UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

GERAÇÃO DE DADOS SINTÉTICOS EM TRANSPORTES APLICANDO O MÉTODO DE AJUSTE ITERATIVO (IPF)

FELIPE GONÇALVES MARTINS

ORIENTADORA: FABIANA SERRA DE ARRUDA

MONOGRAFIA DE PROJETO FINAL EM TRANSPORTES

BRASÍLIA / DF: DEZEMBRO - 2017

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

GERAÇÃO DE DADOS SINTÉTICOS EM TRANSPORTES

APLICANDO O MÉTODO DE AJUSTE ITERATIVO (IPF)

FELIPE GONÇALVES MARTINS

MONOGRAFIA DE PROJETO FINAL SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM ENGENHARIA CIVIL.

APROVADA POR:

_____

Profª. FABIANA SERRA DE ARRUDA, Doutora (ENC - UnB)
(ORIENTADORA)

_____

Profª. MICHELLE ANDRADE, Doutora (ENC - UnB)
(EXAMINADORA INTERNA)

_____

Prof. PASTOR WILLY GONZALES TACO, Doutor (ENC - UnB)
(EXAMINADOR INTERNO)

DATA: BRASÍLIA/DF, 11 DE DEZEMBRO DE 2017.

FICHA CATALOGRÁFICA

MARTINS, FELIPE GONÇALVES
Metodologia para geração de população sintética em modelos de planejamento [Distrito Federal] 2017.
xii, 80 p., 210x297 mm (ENC/FT/UnB, Bacharel, Engenharia Civil, 2017)
Monografia de Projeto Final - Universidade de Brasília. Faculdade de Tecnologia. Departamento de Engenharia Civil e Ambiental.
1. Planejamento de Transportes 2. População sintética
3. Tratamento de dados 4. Modelos baseados em atividades
I. ENC/FT/UnB II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

MARTINS, F.G. (2017). Metodologia para geração de população sintética em modelos de planejamento. Monografia de Projeto Final, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 80 p.

CESSÃO DE DIREITOS

NOME DO AUTOR: Felipe Gonçalves Martins
TÍTULO DA MONOGRAFIA DE PROJETO FINAL: Metodologia para geração de população sintética em modelos de planejamento
GRAU / ANO: Bacharel em Engenharia Civil / 2017

_____

Felipe Gonçalves Martins
felipe.gmartins@gmail.com

AGRADECIMENTOS

À professora e orientadora Fabiana, por ser fonte de inspiração e conhecimento na área de Engenharia de Transportes e por ter me incentivado em todos os momentos da realização deste trabalho.

Aos demais professores do PPGT, incluindo as professoras Yaeko e Michelle por terem me mostrado o quanto a área de Engenharia de Transportes é importante e pode ser divertida de se trabalhar. E o professor Pastor, pelas palavras de encorajamento e contribuição na apresentação do projeto desta pesquisa.

Aos meus amigos Víctor e Vinícius por terem me auxiliado e apoiado nos momentos de elaboração deste trabalho e principalmente pelo apoio durante toda a graduação.

Aos meus pais, Cristiane e Daniel, e minha irmã Danielle por serem a minha base de tudo.

RESUMO

A utilização de populações sintéticas é uma das formas de transpor limitações relacionadas à obtenção de dados acerca de toda a população para os modelos de planejamento de transporte. São necessários dados desagregados para fornecer informações em nível detalhado sobre uma amostra da população e dados agregados para serem utilizados como controle e limite para a geração da população sintética.

No contexto Brasileiro poucos foram os estudos de geração de população sintética para aplicação em modelos de microssimulação de demanda de transporte. Na situação exposta, fez-se necessário o desenvolvimento de uma metodologia que tornasse possível a criação de uma população sintética para a cidade de São Paulo com dados fornecidos pelo Instituto Brasileiro de Geografia e Estatística e pela Companhia do Metropolitano de São Paulo.

O tratamento inicial dos dados foi feito através do método de ajuste proporcional iterativo (IPF), que gerou os dados de entrada para o código de geração da população sintética, desenvolvido em linguagem Python 3. A população sintética obtida foi superdimensionada em relação à população existente, porém sendo similar da população real em estudo.

ABSTRACT

The use of synthetic populations is one of the ways to overcome limitations related to obtaining population-wide data for transport planning models. Disaggregated data are required to provide detailed information on a population sample and aggregate data are required to be used as control for the generation of the synthetic population.

In the Brazilian context, few synthetic population generation studies for application in transport demand microsimulation models were carried out. In the exposed situation, it was necessary to develop a methodology that would make the creation of a synthetic population possible for the city of São Paulo using data provided by the Brazilian Institute of Geography and Statistics and by the São Paulo Metro Company.

The initial treatment of data was done using the iterative proportional adjustment method (IPF), which generated the input data for the synthetic population generation code, developed in Python 3 coding language. The obtained synthetic population was oversized in relation to the existing population, but being similar from the real population studied.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

1. INTRODUÇÃO

1.1. APRESENTAÇÃO DO TEMA

Os problemas relacionados ao transporte começaram a surgir no mundo todo, segundo Ortúzar e Willumsen (2004), em razão do crescimento econômico mundial, que provocou um surgimento de demandas que evoluíram aceleradamente em relação às ofertas de transporte. A estabilidade econômica e a qualidade de vida dos habitantes de uma região são dependentes de um sistema de transporte confiável e em bom funcionamento (ITE, 2016). Como resposta aos problemas causados pelo avanço das demandas, nos anos 60 o planejamento de transportes começou a ser sistematizado de forma que pudessem ser solucionadas as falhas dos modelos que se desenvolviam (BANISTER, 2002).

Os modelos atuais de demanda de transporte de forma geral requerem e usam dados oriundos de Pesquisas Origem e Destino (O/D), que fornecem informações acerca do comportamento e das escolhas individuais em relação às origens e destinos das viagens, escolha modal, motivos das viagens e rotas escolhidas, e que são relacionadas com características socioeconômicas, entre outros. A obtenção de dados desagregados de pesquisas O/D pode ser um desafio durante o processo de modelagem, pois frequentemente são levantamentos que demandam bastante recursos financeiros e de tempo, portanto possuem um maior intervalo entre sua realização. Uma das formas de compensar a falta de dados é a criação de dados sintéticos, representativos do sistema real (PIANUCCI, 2016).

De acordo com Bogle e Mehrotra (2016), dados sintéticos, quando compostos de observações que podem substituir as observações reais são uma alternativa para serem utilizados como entrada de dados para modelagem. Uma das aplicações para os métodos de geração de dados sintéticos é a geração de indivíduos sintéticos, compondo uma população sintética. Os trabalhos acerca de populações sintéticas vem sendo estudados para aplicação em modelos matemáticos desde a década de 40, quando Deming e Stephan (1940) desenvolveram o método de ajuste iterativo proporcional (IPF). Esse método é utilizado até os dias de hoje devido à sua eficácia na geração de dados sintéticos (FRICK; AXHAUSEN, 2003).

A abordagem convencional para criação de uma população sintética para aplicação em modelos de transporte, para um ano base determinado, foi desenvolvida por Beckman et al. (1996), em que dados de bases de dados desagregados, como exemplo: pesquisas domiciliares, diários de atividades, são assimilados à bases de dados agregados, como dados do censo realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

O uso da população sintética vem sendo usado em microssimulações e tem se mostrado uma alternativa para criação de micro dados, que são dados que apresentam grande nível de informações em nível individual desagregado. No Brasil ainda são poucos os trabalhos voltados à geração de população sintética aplicada a modelos de planejamento de transporte, tendo destaque o trabalho desenvolvido por Pianucci (2016). Assim, o desafio das pesquisas nacionais está em desenvolver ferramentas para geração de população sintética de forma a garantir bons resultados da modelagem a menores custos.

1.2. OBJETIVO

Este trabalho tem como objetivo principal a geração de dados sintéticos em transportes aplicando o Método de Ajuste Iterativo (IPF)

De forma complementar, este trabalho possui os seguintes objetivos específicos:

a) Desenvolvimento de código em linguagem Python para geração de dados sintéticos;
b) Validação do método desenvolvido para geração da população sintética para a cidade de São Paulo/SP.

1.3. JUSTIFICATIVA

Há uma necessidade de obtenção de micro dados para o desenvolvimento das pesquisas acerca de geração de população sintética. Esse tipo de pesquisa é subsidio para complementar o entendimento do comportamento de viagem dos usuários dos sistemas de transporte, e também para fornecer informação sobre as características socioeconômicas e suas interdependências. Este trabalho é uma das formas de diminuir a dificuldade de se obter esses dados, pois propõe uma metodologia para geração de populações sintéticas sem custo, pensando ainda nas limitações de tempo e de contribuição de terceiros.

As pesquisas O/D de forma geral fornecem informações suficientes para os estudos das populações sintéticas, porém a sua elaboração é bastante onerosa, portanto são necessários que sejam criados caminhos que possibilitem o uso dos dados de amostras da população, de forma a reduzir custos e tempo de execução sem prejudicar os resultados dos estudos de planejamento de transportes.

Os modelos de transporte necessitam constantemente de dados acerca da população e suas características, pois tratam do planejamento relacionado ao crescimento populacional e à mudança dos padrões de comportamento de viagens e dos atributos individuais. As populações sintéticas servem de alicerce para este tipo de planejamento, pois fornecem meios de tratar cada indivíduo independentemente.

Este trabalho é uma forma de complementar os resultados obtidos no trabalho desenvolvidos anteriormente por Miranda (2017), de modo a ampliar o conhecimento sobre o tema. A junção

desses trabalhos cria uma base acerca do tema para que outros discentes da Universidade de Brasília e de outras universidades brasileiras possam se aprofundar no assunto tendo acesso a referencias que tratam da realidade brasileira.

2. FUNDAMENTAÇÃO TEÓRICA

O desenvolvimento e a industrialização dos países intensificou as complicações relativas ao transporte, segundo Ortúzar e Willumsen (2004). O incremento dos fluxos nas rodovias e nas demandas de transporte resultaram, de maneira abrangente, em problemas ambientais e de tráfego, como congestionamentos e atrasos. Banister (2002) descreve como a evolução da posse de carros particulares entre as décadas de 60 a 90 trouxe uma maior mobilidade aos indivíduos, de forma que acarretou uma mudança comportamental e originou demandas de viagens antes não comuns, tendo como exemplo viagens de lazer aos fins de semana. As estimativas de demandas de transporte futuras até meados da década de 50 se baseavam em fatores de crescimento uniformes. Entretanto, o crescimento das áreas metropolitanas e da importância do planejamento de transportes estimulou o uso de métodos de previsão mais formalizados (ITE, 2016).

Os modelos de transporte são apresentados por Ortúzar e Willumsen (2004) como a representação de parte do mundo real. Os modelos não são o planejamento de transportes em si, mas uma forte ferramenta que auxilia no processo. Os autores descrevem a importância de modelos matemáticos e nos avanços que os sucederam nas áreas de tomada de decisão e planejamento.

Para Pianucci (2016) os modelos de demanda de transporte surgiram da necessidade de atender às demandas de transporte crescentes. Segundo Bilt (2002, apud PIANUCCI 2016) os modelos desagregados são necessários para a análise dos problemas derivados da demanda crescente por transporte. Um exemplo de modelos desagregados são os modelos baseados em atividades.

Os modelos baseados em atividades carecem de um elevado nível de detalhes em seus dados de entrada, principalmente ao se tratar dos dados acerca dos usuários dos sistemas de transportes, que são vistos como viajantes individuais (BECKMAN; BAGGERLY; MCKAY, 1996). Comumente, a geração de microdados que representem a população em estudo é uma das primeiras etapas da criação de modelos de microssimulacão (GUO; BHAT, 2007).

A obtenção de dados desagregados com informações demográficas individuais levanta dois obstáculos que devem ser vencidos, de forma a possibilitar o desenvolvimento de modelos para

microssimulação. O primeiro deles é que os dados dos indivíduos são geralmente restritos, pois fornecem informações que dizem respeito à privacidade individual das pessoas entrevistadas (ANDERSON, 2014). O segundo obstáculo é que a elaboração e realização das pesquisas para fornecer dados desagregados são bastante onerosas e demandam muito tempo. Portanto, são necessárias ferramentas para a obtenção desses dados, sendo uma delas a utilização de dados sintéticos (PIANUCCI, 2016). Desta forma, métodos para a geração de uma população sintética para modelos de microssimulacão são apresentados a seguir.

2.1.GERAÇÃO DA POPULAÇÃO SINTÉTICA

A geração de dados sintéticos é uma maneira de contornar as dificuldades relacionadas à obtenção e divulgação de informações de um banco de dados. Os dados sintéticos são os criados para a substituição de valores em uma base de dados ou para a adoção de valores, caso não exista informação a seu respeito, levando em consideração a probabilidade de escolha desses novos valores (GRAHAM; YOUNG; PENNY, 2008).

Pianucci (2016) define população sintética como a representação estatística de uma população real. A população sintética, assim como a população real, deve fornecer dados socioeconômicos e a localização das atividades realizadas por cada indivíduo, sendo assim, possível a sua utilização em modelos de microssimulação de demanda de transportes. O conceito da criação de uma população sintética é a combinação de dados censitários desagregados com dados censitários agregados, de maneira que, o resultado seja um grupo de indivíduos cujos atributos sejam estatisticamente similares aos obtidos com os dados desagregados e cuja quantidade seja similar aos dados agregados (MÜLLER; AXHAUSEN, 2010).

Uma característica das populações sintéticas é a preservação da confidencialidade dos indivíduos, pois ela é gerada através da integração de diversas bases de dados, de forma que são criados atributos para os indivíduos sintéticos, de maneira condizente com a realidade, de forma que a população sintética e a real sejam estatisticamente indistinguíveis (ADIGA et al., 2015).

Na literatura são encontrados alguns métodos para geração de população sintética. Pianucci (2016) apresenta três desses métodos, que podem ser utilizados individualmente ou

complementarmente:o Método de Ajuste Proporcional Iterativo (IPF - *Iterative Proportional Fitting), o* Método Monte Carlo (MMC) e o Método de Otimização Combinatória (CO), sendo os dois primeiros os mais encontrados na literatura. No item 2.2. é detalhado o processo de geração de população sintética.

2.2.MÉTODOS PARA GERAÇÃO DE POPULAÇÃO SINTÉTICA

A geração de população sintética, é dividida em duas etapas: o ajuste e a geração dos domicílios sintéticos. Para a etapa de ajuste o método, Beckman et al. (1996), em seu trabalho, combina dados agregados de um censo com amostras de dados desagregados para gerar uma população sintética pelo método de ajuste proporcional iterativo (IPF - *Iterative Proportional Fitting)*. O método IPF é consolidado na literatura como o método convencional para criação das populações sintéticas. A partir do resultado obtido com o método IPF foi desenvolvida uma metodologia para geração de domicílios sintéticos utilizando o Método Monte Carlo (MMC), como feito no trabalho de Pianucci (2016), que é utilizado na etapa de geração dos domicílios sintéticos. Ambas as etapas e os respectivos métodos são descritos neste capítulo (MA, 2011; MÜLLER; AXHAUSEN, 2011; PIANUCCI, 2016).

Para melhor entendimento dos processos para sintetizar dados de uma população devem ser conhecidos conceitos básicos de análise de dados categóricos, caso dos dados utilizados como base nos estudos de população. A classificação de indivíduos de uma população pode ser feita de diversas maneiras, como por exemplo classificação quanto ao gênero, idade, renda, etc. As categorias devem possuir duas características: elas devem ser exaustivas, de forma que é possível classificar todos os indivíduos da população e deve ser mutuamente exclusiva, de maneira que cada indivíduo se encaixa unicamente em uma categoria (EVERITT, 1977).

2.2.1. TABELAS DE CONTINGÊNCIA

Uma das formas de representar a relação entre variáveis categóricas é por meio das tabelas de contingência. Sendo duas variáveis X e Y com I e J categorias respectivamente, são possíveis $I \times J$ combinações de classes para um indivíduo. As tabelas de contingência são tabelas retangulares com I linhas para representar a variável X e J colunas para representar a variável

Y, em que aa células da tabela representam a frequência de ocorrência conjunta das $I \times J$ combinações (AGRESTI, 2002).

Para as tabelas de contingência são adotadas as seguintes formulações. Adota-se π_{ij} como a probabilidade de ocorrência conjunta das variáveis X e Y na linha i e coluna j. As distribuições marginais são as somas de cada linha e cada coluna, resultando na soma das distribuições conjuntas. A notação adotada para a soma da linha e a soma da coluna é respectivamente π_{i+} e π_{+j}. Uma representação geral de uma tabela de contingência, com as notações descritas é apresentada na Figura 2-1 abaixo.

$$\pi_{i+} = \sum_j \pi_{ij} \quad \text{(Eq. 1)}$$

$$\pi_{+j} = \sum_i \pi_{ij} \quad \text{(Eq. 2)}$$

		Y				Total
		1	2	...	J	
X	1	π_{11}	π_{12}	...	π_{1J}	π_{1+}
	2	π_{21}	π_{22}	...	π_{2J}	π_{2+}
	3	π_{31}	π_{32}	...	π_{3J}	π_{3+}
	...	...	...	...	...	...
	I	π_{I1}	π_{I2}	...	π_{IJ}	π_{I+}
Total		π_{+1}	π_{+2}	...	π_{+J}	

Figura 2-1: Tabela de Contingência I x J

2.2.2. ETAPA DE AJUSTE

De acordo com Müller e Axhausen (2010) a etapa de ajuste consiste em adequar amostras desagregadas de acordo com restrições dos dados agregados, para geração de indivíduos sintéticos. Para a sintetização de indivíduos os dados desagregados (amostra) fornecem informações demográficas de parte da população e as variáveis de controle são escolhidas de acordo com a amostra. Então, para cada atributo demográfico a proporção de indivíduos é obtida, assim como as suas distribuições de probabilidade conjunta.

2.2.2.1. MÉTODO IPF

O método IPF foi desenvolvido por Deming e Stephan (1940) para ajustar dados, quanto à sua consistência, através da utilização de dados obtidos de outras fontes ou deduções embasadas em teorias. Os autores exemplificam a aplicação do método IPF com o problema encontrado com os censos populacionais realizados nos países, em que após a sua elaboração são conhecidos os valores totais de cada característica populacional, porém são desconhecidas as distribuições de probabilidade conjunta dessas mesmas características.

Beckman et al. (1996) foram os precursores da utilização do método de ajuste proporcional iterativo aplicado às necessidades dos modelos de microssimulação de viagens baseados em atividades. O método IPF é bastante difundido na literatura como método eficaz para aplicação em modelos de microssimulação de transportes (ARENTZE et al., 2014; FRICK; AXHAUSEN, 2003, 2004; GUO; BHAT, 2007).

No método IPF os dados agregados obtidos são chamados de variáveis de controle, pois fornecem informações sociodemográficas e a sua distribuição na população, e sendo sua unidade espacial de coleta de dados considerada grande para as microssimulações de transportes é necessário a coleta de dados desagregados. Os dados desagregados fornecem o nível de detalhe de informações necessário para as microssimulações, porém de forma geral o número de amostras em relação às bases de dados agregados é bastante reduzido, devido o alto custo de coleta de dados desagregados (GUO; BHAT, 2007).

A Figura 2-2 demonstra os dois tipos de dados utilizados no método IPF, em que a matriz do universo representa a base de dados agregados, onde são conhecidos os valores marginais totais de cada característica demográfica (N_i e N_j) e a matriz da amostra representa a base de dados desagregados onde são conhecidos os valores das frequências de cada célula (n_{ij}). Sendo assim o objetivo da utilização do método é obter os valores das frequências de cada célula (N_{ij}) para a matriz do universo (DEMING; STEPHAN, 1940).

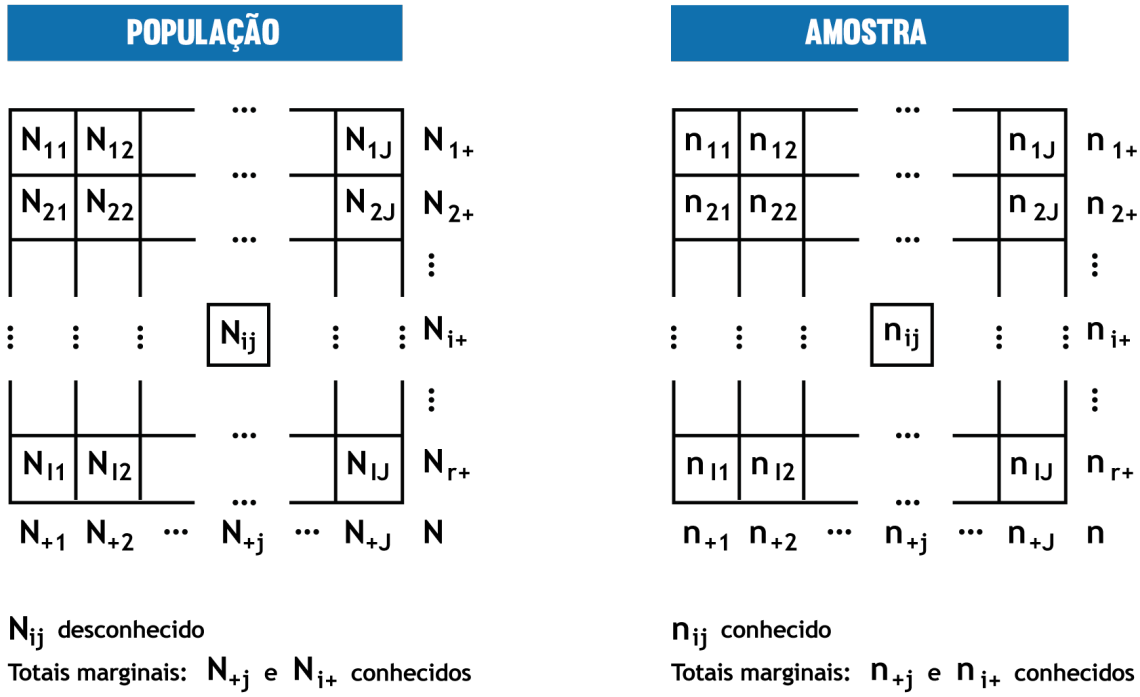

Figura 2-2: Tabelas de contingência da população e da amostra (BECKMAN, 1996)

O processo de sintetização de indivíduos é realizado ao repetir o procedimento descrito a seguir para cada área de estudo. São estimadas as distribuições de probabilidade conjunta para as variáveis de controle, de forma que os totais marginais estejam de acordo com a base de dados agregados para a área, sendo preservada a correlação entre as variáveis obtidas pela base de dados desagregados (GUO; BHAT, 2007).

No trabalho de Guo e Bhat (2007) o método é formulado e desenvolvido da forma descrita a seguir. O processo é iniciado ao igualar a probabilidade de cada célula da matriz (p_{ij}) à proporção das observações da amostra (π_{ij}).

$$p_{ij}^{(0)} = \pi_{ij} \quad \text{(Eq. 3)}$$

$$\pi_{ij} = n_{ij}/n \quad \text{(Eq. 4)}$$

Sendo $p_{ij}^{(t)}$ a proporção estimada nas células da matriz na interação t do método IPF, Fienberg (1970) em seu estudo define as iterações do IPF para ajuste de cada célula do IPF da seguinte maneira:

$$p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \frac{p_{i+}}{p_{i+}^{(2t-1)}} \quad \text{(Eq. 5)}$$

$$p_{ij}^{(2t+1)} = p_{ij}^{(2t)} \frac{p_{+j}}{p_{+j}^{(2t)}} \quad \text{(Eq. 6)}$$

O processo é repetido até que a mudança relativa dos valores para as distribuições seja considerada mínima. Beckman et al. (1996) estima que são necessárias entre 10 e 20 iterações para considerar os resultados satisfatórios. Todo o processo iterativo apresentado acima é ilustrado na Figura 2-3.

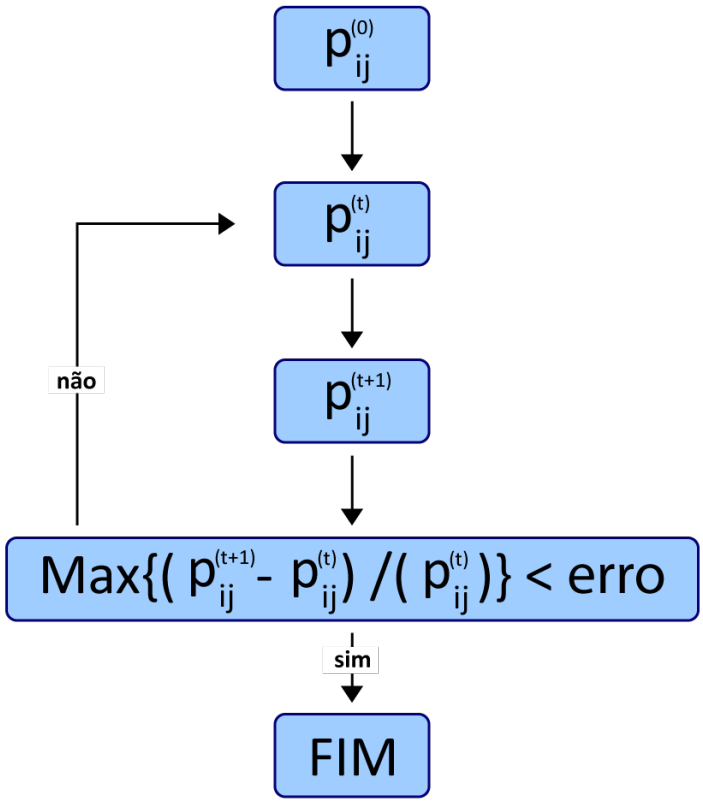

Figura 2-3: Processo iterativo do método IPF (MÜLLER; AXHAUSEN, 2003)

Problema das células com valores nulos

O método apresenta limitações, especialmente em áreas de estudo pequenas ou com amostras bastante reduzidas. Nessas situações possibilita que algumas das células das tabelas de contingência possuam valor nulo, sendo necessário que sua substituição por valores arbitrários desprezíveis, diferentes de (FRICK; AXHAUSEN, 2003). Porém, a adoção de valores pode influenciar nas probabilidades e possui pouca fundamentação teórica, fazendo-se necessária a adoção de medidas que minimizem as tendências de influencia no valor resultante (MOECKEL et. al, 2003; MÜLLER, AXHAUSEN, 2010).

2.2.3. GERACÃO DOS DOMICÍLIOS SINTÉTICOS

Segundo Müller e Axhausen (2010) a etapa de ajuste gera uma representação agregada da população em estudo. Para que os indivíduos sejam desagregados são seguidos três passos: a distribuição conjunta é arredondada para números inteiros, os domicílios são escolhidos da base

de dados da amostra para alocação, de acordo com as novas distribuições e algumas vezes a localização geográfica dos domicílios é refinada.

Na etapa de geração de domicílios sintéticos são associados a cada domicílio, indivíduos com características demográficas, seguindo as distribuições calculadas na etapa de ajuste. Para tal, são realizadas escolhas probabilísticas de características para cada indivíduo, assim como a composição dos domicílios. Uma das formas de geração dos domicílios sintéticos é através do Método Monte Carlo, descrito a seguir (MÜLLER; AXHAUSEN, 2010)

2.2.3.1. MÉTODO MONTE CARLO

Segundo Moeckel et al. (2003) o Método Monte Carlo (MMC) é um método bastante utilizado nos estudos de população sintética, como forma de complementar os resultados gerados para as tabelas de contingência, geralmente obtidas pelo método IPF. Esse método permite a geração de dados sintéticos a partir de registros populacionais. O MMC estabelece uma ordem de escolha de características de um indivíduo de modo que sejam levadas em consideração as relações de dependência entre elas.

O Método Monte Carlo é utilizado também para geração de uma população sintética com atenção para que os dados sejam coerentes com a realidade. Um exemplo disto, está representado na Figura 2-4, em que é evitada a ocorrência de um domicílio sintético composto apenas por crianças, desta forma os domicílios sintéticos já gerados são modificados e adequados até que todos possuam resultados plausíveis.

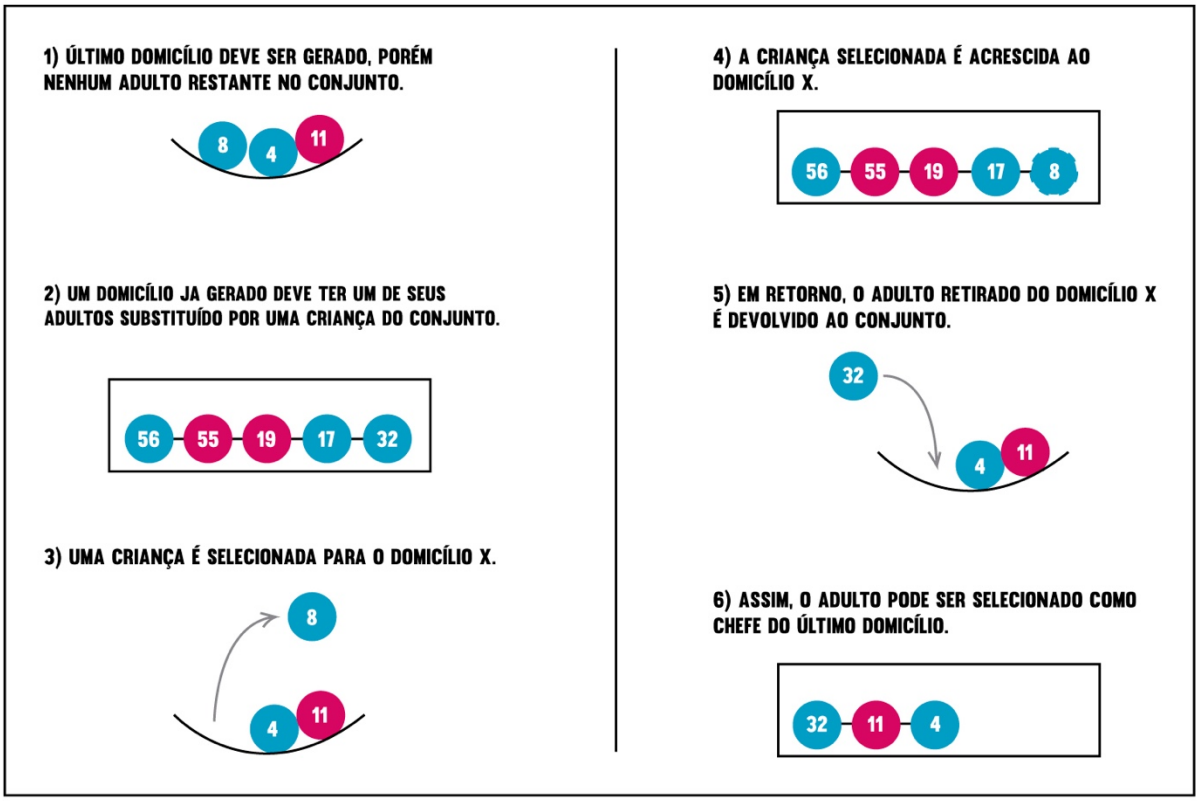

Figura 2-4: Exemplo de aplicação do Método Monte Carlo (MOECKEL et al., 2003)

2.3. APLICAÇÃO DOS MÉTODOS DE GERAÇÃO DE POPULACÃO SINTÉTICA

Beckman et al. (1996) foram os precursores dos estudos de geração de população sintética e aplicaram o método IPF de Deming e Stephan (1940) para criação de população sintética para o condado de Tarrant, no estado do Texas, Estados Unidos, com dados da Amostra de Uso Público do censo americano. O trabalho apresenta os procedimentos para ajustar a distribuição de probabilidade conjunta do número de veículos em um domicílio com o número de moradores, de acordo com os totais marginais obtidos no censo. Os autores obtiveram uma população sintética de 5628 indivíduos, para uma região cuja população real era de 5592 indivíduos.

Moeckel et al. (2003) apresentam em seu estudo a utilização dos métodos IPF e Monte Carlo, para criar uma população sintética para a cidade de Netanya em Israel, incluindo a utilização das ferramentas de GIS para gerar dados espaciais desagregados. Os autores tornaram fixos o número de habitantes da cidade e o número de domicílios, para que a população sintética fosse idêntica em quantidade à população existente. Embora não realizado nenhum tipo de validação

da população sintética com dados reais, os autores afirmam que a utilização do Método Monte Carlo é suficiente para considerar os resultados satisfatórios.

Frick e Axhausen (2003) utilizam o método IPF para ajuste da tabela de contingência com as variáveis de gênero e idade, a partir de dados do censo, microcenso e das amostras de uso público da Suíça, aplicados para a cidade de Zurique. Após a aplicação do método, os autores obtiveram uma tabela em que a soma das colunas e das linhas resultava nos valores totais fornecidos pelo censo.

Müller e Axhausen (2010) apresentam um trabalho com diretrizes e conhecimentos sobre a geração de população sintética adquiridos ao longo dos anos e com base nos estudos realizados por outros pesquisadores. Os autores descrevem os métodos IPF e Monte Carlo como eficazes para aplicação em modelos de microssimulação de transporte, como o *MATSim*.

No Brasil, Pianucci (2016) dá início à aplicação das técnicas já utilizadas internacionalmente para gerar uma população sintética para a cidade de São Carlos/SP, a partir dos dados fornecidos online pelo IBGE e dos dados coletados na Pesquisa O/D realizada na cidade no ano de 2009. A autora usa o Método Monte Carlo como base para o estudo e após obter a população simula as demandas de transporte da cidade com auxilio das Redes Neurais Artificiais (RNA). A quantidade de domicílios e habitantes foi fixada com base no censo, sendo 68883 domicílios sintéticos e 212263 indivíduos. A distribuição das variáveis de idade e número de moradores por domicílio se mostraram indiferenciáveis dos dados reais, havendo pequena discrepância apenas em relação à variável de situação domiciliar.

3. MÉTODO E APLICAÇÃO

No capítulo anterior foi apresentado o referencial teórico que serviu de base para subsidiar o método adotado nesta pesquisa. Neste capítulo serão descritos os procedimentos adotados, as ferramentas utilizadas e as fontes de dados consultadas para obtenção dos dados. Foram utilizados os *softwares*: Microsoft Office Excel 2016 e OpenOffice, em conjunto com programas de cálculo desenvolvidos em linguagem Python, para processar os dados obtidos do Censo Demográfico do IBGE e da Pesquisa Origem e Destino da Companhia do Metropolitano de São Paulo.

O processo adotado neste trabalho foi dividido em etapas sequenciais, conforme ilustrado na Figura 3-1. Cada uma das etapas é detalhada nos próximos tópicos deste capítulo.

Figura 3-1: Esquema sequencial do método proposto

3.1. OBTENÇÃO E TRATAMENTO DE DADOS

A primeira etapa da geração de população sintética é a obtenção dos dados. Em princípio são necessários dados agregados, geralmente de censos demográficos, e dados desagregados

obtidos de pesquisas por amostragem. A geração da população sintética requer dados agregados que fornecem características socioeconômicas de uma população. Esses dados são agregados geralmente em nível municipal e algumas vezes por bairros de um município.

A cidade de São Paulo, por ser a maior cidade brasileira, dispõe de uma grande quantidade de informação sobre sua população, fator decisivo para a sua escolha como objeto de estudo desta pesquisa. Tendo em vista as pesquisas existentes no Brasil e em especial em São Paulo, optou-se por usar a Pesquisa O/D da cidade por se tratar de uma base de dados de pesquisa realizada em 2007 com informações suficientes para este trabalho e o seu banco de dados é público. A Figura 3-2 esquematiza as fontes de dados e os dados coletados.

Para geração de população sintética para a cidade de São Paulo/SP foram utilizados dados de duas fontes principais. Os dados agregados, com atributos referentes à toda população do município foram obtidos junto ao Instituto Brasileiro de Geografia e Estatística (IBGE), em sua base de dados disponibilizada em sua página na internet. Os dados desagregados foram obtidos na página da internet da Companhia do Metropolitano de São Paulo, que realizou um levantamento de dados acerca de uma amostra da população da região metropolitana de São Paulo.

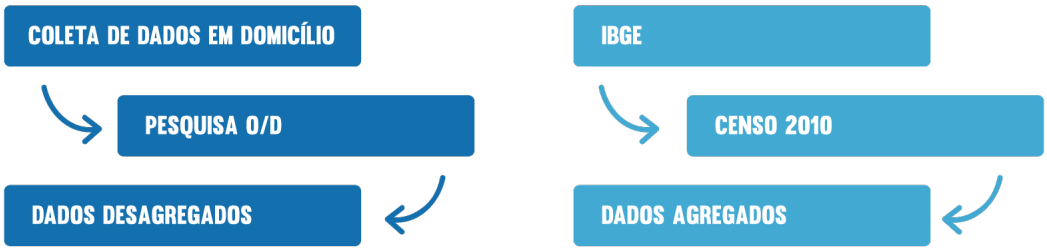

Figura 3-2: Fontes de dados demográficos

É importante salientar que para que seja possível sintetizar a população de uma região, os diferentes bancos de dados devem apresentar variáveis socioeconômica semelhantes, mesmo que apresentadas e classificadas de forma distinta, para que seja possível a aplicação dos métodos, principalmente o IPF. As variáveis consideradas importantes são as que fornecem informações acerca de idade, gênero, renda e padrões de viagem.

Diante o exposto, é importante que os dados sejam preliminarmente visualizados e estudados de forma o entendimento necessário torne possível uma reclassificação das variáveis posteriormente, na etapa de elaboração das tabelas de contingência.

3.1.1. DADOS DESAGREGADOS

O processo de geração de população sintética requer dados desagregados que forneçam informação referente às características individuais e domiciliares de uma amostra da população. Esses dados foram então analisados para que possa ser verificada a ocorrência das mesmas variáveis que ocorrem na base de dados agregada

A cada 10 anos é realizada uma pesquisa de origem e destino na região metropolitana de São Paulo pela Companhia do Metropolitano da cidade, onde são coletadas informações de uma amostra da população, incluindo um diário de atividades dos entrevistados. A primeira delas foi realizada no ano de 1967 e vem sendo realizada desde então. Para esta pesquisa, são utilizados os dados referentes à Pesquisa O/D do ano de 2007, na qual foram visitados e entrevistados 105293 indivíduos em 30 000 domicílios. Esses dados são disponibilizados *on-line* e são abertos ao público.

É importante que os dados da amostra forneçam informações socioeconômicas dos entrevistados e que seja possível o seu tratamento de forma que se possa visualizar a relação entre as suas variáveis e as suas probabilidades de ocorrência conjunta. A base de dados da Pesquisa O/D de São Paulo é fornecida na forma de tabela onde cada linha representa um indivíduo entrevistado, associado a um domicílio e cada coluna representa uma variável.

Para trabalhar com os dados da Pesquisa O/D é necessário que os arquivos estejam em formato compatível para utilização no Excel 2016, com as extensões .xls ou .xlsx. Os dados fornecidos pela Companhia do Metropolitano de São Paulo para o ano de 2007 se encontram em formato dBASE (.dbf), que é um formato próprio para arquivos de bancos de dados, utilizados em diversos *softwares* de gerenciamento de banco de dados. Portanto, foi necessária conversão do arquivo para um formato compatível com o Excel 2016. Para conversão do formato do arquivo fui utilizado o software OpenOffice que possui suporte para os arquivos .dbf e então a base de dados foi convertida para .xlsx.

Posteriormente, de forma a ter um melhor conhecimento dos dados cabe realizar uma visualização das informações, através da análise dos documentos e relatórios emitidos pela companhia, pela montagem de gráfico de distribuição e ocorrência de dados e pelo cálculo de parâmetros estatísticos como: média, desvio padrão e percentis.

A região metropolitana de São Paulo foi dividida para se obter os dados de forma mais desagregada possível. Primeiramente, os municípios foram agrupados em sub-regiões. Após a divisão em sub-regiões a região metropolitana foi dividida em 460 zonas de pesquisa, sendo 320 dessas dentro do município de São Paulo (em azul), conforme ilustrado na Figura 3-3. Por critérios de simplificação neste estudo serão utilizados apenas os dados referentes ao município de São Paulo, que se encontra na Sub-região Centro nas zonas 1 a 320.

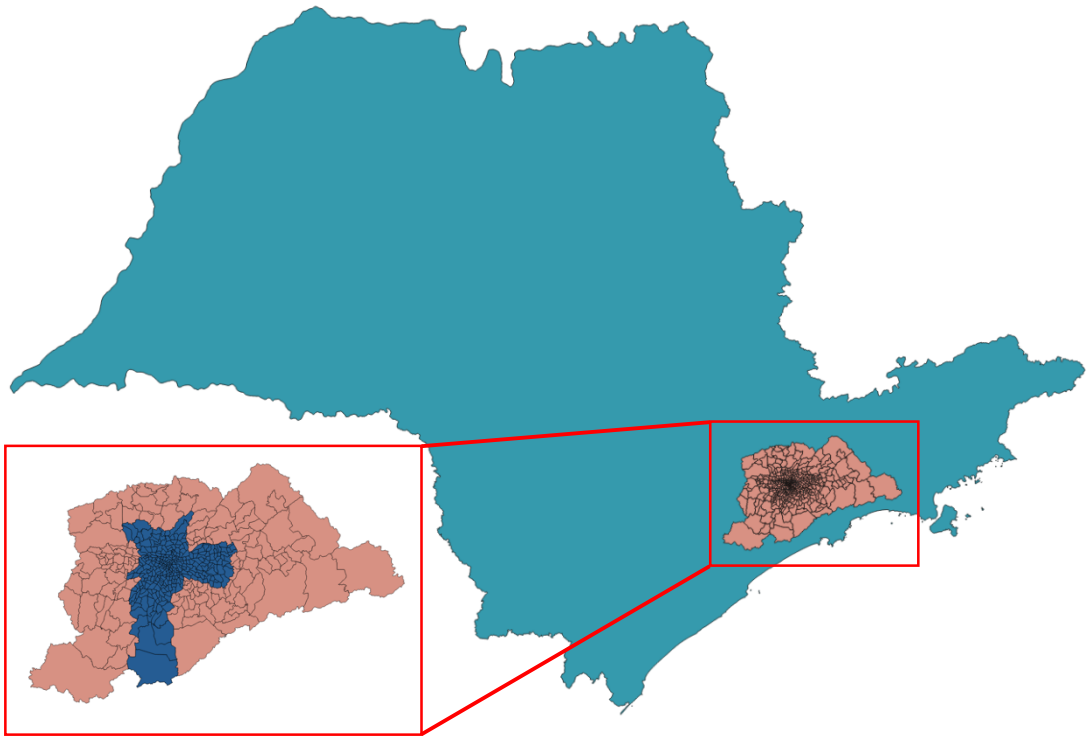

Figura 3-3: Zonas de Coleta de Dados da Pesquisa O/D no estado de São Paulo

A Pesquisa O/D apresenta dados acerca dos domicílios e de todos os indivíduos que neles residem. Foram coletadas informações sociodemográficas a nível individual (gênero, idade, escolaridade, ocupação, etc...) e a nível domiciliar (renda total do domicílio, quantos indivíduos o habitam, propriedade de automóvel, etc...),. Também foram aplicados diários de atividade

onde foram registradas as viagens realizadas pelos indivíduos do domicílio, com informações sobre a localização das atividades realizadas, modo de transporte utilizado, horário de realização da viagem e outras informações. No Apêndice A se encontram as variáveis que compõe a Pesquisa O/D da região metropolitana de São Paulo.

As variáveis da base de dados da Pesquisa O/D podem ser divididas em duas categorias: variáveis numéricas (ou quantitativas) e variáveis discretas (ou qualitativas), de maneira que há uma diferenciação na forma de visualização dos dados para cada um dos tipos.

Variáveis numéricas

As variáveis quantitativas foram analisadas por meio dos parâmetros estatísticos: média aritmética, desvio padrão e de quartis, com o auxilio do *software* Microsoft Office Excel 2016. A Tabela 3-1 apresenta as variáveis quantitativas e os parâmetros estatísticos calculados.

Tabela 3-1: Parâmetros estatísticos das variáveis quantitativas

	Média	Desvio	Min	Max	Q25	Q50	Q75
Número de moradores (docmicílio)	3,86	1,85	1	24	3	4	5
Total de famílias	1,07	0,38	1	10	1	1	1
Número de moradores (família)	3,67	1,53	1	14	3	4	4
Renda da família	3729,45	3744,53	0	46000	1362,4	2500	4712,77
Idade	36,22	19,91	1	99	21	34	50
Renda Individual	759,27	1769,19	0	40000	0	0	800

Conforme os dados da Tabela 3-1, os domicílios podem ter um máximo de 24 moradores em um único domicílio, assim como pode também ser habitado por até 10 famílias, tendo estas uma média de 3,67 membros. A idade média dos entrevistados é de aproximadamente 36 anos, com valores mínimo e máximo de 1 e 99 anos respectivamente. Percebe-se que 50% ou mais dos entrevistados não possuíam renda individual ou não responderam à pergunta da pesquisa.

Variáveis discretas

As variáveis qualitativas foram preparadas para visualização de forma distinta. Foram calculadas para algumas delas a distribuição das suas categorias em relação ao número total de observações, conforme apresentado abaixo nas figuras. Da mesma forma, o *software* utilizado foi o Excel 2016, para elaboração de gráficos do tipo pizza e histogramas.

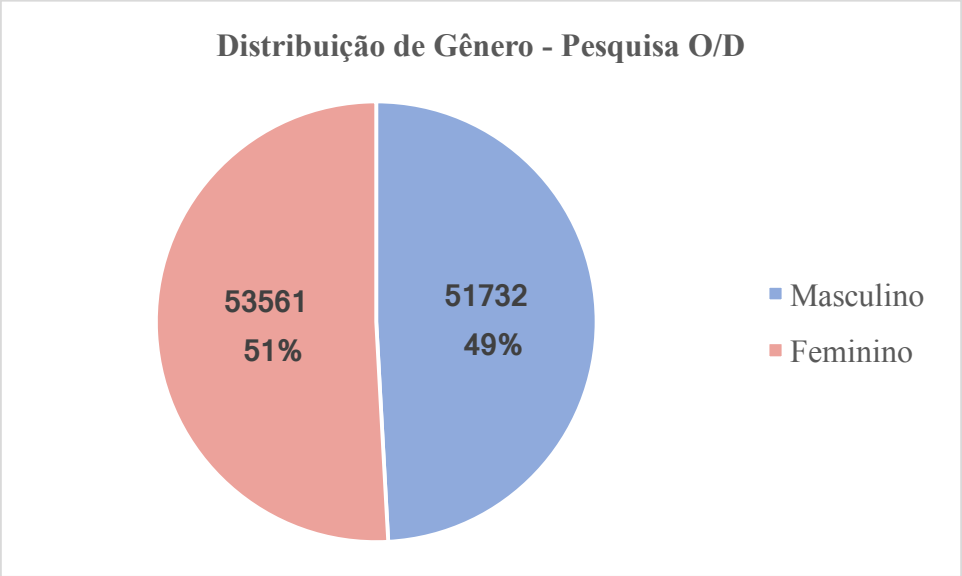

Figura 3-4: Gráfico da distribuição de gêneros na amostra da Pesquisa O/D

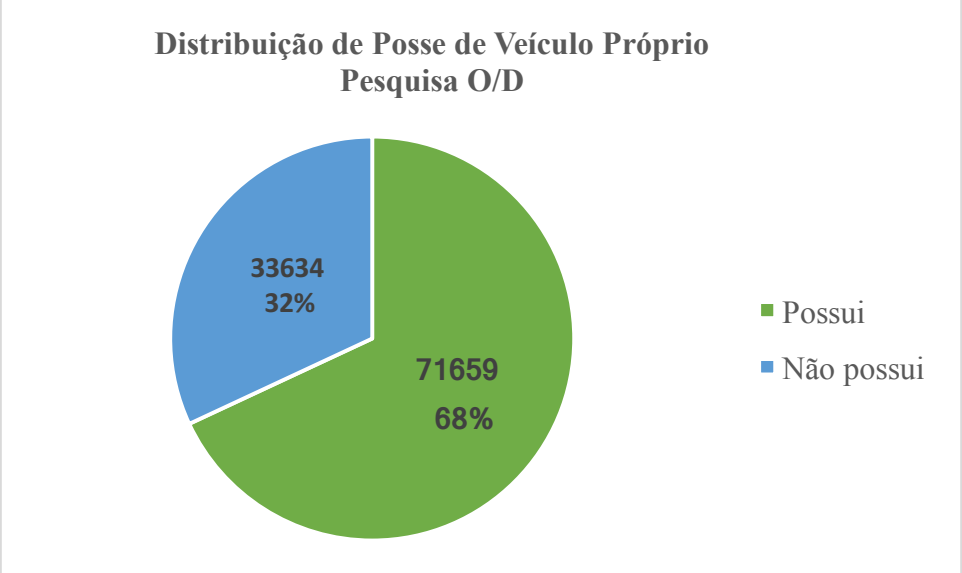

Figura 3-5: Gráfico da distribuição da posse de veículo próprio na amostra da Pesquisa O/D

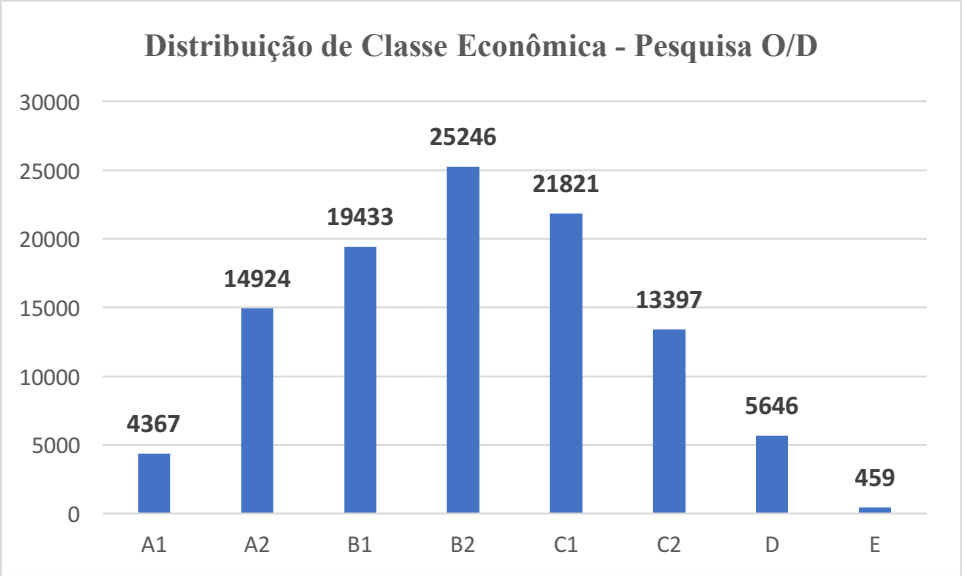

Figura 3-6: Gráfico da distribuição de classes econômicas na amostra da Pesquisa O/D

Percebe-se da Figura 3-4 que o número de homens e mulheres na amostra é relativamente bem distribuído, de forma que os dados não possuem uma tendência de respostas devido ao gênero dos respondentes.

Da Figura 3-5 é possível observar que 68% dos domicílios possuíam veículos de transporte individual como automóvel e motocicleta.

No histograma da Figura 3-6 é mostrada distribuição das classes financeiras dos domicílios da amostra, sendo a classe B2 a que ocorre com maior frequência, enquanto a classe E ocorre com menor frequência.

3.1.2. DADOS AGREGADOS

Os dados agregados são necessários para fornecer os valores totais das observações encontradas na população. As informações foram obtidas na base de dados do Censo Demográfico de 2010 e está disponível na página do IBGE. As variáveis selecionadas devem estar em consonância com as variáveis da base de dados desagregada, uma vez que os dados da população serão o parâmetro para ajuste dos dados da amostra. O censo contém informações diversas e não são todas necessárias para a criação de população sintética para modelos de demanda.

As variáveis semelhantes à disponíveis na Pesquisa O/D foram selecionadas do referido censo e estão detalhadas no Apêndice B e um tabela resumo é apresentada na Tabela 3-2.

Tabela 3-2: Variáveis socioeconômicas do Censo Demográfico de 2010

Tabela	Variável	Tipo
4.20.1.1	Gênero	Numérica
4.20.1.2	Grupos de idade	Numérica
4.20.5.2	Condição de ocupação do domicílio	Numérica
4.20.7.1	Classes de rendimento	Numérica
4.20.7.2	Classes de rendimento (domicílio)	Numérica
2.20.5.2	População que frequenta escola ou creche	Numérica
2.20.6.1	Condição de atividade	Numérica
2.20.7.3	Tempo de deslocamento até o trabalho	Numérica

Os dados fornecidos pelo IBGE já se encontram em formato .xls, que é a extensão nativa do Microsoft Office Excel 2016, portanto não é necessário qualquer tratamento de dados em relação ao formato dos arquivos.

Como apontado por Pianucci (2016), o Censo Demográfico não apresenta variáveis relacionadas às viagens dos indivíduos, exceto o tempo de deslocamento até o trabalho, portanto, a estimativa das demais variáveis, como modo de transporte utilizado e atividades realizadas para a população sintética ainda se baseia apenas nos dados da amostra.

Os dados coletados no banco de dados do censo demográfico de 2010 do IBGE se encontram com valores marginais totais, dessa maneira não é possível a aplicação de parâmetros estatísticos para análise das variáveis. Assim, como feito para os dados qualitativos da amostra da Pesquisa O/D, as distribuições de ocorrência dos dados agregados podem ser visualizadas com o auxílio de gráficos do tipo pizza e histogramas, assim como tabelas que apresentem o número total de ocorrências por classe e a sua probabilidade de ocorrência em relação ao universo.

3.1.3. TABELAS DE CONTINGÊNCIA

Após a visualização e tratamento de dados o processo realizado foi a elaboração das tabelas de contingência para encontrar a frequência de ocorrência simultânea das classes de duas ou mais variáveis. O processo é iniciado pela reclassificação dos dados em ambas as bases de dados para que seja possível elaborar as tabelas de contingência.

3.1.3.1. DEFINIÇÃO DAS VARIÁVEIS

A definição das variáveis para elaboração das tabelas de contingência foi baseada nos estudos de Saadi et al. (2016), Pianucci (2016) e Frick e Axhausen (2003). Ambos os bancos de dados devem conter as mesmas variáveis, com classes semelhantes para que seja possível a aplicação do método IPF posteriormente.

Saadi et al. (2016) em seu trabalho, analisou algumas combinações de variáveis, que foram levadas em consideração para definição neste trabalho. Nesta pesquisa foram escolhidas as seguintes variáveis para comparação, por se tratarem de variáveis existentes nas duas bases de dados. As demais variáveis foram analisadas de acordo com sua frequência de ocorrência, devido à sua ausência em um dos dois bancos de dados.

Variáveis a nível individual

- Idade x Gênero
- Idade x Renda
- Condição de atividade x Gênero

3.1.3.2.ELABORAÇÃO DAS TABELAS DE CONTINGÊNCIA

Todas as tabelas mostradas a seguir foram elaboradas da seguinte forma: os valores referentes às células n_{ij} foram obtidos da Pesquisa O/D e os valores totais marginais foram obtidos do Censo de 2010, portanto as tabelas não são matematicamente fieis, pois ainda precisam passar pela etapa de ajuste.

Para preenchimento das células n_{ij} das matrizes de contingência foi utilizado o *software* Microsoft Office Excel 2016, com o auxílio da função "=cont.ses( )", que contabiliza o número de ocorrências de acordo com critérios lógicos estabelecidos. Na fórmula abaixo é descrita uma fórmula genérica utilizada. Considerando as variáveis X e Y e suas respectivas classes X_1, X_2, ... X_I e Y_1, Y_2, ... Y_J

$$= cont.ses(Intervalo\ de\ X;\ X_i; Intervalo\ de\ Y;\ Y_j)$$

A fórmula acima retorna o valor de ocorrências em que, dentro do total de ocorrências da amostra, a variável X pertence à classe X_i e a variável Y pertence à classe Y_j, simultaneamente, preenchendo assim a tabela.

As variáveis da amostra consideradas para as tabelas são as mesmas encontradas no censo, porém encontram-se como variáveis contínuas enquanto no censo se encontram de forma discreta classificada. Dessa forma foram adotadas as classificações para as variáveis, conforme Tabela 3-3.

Tabela 3-3: Classificação de variáveis compatibilizada entre bancos de dados

Variável	Classes
Gênero	Masculino
	Feminino
Grupos de idade	0 a 4 anos
	5 a 9 anos
	10 a 14 anos
	15 a 19 anos
	20 a 24 anos
	25 a 29 anos
	30 a 34 anos
	35 a 39 anos
	40 a 44 anos
	45 a 49 anos
	50 a 54 anos
	55 a 59 anos
	60 a 64 anos
	65 a 69 anos
	70 a 74 anos
	75 a 79 anos

	80 a 84 anos
	85 a 89 anos
	90 a 94 anos
	95 a 99 anos
	Mais de 100 anos
Classes de rendimento	Até ½ salário mínimo
	Mais de 1/2 a 1 salário mínimo
	Mais de 1 a 2 salários mínimos
	Mais de 2 a 5 salários mínimos
	Mais de 5 a 10 salários mínimos
	Mais de 10 a 20 salários mínimos
	Mais de 20 salários mínimos
	Sem rendimento (2)
Condição de atividade	Ocupadas
	Desocupadas
	Não economicamente ativas
Tempo de deslocamento até o trabalho	Até 5 minutos
	De 6 minutos até meia hora
	Mais de meia hora até uma hora
	Mais de uma hora até duas horas
	Mais de duas horas

As classes adotadas são compatíveis com o banco de dados do Censo Demográfico, porém para a Pesquisa O/D foram necessários ajustes e criação de classes de acordo com cada variável. A elaboração de cada tabela de contingência e suas hipóteses são explicadas individualmente a seguir.

No Apêndice C são apresentadas todas as tabelas de contingência para a amostra e a população separadamente e para todos os casos.

a) Idade x Gênero

A matriz Idade x Gênero é a mais comum de ser encontrada na literatura. Essa relação se encontra nos trabalhos de: Saadi et al. (2016), Pianucci (2016) e Frick e Axhausen (2003). As distribuições de probabilidade conjunta entre categorias de idade e gênero são também disponíveis na base de dados do censo do IBGE, não sendo necessária a criação de dados sintéticos. Porém, para melhor validar o método os dados sintéticos foram criados para que o

resultado encontrado na etapa de ajuste, ao combinar dados totais marginais do censo com as proporções da amostra, fosse comparado com os dados reais fornecidos.

A amostra tem a sua variável de idade de forma contínua, ou seja, apresenta a idade real de cada indivíduo, sem nenhum tipo de classificação. Para adequar as duas bases de dados serão agrupados os indivíduos dentro das categorias listadas na Tabela 3-3.

A Tabela 3-1 abaixo encontra-se em situação para que seja feito o seu ajuste, pois nota-se que os totais das linhas e das colunas não correspondem à soma das células.

Tabela 3-4: Tabela Idade x Gênero para ajuste

Variáveis		**Gênero**		
		Homem	Mulher	**Total**
Idade	0 a 4 anos	3277	3231	710927
	5 a 9 anos	5336	5160	758279
	10 a 14 anos	6329	6177	867430
	15 a 19 anos	7952	7261	842257
	20 a 24 anos	8898	8956	991659
	25 a 29 anos	9200	9750	1074582
	30 a 34 anos	7771	9228	1010076
	35 a 39 anos	7195	8575	888685
	40 a 44 anos	7228	8294	812979
	45 a 49 anos	7087	7931	742720
	50 a 54 anos	6412	7375	667658
	55 a 59 anos	4937	5615	548113
	60 a 64 anos	4044	4577	423055
	65 a 69 anos	2965	3637	302338
	70 a 74 anos	2320	2665	237301
	75 a 79 anos	1562	2070	170969
	80 a 84 anos	971	1297	119511
	85 a 89 anos	376	566	57205
	90 a 94 anos	136	242	21234
	95 a 99 anos	25	70	5498
	> 100 anos	0	0	1027
	Total	5328632	5924871	11253503

b) Idade x Renda

A variável “Renda” na amostra se encontra de forma contínua, sem classificação. Para compatibilização com as categorias pré-definidas, foi necessário consultar o salário mínimo do ano de 2010 para modificar a variável. O salário mínimo da época foi decretado pela Lei nº 12.255 de 15 de junho de 2010 e possuía a importância de R$ 510,00. E então o resultado encontra-se na Tabela 3-5.

Tabela 3-5: Tabela Idade x Renda para ajuste

Variáveis		Renda								
		Até 1/2	Mais de 1/2 a 1	Mais de 1 a 2	Mais de 2 a 5	Mais de 5 a 10	Mais de 10 a 20	Mais de 20	Sem rendimento	**Total**
Idade	0 a 4 anos	7	1	1	1	0	0	0	2126	710927
	5 a 9 anos	24	22	8	4	0	0	0	5205	758279
	10 a 14 anos	110	29	6	0	0	0	0	6484	867430
	15 a 19 anos	262	1190	784	84	12	0	0	6301	842257
	20 a 24 anos	114	1388	2604	882	123	22	0	5021	991659
	25 a 29 anos	84	958	2408	1847	659	99	6	4837	1074582
	30 a 34 anos	105	780	2010	1650	724	222	27	4499	1010076
	35 a 39 anos	88	604	1632	1633	790	259	57	4192	888685
	40 a 44 anos	67	571	1353	1555	858	314	61	4206	812979
	45 a 49 anos	65	529	1094	1463	978	408	67	3933	742720
	50 a 54 anos	42	386	993	1248	809	365	96	3645	667658
	55 a 59 anos	33	325	655	981	675	259	67	2447	548113
	60 a 64 anos	25	299	550	701	446	244	88	1927	423055
	65 a 69 anos	12	327	345	498	315	147	33	1363	302338
	70 a 74 anos	5	273	260	387	195	103	43	874	237301
	75 a 79 anos	0	216	171	216	114	86	20	509	170969
	80 a 84 anos	0	92	80	99	62	49	4	298	119511
	85 a 89 anos	0	33	21	32	15	13	2	94	57205
	90 a 94 anos	0	8	4	5	8	3	1	40	21234
	95 a 99 anos	0	0	0	1	0	1	0	3	5498
	> 100 anos	0	0	0	0	0	0	0	0	1027
	Total	102784	1119360	2340278	1659072	679919	291102	132882	3458900	11253503

c) Condição de atividade x Gênero

A variável condição de atividade é classificada de forma diferente em ambas as bases de dado. Uma relação entre as classes da amostra e as classes adotadas está apresentada na Tabela 3-6. A classe de estudante foi divida em duas classes, pois englobava indivíduos com mais de 15 anos, que segundo a classificação do IBGE deixam de ser não economicamente ativos e passam a ser desocupados.

A Tabela 3-7 mostra a tabela de contingência obtida para as variáveis "condição de atividade" e gênero para seguir para a etapa de ajuste.

Tabela 3-6: Reclassificação da variável "condição de atividade" da amostra

Classe da amostra	Nova classe	
1 - Tem trabalho	Ocupada	
2 - Faz bico	Ocupada	
3 - Em Licença Médica	Não economicamente ativo	
4 - Aposentado/Pensionista	Não economicamente ativo	
5 - Sem Trabalho	Desocupado	
6 - Nunca Trabalhou	Desocupado	
7 - Dona de Casa	Desocupado	
8 - Estudante	< 15 anos	Não economicamente ativo
	≥ 15 anos	Desocupado

Tabela 3-7: Tabela Condição de Atividade x Gênero para ajuste

Variáveis		Gênero		
		Homem	Mulher	**Total**
Condição de atividade	Ocupadas	36146	30391	6383421
	Desocupadas	6625	13327	516389
	Não economicamente ativas	8961	9843	4353693
	Total	5328632	5924871	11253503

3.1.4. ETAPA DE AJUSTE

Com as tabelas de contingência definidas na etapa anterior foi necessário o ajuste dos valores das células n_{ij}, obtidas da amostra. Para esta etapa foi utilizado o método IPF de Deming e

Stephan (1940), que é um processo iterativo de ajuste dos valores das células de acordo com o valor total marginal esperado. As equações utilizadas foram as equações 4, 5, 6, 7 e 8 abaixo, sendo as três primeiras já apresentadas no capítulo 2 deste trabalho.

$$p_{ij}^{(0)} = n_{ij} \quad \text{(Eq. 7)}$$

$$p_{ij}^{(2t)} = p_{ij}^{(2t-1)} \frac{p_{i+}}{p_{i+}^{(2t-1)}} \quad \text{(Eq. 5)}$$

$$p_{ij}^{(2t+1)} = p_{ij}^{(2t)} \frac{p_{+j}}{p_{+j}^{(2t)}} \quad \text{(Eq. 6)}$$

$$\pi_{ij} = n_{ij}/n \quad \text{(Eq. 4)}$$

$$p_{ij}^{(final)} = \pi_{ij} \quad \text{(Eq. 8)}$$

A ordem de aplicação das equações foi diferente da utilizada por Deming e Stephan (1940), seguindo a ordem apresentada acima, porém sem prejuízo do valor final obtido. Essa inversão das da ordem das equações foi adotada para se trabalhar com valores absolutos de frequência de classe, deixando apenas o valor $p_{ij}^{(final)}$ em forma percentual, para aplicação na etapa de geração dos domicílios sintéticos.

O processo foi iniciado, com a equação 7, considerando o valor inicial de cada célula (t=0) igual ao valor obtido da amostra, apresentados na etapa anterior de elaboração das tabelas. Para as iterações pares (2t), na equação 5, o resultado obtido na iteração anterior (2t-1) foi ajustado de acordo com a soma das linhas. Para as iterações ímpares (2t+1), na equação 6, o resultado obtido na iteração anterior (2t) foi ajustado de acordo com a soma das colunas.

Esses ajustes de colunas e linhas ocorrem pela multiplicação do valor da iteração anterior pela razão entre o valor de soma esperado (totais marginais obtidos do Censo de 2010) e o valor de soma obtido na iteração anterior. As iterações foram interrompidas quando a diferença entre o valor das células n_{ij} entre duas iterações foi menor que 0,01.

Por fim, as equações 4 e 8 foram utilizadas para obter a distribuição de ocorrência conjunta para a etapa seguinte.

As Tabelas 3-8, 3-9 e 3-10 abaixo mostram os resultados obtidos com o método IPF aplicado a cada tabela de contingência. Na lista abaixo é mostrado o número de iterações necessários para o ajuste de cada tabela.

- Tabela 3-8 – 3 iterações;
- Tabela 3-9 – 30 iterações;
- Tabela 3-10 – 5 iterações.

Para aplicação do método IPF foi utilizado o Excel 2016, com a elaboração de planilhas de cálculo, não sendo necessário a programação de código para tal, pois o número de iterações se mostrou baixo, com exceção da tabela Idade x Renda.

Tabela 3-8: Tabela Idade x Gênero ajustada pelo método IPF

Variáveis		**Gênero**		
		Homem	Mulher	**Total**
Idade	0 a 4 anos	353213,93	357713,06	710927
	5 a 9 anos	380418,29	377860,70	758279
	10 a 14 anos	433176,24	434253,76	867430
	15 a 19 anos	434623,56	407633,43	842257
	20 a 24 anos	487576,91	504082,09	991659
	25 a 29 anos	514508,02	560073,98	1074582
	30 a 34 anos	455042,56	555033,44	1010076
	35 a 39 anos	399559,02	489125,98	888685
	40 a 44 anos	373158,40	439820,60	812979
	45 a 49 anos	345534,50	397185,50	742720
	50 a 54 anos	306065,53	361592,48	667658
	55 a 59 anos	252794,48	295318,52	548113
	60 a 64 anos	195629,16	227425,84	423055
	65 a 69 anos	133780,54	168557,47	302338
	70 a 74 anos	108858,60	128442,40	237301
	75 a 79 anos	72407,31	98561,70	170969
	80 a 84 anos	50383,86	69127,14	119511
	85 a 89 anos	22466,83	34738,17	57205
	90 a 94 anos	7509,19	13724,81	21234

	95 a 99 anos	1418,46	4079,54	5498
	> 100 anos	506,62	520,38	1027
	Total	5328632	5924871	11253503

Tabela 3-9: Tabela Idade x Renda ajustada pelo método IPF

Variáveis		Renda								
		Até 1/2	Mais de 1/2 a 1	Mais de 1 a 2	Mais de 2 a 5	Mais de 5 a 10	Mais de 10 a 20	Mais de 20	Sem rendimento	Total
Idade	0 a 4 anos	4189,85	1140,38	1378,04	1123,13	0,90	0,99	2,07	610276,20	710927
	5 a 9 anos	6113,73	10677,42	4691,86	1911,99	0,39	0,42	0,88	635884,81	758279
	10 a 14 anos	25225,91	12670,71	3167,86	0,43	0,35	0,38	0,79	713115,80	867430
	15 a 19 anos	25473,41	220435,41	175493,89	15324,81	1763,52	0,16	0,34	293804,23	842257
	20 a 24 anos	7538,16	174863,39	396425,84	109435,84	12293,58	2409,57	0,23	159225,94	991659
	25 a 29 anos	5442,69	118262,84	359212,00	224559,41	64540,46	10624,91	1342,34	150304,87	1074582
	30 a 34 anos	7078,11	100177,81	311949,67	208709,60	73769,92	24787,75	6284,46	145447,78	1010076
	35 a 39 anos	5718,32	74777,54	244155,19	199114,10	77593,47	27876,69	12789,00	130638,05	888685
	40 a 44 anos	4216,20	68459,17	196022,03	183614,72	81610,62	32728,97	13254,18	126934,30	812979
	45 a 49 anos	3956,69	61351,13	153318,92	167106,28	89984,89	41137,15	14082,16	114816,68	742720
	50 a 54 anos	2618,11	45843,17	142510,93	145976,69	76225,40	37686,64	20662,66	108968,02	667658
	55 a 59 anos	2294,09	43045,63	104833,12	127966,48	70927,28	29823,08	16082,30	81581,90	548113
	60 a 64 anos	1677,23	38218,43	84952,46	88247,27	45227,27	27114,32	20385,06	62000,84	423055
	65 a 69 anos	819,11	42526,44	54217,81	63785,54	32500,17	16620,18	7777,73	44619,15	302338
	70 a 74 anos	357,84	37224,31	42839,94	51970,45	21094,17	12209,79	10625,76	29997,81	237301
	75 a 79 anos	0,08	33274,29	31831,91	32771,05	13932,33	11517,56	5583,58	19737,27	170969
	80 a 84 anos	0,12	20771,55	21826,42	22013,94	11105,46	9617,99	1636,70	16936,02	119511
	85 a 89 anos	0,19	11691,79	8990,79	11166,02	4216,21	4004,22	1284,18	8383,19	57205
	90 a 94 anos	0,25	3826,83	2312,17	2355,60	3036,01	1247,61	866,92	4816,41	21234
	95 a 99 anos	0,96	1,83	2,21	1800,52	1,45	1589,37	3,31	1380,55	5498
	> 100 anos	62,95	119,94	144,93	118,12	95,15	104,27	217,36	30,19	1027
	Total	102784	1119360	2340278	1659072	679919	291102	132882	3458900	11253503

Tabela 3-10: Tabela Condição de atividade x Gênero ajustada pelo método IPF

Variáveis		Gênero		
		Homem	Mulher	**Total**
Condição de atividade	Ocupadas	3247523,90	3135897,10	6383421
	Desocupadas	155993,91	360395,09	516389
	Não economicamente ativas	1925114,19	2428578,81	4353693
	Total	5328632	5924871	11253503

3.1.5. ETAPA DE GERAÇÃO DOS DOMICÍLIOS SINTÉTICOS

A etapa de geração dos domicílios sintéticos foi desenvolvida para este trabalho, com base no trabalho de Moeckel et al., (2003). Com as distribuições de ocorrência conjunta obtidas no método IPF foi possível gerar indivíduos com características próprias para cada domicílio sintético. O processo adotado para gerar cada domicílio é iterativo e está ilustrado na Figura 3-7.

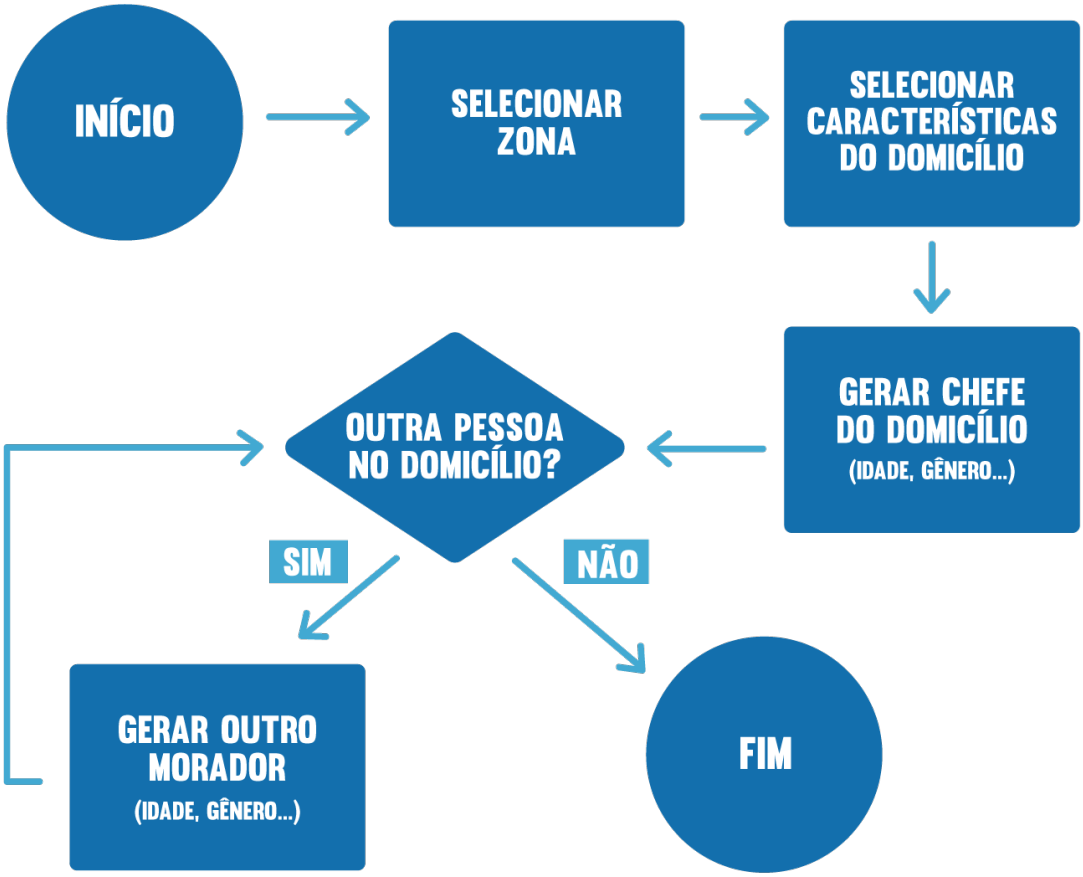

Figura 3-7: Método iterativo para geração de domicílio sintética (MOECKEL et al., 2003)

Para realizar esse processo iterativo foi utilizado o *software* PyCharm para desenvolvimento de um programa em linguagem Python 3. O código desenvolvido está disponível no Apêndice D.

Antes de gerar cada domicílio sintético, foi fixado o número de domicílios na cidade de São Paulo de acordo com o censo de 2010. No município existiam 3470566 domicílios particulares ocupados e esse foi o número adotado para esta pesquisa.

Características do domicílio

a) Posse de veículo

Para os domicílios foi definido como característica se possuíam ou não veículo próprio. Esta variável se encontrava presente apenas na Pesquisa O/D, portanto a distribuição de ocorrência foi baseada apenas na amostra, conforme Tabela 3-11.

Tabela 3-11: Distribuição de posse de veículo próprio na amostra

Posse de veículo	Distribuição
Sim	68,06%
Não	31,94%

b) Número de moradores em cada domicílio

O número de moradores em cada domicílio foi defino de acordo com a distribuição de ocorrência encontrada na Pesquisa O/D, uma vez que esta variável não se encontra no Censo Demográfico. Na amostra o número variava entre 1 e 24 moradores, porém para as classes maiores que 14 moradores a probabilidade de ocorrência era menor que 0,01%, portanto foram desconsideradas. A distribuição adotada é mostrada na Tabela 3-12.

Tabela 3-12: Distribuição do número de moradores na amostra

Número de moradores	Distribuição
1	5,20%
2	15,80%
3	23,70%
4	27,90%
5	14,70%
6	6,40%
7	3,00%
8	1,20%
9	0,80%
10	0,70%
11	0,30%
12	0,10%
13	0,10%

Características dos indivíduos

Definidos o número de moradores do domicílio, cada morador é gerado individualmente com as seguintes características:

a) Gênero

A escolha do gênero dos indivíduos foi considerada independente de qualquer outra variável. O censo de 2010 fornece a distribuição de homens e mulheres dentre os 11 253 503 milhões de habitantes. A distribuição inserida no código consta na Tabela 3-13.

Tabela 3-13: Distribuição do gênero na população

Gênero	Distribuição
Homens	47,35%
Mulheres	52,65%

b) Idade

O Censo Demográfico fornece a distribuição de grupos de idade na população divida por gênero, portanto a tabela de contingência Idade x Gênero ajustada pelo método IPF não foi utilizada para determinação da idade. Para a seleção do grupo de idade do indivíduo foi feita com a probabilidade de ocorrência do grupo de idade, dado que o gênero é definido, uma vez que o gênero foi definido anteriormente. A categoria de pessoas com mais de 100 anos não foi considerada nessa escolha, pois sua probabilidade de ocorrência era menor que 0,01%. As probabilidades consideradas estão na Tabela 3-14.

Caso o indivíduo fosse o primeiro do domicílio (chefe), foi considerado que esse deveria ser um adulto, ou seja, ter mais de 20 anos. Desta forma foi eliminada a probabilidade de ocorrência de domicílios compostos apenas por crianças e adolescentes, situação que normalmente não é observada com frequência na população.

Tabela 3-14: Distribuição de idade na população de acordo com gênero

Grupo de Idade	Distribuição	
	Homem	Mulher
0 a 4 anos	6,79%	5,89%
5 a 9 anos	7,24%	6,29%
10 a 14 anos	8,23%	7,24%
15 a 19 anos	7,89%	7,12%
20 a 24 anos	9,18%	8,48%
25 a 29 anos	9,75%	9,37%
30 a 34 anos	9,03%	8,93%
35 a 39 anos	7,94%	7,86%
40 a 44 anos	7,23%	7,22%
45 a 49 anos	6,42%	6,76%
50 a 54 anos	5,66%	6,17%
55 a 59 anos	4,58%	5,14%
60 a 64 anos	3,43%	4,05%
65 a 69 anos	2,38%	2,96%

70 a 74 anos	1,79%	2,40%
75 a 79 anos	1,21%	1,80%
80 a 84 anos	0,78%	1,32%
85 a 89 anos	0,33%	0,67%
90 a 94 anos	0,11%	0,26%
95 a 99 anos	0,03%	0,07%

c) Condição de atividade

O censo de 2010 não fornece nenhum tipo de distribuição de probabilidade conjunta entre a variável de condição de atividade e outra variável. Com os dados da amostra, na etapa de ajusta das tabelas de contingência foi obtida a tabela Condição de atividade x Gênero, que foi utilizada na escolha da condição de atividade dos indivíduos.

Tabela 3-15: Distribuição da condição de atividade ajustada pelo IPF de acordo com gênero

Condição de atividade	Gênero	
	Homem	Mulher
Ocupadas	60,94%	52,93%
Desocupadas	2,93%	6,08%
Não economicamente ativas	36,13%	40,99%

Domicílios e população sintéticos

Com todas as variáveis definidas e as suas probabilidades de ocorrência os domicílios sintéticos foram gerados, assim como os seus moradores. Cada indivíduo foi armazenado em uma linha de um arquivo de texto (.txt), com um código de identificação, que se tornou uma base de dados que pode ser importada para o Excel 2016 para posterior análise dos dados. A Tabela 3-16 mostra um exemplo de linhas do arquivo de texto obtido ao final da geração da população sintética.

Tabela 3-16: Exemplo de linhas do arquivo de texto

ID_domicílio	Veículo	ID_individuo	Idade	Gênero	Atividade
1	Sim	1	20 a 24 anos	Homem	Ocupado
1	Sim	2	25 a 29 anos	Mulher	Ocupada
2	Não	3	30 a 34 anos	Homem	Desocupado

As tabelas 3-17, 3-18, 3-19 e 3-20 mostram a frequência de ocorrência das variáveis em toda a população sintética.

Tabela 3-17: Frequência de ocorrência dos grupos de idade na população sintética

Grupo de Idade	Frequência
0 a 4 anos	621788
5 a 9 anos	663116
10 a 14 anos	758247
15 a 19 anos	734613
20 a 24 anos	1292109
25 a 29 anos	1401704
30 a 34 anos	1316893
35 a 39 anos	1159278
40 a 44 anos	1061261
45 a 49 anos	968268
50 a 54 anos	869177
55 a 59 anos	714223
60 a 64 anos	551626
65 a 69 anos	394359
70 a 74 anos	309512
75 a 79 anos	222305
80 a 84 anos	155132
85 a 89 anos	74469
90 a 94 anos	27500
95 a 99 anos	7212
Total	13302792

Tabela 3-18: Frequência da ocorrência dos gêneros na população sintética

Gênero	Frequência
Masculino	6300050
Feminino	7002742
Total	13302792

Tabela 3-19: Frequência da ocorrência da condição de atividade na população sintética

Condição de atividade	Frequência
Ocupado	7543372
Desocupado	611423
Não economicamente ativo	5147997
Total	13302792

Tabela 3-20: Frequência da ocorrência da posse de veículo nos domicílios sintéticos

Veículo no domicílio	Frequência
Sim	2359919
Não	1110647
Total	3470566

4. RESULTADOS

Este capítulo apresenta os resultados e análise da etapa de ajuste das tabelas de contingência e etapa de geração dos domicílios sintéticos, através dos métodos descritos no capítulo anterior.

4.1. ETAPA DE AJUSTE

Na etapa de ajuste das tabelas de contingência foram obtidas as distribuições de probabilidade conjunta em tabelas de contingência de duas dimensões. As tabelas resultantes foram apresentadas no capítulo anterior.

Para avaliar a eficácia do método IPF foi realizado o ajuste para a tabela Idade x Gênero, mesmo que a informação já fosse presente no Censo Demográfico de 2010. Nas Figuras 4-1 e 4-2 percebe-se que houve pouca discrepância entre a frequência de ocorrência da variável idade para os dois gêneros, levando a duas análises:

- A amostra de fato é representativa da população, uma vez que o método IPF é bastante influenciado pelos valores de entrada; e
- O método IPF é eficaz para as classes com ocorrência mais frequente. Para as idades entre 0 e 69 anos os erros relativos entre as estimativas do método e do valor observado no censo foram menores que 7%. Os erros relativos entre os valores da população e do método estão disponíveis na Tabela 4-1.

Tabela 4-1: Erros relativos entres as classes de idade (Censo x IPF)

Idade	Erro relativo	
	Homem	Mulher
0 a 4 anos	2,35%	2,43%
5 a 9 anos	1,36%	1,41%
10 a 14 anos	1,18%	1,21%
15 a 19 anos	3,35%	3,34%
20 a 24 anos	0,38%	0,37%
25 a 29 anos	1,00%	0,93%
30 a 34 anos	5,45%	4,96%
35 a 39 anos	5,55%	5,04%
40 a 44 anos	3,12%	2,81%
45 a 49 anos	1,02%	0,87%
50 a 54 anos	1,40%	1,15%
55 a 59 anos	3,66%	2,94%
60 a 64 anos	6,89%	5,26%
65 a 69 anos	5,32%	3,86%
70 a 74 anos	14,33%	9,60%
75 a 79 anos	12,57%	7,58%
80 a 84 anos	21,98%	11,61%
85 a 89 anos	26,67%	11,98%
90 a 94 anos	27,77%	10,63%
95 a 99 anos	11,69%	3,51%

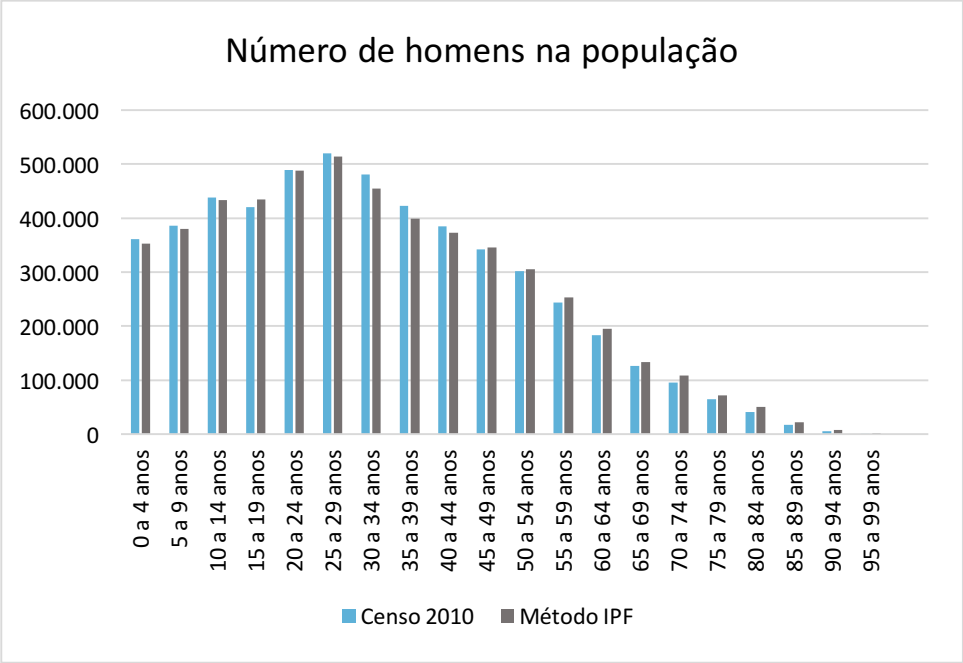

Figura 4-1: Histograma do número de homens na população de São Paulo (Censo x IPF)

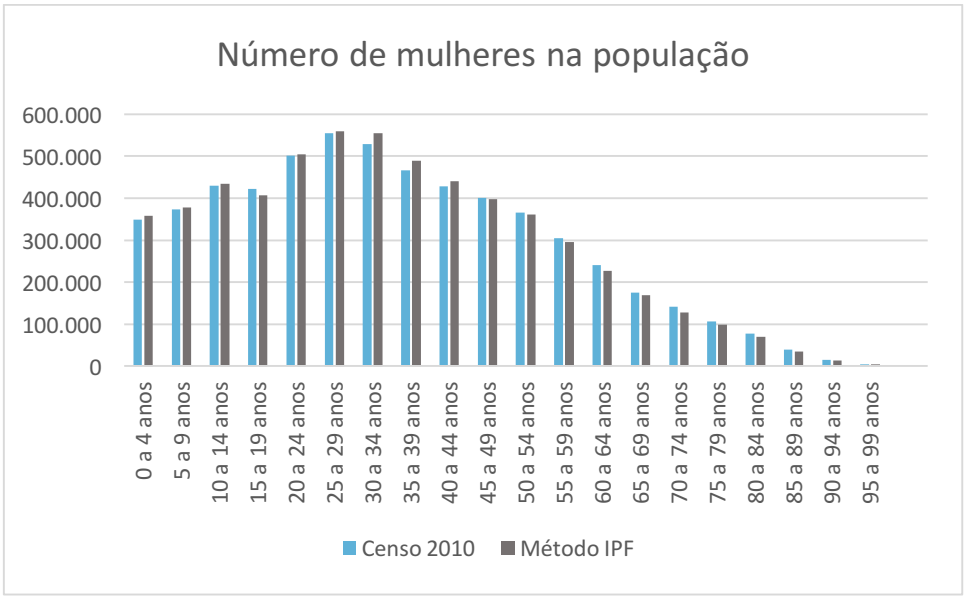

Figura 4-2: Histograma do número de mulheres na população de São Paulo (Censo x IPF)

4.2. ETAPA DE GERAÇÃO DOS DOMICÍLIOS SINTÉTICOS

Na etapa de geração dos domicílios sintéticos foram obtidos os domicílios sintéticos habitados. Os valores absolutos de ocorrência de cada variável foram apresentados no capítulo anterior.

Número de indivíduos

Percebe-se inicialmente que o número de indivíduos sintéticos obtidos é superior ao número de indivíduos existentes no censo de 2010, conforme Figura 4-3. O erro relativo entre os dois valores é de 18,21%. Esse resultado pode ser analisado da seguinte forma:

- As probabilidades de ocorrência do número de moradores na amostra não são representativas;
- Com a consideração da cidade como uma grande zona, pode ocorrer de áreas com uma densidade baixa de habitantes/domicílio receberem mais ocupantes que o observado na realidade.

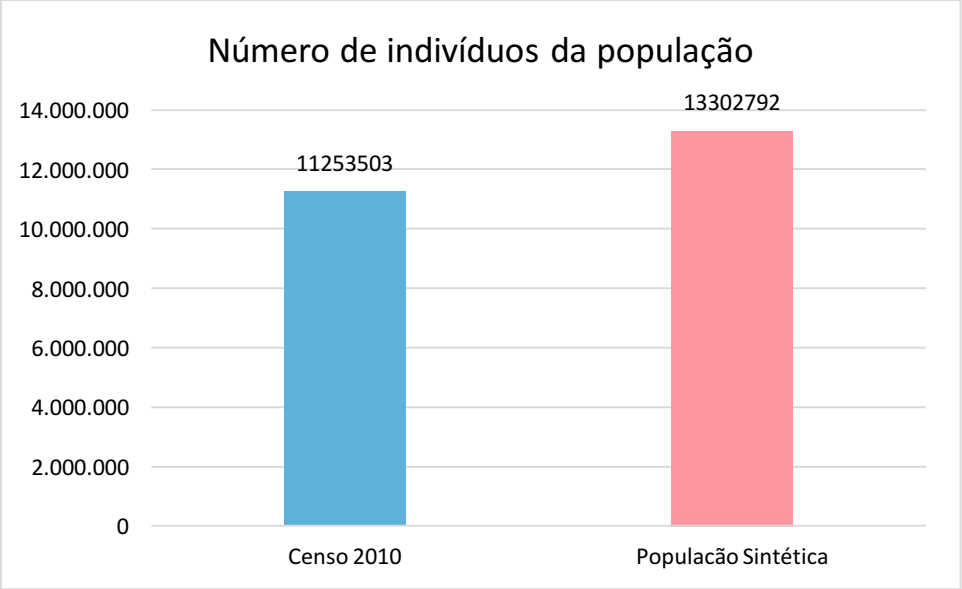

Figura 4-3: Número de indivíduos da população de São Paulo (Censo x Pop. Sintética)

Domicílios sintéticos

O número de iterações para obtenção da população sintética foi fixado como a quantidade de domicílios particulares ocupados (3470566), portanto o número de domicílios sintéticos é igual ao do censo.

Posse de veículo próprio

A probabilidade de ocorrência da posse de veículo particular em cada domicílio foi obtida da Pesquisa O/D, uma vez que essa variável não consta no censo de 2010. Portanto, a comparação do número absoluto obtido entre a pesquisa e a população sintética não fornece nenhuma base para análise, sendo então analisada apenas a distribuição de ocorrência da variável. Observou-

se na Figura 4-4 que a distribuição entre a Pesquisa O/D e a população sintética foi mantida relativamente igual, com diferença de 0,06%.

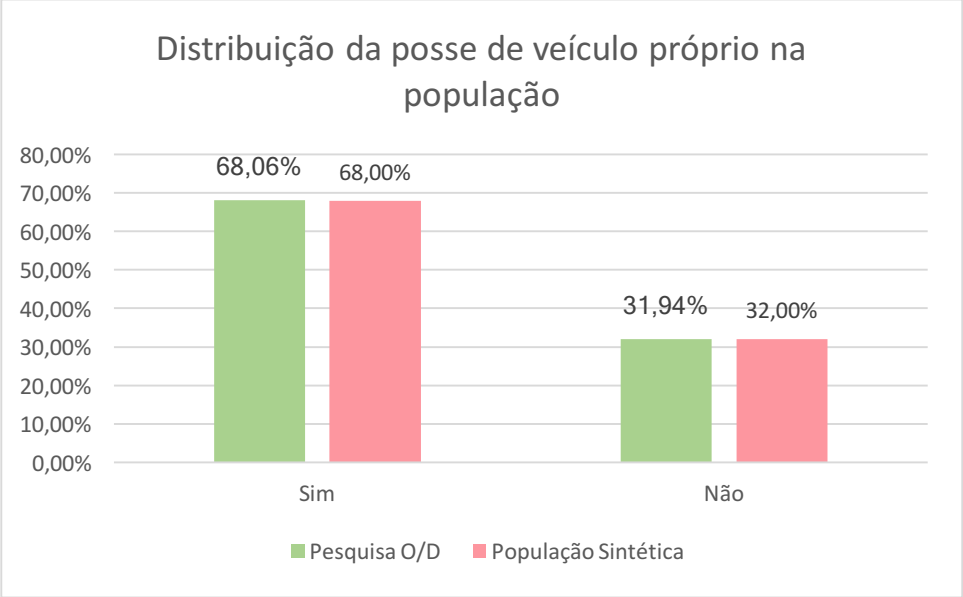

Figura 4-4: Distribuição da posse de veículo próprio (Pesquisa O/D x Pop. Sintética)

Gênero

As probabilidades de ocorrência do gênero da população do censo e da população sintética (Figura 4-5) são relativamente iguais, com diferença de apenas 0,01%, uma vez que a sua distribuição foi obtida do próprio censo e não do método IPF.

Do gráfico da Figura 4-6 percebe-se um erro relativo de aproximadamente 18%, apresentado na (Tabela 4-2) entre o número absoluto de indivíduos por categoria de gênero. Essa diferença é devida ao total de indivíduos na população sintética em relação ao número de indivíduos do censo, e não em relação à distribuição de probabilidade.

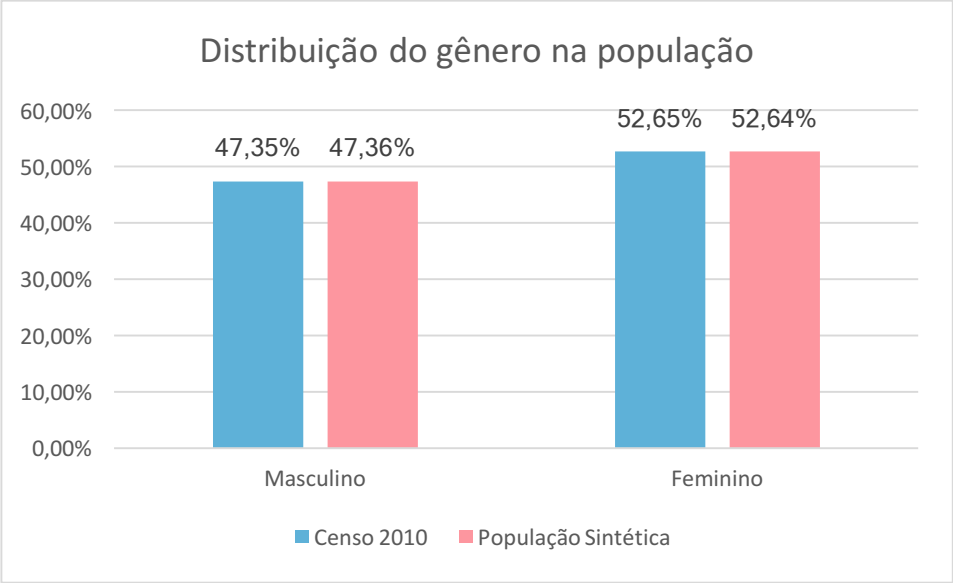

Figura 4-5: Distribuição do gênero na população (Censo x Pop. sintética)

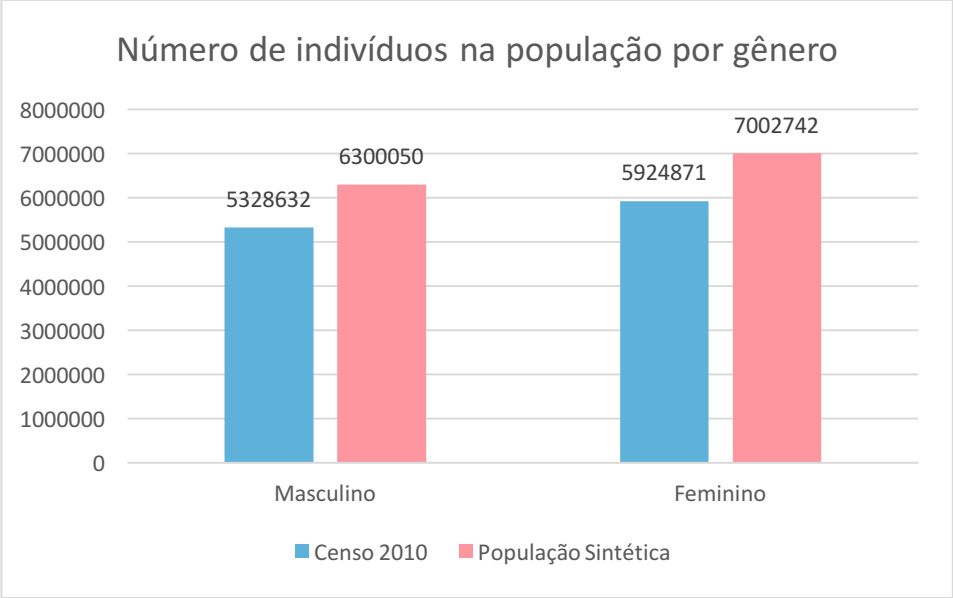

Figura 4-6: Número de indivíduos na população por gênero (Censo x Pop. sintética)

Tabela 4-2: Erro relativo da ocorrência de gênero na população

Gênero	Erro relativo
Masculino	18,23%
Feminino	18,19%

Idade

A distribuição dos grupos de idade na população sintética e na população do censo não é compatível. Percebe-se da imagem (Figura 4-7) que para as categorias que englobam os indivíduos que possuem entre 0 e 19 anos a frequência de ocorrência é maior nos dados do censo que na população sintética. Tal fato se deve à imposição feita na etapa de aplicação, onde a restrição de que o primeiro indivíduo do domicílio deveria ter o primeiro indivíduo como adulto, dessa forma foi criado uma tendência para escolha das idades dos indivíduos.

Conforme descrito acima e demonstrado na Figura 4-8 e Tabela 4-3, a hipótese adotada fez os erros relativos entre o número de ocorrência ser negativos nas categorias: 0 a 4 anos, 5 a 9 anos, 10 a 14 anos e 15 a 19 anos. Desta maneira o número de indivíduos com idade maior ou igual a 20 anos foi superestimada.

Diante o exposto, cabe uma revisão na forma de impedir que sejam gerados domicílios sintéticos apenas com crianças e adolescentes.

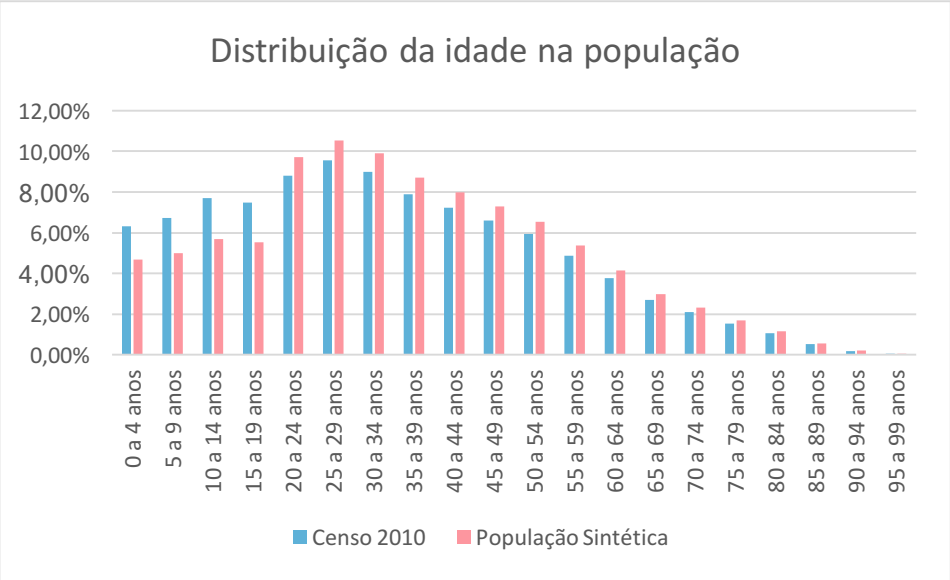

Figura 4-7: Distribuição da idade na população (Censo x Pop. Sintética)

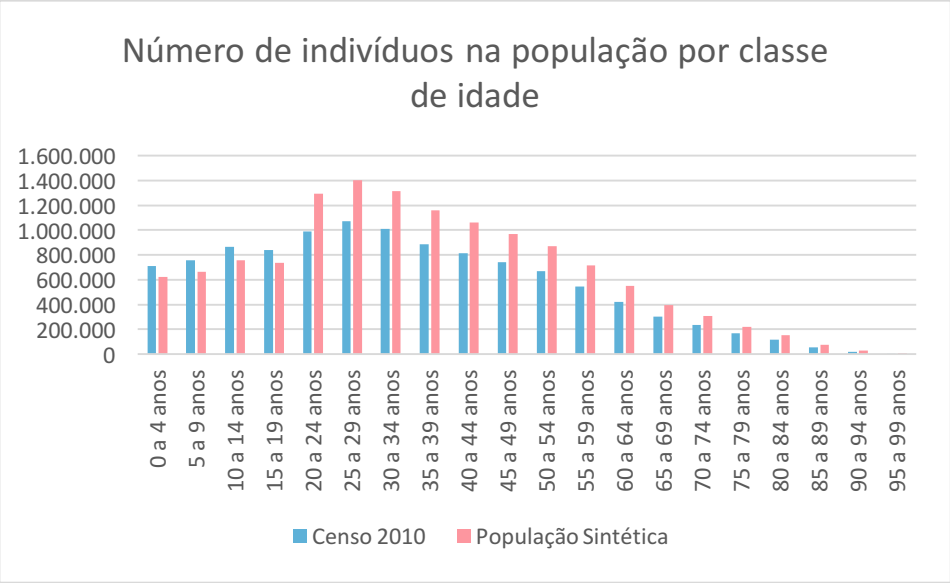

Figura 4-8: Número de indivíduos na população por grupo de idade (Censo x Pop. Sintética)

Tabela 4-3: Erro relativo do resultado da ocorrência de idade na população

Grupo de Idade	Erro relativo
0 a 4 anos	12,54%
5 a 9 anos	12,55%
10 a 14 anos	12,59%
15 a 19 anos	12,78%
20 a 24 anos	30,30%
25 a 29 anos	30,44%
30 a 34 anos	30,38%
35 a 39 anos	30,45%
40 a 44 anos	30,54%
45 a 49 anos	30,37%
50 a 54 anos	30,18%
55 a 59 anos	30,31%
60 a 64 anos	30,39%
65 a 69 anos	30,44%
70 a 74 anos	30,43%
75 a 79 anos	30,03%
80 a 84 anos	29,81%
85 a 89 anos	30,18%
90 a 94 anos	29,51%
95 a 99 anos	31,17%

Condição de atividade

A condição de atividade, conforme Figura 4-9, apresentou uma distribuição de ocorrência similar entre o censo e a população sintética, com diferença de apenas 0,01%. Esse resultado afirma a validade do método IPF enquanto método para ajuste das distribuições de probabilidade conjunta entre as variáveis.

Os valores absolutos, demonstrados na Figura 4-10, apresentam erros relativos positivos (Tabela 4-4) entre o número esperado e as ocorrências sintéticas maiores que 35%, devido à superestimação do número de indivíduos sintéticos conforme descrito no tópico acerca do número de indivíduos.

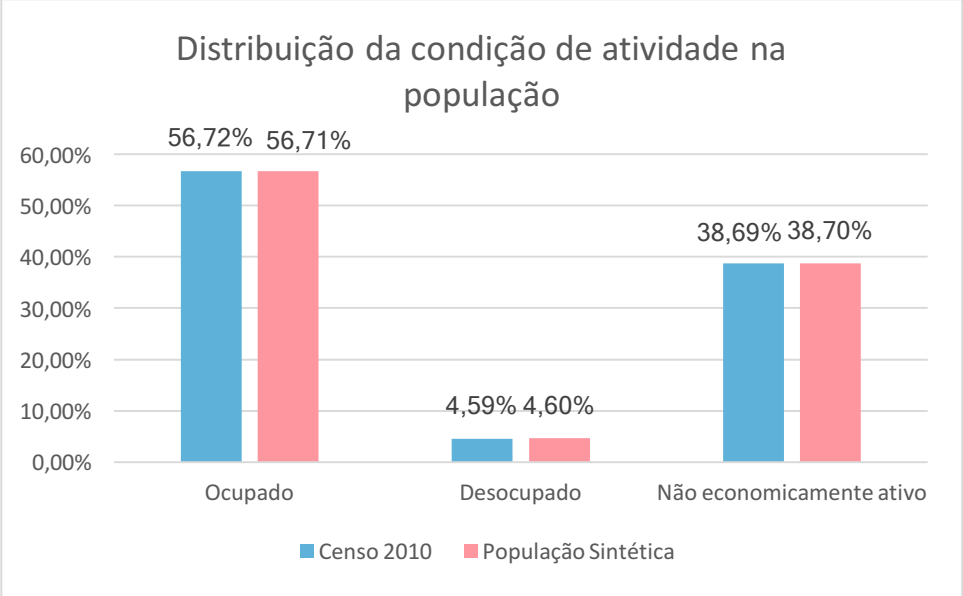

Figura 4-9: Distribuição da condição de atividade na população (Censo x Pop. Sintética)

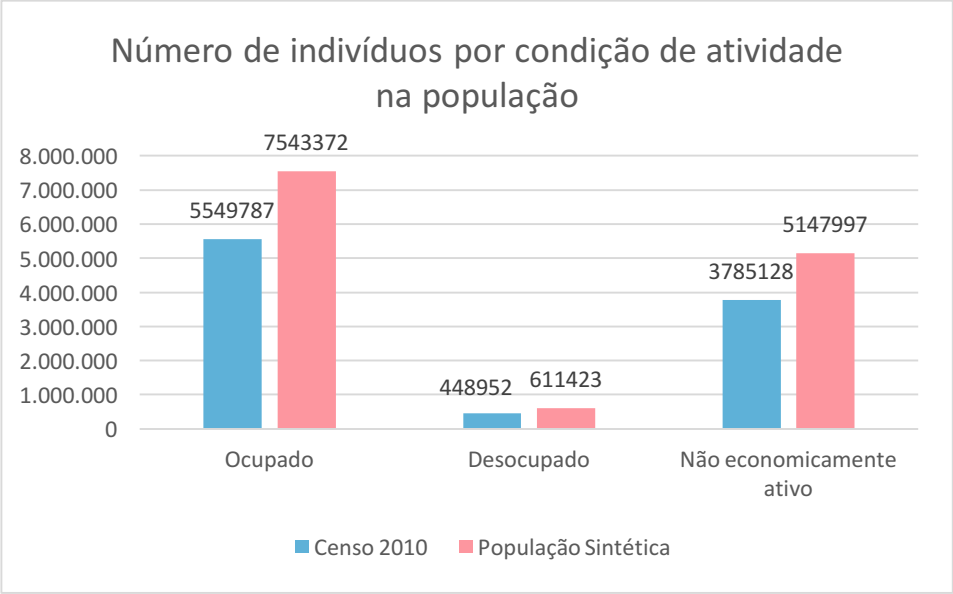

Figura 4-10: Número de indivíduos na população por condição de atividade (Censo x Pop. Sintética)

Tabela 4-4: Erro relativo do resultado da condição de atividade na população

Condição de atividade	Erro relativo
Ocupado	35,92%
Desocupado	36,19%
Não economicamente ativo	36,01%

5. CONCLUSÕES E DISCUSSÕES

5.1. CONCLUSÃO DO ESTUDO

Este trabalho teve o propósito de apresentar uma metodologia para geração de população sintética para o município de São Paulo, com base em dados do Instituto Brasileiro de Geografia e Estatística da Companhia do Metropolitano de São Paulo, cujos dados são de fácil acesso. O processo foi baseado na obtenção de dados agregados e desagregados, aplicação e análise de métodos apresentados na literatura como eficazes para obtenção de uma população sintética representativa da população real em estudo.

Ambas as bases de dados fornecem seus dados de forma simples em suas páginas na internet. Houve uma compatibilização das diferentes classificações que são utilizadas nas bases de dados de forma a ser possível a integração dos dados. Tal compatibilização se mostrou possível, pois O IBGE e a Companhia do Metropolitano utilizam variáveis iguais, como idade e gênero ou similares como as de condição de atividade, para a caracterização da população em questão.

Os resultados obtidos pelo método IPF se mostraram satisfatórios, principalmente para os casos onde o número de células para ajuste não era demasiado, como o caso da tabela Gênero x Condição de Atividade, apresentando nenhuma ou mínima variação em relação aos dados censitários. Nos casos de matrizes extensas, como a tabela Idade x Gênero o método apresentou maior variação em relação ao censo, porém por se tratar de uma estimativa baseada em uma amostra de menos de 1% da população original os resultados se mostraram satisfatórios.

O método desenvolvido para a geração dos indivíduos e domicílios sintéticos pode ser analisado em termos de valores absolutos e valores de proporção. A população sintética obtida foi 18,21% maior que a população do município de São Paulo no ano de 2010, com excedente de 2049289 indivíduos. Esse é um número bastante elevado de indivíduos para serem considerado a mais nos modelos de planejamento. Porém, ao analisar a distribuição das variáveis entre a população sintética e a população real, observou-se que as proporções de ocorrência de cada característica socioeconômica se mantiveram praticamente idênticas, com variações de no máximo 0,06%, com exceção da variável dos grupos de idade na população, que apresentou erros relativos

maiores, devido à tentativa de obter resultados lógicos, sem a ocorrência de domicílios compostos apenas de crianças e adolescentes.

5.2. LIMITAÇÕES ENCONTRADAS

Os dados fornecidos pelo IBGE não abrangem variáveis relacionadas ao comportamento de viagem da população, desta forma, como são facilmente encontradas outras fontes de dados que forneçam informações publicamente acerca do comportamento de viagem de toda população, os estudos de geração de populações sintéticas tornam-se limitados no âmbito nacional.

O Censo Demográfico utilizado nesta pesquisa data de 2010, portanto os resultados obtidos encontram-se defasados temporalmente, assim como a Pesquisa O/D de São Paulo que é do ano de 2007. Embora a versão do ano de 2017 estivesse sendo elaborada os seus dados não foram disponibilizados publicamente até o término desta pesquisa. O IBGE fornece estimativas populacionais para os anos entre censos, porém essas estimativas fornecem puramente a quantidade de habitantes do município e não engloba mudanças que podem ter ocorrido ao decorrer dos 7 anos entre a realização do censo e a elaboração deste trabalho.

5.3. SUGESTÕES PARA TRABALHOS FUTUROS

Com os aprendizados e limitações encontradas sugere-se para futuros trabalhos:

- Otimização do código utilizado para geração dos domicílios sintéticos, de forma a torná-lo de mais rápido processamento e de maneira que possam ser inseridos dados de fontes externas, como planilhas e arquivos de texto;
- Inclusão de variáveis relacionadas à viagens diárias para análises de comportamento de viagem;
- Aplicação da população sintética gerada em modelos de microssimulação de demanda de transportes, para melhor validação de seu comportamento enquanto representação fidedigna da população real tomada como base;
- Escolha de municípios com menor quantidade de habitantes e menor complexidade como a cidade de São Paulo;

- Elaboração de coleta de dados própria, para obtenção de todas as variáveis consideradas importantes para a aplicação da população sintética em modelos de microssimulação.

REFERÊNCIA BIBLIOGRÁFICA

ADIGA, A. et al. Generating a synthetic population of the United States. Virginia Tech, p. 1–9, 2015.

AGRESTI, A. **Categorial Data Analysis**. 2 ed. Estados Unidos: John Wiler & Sons, Inc. 2002. 721 p.

ANDERSON, P.; FAROOQ B.; EFTHYMIOU D.; M. BIERLAIRE. Association Generation in Synthetic Population for Transportation Applications: Graph-Theoretic Solution. **Transportation Research Record, Journal of the Transportation Research Board**, Washington D.C., Vol. 2429, pp. 38-50, 2014. v. 2429, p. 38–50, 2014.

ARENTZE, T.; TIMMERMANS, H.; HOFMAN, F. Creating Synthetic Household Populations Problems and Approach. **Transportation Research Record**, p. 85–91, 2014.

BANISTER, D. **Transport Planning**. 2 ed. Londres, Reino Unido: Spoon Press, 2002.

BECKMAN, R. J.; BAGGERLY, K. A.; MCKAY, M. D. Creating Synthetic Baseline Populations. **Transportation Research Part A: Policy and Practice,** Elsevier, v. 30, n. 6, 1996.

BOGLE, B. M.; MEHROTRA, S. A Moment Matching Approach for Generating Synthetic Data. **Big Data**. v. 4, n. 3, p. 160–178, 2016.

DEMING, W. E.; STEPHAN, F. F. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. **The Annals of Mathematical Statistics**, v. 11, n. 4, p. 427–444, 1940.

EVERITT, B. S. **The Analysis of Contigency Tables**. Dissertacão de Mestrado. Institue of Psychiatry. Londres, Reino Unido. 135 p. 1997

FRICK, M.; AXHAUSEN, K. W. Generating Synthetic Populations using Iterative

Proportional Fitting (IPF) and Monte Carlo Techniques. In: 3rd SWISS TRANSPORT RESEARCH CONFERECE. Monte Verità, Suiça. **Anais…** 25 p. 2003.

FRICK, M.; AXHAUSEN, K. W. Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. In: 4rd SWISS TRANSPORT RESEARCH CONFERECE. Monte Verità, Suiça. **Anais…** 28 p. 2004.

GRAHAM, P.; YOUNG, J.; PENNY, R. **Methods for Creating Synthetic Data**. Wellington, New Zeland: The Official Statistics System, 2008.

GUO, J. Y.; BHAT, C. R. Population synthesis for microsimulating travel behavior. **Transportation Research Record:** Journal of the Transportation Research Board. p. 1–25, 2007.

INSTITUTE OF TRANSPORTATION ENGINEERS. **Transport Planning**. 4 Edição ed. Hoboken, New Jersey: Wiley, 2016.

MA, L. **Generating Disaggregate Population Characteristics for Input to Travel-Demand Models**. Tese de Doutorado. University of Florida, 2011.

MOECKEL, R.; WEGENER, M.; SPIEKERMANN K. Creating a Synthetic Population. In: 8TH INTERNATIONAL CONFERENCE ON COMPUTERS IN URBAN PLANNING AND URBAN MANAGEMENT. Sendai, Japão. **Anais…** p. 1–18, 2003.

MÜLLER, K.; AXHAUSEN, K. W. **Population synthesis for microsimulation : State of the art**. 90TH ANNUAL MEETING OF TRANSPORTATION RESEARCH BOARD. **Anais**...Washington, DC: 2010

MÜLLER, K.; AXHAUSEN, K. W. Hierarchical IPF: Generating a synthetic population for Switzerland. In: 51 ST CONGRESS OF THE EUROPENA REGIONAL SCIENCE ASSOCIATION. **Anais…** Barcelona, Espanha. 22 p. 2011.

ORTÚZAR, J. DE D.; WILLUMSEN, L. G. **Modelling Transport**. 3. ed. Chicester, England: Wiley, 2004.

PIANUCCI, M. N. **Uma proposta para a obtenção da população sintética através de dados agregados para modelagem de geração de viagens por domicílio**. 186 p. Tese de Doutorado. Escola de Engenharia de São Carlos, Universidade de São Paulo, 2016.

SAADI, I. et al. Hidden Markov Model-based population synthesis. **Transportation Research Part B: Methodological,** Elsevier v. 90, p. 1–21, 2016.

APÊNDICE A – Variáveis do Banco de Dados da Pesquisa O/D de São Paulo/SP

Variável	Conteúdo	Tipo de variável
ZONA	Zona do Domicílio	Discreta
MUNI_DOM	Município de Domicílio	Discreta
CO_DOM_X	Coordenada X Domicílio	Numérica
CO_DOM_Y	Coordenada Y Domicílio	Numérica
ID_DOM	Identifica Domicílio	Numérica
F_DOM	Identifica Primeiro Registro do Domicílio	Discreta
FE_DOM	Fator de Expansão do Domicílio	Numérica
DOM	Número do Domicílio	Numérica
CD_ENTRE	Código de Entrevista	Discreta
DATA	Data da Entrevista	Numérica
TIPO_DOM	Tipo de Domicílio	Discreta
NO_MORAD	Total de Moradores no Domicílio	Numérica
TOT_FAM	Total de Familias no Domicílio	Numérica
ID_FAM	Identifica Família	Numérica
F_FAM	Identifica Primeiro Registro da Família	Discreta
FE_FAM	Fator de Expansão da Família	Numérica
FAMILIA	Número da Família	Numérica
NO_MORAF	Total de Moradores na Família	Numérica
CONDMORA	Condição de Moradia	Discreta
NAO_DCL_IT	Declaração de Itens de Conforto	Discreta
CRITERIO_B	Critério de Classificação Econômica	Discreta
ANO_AUTO1	Ano Fabricação - Auto 1	Numérica
ANO_AUTO2	Ano Fabricação - Auto 2	Numérica
ANO_AUTO3	Ano Fabricação - Auto 3	Numérica
RENDA_FA	Renda Familiar	Numérica
CD_RENFA	Código de Renda Familiar	Discreta
ID_PESS	Identifica Pessoa	Numérica
F_PESS	Identifica Primeiro Registro da Pessoa	Discreta
FE_PESS	Fator de Expansão da Pessoa	Numérica

PESSOA	Número da Pessoa	Numérica
SIT_FAM	Situação Familiar	Discreta
IDADE	Idade	Numérica
SEXO	Gênero	Discreta
ESTUDA	Estuda Atualmente?	Discreta
GRAU_INS	Grau de Instrução	Discreta
CD_ATIVI	Condição de Atividade	Discreta
CO_REN_I	Condição de Renda Individual	Discreta
VL_REN_I	Renda Individual	Numérica
ZONA_ESC	Zona da Escola	Discreta
MUNIESC	Município da Escola	Discreta
CO_ESC_X	Coordenada X Escola	Numérica
CO_ESC_Y	Coordenada Y Escola	Numérica
TIPO_ESC	Tipo de Escola	Discreta
ZONATRA1	Zona do Primeiro Trabalho	Discreta
MUNITRA1	Município do Primeiro Trabalho	Discreta
CO_TR1_X	Coordenada X 1º Trabalho	Numérica
CO_TR1_Y	Coordenada Y 1º Trabalho	Numérica
TRAB1_RE	Primeiro Trabalho é igual a Residência ?	Discreta
TRABEXT1	Realiza Trabalho Externo-1º Trabalho	Discreta
OCUP1	Ocupação do 1º Trabalho	Discreta
SETOR1	Setor de Atividade do 1º Trabalho	Discreta
VINC1	Vínculo Empregatício do 1º Trabalho	Discreta
ZONATRA2	Zona do Segundo Trabalho	Discreta
MUNITRA2	Município do Segundo Trabalho	Discreta
CO_TR2_X	Coordenada X 2º Trabalho	Numérica
CO_TR2_Y	Coordenada Y 2º Trabalho	Numérica
TRAB2_RE	Segundo Trabalho é igual a Residência ?	Discreta
TRABEXT2	Realiza Trabalho Externo 2º Trabalho	Discreta
OCUP2	Ocupação do 2º Trabalho	Discreta
SETOR2	Setor de Atividade do 2º Trabalho	Discreta
VINC2	Vínculo Empregatício do 2º Trabalho	Discreta
N_VIAG	Número da Viagem	Numérica

FE_VIA	Fator de Expansão da Viagem	Numérica
DIA_SEM	Dia da Semana	Discreta
TOT_VIAG	Total de Viagens da Pessoa	Numérica
ZONA_O	Zona de Origem	Discreta
MUNI_O	Município de Origem	Discreta
CO_O_X	Coordenada X Origem	Numérica
CO_O_Y	Coordenada Y Origem	Numérica
ZONA_D	Zona de Destino	Discreta
MUNI_D	Município de Destino	Discreta
CO_D_X	Coordenada X Destino	Numérica
CO_D_Y	Coordenada Y Destino	Numérica
ZONA_T1	Zona da 1ª Transferência	Discreta
MUNI_T1	Município 1ª Transferência	Discreta
CO_T1_X	Coordenada X 1ª Transferência	Numérica
CO_T1_Y	Coordenada Y 1ª Transferência	Numérica
ZONA_T2	Zona da 2ª Transferência	Discreta
MUNI_T2	Município 2ª Transferência	Discreta
CO_T2_X	Coordenada X 2ª Transferência	Numérica
CO_T2_Y	Coordenada Y 2ª Transferência	Numérica
ZONA_T3	Zona da 3ª Transferência	Discreta
MUNI_T3	Município 3ª Transferência	Discreta
CO_T3_X	Coordenada X 3ª Transferência	Numérica
CO_T3_Y	Coordenada Y 3ª Transferência	Numérica
MOTIVO_O	Motivo na Origem	Discreta
MOTIVO_D	Motivo no Destino	Discreta
SERVIR_O	Servir Passageiro na Origem	Discreta
SERVIR_D	Servir Passageiro no Destino	Discreta
MODO1	Modo 1	Discreta
MODO2	Modo 2	Discreta
MODO3	Modo 3	Discreta
MODO4	Modo 4	Discreta
H_SAIDA	Hora Saída	Numérica
MIN_SAIDA	Minuto Saída	Numérica

ANDA_O	Tempo Andando na Origem	Numérica
H_CHEG	Hora Chegada	Numérica
MIN_CHEG	Minuto Chegada	Numérica
ANDA_D	Tempo Andando no Destino	Numérica
DURACAO	Duração da Viagem (em minutos)	Numérica
MODOPRIN	Modo Principal	Discreta
TIPOVG	Tipo de Viagem	Discreta
PAG_VIAG	Quem Pagou a Viagem	Discreta
TP_ESAUTO	Tipo de Estacionamento Automóvel	Discreta
VL_EST	Valor do Estacionamento Automóvel	Numérica
PE_BICI	Por Que Viajou A Pé ou Bicicleta	Discreta
TP_ESBICI	Estacionamento Bicicleta	Discreta
ID_ORDEM	Número de Ordem do Registro	Numérica
DISTANCIA	Distância em metros*	Numérica

APÊNDICE B – Variáveis do Censo Demográfico de 2010

Tabela	Variável	Classes
4.20.1.1	Gênero	Masculino
		Feminino
4.20.1.2	Grupos de idade	0 a 4 anos
		5 a 9 anos
		10 a 14 anos
		15 a 19 anos
		20 a 24 anos
		25 a 29 anos
		30 a 34 anos
		35 a 39 anos
		40 a 44 anos
		45 a 49 anos
		50 a 54 anos
		55 a 59 anos
		60 a 64 anos
		65 a 69 anos
		70 a 74 anos
		75 a 79 anos
		80 a 84 anos
		85 a 89 anos
		90 a 94 anos
		95 a 99 anos
4.20.5.1	Quantidade de domicílios particulares	Total
4.20.5.2	Condição de ocupação do domicílio	Próprio
		Alugado
		Cedido
		Outra
4.20.7.1	Classes de rendimento	Até 1/2
		Mais de 1/2 a 1
		Mais de 1 a 2
		Mais de 2 a 5
		Mais de 5 a 10
		Mais de 10 a 20
		Mais de 20
		Sem rendimento (2)
2.20.5.2	População que frequenta escola ou creche	0 a 3 anos
		4 ou 5 anos
		6 anos

		7 a 14 anos
		15 a 17 anos
		18 a 19 anos
		20 a 24 anos
		25 ou mais anos
2.20.6.1	Condição de atividade	Ocupadas
		Desocupadas
		Não economizamento ativas
2.20.7.3	Tempo de deslocamento até o trabalho	Até 5 minutos
		De 6 minutos até meia hora
		Mais de meia hora até uma hora
		Mais de uma hora até duas horas
		Mais de duas horas
2.20.9.1	Rendimento nominal	Até 1
		Mais de 1 a 2
		Mais de 2 a 5
		Mais de 5 a 10
		Mais de 10 a 20
		Mais de 20
		Sem rendimento (2)

APÊNDICE C – Tabelas de Contingência da Amostra e da População

Tabela C 1: Tabela Idade x Gênero para a amostra

Variáveis		Gênero		
		Homem	Mulher	**Total**
Idade	0 a 4 anos	3277	3231	6508
	5 a 9 anos	5336	5160	10496
	10 a 14 anos	6329	6177	12506
	15 a 19 anos	7952	7261	15213
	20 a 24 anos	8898	8956	17854
	25 a 29 anos	9200	9750	18950
	30 a 34 anos	7771	9228	16999
	35 a 39 anos	7195	8575	15770
	40 a 44 anos	7228	8294	15522
	45 a 49 anos	7087	7931	15018
	50 a 54 anos	6412	7375	13787
	55 a 59 anos	4937	5615	10552
	60 a 64 anos	4044	4577	8621
	65 a 69 anos	2965	3637	6602
	70 a 74 anos	2320	2665	4985
	75 a 79 anos	1562	2070	3632
	80 a 84 anos	971	1297	2268
	85 a 89 anos	376	566	942
	90 a 94 anos	136	242	378
	95 a 99 anos	25	70	95
	> 100 anos	0	0	0
	Total	94021	102677	196698

Tabela C 2: Tabela Idade x Gênero para a população

Variáveis		Gênero		
		Homem	Mulher	**Total**
Idade	0 a 4 anos			710927
	5 a 9 anos			758279
	10 a 14 anos			867430
	15 a 19 anos			842257
	20 a 24 anos			991659
	25 a 29 anos			1074582
	30 a 34 anos			1010076
	35 a 39 anos			888685
	40 a 44 anos			812979
	45 a 49 anos			742720
	50 a 54 anos			667658
	55 a 59 anos			548113
	60 a 64 anos			423055
	65 a 69 anos			302338
	70 a 74 anos			237301
	75 a 79 anos			170969
	80 a 84 anos			119511
	85 a 89 anos			57205
	90 a 94 anos			21234
	95 a 99 anos			5498
	> 100 anos			1027
	Total	5328632	5924871	11253503

Variáveis		Renda								
		Até 1/2	Mais de 1/2 a 1	Mais de 1 a 2	Mais de 2 a 5	Mais de 5 a 10	Mais de 10 a 20	Mais de 20	Sem rendimento	**Total**
Idade	0 a 4 anos	7	1	1	1	0	0	0	2126	2136
	5 a 9 anos	24	22	8	4	0	0	0	5205	5263
	10 a 14 anos	110	29	6	0	0	0	0	6484	6629
	15 a 19 anos	262	1190	784	84	12	0	0	6301	8633
	20 a 24 anos	114	1388	2604	882	123	22	0	5021	10154
	25 a 29 anos	84	958	2408	1847	659	99	6	4837	10898
	30 a 34 anos	105	780	2010	1650	724	222	27	4499	10017
	35 a 39 anos	88	604	1632	1633	790	259	57	4192	9255
	40 a 44 anos	67	571	1353	1555	858	314	61	4206	8985
	45 a 49 anos	65	529	1094	1463	978	408	67	3933	8537
	50 a 54 anos	42	386	993	1248	809	365	96	3645	7584
	55 a 59 anos	33	325	655	981	675	259	67	2447	5442
	60 a 64 anos	25	299	550	701	446	244	88	1927	4280
	65 a 69 anos	12	327	345	498	315	147	33	1363	3040
	70 a 74 anos	5	273	260	387	195	103	43	874	2140
	75 a 79 anos	0	216	171	216	114	86	20	509	1332
	80 a 84 anos	0	92	80	99	62	49	4	298	684
	85 a 89 anos	0	33	21	32	15	13	2	94	210
	90 a 94 anos	0	8	4	5	8	3	1	40	69
	95 a 99 anos	0	0	0	1	0	1	0	3	5
	> 100 anos	0	0	0	0	0	0	0	0	0
	Total	1043	8031	14979	13287	6783	2594	572	58004	105293

Variáveis		Renda								
		Até 1/2	Mais de 1/2 a 1	Mais de 1 a 2	Mais de 2 a 5	Mais de 5 a 10	Mais de 10 a 20	Mais de 20	Sem rendimento	**Total**
Idade	0 a 4 anos									710927
	5 a 9 anos									758279
	10 a 14 anos									867430
	15 a 19 anos									842257
	20 a 24 anos									991659
	25 a 29 anos									1074582
	30 a 34 anos									1010076
	35 a 39 anos									888685
	40 a 44 anos									812979
	45 a 49 anos									742720
	50 a 54 anos									667658
	55 a 59 anos									548113
	60 a 64 anos									423055
	65 a 69 anos									302338
	70 a 74 anos									237301
	75 a 79 anos									170969
	80 a 84 anos									119511
	85 a 89 anos									57205
	90 a 94 anos									21234
	95 a 99 anos									5498
	> 100 anos									1027
	Total	102784	1119360	2340278	1659072	679919	291102	132882	3458900	11253503

Variáveis		Gênero		
		Homem	Mulher	**Total**
Condição de atividade	Ocupadas	36146	30391	66537
	Desocupadas	6625	13327	19952
	Não economicamente ativas	8961	9843	18804
	Total	51732	53561	105293

Variáveis		Gênero		
		Homem	Mulher	**Total**
Condição de atividade	Ocupadas			6383421
	Desocupadas			516389
	Não economicamente ativas			4353693
	Total	5328632	5924871	11253503

APÊNDICE D – Código de Programação em Python para Geração da População Sintética

```
import numpy as np

#define a funcão de criação da população sintética
def popsint():

    #cria arquivo .txt e insere títulos das colunas
    with open('teste.txt', 'w') as f:

f.write('domicilio;veiculos_domicilio;id_pessoa;idade;sexo;atividade\n')

    domicilio = 0
    id_pessoa = 0
    individuos = 0

    # cria contador de cada categoria
    c_sexo = 2 * [0]
    c_idade = 20 * [0]
    c_atividade = 3 * [0]
    c_veiculo_dom = 2*[0]

    #processo iterativo - geracão de domicílios
    while domicilio < 3470566:

        domicilio += 1

        #número de moradores no domicílio
        num_moradores = np.random.choice([1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14], \
                                         p=[0.052, 0.158, 0.237, 0.279,
0.147, 0.064, 0.030, \
                                            0.012, 0.008, 0.007, 0.003,
0.001, 0.001, 0.001])

        chefe = 0

        #domicílio possui veículo individual?
        veiculo_dom = np.random.choice([1,2], p=[0.6806, 0.3194])
        c_veiculo_dom[veiculo_dom-1] +=1

        #processo iterativo - geração de moradores
        for i in range(num_moradores):
            id_pessoa +=1

            #escolha do gênero
            sexo = np.random.choice([1, 2], p=[0.4735, 0.5265])

            #Se o indivíduo não for o primeiro morador (chefe)
            if chefe != 0:

                #dado que é homem, escolher idade e condição de atividade
                if sexo == 1:
                    idade =
np.random.choice([1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20],\
                                             p=[0.0679, 0.0724, 0.0823,
```

```
0.0789, 0.0918, 0.0975, 0.0903, 0.0794,\
                                                0.0723, 0.0642, 0.0566,
0.0458, \
                                                0.0343, 0.0238, 0.0179,
0.0121, 0.0078, 0.0033, 0.0011, 0.0003])

                    atividade = np.random.choice([1, 2, 3], p=[0.6094,
0.0293, 0.3613])

                # dado que é mulher, escolher idade e condição de atividade
                else:
                    idade =
np.random.choice([1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20],\
                                                p=[0.0589, 0.0629, 0.0724,
0.0712, 0.0848, 0.0937, 0.0893, 0.0786,\
                                                0.0722, 0.0676, 0.0617,
0.0514, \
                                                0.0405, 0.0296, 0.0240,
0.0180, 0.0132, 0.0067, 0.0026, 0.0007])
                    atividade = np.random.choice([1, 2, 3], p=[0.5293,
0.0608, 0.4099])

            #Se o indivíduo for o primeiro morador (chefe), sua idade é
maior ou igual que 19 anos
            else:

                # dado que é homem, escolher idade e condição de atividade
                if sexo == 1:
                    idade = np.random.choice([5, 6, 7, 8, 9, 10, 11, 12,
13, 14, 15, 16, 17, 18, 19, 20], \
                                                p=[0.1315, 0.1396, 0.1293,
0.1137, 0.1035, 0.0919, 0.0811, 0.0655, \
                                                0.0492, 0.0341, 0.0256,
0.0173, \
                                                0.0111, 0.0048, 0.0016,
0.0002])
                    atividade = np.random.choice([1, 2, 3], p=[0.6094,
0.0293, 0.3613])

                # dado que é mulher, escolher idade e condição de atividade
                else:
                    idade = np.random.choice([5, 6, 7, 8, 9, 10, 11, 12,
13, 14, 15, 16, 17, 18, 19, 20], \
                                                p=[0.1153, 0.1275, 0.1215,
0.107, 0.0983, 0.0921, 0.0841, 0.0699,\
                                                0.0552, 0.0403, 0.0327,
0.0245,\
                                                0.0180, 0.0091, 0.0035,
0.001])
                    atividade = np.random.choice([1, 2, 3], p=[0.5293,
0.0608, 0.4099])

            chefe += 1

            #contadores das classes
            c_idade[idade-1] +=1
            c_sexo[sexo-1] +=1
            c_atividade[atividade-1] +=1
            individuos +=1
```

```
            #preencher arquivo .txt
            with open('teste.txt', 'a') as f:
                f.write('%i;%i;%i;%i;%i;%i\n' %(domicilio, veiculo_dom,
id_pessoa, idade, sexo, atividade))

    print (c_idade)
    print (c_sexo)
    print (individuos)
    print (c_atividade)
    print (c_veiculo_dom)

#Chama a função popsint
popsint()
```