# PROCEDIMENTOS PARA GERAÇÃO DE POPULAÇÕES SINTÉTICAS APLICADA À MODELAGEM DE TRANSPORTES: UMA REVISÃO DOS MÉTODOS DE RECONSTRUÇÃO SINTÉTICA

**Rodrigo Ajauskas**
**Orlando Strambi**
Universidade de São Paulo
Escola Politécnica

**RESUMO**
Modelos baseados em atividades (ABMs) vêm se destacando na estimativa da demanda frente a intervenções na oferta de transportes. Estes exigem dados desagregados sobre a população que, apesar de serem coletados por institutos nacionais de geografia e estatística, como no caso do Brasil, são disponibilizados apenas de forma agregada por questões de privacidade. Dessa forma, é necessário recriar esta população com suas características, em um processo conhecido na literatura por geração de populações sintéticas. Destacam-se três linhas de abordagens para esta tarefa: (i) reconstrução sintética, (ii) otimização combinatória e (iii) aprendizagem estatística. O presente artigo foca no desenvolvimento da primeira linha – reconstrução sintética –, que se destaca para a aplicação em modelagem de transportes, apresentando as suas principais evoluções ao longo das duas últimas décadas.

**ABSTRACT**
Activity-based models are increasingly seen as the preferred modelling approach to estimate travel demand in face of changes in transport supply. They require disaggregate data about the population that, in spite of being regularly collected by national and regional statistical agencies, are made publicly available only at an aggregate level, for the sake of privacy. It is thus necessary to recreate the characteristics of each agent in the population in a process known as Synthetic Population Generation. Three main general approaches are identified: (i) synthetic reconstruction, (ii) combinatorial optimization, and (iii) statistical learning. This article focus on the development of the first approach – synthetic reconstruction –, frequently applied to transportation modelling, presenting its evolution in the last two decades.



## 1. INTRODUÇÃO

Modelos baseados em atividades (ABM, do inglês "Activity-Based Models") vêm ao longo dos últimos anos se popularizando para a estimação de demanda de transportes e são considerados na literatura como as ferramentas mais adequadas para a avaliação de intervenções na oferta de transportes.

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, veículos e 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. Dessa forma, é necessário recriar esta população com suas características, em um processo conhecido na literatura como geração de populações sintéticas (em inglês, "Synthetic Population Generation").

A população sintética, que é o produto a ser gerado, é uma base de dados em que cada linha corresponde a um agente (indivíduo ou domicílio) e suas respectivas características. A ideia básica do procedimento é que esta população esteja no grupo das "melhores estimativas" da população real, respeitando as informações agregadas (Ryan *et al.*, 2009).

O processo para reconstrução da população normalmente se baseia principalmente em dados disponibilizados pelos institutos nacionais de estatísticas, nas formas de:

(i) Totais marginais (por categoria) para diversas variáveis agregadas por zonas; e
(ii) Amostras da população com informações desagregadas por indivíduo, no nível de região (agregação de zonas).

Há uma série de métodos para a realização do processo de geração da população sintética, que podem ser divididos em três classes principais: (i) reconstrução sintética, (ii) otimização combinatória e (iii) aprendizagem estatística. Os métodos de reconstrução sintética são os mais utilizados para a geração de populações sintéticas e procedimentos para superar as suas principais limitações foram desenvolvidos por diversos autores ao longo das últimas duas décadas. Nos últimos anos, a aplicação de técnicas de aprendizagem estatística tornou-se mais comum para o problema de gerar populações sintéticas.

O presente artigo apresenta uma revisão da literatura de procedimentos de geração de populações sintéticas, especialmente dos trabalhos mais relevantes na área de planejamento de transportes. Inicia com uma visão geral dos principais autores e linhas de desenvolvimento. Na sequência, as evoluções a partir do procedimento clássico de Beckman *et al.* (1996) são apresentadas, sendo identificadas as suas motivações e contribuições chave. Por fim são discutidos os trabalhos produzidos no Brasil que tratam do tema.

## 2. REVISÃO BIBIOGRÁFICA

### 2.1. Introdução

São identificadas três classes principais de procedimentos para a geração de populações sintéticas: (i) reconstrução sintética, (ii) otimização combinatória e (iii) aprendizagem estatística. Apresenta-se brevemente a seguir o trabalho seminal 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 enquadra-se na categoria de algoritmos denominados de reconstrução sintética (“Synthetic Reconstruction”, SR);
2. Voas e Williamson (2000): utilizaram a técnica “Simulated Annealing” de otimização combinatória (“Combinatorial Optimization”, CO). Nesta abordagem, assume-se que a amostra é suficientemente grande em relação à população da zona que deve ser sintetizada e que esta zona é uma parcela da população desta amostra (Harland et. al, 2012);
3. Farooq *et al.* (2013): aplicaram técnicas de aprendizagem estatística (“Statistical Learning”, SL), buscando resolver os principais problemas identificados com métodos de SR e CO. 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.

Observa-se na literatura uma predominância na utilização de métodos de reconstrução sintética, especialmente para aplicação em modelagem de transportes. Métodos baseados em otimização combinatória são mais observados em aplicações na área de geografia, por

exemplo, em questões relacionadas à saúde, saneamento e avaliação de políticas de impostos. 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. Isso posto, optou-se por focar os esforços do presente artigo nas abordagens de reconstrução sintética, que vêm sendo aprimoradas há mais de duas últimas décadas por diversos autores.

Ao longo deste documento são utilizados alguns termos com significados similares: "matriz semente", "tabela de contingência" e "distribuição conjunta" possuem significados parecidos, porém se encaixam melhor em contextos diferentes. Matrizes sementes são normalmente preparadas a partir de "microdados" ou "amostras" – que são parcelas da população total para as quais são apresentados dados desagregados. "Totais marginais" e "totais de controle" também podem ser entendidos como sinônimos. Além disso, é importante destacar que o termo IPF, que a rigor se refere a apenas uma etapa do procedimento de geração de população sintética, é utilizado por vezes ao longo do texto para se referir a todo procedimento de reconstrução sintética. Isto ocorre porque a etapa de ajustamento, em que ele é utilizado, pode ser considerada a principal do procedimento.

## 2.2. Reconstrução Sintética: o IPF de Beckman

O trabalho de Beckman *et al.* (1996) é um dos mais reconhecidos para a geração de populações sintéticas, por ter sido o primeiro a aplicar o procedimento conhecido por "Iterative Proportional Fitting" (IPF) para resolver o problema de geração 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 1. No exemplo, indivíduos são classificados segundo dois atributos: (i) status de trabalho, com três categorias: não trabalha (○), trabalha em tempo parcial (⦿) e trabalha em tempo integral (●); 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 ($\boldsymbol{\ell}_i$) e coluna ($\mathbf{c}_j$) são representados à direita e abaixo, respectivamente. Os fatores de ajuste $\boldsymbol{\ell}_i/\mathbf{n}_{i*}^{(k)}$ e $\mathbf{c}_j/\mathbf{n}_{*j}^{(k)}$, utilizados para atualizar os valores da tabela de contingência alternadamente ao longo das iterações, são indicados na última linha e coluna, respectivamente.

(a) Tabela de contingência inicializada para o IPF

| k = 0 | | *j:* idade 0-14 | 15-34 | 35-64 | 65+ | $n_{i*}^{(k)}$ | $\ell_i$ | $\ell_i / n_{i*}^{(k)}$ |
|---|---|---|---|---|---|---|---|---|
| ***i:* trabalho** | ○ | 73 | 23 | 35 | 74 | 205 | 124 | 0,60 |
| | ⊙ | 0 | 42 | 17 | 15 | 74 | 83 | 1,12 |
| | ● | 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 0-14 | 15-34 | 35-64 | 65+ | $n_{i*}^{(k)}$ | $\ell_i$ | $\ell_i / n_{i*}^{(k)}$ |
|---|---|---|---|---|---|---|---|---|
| ***i:* trabalho** | ○ | 44,2 | 13,9 | 21,2 | 44,8 | 124 | 124 | 1,00 |
| | ⊙ | 0,0 | 47,1 | 19,1 | 16,8 | 83 | 83 | 1,00 |
| | ● | 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 | | | |

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

| k = 2 | | *j:* idade 0-14 | 15-34 | 35-64 | 65+ | $n_{i*}^{(k)}$ | $\ell_i$ | $\ell_i / n_{i*}^{(k)}$ |
|---|---|---|---|---|---|---|---|---|
| ***i:* trabalho** | ○ | 88,0 | 10,9 | 15,6 | 68,0 | 182,5 | 124 | 0,68 |
| | ⊙ | 0,0 | 37,0 | 14,0 | 25,6 | 76,5 | 83 | 1,08 |
| | ● | 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 0-14 | 15-34 | 35-64 | 65+ | $n_{i*}^{(k)}$ | $\ell_i$ | $\ell_i / n_{i*}^{(k)}$ |
|---|---|---|---|---|---|---|---|---|
| ***i:* trabalho** | ○ | 88,0 | 1,8 | 2,5 | 31,7 | 124 | 124 | 1,00 |
| | ⊙ | 0,0 | 25,2 | 9,3 | 48,5 | 83 | 83 | 1,00 |
| | ● | 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 1**: Exemplo de aplicação do método IPF de Beckman *et al.* (1996)
Fonte: Müller (2017). Adaptado pelos autores.

O procedimento indicado no exemplo da Figura 1, na prática, é realizado de maneira análoga, simultaneamente para as "n" variáveis a serem utilizadas como controles.

Moreno e Moeckel (2018) destacam que, enquanto para a fase de ajustamento há uma série de

procedimentos alternativos empregados (que serão discutidos ao longo deste artigo), 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 (Ye *et al.*, 2009). Fogem desta regra os métodos baseados em otimização combinatória e alguns dos métodos de aprendizagem estatística, conforme apresentado no levantamento realizado por Moreno e Moeckel (2018).

Após o trabalho seminal de Beckman *et al.* (1996), outros geradores de populações sintéticas foram criados. Bowman (2004) elaborou 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 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. As Evoluções do IPF

O procedimento elaborado por Beckman *et al.* (1996) apresentava 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.

#### *2.4.1. 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

características não esteja representado nesta amostra, e que assim seu valor seja zero na tabela de contingência, apesar de 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, apenas 609 apresentavam valores diferentes de zero.

No procedimento IPF, as células inicializadas com valor zero jamais mudam de valor. 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 diversos valores para variáveis quantitativas), 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 2. 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.

*2.4.2. 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.

(a) Amostra da zona

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

(b) Amostra da região

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

(c) Probabilidades da zona, valor zero destacado

| | Zona n | Renda Alta | Renda 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 | Renda 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 | Renda 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 | Renda Baixa | |
|---|---|---|---|---|
| Tamanho | 1 | 0,24 | 0,06 | |
| | 2 | 0,16 | 0,31 | |
| | >3 | 0,16 | 0,08 | |
| | | | Total | 1,00 |

**Figura 2**: Abordagem de Ye *et al.* (2009) para o problema de valores nulos na amostra
Fonte: Ye *et al.* (2009). Adaptado pelos autores.

Outro caminho explorado para resolver o problema do uso de controles em mais de um nível hierárquico foi o de otimização 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. Estas técnicas não utilizam IPF, entretanto partem do mesmo principio de rebalancear a matriz semente – 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.

Outra 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 Pritchhard 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 3, adaptada a partir de Müller (2017), exemplifica o procedimento do IPU. No exemplo, **y** corresponde aos registros de domicílios com as características indicadas; na sequência o atributo de posse (⊡) ou não (□) de automóvel é indicado na forma de duas "dummies". Os domicílios também são categorizados em função do número de indivíduos que não trabalha (○) e que trabalha (●). 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 categoria **A,** $\mathbf{c_A}$, e a frequência calculada daquela iteração, $\mathbf{F_A^{(k)}}$. 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** | **Posse Auto** ⊡ | □ | **núm. dom.** | **Trabalho** ○ | ● | **num. ind.** |
|---|---|---|---|---|---|---|
| 1-22 | 0 | 1 | 22 | 2 | 1 | 66 |
| 23-43 | 0 | 1 | 21 | 1 | 1 | 42 |
| 44-64 | 0 | 1 | 21 | 3 | 0 | 63 |
| 65-80 | 1 | 0 | 16 | 2 | 0 | 32 |
| 81-96 | 1 | 0 | 16 | 2 | 1 | 48 |
| 97-108 | 1 | 0 | 12 | 1 | 0 | 12 |
| 109-119 | 0 | 1 | 11 | 2 | 0 | 22 |
| 120-128 | 0 | 1 | 9 | 1 | 0 | 9 |
| 129-136 | 1 | 0 | 8 | 1 | 2 | 24 |
| 137-144 | 0 | 1 | 8 | 1 | 2 | 24 |
| 145-151 | 1 | 0 | 7 | 1 | 1 | 14 |
| 152-158 | 1 | 0 | 7 | 3 | 0 | 21 |
| 159-164 | 0 | 1 | 6 | 0 | 1 | 6 |
| 165-170 | 0 | 1 | 6 | 0 | 2 | 12 |
| 171-173 | 1 | 0 | 3 | 0 | 1 | 3 |
| 174-175 | 0 | 1 | 2 | 0 | 3 | 6 |
| 176 | 1 | 0 | 1 | 0 | 2 | 2 |
| **Σ** | 70 | 106 | 176 | 279 | 127 | 406 |

| **k** | 0 | 1 | 2 | 3 | 4 | ∞ |
|---|---|---|---|---|---|---|
| **A** | ○ | ● | ⊡ | □ | ○ | - |
| | 1,00 | 0,74 | 1,66 | 1,66 | 1,64 | 1,60 |
| | 1,00 | 0,74 | 1,66 | 1,66 | 1,64 | 1,60 |
| | 1,00 | 0,74 | 0,74 | 0,74 | 0,74 | 0,33 |
| | 1,00 | 0,74 | 0,74 | 0,39 | 0,39 | 0,19 |
| | 1,00 | 0,74 | 1,66 | 0,86 | 0,86 | 0,95 |
| | 1,00 | 0,74 | 0,74 | 0,39 | 0,39 | 0,19 |
| | 1,00 | 0,74 | 0,74 | 0,74 | 0,74 | 0,33 |
| | 1,00 | 0,74 | 0,74 | 0,74 | 0,74 | 0,33 |
| $\mathbf{f_y^{(k)}}$ | 1,00 | 0,74 | 1,66 | 0,86 | 0,86 | 0,95 |
| | 1,00 | 0,74 | 1,66 | 1,66 | 1,64 | 1,60 |
| | 1,00 | 0,74 | 1,66 | 0,86 | 0,86 | 0,95 |
| | 1,00 | 0,74 | 0,74 | 0,39 | 0,39 | 0,19 |
| | 1,00 | 1,00 | 2,23 | 2,23 | 2,21 | 3,51 |
| | 1,00 | 1,00 | 2,23 | 2,23 | 2,21 | 3,51 |
| | 1,00 | 1,00 | 2,23 | 1,16 | 1,16 | 2,80 |
| | 1,00 | 1,00 | 2,23 | 2,23 | 2,21 | 3,51 |
| | 1,00 | 1,00 | 2,23 | 1,16 | 1,16 | 2,80 |
| $\mathbf{F_A^{(k)}}$ | 279,0 | 101,7 | 86,2 | 146,1 | 255,0 | |
| $\mathbf{c_A}$ | 207 | 227 | 45 | 145 | 207 | |
| $\mathbf{c_A / F_A^{(k)}}$ | 0,74 | 2,23 | 0,52 | 0,99 | 0,81 | |

**Figura 3**: Exemplo de aplicação do método IPU de Ye *et al.* (2009)
Fonte: Müller (2017). Adaptado pelos autores.

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 tipo de domicílio. 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 sejam arredondadas para 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 "n" tipos de domicílios cujos totais apresentam as maiores diferenças tem seus valores aumentados em 1 (sendo "n" a diferença entre os totais das populações obtida e desejada).

Müller e Axhausen (2011) propuseram o algoritmo HIPF, 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 principio 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, independente do número de pessoas de um domicílio de uma determinada categoria (conforme exemplificado na Figura 3), 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 3 – em contraste ao HIPF.

### *2.4.3. Requisitos computacionais*

Em termos computacionais, Pritchhard 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), a cada variável de controle adicionada ao modelo, uma nova dimensão é criada na tabela de contingência, que cresce de maneira exponencial com o número de atributos, exigindo um requerimento maior de memória. Na "sparse list representation" de Pritchhard 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 atributos e suas respectivas categorias podem ser utilizados.

### *2.4.4. Controles em diferentes escalas geográficas*

Conforme observado no item anterior, diversos pesquisadores desenvolveram métodos capazes de utilizar controles nos níveis de indivíduo e domicilio 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.

Vovsha *et al.* (2015) destacaram em seu artigo três novas funcionalidades do gerador de populações sintéticas utilizado no condado de Maricopa (Arizona, Estados Unidos): (i) o uso de totais de controle imperfeitos, possibilitando o uso de graus de confiabilidade distintos para cada total de controle; (ii) a discretização otimizada dos fatores de expansão fracionais, utilizando programação linear; e (iii) o uso de controles em múltiplas escalas geográficas. Em relação ao último ponto, 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.

### 2.4. Literatura Nacional

Foram identificadas poucas aplicações dos procedimentos de geração de populações sintéticas no Brasil, todas elas no âmbito acadêmico. Essas aplicações utilizaram procedimentos distintos das principais abordagens encontradas em planejamento de transportes e que, em geral, apresentam desempenho inferior quando comparados a estas.

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). A modelagem foi elaborada em VBA e utilizou dados agregados por setores censitários e microdados do Censo Demográfico de 2010 (IBGE, 2010).

O Método de Monte Carlo, popular no campo da geografia (Harland *et al.*, 2012), também é conhecido por "probabilidades condicionais", e tem sua população construída atributo a atributo. Na avaliação de Harland *et al.* (2012), que comparou as técnicas de probabilidades condicionais, "deterministic reweighting" e "simulated annealing", a abordagem de probabilidades condicionais apresentou desempenho inferior ao método de "simulated annealing" de Voas e Williamson (2000).

Ribeiro (2011) não tinha como foco a geração de populações sintéticas, entretanto tratou o assunto em sua tese, utilizando 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.

Miranda (2017) gerou uma população sintética para alimentar um modelo em MATSim através de um procedimento baseado em informações espaciais (utilizando dados abertos do "OpenStreetMap" para uso do solo) e já incorporando o plano de atividade dos indivíduos.

## 3. CONCLUSÃO E RECOMENDAÇÕES FINAIS

Conforme apresentado no capítulo anterior, diversos métodos de geração de populações sintéticas foram desenvolvidos desde o trabalho seminal de Beckman *et al.* (1996). O método apresentava limitações que vários autores buscaram superar, seja seguindo a mesma linha de abordagem (reconstrução sintética) ou abordagens alternativas (otimização combinatória e aprendizagem estatística).

Enquanto a linha de otimização combinatória não foi muito explorada para aplicações em planejamento de transportes, observou-se uma evolução nas abordagens de reconstrução sintética, baseadas em IPF – destacando-se a linha de Ye *et al.* (2009) e Konduri *et al.* (2016), com posterior aplicação de Moreno e Moeckel (2018). Também são notáveis os desenvolvimentos utilizando técnicas de aprendizagem estatística, em que se destacam Farooq *et al.* (2013) e, mais recentemente, Sun *et al.* (2018). Estes últimos, porém, ainda encontram-se restritos apenas à geração da população sintética, não tendo sido observadas aplicações posteriores em modelagem de transportes. Similarmente, procedimentos baseados em otimização combinatória são pouco aplicados na área, apesar serem utilizados há cerca de duas décadas. Sendo assim, as abordagens baseadas em IPF mostram-se como um caminho seguro e consolidado para a área, tendo sido observadas aplicações em diversos ABMs, após o desenvolvimento de 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. Visto que a geração da população sintética é um passo chave para a aplicação de ABMs – que vem se mostrando como uma alternativa promissora ao Modelo de Quatro Etapas para a estimação de demanda de transportes em áreas urbanas – considera-se que o desenvolvimento de um método amigável e adaptado ao contexto nacional seja pertinente para contribuir com o desenvolvimento da área no Brasil.

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

**REFERÊNCIAS BIBLIOGRÁFICAS**

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.*

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.

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.

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).

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

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.

MARG (2016). PopGen: Synthetic Population Generator. Mobility Analytics Research Group. Disponível em: <http://www.mobilityanalytics.org/popgen.html>. Acesso em: 9 abr. 2019.

Miranda, D. F. (2018). Metodologia de tratamento de dados para simulação de modelo baseado em atividades usando o software MATSim. Trabalho de Conclusão de Curso (Graduação). Universidade de Brasília.

Moreno, A., e Moeckel, R. (2018). Population synthesis handling three geographical resolutions. *ISPRS International Journal of Geo-Information*, *7*(5), 174.

Miyamoto, K., Sugiki, N., Otani, N., e Vichiensan, V. (2010). Agent-based estimation method of household microdata for base year in land use microsimulation. *TRB 89th Annual Meeting Compendium of Papers*.

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*.

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. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685-704.

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.

Smith, L., Beckman, R., e Baggerly, K. (1995). *TRANSIMS: Transportation analysis and simulation system*. Los Alamos National Lab., NM (United States).

Sun, L., Erath, A. e Ming C. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological. 114*.

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.

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).

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*.

Rodrigo Ajauskas (rodrigo.ajauskas@usp.br)
Orlando Strambi (ostrambi@usp.br)
Departamento de Engenharia de Transportes, Escola Politécnica, Universidade de São Paulo
Av. Professor Almeida Prado, Travessa 2, No. 83 – São Paulo, SP, Brasil