TRANSPORTES
revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
anpet.org.br

Procedimento para geração de populações sintéticas com base em dados disponíveis no Brasil

Synthetic population generation procedure based on Brazilian data

Rodrigo Ajauskas[1], **Orlando Strambi**[1]

[1]Universidade de São Paulo, São Paulo, São Paulo, Brasil

Contato: rodrigo_ajauskas@hotmail.com, (RA); ostrambi@usp.br, (OS)

Recebido:
5 de maio de 2021
Revisado:
12 de junho de 2024
Aceito para publicação:
13 de junho de 2024
Publicado:
9 de setembro de 2024
Editor de Área:
Bruno Vieira Bertoncini, Universidade Federal do Ceará, Brasil

Palavras-chave:
Populações sintéticas.
Geradores de populações.
Sintetizadores de populações.
Maximização de entropia.

Keywords:
Synthetic population.
Synthetic population generators.
Population synthesizers.
Entropy maximization.

DOI: 10.58922/transportes.v32i3.2617

RESUMO

Este trabalho apresenta um gerador de populações sintéticas adaptado para o Brasil e sua aplicação para a Região Metropolitana de São Paulo (RMSP). Populações sintéticas são utilizadas em modelos desagregados de previsão de demanda; resultam da estimação de informações desconhecidas em escalas geográficas desagregadas, tendo como base informações agregadas conhecidas e (uma amostra de) microdados, ambos disponibilizados pelos Censos. Considerando as diferentes abordagens teóricas e a disponibilidade de códigos, selecionou-se o gerador PopulationSim, pertencente à categoria de procedimentos de reconstrução sintética. Desenvolveu-se uma extensão, chamada de PopulationSimBR, para facilitar a aplicação do gerador em regiões no Brasil. Na aplicação realizada para a RMSP foram também utilizados dados da Pesquisa OD de São Paulo. Os resultados apresentam indicadores de qualidade superior aos encontrados na literatura, o que sugere que o PopulationSim pode ser utilizado no Brasil, assim como a base de dados gerada para a RMSP.

ABSTRACT

This paper presents both a population synthesizer adapted to Brazil and its application to the Metropolitan Area of São Paulo (RMSP). Synthetic populations are used in disaggregate travel demand models; they result from estimating unknown information at a fine geographical level based on available aggregated information and (a sample) of microdata, both made available by the Census. Considering the theoretical approaches and the availability of codes, we selected PopulationSim, synthesizer, belonging to the category of synthetic reconstruction synthesizers. An extension, called PopulationSimBR, was developed to facilitate the use of this synthesizer in different regions of Brazil. OD survey data files were used in addition to Census data for the application to the RMSP. Validation metrics show that results compare favorably to those reported in the literature and suggest that PopulationSim can be used in Brazil, as well as the synthetic population generated for the RMSP.

1. INTRODUÇÃO

Geradores de populações sintéticas têm se tornado uma ferramenta importante para os planejadores de transportes em seus esforços de modelagem da demanda (Ramadan e Sisiopiku, 2019). Isto se deve à crescente importância dos modelos de transportes baseados em atividades (ABM, do inglês *Activity-Based Models*), que utilizam populações sintéticas como dado de entrada.

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 estrutura 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. 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*).

A pesquisa na área de geradores de populações sintéticas tem como objetivo estimar informações desconhecidas em escalas geográficas desagregadas, com base em informações agregadas conhecidas. O processo para geração da população sintética normalmente se baseia em dados disponibilizados por institutos nacionais de estatística, nas formas de:

- Amostras da população (de ao menos 5% do total, no caso do Brasil), referidos como microdados, com informações completas dos domicílios e indivíduos, porém sem precisão espacial (na escala das chamadas áreas de ponderação do Censo); 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 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 considerar os totais dados para áreas maiores do que aquelas em que são fornecidos os microdados da amostra (como por exemplo, municípios). 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 (Ramadan e Sisiopiku, 2019).

Observou-se, neste panorama de desenvolvimento de diversos procedimentos para geração de populações sintéticas, que não havia um procedimento devidamente adaptado para ser aplicado com dados disponíveis no Brasil, independentemente da região. Tampouco havia 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, considerou-se que o desenvolvimento de um procedimento de fácil utilização e adaptado ao contexto nacional (isto é, práticas de modelagem e com os dados disponíveis no Brasil) seria pertinente para contribuir com o desenvolvimento da área no país.

Dessa forma, o presente trabalho tem como objetivos apresentar um pacote de códigos desenvolvido para a aplicação de um gerador de populações sintéticas no Brasil, além da validação dos resultados gerados: uma base de dados completa e desagregada da população da RMSP.

O pacote de códigos desenvolvido, denominado PopulationSimBR, tem como intuito facilitar o processo de geração de populações sintéticas para regiões do Brasil, permitindo a extração e tratamento dos dados do Censo para um formato próprio para alimentar o gerador de populações sintéticas PopulationSim (Paul et al., 2018).

2. REVISÃO DA LITERATURA

Esta seção inicia com uma visão geral dos procedimentos para geração de populações sintéticas; na sequência, apresenta a linha de procedimentos baseados em maximização de entropia e, por fim, discute o uso de programação linear para esses procedimentos, segundo a metodologia de Vovsha et al. (2015).

Destaca-se que a maioria destes procedimentos foi desenvolvida com o objetivo de alimentar modelos de demanda por transportes. Entretanto, considerando os objetivos do presente artigo, estas aplicações não são descritas no presente item.

2.1. Procedimentos para geração de populações sintéticas

Beckman, Baggerly e McKay (1996) utilizaram o método IPF (*Iterative Proportional Fitting*) para a geração da população sintética do simulador TRAMSIMS (Smith, Beckman e Baggerly, 1995). Diversos métodos de geração de populações sintéticas foram desenvolvidos desde este trabalho seminal, cujo procedimento apresentava limitações que outros autores buscaram superar, seja seguindo a mesma abordagem (usualmente denominada reconstrução sintética) ou outras (como otimização combinatória e aprendizagem estatística), conforme descrito em Ajauskas (2021a)

Nas últimas duas décadas houve uma evolução das abordagens de reconstrução sintética (utilizando IPF ou maximização de entropia) – destacando-se as linhas de (i) *Iterative Proportional Updating* (IPU), 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. (2018).

Também são notáveis os desenvolvimentos mais recentes baseados em simulação ou utilizando técnicas de aprendizagem estatística, em que se destacam Farooq et al. (2013) e, mais recentemente, Sun, Erath e Ming (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.

No Brasil, foram identificadas poucas aplicações dos procedimentos de geração de populações sintéticas. Pianucci (2016) desenvolveu 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, utilizou o Método de Monte Carlo, introduzido por Birkin e Clarke (1988). Ribeiro (2011) não tinha como foco a geração de populações sintéticas, entretanto tratou o assunto em sua tese em que desenvolveu um modelo de geração e distribuição de viagens intraurbanas, utilizando o procedimento proposto por Miyamoto et al. (2010) – que também é um Método de Monte Carlo. Mais recentemente, Sallard et al. (2020) aplicaram um procedimento de geração de populações sintéticas para São Paulo em que a expansão é feita diretamente com o fator de expansão disponibilizado pelo Censo, sendo a localização de cada residência definida com base em uma correspondência com as amostras da Pesquisa OD 2017 de São Paulo, que apresentam esta informação.

Uma revisão dos principais métodos de reconstrução sintética baseadas em IPF pode ser encontrada em Ajauskas e Strambi (2019). Como complemento a esse trabalho, é apresentada nos itens seguintes uma breve revisão sobre os métodos baseados no conceito de maximização de entropia, em especial o proposto por Vovsha et al. (2015), que utiliza programação linear e constitui a base teórica para o gerador de populações sintéticas PopulationSim.

Uma revisão atual e mais abrangente dos métodos de geração de populações sintéticas disponíveis pode ser encontrada em Ramadan e Sisiopiku (2019).

2.2. Procedimentos baseados em maximização da entropia

O princípio de maximização da entropia foi inicialmente aplicado na geração de populações sintéticas por Bar-Gera et al. (2009) e Lee e Fu (2011), que buscavam tratar o problema do uso de controles de atributos tanto de indivíduos como domicílios, enfrentado pelos demais geradores de populações sintéticas da época, através da resolução de um problema de otimização sujeito a restrições. Bar-Gera et al. (2009) também introduziram o princípio de flexibilização dos totais, ao permitir, em seu algoritmo, um desvio entre os totais de controle e os totais agregados obtidos da sintetização.

Apesar de não utilizarem IPF, as técnicas baseadas em maximização de entropia também partem do princípio de rebalancear uma matriz semente, que representa a distribuição conjunta dos atributos obtida a partir da amostra do Censo. Assim, também podem ser enquadradas na categoria de reconstrução sintética.

Posteriormente, Barthelemy e Toint (2013) utilizaram maximização de entropia em um método que, diferente da grande maioria 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, combinar estes indivíduos para realizar a composição dos domicílios.

2.3. Maximização de entropia utilizando programação linear

Vovsha et al. (2015) propuseram um procedimento baseado em maximização da entropia utilizando programação linear. Os autores destacaram em seu artigo três funcionalidades principais 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.

O procedimento de Vovsha et al. (2015) resolve problemas de programação linear com restrições em dois de três dos seus processos: (i) o balanceamento, em que os pesos fornecidos pelo censo para cada domicílio na amostra são redefinidos e (ii) a "integralização", em que os pesos fracionais resultantes são transformados em números inteiros. O terceiro processo, fatoração, é aplicado quando se deseja utilizar totais de controle em uma escala maior do que aquela em que são apresentados os microdados (no caso do Censo brasileiro, as áreas de ponderação).

Paul et al. (2018) 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" simultaneamente para todas as zonas – ao invés de sequencialmente, como é feito por Vovsha et al. (2015).

O procedimento de Paul et al. (2018) é o adotado pelo gerador de populações sintéticas PopulationSim, que foi o utilizado na aplicação do presente trabalho.

2.4. Geradores de populações sintéticas disponíveis

São apresentadas nesta seção seis implementações de geradores de populações sintéticas de código aberto, listados na Tabela 1 e brevemente discutidos na sequência.

A revisão destes geradores teve como objetivo identificar aquele mais apropriado para uso e eventual adaptação para o Brasil. Após avaliar mais de uma dezena de geradores de populações sintéticas, Lim (2020) afirma que muitos procedimentos estão "escondidos" em algoritmos praticamente inacessíveis devido à baixa qualidade dos códigos, dificultando assim a utilização por terceiros. Por este motivo, a escolha de um gerador apropriado é de extrema importância para o trabalho.

Tabela 1: Geradores de populações sintéticas de códigos aberto identificados.

Projeto ou Autor(es)	Método(s)*	Linguagem de programação
PopGen (MARG, 2016)	IPU / Ent	Python
SILO (Moreno e Moeckel, 2018)	IPU	Java
Müller (2017)	ML IPF	R
Sun et al. (2018)	HMM	MATLAB
PopSyn III (Vovsha et al., 2015)	Ent	R
PopulationSim (Paul et al., 2018)	Ent	Python

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

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, 1.1 e 2.0. A primeira versão (1.1) possui uma interface gráfica para o usuário e é capaz de baixar e processar dados do censo dos EUA automaticamente, além de possibilitar ajustes para superar inconsistência entre totais de controle nos níveis de indivíduo e domicílio. Entretanto, permite o uso de 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.

O projeto SILO (*Simple Integrated Land Use Orchestrator*), inclui o gerador de populações sintéticas baseado em IPU e aperfeiçoado por Moreno e Moeckel (2018) e foi elaborado na linguagem Java. Müller (2017) realizou a implementação de diversos procedimentos utilizando a linguagem de programação R: IPF hierárquico (HIPF), IPU, maximização de entropia e *generalized raking*. Sun et al. (2018) utilizaram uma estrutura hierárquica contendo três modelos baseados em técnicas de aprendizagem estatística e disponibilizaram seus códigos, escritos para MATLAB, em repositório do GitHub.

Outros dois geradores, que utilizam procedimentos baseados em maximização de entropia, foram identificados: o PopSyn III, elaborado a partir do algoritmo de Vovsha et al. (2015), e o PopulationSim, apresentado em Paul et al. (2018).

2. POPULATIONSIM

2.1. Escolha do *software*

Após a revisão de diversas metodologias e dos códigos disponíveis, o PopulationSim foi escolhido como o código principal a ser utilizado no presente trabalho. A escolha deste *software* se baseou em uma

série de critérios, como: a qualidade do algoritmo, a disponibilidade de documentação, a plataforma de compartilhamento de código, a linguagem de programação utilizada e a qualidade do código.

A qualidade do algoritmo está vinculada ao fato deste ser uma evolução do artigo seminal de Vovsha et al. (2015), cujo 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 diferentes escalas podem divergir.

Em relação à documentação, o PopulationSim dispõe de um artigo científico (Paul et al., 2018), um memorando técnico (RSG, 2017) e um site de documentação (PopulationSim, 2020). Além disso, o código é hospedado na plataforma GitHub, que possibilita a realização de discussões sobre possíveis problemas e melhorias para o software, além de registrar o 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.

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.

Também contribuiu para a escolha o fato de que o PopulationSim possui desenvolvedores atualmente ativos e uma comunidade que o utiliza, havendo assim potencial para melhorias.

2.2. Sobre o procedimento

O PopulationSim é uma plataforma aberta para a geração de populações sintéticas (PopulationSim, 2020) baseada na metodologia apresentada por Paul et al. (2018) e faz parte de um projeto mais abrangente de um software de modelagem baseada em atividades: o ABM ActivitySim, projeto patrocinado por um consórcio de agências de transportes de 5 estados dos EUA (ActivitySim, 2020).

A Figura 1 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.

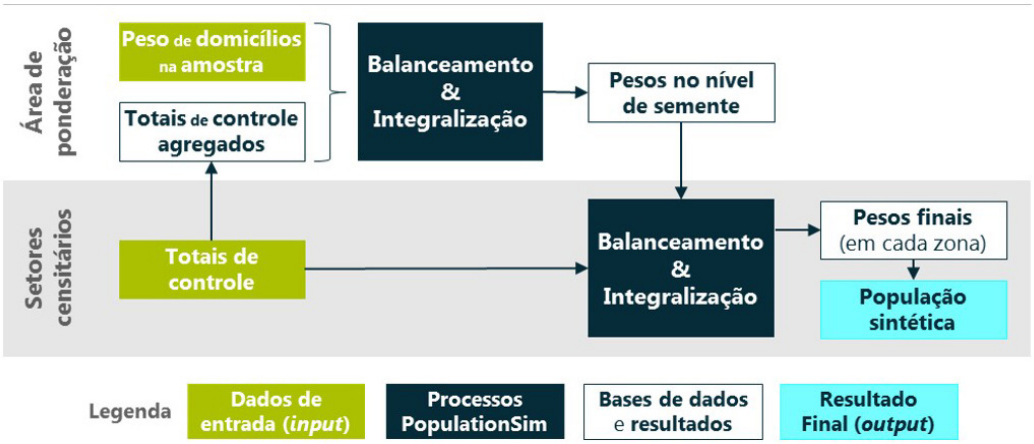

Figura 1. Processos do PopulationSim para duas escalas geográficas (adaptado de Paul et al., 2018).

O procedimento, para este cenário de dados de entrada em apenas duas escalas, inicia com a agregação dos totais de controle (disponíveis na escala dos setores censitários do Censo) para a escala em que são disponibilizados os microdados (áreas de ponderação no 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.

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, no nível de setor censitário) como restrições do problema de programação linear. Como resultado deste procedimento, têm-se os pesos finais dos domicílios que compõem cada setor censitário.

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 (Paul et al., 2018).

2.3. Expansão do populationsim para o Brasil

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 níveis de agregação diferentes, é necessário extrair os dados brutos da base de dados nacional e tratá-los, caso se deseje aplicar o procedimento do PopulationSim no Brasil.

Para isso, dois códigos de extração e tratamento de dados foram inicialmente desenvolvidos: um para a criação das tabelas de totais de controle no formato utilizado no PopulationSim; e outro para a criação da matriz semente com base nos microdados. Ambos utilizam bases de dados do Censo 2010 como dado de entrada. Um terceiro código foi desenvolvido para facilitar a especificação dos municípios de interesse para a sintetização da população por parte do usuário.

Alguns dos atributos presentes na base de microdados do Censo 2010 não possuem totais correspondentes na base de agregados por setor censitário. Isto ocorre uma vez que o questionário aplicado na amostra é mais extenso que o questionário básico utilizado no censo. Informações de outras fontes podem ser utilizadas para obter totais de controle para atributos que não estão presentes na base de agregados por setor censitário. Para a RMSP, a Pesquisa OD 2017 de São Paulo (ODSP 2017) foi identificada como possível fonte adicional, por conter, entre outras, informações sobre grau de instrução dos indivíduos e posse de automóvel nos domicílios, e se tratar de uma fonte de dados mais atual que o Censo. Apesar de apresentar as informações para escalas geográficas diferentes (zonas OD, posteriormente agregadas para distritos e municípios, conforme seção seguinte), o PopulationSim permite o uso de totais de controle em múltiplas escalas, inclusive maiores que aquela da matriz semente. Dessa forma, um terceiro código foi desenvolvido para aplicação no Brasil – este mais especificamente para a RMSP – de extração e compatibilização das informações da ODSP 2017 para o formato do PopulationSim.

Por fim, foi também elaborado um código para a validação dos resultados, em que são realizados os cálculos de métricas de qualidade de ajuste para a verificação da aderência entre os totais de controle e as contagens da população sintetizada.

Os códigos fazem parte do pacote intitulado PopulationSimBR, desenvolvido no presente trabalho.

2.4. Aplicação na RMSP

A Região Metropolitana de São Paulo (RMSP) 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.

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 (ODSP 2017). Na aplicação, foram utilizados controles em três escalas geográficas distintas (setores censitários e áreas de ponderação, do Censo 2010; e distritos/municípios, para dados da ODSP 2017) e atributos tanto de domicílios como de indivíduos, indicados na Tabela 2.

Tabela 2: Atributos selecionados para aplicação na RMSP e escala de seus controles.

Escala	Nível	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[1]	6	
	Indivíduo	**SEX**	Sexo do indivíduo	2	
		AGEP	Idade do indivíduo	8	
Distrito/ Município	Domicílio	**HHAUT**	Posse de automóvel no domicílio	2	6
	Indivíduo	**EDUC**	Grau de instrução	4	
- (não controlados)	Domicílio	**HHCRIAN**	Presença de criança no domicílio	-	-
	Indivíduo	**OCUP**	Categoria de ocupação no trabalho	-	

[1] SM = Salários mínimos de julho de 2010.

Os números de categorias dos atributos foram definidos pelo Censo, com exceção dos atributos AGEP (em que o número de categorias foi reduzido de 14 para 8) e HHSIZE (redução de 10 para 6 categorias).

Conforme apresentado na Tabela 2, dois dos atributos na presente aplicação não são controlados; isto é, são incluídos na criação da matriz semente – e consequentemente fazem parte da população sintética gerada – porém não influenciam os processos de cálculo dos pesos (balanceamento). O primeiro deles, HHCRIAN, se trata de uma variável derivada: ela é obtida através da verificação da presença ou não de indivíduos com menos de 18 anos no domicílio. Já o segundo, OCUP, está presente na base de microdados, porém não apresenta totais de controle no Censo. Estes dois atributos foram selecionados por se tratar de informações em geral relevantes para uso em modelos de estimação de demanda por transportes, principal foco desta aplicação. Uma alternativa avaliada para obtenção de totais de controle do atributo de ocupação do indivíduo foi o uso da base de dados da RAIS (Relação Anual de Informações Sociais, do Ministério do Trabalho e Emprego), inviabilizado pelo fato de esta indicar apenas o município de trabalho, e não de moradia, dos indivíduos.

Os controles obtidos da ODSP 2017, apesar de serem disponibilizados na escala de zonas OD (517 unidades na RMSP), tiveram que ser agregados para a escala de distritos no município de São Paulo (96 unidades) de forma a serem compatíveis com a agregação das áreas de ponderação. Nos demais municípios da região metropolitana, os próprios municípios foram utilizados como escala de controle (38 unidades), devido a incompatibilidades verificadas entre os distritos em alguns destes municípios e as zonas OD.

As tabelas a seguir apresentam o número de zonas de cada escala utilizada na presente aplicação na RMSP (Tabela 3), assim como o número de domicílios e pessoas, no total da população e na amostra (Tabela 4). O número de setores censitários indicado na Tabela 3 corresponde apenas àqueles que continham ao menos um domicílio particular.

Tabela 3: Número de setores censitários, 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	11.536	323	38[2]
Total	*29.869*	*633*	*134*

[1] Distritos do município de São Paulo. [2] Municípios da RMSP (exceto São Paulo).

Tabela 4: 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%*

Adaptado de IBGE (2010).

Foram sintetizados, na aplicação realizada na RMSP, apenas domicílios particulares e seus respectivos moradores. Considera-se, porém, que esta limitação não gere prejuízos para aplicações em modelagem de demanda por transportes – visto que os moradores de domicílios coletivos (que inclui presídios, quartéis, asilos, orfanatos, conventos e hospitais) representam uma parcela reduzida da população total (menos de 0,5% no caso do município de São Paulo) e, geralmente, realizam poucas viagens intraurbanas em comparação com o restante da população.

As bases de dados com a população sintética da RMSP (indivíduos e domicílios) podem ser baixadas em Ajauskas (2021b).

3. VALIDAÇÃO DOS RESULTADOS

Müller e Axhausen (2010) Aaplicaram 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 frente às suas distribuições conjuntas.

Entretanto, diferentemente da Suíça e dos países escandinavos (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 a existência de geradores de populações sintéticas. Dessa forma, face à indisponibilidade de uma base real desagregada, a validação dos resultados torna-se um desafio.

O procedimento adotado na maioria das aplicações pesquisadas, e também no presente trabalho, 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". Para a presente aplicação, foram selecionados quatro indicadores para avaliar a qualidade dos resultados: (i) o erro médio relativo, chamado de EMR; (ii) o desvio padrão do erro relativo, "DesvPad"; (iii) a raiz do erro quadrático médio, RMSE; e (iv) o qui-quadrado, "Qui-quad".

São apresentados nos próximos quatro subitens: a comparação entre os totais de controle e o total sintetizado, agregados para toda a RMSP e para cada variável (7.1); as métricas de qualidade ajuste supracitadas por variável, analisadas por setor censitário (7.2) e por distrito/município (7.3); e uma análise da distribuição dos valores obtidos para a estatística do qui-quadrado nos setores censitários (7.4).

3.1. Totais de controle *vs.* valores sintetizados

Uma primeira avaliação feita nesta seção é a comparação dos totais por variável (categorias dos atributos) para toda a RMSP através do cálculo (i) da diferença absoluta entre o valor sintetizado e o total de controle e (ii) da diferença relativa, ambas apresentadas na Tabela 5.

Quanto à população total gerada, observa-se um ajuste quase perfeito do número de domicílios (NUM_HH), com uma diferença de apenas 11 unidades dentre mais de 6 milhões de domicílios sintetizados. Já para indivíduos, o sintetizador subestimou a população total em aproximadamente 35 mil pessoas, o que corresponde a menos 0,2% do total. Em comparação, observou-se em Ye et al. (2009), que aplicou o seu procedimento para gerar uma população sintética para o condado de Maricopa (Arizona, EUA), e Lim (2020), em aplicações nas três maiores regiões metropolitanas da Austrália, subestimações de 3 a 5% do total de indivíduos.

A Tabela 5 mostra que as diferenças relativas não superaram 0,5% em nenhuma das variáveis controladas na escala de setores censitários. Para as variáveis controladas no nível meta (categorias dos atributos "posse ou não de automóvel" e "grau de instrução"), as diferenças relativas também foram pequenas: não superaram 1%, exceto na variável EDUC4 ("grau de instrução: superior completo"), que foi subestimada em 2% devido à sintetização de cerca de 55 mil indivíduos a menos com estas características do que o valor de referência utilizado (controle).

Em Lim (2020), após um passo de pós-processamento com o objetivo de melhorar o ajuste dos indivíduos, foram obtidos erros relativos de até 0,4%, valor superado por algumas poucas variáveis na aplicação aqui apresentada. 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).

Tabela 5: Diferença relativa por variável.

Atributo	Variável	Controle	Sintetizado	Diferença Absoluta	Diferença Relativa
Num. domicílios	NUM_HH	6.089.847	6.089.836	-11	0,00%
Tamanho do domicílio	HHSIZE1	766.227	765.055	-1.172	-0,15%
	HHSIZE2	1.387.336	1.385.007	-2.329	-0,17%
	HHSIZE3	1.558.274	1.558.352	78	0,01%
	HHSIZE4	1.318.714	1.319.316	602	0,05%
	HHSIZE5	618.818	619.899	1.081	0,17%
	HHSIZE6	440.478	442.207	1.729	0,39%
Idade do responsável	HHAGE1	885.428	885.387	-41	0,00%
	HHAGE2	2.137.447	2.137.444	-3	0,00%
	HHAGE3	1.824.111	1.824.248	137	0,01%
	HHAGE4	1.242.895	1.242.757	-138	-0,01%
Renda per capita em salários mínimos	HHINC1	1.247.912	1.247.555	-357	-0,03%
	HHINC2	1.541.695	1.541.771	76	0,00%
	HHINC3	1.619.191	1.619.546	355	0,02%
	HHINC4	593.607	593.823	216	0,04%
	HHINC5	493.666	493.779	113	0,02%
	HHINC6	593.780	593.362	-418	-0,07%
Sexo do indivíduo	SEXM	9.385.162	9.365.940	-19.222	-0,21%
	SEXF	10.223.963	10.207.330	-16.633	-0,16%
Idade do indivíduo	AGEP1	1.302.487	1.302.106	-381	-0,03%
	AGEP2	3.002.098	2.995.703	-6.395	-0,21%
	AGEP3	1.547.984	1.542.581	-5.403	-0,35%
	AGEP4	3.606.806	3.589.559	-17.247	-0,48%
	AGEP5	3.297.504	3.285.978	-11.526	-0,35%
	AGEP6	2.711.746	2.705.579	-6.167	-0,23%
	AGEP7	2.051.423	2.047.954	-3.469	-0,17%
	AGEP8	2.106.742	2.103.810	-2.932	-0,14%
Posse ou não de automóvel	HHAUT0	2.855.496	2.861.288	5.792	0,20%
	HHAUT1	3.234.223	3.228.548	-5.675	-0,18%
Grau de instrução	EDUC1	7.056.528	7.104.861	48.333	0,68%
	EDUC2	2.851.352	2.859.879	8.527	0,30%
	EDUC3	6.805.785	6.769.051	-36.734	-0,54%
	EDUC4	2.895.216	2.839.479	-55.737	-1,96%
População total (soma indivíduos)		*19.608.881*	19.573.270	-35.611	*-0,18%*

3.2. Avaliação por variável para setores censitários

Neste subitem é examinada a qualidade de ajuste entre totais de controle e sintetizado para cada variável (categorias dos atributos), a partir da contagem de suas incidências em cada setor censitário. Isso é feito inicialmente através do cálculo de quatro métricas que comparam a tabela de

totais de controle (*input*) e a tabela de contagem dos resultados sintetizados (*output*). Estas tabelas apresentam as 27 variáveis nas linhas e os quase 30 mil setores censitários nas colunas. As variáveis relacionadas aos atributos "posse ou não de automóvel" (HHAUT) e "grau de instrução" (EDUC) não são apresentadas por não possuírem controles no nível de setores censitários. Os resultados podem ser observados na Tabela 6.

Tabela 6: Métricas de validação por variável para setores censitários da RMSP.

		Para setores censitários (N=29.869)			
Atributo	**Variável**	**EMR**	**DesvPad**	**RMSE**	**Qui-quad**
Num. dom.	NUM_HH	0,0%	0,0%	0,00	0,0
Tamanho do domicílio	HHSIZE1	0,0%	5,5%	0,44	216,6
	HHSIZE2	0,0%	5,3%	0,87	465,1
	HHSIZE3	0,3%	7,9%	0,61	253,7
	HHSIZE4	0,0%	4,4%	0,58	276,0
	HHSIZE5	0,1%	5,3%	0,46	307,7
	HHSIZE6	-0,4%	10,7%	0,60	516,7
Idade do responsável	HHAGE1	0,0%	4,6%	0,42	216,0
	HHAGE2	0,0%	5,2%	0,61	216,8
	HHAGE3	0,0%	4,7%	0,61	244,9
	HHAGE4	-0,1%	5,0%	0,49	200,4
Renda per capita em salários mínimos	HHINC1	0,1%	6,2%	0,53	245,0
	HHINC2	0,2%	6,2%	0,55	226,7
	HHINC3	0,4%	9,4%	0,58	257,2
	HHINC4	0,2%	6,4%	0,38	215,2
	HHINC5	0,2%	6,2%	0,34	163,8
	HHINC6	-0,7%	10,7%	0,44	126,4
Sexo do indivíduo	SEXM	0,0%	6,5%	3,28	1.017,1
	SEXF	0,1%	6,6%	3,51	1.069,6
Idade do indivíduo	AGEP1	0,3%	7,1%	1,03	828,3
	AGEP2	0,5%	14,4%	1,71	941,3
	AGEP3	-0,1%	7,8%	1,03	720,7
	AGEP4	0,3%	11,9%	1,83	836,3
	AGEP5	0,3%	12,7%	1,60	707,3
	AGEP6	0,1%	9,3%	1,39	699,8
	AGEP7	0,0%	8,5%	1,11	633,6
	AGEP8	-0,3%	6,2%	1,08	583,1

Ao se avaliar os resultados da Tabela 6, observa-se:

- EMR: os valores do erro médio relativo não superaram 0,4% (em módulo), exceto para a variável HHINC6, o que sugere um equilíbrio entre erros de sinal positivo (superestimando os valores controlados nas estimativas) e de sinal negativo (subestimações);
- DesvPad: observa-se a maior parte das variáveis apresentando valores entre 4 e 8%, com o máximo de 14,4% na variável AGEP2. Em comparação, para o desvio padrão do erro relativo, Paul et al. (2018), obtiveram valores acima de 25% para algumas das variáveis;

- 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 contagem (por ter apenas duas categorias), para as quais o RMSE alcança o valor de mais de 3. Um valor mais alto para estas categorias 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 por setor censitário (exceto para as variáveis HHINC5 e HHINC6, acima de 300). Em comparação, Vovsha et al. (2015) obtiveram, em média, valores de RMSE pouco superiores que a presente aplicação para unidades geográficas compatíveis com as da presente aplicação (setores censitários);
- Qui-quad: foram obtidos excelentes resultados, visto que o valor crítico da estatística qui-quadrado para 29.873 graus de liberdade e 99,9% de probabilidade é 29.307, enquanto, segundo a Tabela 6, o valor do qui-quadrado não supera 1.100 para nenhuma variável. Isso significa que a hipótese de que as distribuições dos totais de controle e sintetizado para uma dada variável entre setores censitários sejam estatisticamente iguais não pode ser rejeitada.

Adicionalmente, apresentam-se na Tabela 7 as distribuições dos erros relativos de cada variável para os setores censitários.

Tabela 7: Parcela de observações por intervalo de erro relativo por variável.

	Valores por intervalo de erro relativo										Set. Cens. >10 obs.	
Variável	< -25%	-25% -10%	-10% -5%	-5% 0	Erro =0	0 5%	5% 10%	10% 25%	> 25%	>5% em módulo	Abs.	Relat.
NUM_HH	0,0%	0,0%	0,0%	0%	100%	0%	0,0%	0,0%	0,0%	0,0%	29.504	99%
HHSIZE1	0,0%	0,0%	0,3%	3%	93%	3%	1,0%	0,1%	0,0%	1,5%	25.326	85%
HHSIZE2	0,0%	0,2%	0,1%	1%	88%	10%	0,8%	0,0%	0,0%	1,0%	28.190	94%
HHSIZE3	0,0%	0,0%	0,1%	1%	85%	13%	1,1%	0,1%	0,0%	1,2%	28.525	96%
HHSIZE4	0,0%	0,0%	0,1%	1%	83%	14%	1,4%	0,0%	0,0%	1,6%	28.086	94%
HHSIZE5	0,0%	0,0%	0,2%	3%	84%	9%	3,7%	0,1%	0,0%	4,0%	23.603	79%
HHSIZE6	0,0%	0,0%	0,6%	6%	73%	11%	8,6%	0,4%	0,0%	9,6%	18.298	61%
HHAGE1	0,0%	0,0%	0,2%	2%	88%	7%	2,1%	0,0%	0,0%	2,4%	25.140	84%
HHAGE2	0,0%	0,0%	0,1%	1%	81%	17%	0,8%	0,0%	0,0%	0,9%	28.961	97%
HHAGE3	0,0%	0,0%	0,1%	1%	81%	17%	0,9%	0,0%	0,0%	1,0%	28.761	96%
HHAGE4	0,0%	0,0%	0,1%	1%	87%	11%	1,1%	0,0%	0,0%	1,3%	27.033	91%
HHINC1	0,0%	0,0%	0,2%	2%	85%	11%	1,8%	0,0%	0,0%	2,0%	24.378	82%
HHINC2	0,0%	0,0%	0,1%	1%	84%	14%	1,2%	0,0%	0,0%	1,2%	25.770	86%
HHINC3	0,0%	0,0%	0,1%	1%	82%	15%	1,2%	0,0%	0,0%	1,3%	27.329	91%
HHINC4	0,0%	0,0%	0,2%	3%	88%	6%	2,8%	0,2%	0,0%	3,3%	20.543	69%
HHINC5	0,0%	0,0%	0,2%	2%	89%	7%	2,4%	0,1%	0,0%	2,7%	15.004	50%
HHINC6	0,0%	0,0%	0,2%	1%	88%	9%	1,7%	0,1%	0,1%	2,2%	10.181	34%
SEXM	0,0%	0,0%	0,1%	0%	63%	36%	0,3%	0,1%	0,0%	0,5%	29.615	99%
SEXF	0,0%	0,0%	0,1%	0%	63%	37%	0,3%	0,1%	0,0%	0,5%	29.610	99%
AGEP1	0,0%	0,0%	0,5%	4%	71%	19%	4,9%	0,8%	0,0%	6,2%	27.587	92%
AGEP2	0,0%	0,0%	0,4%	2%	67%	28%	2,3%	0,5%	0,0%	3,2%	29.033	97%
AGEP3	0,0%	0,0%	0,7%	3%	73%	20%	2,6%	0,2%	0,0%	3,5%	28.266	95%
AGEP4	0,0%	0,0%	0,3%	1%	66%	31%	0,9%	0,1%	0,0%	1,4%	29.275	98%
AGEP5	0,0%	0,0%	0,2%	1%	67%	30%	0,8%	0,1%	0,0%	1,1%	29.232	98%
AGEP6	0,0%	0,0%	0,3%	1%	68%	29%	1,0%	0,1%	0,0%	1,5%	29.158	98%
AGEP7	0,0%	0,0%	0,4%	2%	70%	26%	1,4%	0,1%	0,0%	1,9%	28.780	96%
AGEP8	0,0%	0,0%	0,4%	2%	74%	22%	1,5%	0,2%	0,0%	2,1%	28.299	95%

Para baixos valores absolutos das variáveis, os erros relativos podem ser altos, tornando esta métrica muito sensível a pequenas variações nos valores do erro absoluto. Adicionalmente, na

aplicação de um modelo de previsão de demanda por transportes, setores com valores absolutos baixos para uma dada variável tendem a ter pouco impacto no conjunto. Por este motivo, foram selecionados para a análise da distribuição dos erros, para cada variável, apenas os setores com mais de 10 observações dos totais de controle. O total de setores considerado para cada variável é indicado na coluna da anotação [2] da tabela.

Verifica-se que, para todas as variáveis, menos de 10% dos setores censitários possuem desvios maiores que 5% (coluna da anotação [1]). Observa-se que as variáveis no nível do indivíduo (SEX, AGEP) possuem menos zonas com estimação perfeita (erro 0). Apesar disso, a ordem de grandeza dos erros maiores que 5% é similar à das variáveis no nível de domicílio (HHSIZE, HHAGE, HHINC). Destaca-se ainda que, devido a uma premissa estabelecida para o modelo, há um ajuste perfeito entre o observado e o sintetizado na variável NUM_HH, que representa o número de domicílios por setor censitário.

Em relação aos valores extremos, foram observados apenas oito setores censitários com erro relativo superior a 50% em módulo em alguma variável: três observados na variável HHINC6 e os outros cinco distribuídos entre outras cinco variáveis (HHINC3, HCINC5, SEXM, SEXF, AGEP7).

3.3. Avaliação por variável para distritos/municípios

Como complementação do subitem anterior, neste subitem é feita a avaliação das 6 variáveis restantes, provenientes da pesquisa OD e controladas no nível meta, correspondente a distritos no município de São Paulo e os próprios limites municipais nos demais 38 municípios da RMSP. Os resultados são apresentados na Tabela 8.

Tabela 8: Métricas de validação por variável para distritos/municípios da RMSP.

		Para distritos/municípios (N=134)			
Atributo	**Variável**	**EMR**	**DesvPad**	**RMSE**	**Qui-quad**
Posse ou não de automóvel	HHAUT0	0,0%	0,1%	9,4	0,6
	HHAUT1	0,0%	0,1%	9,6	0,6
Grau de instrução	EDUC1	0,0%	0,5%	298,6	99,8
	EDUC2	-0,1%	0,3%	86,3	18,9
	EDUC3	-0,2%	0,4%	229,2	119,3
	EDUC4	-0,2%	1,2%	144,4	137,7

São observados valores de EMR e DesvPad (desvio padrão do erro médio) reduzidos, indicando a ausência de erros sistemáticos (EMR menor que 0,2% em módulo) e exatidão na população sintética agregada (desvio padrão abaixo de 0,5% para todas as variáveis exceto EDUC4, que alcança 1,2% de valor do desvio padrão).

A raiz do erro quadrático médio (RMSE) obtida também foi reduzida ao se considerar os valores absolutos destas variáveis. No caso da EDUC1, que obteve um valor de quase 300 para o RMSE, são verificados valores de referência de 50 mil (número médio aproximado de indivíduos com este grau de instrução nos distritos/municípios).

Alguns dos resultados obtidos para a estatística qui-quadrado, entretanto, superaram os valores críticos para 134 graus de liberdade. Ao se considerar 99% como nível de confiança, o valor crítico é 98,9 – o que indica que para 3 das 6 variáveis analisadas nesta escala (EDUC1, EDUC3 e EDUC4) a hipótese de independência entre as distribuições não pode ser descartada. Ao se considerar 95%, o valor crítico passa a ser 108,3, e esta hipótese poderia ser descartada para a variável EDUC1.

Ao explorar os resultados do qui-quadrado de cada célula (combinação entre variáveis e cada distrito/município), observou-se que dois distritos do município de São Paulo eram responsáveis por mais da metade do valor da estatística qui-quadrado das variáveis EDUC3 e EDUC4 (graus de instrução mais elevados): Iguatemi e Cidade Tiradentes, ambos na Zona Leste do município. A Tabela 9 indica a contribuição destes distritos para o valor do qui-quadrado para as variáveis EDUC3 e EDUC4, assim como os valores de seus erros.

Tabela 9: Análise dos resultados dos distritos de Iguatemi para EDUC3 e de Cidade Tiradentes para EDUC4.

Distrito	Variável	Total de Controle	Contagem do Sintetizado	Qui-quad	Erro absoluto	Erro relativo
Iguatemi	EDUC3	49.806	47.854	79,6	-1.952	-3,9%
Cid. Tiradentes	EDUC4	15.114	14.104	72,3	-1.010	-6,7%

Observa-se que, apesar de representarem mais de 50% do valor do qui-quadrado para as respectivas variáveis, ambos distritos apresentam erros absolutos de menos de 2.000 pessoas nestas categorias e erros relativos de menos de 10%. Ao se analisar os valores dos totais de controle destes distritos para o atributo de grau de instrução (EDUC 1 a 4), observou-se que ambos apresentam as porcentagens mais altas de indivíduos de "grau de instrução: fundamental incompleto" (EDUC1) do município, com 48% dos seus residentes nesta categoria. O fato destes dois distritos ocuparem a 1ª e 2ª posição dentre os 96 distritos do município pode ter influenciado a subestimação de duas das três variáveis complementares (EDUC3 e EDUC4).

A título de comparação, Moreno e Moeckel (2018) indicaram erros relativos médios da ordem de 3% para os totais por variável na escala de boros e municípios da Alemanha. Esta magnitude de erros é compatível com os dez piores distritos/municípios da presente aplicação.

3.4. Avaliação por setor censitário

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 setor censitário. Esta análise se baseia na agregação dos valores do qui-quadrado calculados para as 27 variáveis avaliadas e suas comparações com os valores críticos do qui-quadrado.

A Figura 2 exibe um histograma dos valores obtidos para o qui-quadrado nos 29.869 setores censitários da RMSP que tiveram suas populações sintetizadas. Observa-se que a grande maioria dos valores da estatística do qui-quadrado obtidos são menores do que 1.

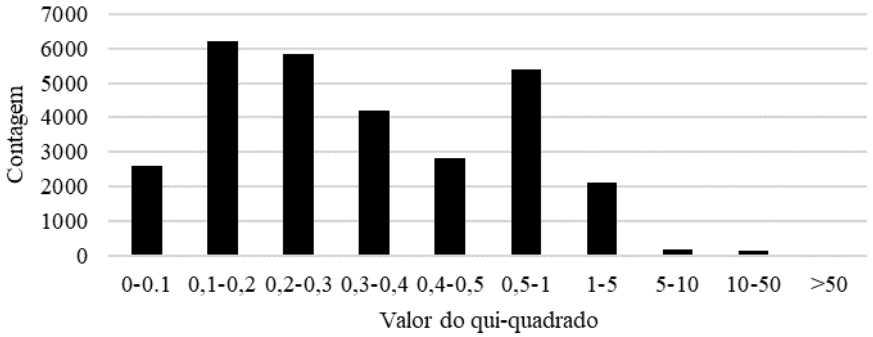

Figura 2. Distribuição dos valores do qui-quadrado para os setores censitários.

A Tabela 10 apresenta a contagem de setores censitários segundo o valor-p da estatística qui-quadrado da comparação dos valores controle x sintetizado para o conjunto de 27 variáveis. Nota-se que a grande maioria dos setores censitários (99,45%) apresentou valor-p maior do que 99,9%, o que indica uma ótima qualidade de ajuste entre as variáveis observadas (controladas) e estimadas (sintetizadas). Para apenas 103 dos 29.869 setores censitários (0,34% do total) o valor-p foi inferior a 95% e, na maior parte desses casos, o valor do qui-quadrado não seria suficiente para rejeitar a hipótese de homogeneidade das distribuições.

Tabela 10: Contagem de setores censitários segundo o valor-p da estatística qui-quadrado.

Valor-p	Valor crítico (d.f. = 27)	Contagem	Parcela
>99,9%	9,8 (99,9%)	29.705	99,45%
95-99,9%	12,9 (95%)	61	0,20%
<95%	-	103	0,34%
Total	*Total*	*29.869*	*100,00%*

3.5. Considerações sobre a validação

Consideradas as limitações para a realização da validação intrínsecas à ausência de uma *ground truth*, pode-se considerar que a população gerada é suficientemente próxima da real para fins de aplicação em modelos e análises desagregados ao apresentar (i) resultados satisfatórios para o teste de qualidade de ajuste com a estatística qui-quadrado e (ii) valores que podem ser considerados razoáveis para as demais métricas de avaliação utilizadas, ao comparar com outras aplicações de geradores de populações sintéticas.

É importante destacar que 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 ajuste de diferentes variáveis, através da alteração do parâmetro de "importância" atribuída aos atributos no PopulationSim.

4. CONSIDERAÇÕES FINAIS

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 revisão da literatura permitiu identificar, além das distintas linhas de abordagem, também 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, diversos códigos adicionais foram desenvolvidos para tornar o procedimento replicável para qualquer área no Brasil. A extensão criada no presente trabalho permite que um usuário com conhecimentos básicos de computação gere populações sintéticas para áreas de seu interesse, e que usuários com conhecimentos mínimos de programação utilizem o procedimento de geração da população sintética para atributos e fontes de dados

de sua preferência. Foram também desenvolvidos códigos para a verificação da qualidade da população sintética gerada, através da aplicação de distintos 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. Os códigos gerados neste trabalho, assim como o PopulationSim, são de código aberto e encontram-se disponíveis em um repositório *online* (Ajauskas, 2021b), assim como a população sintética gerada para a RMSP. São também disponibilizados um tutorial de instalação e uso e uma documentação dos códigos produzidos.

A aplicação dos códigos desenvolvidos e do PopulationSim para a geração da população sintética da Região Metropolitana de São Paulo utilizou dados de duas fontes distintas (Censo 2010 e Pesquisa OD 2017 de São Paulo) e 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.

As avaliações dos resultados gerados, tanto finais como preliminares, permitiu a identificação de alguns pontos de atenção quanto à qualidade dos resultados. 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 criados procedimentos de imputação com base em outras informações disponíveis. Também foram identificados possíveis erros na base de totais do Censo através do cálculo de indicadores; estes problemas foram corrigidos com a eliminação de setores que provavelmente apresentavam erros em suas bases.

Uma limitação conceitual, que afeta não apenas o PoputationSim, 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. Esta limitação, entretanto, não pode ser superada no PopulationSim por se tratar de uma característica intrínseca dos procedimentos de reconstrução sintética.

Por outro lado, identificam-se cinco frentes principais para melhoria e desenvolvimento do presente trabalho ou de aplicações futuras: (i) melhorias do software PopulationSimBR, através da criação de interface gráfica, por exemplo; (ii) a aplicação da população gerada em um modelo "final", como um ABM; (iii) a realização de análises de sensibilidade dos parâmetros do modelo (número e tamanho de categorias, valores do parâmetro "importância" etc.); (iv) o uso de variáveis cruzadas, disponibilizadas pelo Censo, como controles; e (v) a avaliação da qualidade de ajuste das variáveis não controladas pelo modelo.

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

AGRADECIMENTOS

O segundo autor agradece ao CNPq pela bolsa de Produtividade em Pesquisa.

REFERÊNCIAS

ActivitySim (2020) ActivitySim: An open platform for activity-based travel modeling. Disponível em: <https://activitysim.github.io/> (acesso em 13/06/2024).

Ajauskas, R. (2021a) *Procedimento para geração de populações sintéticas com base em dados disponíveis no Brasil.* Dissertação (mestrado). Universidade de São Paulo. São Paulo. DOI: 10.11606/D.3.2021.tde-04112021-120207.

Ajauskas, R. (2021b) *Documentação PopulationSimBR: a extensão do PopulationSim para o Brasil.* Disponível em: <https://populationsimbr.readthedocs.io/> (acesso em 13/06/2024).

Ajauskas, R. e O. Strambi (2019) Procedimentos para geração de populações sintéticas aplicada à modelagem de transportes: uma revisão dos métodos de reconstrução sintética. In *33º Congresso de Pesquisa e Ensino em Transporte da ANPET*. Balneário Camboriú: ANPET, p. 25–37.

Bar-Gera, H.; K.C. Konduri; B. Sana et al. (2009) Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board*, Washington D.C., Estados Unidos: Transportation Research Board, p. 09–1354.

Barthelemy, J. e P.L. Toint (2013) Synthetic population generation without a sample, *Transportation Science*, v. 47, n. 2, p. 266-79. DOI: 10.1287/trsc.1120.0408.

Beckman, R.; K. Baggerly e M.D. McKay (1996) Creating synthetic baseline populations, *Transportation Research Part A, Policy and Practice*, v. 30, n. 6, p. 415-29. DOI: 10.1016/0965-8564(96)00004-3.

Birkin, M. e M. Clarke (1988) SYNTHESIS—a synthetic spatial information system for urban and regional analysis: methods and examples, *Environment & Planning A*, v. 20, n. 12, p. 1645-71. DOI: 10.1068/a201645.

Farooq, B.; M. Bierlaire; R. Hurtubia et al. (2013) Simulation based population synthesis, *Transportation Research Part B: Methodological*, v. 58, p. 243-63. DOI: 10.1016/j.trb.2013.09.012.

IBGE (2010) *Censo 2010*. Disponível em: <https://censo2010.ibge.gov.br/> (acesso em 13/06/2024).

Konduri, K.; D. You; V.M. Garikapati et al. (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)*. Transportation Research Board, p. 10–14.

Lee, D. H.; Y. Fu (2011). Cross-entropy optimization model for population synthesis in activity-based microsimulation models. Transportation Research Record, v. 2255, n. 1, p. 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, Queensland, DOI: 10.14264/uql.2020.822.

MARG (2016) *PopGen: Synthetic Population Generator. Mobility Analytics Research Group*. Disponível em: <http://www.mobilityanalytics.org/popgen.html> (acesso em 13/06/2024).

Miyamoto, K.; N. Sugiki; N. Otani et al. (2010) Agent-based estimation method of household microdata for base year in land use microsimulation. In *89th TRB Meeting Compendium of Papers*. Washington D.C.: Transportation Research Board.

Moreno, A. e R. Moeckel (2018) Population synthesis handling three geographical resolutions, *ISPRS International Journal of Geo-Information*, v. 7, n. 5, p. 174. DOI: 10.3390/ijgi7050174.

Müller, K. (2017) *A generalized approach to population synthesis*. Tese (doutorado). ETH Zurich, Zurich. DOI: 10.3929/ethz-b-000171586.

Müller, K., e Axhausen, K. W. (2010) Population synthesis for microsimulation: state of the art. *Arbeitsberichte Verkehrs-und Raumplanung*, v. 638, p. 1-14.

Paul, B.M.; J. Doyle; B. Stabler et al. (2018) Multi-level population synthesis using entropy maximization-based simultaneous list balancing (No. 18-03886). In *97th Annual Meeting of the Transportation Research Board*. Washington D.C.: *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, São Carlos. DOI: 10.11606/T.18.2016.tde-24102016-154347.

PopulationSim (2020) PopulationSim 0.5.1. Disponível em: <https://activitysim.github.io/populationsim/> (acesso em 31/07/2024)

Ramadan, O.E. e V.P. Sisiopiku (2019) A critical review on population synthesis for activity-and agent-based transportation models. In *Transportation Systems Analysis and Assessment*. IntechOpen. DOI: 10.5772/intechopen.86307.

Ribeiro, R.A. (2011) *Modelo baseado em agentes para estimar a geração e a distribuição de viagens intraurbanas.* Tese (Doutorado). Universidade de São Paulo, São Carlos. DOI: 10.11606/T.18.2011.tde-31012012-081352

RSG (2017) *PopulationSim Specification*. Disponível em: <https://activitysim.github.io/populationsim/docs.html>. (acesso em 13/06/2024).

Sallard, A.; Balać, M.; Hörl, S. (2020). A synthetic population for the greater São Paulo metropolitan region. *Arbeitsberichte Verkehrs-und Raumplanung*, v. 1545.

Smith, L., R. Beckman e K. Baggerly (1995) *TRANSIMS: Transportation analysis and simulation system*. New Mexico: Los Alamos National Lab. DOI: 10.2172/88648.

Sun, L.; A. Erath e C. Ming (2018) A hierarchical mixture modeling framework for population synthesis, *Transportation Research Part B: Methodological*, v. 114, p. 199. DOI: 10.1016/j.trb.2018.06.002.

Voas, D. e P. Williamson (2000) An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata, *International Journal of Population Geography*, v. 6, n. 5, p. 349-66. DOI: 10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5.

Vovsha, P.; J.E. Hicks; B.M. Paul et al. (2015) New features of population synthesis. In *94th Annual Meeting of the Transportation Research Board*, Washington D.C., Estados Unidos: Transportation Research Board, p. 15–5343.

Ye, X.; K. Konduri; R.M. Pendyala et al. (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 D.C., Estados Unidos: Transportation Research Board.