Distrito (SP) ou Município (exceto SP), fonte: ODSP

Tabela 19 – Atributos de indivíduos selecionados para a aplicação na RMSP

Atributos da semente	Descrição do atributo	Nível de controle¹
REGION	Id da região (distrito/município, escala meta)	- (id)
PUMA	Id da área de ponderação	- (id)
hhnum2	Id original (Censo) do domicílio	- (id)
hhnum	Id criado (sequencial e a partir de 1)	- (id)
PWGTP	Peso amostral (igual a respectivo domicílio)	- (peso)
SPORDER	Id do indivíduo em seu respectivo domicílio	-
RELP	Grau de parentesco	-
SEX	Sexo	SC
AGEP	Idade	SC
EDUC	Grau de instrução	Dist./Mun.
OCUP	Categoria de ocupação quanto a trabalho	-
ENQUAD	Tipo de enquadramento no trabalho	-

¹ SC = Setor Censitário, fonte: Censo 2010 | Dist./Mun. = Distrito (SP) ou Município (exceto SP), fonte: ODSP

A Tabela 20 indica os prefixos das variáveis utilizadas como totais de controle, extraídos da base de (i) agregados por setor censitário, do Censo ou (ii) das tabelas de totais da ODSP. A

Tabela 20 a seguir é uma versão compacta da tabela apresentada no Apêndice C, que contém, além das informações indicadas a seguir, as descrições das variáveis segundo as fontes originais e as categorias adotadas – além de incluir os atributos da população sintética que não são controlados³⁴.

Tabela 20 – Atributos utilizados como totais de controle para a aplicação na RMSP

Escala ¹	Nível ²	Código do atributo	Descrição do atributo	Número de categorias	Variáveis controladas
Setor Censitário	Domicílio	HHBASE	Número de domicílios	1	27
		HHSIZE	Moradores no domicílio	6	
		HHAGE	Idade do responsável pelo domicílio	4	
		HHINC	Renda per capita do domicílio em SM ²	6	
	Indivíduo	SEX	Sexo do indivíduo	2	6
		AGEP	Idade do indivíduo	8	
Distrito/Município	Domicílio	HHAUT	Posse de automóvel no domicílio	2	
	Indivíduo	EDUC	Grau de instrução	4	
<i>Total</i>	-	-	-	33	33

¹ Fontes para cada escala: Setor Censitário: Censo 2010; Distrito/Município: ODSP 2017

² SM = Salários mínimos de julho de 2010

Observa-se, pela Tabela 20, que:

- Na escala de setor censitário, 27 variáveis (soma das categorias de atributos) controlam 6 elementos da matriz semente (5 atributos mais a contagem do total domicílios, HHBASE);
- Na escala meta de distrito/município, 6 variáveis (soma das categorias de atributos) controlam 2 atributos da matriz semente;
- No total, 33 variáveis são controladas na aplicação. Deve-se notar que outros atributos estão presentes nos domicílios da matriz semente e podem existir também na população sintética final³⁵, embora sem estarem submetidos a controle.

Por fim, são indicados na Tabela 21 os valores das importâncias adotados para as variáveis dos atributos apresentados na Tabela 20.

³⁴ Indivíduo: OCUP (categoria de ocupação); domicílio: HHMOT (posse de motocicleta), HHCRIAN (presença de criança no domicílio).

³⁵ Em função do que for especificado no arquivo de configurações, conforme indicado na descrição do passo 11 do item 5.3.4.

Tabela 21 – Valores adotados para a “importância” das variáveis dos atributos utilizados

Atributo	HHBASE _n	HHSIZE _n	HHAGE _n	HHINC _n	SEX _n	AGEP _n	HHAUT _n	EDUC _n
Importância	10 ⁹	5000	500	500	5000	500	5000	500

Para o valor da importância da contagem de domicílios (HHBASE) foi adotado um valor de ordem de grandeza muito maior que os demais, de forma a se obter um ajuste excelente para esta variável. Dentre as demais variáveis, o tamanho do domicílio (HHSIZE), sexo (SEX) e posse de automóvel (HHAUT) receberam valores de importância maior que as restantes. A escolha de um valor relativamente elevado para as variáveis do HHSIZE tem como objetivo reduzir o erro na contagem de indivíduos, que é gerada a partir da seleção de domicílios classificados por números de indivíduos; já para as variáveis de sexo e posse de auto o objetivo foi reduzir os erros absolutos para atributos divididos em apenas duas categorias (masculino/feminino e possui/não possui automóvel). Para HHAUT, o fato de esta estar relacionada com duas bases de dados distintas e divergentes é um motivo adicional para a escolha de um valor mais elevado da importância para ela.

6.5 CÓDIGOS PARA TRATAMENTO DE DADOS DE ENTRADA

São apresentados neste item os códigos desenvolvidos para tratamento de dados de entrada. Estes códigos têm, como principal função, a extração das variáveis de interesse para os municípios desejados, aplicando filtros e buscando os dados em bases com diferentes formatos. Também transformam os tipos dos dados e criam variáveis adicionais de forma a serem lidos adequadamente pelo PopulationSim.

Foram desenvolvidos três códigos para selecionar e transformar os dados do censo brasileiro para o formato de entrada do PopulationSim. Com estes códigos já é possível aplicar o modelo para uma região brasileira. Adicionalmente, foi desenvolvido um código para o uso de totais de controle no nível meta na RMSP, utilizando como fonte a pesquisa OD de São Paulo de 2017.

Todos estes códigos são hospedados em um repositório *online* hospedado na plataforma GitHub (Ajauskas, 2021a). Detalhamentos técnicos sobre estes códigos são apresentados na documentação online que acompanha esta dissertação (Ajauskas, 2021b) e através de comentários ao longo dos próprios códigos.

6.5.1 Área a sintetizar (base: Censo 2010)

6.5.1.1 Sobre

O primeiro código, “aAreaSintetizar” é responsável por criar o arquivo “geo_cross_walk.csv”.

De forma a facilitar a especificação dos municípios de interesse para a sintetização da população, o usuário lista os municípios para os quais se deseja gerar a população sintética em um arquivo de texto, “municípios.yaml”.

O código lê o arquivo supracitado, identifica os municípios listados e busca os seus códigos de município segundo uma tabela disponível na documentação das áreas de ponderação. Na sequência, consulta a tabela de correspondências geográficas³⁶ para filtrar os setores censitários que compõem estes municípios e identificar as suas respectivas áreas de ponderação e região (que na aplicação do presente trabalho pode ser município ou distrito – identificados como os primeiros 7 e 9 dígitos do setor censitário, respectivamente).

Este arquivo é posteriormente lido pelos códigos de tratamento da matriz semente e dos totais de controle, tanto no nível de setor censitário como de distrito/região, e utilizado como referência para a delimitação dos setores censitários de interesse para a sintetização da população.

6.5.1.2 Alterações em arquivos de referência

Foi observado, ao analisar os resultados gerados por área de ponderação, que o limite de uma delas não coincidia com o limite de distrito, como pode ser visto na Figura 32.

³⁶ Preparada a partir do arquivo “Composição das Áreas de Ponderação.txt” (parte da documentação dos microdados). Por conter informações do Brasil todo, foi copiada, tratada e corrigida (conforme item 6.5.1.2 a seguir) no arquivo criado “comp.csv”, disponível para download no repositório do GitHub do presente projeto (na pasta “bruto”)

Figura 32 – Recortes do setor censitário alocado em área de ponderação incorreta

Ao analisar o arquivo de referência³⁷, foi encontrado um erro na correspondência entre setor censitário e área de ponderação, que foi devidamente corrigido. Esta correção é especialmente importante pelo fato de que o código “aAreaSintetizar” desenvolvido obtém o “id”³⁸ da escala meta (distrito ou município, a depender do município) a partir do “id” do primeiro setor censitário encontrado no arquivo de referência para uma dada área de ponderação.

O arquivo de referência corrigido³⁹ foi disponibilizado junto aos demais arquivos no repositório do GitHub do presente projeto (Ajauskas, 2021a).

6.5.2 Totais de controle (base: Censo 2010)

6.5.2.1 Sobre

O código “cTotaisControle” cria o arquivo “control_totals_taz.csv”, que contém os totais de controle a serem utilizados pelo PopulationSim para a escala de setor censitário.

Para isso, consulta uma série de tabelas da base “Agregados por Setores Censitário” do Censo 2010. O código contém instruções para localizar, renomear, agregar e transformar⁴⁰ as variáveis de interesse.

³⁷ “Composição das Áreas de Ponderação.txt”, disponível na documentação dos microdados do Censo 2010

³⁸ Código de identificação único

³⁹ O arquivo de referência criado recebeu o nome de “comp.txt”

⁴⁰ Quanto ao tipo de variável – principalmente transformando textos em números inteiros

6.5.2.2 Preenchimentos e modificações com base em premissas

Além das tarefas de extração de dados, o código também preenche dados faltantes (cujos resultados são indicados com um “X” na base de agregados por setor censitário) ou de setores mistos (que contém tanto domicílios particulares como coletivos) a partir de critérios apresentados a seguir.

Dados faltantes: XFLAG

Para o primeiro caso, em que não são apresentados controles para o setor censitário, foi criada uma função para preencher os controles considerando que eles seguiriam uma distribuição dada pela média dos demais setores presentes na base de dados de totais de controle (isto é, dos demais setores censitários que terão suas populações sintetizadas, independente da área de ponderação).

A Figura 33 indica exemplos deste preenchimento para três zonas distintas. É possível observar que todas seguem aproximadamente a mesma distribuição para as categorias do atributo preenchidas, com diferenças devido à (i) arredondamento, sempre feito para o menor número inteiro, e (ii) última categoria, que é calculada de forma a satisfazer o total (através da subtração entre total e a somatória dos valores para as demais categorias do atributo).

Figura 33 – Exemplos de variáveis geradas para três zonas distintas: mesma distribuição

A função criada para o preenchimento de valores indisponíveis (“X”) em função da proporção observada em outros setores censitários foi chamada de “propx” no código desenvolvido e atua nos registros de domicílios cuja variável XFLAG é igual a 1⁴¹.

Setores mistos: MIXFLAG

O segundo caso, para tratamento de setores “mistos”, é necessário devido aos fatos de que a aplicação foi desenvolvida para processar apenas moradores de domicílios particulares – isto é, excluindo domicílios coletivos – e de que a maioria dos totais de controle não são apresentados em função do tipo domicílio (isto é, não existe uma variável que indique o número de moradores de domicílios particulares permanentes com idade entre 21 e 30 anos, por exemplo). Dos atributos considerados, estão disponíveis separadamente para moradores de domicílios particulares apenas sexo (atributo de indivíduo) e número de moradores (atributo de domicílio).

Sendo assim, para os demais atributos utilizados, é necessário ajustar os totais de controle de forma a coincidirem com a quantidade de domicílios particulares de cada setor censitário. Isso é feito através do cálculo e aplicação de um fator de redução, dado pela razão entre domicílios particulares e o total de domicílios de cada setor censitário, que multiplica cada atributo de domicílio. O mesmo é feito para indivíduos, considerando a razão entre moradores de domicílios particulares e todos os residentes do setor censitário.

A Figura 34 exibe um exemplo dos resultados antes e depois da correção descrita. As diferenças entre as proporções destas distribuições ocorrem devido aos mesmos dois motivos apresentados no parágrafo anterior (arredondamento e cálculo da última categoria).

⁴¹ Esta variável recebe valor 1 caso a variável HHSIZE1 esteja vazia após a sua transformação para valor numérico (que, em caso de erro, deixa a célula vazia).

Figura 34 – Exemplos de variáveis ajustadas: pré e pós correção

A função criada para reduzir os valores das variáveis de setores censitários mistos foi chamada de “redmix” no código desenvolvido e atua apenas nos registros de domicílios cuja variável MIXFLAG é igual a 1. Esta variável recebe valor 1 quando em uma dada zona, tanto a variável POPDC (população em domicílios coletivos), como a variável POPBASE (população em domicílios particulares), são diferentes de zero.

Destaca-se que se o setor censitário for composto exclusivamente por domicílios coletivos (isto é, HHBASE = 0) ele é excluído automaticamente pelo PopulationSim – sendo assim não há a necessidade de excluir estes setores da base original. A Figura 35 indica um exemplo de setor censitário exclusivamente formado por domicílio(s) coletivo(s) – no caso, referente a um centro de detenção no município de Guarulhos.

Figura 35 – Setor censitário exclusivamente formado por domicílio coletivo (em vermelho)

Fonte: Metrô (2020). Adaptado pelo autor.

População de domicílios particulares incorreta: relação POPBASE/HHBASE

Após uma primeira rodada de validação de resultados, observou-se que alguns setores censitários apresentavam erros (em especial o valor da estatística qui-quadrado) muito maiores

que os demais. Ao se examinar estes setores, observou-se que a razão POPBASE/HHBASE (razão entre a população residente em domicílios particulares e o total de domicílios particulares – ou seja, o número médio de indivíduos por domicílio particular) era demasiadamente alta em alguns destes, gerando erros que prejudicavam significativamente a qualidade de ajuste de algumas variáveis.

Entendeu-se que estes setores censitários apresentavam erros em seus cadastros e, por este motivo, foi incluída uma função no código para excluir os setores censitários que apresentavam média de mais de 30 residentes por domicílio. Na RMSP, a aplicação deste critério teve como consequência a exclusão de seis setores censitários.

Observou-se que alguns desses setores poderiam conter uns poucos domicílios particulares, cuja população residente não era possível distinguir. Dessa forma, esta alteração tem como impacto uma pequena redução da população sintética gerada (estimado em dezenas de pessoas na RMSP) em relação ao total efetivo de moradores de domicílios particulares destas áreas.

6.5.3 Matriz semente (base: Censo 2010)

6.5.3.1 Sobre

O código “bMatrizSemente” realiza a criação de dois arquivos de entrada do PopulationSim: “seed_household.csv” e “seed_persons.csv”, ambos gerados a partir da base de microdados do Censo. Por se tratar de uma base com muitas informações (atributos e registros), o Censo a disponibiliza em formato compactado e apropriado para o uso em *softwares* de análise estatística – e não pode ser facilmente lido através de editores de planilha ou bloco de notas.

O código inicia com a leitura da planilha de documentação dos microdados, de onde são criados “dicionários” contendo, para cada variável: sua localização no arquivo (dígito de início e de fim), seu código e seus números de dígitos inteiros e decimais. Na sequência, as bases de dados principais são lidas e conectadas com os dicionários criados. Então, as variáveis são renomeadas e transformadas. Algumas outras são criadas a partir de condições (por exemplo, presença de criança no domicílio e idade do responsável pelo domicílio). Esse procedimento é feito separadamente para a base de domicílio e para a base de indivíduos.

A base de dados de domicílios é gerada apenas para domicílios particulares, a partir da aplicação de um filtro que elimina da base os domicílios de outras espécies. A base de indivíduos, apesar

de manter todos os registros, é utilizada apenas para os moradores de domicílios particulares (isto é, aqueles filtrados na base de domicílios) no PopulationSim.

6.5.3.2 Modificações com base em premissas

Algumas variáveis apresentam a possibilidade de não terem informações para determinados domicílios ou indivíduos. De forma a não reduzir o tamanho da amostra e assim possivelmente subestimar alguns tipos de domicílio ou indivíduos por causa de uma única variável, foram assumidos valores considerados prováveis para estas variáveis.

Isso foi feito para duas variáveis: (i) para grau de instrução (EDUC), os registros com valor “5” (“não determinado”) foram considerados como “1” (“ensino fundamental incompleto”); e (ii) para posse de automóvel, os domicílios cujo valor estava em branco foram considerados com o valor 0 (“não” para a “existência de automóvel para uso particular”).

Para as aplicações realizadas, verificou-se que isso ocorre em uma pequena parcela de registros. As variáveis modificadas, de códigos V6400 e V0222 na base de microdados do Censo 2010, são destacadas na Figura 36.

V6400	<p>NÍVEL DE INSTRUÇÃO:</p> <p>1- Sem instrução e fundamental incompleto</p> <p>2- Fundamental completo e médio incompleto</p> <p>3- Médio completo e superior incompleto</p> <p>4- Superior completo</p> <p>5- Não determinado</p>
V0222	<p>AUTOMÓVEL PARA USO PARTICULAR, EXISTÊNCIA:</p> <p>1- Sim</p> <p>2- Não</p> <p>Branco</p>

Figura 36 – Recortes da documentação dos microdados para variáveis modificadas com base em premissas

Fonte: IBGE (2010). Adaptado pelo autor.

6.5.4 **Totais de controle meta 1 (base: ODSP 2017)**

6.5.4.1 Sobre

O conjunto de informações coletadas pela Pesquisa OD de São Paulo inclui o número de automóveis disponíveis por domicílio em cada zona OD. Este dado, já expandido, é

disponibilizado em sua tabela 7, "Famílias por Número de Automóveis Particulares e Zona de Residência". A pesquisa também apresenta informações sobre o grau de instrução dos indivíduos, na tabela 3, "População por Grau de Instrução e Zona de Residência – 2017". Estas tabelas fazem parte do arquivo "Dados Gerais OD2017.xlsx".

A posse ou não de automóvel, assim como o grau de instrução, são informações também coletadas no questionário da amostra do Censo e apresentadas nos microdados. Entretanto, por se tratar de informações coletadas apenas para a amostra, o Censo não possui totais de controle para elas na base de "agregados por setor censitário".

De forma a incorporar no modelo um dado mais recente e utilizar a funcionalidade de "meta controles" do PopulationSim, os totais expandidos da pesquisa OD de 2017 foram utilizados como totais de controle para os atributos de posse de automóvel do domicílio e grau de instrução (do indivíduo).

Conforme indicado no item 6.3.2, os dados da pesquisa ODSP, disponibilizados para as zonas OD (criadas para esta pesquisa), são agregados para distritos (no município de São Paulo) e municípios (nos demais municípios da RMSP).

O código elaborado agrega as categorias de posse de automóvel e grau de instrução originais da ODSP para que correspondam às categorias dos respectivos atributos na base de microdados do Censo:

- Número de automóveis (com 4 categorias na pesquisa OD) foi transformado para posse ou não de automóvel (as 2 categorias presentes nos microdados). Para registros em branco nos microdados, considerou-se que o domicílio não possuía automóvel;
- Grau de instrução (com 5 categorias na pesquisa OD) teve as categorias "Não Alfabetizado/Fundamental I Incompleto" e "Fundamental I Completo/ Fundamental II Incompleto" agregadas para corresponder à categoria "Sem instrução e fundamental incompleto" dos microdados (com 4 categorias para este atributo).

6.5.4.2 Premissas

A variável de número de automóveis, na tabela da ODSP, é apresentada por família – e não por domicílio, como é feito no Censo. Admitiu-se que a diferença entre a posse de automóveis em famílias e em domicílios não prejudicaria a qualidade dos resultados significativamente.

Por fim, visto que o ano da ODSP (2017) não é o mesmo de realização do Censo (2010), os números extraídos são multiplicados por um fator de ajuste, obtido a partir das variáveis HHBASE (para domicílios) e POPBASE (para indivíduos), que iguala a soma das contagens por categoria obtidas da OD ao total de domicílios particulares (ou indivíduos, a depender do atributo) presentes no Censo.

6.5.4.3 Posse de automóvel por município/distrito

De forma a ilustrar a criação da variável de posse de automóvel, as figuras a seguir (Figura 37 e Figura 38) indicam a parcela de domicílios que possuem um automóvel segundo informações da OD 2017.

Figura 37 – Mapa indicando a parcela dos domicílios com posse de auto na RMSP (parte 1/2: nos 38 municípios da RMSP exclusive SP)

Fonte: Metrô (2019) e IBGE (2020). Elaboração própria.

Figura 38 – Mapa indicando a parcela dos domicílios com posse de carro na RMSP (parte 2/2: nos 96 distritos do município de São Paulo)
 Fonte: Metrô (2019) e IBGE (2020). Elaboração própria.

6.6 CÓDIGOS PARA VALIDAÇÃO

6.6.1 Código original modificado

Conforme indicado no capítulo anterior, que trata do PopulationSim, no código original (“validation.ipynb”) são calculadas métricas de validação (média do erro relativo, desvio padrão do erro relativo e RMSE), além de gráficos.

Este código de validação *vanilla*⁴² do PopulationSim foi modificado para (i) traduzir e adaptar os gráficos para escalas consideradas mais informativas, (ii) corrigir algumas linhas do código e (iii) melhorar os gráficos de saída em função dos dados obtidos⁴³. Este código, assim como a relação das correções realizadas, é indicado na seção de apêndices do presente documento.

6.6.2 Código adicional

Além do código modificado, um novo código foi criado (“zVerificaçãoComMeta.py”) de forma a (i) facilitar a leitura dos resultados gerados em softwares de planilhas eletrônicas e geoprocessamento (como Microsoft Excel e QGIS, respectivamente) e (ii) incorporar o cálculo de uma métrica de validação adicional, a estatística de teste qui-quadrado.

O código, “zVerificaçãoComMeta.py”, lê os dados dos arquivos de sumário de saída do PopulationSim “final_summary_TAZ.csv” e “final_summary_TAZ_PUMA.csv” (a segunda apenas se houver controle na escala meta). A partir deles, calcula as quatro métricas escolhidas para validação e organiza os resultados parciais em múltiplas abas de uma planilha “.xlsx” de saída. O código também gera arquivos “.csv” com estes resultados para facilitar a avaliação da qualidade dos resultados por zona (setor censitário ou área de ponderação) através do uso de softwares de geoprocessamento.

Células (contagem de domicílios/indivíduos para uma dada variável em uma determinada zona) com valor menor ou igual a 5 são desconsideradas da estatística de teste qui-quadrado, devido às distorções geradas por esta métrica para valores pequenos (conforme item 2.5.3). Por estes casos serem raros (ocorrerem em menos de 0,1% das células), entende-se que isso não gera prejuízos para a análise.

⁴² Termo utilizado para se referir a softwares que não são modificados em relação à sua forma original

⁴³ O código original, na data de acesso, (i) lia arquivos com nomes distintos daqueles gerados pelo PopulationSim, que foram modificados e (ii) apresentava erro no cálculo da métrica RMSE (“sum()” adicional, que foi simplesmente eliminado). Adicionalmente, (iii) os parâmetros de geração dos gráficos foram alterados.

7. APLICAÇÃO NA RMSP

O presente capítulo apresenta alguns aspectos referentes à aplicação do PopulationSim na Região Metropolitana de São Paulo (RMSP). São apresentados números-chave da RMSP, dados de entrada principais e o tempo de processamento obtido – além de considerações sobre a população sintética quanto ao ano de referência e à abrangência. Os resultados da aplicação, assim como a validação, são apresentados apenas no capítulo seguinte.

7.1 VISÃO GERAL DA RMSP

A Região Metropolitana de São Paulo (RMSP), instituída em lei federal de 1973, é atualmente composta por 39 municípios. Em 2010, segundo dados do Censo, apresentava uma população total de 19.683.975 habitantes – dos quais 11.253.503 (57,2%) habitavam o município de São Paulo. Segundo estimativa do IBGE, a população da RMSP em 2020 ultrapassa a marca de 21 milhões de habitantes. A Figura 39 indica os limites do município de São Paulo e da RMSP sobre uma imagem de satélite da região.

Figura 39 – Limites do município de São Paulo e de sua região metropolitana
Fonte: IBGE (2010) e Mapbox. Elaboração própria.

O município de São Paulo é dividido em 96 distritos, 310 áreas de ponderação e 18.953 setores censitários. Ao se analisar a composição das áreas de ponderação quanto ao número de setores censitários, nota-se que a maioria delas (69%) no município possui entre 41 e 70 setores censitários, como pode ser observado na Figura 40.

Figura 40 – Histograma do número de setores censitários por área de ponderação no município de São Paulo

Fonte: IBGE (2010) e Mapbox. Elaboração própria.

Os demais municípios da RMSP são divididos em 11.580 setores censitários, totalizando cerca de 30 mil setores censitários na região metropolitana. Já na escala meta, há um total de 134 zonas (96 distritos de São Paulo mais 38 municípios), indicados na Figura 41.

Figura 41 – 134 áreas da escala meta utilizadas

Fonte: Metrô (2020)

7.2 DADOS DE ENTRADA

7.2.1 Zonas e população

São indicados na Tabela 22 os números de setores censitários, áreas de ponderação e distritos/municípios (escala meta) na RMSP.

Tabela 22 – Número de setores censitários⁴⁴, áreas de ponderação e distritos/municípios (zonas da escala meta) na RMSP

Município/ Região	Setores censitários	Áreas de ponderação	Distritos/ Municípios
São Paulo	18.333	310	96 ¹
RMSP exc. SP	11.538	323	38 ²
<i>Total</i>	<i>29.871</i>	<i>633</i>	<i>134</i>

¹ Distritos do município de São Paulo | ² Municípios

A amostra disponibilizada para o município de São Paulo é de aproximadamente 5% da população, conforme previsto na Tabela 3 para municípios com mais de 500 mil habitantes. Para os demais municípios da RMSP estão disponíveis amostras de 20% (para aqueles com até 20 mil habitantes), 10% (população entre 20 mil e 500 mil habitantes) ou também 5% (caso dos municípios de Osasco, Santo André, São Bernardo do Campo, Guarulhos – além de São Paulo).

A Tabela 23 indica os totais (i) no município de São Paulo e (ii) nos demais municípios da RMSP, exclusive São Paulo. É importante observar que os números, tanto os totais como a amostra, correspondem a domicílios particulares – excluindo domicílios coletivos e aqueles de setores censitários com totais inconsistentes (conforme item 6.5.2.2). Isso porque a presente aplicação se restringe a essa população, conforme justificado no item 7.4.2. Para o caso do município de São Paulo, a população “descontada” do total – isto é, que não reside em domicílios particulares – é de 41.238 pessoas, aproximadamente 0,4% do total de moradores do município.

⁴⁴ Que foram sintetizados (ou seja, aqueles que continham ao menos um domicílio particular).

Tabela 23 – Número de registros das amostras de domicílios e pessoas na RMSP em relação ao total (apenas domicílios particulares e seus respectivos moradores)

Município/ Região	Base	Total	Amostra	Parcela do total
São Paulo	Domicílios	3.574.286	174.162	4,9%
	Pessoas	11.212.265	552.037	4,9%
RMSP exc. SP	Domicílios	2.515.561	197.658	7,9%
	Pessoas	8.396.860	664.574	7,9%
Total (RMSP)	Domicílios	6.089.847	371.820	6,1%
	Pessoas	19.609.125	1.216.611	6,2%

Fonte: IBGE (2010). Processado pelo autor.

Observa-se na Tabela 23 que, por ser formada por diversos municípios de população menor que 500 mil habitantes (cuja amostra é de 10% ou mais, conforme Tabela 3), a amostra nos 38 municípios da RMSP é, na média, de quase 8%.

7.2.2 Configurações (settings.yaml)

As configurações de execução do PopulationSim (especificadas no arquivo “settings.yaml”) foram mantidas como o padrão do original, com exceção do máximo fator de expansão, reduzido para o valor de 3⁴⁵ de forma a buscar uma heterogeneidade maior na população sintética gerada do que aquela que seria obtida para o valor padrão do PopulationSim (30). Um recorte do arquivo “settings.yaml” com as configurações utilizadas é indicado na Figura 42.

```
# Algorithm/Software Configuration
# -----
INTEGERIZE_WITH_BACKSTOPPED_CONTROLS: True
SUB_BALANCE_WITH_FLOAT_SEED_WEIGHTS: False
GROUP_BY_INCIDENCE_SIGNATURE: True
USE_SIMUL_INTEGERIZER: True
USE_CVXPY: False
max_expansion_factor: 3
```

Figura 42 – Configurações de execução do PopulationSim utilizadas

7.2.3 Variáveis e importâncias (controls.csv)

O arquivo “controls.csv”, em que as variáveis da matriz semente e dos controles são conectadas (através do uso de expressões para a realização das contagens), foi adaptado para a aplicação de acordo com as categorias dos atributos definidas no Apêndice C. Além disso, os valores das importâncias foram definidos buscando (i) o acerto do número total de domicílios, através da

⁴⁵ Um fator de expansão de 10 foi utilizado nos municípios de Santo André e Poá, na aplicação na RMSP. Isto porque estes municípios apresentaram qualidade de ajuste reduzida para o processamento utilizando fator de expansão 3 – o que foi solucionado ao permitir um fator de expansão maior, conforme discutido no item 8.4.2.

inserção de um valor muito mais alto para a importância da variável HHBASE, (ii) maior importância para os atributos de tamanho de domicílio (HHSIZE), seguindo o valor padrão do PopulationSim; e de posse de automóvel (HHAUT), por se tratar de um atributo com alteração significativa entre os microdados expandidos do Censo 2010 e os totais da ODSP 2017. O arquivo “controls.csv” completo, contendo estas informações, é apresentado na Figura 43.

```
target,geography,seed_table,importance,control_field,expression
num_hh,TAZ,households,1000000000,HHBASE,households.WGTP >= 1
hh_size_1,TAZ,households,5000,HHSIZE1,households.NP == 1
hh_size_2,TAZ,households,5000,HHSIZE2,households.NP == 2
hh_size_3,TAZ,households,5000,HHSIZE3,households.NP == 3
hh_size_4,TAZ,households,5000,HHSIZE4,households.NP == 4
hh_size_5,TAZ,households,5000,HHSIZE5,households.NP == 5
hh_size_6,TAZ,households,5000,HHSIZE6,households.NP >= 6
hh_age_1,TAZ,households,500,HHAGE1,(households.AGEHOH>9) & (households.AGEHOH<=29)
hh_age_2,TAZ,households,500,HHAGE2,(households.AGEHOH>29) & (households.AGEHOH<=44)
hh_age_3,TAZ,households,500,HHAGE3,(households.AGEHOH>44) & (households.AGEHOH<=59)
hh_age_4,TAZ,households,500,HHAGE4,(households.AGEHOH>59) & (households.AGEHOH<=999)
hh_inc_1,TAZ,households,500,HHINC1,(households.HHINC>-99) & (households.HHINC<= 0.5)
hh_inc_2,TAZ,households,500,HHINC2,(households.HHINC>0.5) & (households.HHINC<=1)
hh_inc_3,TAZ,households,500,HHINC3,(households.HHINC>1) & (households.HHINC<=2)
hh_inc_4,TAZ,households,500,HHINC4,(households.HHINC>2) & (households.HHINC<=3)
hh_inc_5,TAZ,households,500,HHINC5,(households.HHINC>3) & (households.HHINC<=5)
hh_inc_6,TAZ,households,500,HHINC6,(households.HHINC>5) & (households.HHINC<=9999999)
p_sex_masc,TAZ,persons,5000,SEX1,persons.SEX == 1
p_sex_femi,TAZ,persons,5000,SEX2,persons.SEX == 2
p_age_01,TAZ,persons,500,AGEP1,persons.AGEP <= 4
p_age_02,TAZ,persons,500,AGEP2,(persons.AGEP > 5) & (persons.AGEP <= 14)
p_age_03,TAZ,persons,500,AGEP3,(persons.AGEP >= 15) & (persons.AGEP <= 19)
p_age_04,TAZ,persons,500,AGEP4,(persons.AGEP >= 20) & (persons.AGEP <= 29)
p_age_05,TAZ,persons,500,AGEP5,(persons.AGEP >= 30) & (persons.AGEP <= 39)
p_age_06,TAZ,persons,500,AGEP6,(persons.AGEP >= 40) & (persons.AGEP <= 49)
p_age_07,TAZ,persons,500,AGEP7,(persons.AGEP >= 50) & (persons.AGEP <= 59)
p_age_08,TAZ,persons,500,AGEP8,(persons.AGEP >= 60) & (persons.AGEP <= 999999999)
hh_aut_0,REGION,households,5000,HHAUT0_,households.HHAUT == 0
hh_aut_1,REGION,households,5000,HHAUT1_,households.HHAUT == 1
p_educ_1,REGION,persons,500,EDUC1_,persons.EDUC == 1
p_educ_2,REGION,persons,500,EDUC2_,persons.EDUC == 2
p_educ_3,REGION,persons,500,EDUC3_,persons.EDUC == 3
p_educ_4,REGION,persons,500,EDUC4_,persons.EDUC == 4
```

Figura 43 – Arquivo “configs.csv” para a aplicação na RMSP

7.2.4 Totais de controle

Os valores obtidos a partir da base de “agregados por setor censitário”, somados para toda a região metropolitana para cada categoria de cada atributo controlado nesta escala (total de 27), são indicados na Tabela 24 – assim como a participação de cada categoria.

Tabela 24 – Totais de controle da RMSP na escala de setor censitário

Atributo	Variável	Total de controle	Parcela
Num. dom.	NUM_HH	6.089.847	100%
Tamanho do domicílio	HHSIZE1	766.227	13%
	HHSIZE2	1.387.336	23%
	HHSIZE3	1.558.274	26%
	HHSIZE4	1.318.714	22%
	HHSIZE5	618.818	10%
	HHSIZE6	440.478	7%
Idade do responsável	HHAGE1	885.428	15%
	HHAGE2	2.137.447	35%
	HHAGE3	1.824.111	30%
	HHAGE4	1.242.895	20%
Renda per capita em salários mínimos	HHINC1	1.247.912	20%
	HHINC2	1.541.695	25%
	HHINC3	1.619.191	27%
	HHINC4	593.607	10%
	HHINC5	493.666	8%
	HHINC6	593.780	10%
Sexo do indivíduo	SEX1	9.385.162	48%
	SEX2	10.223.963	52%
Idade do indivíduo	AGEP1	1.302.487	7%
	AGEP2	3.002.098	15%
	AGEP3	1.547.984	8%
	AGEP4	3.606.806	18%
	AGEP5	3.297.504	17%
	AGEP6	2.711.746	14%
	AGEP7	2.051.423	10%
	AGEP8	2.106.742	11%

Fonte: IBGE (2010). Processado pelo autor.

São utilizados, além dos totais de controle na escala de setor censitário apresentados na Tabela 24: as duas bases de microdados (pessoas e domicílios) com o número de registros indicados na Tabela 23; o arquivo de correspondência geográfica e os totais de controle da pesquisa OD – esta última apresentada na Tabela 25 para as 6 categorias das variáveis controladas.

Tabela 25 – Totais de controle da RMSP na escala meta

Atributo	Variável	Total de controle	Parcela
Posse ou não de automóvel	HHAUT0	2.855.498	47%
	HHAUT1	3.234.215	53%
Grau de instrução	EDUC1	7.056.532	36%
	EDUC2	2.851.342	15%
	EDUC3	6.805.774	35%
	EDUC4	2.895.207	15%

Fontes: ODSP (2017) e IBGE (2010). Processado pelo autor.

Conforme indicado na descrição do código para extração e tratamento dos dados da ODSP (item 0), os valores dos atributos de posse de automóvel (prefixo “HHAUT”) e grau de instrução (prefixo “EDUC”) são reduzidos para se ajustarem à população total, dada pelos totais de controle do Censo 2010.

7.2.5 Microdados expandidos x controles meta

Da mesma forma que Vovsha *et al.* (2015), a presente aplicação utilizou totais de controle no nível meta que diferem substancialmente dos microdados expandidos, como pode se observar na Tabela 26 e na Figura 44.

Tabela 26 – Comparação entre microdados expandidos e totais de controle meta para variáveis controladas nesta escala

Atributo	Variável	Microdados expandidos (Censo 2010)	Parcela do atributo	Total de controle meta (OD 2017)	Parcela do atributo
Posse ou não de automóvel	HHAUT0	2,957,923	48%	2,855,498	47%
	HHAUT1	3,201,784	52%	3,234,215	53%
Grau de instrução	EDUC1	9,588,494	49%	7,056,532	36%
	EDUC2	3,206,080	16%	2,851,342	15%
	EDUC3	4,651,596	24%	6,805,774	35%
	EDUC4	2,237,804	11%	2,895,207	15%

Figura 44 – Comparação da parcela de cada categoria para cada atributo em relação ao total, para microdados expandidos e totais de controle meta

Observa-se uma leve tendência de aumento na posse de automóvel dos domicílios. Quanto ao grau de instrução, nota-se um aumento significativo da parcela da população nas categorias 3 e 4 (ensino médio completo e superior completo, respectivamente) e uma redução expressiva da categoria 1 (sem instrução ou fundamental incompleto).

7.3 PROCESSAMENTO

A população sintética para a região metropolitana foi gerada através de duas aplicações: uma para o município de São Paulo e outra para os demais municípios da região metropolitana. Optou-se por esta aplicação dividida em duas partes pelos seguintes motivos: (i) os “agregados por setores censitários” da capital (município de São Paulo) são fornecidos em tabelas separadas do restante do estado no Censo 2010, (ii) para os controles meta (oriundos da ODSP), as zonas utilizadas são distritos no município de São Paulo e os próprios municípios no restante da RMSP e (iii) o tempo de processamento obtido para cada parte foi de 6 a 10 horas, sendo considerado adequado para o processamento entre um dia e outro. Como consequência deste processamento em duas partes, as métricas de validação (apresentadas no capítulo 8) foram agregadas manualmente para o resultado conjunto da RMSP.

Um computador com processador Intel Core i5-3470 CPU @ 3,20 GHz e 12 GB de memória RAM foi utilizado para a geração da população sintética. O tempo total de processamento dos códigos de tratamento de dados de entrada foi de cerca de 30 minutos (dos quais a grande maioria foi para o tratamento dos dados da matriz semente), enquanto o tempo de execução do PopulationSim foi de aproximadamente 16 horas (das quais 10 horas se referem à geração da

população do município de São Paulo e as outras 6 horas para os demais 38 municípios que compõem a Região Metropolitana). Os códigos de validação foram processados em alguns segundos.

7.4 DEMAIS CONSIDERAÇÕES SOBRE A APLICAÇÃO

7.4.1 Data-base e projeções

O Censo mais recente disponível durante a elaboração do presente trabalho apresentava dados coletados em 2010. Já a pesquisa OD utilizada para controlar alguns dos atributos foi realizada em 2017 (e completada em 2018). O uso de totais de controle com informações de 2017 na escala meta tende a gerar uma população com características mais próximas daquela observada naquele ano, especialmente para as 6 variáveis controladas nesta escala. Entretanto, visto que tanto os microdados como os totais de controle no nível mais desagregado (setor censitário, com 27 variáveis controladas) apresentam 2010 como ano-base, pode-se dizer que a população sintetizada na presente aplicação se refere a este ano.

A documentação da aplicação do PopulationSim em South Oregon (SOABM, 2020b) apresenta uma seção específica para a modelagem de populações em anos futuros.

7.4.2 Apenas para domicílios particulares

A unidade a ser replicada e alocada em zonas em setores censitários pelo PopulationSim é o domicílio, assim como é feito pela maioria dos outros procedimentos de geração de população sintética identificados no Capítulo 2.

A documentação do Censo define domicílio como “um local estruturalmente separado e independente que se destina a servir de habitação a uma ou mais pessoas, ou que esteja sendo utilizado como tal”, caracterizando a “separação” supracitada como “quando o local de habitação for limitado por paredes, muros ou cercas e coberto por um teto” – e a independência como “quando o local de habitação tem acesso direto (à rua), permitindo a seus moradores entrar e sair sem necessidade de passar por locais de moradia de outras pessoas”.

O Censo divide os domicílios em duas “espécies”: particulares e coletivos. Segundo a documentação dos “agregados por setor censitário”, domicílios particulares são aqueles onde “o relacionamento entre seus ocupantes era ditado por laços de parentesco, de dependência doméstica ou por normas de convivência”.

Já domicílio coletivo é definido como "uma instituição ou estabelecimento onde a relação entre as pessoas que nele se encontravam, moradoras ou não, era restrita a normas de subordinação administrativa, como em hotéis, motéis, 'camping', pensões, penitenciárias, presídios, casas de detenção, quartéis, postos militares, asilos, orfanatos, conventos, hospitais e clínicas (com internação), alojamento de trabalhadores ou de estudantes etc.".

A presente aplicação gera apenas domicílios particulares e seus respectivos moradores. Via de regra, os moradores de domicílios coletivos representam uma pequena parcela da população⁴⁶ e tendem a realizar menos viagens intraurbanas do que moradores de domicílios particulares. Por estes motivos, entende-se que a não consideração destes na modelagem não impactaria significativamente a aplicação final em um modelo de demanda por transportes.

A aplicação do PopulationSim em South Oregon (SOABM, 2020a) também discute este assunto, gerando a população sintética para domicílios coletivos (*group quarters*) separadamente e com um número reduzido de atributos controlados em relação à aplicação para domicílios particulares.

⁴⁶ Para a maioria dos municípios do Brasil. Entretanto, para municípios pequenos e com muitos moradores de domicílios coletivos, a diferença pode ser significativa. No caso dos municípios de Guarulhos e São Paulo, menos de 1% do total da população residia em domicílios desta espécie.

8. VALIDAÇÃO

No presente capítulo a população sintética gerada para a RMSP através da aplicação do PopulationSim (e após a execução dos códigos de tratamento desenvolvidos neste trabalho) é examinada. A validação dos resultados gerados foi feita com base nos procedimentos e métricas indicados no item 0.

8.1 COMPARAÇÃO DE TOTAIS AGREGADOS

8.1.1 Totais de controle x sintetizado

Uma primeira avaliação feita neste capítulo é a comparação dos totais por variável (categorias dos atributos utilizados) para toda a RMSP através do cálculo da diferença relativa (diferença entre sintetizado e controle sobre o controle), apresentada na Tabela 27.

Tabela 27 – Comparação entre totais de controle e sintetizado por variável para a RMSP

Atributo	Variável	Controle	Sintetizado	Dif. Relativa
Num. dom.	NUM_HH	6.089.847	6.089.839	0,00%
Tamanho do domicílio	HHSIZE1	766.227	764.495	-0,23%
	HHSIZE2	1.387.336	1.384.752	-0,19%
	HHSIZE3	1.558.274	1.558.439	0,01%
	HHSIZE4	1.318.714	1.319.442	0,06%
	HHSIZE5	618.818	620.040	0,20%
	HHSIZE6	440.478	442.671	0,50%
Idade do responsável	HHAGE1	885.428	885.337	-0,01%
	HHAGE2	2.137.447	2.137.331	-0,01%
	HHAGE3	1.824.111	1.824.027	0,00%
	HHAGE4	1.242.895	1.242.919	0,00%
Renda per capita em salários mínimos	HHINC1	1.247.912	1.247.143	-0,06%
	HHINC2	1.541.695	1.541.237	-0,03%
	HHINC3	1.619.191	1.619.758	0,04%
	HHINC4	593.607	594.128	0,09%
	HHINC5	493.666	493.881	0,04%
	HHINC6	593.780	593.692	-0,01%
Sexo do indivíduo	SEXM	9.385.162	9.363.853	-0,23%
	SEXF	10.223.963	10.206.216	-0,17%
Idade do indivíduo	AGEP1	1.302.487	1.300.757	-0,13%
	AGEP2	3.002.098	2.993.615	-0,28%
	AGEP3	1.547.984	1.542.119	-0,38%
	AGEP4	3.606.806	3.591.693	-0,42%
	AGEP5	3.297.504	3.286.757	-0,33%
	AGEP6	2.711.746	2.705.165	-0,24%
	AGEP7	2.051.423	2.046.898	-0,22%
	AGEP8	2.106.742	2.103.065	-0,17%
Posse ou não de automóvel	HHAUT0	2.855.496	2.861.326	0,20%
	HHAUT1	3.234.223	3.228.513	-0,18%
Grau de instrução	EDUC1	7.056.528	7.109.025	0,74%
	EDUC2	2.851.352	2.859.378	0,28%
	EDUC3	6.805.785	6.764.786	-0,60%
	EDUC4	2.895.216	2.836.880	-2,01%
População total (soma indivíduos)		19.608.881	19.570.069	-0,20%

Quanto à população total gerada, observa-se um ajuste quase perfeito do número de domicílios (NUM_HH), com uma diferença de apenas 8 unidades dentre mais de 6 milhões de domicílios sintetizados. Já para indivíduos, o sintetizador subestimou a população total em aproximadamente 38 mil pessoas, o que corresponde a 0,2% do total (calculado a partir da soma de indivíduos para qualquer um dos atributos de indivíduo e indicado na última linha da Tabela 27).

Observa-se que as diferenças relativas não superaram 0,5% para as variáveis controladas na escala de setores censitários. Para as variáveis controladas no nível meta (posse de automóvel e grau de instrução), as diferenças relativas também foram pequenas neste nível de agregação: não superaram 1%, exceto na variável EDUC4 (grau de instrução: superior completo), que foi subestimada em 2% (devido à sintetização de cerca de 60 mil indivíduos a menos com estas características do que o valor de referência controlado).

Para as algumas das variáveis controladas no nível meta, conforme destacado no item 7.2.5, as diferenças entre microdados expandidos e controle foi significativa. Caso os totais de controle meta não fossem utilizados, poderia se esperar a sintetização de 4% a menos de indivíduos na categoria EDUC4 (11% ao invés de 15% dos indivíduos, conforme Tabela 26, caso variável apresentasse alto grau de correlação com as demais controladas).

O gerador de populações sintéticas, ao tentar atender controles não utilizados para o cálculo do peso amostral, pode gerar fatores de expansão “não-uniformes” – isto é, replicando demasiadamente os domicílios que apresentam uma determinada característica que se distancia da expansão direta dos microdados naquela área de ponderação.

Ao limitar o fator de expansão em 3 (exceto nos municípios de Santo André e Poá, limitados em 10), estes totais podem não ser alcançados, respeitando a limitação de não replicar os domicílios 3 vezes mais do que o sugerido pelo peso amostral.

8.1.2 Variáveis combinadas

É feita também a verificação do ajuste entre combinações de variáveis e totais não utilizados de variáveis combinadas (como, por exemplo, o cruzamento entre idade e sexo, disponível na base de agregados por setor censitário), conforme sugerido por Voas e Williamson (2000).

Os resultados da comparação entre (i) os totais não utilizados da base de agregados por setor censitário do Censo e (ii) a contagem de indivíduos da população sintética para duas condições

(idade e sexo) são indicados na Tabela 28. Foram avaliadas quatro variáveis combinadas, dadas pelo cruzamento de duas variáveis de idade (AGEP4: 20 a 29 anos de idade e AGEP7: 50 a 59 anos de idade) com as duas categorias do atributo de sexo.

Tabela 28 – Cruzamento de variáveis da base gerada (sintetizado) comparados com agregados por setores censitários

Categoria de idade	Categoria de Sexo	Parcela da categoria de idade		Diferença
		Totais Agregados SC	Sintetizado	
AGEP4	SEXM	48,3%	49,2%	0,9%
	SEXF	51,7%	50,8%	-0,9%
AGEP7	SEXM	45,9%	45,8%	-0,1%
	SEXF	54,1%	54,2%	0,1%

Observa-se na Tabela 28 que, para as combinações de variáveis analisadas, a contribuição de cada variável combinada não superou 1% de diferença entre o total não utilizado como controle e o total sintetizado. Entretanto, é importante destacar que pelo fato de a variável sexo apresentar uma divisão quase igualitária entre as suas categorias (masculino e feminino), diferenças pequenas já poderiam ser esperadas.

8.2 FATORES DE EXPANSÃO

De forma a examinar a uniformidade dos fatores de expansão obtidos do procedimento (razão entre os pesos finais inteiros e os pesos amostrais), são indicados a seguir (Tabela 29 e Figura 45) os fatores de expansão acumulados para a RMSP.

Destaca-se que o parâmetro do fator de expansão máximo foi definido como igual a 3 (exceto em Santo André e Poá). Entretanto, devido ao procedimento de cálculo utilizado pelo PopulationSim, este limite teórico é ultrapassado em uma parcela dos domicílios da amostra (aproximadamente de 6%).

Tabela 29 – Contagem de domicílios por faixa de fator de expansão na RMSP

Fator de expansão	Contagem	Parcela
= 0	92.296	27,2%
>0 a 0,5	66.096	19,5%
0,5 a 1	41.642	12,3%
1 a 1,5	36.769	10,8%
1,5 a 2	33.205	9,8%
2 a 2,5	27.095	8,0%
2,5 a 3,1	23.739	7,0%
3,1 a 5	15.627	4,6%
5 a 10	2.830	0,8%
>10	436	0,1%
<i>Total</i>	<i>339.735</i>	<i>100%</i>

Figura 45 – Histograma dos fatores de expansão obtidos

Observa-se que mais de 25% dos domicílios da amostra apresentaram fator de expansão igual a 0 – isto é, não foram utilizados para a composição da população sintética de nenhum setor censitário. Cerca de 23% dos fatores de expansão obtidos estão entre os valores de 0,5 e 1,5 – o que indica que os pesos amostrais destes domicílios foram relativamente pouco alterados frente aos demais (em 50% para mais ou para menos).

Por fim, é importante mencionar que algumas centenas de domicílios apresentaram fatores de expansão maiores que 10, algo não desejável devido à possível baixa heterogeneidade da população gerada. Entretanto, os 436 domicílios cujo fator de expansão obtido foi maior que 10 representam apenas 1,4% dos domicílios da população sintética gerada. Ou seja, apesar da expansão elevada em relação ao peso amostral, estes domicílios ainda representam uma pequena parcela do total de domicílios sintetizados.

Ressalta-se ainda que, por ter sido utilizada a funcionalidade de agrupar domicílios em função das categorias de controle utilizadas, os números apresentados se referem a grupos de domicílios (para os quais os pesos amostrais são somados) e não diretamente aos domicílios da amostra. Observou-se, na aplicação realizada, que os 367.308 domicílios disponíveis na amostra foram transformados em 339.735 grupos de domicílios e utilizados para a sintetização – equivalente a uma redução de 7,5%. Este procedimento é realizado para reduzir a quantidade de domicílios processados e consequentemente o tempo de processamento.

8.3 AVALIAÇÃO POR VARIÁVEL (CATEGORIA DOS ATRIBUTOS)

No presente item é examinada a qualidade de ajuste entre totais de controle e sintetizado para cada variável (total de categoria dos atributos) utilizada. As variáveis são avaliadas tanto para agregações no nível de setores censitários como para áreas de ponderação, através das quatro métricas definidas no item 4.5.2: (i) erro médio relativo – EMR, (ii) desvio padrão do EMR, (iii) raiz do erro quadrático médio – RMSE e (iv) valor da estatística do qui-quadrado – reduzido para “Qui-quad” na tabela.

As variáveis relacionadas aos atributos HHCRIAN (presença ou não de criança no domicílio) e OCUP (categoria de ocupação no trabalho) não são apresentadas nos itens a seguir por não serem controladas – e, sendo assim, não apresentam totais de controle para serem utilizados como referência.

Para setores censitários, a tabela também não apresenta valores para as variáveis HHAUT e EDUC, visto que estas são controladas apenas no nível meta e não para setores censitários. Sendo assim, é composta apenas pelas 27 variáveis controladas nesta escala.

A avaliação das variáveis cujo controle é disponibilizado no nível meta (no caso, distritos/municípios) é realizada através de estimações feitas pelo PopulationSim para o nível de área de ponderação – ou seja, trata-se de um dado processado, estimado no processo de fatoração dos controles meta.

8.3.1 Para setores censitários

A Tabela 30 indica os resultados agregados (médias ou somatórias, a depender da métrica) para os 29.873 setores censitários sintetizados⁴⁷ da RMSP.

Tabela 30 – Métricas de validação para setores censitários da RMSP

Atributo	Variável	Para setores censitários (N=29869)			
		EMR	DesvPad	RMSE	Qui-quad
Num. dom.	NUM_HH	0,0%	0,0%	0,0	0,0
Tamanho do domicílio	HHSIZE1	0,0%	4,3%	0,4	170,5
	HHSIZE2	-0,1%	4,0%	0,9	404,7
	HHSIZE3	0,3%	5,5%	0,6	187,9
	HHSIZE4	0,0%	3,2%	0,6	194,8
	HHSIZE5	0,1%	3,8%	0,5	222,6
	HHSIZE6	-0,3%	7,3%	0,6	367,5
Idade do responsável	HHAGE1	0,0%	3,4%	0,4	166,9
	HHAGE2	0,0%	3,2%	0,7	182,8
	HHAGE3	0,0%	3,1%	0,7	198,1
	HHAGE4	-0,1%	3,4%	0,5	158,4
Renda per capita em salários mínimos	HHINC1	0,1%	4,1%	0,6	206,6
	HHINC2	0,1%	4,8%	0,6	177,3
	HHINC3	0,3%	5,9%	0,6	188,4
	HHINC4	0,2%	4,9%	0,4	176,8
	HHINC5	0,2%	4,5%	0,4	143,0
	HHINC6	-0,6%	8,4%	0,6	160,9
Sexo do indivíduo	SEXM	0,0%	4,5%	3,2	740,5
	SEXF	0,0%	4,7%	3,3	765,0
Idade do indivíduo	AGEP1	0,1%	4,8%	1,0	542,3
	AGEP2	0,3%	10,5%	1,8	704,0
	AGEP3	-0,1%	5,8%	1,0	517,6
	AGEP4	0,2%	8,5%	1,8	622,7
	AGEP5	0,1%	7,9%	1,6	525,0
	AGEP6	0,1%	6,6%	1,4	498,9
	AGEP7	-0,1%	5,9%	1,1	439,0
	AGEP8	-0,3%	4,4%	1,1	427,8

Ao se avaliar os resultados da Tabela 30, observa-se:

- EMR:
 - Para o erro médio relativo (EMR), os valores não superaram 0,4% (em módulo),

⁴⁷ Excluídos setores censitários sem domicílios particulares ou com provável erro de cadastro (conforme item 6.5.3.2)

exceto para a variável HHINC6. Visto que esta teve uma diferença de apenas 0,01% entre total de controle e sintetizado, seu valor de EMR de -0,6% sugere uma pequena subestimação de domicílios com esta característica em setores censitários em que ela é mais rara (menor incidência e consequentemente maiores erros relativos);

- DesvPad:
 - Para o desvio padrão do erro médio relativo (DesvPad), observa-se a maior parte das variáveis apresentando valores entre 4 e 8%, com o máximo de 10,5% na variável AGE2;
- RMSE:
 - Os valores da raiz do erro quadrático médio (RMSE) não ultrapassam 2 – exceto para as variáveis de sexo, que são as que tem a maior incidência, para as quais o RMSE alcança o valor de 3. Um valor mais alto para estas era esperado, visto que o RMSE é uma métrica que depende dos valores absolutos.
 - A ordem de grandeza dos valores encontrados para o RMSE indica bons resultados, visto que os totais destas categorias são, na média, acima de 500 (exceto para as variáveis HHINC5 e HHINC6, acima de 300).
 - Observa-se que a variável HHINC6, que apresentou resultados piores para as demais métricas, apresenta um valor médio para o RMSE (em comparação com as demais); isto porque a categoria à qual ela se refere apresenta uma contribuição pequena (cerca de 10% dos domicílios se enquadram nesta categoria);
- Qui-quad:
 - Por fim, para os valores da estatística do qui-quadrado, foram obtidos resultados excelentes – considerando que para 29.873 graus de liberdade e 99,9% de probabilidade, o valor-p de referência é de 29.307 – e que o valor da estatística de nenhuma das variáveis alcança a ordem de grandeza de milhares, não rejeitando a hipótese de que as distribuições dos totais de controle e sintetizado para uma dada variável entre os milhares de setores censitários sejam estatisticamente iguais.

8.3.2 Para áreas de ponderação

A Tabela 31 apresenta os resultados para todas as 33 variáveis controladas, seja na escala de setor censitário ou distrito/município, para as 633 áreas de ponderação da RMSP.

Tabela 31 – Métricas de validação para áreas de ponderação da RMSP

Atributo	Variável	Para áreas de ponderação (N=633)			
		EMR	DesvPad	RMSE	Qui-quad
Num. dom.	NUM_HH	0,0%	0,0%	0,0	0,0
Tamanho do domicílio	HHSIZE1	-0,3%	0,9%	1,4	56,4
	HHSIZE2	-0,2%	1,1%	4,5	255,7
	HHSIZE3	0,0%	0,4%	1,7	30,8
	HHSIZE4	0,1%	0,4%	1,3	22,7
	HHSIZE5	0,2%	0,6%	0,9	23,9
	HHSIZE6	0,3%	1,9%	1,6	97,3
Idade do responsável	HHAGE1	0,0%	0,3%	0,5	4,7
	HHAGE2	0,0%	0,2%	0,7	4,7
	HHAGE3	0,0%	0,2%	0,7	6,0
	HHAGE4	0,0%	0,3%	0,6	5,5
Renda per capita em salários mínimos	HHINC1	-0,1%	0,5%	1,4	30,5
	HHINC2	0,0%	0,3%	1,0	14,5
	HHINC3	0,1%	0,4%	1,0	13,9
	HHINC4	0,1%	0,9%	0,7	26,4
	HHINC5	0,1%	1,4%	0,6	29,6
	HHINC6	-1,0%	7,8%	1,4	125,8
Sexo do indivíduo	SEXM	-0,2%	0,6%	15,7	427,5
	SEXF	-0,2%	0,7%	17,1	471,2
Idade do indivíduo	AGEP1	0,0%	0,9%	2,9	107,5
	AGEP2	-0,2%	0,8%	6,6	220,2
	AGEP3	-0,2%	0,6%	2,8	75,5
	AGEP4	-0,2%	0,8%	7,1	233,1
	AGEP5	-0,2%	0,8%	6,6	214,9
	AGEP6	-0,1%	0,8%	5,2	167,7
	AGEP7	-0,2%	0,7%	3,5	106,6
	AGEP8	-0,2%	0,7%	3,3	92,4
Posse ou não de automóvel	HHAUT0	0,0%	0,1%	0,6	3,0
	HHAUT1	0,0%	0,1%	0,7	2,6
Grau de instrução	EDUC1	0,1%	1,0%	16,4	613,5*
	EDUC2	-0,1%	0,8%	5,5	167,7
	EDUC3	-0,3%	0,8%	13,4	458,9
	EDUC4	-0,5%	1,5%	6,2	333,8

¹ Valores marcados com um asterisco (*) apresentam qualidade de ajuste ruim

Ao se avaliar os resultados da Tabela 31, observa-se:

- EMR:
 - Para o erro médio relativo (EMR), os valores não superaram, em módulo, 1%. Assim como observado na análise por setor censitário, a variável HHINC6 apresenta o maior desvio dentre as 33 variáveis analisadas, por subestimar a incidência em áreas de ponderações menores;
- DesvPad:
 - Para o desvio padrão do erro médio relativo (DesvPad), observa-se a maior parte das variáveis com valores abaixo de 1%. É possível notar que a variável HHINC6 apresenta valor muito acima das demais para esta variável: quase 8%;
- RMSE:
 - Os valores da raiz do erro quadrático médio (RMSE) superaram o valor de 10 para quatro variáveis na escala de área de ponderação:
 - As duas de sexo (SEXM e SEXF), que apresentam contagem alta por se tratar de variáveis de indivíduos e de um atributo com apenas duas categorias, e
 - Duas das quatro variáveis do atributo de grau de instrução (EDUC1 e EDUC3), as quais apresentaram os maiores desvios “teóricos” entre microdados expandidos e totais de controle meta (conforme item 7.2.5);
- Qui-quad:
 - Por fim, para os valores da estatística do qui-quadrado, foi possível descartar a hipótese nula de independência entre os vetores (o que sugere evidência de dependência entre as distribuições em cada zona) de todas as variáveis, para 99% de nível de confiança, exceto para EDUC1. O valor observado (613,5) é maior que o valor de referência para 99% (553,2).

8.4 AVALIAÇÃO POR ZONA (SETOR CENSITÁRIO E ÁREA DE PONDERAÇÃO)

De forma a identificar o desempenho do procedimento da perspectiva espacial (e consequentemente em função de características predominantes de determinadas zonas), é também realizada uma análise por setores censitários e área de ponderação. Esta análise se baseia na agregação dos valores do qui-quadrado calculados para as 27 ou 33 variáveis avaliadas (setores censitários e áreas de ponderação, respectivamente) e suas comparações com os valores-p de referência.

Enquanto a análise para setores censitários é realizada apenas através de tabelas e histogramas, a análise por área de ponderação permitiu a visualização dos resultados através de mapas. Adicionalmente, foram avaliadas as cinco áreas de ponderação com os valores mais altos calculados para o qui-quadrado quanto aos erros relativos obtidos para cada variável.

8.4.1 Para setores censitários

São apresentados, na Tabela 32, a distribuição dos valores do qui-quadrado e dos valores-p obtidos por setor censitário. O valor de referência, para 99,9% de nível de confiança e 27 graus de liberdade, é de 9,8.

Tabela 32 – Valores da estatística de teste qui-quadrado para setores censitários da RMSP

Qui-quad	Contagem	Parcela
0-0,2	4,915	16.5%
0,2-0,4	7,886	26.4%
0,4-0,6	5,474	18.3%
0,6-0,8	3,775	12.6%
0,8-1	2,187	7.3%
1-2	4,756	15.9%
2-5	527	1.8%
5-10	178	0.6%
10-50	159	0.5%
> 50	12	0.04%
<i>Total</i>	<i>29,869</i>	<i>100.0%</i>

Na Figura 46, é apresentado o histograma destes valores-p, considerando 27 graus de liberdade (número de atributos controlados).

Figura 46 – Histograma dos valores da estatística de teste qui-quadrado obtidos por setor censitário (somados as 27 variáveis – valor de referência: 9,8 para 99,9% de nível de confiança).

Tabela 33 – Valores da estatística de teste qui-quadrado para setores censitários da RMSP

Valor-p ⁴⁸	Contagem	Parcela
>99,9%	29.697	99,4%
95-99,9%	63	0,2%
<95%	109	0,4%
Total	29.869	100,0%

Observa-se que a grande maioria dos setores censitários (99,4%) apresentou valor-p maior do que 99,9%, o que indica uma ótima qualidade de ajuste entre as variáveis observadas (controladas) e estimadas (sintetizadas). O valor médio (por setor censitário) do valor-p foi de apenas 0,60.

Ao comparar estes valores obtidos com os valores de referência, conforme Tabela 33, observa-se que apenas para 109 dos 29.869 setores censitários (0,4% do total) não foi possível rejeitar a hipótese nula de independência entre as distribuições analisadas, considerando valor-p de 95% como referência.

Apenas 1 destes 109 setores censitários apresentavam valor da estatística do qui-quadrado acima de 100 (104,8). Este setor censitário apresentava apenas 3 domicílios e 9 moradores.

⁴⁸ Valores-p para 27 graus de liberdade: (i) 99,9%: 9,8; (ii) 99%: 12,9; (iii): 95%: 16,2.

8.4.2 Para áreas de ponderação

Por haver pouco mais de 600 áreas de ponderação na RMSP, é possível visualizar os resultados através de mapas (algo que não é razoável no caso dos setores censitários), além de gráficos.

Em uma primeira avaliação de resultados, foram observadas algumas áreas *outliers* quanto à qualidade de ajuste segundo a estatística do qui-quadrado. Estas áreas apresentavam resultados ruins especialmente para os atributos HHSIZE2 e EDUC 1 a 4. Para tentar obter resultados melhores para estas áreas (uma área no município de Santo André e as cinco áreas do município de Poá), um limite de fator de expansão de 10 foi utilizado para a geração da população sintética destes dois municípios. Com isso, foram obtidos resultados melhores para estas áreas – e consequentemente para as avaliações por variável apresentadas anteriormente.

São apresentados, nas Figura 47 e Figura 48 a seguir, os valores da estatística do qui-quadrado agregados por área de ponderação para as 33 variáveis controladas, classificados em função de seus valores-p para 33 graus de liberdade⁴⁹.

Figura 47 – Valor-p por área de ponderação na RMSP
(parte 1/2: nos municípios da RMSP exclusive SP)

⁴⁹ Valores-p para 33 graus de liberdade: (i) 99,9%: 13,4; (ii) 99%: 17,1; (iii): 95%: 20,9; (iv) 90%: 18,1; (v) 70%: 22,7.

Figura 48 – Valor-p por área de ponderação na RMSP
(parte 2/2: no município de São Paulo)

Assim como foi feito no item anterior, são apresentados a seguir um histograma dos valores do qui-quadrado (Figura 49 e Tabela 34) e dos valores-p (Tabela 35) obtidos para eles por área de ponderação.

Tabela 34 – Valores da estatística de teste qui-quadrado para áreas de ponderação da RMSP

Qui-quad	Contagem	Parcela
0-0.2	17	2.7%
0.2-0.4	79	12.5%
0.4-0.6	46	7.3%
0.6-0.8	36	5.7%
0.8-1	26	4.1%
1-2	93	14.7%
2-5	130	20.5%
5-10	88	13.9%
10-50	106	16.7%
>50	12	1.9%
<i>Total</i>	633	100.0%

Figura 49 – Histograma dos valores da estatística de teste qui-quadrado obtidos por área de ponderação (valor de referência: 9,8 para 99,9% de nível de confiança).

Tabela 35 – Valores da estatística de teste qui-quadrado para áreas de ponderação da RMSP

Valor-p ⁵⁰	Contagem	Parcela
>99,9%	545	86,1%
95-99,9%	41	6,5%
<95%	47	7,4%
<i>Total</i>	633	100,0%

⁵⁰ Valores-p para 33 graus de liberdade: (i) 99,9%: 13,4; (ii) 99%: 17,1; (iii): 95%: 20,9.

Observa-se que 47 das 633 áreas de ponderação analisadas apresentam valores-p abaixo de 95%. Apesar disso, são observadas nestas áreas de ponderação valores relativamente baixos para o erro médio relativo. A Tabela 36 (na página seguinte) apresenta o erro relativo (razão entre o erro absoluto e o total de controle) para as cinco áreas de ponderação críticas para a estatística do qui-quadrado, para as quais o valor-p não supera 50% (estatística do qui-quadrado $> 43,7$).

A diferença relativa de 100% na variável HHINC6 da APOND 4 se deve ao fato de que foram sintetizados 0 domicílios que seriam classificados nesta categoria na população sintética para os setores censitários que compõem esta área de ponderação – enquanto o total de controle especificava o valor de 75. Esta foi a única célula de valor igual a zero encontrada na tabela de sintetizados agregados. O motivo para este erro não foi identificado.

Para as demais quatro áreas de ponderação analisadas, observa-se: (i) erros acima de 10% para duas ou três das 33 variáveis, caso das APONDS 1 e 5 ou (ii) erros mais homogêneos, de alguns pontos percentuais. As médias dos módulos dos EMRs de cada área de ponderação não superaram o valor de 3% em quatro das cinco áreas de ponderação analisadas, o que indica um erro relativamente baixo para estas áreas que se destacaram negativamente.

Tabela 36 – Erro relativo para as 5 áreas de ponderação críticas para o teste do qui-quadrado

Métrica/ Variável	Erro relativo por variável por área de ponderação⁵¹				
	APOND 1	APOND 2	APOND 3	APOND 4	APOND 5
Qui-quad	298,7	108,8	86,9	80,9	76,4
NUM_HH	0%	0%	0%	0%	0%
HHSIZE1	12%	-2%	0%	0%	0%
HHSIZE2	-31%	-1%	0%	0%	0%
HHSIZE3	9%	0%	0%	1%	0%
HHSIZE4	8%	1%	0%	0%	0%
HHSIZE5	7%	2%	0%	0%	0%
HHSIZE6	5%	3%	0%	-2%	0%
HHAGE1	0%	0%	0%	0%	0%
HHAGE2	0%	0%	2%	0%	0%
HHAGE3	0%	0%	-2%	0%	0%
HHAGE4	0%	0%	-1%	0%	0%
HHINC1	0%	-7%	-2%	1%	-3%
HHINC2	0%	-5%	-1%	1%	-3%
HHINC3	0%	0%	2%	1%	-2%
HHINC4	0%	3%	7%	1%	-1%
HHINC5	0%	4%	3%	-9%	28%
HHINC6	0%	4%	0%	-100%	28%
SEXM	0%	0%	0%	0%	0%
SEXF	0%	0%	0%	0%	0%
AGEP1	0%	-5%	-2%	1%	-1%
AGEP2	0%	-5%	-3%	0%	-1%
AGEP3	0%	1%	1%	0%	0%
AGEP4	0%	2%	3%	1%	1%
AGEP5	1%	1%	1%	0%	1%
AGEP6	1%	2%	1%	0%	1%
AGEP7	1%	2%	0%	0%	1%
AGEP8	0%	1%	-2%	1%	0%
HHAUTO	0%	0%	0%	0%	0%
HHAUT1	0%	0%	0%	0%	0%
EDUC1	0%	6%	5%	1%	4%
EDUC2	0%	1%	2%	0%	2%
EDUC3	0%	-1%	-5%	-1%	2%
EDUC4	0%	-2%	-4%	2%	-11%
<i>Média²</i>	<i>2,4%</i>	<i>1,9%</i>	<i>1,5%</i>	<i>3,8%</i>	<i>2,7%</i>

⁵¹ Os “ids” das áreas de ponderação analisadas são, respectivamente: 3547809005022, 3547304003004, 3522505005011, 550308005275, 3522505005011 | ² Média do módulo dos erros médios relativos

8.5 CONSIDERAÇÕES SOBRE A VALIDAÇÃO

Consideradas as limitações para a realização da validação intrínsecas à ausência de uma *ground truth* (discutida no Capítulo 3), pode-se considerar que a população gerada é suficientemente próxima da real para fins de aplicação em modelos desagregados ao apresentar (i) resultados satisfatórios para o teste de qualidade de ajuste com a estatística do qui-quadrado e (ii) valores que podem considerados razoáveis para as demais métricas de avaliação utilizadas, ao comparar com outras aplicações de geradores de populações sintéticas:

- Para a comparação entre total sintetizado versus controlado:
 - Foram observados erros abaixo de 2 pontos percentuais em cada variável e de apenas 0,2% para o total de indivíduos. Em comparação, observou-se em Ye et al. (2009) uma subestimação de 4,6% do total de indivíduos,
 - Em Lim (2020), após um passo de pós-processamento, foram obtidos erros relativos de até 0,4%, valor superado por algumas poucas variáveis na aplicação realizada. Destaca-se que, diferente da aplicação de Lim (2020), a aplicação realizada no presente trabalho utilizou dados de duas fontes distintas (Censo e Pesquisa OD);
- Quanto às métricas de qualidade de ajuste:
 - Vovsha et al. (2015) obtiveram, em média, valores de RMSE pouco superiores que a presente aplicação, para as TAZ (com contagens compatíveis às da presente aplicação). Para o desvio padrão do erro relativo, Paul et al. (2018), obtiveram valores acima de 25% para algumas das variáveis – em comparação, na presente aplicação o máximo desvio padrão observado foi de 10%,
 - Moreno e Moeckel (2018) indicaram erros relativos da ordem de 3% para os totais por categoria das variáveis na escala de boros e municípios da Alemanha. Estes erros são compatíveis com as cinco piores áreas de ponderações (dentre mais de 600) da população sintética gerada na presente aplicação.

Ao analisar isoladamente os valores obtidos para as áreas de ponderação com os piores resultados para a estatística do qui-quadrado, observou-se que – apesar de não ser possível descartar a hipótese nula de independência (valor-p menor que 50%) – foram observados erros

relativos de cerca de 3% em média, que podem ser considerados admissíveis. Apesar de não ter sido exposta a mesma análise para setores censitários, os erros relativos observados não superam 10%, em média, nos setores com os piores desempenhos (destaca-se que apenas 12 setores apresentaram qui-quadrado maior que 50, conforme Tabela 32, e apenas 1 destes apresentou valor da estatística do qui-quadrado acima de 100).

O cálculo das métricas de validação deve servir não apenas para a verificação de conformidade do produto gerado, mas também como uma referência para a realização de eventuais *trade-offs* quanto à qualidade de ajustes de diferentes variáveis e/ou zonas. Isso pode ser feito através da alteração de dois parâmetros do PopulationSim: (i) o valor da importância de cada variável, definido no arquivo de configurações e (ii) da definição do fator de expansão máximo:

- O valor da importância das variáveis pode ser modificado caso se deseje obter estimativas mais precisas para determinados atributos. Recomenda-se que ele seja calibrado em função da importância de cada atributo para a aplicação final da população sintética e a partir dos resultados das métricas de validação obtidos. Dada a diferença das fontes, os controles no nível meta tendem a apresentar qualidades de ajuste inferiores em relação aos controles em níveis inferiores (especialmente caso estes tenham a mesma fonte da matriz semente, como no caso desta aplicação).
- A definição de um limite superior para o fator de expansão tem como benefício o aumento da diversidade da população sintética, evitando que alguns domicílios sejam replicados desproporcionalmente na população gerada. Entretanto, em determinados casos, pode haver uma falta de domicílios com determinadas características na amostra – e o uso de um fator máximo baixo pode comprometer a qualidade dos resultados.

9. CONCLUSÃO

9.1 SÍNTESE DO TRABALHO E CONCLUSÕES

Após a verificação da dificuldade para a aplicação de geradores de populações sintéticas, algo destacado por diversos autores, e devido ao fato de não haver geradores adaptados para o Brasil, o presente trabalho teve como objetivos principais a disponibilização de um gerador de populações sintéticas adaptado para o Brasil e a produção de uma população sintética para Região Metropolitana de São Paulo.

A literatura sobre geradores de populações sintéticas comumente classifica os geradores em três linhas de abordagem distintas: reconstrução sintética, otimização combinatória e aprendizagem estatística. Observou-se que os procedimentos de reconstrução sintética eram mais consolidados e se destacavam para a posterior utilização das populações geradas em modelos desagregados de demanda por transportes. A revisão da literatura permitiu também identificar os códigos disponíveis para possível utilização e a seleção de um deles para adaptar ao Brasil e aplicar. O PopulationSim, procedimento de reconstrução sintética baseado em maximização de entropia, foi selecionado.

Após a seleção do PopulationSim, para possibilitar a aplicação do procedimento no Brasil foram desenvolvidos três códigos de tratamento de dados de entrada do Censo 2010 (filtragem da área de interesse, criação da base de totais de controle e criação da matriz semente), além de um código complementar para a incorporação de informações de outra fonte (Pesquisa OD 2017 de São Paulo). Estes códigos, além de realizarem a filtragem, a extração e o tratamento dos dados de entrada, corrigem algumas inconsistências e erros encontrados nas bases originais.

Com isso, um usuário com conhecimentos básicos de computação é capaz de gerar uma população sintética para qualquer área do Brasil para os atributos presentes neste trabalho. Usuários com algum conhecimento em programação podem adaptar o procedimento para gerar populações sintéticas dotadas de múltiplos atributos distintos daqueles selecionados neste trabalho, mas presentes na base de microdados do Censo.

Adicionalmente, foram desenvolvidos códigos para a verificação da qualidade da população sintética gerada, através de procedimentos e diversas métricas de qualidade de ajuste, fornecendo insumos para que o usuário modifique os parâmetros do modelo de acordo com as suas necessidades e identifique eventuais erros nas bases de dados de entrada.

A aplicação dos códigos desenvolvidos e do PopulationSim para a geração da população sintética da Região Metropolitana de São Paulo possibilitou, além da verificação do funcionamento dos algoritmos, a avaliação da qualidade da população gerada e a disponibilização desta base “pronta para uso” para potenciais aplicações posteriores.

A população sintética gerada para a Região Metropolitana de São Paulo utilizou dados de duas fontes distintas: Censo 2010 e Pesquisa OD 2017 de São Paulo. Na aplicação, foram utilizadas três escalas geográficas distintas e controles de variáveis tanto no nível de domicílios como de indivíduos.

As avaliações dos resultados gerados, tanto finais como preliminares, permitiu a identificação de alguns pontos de atenção quanto à qualidade dos resultados. Estes pontos são discutidos a seguir, divididos em duas partes: (i) qualidade dos dados de entrada e (ii) parâmetros e funcionamento dos algoritmos.

Como em todo modelo, a consistência e confiabilidade nos dados de entrada é vital para a qualidade dos resultados produzidos pelo gerador de populações sintéticas. Nesta frente, destacou-se a identificação de lacunas (dados faltantes) não previstas nas bases de totais do Censo, para as quais foram criadas funções de preenchimento com base em outras informações disponíveis. Também foram identificados possíveis erros na base de totais do Censo através do cálculo de indicadores, solucionados através da eliminação de setores que potencialmente continham erros em suas bases, assim como em um arquivo de correspondência geográfica.

Outros dois fatores identificados que poderiam prejudicar a qualidade dos resultados são os parâmetros utilizados pelo algoritmo e os próprios códigos utilizados, tanto aqueles do gerador (PopulationSim), como os da extensão elaborados no presente trabalho. Quanto aos parâmetros de entrada, identificou-se, para alguns casos, uma potencial influência do valor máximo definido para o fator de expansão sobre a qualidade dos resultados. Quanto aos códigos, não foram identificados erros – entretanto, a existência de alguns *outliers* pode sugerir pontos a serem melhorados.

Todos os códigos gerados neste trabalho, assim como o PopulationSim, são de código aberto e encontram-se disponíveis em um repositório *online* (Ajauskas, 2021a). A base de dados da população sintética gerada para a RMSP também é disponibilizada neste repositório. Adicionalmente, foram elaborados, além do presente documento, um tutorial de instalação e

uso e uma documentação *online* para facilitar o uso dos códigos desenvolvidos por terceiros (Ajauskas, 2021b).

Espera-se que o pacote de extensão do PopulationSim para o Brasil, intitulado “PopulationSimBR”, auxilie o desenvolvimento da área de modelos baseados em atividades no Brasil – e que assim contribua, no futuro, para a elaboração de políticas públicas e nos processos de tomada de decisão de intervenções relacionadas à mobilidade e transportes no país.

9.2 LIMITAÇÕES DO PROCEDIMENTO

Uma limitação importante deste trabalho é a ausência de uma base de referência – *ground truth* – para avaliação da qualidade dos resultados obtidos. Esta é, na verdade, uma limitação comum a praticamente todas as aplicações dos geradores de populações sintéticas, uma vez que dados detalhados da população não são normalmente abertos, visando a proteger a privacidade das pessoas. A alternativa às vezes adotada é a comparação entre procedimentos distintos, avaliando-se comparativamente o atendimento aos controles. São relativamente poucas as comparações encontradas na literatura, uma vez que a maior parte dos artigos se limita a apresentar um dado procedimento. O esforço necessário para acessar e dominar o uso de cada gerador, prover os dados necessários e realizar os processamentos para uma área com a dimensão da Região Metropolitana de São Paulo, é considerável e não foi realizado no escopo desta dissertação.

Uma forma de simular a existência de uma base real de referência desagregada para avaliar a qualidade de um gerador de populações sintéticas é admitir que os microdados em si representem uma população completa de uma determinada área, da qual se extrai uma amostra para fazer o papel de semente. Os microdados originais forneceriam, neste caso, os totais de controle (além de eventuais controles meta adicionados) e a comparação poderia se dar entre cada domicílio e indivíduo sintetizado e aqueles presentes nos microdados/população. Ainda assim, este tipo de aproximação é limitado pelo tamanho da amostra dos microdados e do fato de que os controles seriam estimativas, sujeitas a erro.

Outro aspecto não tratado aqui se refere a aplicações visando projeções futuras da população. Embora do ponto de vista do procedimento não haja diferenças importantes, o fato de que os controles representam previsões sobre os diferentes atributos e sua distribuição no espaço, deve-se esperar a ocorrência mais frequente de inconsistências nessas previsões. Neste sentido, o

PopulationSim apresenta vantagens em relação a outros sintetizadores, uma vez que sua formulação permite a flexibilização do atendimento aos controles.

Uma limitação conceitual, que afeta não apenas o PopulationSim, mas os procedimentos de reconstrução sintética em geral, refere-se à crítica feita por Farooq *et al.* (2013): pelo fato de estas técnicas simplesmente clonarem os domicílios disponíveis na amostra (de 7% na RMSP, mas que pode ser de 1% em outros países), a população sintética tem sua diversidade limitada. Combinações menos usuais de atributos na população podem não aparecer na amostra, embora possam existir, ainda que em pequena proporção, na população.

9.3 TRABALHOS FUTUROS

Identificam-se cinco frentes principais para melhoria e desenvolvimento do presente trabalho, descritas a seguir.

9.3.1 Melhorias do *software*

O presente trabalho, por se tratar também do desenvolvimento da extensão de um *software*, pode ser melhorado do ponto de vista de facilitar o seu uso por parte de um novo usuário. Dentre as melhorias a se fazer, entende-se que as seguintes sejam prioritárias:

- Criação de uma interface gráfica para seleção de (i) áreas e (ii) variáveis do Censo;
- Criação de um instalador mais amigável, sem exigir instalação do Anaconda;
- Criação de um módulo de projeção temporal de variáveis.

9.3.2 Aplicação da população gerada em um modelo “final”

O produto do presente trabalho por si só não permite a extração de conclusões que possam subsidiar a tomada de decisão, e sim se configura como um primeiro passo para modelos mais acurados de previsão de demanda por transportes ou outras aplicações, citadas na introdução. Dessa forma, uma continuação natural deste trabalho seria a aplicação em algum modelo, como um ABM, com a verificação da possível melhoria das projeções de demanda ao se utilizar o PopulationSim ao invés de procedimentos mais simples de geração de populações sintéticas ou o uso de estimativas agregadas da população segundo categorias sociodemográficas e econômicas.

9.3.3 Análise de sensibilidade de parâmetros do modelo

Conforme indicado no Capítulo 5, o PopulationSim permite a definição de “importâncias” para cada categoria de cada atributo utilizado como controle. Visto que alguns atributos são mais importantes que outros (a depender da natureza da modelagem a realizar), sugere-se que sejam consultados estudos que apresentem as contribuições de diversos atributos para a estimação de variáveis de interesse (como a geração de viagens) e que uma análise de sensibilidade seja feita para verificar os eventuais ganhos obtidos.

9.3.4 Uso de variáveis cruzadas como controles

O Censo 2010 disponibiliza, em sua base de “agregados por setor censitário”, totais que se referem à combinação de alguns atributos, em particular o cruzamento entre categorias de sexo e idade (homens com idade entre 21 e 25 anos, por exemplo). Apesar de se considerar que estes totais combinados sejam informações importantes para a qualidade da população gerada, estes não foram utilizados na aplicação realizada devido ao número de variáveis que seriam adicionadas no modelo – e ao aumento da complexidade que isso geraria para as discussões sobre validação.

Todavia, recomenda-se que aplicações com objetivos além da própria geração da população sintética (como aplicação em modelos de demanda ou avaliação da distribuição da população no território) incorporem estas variáveis combinadas no modelo, caso sejam de interesse para a aplicação final. Isso exigiria, além da obtenção dessas variáveis na base de totais de controle, a criação de variáveis auxiliares que representem as categorias combinadas na matriz semente.

9.3.5 Avaliação da qualidade de ajuste para variáveis não controladas

Voas e Williamson (2000) sugeriram que se avaliasse a qualidade de ajuste obtida para variáveis não controladas no processo de geração da população sintética. Para isso, seria necessário incorporar à matriz semente um atributo cujo total de controle estivesse disponível na base de agregados por setor censitário. A qualidade de ajuste para estas variáveis (categorias deste atributo) dependeria, principalmente, do grau de correlação entre o atributo não controlado e os demais atributos controlados.

REFERÊNCIAS

ActivitySim (2020). ActivitySim: An open platform for activity-based travel modeling. Disponível em <<https://activitysim.github.io/>>. Acesso em: 23 dez. 2020.

Ajauskas, R. (2021a). PopulationSimBR. Repositório do GitHub. Disponível em <<https://github.com/ajauskas/PopulationSimBR>>. Acesso em: 11 jan. 2021.

Ajauskas, R. (2021b). Documentação PopulationSimBR: a extensão do PopulationSim para o Brasil. Disponível em <<https://populationsimbr.readthedocs.io/>>. Acesso em: 11 jan. 2021.

Arentze, T., Timmermans, H., e Hofman, F. (2007). Creating synthetic household populations: problems and approach. *Transportation Research Record*, 2014(1), 85-91.

Auld, J., e Mohammadian, A. (2010). Efficient methodology for generating synthetic populations with multiple control levels. *Transportation Research Record*, 2175(1), 138-147.

Auld J. A., Mohammadian A. K. e Wies K. (2008), Population synthesis with control category optimization, *10th International Conference on Application of Advanced Technologies in Transportation*, 2008. Atenas, Grécia.

Bar-Gera, H., Konduri, K. C., Sana, B., Ye, X., e Pendyala, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods. *Transportation Research Board 88th Annual Meeting Transportation Research Board*, (09-1354).

Barthelemy, J., e Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, 47(2), 266-279.

Beckman, R. J., Baggerly, K. A., e McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6), 415-429.

Bierlaire M., Farooq, B., Hurtubia, R., e Flötteröd, G. (2013). Simulation based Population Synthesis. Apresentação. Disponível em: <<https://infoscience.epfl.ch/record/195968/files/MIT2013.pdf>>. Acesso em: 9 abr. 2019

Borysov, S. S., Rich, J., e Pereira, F. C. (2018). Scalable Population Synthesis with Deep Generative Modeling. arXiv preprint arXiv:1808.06910.

Bowman, J. L. (2004). A comparison of population synthesizers used in microsimulation models of activity and travel demand. Unpublished working paper. Disponível em:

<http://jbowman.net/papers/2004.Bowman.Comparison_of_PopSyns.pdf>. Acesso em: 9 abr. 2019.

BRASIL. Decreto nº 76.900, de 23 de dezembro de 1975. Institui a Relação Anual de Informações Sociais – RAIS e dá outras providências.

Casati, D., Müller, K., Fourie, P. J., Erath, A., e Axhausen, K. W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record*, 2493(1), 107-116.

Deming, W. E., e Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4), 427-444.

Farooq, B., Bierlaire, M., Hurtubia, R., e Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58, 243-263.

Guo, J. Y., e Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014(1), 92-101.

Harland, K., Heppenstall, A., Smith, D., e Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, 15(1).

Hörl, S., e Balac, M. (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, 130, 103291.

IBGE (2010). Censo 2010. Disponível em: <<https://censo2010.ibge.gov.br/>>. Acesso em: 9 abr. 2019.

Jain, S., Ronald, N., e Winter, S. (2015). Creating a synthetic population: A comparison of tools. In *Proceedings of the 3rd Conference Transportation Reserch Group, Kolkata, India* (pp. 17-20).

Konduri, K., You, D., Garikapati, V. M., e Pendyala, R. (2016). Application of an Enhanced Population Synthesis Model that Accommodates Controls at Multiple Geographic Resolutions. In *Proceedings of the 95th Annual Meeting of the Transportation Research Board, Washington, DC, USA* (pp. 10-14).

- Lee, D. H., e Fu, Y. (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. *Transportation Research Record*, 2255(1), 20-27.
- Lim, P. P. (2020). Population synthesis for travel demand modeling in Australian capital cities. Tese (Doutorado). Institute for Social Science Research, University of Queensland.
- Lovelace, R, Birkin, M, Ballas, D & van Leeuwen, E 2015, 'Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique', *Journal of Artificial Societies and Social Simulation*, vol. 18, no. 2, pp. 21.
- MARG (2016). PopGen: Synthetic Population Generator. Mobility Analytics Research Group. Disponível em: <<http://www.mobilityanalytics.org/popgen.html>>. Acesso em: 9 abr. 2019.
- MATSIM.org (2019). Disponível em <<https://matsim.org/>>. Acesso em: 9 abr. 2019.
- Miranda, D.F. (2019). Reviewing Synthetic Population Generation For Transportation Models Over The Decades. Trabalho final de disciplina. Universidade de Brasília.
- Moreno, A., e Moeckel, R. (2018). Population synthesis handling three geographical resolutions. *ISPRS International Journal of Geo-Information*, 7(5), 174.
- Müller, K. (2016) *MultiLevelIPF: Implementation of algorithms that extend IPF to nested structures*. R package version 0.3-2. Disponível em: <<http://krmlr.github.io/MultiLevelIPF>>. Acesso em: 9 abr. 2019.
- Müller, K. (2017). A generalized approach to population synthesis. Tese (Doutorado) – ETH Zurich.
- Müller, K., e Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. *Arbeitsberichte Verkehrs-und Raumplanung*, 638.
- Müller, K., e Axhausen, K. W. (2011). Hierarchical IPF: Generating a synthetic population for Switzerland. [*Arbeitsberichte Verkehrs-und Raumplanung*], 718.
- Ortúzar, J.d.D., e Willumsen, L. G. (2011). Modelling Transport – Fourth Edition. Reino Unido, John Willey e Sons.
- Ory, D. (2017). A first step toward creating a digital planning laboratory is populating it: Introducing Doppelgänger, Model Lab's open-source population synthesizer. *Sidewalk Talk*.

Disponível em: <<https://medium.com/sidewalk-talk/a-first-step-toward-creating-a-digital-planning-laboratory-is-populating-it-beeb87d485f1>>. Acesso em: 9 abr. 2019.

Paul, B. M., Doyle, J., Stabler, B., Freedman, J., e Bettinardi, A. (2018a). Multi-level population synthesis using entropy maximization-based simultaneous list balancing (No. 18-03886). *Transportation Research Board 97th Annual Meeting Transportation Research Board*.

Paul, B. M., Doyle, J., Stabler, B., Freedman, J., e Bettinardi, A. (2018b). Multi-level population synthesis using entropy maximization-based simultaneous list balancing. 2018 TRB poster. Disponível em: <https://github.com/ActivitySim/populationsim/raw/master/papers/TRBPoster_PaulDoyleStablerFreedmanBettinardi_v5%5B1%5D.pptx>. Acesso em: 7 jan. 2021.

Pendyala, R. M., You, D., Garikapati, V. M., Konduri, K. C., e Zhou, X. (2017). Paradigms for integrated modeling of activity-travel demand and network dynamics in an era of dynamic mobility management. *Transportation Research Board 96th Annual Meeting Transportation Research Board*.

Pianucci, M. N. (2016). Uma proposta para a obtenção da população sintética através de dados agregados para modelagem de geração de viagens por domicílio. Tese (Doutorado). Universidade de São Paulo.

Pritchard, D. R., e Miller, E. J. (2009). Advances in agent population synthesis and application in an integrated land use and transportation model. *Transportation Research Board 88th Annual Meeting Transportation Research Board*, (09-1686).

Sidewalk Labs (2018). Replica: A next-generation urban planning tool. Vídeo. Disponível em <<https://www.youtube.com/watch?v=YKIrSUCeOtU>>. Acesso em: 9 abr. 2019.

Ribeiro, R. A. (2011). Modelo baseado em agentes para estimar a geração e a distribuição de viagens intraurbanas. Tese (Doutorado). Universidade de São Paulo.

Ryan, J., Maoh, H., e Kanaroglou, P. (2009). Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, 41(2), 181-203.

- Saadi, I., Mustafa, A., Teller, J., Farooq, B., e Cools, M. (2016). Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, 90, 1-21.
- Saadi, I., Farooq, B., Mustafa, A., Teller, J., e Cools, M. (2018). An efficient hierarchical model for multi-source information fusion. *Expert Systems with Applications*, 110, 352-362.
- Sallard, A., Balać, M., e Hörl, S. (2020). A synthetic population for the greater São Paulo metropolitan region. *Arbeitsberichte Verkehrs-und Raumplanung*, 1545.
- Smith, L., Beckman, R., e Baggerly, K. (1995). *TRANSIMS: Transportation analysis and simulation system*. Los Alamos National Lab., NM (United States).
- SOABM (2020a). ODOT Southern Oregon Activity Based Model - Running the Population Synthesizer. Disponível em: <<https://github.com/RSGInc/SOABM/wiki/Running-the-Population-Synthesizer>>. Acesso em: 15 dez. 2020.
- SOABM (2020b). ODOT Southern Oregon Activity Based Model - Future year population. Disponível em: <<https://github.com/RSGInc/SOABM/wiki/Future-Year-Population>>. Acesso em: 16 dez. 2020.
- STM e Metrô, 2019 (Secretaria dos Transportes Metropolitanos do Estado de São e Companhia do Metropolitano de São Paulo). OD - A Mobilidade Urbana da Região Metropolitana de São Paulo em Detalhes. Versão 4, 2019.
- Sun, L., e Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, 61, 49-62.
- Sun, L., Erath, A. e Ming C. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B Methodological*. 114.
- Templ, M. (2014). Simulation of Synthetic Populations. Apresentação. Disponível em <https://www.researchgate.net/publication/262451017_Simulation_of_Synthetic_Populations>. Acesso em: 9 abr. 2019
- Voas, D., e Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(5), 349-366.

Voas, D., e Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 5(2), 177-200.

Vovsha, P., Hicks, J. E., Paul, B. M., Livshits, V., Maneva, P., e Jeon, K. (2015). New features of population synthesis. *Transportation Research Board 94th Annual Meeting Transportation Research Board*, (15-5343).

Williamson, P., Birkin, M., e Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, 30(5), 785-816.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., e Waddell, P. (2009). 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, Washington, DC*.

Ye, X., Pendyala, R. M., Konduri, K., e Sana, B (sem data). A Synthetic Population Generator that Matches Both Household and Person Attribute Distributions. Disponível em: <<https://slideplayer.com/slide/6658441/>>. Acesso em: 9 abr. 2019.

APÊNDICE A - CONTEÚDO DIGITAL

Este apêndice tem como objetivo apresentar e descrever brevemente o conteúdo complementar a esta dissertação, disponibilizado através de um repositório de códigos (Ajauskas, 2021a) e um site contendo a documentação (Ajauskas, 2021b).

Repositório

O repositório de códigos contém também os arquivos pode ser acessado através do seguinte link: <https://github.com/ajauskas/PopulationSimBR> (Ajauskas, 2021a). A Figura 50 é uma captura de tela da página inicial do repositório.

Figura 50 – Recorte da tela do repositório de códigos “PopulationSimBR”

São disponibilizados no repositório de códigos:

- Códigos de tratamento de dados de entrada:
 - a. Área a Sintetizar,
 - b. Totais de controle,
 - c. Matriz semente,
 - d. Meta controles: ODSP 2017,
 - e. Meta controles: RAIS 2010;
- Códigos de validação:
 - Qualidade do ajuste,
- Arquivos de base do PopulationSimBR:
 - Layout_microdados_Amostra.xls,
 - comp.csv,
 - ibgeod_cross_walk.csv;
- Arquivos necessários (pastas /bruto, /data e /config) para a realização do tutorial;
- População sintética da RMSP (domicílios e indivíduos).

Documentação

A documentação descreve os códigos desenvolvidos, de maneira menos teórica e mais aplicada do que o conteúdo disponibilizado no presente trabalho, e pode ser acessada através do seguinte link: <https://populationsimbr.readthedocs.io/> (Ajauskas, 2021b). A Figura 51 indica um recorte da página inicial do *site*.

Figura 51 – Recorte da tela da página de documentação do “PopulationSimBR”

É composta, atualmente, pelas seguintes seções:

- Guia de instalação e uso
- Códigos desenvolvidos
 - Códigos de tratamento de dados de entrada
 - Área a Sintetizar
 - Totais de controle
 - Matriz semente
 - Meta controles: ODSP 2017
 - Meta controles: RAIS 2010
 - Códigos de validação
 - Qualidade do ajuste
- Tutorial: aplicação na região oeste do Acre
 - Introdução
 - Passo a passo
 - Resultados e Validação
- População Sintética da RMSP

APÊNDICE B - CÓDIGOS DE GERAÇÃO DE POPULAÇÕES SINTÉTICAS

São apresentados neste apêndice os *links* para acesso à documentação e/ou códigos dos geradores de populações sintética em código aberto encontrados (listados na Tabela 37).

Tabela 37 – Lista de geradores de populações sintéticas de códigos código aberto identificados

Id	Projeto ou Autor(es)	Método(s)*	Linguagem de programação
1	Müller (2017)	ML IPF	R
2	SILO	IPU	Java
3	PopGen 2.0 (Maricopa)	IPU / Ent	Python
4	SynthPop	IPU	Python
5	Sun <i>et al.</i> (2018)	HMM	MATLAB
6	PopSyn III	Ent	R
7	PopulationSim	Ent	Python

* ML IPF: Multi-level Iterative Proportional Fitting / IPU: Iterative Proportional Updating / HMM: Hierarchical Mixture Model / Ent: Maximização de Entropia

Links:

- [1] Müller (2017): <https://github.com/krlmlr/MultiLevelIPF>
- [2] SILO: <https://silo.zone/synPopDE.html> & <https://github.com/msmobility/silo/>
- [3] PopGen 2.0: <https://www.mobilityanalytics.org/popgen.html> (código via download no site)
- [4] PopGenPy: <https://urbansim.com/udst> & <https://github.com/UDST/synthpop>
- [5] Sun *et al.* (2018): https://github.com/mcgill-smart-transport/population_synthesis_hierarchical
- [6] PopSyn III: <https://github.com/BayAreaMetro/popsyn3>
- [7] PopulationSim: <https://activitysim.github.io/populationsim/index.html> & <https://github.com/ActivitySim/populationsim>

APÊNDICE C - ATRIBUTOS UTILIZADOS NA APLICAÇÃO NA RMSP

Nível	Atributo	Categorias (quantidade e descrição)	Importância (PopulationSim)	Semente: tabela e variável(is) nos Microdados do IBGE	Controles: tabela e variável(is) nos Agregados por Set. Cens. IBGE ou ODSP
Domicílio	NUM_HH Número de domicílios	1 (Total)	999999	- (não há) Contagem de registros (linhas) nos microdados	Domicilio01 - V002 <i>Domicílios particulares permanentes</i>
	HHSIZE Moradores no domicílio	6 1; 2; 3; 4; 5; 6+	5000	Domicílios - V0401 <i>QUANTAS PESSOAS MORAVAM NESTE DOMICÍLIO EM 31 DE JULHO DE 2010</i>	Domicilio01 - V050 a V059 <i>Domicílios particulares permanentes com [N] moradores</i>
	HHAGE Idade do responsável pelo domicílio	4 10-29; 30-44; 45-59; 60+	500	Pessoas - V6036 (dos indivíduos cujo RELP = 1') <i>VARIÁVEL AUXILIAR DA IDADE CALCULADA EM ANOS</i>	Responsavel02 - V002 a V092 <i>Pessoas Responsáveis com [N] anos de idade</i>
	HHINC Renda per capita em SMs do domicílio	6 <0,5; 0,5-1; 1-2; 2-3; 3-5; 5+ SMs ²	500	Domicílios - V6532 <i>RENDIMENTO DOMICILIAR PER CAPITA, EM Nº DE SALÁRIOS MÍNIMOS, EM JULHO DE 2010</i>	DomicilioRenda - V005 a V014 <i>Domicílios particulares com rendimento nominal mensal domiciliar per capita de [N] a [N] salário mínimo</i>
	HHAUT Posse de automóvel no domicílio	2 0 (Não); 1 (Sim)	5000	Domicílios - V0222 <i>V0222: AUTOMÓVEL PARA USO PARTICULAR, EXISTÊNCIA</i>	ODSP - Tabela 7, valores das colunas B (0) e C+D+E (1) <i>Tabela 7 - Famílias por Número de Automóveis Particulares e Zona de Residência</i>
	HHMOT Posse de moto no domicílio	2 0 (Não); 1 (Sim)	-	Domicílios - V0221 <i>V0221: MOTOCICLETA PARA USO PARTICULAR, EXISTÊNCIA</i>	- (não há) Não disponível no IBGE ou na ODSP
	HHCRIAN Presença ou não de criança no domicílio	2 0 (Não); 1 (Sim)	-	<i>Boolean criado a partir de Pessoas - V6036 (dos domicílios em que há ao menos um indivíduo c/ AGEP < 18)</i> <i>VARIÁVEL AUXILIAR DA IDADE CALCULADA EM ANOS</i>	- (não há) Não disponível no IBGE ou na ODSP
Indivíduo	SEX Sexo do indivíduo	2 1 (Masculino); 2 (Feminino)	5000	Pessoas - V0601 <i>SEXO: 1- Masculino 2- Feminino</i>	Pessoa11 - V002 / Pessoa12 - V002 <i>Homens/Mulheres residentes em domicílios particulares</i>
	AGEP Idade do indivíduo	8 Ver nota ⁴	500	Pessoas - V6036 <i>VARIÁVEL AUXILIAR DA IDADE CALCULADA EM ANOS</i>	Pessoa03 - V007 a V086 <i>Pessoas Residentes de [N] a [N] anos de idade e cor ou raça – [raça] (observação: agregado por raça)</i>
	EDUC Grau de instrução	4 Ver nota ⁵	500	Pessoas - V6400 <i>NÍVEL DE INSTRUÇÃO: 1- Sem instrução e fundamental incompleto 2- Fundamental completo e médio incompleto [...]</i>	ODSP - Tabela 3, valores das colunas B+C, D, E e F <i>Tabela 3 - População por Grau de Instrução e Zona de Residência</i>
	OCUP Ocupação do indivíduo	11 10 classes + vazio (sem ocupação) Ver nota ⁶	-	Pessoas - V6471 <i>OCUPAÇÃO – A relação de códigos encontra-se no arquivo: “Ocupação COD_Estrutura 2010.xls”</i>	- (não há) Não disponível no IBGE ou ODSP (RAIS exigiria premissa)

Observações e notas:

¹ Grau de parentesco = pessoas responsáveis pelo domicílio | ² SMs = Salários mínimos, de julho de 2010 | ³ Cada categoria, para o procedimento, passa a funcionar como variável de controle

⁴ 1ª categoria: 0-4 anos / 2ª: 5-14 / 3ª: 15-19 / 4ª a 7ª: 10 em 10 anos / 8ª: 60+ | ⁵ 1 (Fundamental Incompleto), 2 (Fundamental), 3 (Médio), 4 (Superior) | ⁶ Grandes grupos da Classificação Brasileira de Ocupações 2002 (CBO), de 1 dígito

Células em verde indicam fontes diferentes do Censo 2010; células cinzas indicam a não utilização de controles para o atributo