UNIVERSIDADE DE SÃO PAULO
Escola de Engenharia de São Carlos
Departamento de Engenharia de Transportes
Programa de Pós-Graduação em Engenharia de Transportes

Marcela Navarro Pianucci

Uma proposta para a obtenção da população sintética através de dados agregados para modelagem de geração de viagens por domicílio

São Carlos
2016

Marcela Navarro Pianucci

Uma proposta para a obtenção da população sintética através de dados agregados para modelagem de geração de viagens por domicílio

Tese de Doutorado submetida à Escola de Engenharia de São Carlos, da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de *Doutor em Ciências*, Programa de Pós-Graduação em Engenharia de Transportes. Área de Concentração: Infraestrutura de Transportes

Orientador: Prof. Associado Paulo César Lima Segantine

São Carlos
2016

Este documento é a versão corrigida da tese. A versão original está depositada no Serviço de Pós-Graduação da EESC-USP.

Navarro Pianucci, Marcela
N314u Uma proposta para a obtenção da população sintética através de dados agregados para modelagem de geração de viagens por domicílio / Marcela Navarro Pianucci; orientador Paulo César Lima Segantine. São Carlos, 2016.

Tese (Doutorado) - Programa de Pós-Graduação em Engenharia de Transportes e Área de Concentração em Infraestrutura de Transportes -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2016.

1. População Sintética. 2. Redes Neurais Artificiais. 3. Método Monte Carlo. 4. Demanda por transportes. I. Título.

FOLHA DE JULGAMENTO

Candidata: Tecnóloga **MARCELA NAVARRO PIANUCCI.**

Título da tese: "Uma proposta para a obtenção da população sintética através de dados agregados para modelagem de geração de viagens por domicílio"

Data da defesa: 16/09/2016.

Comissão Julgadora:	**Resultado:**
Prof. Associado **Paulo Cesar Lima Segantine (Orientador)** (Escola de Engenharia de São Carlos/EESC)	Aprovada
Prof. Dr. **André Luiz Barbosa Nunes da Cunha** (Escola de Engenharia de São Carlos/EESC)	APROVADA
Prof. Dr. **Bruno Vieira Bertoncini** (Universidade Federal do Ceará/UFC)	Aprovada
Profa. Dra. **Anabela dos Santos Aleixo Simões** (FMH-Universidade de Lisboa/Portugal)	Aprovada
Profa. Dra. **Suely da Penha Sanches** (Universidade Federal de São Carlos/UFSCar)	APROVADA

Coordenadora do Programa de Pós-Graduação em Engenharia de Transportes:
Prof. Dra. **Ana Paula Camargo Larocca**

Presidente da Comissão de Pós-Graduação:
Prof. Associado **Luis Fernando Costa Alberto**

Aos meus pais, Cristina e Carlos,
pela vida e oportunidades que me deram.

Agradecimentos

A Deus, por tudo que tem me proporcionado.

Ao Érico Guerreiro, pela dedicação, incentivo e companheirismo, sempre ao meu lado neste trabalho.

À minha querida família: minha irmã Marília, minha mãe Cristina e meu pai Carlos, pelo apoio, incentivo e compreensão.

Agradeço especialmente ao professor Paulo César Lima Segantine pela amizade e paciência ao longo de cinco anos. Um exemplo de ética, comprometimento e competência com seu trabalho.

Aos professores, Cira Souza Pitombo e André Cunha pela disponibilidade de sempre e pelas ajudas fundamentais para o desenvolvimento desta tese. Agradeço sinceramente pela amizade, confiança e por acreditarem em mim.

Ao professor Bruno Vieira Bertoncini (Universidade Federal do Ceará) pelas sugestões ao trabalho no exame de qualificação e no encontro da ANPET.

Ao professor Antônio Nelson Rodrigues da Silva pela cessão dos dados da pesquisa origem e destino 2007/2008 da cidade de São Carlos-SP.

À professora Ana Paula Larocca pela amizade, conselhos e carinho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de doutorado.

À todos os colegas do STT pela agradável convivência, em especial aos queridos amigos Andressa, Artur, Diego, Fernando, Luiz Henrique, Luiz Miguel, Magaly, Monique, Nil, Piva,

Simone, Thaís e Vivian. O companheirismo e a amizade de vocês foram essenciais. Agradeço especialmente a Gabi pela amizade, carinho e confiança.

A todos os funcionários e professores do STT, pelo apoio e ensinamentos tão preciosos.

Aos amigos Maza e Widmer pela atenção, cuidado, carinho e amizade.

Finalmente, a todos que contribuíram de alguma forma para o desenvolvimento e conclusão deste trabalho.

"O segredo de progredir é começar."
(Mark Twain)

Resumo

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.

A estimativa de viagens por domicílio é fundamental para a tomada de decisões relativas ao planejamento urbano e de transportes. Em geral, a obtenção destas informações é por meio de modelos tradicionais como o modelo clássico de quatro etapas, e a primeira etapa do modelo é a geração de viagens. Entretanto, modelos clássicos apresentam inúmeras falhas, muitas delas relacionadas as suposições prévias matemáticas (normalidade ou continuidade da variável dependente). Desta forma, surge a necessidade de testar outras técnicas de modo a minimizar as falhas apresentadas pelos modelos clássicos e utilizá-las como uma ferramenta auxiliar, como por exemplo, as Redes Neurais Artificiais (RNAs), que podem ser aplicáveis na modelagem de problemas complexos e não lineares na área de engenharia de transportes, pois apresentam capacidade de aprendizagem, adaptação e generalização. Assim, para estimar viagens por domicílio, seja pela modelagem tradicional ou pela modelagem RNA são necessários dados desagregados dos domicílios, incluindo dados dos indivíduos, como as atividades diárias que exercem e dados sociodemográficos, etc. Esses dados são geralmente obtidos por uma Pesquisa O/D, que fornece um banco de dados detalhado sobre o comportamento de viagem da população de uma cidade. No entanto, a maioria das cidades enfrenta problemas para a aquisição desses dados, uma vez que este tipo de pesquisa possui alto custo de preparação, execução, processamento e análise. Portanto, percebe-se a necessidade de novas propostas de procedimentos que forneçam dados confiáveis e com baixo custo, para estimar a demanda por

viagens, capazes de gerar resultados com rapidez, qualidade e acurácia e sem a necessidade dos dados provenientes de uma Pesquisa O/D. Devido a dificuldade de aquisição de dados desagregados, foi proposto neste trabalho, a geração da população sintética com dados agregados a partir da aplicação do Método Monte Carlo. Este trabalho tem por objetivo gerar uma população sintética baseada em dados censitários agregados e testar a adequabilidade das RNAs para estimar viagens produzidas por domicílio. Neste estudo, a modelagem tradicional foi utilizada para comparar os resultados obtidos com a modelagem RNA, pois o objetivo não foi checar minuciosamente a qualidade dos modelos lineares, e sim, testar a adequabilidade das RNAs para estimar viagens por domicílio. A abordagem tradicional se baseou em um modelo de regressão linear enquanto que a abordagem de redes neurais consistiu da rede perceptron multicamadas. Na execução do trabalho foram calibrados quatro modelos (dois de cada abordagem) com os dados desagregados da Pesquisa O/D e foram comparados os resultados obtidos de cada abordagem. Ao final do trabalho, foi possível escolher o modelo mais adequados de cada abordagem e em seguida, foram utilizados para prever viagens produzidas por domicílio com os dados obtidos pela população sintética. Os resultados indicaram que 70% das variáveis obtidas na população sintética foram consideradas aptas para o estudo e que a estimativa de viagens por domicílio da população sintética obtida em ambos os modelos (Modelo 3-RNA) e (Modelo 4-RLM) obtiveram uma boa previsão, ou seja, mais de 70% das viagens produzidas por domicílio da população sintética foram consideradas válidas. Isso demonstrou que, o uso de da modelagem RNA é uma técnica alternativa eficiente e promissora na área de planejamento de transportes, especificamente para a previsão de viagens produzidas por domicílio.

Palavras-chave: População Sintética. Redes Neurais Artificiais. Método Monte Carlo. Demanda por transportes.

Abstract

Pianucci, M. N. **A proposal to obtain a synthetic population through aggregated data to model the number of trip productions per household**. 186 p. Ph.D. Thesis – São Carlos School of Engineering, University of São Paulo, 2016.

The estimated number of household travels is essential in the decision-making process related to urban and transportation planning. Usually, this information is obtained through traditional models, such as four-step classic model, for example, which has trip generation as a first step. However, classic models feature numerous failures. Many of these failures are related to mathematical prior assumptions (normality or continuity of the dependent variable). Thus, it is important to test other techniques in order to reduce the failures and use these techniques as an auxiliary tool, i.e. Artificial Neural Networks (ANN). ANN are applicable in the modeling of complex and nonlinear transportation problems, due to its learning, adaptation and generalization capacities. Thus, to estimate the number of household travel, either by traditional or by ANN models, it is required disaggregated data of the households. It might include information of individuals, as daily activities and sociodemographic information. Usually, these data are obtained by a O/D survey, which provides a detailed database of the population travel behavior of the city. However, the obtainment of this information leads to high costs of preparation, execution, processing and analysis of the data. Thus, most cities have faced problems to attain this information. Therefore, new methods of estimation providing reliable data and low cost, are required. It will enable to estimate the demand of travel, rapidly with quality and accuracy, without the need of data provided through an O/D survey. Due to the difficulty of acquiring disaggregated data, this study proposes the generation of synthetic population through aggre-

gated data by applying the method of Monte Carlo. This study aims to generate a synthetic population based on aggregated census data, and test the suitability of ANN to estimate the number of household travels. Since the aim was not thoroughly check the quality of linear models, instead, test the suitability of ANN to estimate the number of household travels, obtained results of traditional and ANN models were compared. The traditional approach was based on a linear regression while the neural network consisted of Multilayer Perceptron network. Four models (two of each approach) were proposed and calibrated with disaggregated data of an O/D Survey. Then, the results were compared. It enabled to choose the most appropriate model of each approach. Hence, these models were used to forecast the number of trip productions per household, using the data obtained by the synthetic population proposed. The results indicated that 70% of the variables obtained through the synthetic population, were considered suitable for the study. Besides, the estimated number of household travels of the synthetic population obtained for both models (Model 3-RNA and 4-MLR model) presented a good forecast, indicating that more than 70% of household travels of the synthetic population were considered valid. Finally, it is concluded that the use of ANN modeling is an effective and promising alternative technique in the transportation-planning field, specifically to forecast the number of trip productions per household.

Keywords: Synthetic Population. Artificial Neural Networks. Monte Carlo method. demand for transport.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

INTRODUÇÃO

O objetivo desta introdução é responder às indagações básicas para o desenvolvimento da pesquisa: (1) Qual o tema do trabalho? (breve revisão bibliográfica de forma a contextualizar o tema), (2) Quais os objetivos do trabalho?, (3) Como atingir estes objetivos? (descrição sucinta do método) e (4) Porque eles são importantes? (contribuições para a área de transportes).

1.1 Uma breve revisão bibliográfica - Contextualizando o tema

O conhecimento da demanda por transportes de uma cidade é fundamental para entender os motivos da demanda e como eles interagem e afetam a evolução do volume de tráfego. Pois, no processo de planejamento dos transportes é verificado o valor das demandas atuais e realizado uma previsão dos valores futuros e, a partir daí, é recomendado a ampliação ou criação de infraestruturas como, por exemplo, a implantação numa cidade de uma nova linha de ônibus ou um novo modo de transporte coletivo. No entanto, a previsão da demanda de viagens não é uma tarefa tão simples como parece, pois envolve a solução de problemas complexos, tais como: recursos escassos e ausência de dados.

Para auxiliar essa tarefa e realizar estimativas precisas da demanda por viagens, normalmente, são usados modelos de transportes. A aplicação de modelos auxilia nos processos de decisão e planejamento, principalmente quando se trata de fatos futuros, ou seja, estimar a variação ocorrida no volume de tráfego, ao longo do tempo e projetá-la para uma data futura. Um modelo serve, por exemplo, para especificar uma função de demanda que represente o volume de viagens que ocorrem entre um par de origem e destino por um determinado modo de viagem, em função de suas características.

Os modelos de transporte geralmente necessitam de uma base de dados fornecida pelas pesquisas denominadas de Pesquisas Origem e Destino (O/D) para estimativas de parâmetros. Essas pesquisas, geralmente são realizadas por meio de entrevista domiciliar, entrevista telefônica ou por meio da Internet, pesquisa *Cordon-line* ou *Screen-line*, pesquisa embarcada e pesquisa baseada em atividades diárias. No entanto, as pesquisas domiciliares são utilizadas para conhecer informações relevantes sobre o comportamento da população relativo a viagens, os modos de transporte utilizados, os motivos das viagens, os polos de geração e atração de viagens e até dados socioeconômicos (SILVA, 2008). Com esses dados é realizada a calibração dos modelos, cujo objetivo é estimar os parâmetros para reproduzir os valores da variável dependente quando se substitui nos modelos as variáveis independentes correspondentes.

Apesar da sua grande importância, a Pesquisa O/D, normalmente não é realizada periodicamente nas cidades, pois se trata de uma atividade de alto custo de preparação, execução, processamento e análise dos dados. Esses fatos dificultam o planejamento, pois os dados, geralmente, encontram-se defasados no tempo.

Uma maneira de resolver o problema da falta de dados e da sua periodicidade é lançar mão de dados sintéticos. Uma forma de obter esses dados é através da geração de uma população sintética. Entende-se por população sintética um conjunto de dados referentes a uma população artificial que, em termos estatísticos, representa a população real do fenômeno a ser simulado. As populações sintéticas podem ser geradas a partir de dados agregados ou desagregados de um censo, tentando obter a maior precisão possível com a população verdadeira. No Brasil os dados agregados contêm informações relativas ao morador e ao uso do solo, porém essas informações estão agregadas por setores censitários. Esses dados estão disponíveis no site do Instituto Brasileiro de Geografia e Estatística (IBGE) e são atualizados a cada dez anos.

A população sintética pode ser gerada por diversas técnicas, sendo que cada uma possui suas limitações. Nos últimos anos, três métodos vêm sendo muito utilizadas para a geração deste tipo de população, são eles: (1) Método de ajuste proporcional iterativo (IPF, de acordo com as iniciais de seu nome em inglês - *Iterative Proportional Fitting*) desenvolvida por Deming e Stephan (1940), (2) Método Monte Carlo (MMC), probabilidade condicional detalhado no trabalho de Birkin e Clarke (1988) e (3) Método da otimização combinatória (CO), de acordo com as iniciais de seu nome em inglês - *Combinatorial Optimisation*) aplicados nos trabalhos de Williamson, Birkin e Rees (1998) e Voas e Williamson (2001).

Entretanto, mesmo com as dificuldades na obtenção dos dados e da necessidade de obtenção de dados desagregados por indivíduo ou domicílio, o advento da microssimulação foi importante para a modelagem desagregada da demanda por viagens, pois permite uma simulação mais detalhada e precisa da previsão de viagens quando comparadas aos modelos tradicionais de agregação.

Com base na literatura, o uso de RNAs vem se tornando nas últimas décadas cada vez mais popular no planejamento de transportes dado o avanço computacional. As RNAs são apli-

cáveis em diversas áreas da Engenharia de Transportes, incluindo modelagem de problemas complexos, como, por exemplo, a geração de viagens, isso devido à sua grande habilidade de classificar e reconhecer padrões em banco de dados.

Desta forma, será testado nesta pesquisa, um procedimento por meio de dados agregados para a geração da população sintética associada à técnica de modelagem de redes neurais artificiais (RNAs) para a previsão da produção de viagens por domicílio.

Os modelos serão gerados utilizando esta técnica por apresentarem melhor desempenho que os modelos tradicionais, como relatados nos trabalhos de Bocanegra (2002) e Lin, Zito e Taylor (2005). Além disso, a técnica de RNAs é mais robusta, pois não necessita testar as suposições estatísticas (normalidade, homocedasticidade, linearidade e ausência de erros correlacionados) para a modelagem, ao contrário da técnica de modelos tradicionais, como a regressão linear múltipla (RLM), que essas suposições afetam diretamente a qualidade do modelo. Vale ressaltar que, as viagens produzidas por domicílio também serão estimadas pela técnica de regressão linear múltipla apenas como procedimento de validação do método proposto.

1.2 Objetivos

O objetivo principal da pesquisa proposta é a geração de uma população sintética baseada em dados censitários agregados para modelar número de viagens produzidas por domicílio.

Há um objetivo secundário, associado ao principal, que é testar a adequabilidade da técnica das RNAs para estimar demanda por viagens para amostra real e sintética.

1.3 Síntese do método e estrutura do trabalho

O problema de pesquisa desta tese está vinculado às limitações identificadas na aplicação prática dos modelos de transporte no processo de tomada de decisões no planejamento. A dificuldade de obter dados totalmente desagregados devido ao alto custo, à falta de periodicidade das pesquisas de tráfego, Pesquisas O/D ou até mesmo da inexistência e o uso de variáveis que possuam relação causal com o fenômeno são os motivos que perturbam o emprego desses modelos na área de planejamento de transportes.

Baseando-se na revisão preliminar da literatura e no problema de pesquisa, chegou-se à seguinte pergunta de pesquisa:

Como gerar dados totalmente desagregados e utilizá-los para modelar número de viagens produzidas por domicilio?

Analisando-se esta pergunta é adotada uma sequência de processos a fim de verificar as duas hipóteses desta tese, as quais são anunciadas da seguinte forma:

- É possível gerar uma população sintética através de dados agregados.

• É possível utilizar os dados sintéticos para estimar número de viagens produzidas por domicílio.

Este texto, aqui apresentado, é formado por sete capítulos vide ilustração da Figura 1.1, além desta Introdução, conforme descrito a seguir:

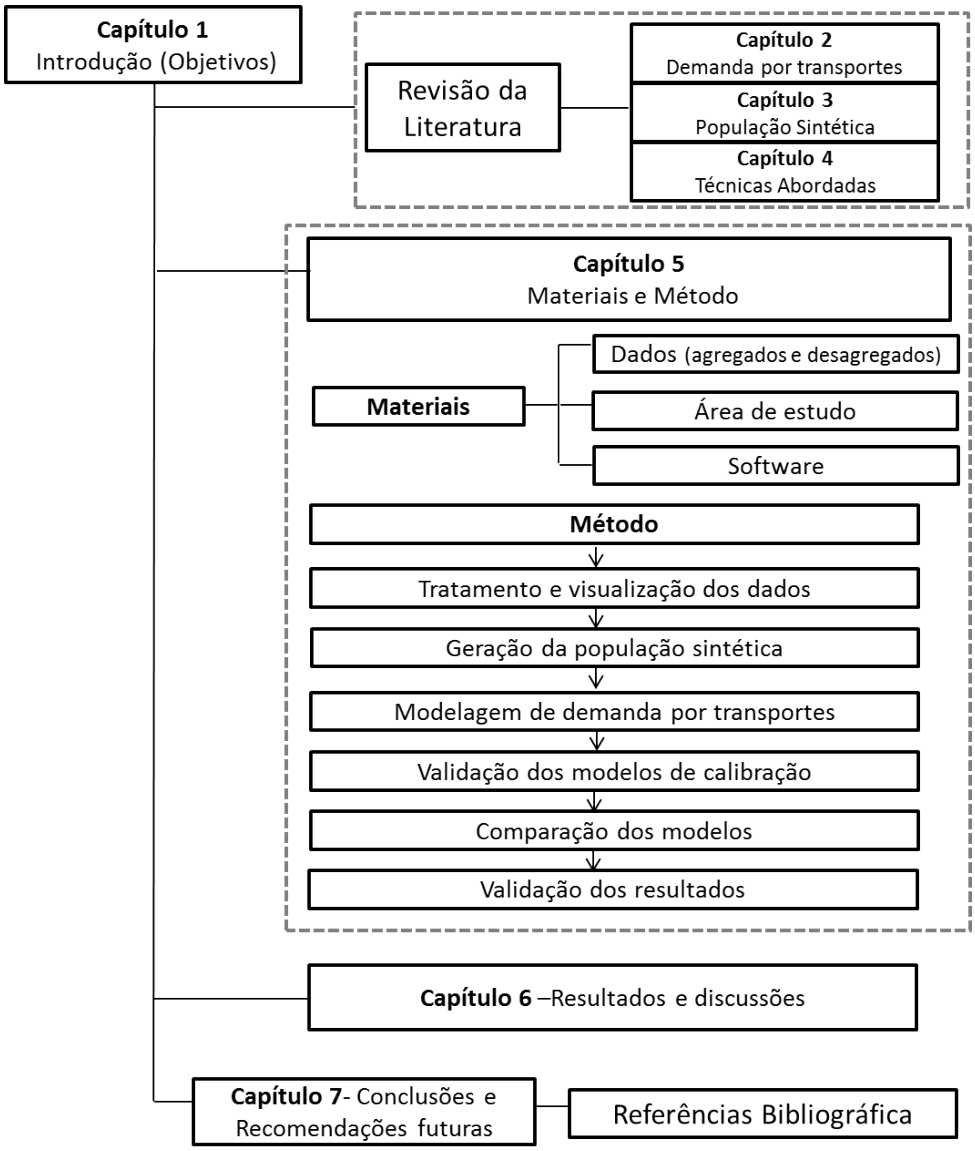


Figura 1.1 – Ilustração da estrutura do formato desta tese.

Os Capítulos 2, 3 e 4 trazem a etapa de revisão da literatura, que corresponde à elaboração do corpo conceitual do desenvolvimento desta tese. O Capítulo 2 mostra, sucintamente, uma contextualização e conceitos básicos sobre demanda por viagens. Também é apresentado o modelo de geração de viagens através de trabalhos publicados na literatura. O Capítulo 3 apresenta as técnicas de geração da população sintética e os trabalhos que a utilizaram na solução de problemas de transportes. Por fim, no Capítulo 4 são apresentadas as duas técnicas utilizadas para o desenvolvimento desta tese: Monte Carlo e Redes Neurais Artificiais, além de uma comparação das técnicas: Estatística e RNAs.

O Capítulo 5 apresenta as etapas que levaram a verificação da hipótese inicial. Neste capítulo são explanados os materiais, o software e o método utilizados para a obtenção da estimativa da variável objeto de estudo (*viagens produzidas por domicílio*), por meio de modelagem de redes neurais artificiais e tradicional (regressão).

Em seguida, a análise e discussão dos resultados é apresentada no Capítulo 6. Neste capítulo são obtidas as viagens produzidas por domicílio da população sintética através dos modelos (RNA e RLM) obtidos no Capítulo 5. Por meio de técnicas estatísticas são comparados os resultados e validadas as viagens por domicílio da população sintética estimadas pelos modelos de RNAs e de RLM.

Finalmente, o Capítulo 7 apresenta as conclusões obtidas neste estudo com o objetivo de corroborar as hipóteses desta pesquisa. Também são descritas as recomendações para trabalhos futuros. São apresentadas no final deste texto as referências bibliográficas que fundamentaram o estudo.

1.4 Justificativa

Para modelar a demanda por viagens são necessários dados dos modos e dos motivos de deslocamentos dos habitantes, as atividades diárias que exercem e os dados sociodemográficos, entre outros. Geralmente, esses dados são obtidos por uma Pesquisa O/D, que fornece um banco de dados detalhado sobre o comportamento de viagem da população de uma cidade.

No entanto, a maioria das cidades brasileiras enfrenta problemas para a aquisição desses dados, uma vez que este tipo de pesquisa requer alto custo de preparação, execução, processamento e análise. Como resultado, os municípios, quando realizam tais pesquisas, fazem sempre com baixa frequência, pois geralmente os recursos para o planejamento são escassos. Devido à falta de dados atualizados, os planejadores de transporte fazem uso de dados antigos que podem conduzir a erros e não representar mais a realidade urbana para o momento de interesse.

Portanto, percebe-se a necessidade de novas propostas de procedimento de obtenção de dados, que forneçam dados confiáveis e com baixo custo, para estimar a demanda por viagens, capazes de gerar resultados com rapidez, qualidade e acurácia e sem a necessidade dos dados provenientes de Pesquisa O/D.

Apesar do IBGE disponibilizar a cada dez anos dados sociodemográficos do morador, esses dados estão agregados por setores censitários. Embora o IBGE também fornecer dados desagregados por domicílio, esse dados são referentes a uma pequena parcela da população, apenas 10% da população total. Uma solução para a obtenção de dados totalmente desagregados da população total é gerar uma população sintética a partir desses dados agregados.

Para se estimar demanda por viagens, os modelos tradicionais que utilizam a técnica de RLM são os mais utilizados, porém não seria a técnica mais indicada para a previsão de produção

de viagens, pois apresentam inúmeras falhas e muitas delas relacionadas a suposições prévias matemáticas (normalidade, homocedasticidade, linearidade). Pois, por exemplo, viagens são dados de contagem e não atendem à suposição de normalidade ou a relação existente entre produção de viagens e variáveis explicativas, como a população, não é linear, etc.

Nas últimas décadas, com o avanço computacional, as RNAs são aplicáveis na modelagem de problemas complexos e não lineares de transportes devido a capacidade de aprendizagem, adaptação e generalização. Além disso, esta técnica não necessita de análise prévia das suposições apresentadas pela técnica de regressão linear. Portanto, neste trabalho a estimativa do número de viagens produzidas por domicílio da população sintética será obtida utilizando as duas técnicas RNA e RLM, lembrando que, a técnica de regressão será utilizada apenas para verificar se a técnica de RNA é adequada para a previsão do número de viagens produzidas por domicílio.

O trabalho aqui exposto procura contribuir no fato de que, por meio da geração da população sintética com dados agregados facilmente disponíveis, será possível obter dados desagregados da população total para o ano do último censo. A obtenção desses dados desagregados permite modelar o número de viagens produzidas por domicílio da população da área de estudo através do uso da técnica de RNAs. Vale ressaltar ainda a facilidade da utilização do método proposto em outras áreas de estudo semelhantes à cidade de São Carlos-SP, sobretudo àquelas que não possuem Pesquisa O/D.

Capítulo 2

DEMANDA POR TRANSPORTES

Este capítulo apresenta uma revisão bibliográfica relativa à demanda por transportes. Inicialmente, trata da importância de tais estudos e define, sucintamente, o Modelo Quatro Etapas. Posteriormente é descrita a primeira etapa do modelo, etapa de geração de viagens, as técnicas mais utilizadas e os trabalhos recentes que incluem a utilização de técnicas mais avançadas comparadas à RLM e Classificação Cruzada.

2.1 Contextualização e conceitos básicos

A análise adequada da demanda por transporte é um dos grandes desafios do Brasil e do mundo, pois a mesma é totalmente dependente das características físicas e socioeconômicas da região a ser estudada e quaisquer mudanças nessas variáveis têm efeitos diretos sobre os deslocamentos dos indivíduos. Portanto, a análise e previsão da demanda têm como objetivo: auxiliar nas tomadas de decisão quanto às mudanças que se fazem necessárias no sistema de transporte.

Segundo Ortúzar e Willumsen (2011), os problemas e as técnicas de planejamento de transportes sofreram profundas mudanças a partir da década de 80. Nas grandes cidades dos países desenvolvidos nas décadas de 60 e 70, problemas como congestionamentos, poluições ambientais e acidentes viários já existiam e permanecem até os dias de hoje. Isso caracteriza que um planejamento deficiente, de curto prazo, com investimentos limitados e pouco conhecimento de planejamento estratégico e tomada de decisão são complexos de se resolver e devem ser planejados com os devidos cuidados com o intuito de evitar desperdícios de recursos públicos.

Com a evolução das técnicas de planejamento na década de 1980, as estratégias de modelagem de transportes e as tomadas de decisão tornaram-se ainda mais importantes no processo

de planejamento urbano. O aparecimento da computação de alta capacidade de processamento a baixo custo, deixou de ser problema para a modelagem de problemas de transportes.

Para ajudar nos processos de decisão e planejamento, principalmente quando se trata de análise de fatos futuros, utilizam-se "modelos". Um modelo representa, de maneira simplificada, um fenômeno real, e procura se concentrar em certos elementos para facilitar uma análise particular. Muitos experimentos e incrementos de técnicas foram desenvolvidos na área de transportes para chegar a uma estrutura de modelo a qual foi chamada de Modelo Clássico de Transportes (ORTÚZAR; WILLUMSEN, 2011).

Dentro do modelo clássico da demanda por transporte, o chamado Modelo Quatro Etapas, a geração de viagens constitui a primeira das etapas. Nesta etapa são determinados, através das variáveis socioeconômicas da população, os totais de viagens produzidas e atraídas em cada zona da área de estudo. Essa etapa pode ter dois enfoques: o agregado e o desagregado. As três etapas seguintes do modelo distribuem, respectivamente, as viagens para seus destinos particulares, após é modelada a escolha do modo e por fim, as viagens de cada modo são alocadas a rede viária (ORTÚZAR; WILLUMSEN, 2011).

Os primeiros modelos de análise da demanda por transportes foram desenvolvidos para promover uma infraestrutura que atendesse a uma demanda crescente. Hoje, o enfoque é outro, pois com o aumento da demanda por transporte motorizado, principalmente por automóveis, surgiu a necessidade de melhoramento da infraestrutura e gerar locais de estacionamento, e como consequências degradaram a qualidade de vida em muitas áreas urbanas. Portanto, começa-se a pensar em restringir ao invés de atender a demanda.

Devido a isto, surgem necessidades de encontrar uma forma de alterar os hábitos da população, como por exemplo, diminuir o uso do transporte particular e aumentar o uso do transporte público decorrentes aos problemas já gerados como por exemplo, o congestionamento, a emissão de poluentes, etc.

As viagens que o indivíduo faz estão relacionadas com o tipo de domicílio e o local em que este mora. As necessidades básicas do domicílio (estudo, compras, trabalho, etc.) são encontradas em todas as famílias o que diferencia é se no domicílio a pessoa mora sozinha, na qual ela tem a liberdade de organizar sua agenda, se tem a presença de idosos, que neste caso, idosos tendem a reduzir a mobilidade e de crianças, que têm suas necessidades próprias mas, ao mesmo tempo, dependem da movimentação dos adultos. Desta forma, uma população apresenta uma heterogeneidade nos padrões de atividades e viagens.

Assim, as características de cada domicílio não podem ser utilizadas através da modelagem agregada. Embora estes tipos de modelagens sejam úteis para o entendimento do sistema de transportes como um todo, não se consegue obter informações relativas aos usuários deste sistema, que são características importantes e melhor representam a ampla variabilidade dos comportamentos dos domicílios residentes numa mesma zona. Portanto, de acordo com Bilt

(2002), só é possível analisar os problemas relacionados à demanda crescente por transporte por meio da utilização de modelos desagregados devido à complexidade dos fatores que a influenciam.

2.2 Modelos de geração de viagens

Esta tese aborda somente a primeira etapa do Modelo Quatro Etapas, a etapa de geração de viagens, por isso uma ênfase é dada a esses modelos, sendo apresentada nesta seção uma breve fundamentação teórica e trabalhos desenvolvidos a respeito dos modelos de geração de viagens. Esta fase é essencial para o planejamento de transportes, pois é nela que se analisam os principais fatores que determinam a geração de viagens em uma zona ou domicílio.

Os modelos de geração de viagens são usados para estimar o número total de viagens (atraídas e produzidas) numa zona de tráfego, em função das variáveis capazes de explicá-las. Basicamente, são variáveis explicativas quanto ao uso do solo, das características socioeconômicas nas zonas da área de estudo e das características do sistema de transportes (MORLOK, 1978).

Segundo Donnelly et al. (2010) a primeira fase do Modelo Quatro Etapas é a geração de viagens e a abordagem tradicional, que é baseada nas viagens, ainda é a prática padrão no planejamento de transportes, apesar de existirem abordagens mais avançadas, que examinam de forma mais realistas, os comportamentos nos estudos de demanda de viagens, como por exemplo, os modelos de viagens baseados em atividades.

Os modelos de geração de viagens são elaborados de relações funcionais entre as viagens geradas e as variáveis explicativas, de modo que a demanda por viagem em um cenário futuro possa ser estimada com precisão através do conhecimento das variáveis explicativas nesse cenário (AMAVI et al., 2014).

Embora na literatura encontram-se sugeridas diferentes técnicas para esta etapa, os modelos mais usuais para representação da geração de viagens são, nomeadamente: (1) Fator de Crescimento, (2) Classificação Cruzada e (3) Regressão Linear Múltipla. Segundo Chang et al. (2014) os dois últimos modelos têm apresentado um nível aceitável de desempenho no planejamento de transportes.

Apesar do método de Classificação Cruzada ter um desempenho cabível na solução de problemas de transportes, o método mais popular utilizado em modelos de previsão da geração de viagens é a Regressão Linear Múltipla (RLM).

A escolha adequada do modelo na etapa de geração de viagens é fundamental, pois é a primeira etapa do Modelo Quatro Etapas e erros aqui são levados durante todo o procedimento e pode invalidar o trabalho em etapas posteriores (ORTÚZAR; WILLUMSEN, 2011). A Tabela 2.1 apresenta as características de cada método, os dados que podem ser utilizados e os problemas apresentados na modelagem de geração de viagens.

Tabela 2.1 – Características dos três modelos mais utilizados na geração de viagens.

Modelos	Representação	Dados	Problemas	Referências
Fator de Crescimento	Dados atuais multiplicados pelo Fator de Crescimento. Modelo Fratar $Q_{ij}^{t} = Q_{ij}^{0} \times F_i \times F_j \times L_i$ Onde: Q_{ij}^{t} = nº de viagens no ano t de i para j; Q_{ij}^{0}=nº de viagens atuais; F_i= fator de crescimento da zona de origem i; F_j= fator de crescimento da zona de destino j; L_i= fator de ajuste das origens.	Agregados	Estimativa do fator de crescimento, pois em alguns casos pode superestimar o número total de viagens.	Ortúzar e Willumsen (2011)
Classificação Cruzada	As viagens são agrupadas e relacionadas a estrutura e condições socioeconômicas de cada família.	Desagregados por categorias de domicílios.	O método necessita de grande quantidade de dados e falta de um método eficaz para a escolha das variáveis de classificação, ou para escolher melhores agrupamentos de uma determinada variável, tornando a formação dos estratos domiciliares muitas vezes arbitrárias.	Campos (2007) Ortúzar e Willumsen (2011)
Regressão Linear Múltipla	Relação linear entre o nº de viagens existentes (variável dependente) e os vários fatores que influenciam as viagens (variáveis independentes). $q_n^* = \beta^T x_n + \varepsilon_n, \quad n=1,2,\ldots,N$ Onde: q*=taxa de geração de viagens; n=índice a enésimas observações (qual é o domicílio no estudo); β=vetor dos parâmetros a serem estimados; x= vetor das variáveis independentes e ε= um erro aleatório.	Agregados por zonas de tráfego ou desagregados por domicílios ou indivíduos.	Necessidade de testar suposições estatísticas (normalidade, homocedasticidade, linearidade e ausência de erros correlacionados).	Mello (1975) Chang *et al.* (2014) Hair (2004)

Segundo Rassafi, Rezaei e Hajizamani (2012), a realização de um modelo de previsão de viagens inicia-se através de um processo de calibração para estimar os parâmetros (coeficientes das variáveis independentes). Esses parâmetros são estimados com dados atuais da variável dependente e variáveis independentes. Após a obtenção dos valores previstos das variáveis independentes são obtidas as viagens futuras. Esta previsão pode ser feita com o pressuposto de que a relação entre variáveis explicativas e variáveis dependentes permanece inalterada para o ano de planejamento. Para avaliar o desempenho do modelo obtido é necessário testar e validar, e então o modelo estará pronto para gerar os valores futuros da variável dependente.

Alguns fatores devem ser considerados quando se utiliza a regressão linear para a modelagem de geração de viagens. No caso do uso de modelos de regressão na previsão de viagens por

zonas, o modelo só conseguirá explicar a variação no comportamento das viagens entre zonas. Outro fator é em relação as zonas que não contenham informações sobre certas variáveis dependentes, por exemplo, a produção de viagens de base domiciliar em zonas não residenciais são chamadas de "zonas nulas" e devem ser excluídas da análise, pois zonas que não fornecem dados úteis tendem a produzir coeficientes que superestimam a precisão dos dados estimados pela regressão (ORTÚZAR; WILLUMSEN, 2011).

Além disso, outros fatores importantes a serem considerados no uso da regressão na modelagem da geração de viagens são citados na literatura, como: o número de viagens é tratado como uma variável aleatória contínua, embora seja um dado discreto, a variável dependente pode assumir valores negativos devido à suposição de distribuição normal para queda das taxas de viagens e o modelo não representa o comportamento dos viajantes, pois corresponde a uma relação estatística entre a variável dependente e um conjunto de variáveis independentes (Schmöcker et al. (2005); Badoe (2007); Roorda et al. (2010)).

De acordo com Rassafi, Rezaei e Hajizamani (2012) apesar das diversas aplicações da RLM nos modelos de geração de viagens, ela apresenta limitações, como: são necessários outros métodos para prever os valores das variáveis independentes para o período de previsão e na medida em que as previsões dos valores das variáveis independentes são errôneas, as previsões baseadas em regressão serão imprecisas, ou seja, são necessários dados exatos das variáveis para a previsão futura e não há garantia de que a relação passada entre a variável dependente e independentes continuará inalterada no futuro.

Nos modelos matemáticos de geração de viagens, cada motivo da viagem é associado às características demográficas, tais como: população, domicílios, emprego, posse de veículos e renda. Essas informações podem ser obtidas a partir de pesquisas domiciliares ou relatórios de censo. Para informações futuras são necessárias projeções (CHANG et al., 2014).

Assim, nota-se que não só a escolha adequada da técnica para a modelagem de geração de viagens é importante, mas também a escolha dos dados para o desenvolvimento dos modelos. A variável dependente para a geração de viagens é sempre viagens de veículos, mas as variáveis independentes irão variar dependendo do tipo de uso do solo e do tipo e local do domicílio (MOORE, 2013).

As abordagens dos primeiros modelos eram agregações zonais, porém, hoje vem sendo muito utilizado os modelos desagregados (modelos de geração de viagens por domicílios). A Tabela 2.2, apresentada a seguir, traz uma síntese das variáveis frequentemente utilizadas na modelagem tradicional de geração de viagens (ORTÚZAR; WILLUMSEN, 2011).

Entretanto, são diversos trabalhos de geração de viagens encontrados na literatura que enfatizam a importância da escolha das variáveis e da técnica adequada para o desenvolvimento dos modelos de previsão de viagens.

Tabela 2.2 – Variáveis mais utilizadas nos modelos de geração de viagens.

Variáveis explicativas	Produção de viagens	Atração de viagens
Análise agregada	população densidade renda média população ocupada número de estudantes taxa de motorização	empregos acessibilidade escolas presença de polos geradores de viagem (PGV)
Análise desagregada	renda posse de carro tamanho da família estrutura familiar	

Segundo, Amavi et al. (2014) apresentam em seu trabalho que algumas variáveis são fundamentais para a determinação do número de viagens por domicílio. Por exemplo, a renda familiar é uma das características mais importantes para determinar o número de viagens por domicílio e o modo utilizado, pois quanto maior a renda domiciliar mais viagens em um determinado período de tempo e maior o número de viagens de carro; o tamanho da família tem um efeito positivo, pois há mais viagens conforme o aumento da família; a posse de carro está relacionada com a renda e o tamanho da família; a ocupação dos membros do domicílio também influencia a geração de viagens, pois a ocupação do chefe determina o nível de renda do domicílio e quanto mais membros trabalhando maior número de viagens podem ser geradas no domicílio. Portanto, pode-se concluir que a geração de viagens varia de acordo com as características dos membros que compõe a família.

Chang et al. (2014) propuseram um modelo de geração de viagens para a região metropolitana da cidade de Seul, Coréia do Sul e utilizaram dados de viagens domiciliares (características do domicílio, do transporte e do uso do solo) obtidos por uma pesquisa de tráfego. Compararam a abordagem tradicional (RLM) com métodos mais avançados (Modelo Tobit, Modelo Poisson, Modelo Logit Ordenado e Análise de Classificação Múltipla). Como resultados, os métodos tradicionais apresentaram desempenhos aceitáveis para o uso no planejamento, apesar de suas limitações. Foram identificados alguns problemas na regressão como: a probabilidade da taxa de viagem negativa, a natureza contínua das taxas de viagens e a falta de incorporação das características do comportamento do viajante. Os autores concluíram que os métodos avançados não necessariamente fornecem um melhor desempenho que os métodos tradicionais na estimativa de geração de viagens.

Outro exemplo é o trabalho desenvolvido por Rassafi, Rezaei e Hajizamani (2012), que propuseram um método alternativo (Sistema *Fuzzy*) para estimar viagens. Esta proposta foi devida às limitações do método convencional, tais como: a enorme dependência em relação a estimativa exata das variáveis independentes no futuro e as suposições sobre os dados levantaram questões desafiadoras de sua aplicação e motivaram a aplicação desse método alternativo. Como

resultados, a variável mais importante na previsão de geração de viagens obtida pelo Sistema *Fuzzy* é a população, sendo a mesma já considerada pelos modelos convencionais. Os autores concluíram que o método proposto, em geral, funciona tão bem quanto a RLM, pois os coeficientes de determinação da RLM e do Sistema *Fuzzy* foram desejáveis, mas, o erro quadrático médio no Sistema *Fuzzy* indicou um melhor desempenho para o método proposto.

Nesta revisão foi apresentado que os modelos de geração de viagens podem ser aplicados por diversas técnicas, sendo a Regressão Linear e a Classificação Cruzada as mais utilizadas no passado e aplicadas até hoje. Vários trabalhos encontrados na literatura, os autores afirmam que, apesar da regressão possuir limitações na aplicação em modelos de geração de viagens, apresenta desempenho aceitável para o uso no planejamento e que modelos avançados nem sempre fornecem um melhor desempenho que os modelos tradicionais na estimativa de geração de viagens.

Contudo, esta revisão foi relevante à pesquisa, pois neste trabalho de tese foram usados os modelos tradicionais (regressão) para estimar viagens por domicílios e os resultados foram comparados com os resultados obtidos pelos modelos de redes neurais artificiais (técnica matematicamente mais avançada) e após, foi verificado se a técnica de RNAs é uma alternativa adequada para fins de previsão de geração de viagens por domicílio.

Capítulo 3

POPULAÇÃO SINTÉTICA

Este capítulo apresenta a importância e a necessidade de criar dados sintéticos da população para o desenvolvimento de modelos desagregados de transportes. Trata, especificamente, da apresentação dos modelos de microssimulação desenvolvidos para as diferentes áreas da Engenharia de Transportes e também são abordadas as diversas técnicas de geração da população sintética encontradas na literatura.

3.1 Modelos baseados em microssimulação

Nos últimos anos, muitas pesquisas têm se concentrado no desenvolvimento de modelos desagregados de demanda de viagem, pois permitem uma simulação mais detalhada e precisa da previsão de viagens do que os modelos tradicionais de agregação. No entanto, esses modelos também requerem dados de entrada desagregados (MÜLLER; AXHAUSEN, 2011b).

Para se realizar a previsão de viagens, seja por modelos agregados ou desagregados, são necessários dados da população. Embora muitos países tenham tais dados recolhidos através de censos domiciliares, pouco está disponível à nível de domicílios, devido às restrições de privacidade. Portanto, surge a necessidade de criar uma base de microdados e a microssimulação oferece várias metodologias para a geração desses dados.

Modelos agregados procuram prever a demanda de viagens da população de uma determinada zona a partir das características socioeconômicas totais ou médias da população (BIRKIN; CLARKE, 1988). Apesar de serem modelos tradicionais e muito utilizados no planejamento de transportes, eles não representam a grande variabilidade de comportamentos de indivíduos e domicílios residentes em uma mesma zona.

Por outro lado, os modelos desagregados utilizam domicílios ou em alguns casos indivíduos, como unidades de análise, por isso esses modelos geram uma representação mais detalhada e precisa da previsão da demanda de viagens que os modelos agregados. Assim, vários estudos apontam a necessidade de adotar o enfoque desagregado para prever demanda de viagens a partir do uso de modelos de microssimulação (MILLER, 1997).

As vantagens dos modelos desagregados de demanda de viagens perante os modelos agregados, são: redução dos erros de agregação, segurança na sensibilidade das alterações demográficas com o envelhecimento da população, captura de respostas diferenciadas dos viajantes para as ações políticas e endereçamento de viagens com necessidades particulares (MA, 2011).

Portanto, devido as melhorias apontadas na previsão de demanda por modelos desagregados, o uso da microssimulação nas pesquisas de transporte apresenta uma tendência crescente, pois a estrutura de modelagem, impulsionada pelo avanço computacional e maior disponibilidade de dados de tráfego, bem como os componentes do sistema e suas respectivas interações podem ser representados com elevado grau de detalhamento.

A microssimulação procura reproduzir ou prever o estado de um sistema dinâmico e complexo, como é o caso dos problemas de transportes, através da simulação do comportamento dos agentes individuais e das suas relações no sistema (BHAT et al., 2004).

Um modelo de microssimulação para o planejamento de transporte necessita de dados de entrada desagregados das informações socioeconômicas da população da área de estudo. Por exemplo, os atributos dos domicílios (tamanho, composição familiar, renda, tipo de habitação e posse de automóvel) e os atributos dos indivíduos (idade, sexo, ocupação). Se uma amostra completa e atualizada de pessoas ou domicílios não estiver disponível, essa população deve ser sintetizada (MÜLLER; AXHAUSEN, 2011a). Sintetizar a população às vezes é a única solução devido à privacidade e restrições de gastos (MÜLLER; AXHAUSEN, 2012).

Em geral, a microssimulação envolve duas etapas principais: (1) construção de um conjunto de microdados que represente as características dos agentes de decisão, ou seja, uma população sintética e (2) a simulação do comportamento do agente para o analista e atualização das características dos agentes de decisão com base em modelos matemáticos ou modelos baseados em regras (GUO; BHAT, 2008).

Modelos desagregados de demanda de viagem podem ser modelados em nível de domicílios ou indivíduos. Como exemplo, o MATSim-T, modelo de microssimulação baseado em agentes (indivíduos), que simula decisões dos agentes ao longo do tempo, a fim de prever os estados futuros do sistema. Esse modelo apresentou uma simulação mais detalhada e precisa da previsão de demanda de viagens que os modelos tradicionais de agregação. No entanto, são modelos que necessitam de dados de entrada desagregados para a previsão (MÜLLER; AXHAUSEN, 2011b).

Outro exemplo são os modelos de previsão da demanda de viagens com base em atividades, em que os agentes de decisão a serem microssimulados, geralmente são os domicílios (e seus

membros constituintes) que residem em uma área de estudo. A representatividade da população sintética para o ano base da simulação é fundamental para a precisão dos resultados da simulação (GUO; BHAT, 2008).

Entretanto, a maioria dos microdados da população (características demográficas individuais, bem como as características dos domicílios) ou são suprimidos para manter a confidencialidade [1] ou estão incompletos devido ao elevado custo da sua aquisição (MOECKEL; SPIEKERMANN; WEGENER, 2003).

Uma fonte viável para obter os microdados é através de um censo nacional, no entanto, alguns problemas impedem o uso desses dados como entrada para os modelos de microssimulação. Primeiro, o censo completo é raramente disponível e em muitos casos, apenas uma pequena subamostra (amostra de uso público), pode ser acessível. E segundo, o censo é normalmente realizado a cada 10 anos, portanto, a baixa periodicidade e um espaço de tempo muito grande entre as pesquisas restringem a escolha do ano de referência para o modelo microssimulação (MÜLLER; AXHAUSEN, 2011a).

Portanto, a pouca disponibilidade de microdados ou até mesmo a inexistência e a proteção das informações particulares dos moradores são fatores que remetem a se obter dados sintéticos, ou seja, elaborar uma população sintética a fim de obter um conjunto de dados mais detalhados sobre o comportamento individual ou domiciliar que tenha o respaldo estatístico dos dados observados.

3.2 Geração da população sintética

População sintética é uma representação estatística de uma população real, de forma a caracterizar individualmente (pessoas ou domicílios) quanto à sua localização e perfil socioeconômico, permitindo a sua utilização nos modelos de microssimulações que preveem as atividades e viagens de uma população.

O processo de geração dessas populações sintéticas, ou seja, a geração dos indivíduos ou domicílios preserva a confidencialidade dos dados quando os dados do censo real são utilizados e produz atributos reais demográficos dos indivíduos ou domicílios sintéticos (ADIGA et al., 2015). Portanto, a finalidade dos dados sintéticos é permitir a divulgação de dados pelas entidades estatísticas, assegurando a privacidade ao mesmo tempo em que a informação transmitida é confiável, ou seja, os dados que são publicados têm que transmitir ao usuário a mesma informação, ou a mais próxima possível, que é transmitida pelos dados originais (GRAHAM; YOUNG; PENNY, 2008).

De acordo com Müller e Axhausen (2012) o procedimento frequentemente utilizado na geração de população sintética é o método conhecido como ajuste proporcional iterativo (IPF-

[1] Confidencialidade, ou seja, os dados são publicados na forma agregada, normalmente tendo como unidade uma área local do levantamento, como o setor censitário

Iterative Proportional Fitting). Esse método foi inicialmente proposto por Deming e Stephan (1940) como um algoritmo para a estimativa dos valores célula a célula de uma tabela de classificação cruzada (bidimensional), quando são conhecidos apenas os totais marginais. Segundo Bishop, Fienberg e Holland (2007), esse método é um procedimento que combina a informação a partir de dois ou mais conjuntos de dados e pode ser usado para produzir estimativas de populações para os anos que não são realizados o censo. Duguay, Jung e McFadden (1976) utilizou este método em problemas de transporte para sintetizar dados do domicílio.

Birkin e Clarke (1988) desenvolveram um método, conhecido como *Synthesis* para a geração de microdados sintéticos a partir de dados agregados de diferentes fontes, com o objetivo da determinação da distribuição de dados não disponíveis e da geração de dados de entrada para os modelos de microssimulação. No *Synthesis* as distribuições são determinadas sequencialmente pelo método IPF e a hipótese é que algumas características são relevantes para a determinação de outras. O foco principal dos autores é na desagregação espacial dos dados.

Mais tarde, Beckman, Baggerly e McKay (1996) utilizaram esse mesmo método para criar uma população sintética para o TRANSIMS, baseadas em dados reais de um censo. O método consistia em: estimar atividades para todos os indivíduos e domicílios; planejar viagens multimodais que se adequem a estas atividades; e criar simulações dos pedestres, veículos e o trânsito de todo sistema de transporte de uma cidade com milhões de habitantes. Utilizaram tabelas de classificações cruzadas do censo e da amostra de microdados – PUMA (*Public Use Micro Area*). De uma maneira simplificada, o método consiste em estimar para cada zona, grupos de domicílios classificados por características sociodemográficas (número de moradores e relações de parentesco) e como resultado apresenta a estimativa do número total de domicílios separada em classes, por exemplo, domicílios familiares e não familiares.

Basicamente, as técnicas baseadas no IPF apresentam duas etapas: (1) etapa de ajustamento, ou seja, uma tabela de contingência é ajustada utilizando os dados da amostra da região agregada e os totais marginais obtidos pela soma; (2) a segunda etapa gera os domicílios sintetizados utilizando Simulação de Monte Carlo, onde os registros domiciliares individuais são desenhados a partir de um conjunto de dados semente. Isto é feito por clonagem e replicação de agentes a partir de uma pequena amostra para alcançar o tamanho desejado da população (BECKMAN; BAGGERLY; MCKAY, 1996).

De acordo com o trabalho de Barthelemy e Toint (2013), a maioria das bases de dados disponíveis é agregada por unidade de área devido à questão da privacidade das informações. Desta forma, os autores propõem como solução para o problema, a geração de uma população sintética, pois o modelo de demanda por viagens com base em atividades envolve um grande número de agentes e pode ser impossível ou oneroso obter esses dados totalmente desagregados. Portanto, o processo de geração da população sintética inicia com a identificação das variáveis sociodemográficas dos agentes (indivíduos ou domicílios). Em seguida, é estimado o número de agentes em cada grupo sociodemográfico e associado a cada agente uma proba-

bilidade (depende do peso de amostragem do agente e do número de agentes semelhantes a população verdadeira) para ser incluído na população sintética. Após a definição das probabilidades, os agentes são extraídos aleatoriamente a partir da amostra pelo Método Monte Carlo até que o número esperado de agentes é atingido para cada grupo sociodemográfico.

Na literatura pesquisada a abordagem padrão de geração da população sintética é baseada no método desenvolvido por Beckman, Baggerly e McKay (1996) e foram encontrados vários estudos de casos que empregam com sucesso esse método.

Vários exemplos podem ser encontrados no trabalho de Bowman (2004) que apresenta, além do TRANSIMS, outros sete sintetizadores da população e sugere melhorias para o desenvolvimento de novos sintetizadores. O autor também compara esses sintetizadores de população, nos seguintes aspectos: a abordagem de base, ou seja, como são gerados o ano base e as distribuições previstas (demográficas e geográficas) das características da população, o procedimento utilizado para a geração de domicílios a partir da distribuição, os procedimentos de validação e resultados e os programas de implementação de cada um dos sintetizadores de população.

Outro método sintetizador de população que responde a algumas das limitações do método IPF é o método de otimização combinatória (CO-*Combinatorial Optimization*). Semelhante ao método IPF, a otimização combinatória, também necessita da informação sobre as características da população tanto na amostra (dados sementes) e os níveis marginais (tabela de controle). É um algoritmo iterativo, na primeira etapa é escolhido aleatoriamente um conjunto de agregados, em seguida é avaliado se há necessidade de substituir cada um desses agregados por um novo agregado do conjunto de dados inicial, mas apenas se houver uma melhoria no ajustamento, ou seja, se a substituição permitir melhorar o ajustamento dos agregados ela concretiza-se, caso contrário não se concretiza (VOAS; WILLIAMSON, 2000).

Neste método, um peso é adicionado à amostra e para cada zona a população é sintetizada através da replicação da amostra e otimização desses pesos a fim de minimizar a diferença entre os marginais da zona (MA, 2011). Uma técnica de otimização combinatória para produzir a população sintética é a *Simulated Annealing* usada por Williamson, Birkin e Rees (1998) e Voas e Williamson (2000).

Annealing é o processo utilizado para fundir um metal, onde este é aquecido a uma temperatura elevada e em seguida é resfriado lentamente, de modo que o produto final seja uma massa homogênea. Assim, o termo *simulated annealing* surgiu no contexto da mecânica estatística, desenvolvido por Kirkpatrick et al. (1983).

Os métodos de otimização combinatória apresentaram vantagens sobre o método IPF, tais como menor variância dos resultados e menor necessidade de memória no processamento. Por outro lado, os métodos CO necessitam de maior tempo para a convergência da técnica perante o método IPF. No emparelhamento simulado o processo pode ser oneroso e demorado. A

população sintética gerada pelo método CO é baseada na clonagem/replicação, ou seja, mais atributos significam mais restrições resultando em uma otimização complicada (RYAN; MAOH; KANAROGLOU, 2009).

Alguns trabalhos mais recentes publicados sobre a geração de uma população sintética são descritos a seguir. Todos os trabalhos apresentam o mesmo objetivo, gerar uma população sintética com resultados mais próximos da população real, porém cada trabalho propõe um método diferente e foram desenvolvidos em diferentes áreas de estudo e muitos deles aplicados para modelos de demanda por viagens.

Adiga et al. (2015) propuseram a geração de uma população sintética para os Estados Unidos e utilizaram dados da pesquisa (ACS) para criar um conjunto desagregado de agentes com diferentes variáveis demográficas. A metodologia utilizada na geração da população sintética foi baseada no trabalho de Beckman, Baggerly e McKay (1996).

Hafezi e Habib (2014) utilizaram o método *Fitness Based Synthesis* (FBS) para sintetizar uma população, ou seja, produzir indivíduos ou domicílios virtuais através do uso da amostra desagregada e os correspondentes dados agregados. O objetivo de sintetizar a população foi contribuir para um estudo do desenvolvimento do transporte integrado no Canadá. O algoritmo foi projetado para permitir sintetizar simultaneamente, ambos os níveis: indivíduos e domicílios, pois os modelos de microssimulação, baseados em agentes, tentam prever o comportamento dos indivíduos e dos domicílios ao invés do comportamento médio das zonas de tráfego.

Huynh et al. (2013) desenvolveram um algoritmo que é uma variação da abordagem CO para a construção de uma população sintética que utiliza apenas dados agregados das distribuições demográficas de Sydney. A importância da população sintética desenvolvida neste trabalho é na sua capacidade para capturar as relações e as mudanças dos indivíduos dentro um domicílio, pois os comportamentos e as decisões dos indivíduos relacionados com o transporte são altamente influenciados pelo tipo e composição que os indivíduos têm. Para a validação da população sintética utilizaram os dados do censo e concluíram que o algoritmo proposto pode construir uma população sintética (indivíduos e domicílios) realista para fins de modelagem baseada em agentes.

Barthelemy e Toint (2013) apresentaram uma nova técnica de geração da população sintética na Bélgica, utilizando o método de reconstrução sintética para evitar as limitações do método baseado em IPFP (*Iterative Proportional Fitting Procedure*). A ideia desse método é que o gerador é livre de amostra, ou seja, não precisa de uma pesquisa para obter dados, pois geralmente são caras e apresentam problemas como a privacidade e pode lidar com (moderada) inconsistência de dados, pois é comum quando os dados são extraídos a partir de várias fontes. O método foi validado comparando os resultados da geração dos domicílios com o IPFP e concluíram que a metodologia proposta tem potencial para produzir populações sintéticas reais e confiáveis.

Müller e Axhausen (2011b) desenvolveram um artigo do estado da arte da população sintética e apresentaram seis procedimentos para população sintética usado em vários modelos de microssimulação. Portanto, a necessidade de elaborar uma população sintética provém dos modelos baseados em microssimulação. Consequentemente, as técnicas de geração da população sintética são concebidas como uma alternativa viável para a obtenção de microdados e para uso de dados de entrada em modelos desagregados (BECKMAN; BAGGERLY; MCKAY, 1996).

Arentze, Timmermans e Hofman (2008) desenvolveram um método usando o Ajuste Proporcional Iterativo (IPF) para gerar domicílios sintéticos, com base na distribuição de indivíduos, para a aplicação no modelo baseado em atividades. O método foi baseado no uso de relações entre matrizes para converter as distribuições, a fim de converter as distribuições por indivíduos para distribuições por domicílios de modo que as distribuições marginais poderiam ser controladas em ambos os níveis. Os autores concluíram que o método IPF é vantajoso para previsão da demanda por transportes (modelo *Albatross*) devido à sua flexibilidade, pois permite aos usuários definirem cenários populacionais. O modelo *Albatross* é um modelo de atividades baseado em regras lógicas, que representam heurísticas de escolha a partir de um conjunto de dados e foi desenvolvido para o Ministério dos Transportes da Holanda. O método desenvolvido foi inserido como um módulo de geração da população sintética no sistema *Albatross*.

Guo e Bhat (2008) propuseram um método de geração da população sintética, com base nas limitações do método convencional IPF. O algoritmo foi desenvolvido com o uso de estruturas de dados genéricos e operadores. A implementação desse algoritmo permitiu que o usuário ajustasse a escolha das variáveis de controle e definisse a classe das variáveis em tempo de execução, ou seja, variáveis com características socioeconômicas no nível de domicílios ou no nível de indivíduos. Portanto, concluíram que o método proposto foi mais flexível que o convencional, principalmente, quando a geração da população sintética é para diferentes áreas de estudo. O algoritmo proposto indicou, através dos resultados de validação, que a população sintética obtida, está mais próxima da população verdadeira quando comparada àquela obtida pelo método convencional IPF. Os autores afirmam que, apesar de ser desenvolvido para um contexto particular, ou seja, para aplicação em modelos de viagens baseado em atividades, o algoritmo proposto é relevante para outras áreas de estudo com microssimulações.

A geração da população sintética para fins desta tese é fundamental, pois para estimar viagens produzidas por domicílio são necessárias informações detalhadas do domicílio e a geração de viagens é altamente influenciada pelo tipo e composição que o domicílio possui. Além disso, os dados desagregados raramente estão disponíveis para uso, por problemas de privacidade e alto custos para pesquisas.

A área de estudo deste trabalho é a cidade de São Carlos-SP que possui apenas uma base desagregada da Pesquisa O/D realizada no ano de 2007/2008. Portanto, os dados desagregados estão desatualizados e incompletos para a modelagem de viagens produzidas por domicílio.

Desta forma, este trabalho propõe a geração de uma população sintética dos domicílios de São Carlos-SP com base nos dados agregados do censo do IBGE-2010 utilizando o Método Monte Carlo, descrito na próxima seção.

Capítulo 4

TÉCNICAS ABORDADAS

Este capítulo apresenta as duas técnicas utilizadas na elaboração do método proposto: (1) Monte Carlo, usada na geração da população sintética e (2) Redes Neurais Artificiais, utilizada na modelagem de geração de viagens por domicílio. É realizada uma breve apresentação conceitual das técnicas, além da seleção de alguns trabalhos que utilizaram a simulação Monte Carlo e a técnica de RNAs. Para finalizar este capítulo são apresentadas ainda, as vantagens e desvantagens das técnicas de Estatísticas e de Redes Neurais quando se trabalha com dados de transportes.

4.1 Método Monte Carlo (MMC)

O Método Monte Carlo tem se tornado uma das técnicas mais populares para se analisar sistemas complexos (BANKS et al., 2005). Devido à complexidade dos sistemas reais, os modelos de simulação obtêm com mais precisão as características dinâmicas e aleatórias desses sistemas, pois buscam imitar o sistema real em um computador para avaliar como o sistema modelado se apresentaria quando submetido às mesmas condições de contorno (CHWIF; MEDINA, 2014).

A simulação de um modelo segue uma sequência lógica de etapas, iniciando-se com a identificação do problema, seguida pela escolha das variáveis significativas do problema, as etapas de construção e teste do modelo, a realização do experimento juntamente com a obtenção e avaliação dos resultados e, por fim, as possíveis necessidades de alterações no modelo (RENDER; STAIR; HANNA, 2000).

Segundo Reis e Martins (2001) existem dois tipos de modelos de simulação, o determinístico, que pressupõe que os dados são obtidos com certeza, ou seja, não incorpora as probabilidades de que o valor escolhido para a simulação sofra alterações futuras e o probabilístico que

incorpora o comportamento probabilístico no relacionamento interno do sistema, por meio da utilização de técnicas estatísticas e computadores. Os modelos probabilísticos tiveram suas origens no Método Monte Carlo e tem como foco as simulações de fenômenos aleatórios (NASCIMENTO; ZUCCHI, 1997).

O Método Monte Carlo (MMC) é uma parte da matemática experimental que está preocupada em experiências com números aleatórios Hammersley e Handscomb (1964). Além disso, envolve probabilidade para resolução de problemas (GUJARATI; PORTER, 2011).

Segundo Chwif e Medina (2014) o MMC utiliza geradores de números aleatórios para simular sistemas físicos ou matemáticos, nos quais não se considera o tempo explicitamente como uma variável. É um procedimento utilizado para simular numericamente caminhos aleatórios de tempo contínuo/de estado discreto (GILLESPIE, 1978).

Conforme Moore e Weatherford (2005), o MMC é um dos vários métodos para análise da propagação da incerteza, onde sua grande vantagem é determinar como uma variação randomizada (aleatória), já conhecida, ou o erro, afetam o desempenho ou a viabilidade do sistema que esta sendo modelado. Geralmente, é muito utilizado em modelos complexos, ou não lineares e uma simulação pode envolver mais de 10.000 avaliações do modelo estudado. Portanto é uma tarefa difícil e demorada, que no passado só poderia ser realizada por eficientes computadores (HAMMERSLEY; HANDSCOMB, 1964).

Desde então, o Método Monte Carlo tem sido utilizado para modelar uma grande variedade de fenômenos, e evoluiu para várias variantes bem estabelecidas. Em suas primeiras formas, esse método era essencialmente um meio eficiente para estimar numericamente integrais complexas. Neste contexto, poderia ser operado simplesmente calculando as energias de um modelo em estados selecionados aleatoriamente, e ponderando a possibilidade de realização de cada estado de acordo com a equação de energia *Boltzmann* (BATTAILE, 2008).

De acordo com Escudero (1973) a resolução de um problema utilizando o MMC depende do uso de várias séries de tentativas aleatórias e, portanto, a precisão final depende desse número de tentativas e também, do tempo de computação.

A execução do MMC segue as seguintes etapas, conforme Corrar e Theóphilo (2004):

a) definição das variáveis com base em dados passados;

b) identificação das distribuições de probabilidades das variáveis aleatórias do estudo;

c) construção das distribuições de probabilidades acumuladas para cada uma das variáveis;

d) definição dos intervalos de números aleatórios para cada variável;

e) simulação dos experimentos.

Para a obtenção de amostras aleatórias é necessária uma sequência de números aleatórios que podem ser obtidas por um gerador de números aleatórios. Esses números gerados são pseudoaleatórios, pois apresentam aproximações razoáveis de números aleatórios inteiros. Os

dados obtidos podem ser utilizados para amostras aleatórias de alguma população de interesse (BARROS; MAZUCHELI, 2005).

Portanto, devido à complexidade dos problemas reais e a evolução dos sistemas computacionais, a simulação é um instrumento muito utilizado nas mais variadas áreas de conhecimento. A seguir são apresentados trabalhos na área de transportes que utilizaram o MMC.

Gillespie (1978) apresentou um procedimento do Método Monte Carlo orientado por computador para simular numericamente o processo de caminho aleatório.

Watanatada e Ben-Akiva (1979) desenvolveram uma metodologia de agregação espacial baseado em funções matemáticas contínuas para a previsão de demanda de viagens urbanas de passageiros. A aplicação da simulação Monte Carlo foi utilizada para a previsão de demanda de viagens urbanas usando modelos de escolha desagregados. Também desenvolveu relações empíricas aproximadas para examinar as propriedades estatísticas dos vieses e erros aleatórios em previsões de Monte Carlo. Aplicaram o MMC para gerar amostras representativas de indivíduos e alternativas espaciais distribuídas ao longo do espaço urbano.

Williams e Ortúzar (1982) analisaram os processos comportamentais que exigem análises numéricas para a resolução e utilizaram o MMC para gerar o modelo de demanda de viagem.

Brundell-Freij (2000) usou *bootstrapping* e simulações Monte Carlo para analisar o efeito de reamostragem de componentes do modelo aleatório no âmbito de um modelo específico investigar a influência de tal variação na saída da estimação e seleção de modelos.

Kitamura et al. (2000) desenvolveram um micro simulador para a geração de padrões de atividades diárias de viagens. Os autores afirmam que os padrões de viagens diárias de um indivíduo podem ser sintetizados de um modo prático por MMC.

Zhang, Xie e Waller (2011) utilizaram o MMC para simular a incerteza devido a três diferentes níveis de demanda de viagens ou de tráfego de congestionamento (0,5, 1,0 e 1,5 vezes os valores médios do número máximo de potenciais viajantes), 3 graus diferentes de incerteza (coeficiente de valores de variação de 0,1; 0,2 e 0,3), e 3 diferentes fontes de incerteza (demanda, oferta e incertezas de parâmetros). Para cada combinação, 300 simulações foram realizadas. A incerteza foi avaliada em termos de propagação de erros nas várias etapas de modelagem.

Rasouli e Timmermans (2012) discutem sobre o problema da incerteza de dados de entrada nos modelos quatro etapas e investigam essa incerteza de entrada e a propagação de erros. Os valores de entrada para as variáveis têm sido substituídos por distribuições de probabilidade (muitas vezes a distribuição normal), com o desvio-padrão da distribuição capturando a quantidade da incerteza. O MMC se fundamenta nestas distribuições de probabilidades, em seguida, gera diferentes configurações de valores de entrada utilizadas em diferentes corridas do modelo de demanda de viagens.

Kim et al. (2013) propuseram um quadro conceitual para obter a natureza probabilística de tempos de viagem por meio de modelos de simulação de tráfego. Utilizaram a simulação Monte

Carlo para a geração de vários cenários de entrada (trajetórias de veículos individuais), por amostragem, a partir da distribuição de probabilidade conjunta dos componentes do cenário. Cada cenário tem a mesma probabilidade e o processador de trajetórias constrói distribuições de tempos de viagem agregados ao longo de cenários aleatórios. Após um grande número de simulações a análise de confiabilidade baseada em cenários obteve as distribuições globais de tempos de viagem. Portanto, os cenários aleatórios e a probabilidade de cada grupo, foram obtidos a partir da amostragem de Monte Carlo.

Rico, Rodenas e Aranda (2014) utilizaram a abordagem Monte Carlo para resolver a incerteza da demanda e empregaram na simulação da demanda estocástica em problema de rede de carregamento de tráfego. Eles optaram por utilizar o MMC, por ser um método muito flexível que permite o uso de qualquer variável aleatória. Através desse método obtiveram uma amostra aleatória de dados de entrada da demanda de tráfego para a rede Nguyen-Dupuis, composta por 13 nós e 4 pares de origem e destino. Os autores apontaram como principal desvantagem deste método o tempo de execução.

Portanto, observa-se que o MMC é amplamente utilizado no campo da Engenharia de Transportes para várias finalidades, entre elas a geração da população sintética. Nesse sentido, optou-se nesta tese por Monte Carlo para gerar a população sintética com as características dos domicílios da área de estudo, pois vários trabalhos ratificam o uso desse método para gerar domicílios sintéticos, dentre eles é conveniente citar os trabalhos de: Birkin e Clarke (1988), Beckman, Baggerly e McKay (1996), Huang e Williamson (2002), Münnich et al. (2003), Namazi-Rad, Mokhtarian e Perez (2014) e Ma e Srinivasan (2015).

4.2 Redes Neurais Artificiais (RNAs)

A estrutura e o funcionamento do sistema nervoso, especificamente o cérebro humano, serviu de inspiração para o desenvolvimento das redes neurais. O objetivo era simular a capacidade de aprendizado do cérebro humano na aquisição de conhecimento (CARVALHO et al., 2011). Assim, uma rede neural é uma máquina projetada para modelar a maneira que o cérebro realiza uma tarefa, através da programação de um computador (HAYKIN, 1999).

Em 1940 teve início a procura por modelos computacionais do sistema nervoso e os primeiros trabalhos tinham por objetivo compreender o cérebro e utilizar o conhecimento obtido para desenvolver sistemas de aprendizado biologicamente prováveis. A Tabela 4.1 apresenta um histórico da evolução dos trabalhos de RNAs.

A década de 1980 foi de grandes mudanças para esta área de conhecimento, devido ao aparecimento de computadores mais rápidos, o interesse da construção de computadores paralelos, as novas propostas de arquiteturas das RNAs com maior capacidade de representação além, de algoritmos de aprendizado mais sofisticados (HAYKIN, 1999).

Tabela 4.1 – Histórico dos trabalhos de RNA.

ANO	AUTORES	TRABALHO
1943	**McCulloch e Pitts**	**Primeiro modelo artificial de um neurônio biológico**
1948	N. Wiener	Livro Cybernetics de Wiener
1949	J. von Neumann	Palestra de Von Neumann na Universidade de Illinois para a divulgação da teoria de McCulloch e Pitts
1949	**Hebb**	**Regra de aprendizagem para a modificação sináptica**
1951	Minsky e Edmonds	Primeiro computador de Redes Neurais
1958	**Rosenblatt**	**Perceptron**
1960	Widrow e Hoff	Adaline
1969	**Minsky e Papert**	**Perceptrons - livro com resultado negativo sobre a capacidade de representação de uma rede neural de uma camada**
1969	**Bryson e Ho**	**Descoberta do algoritmo de aprendizado backpropagation**
1980	Grossberg	Auto-organização
1982	**Hopfield**	**Redes recorrentes com conexões sinápticas simétricas**
1982	Kohonen	Mapas auto-organizáveis
1985	**Ackley, Hinton e Sejnowsky**	**Máquina de Boltzmann- primeira rede neural de múltiplas camadas bem sucedida**
1986	McClelland e Rumelhart	Reinvenção do backpropagation.
1988	Broomhead e Lowe	Funções de base radial (RBF)- alternativa aos perceptrons de múltiplas camadas

Mais informações sobre o surgimento, evolução nas pesquisas e ideias básicas sobre as Redes Neurais Artificiais são encontradas nos trabalhos de Lippmann (1987); Wasserman (1989); Hertz, Krogh e Palmer (1991) e Hinton (1992). O trabalho de Zhang, Patuwo e Hu (1998) traz o estado da arte sobre o uso das RNA como ferramenta de previsão, sintetiza os trabalhos publicados na área e apresenta os problemas da modelagem por RNA.

Percebe-se na literatura que as pesquisas nesta área estão crescendo a cada ano e vêm sendo aplicadas em várias áreas do conhecimento com resultados satisfatórios. Isso ocorre, pois as RNAs são capazes de processar uma grande quantidade de dados e fazer previsões com precisão. Tem poder de generalização, no que se refere à capacidade que elas possuem de produzir saídas razoáveis, a entradas que nunca lhe tenham sido apresentadas, solucionando problemas complexos (SILVA, 2003).

Na Engenharia de Transportes as RNAs passaram a ser utilizadas com mais frequência e as experiências mostram que, em alguns casos, os resultados são melhores que os modelos estatísticos convencionais, pois podem tratar mais adequadamente as variações no comportamento dos dados. Uma vantagem da utilização da rede neural em transportes é encontrada no trabalho de Carneiro (2003) que afirma que a rede sem um conhecimento inicial das variáveis mais importantes e seus respectivos pesos pode estimar a demanda de passageiros entre dois municípios.

A seguir são apresentados alguns conceitos básicos sobre RNAs.

4.2.1 O neurônio artificial e o modelo não linear

O neurônio é uma unidade de processamento da informação e forma a base para o projeto de RNAs, por isso uma breve explicação é apresentada.

As RNAs são modelos computacionais inspirados nos sistemas biológicos, particularmente, no cérebro humano, pois apresenta uma enorme capacidade de processamento das informações, tais como, o reconhecimento da fala e a segmentação de imagens. Devido a isso, pesquisadores tentam reproduzir as reações do cérebro em máquinas e tem-se obtido resultados promissores, apesar do pouco conhecimento do cérebro humano. Assim, as RNAs são formadas por um conjunto de neurônios artificias que interagem entre si e apresentam comportamentos semelhantes ao funcionamento dos neurônios biológicos (SILVA et al., 2008).

O modelo de um neurônio apresenta três elementos básicos que formam a base para o projeto de uma rede neural artificial e estão destacados em negrito na Figura 4.1.

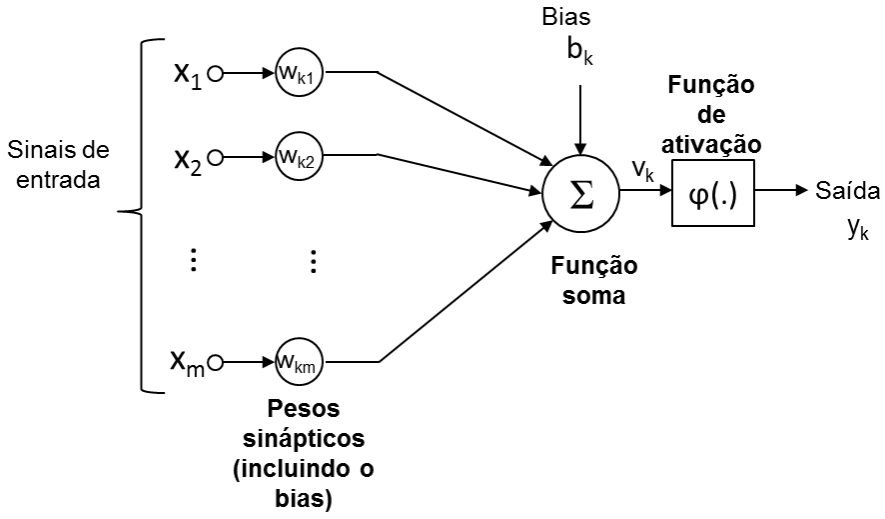
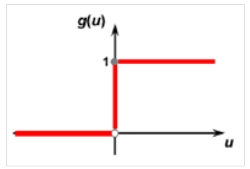
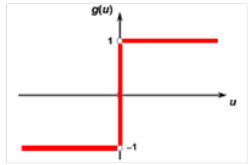
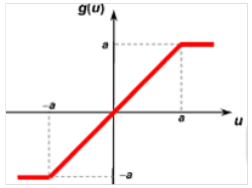
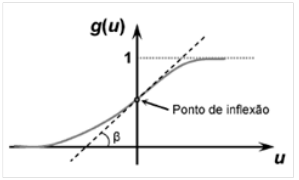
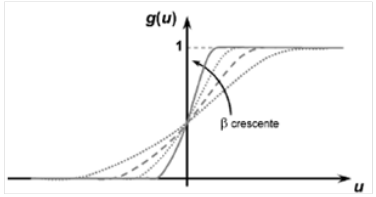


Figura 4.1 – Modelo não linear de um neurônio.

Além disso, é fundamental entender o funcionamento do modelo em termos matemáticos e definir a função de ativação que se aplicada ao modelo. As formulações matemáticas e os diferentes tipos de função de ativação estão apresentadas na Tabela 4.2.

Tabela 4.2 – Resumo dos elementos, funções matemáticas do modelo e tipos de função de ativação.

Modelo não linear de um neurônio

Elementos	**Formulações Matemáticas**	**Tipos de Função de ativação**
Sinapses: são caracterizadas por um peso w, e a função do w é multiplicar o sinal x_j na entrada da sinapse j, conectado ao neurônio k. Se positivo o w, é sinapse excitatória, se negativo o w é sinapse inibitória. Somatório: somam os sinais de entradas, ponderados pelas respectivas sinapses do neurônio, realizadas por combinações lineares. Função de ativação: limita a amplitude da saída de um neurônio. A entrada é normalizada dentro de um intervalo fechado [0,1] ou [-1,1].	1-Saída do combinador linear devido aos sinais de entrada $u_k = \sum_{j=1}^{m} w_{kj} x_j$ 2-Sinal de saída do neurônio $y_k = \varphi(u_k + b_k)$ 3-O uso do bias tem efeito de aplicar uma transformação afim à saída u_k do combinador linear $v_k = u_k + b_k$ 4- O bias é um parâmetro externo do neurônio artificial. Nova formulação para o potencial de ativação $v_k = \sum_{j=0}^{m} w_{kj} x_j$ 5-Nova formulação para a saída do neurônio $y_k = \varphi(v_k)$	Função degrau $\varphi(v) = \begin{cases} 1, se\ v \geq 0 \\ 0, se\ v < 0 \end{cases}$ Função sinal (degrau bipolar) $\varphi(u) = \begin{cases} 1, se\ v > 0 \\ 0, se\ v = 0 \\ -1, se\ v < 0 \end{cases}$ Função rampa simétrica $\varphi(u) = \begin{cases} a, se\ v > a \\ v, se\ -a \leq v \leq a \\ -a, se\ v < a \end{cases}$ Função logística $\varphi(u) = \frac{1}{1 + e^{-\beta . v}}$ Função tangente hiperbólica $\varphi(u) = \frac{1 - e^{-\beta . v}}{1 + e^{-\beta . v}}$	    

Segundo (HAYKIN, 1999) o neurônio artificial simula o comportamento do neurônio biológico e é uma unidade de processamento matematicamente simples. De forma resumida, o neurônio recebe uma ou mais entradas interligados por um grande número de conexões, ou seja, as conexões sinápticas, com outras unidades similares a ele, com seus respectivos pesos e transforma em saídas. Os valores dependem diretamente da somatória ponderada de todas as saídas dos outros neurônios a esse conectado.

Assim, os neurônios de uma rede neural são estruturados de acordo com o algoritmo de aprendizagem utilizado para treinar a rede.

Uma RNA é caracterizada por dois aspectos básicos: arquitetura e aprendizado. A arquitetura ou estrutura da rede está relacionada ao tipo e número de unidades de processamento e à forma como os neurônios estão conectados e o aprendizado define as regras usadas no ajuste dos pesos da rede e qual informação é utilizada pelas regras (HAYKIN, 1999). Esses dois aspectos são descritos nos próximos itens.

4.2.2 Arquitetura de uma rede neural

De acordo com Haykin (1999) a arquitetura das redes se apresenta em três diferentes classes: (1) as redes progressivas de única camada, (2) as redes progressivas de camadas múltiplas (ambas chamadas de redes não recorrentes) e (3) as redes recorrentes. A estrutura (arquitetura) de uma rede neural varia de acordo com as camadas intermediárias, com a quantidade de neurônios, com a função de ativação e com o algoritmo de aprendizagem. Cada uma das classes de arquitetura da rede fundamentalmente diferentes é apresentada a seguir:

1. Redes progressivas de única camada

   Os neurônios geralmente estão organizados em camadas, e a camada de entrada projeta uma camada de saída de neurônios, somente em um sentido, por isso é chamada de progressiva. A camada de entrada não é computada, pois não é realizada qualquer computação, por isso a camada de saída é a única camada da rede, de acordo com a Figura 4.2.

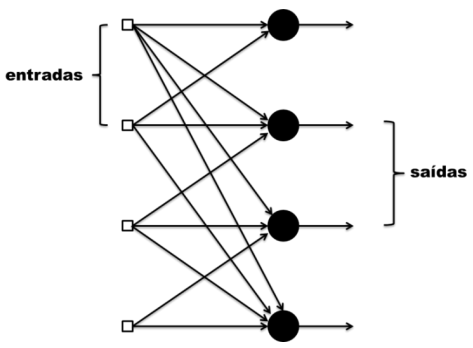


Figura 4.2 – Redes progressivas de única camada.

2. Redes progressivas de múltiplas camadas

   Esta rede se difere da anterior por apresentar uma ou mais camadas escondidas, cuja função é intervir entre a entrada externa e a saída da rede, tornando a rede capaz de extrair estatísticas de ordem mais alta, que as redes de única camada. Essa habilidade de extrair estatísticas de ordem elevada é valiosa quando o tamanho da camada de entrada é grande. A Figura 4.3 mostra uma rede de múltiplas camadas (10-4-2), com 10 nós de entrada, quatro neurônios escondidos e dois neurônios de saída, totalmente conectada, pois cada um dos nós de uma camada da rede está conectado a todos os nós da camada seguinte.

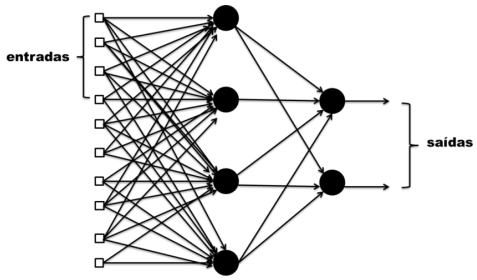


Figura 4.3 – Redes progressivas de múltiplas camadas.

3. Redes recorrentes

   No caso das redes recorrentes, elas se diferem das redes progressivas por ter pelo menos uma realimentação. A rede recorrente pode consistir de uma única camada de neurônios e tem função de retornar o sinal de saída de volta às entradas de todos os neurônios. Se for uma rede com neurônios escondidos, as conexões de retroalimentação originam-se tanto dos neurônios escondidos quanto dos neurônios de saída. Esses loops de retroalimentação têm um grande impacto na capacidade de aprendizado e no desempenho da rede. A Figura 4.4 ilustra esse tipo de rede.

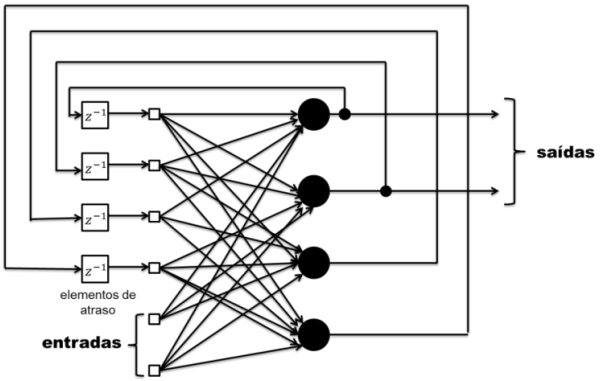


Figura 4.4 – Redes recorrentes.

Essas redes são indicadas para aplicações de processamento de informações sequenciais e na simulação de sistemas dinâmicos, como, por exemplo, o processamento de língua natural e o controle de braços robóticos.

4.2.3 Aprendizado

O ajuste dos parâmetros de uma RNA, ou seja, a definição dos valores dos pesos associados às conexões da rede que fazem com que o modelo obtenha melhor desempenho, geralmente medido pela precisão preditiva, vários algoritmos de treinamento têm sido propostos na literatura. Segundo Carvalho et al. (2011) esses algoritmos podem ser divididos em quatro grupos, que se diferem pela forma de como é estabelecido o ajuste de um peso sináptico de um neurônio:

- ***Correção de erro***: ajuste dos pesos na RNA de forma a reduzir os erros cometidos pela rede. Esses algoritmos são geralmente utilizados em aprendizado supervisionado.
- ***Hebbiano***: baseados na regra de Hebb (se dois neurônios em ambos os lados de uma sinapse estão simultaneamente ativos, a conexão entre eles deve ser reforçada). São frequentemente usados em aprendizagem não supervisionada.
- ***Competitivo***: os neurônios de saída da rede competem entre si para se tornarem ativos (disparar). A diferença do competitivo com o hebbiano é que o algoritmo hebbiano pode ter vários neurônios de saída ativos simultaneamente, e o competitivo somente um neurônio de saída está ativo em um determinado instante. A competição entre os neurônios definem quais devem ter os pesos ajustados. Os neurônios que vencem a competição, geralmente são os que respondem mais fortemente ao objeto apresentado aos seus terminais de entrada.
- ***Termodinâmico (Máquina de Boltzmann)***: algoritmos estocásticos baseados em princípios observados na metalurgia. Os neurônios constituem uma estrutura recorrente e operam de maneira binária (+1 ligado; -1 desligado).

A primeira regra de aprendizagem apresentada foi no trabalho de Hebb (1949) e quase dez anos depois (ROSENBLATT, 1958) desenvolveu a primeira RNA – rede perceptron de camada única desenvolvida O perceptron foi um método inovador de aprendizagem supervisionada e é a forma mais simples de uma rede neural classificar os padrões de 2 classes linearmente separáveis. Basicamente, consiste de um único neurônio com pesos sinápticos ajustáveis. A rede é treinada por um algoritmo supervisionado de correção de erro e usa a função de ativação do tipo degrau ou sinal. Embora seja uma rede simples, de uma camada de neurônios, apresenta uma boa acurácia em problemas de classificação (HAYKIN, 1999).

Apesar de resolver bem problemas de classificação, a rede perceptron de camada única não resolve problemas não lineares. Portanto, surgiu uma generalização do perceptron de camada

única, em que foram adicionadas uma ou mais camadas intermediárias de neurônios para se tornar possível a aproximação de qualquer função. Essa rede com mais camadas é chamada de perceptron multicamadas (MLP-*Multilayer Perceptron*).

As redes MLP têm sido aplicadas com sucesso para resolver problemas difíceis, por meio do treinamento supervisionado com o uso do algoritmo *backpropagation*. Para o desenvolvimento dos modelos de produção de viagens será utilizado neste trabalho a rede perceptron multicamadas, por apresentar um poder computacional muito grande comparado as redes que não possuem camadas intermediárias, e pela grande quantidade de dados de entrada. Além disso, de acordo com Dougherty (1995), as redes de múltiplas camadas, junto ao algoritmo *backpropagation* (retro propagação) são as RNAs mais utilizadas na área de transporte, com bons resultados em estudos semelhantes.

É apresentado a seguir mais detalhadamente o funcionamento das redes MLP e do algoritmo *backpropagation*, pois fundamentam a estrutura metodológica deste trabalho.

4.2.4 Perceptron multicamadas (MLP) e o algoritmo *backpropagation*

Uma rede MLP é formada por um conjunto de nós de entrada, uma ou mais camadas intermediárias (ocultas) e uma camada de saída. Geralmente, é uma rede completamente conectada, de modo que, os neurônios da camada l estão conectados a todos os neurônios da camada $l+1$. O perceptron multicamadas tem três características importantes: a função de ativação não linear, uma ou mais camadas de neurônios ocultos e alto grau de conectividade (CARVALHO et al., 2011).

Segundo Rosa (2011) o modelo de cada neurônio inclui uma função de ativação não linear, como a função sigmoide que é contínua e diferençável, definida normalmente pela função logística. A não linearidade é importante, pois caso as redes utilizassem funções lineares, a relação entrada-saída da rede poderia ser reduzida àquela de um perceptron de camada única. A rede contém uma ou mais camadas ocultas que capacitam a rede a aprender tarefas complexas extraindo as características mais significativas dos padrões de entrada. Por fim, as redes MLP exibe alto grau de conectividade determinada pelas sinapses da rede (HAYKIN, 1999).

Devido à habilidade de aprender da experiência através do treinamento e por meio dessas três características é que a rede MLP deriva seu poder computacional. Porém, essas características tornaram mais difíceis entender o comportamento da rede, como, por exemplo, o processo de aprendizagem mais complicado de ser visualizado. Por isso, foi proposto por Rumelhart et al. (1986) um algoritmo para o treinamento da MLP baseado em gradiente descendente denominado *backpropagation*.

A maior propriedade dos sistemas conexionistas é permitir a generalização da rede neural para classificar entradas nunca vistas e não só aprender classificar as entradas nas quais ela é treinada (ROSA, 2011).

Normalmente a rede MLP é encontrada como mostra a Figura 4.5, com uma camada de entrada, duas camadas ocultas e uma camada de saída e os neurônios das camadas totalmente conectados aos outros neurônios da camada seguinte.

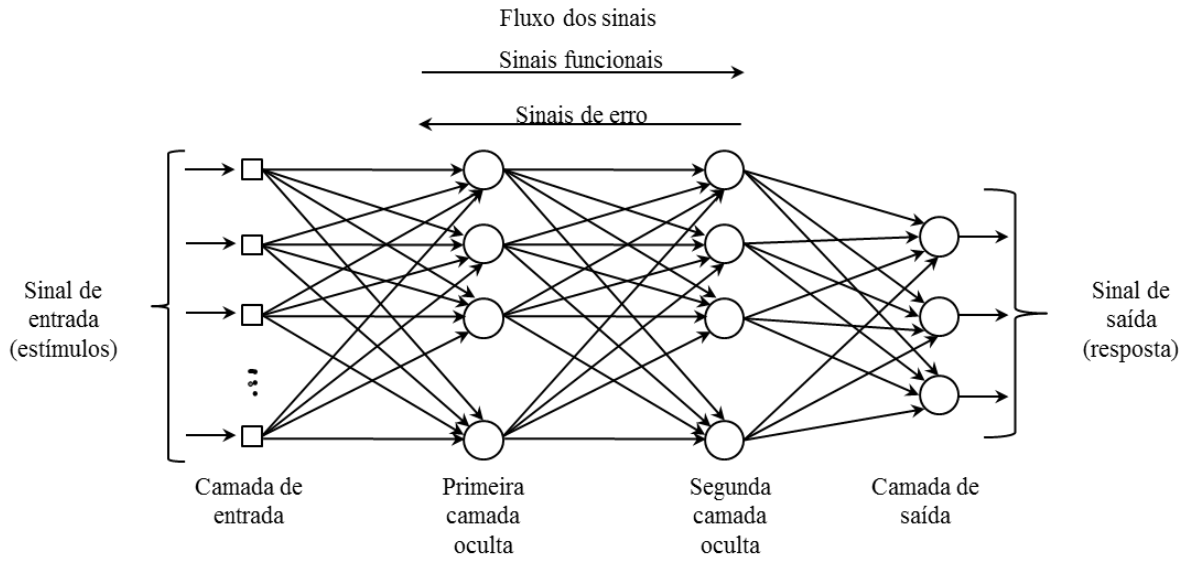


Figura 4.5 – Rede MLP.

O fluxo dos sinais pode ser da esquerda para a direita (sinais funcionais) ou da direita para a esquerda (sinais de erro). O sinal funcional é um sinal de entrada que incide no terminal de entrada da rede e propaga-se de neurônio por neurônio e emerge no terminal de saída e o sinal de erro se origina em um neurônio de saída e se propaga para trás de camada por camada. Cada neurônio oculto de um perceptron multicamadas é projetado para realizar dois cálculos: o cálculo do sinal funcional que aparece na saída do neurônio e o cálculo da estimativa do vetor gradiente (HAYKIN, 1999).

O algoritmo *backpropagation* se inicia com um conjunto de pesos aleatórios e a rede ajusta esses pesos cada vez que vê um par entrada-saída (ROSA, 2011). Este algoritmo é baseado na regra delta e é constituído da iteração de dois estágios: o passo para frente (*forward*) e o passo para trás (*backward*). No primeiro estágio, cada amostra de entrada é apresentada à rede. A amostra é recebida por cada neurônio da primeira camada oculta (ponderação do peso e suas conexões) e cada neurônio aplica a função de ativação a sua entrada total e produz um valor de saída. Esse valor é comparado com o valor desejado da saída desse neurônio obtendo-se o sinal de erro para o neurônio de saída. Os pesos sinápticos neste passo se mantêm inalterados em toda a rede e os sinais funcionais são calculados individualmente, ou seja, neurônio por neurônio. Portanto, a fase de propagação começa na primeira camada oculta (apresentação do vetor de entrada) e termina na camada de saída calculando o sinal de erro de cada neurônio desta camada (CARVALHO et al., 2011).

O valor desse erro é utilizado no próximo estágio para os ajustes dos pesos de entrada.

O passo da retropropagação *backpropagation* começa na camada de saída e recursivamente calcula o gradiente local de cada neurônio. Esse processo recursivo permite que os pesos sinápticos sofram modificações de acordo com a regra delta, ou seja, é medido a distância entre a resposta real e a desejada que produz um sinal de erro que é propagado para trás e são realizados os ajustes apropriados nos pesos das conexões de modo a reduzir essa distância. Esse ajuste pode ser feito através da Equação 4.1.

$$w_{jl}(t+1) = w_{jl}(t) + \eta x^j \delta_l \tag{4.1}$$

sendo w_{jl}: peso entre um neurônio l e o j-ésimo atributo de entrada ou a saída do j-ésimo neurônio da camada anterior;
δ_l: erro associado ao l-ésimo neurônio;
x^j: entrada recebida por esse neurônio; e
η: taxa de aprendizagem.

Para determinar o erro dos neurônios das camadas intermediárias, o algoritmo *backpropagation* estima pela soma dos erros da camada seguinte, ponderados pelo valor do peso associado a essas conexões. A forma de calcular esse erro vai depender da camada em que se encontra o neurônio, como mostra a Equação 4.2.

$$\delta_l = \begin{cases} f'_a e_l, & \textit{se o neurônio estiver na camada de saída} \\ f'_a \sum w_{lk}\delta_k, & \textit{se o neurônio estiver na camada oculta} \end{cases} \tag{4.2}$$

sendo f'_a: derivada parcial da função de ativação do neurônio e
e_l: erro quadrático cometido pelo neurônio de saída quando sua resposta é comparada à desejada.

A taxa de aprendizado tem uma forte influência no tempo necessário de convergência, ou seja, se a taxa for muito pequena, menor as variações nos pesos sinápticos de uma iteração para a outra. Se grande, as modificações nos pesos sinápticos podem tornar a rede instável e dificultam a convergência. Um modo para evitar a instabilidade e aumentar a taxa é modificar a regra delta incluindo um momento, que quantifica o grau de importância da variação de peso do ciclo anterior ao ciclo atual usando a regra delta generalizada (HAYKIN, 1999).

Basicamente, no algoritmo *backpropagation* a aprendizagem deriva das apresentações aleatórias de um conjunto de exemplos de treinamento para a rede perceptron multicamadas em uma base de época, até que os pesos sinápticos e os níveis de bias se estabilizarem e o erro médio quadrado sobre o conjunto de treinamento convergir para um valor mínimo. Essa aprendizagem pode ser de duas formas: *on-line* (a atualização dos pesos é realizada após a

apresentação de cada exemplo de treinamento) e modo por lote (o ajuste dos pesos é realizado após a apresentação de todos os exemplos de treinamento que constituem uma época). Geralmente, o algoritmo *backpropagation* requer muitas épocas (ROSA, 2011).

Segundo Haykin (1999), o modo *online* é o mais popular, pois o algoritmo é mais simples de programar e fornece soluções efetivas a problemas grandes e difíceis.

O processo de aprendizagem termina quando o desempenho da generalização foi adequado ou quando o desempenho comprova que atingiu o máximo. No algoritmo *backpropagation* são iterados até que seja atingido um critério de parada (número máximo de ciclos ou taxa máxima de erro), geralmente esse desempenho é medido por um conjunto de validação.

A seguir, na Figura 4.6 é apresentado um resumo através de um fluxograma com as etapas do algoritmo *backpropagation*, desde a inicialização do algoritmo até a convergência.

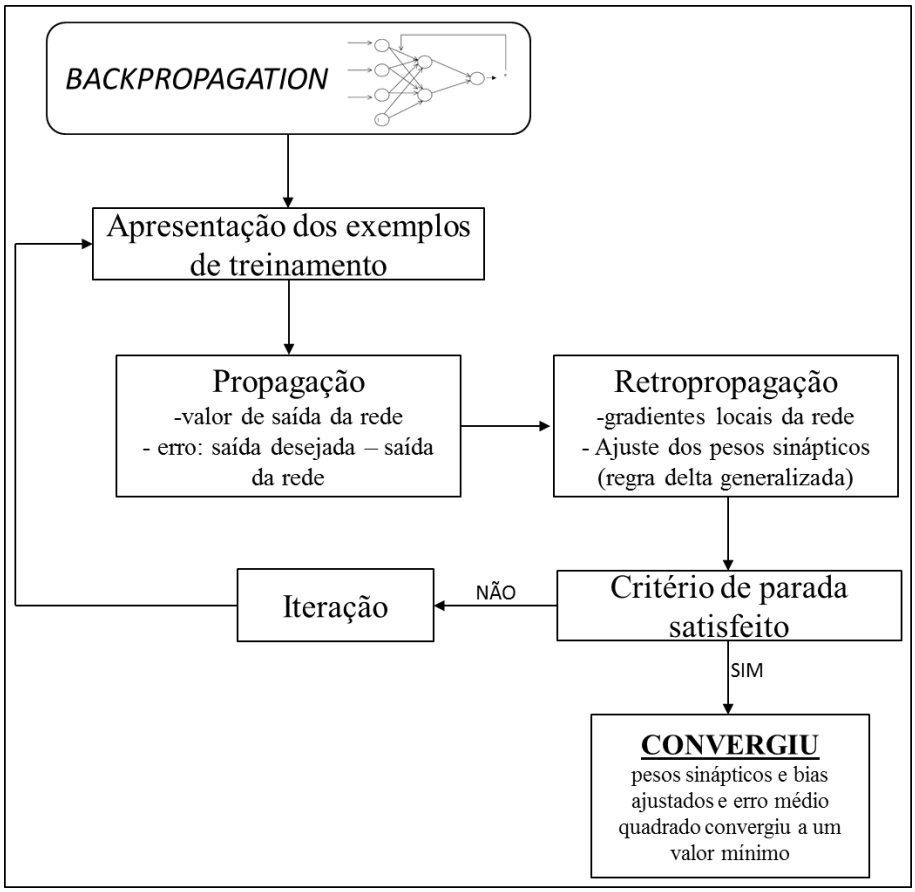


Figura 4.6 – Resumo das etapas do algoritmo *backpropagation*.

4.2.5 Aplicações das Redes Neurais Artificiais na Engenharia de Transportes

Vários modelos de RNAs têm sido propostos desde a década de 80 e a maioria dos modelos utiliza o perceptron multicamadas, redes de Hopfield e redes de auto-organização de Kohonen (HOPFIELD, 1984).

Desde a década de 90, o uso de Redes Neurais Artificiais vem sendo cada vez mais utilizada no Brasil e, desde então, diversas áreas vêm realizando trabalhos com a aplicação desta técnica. A área de Transportes se destaca no uso dessas aplicações, pois são utilizadas na solução de diversos problemas em diversas situações, dentre elas na gerência e manutenção de pavimentos, na Engenharia de Tráfego e no Planejamento de Transportes. Alguns trabalhos são descritos neste capítulo, a seguir.

Raia (2000) desenvolveu um modelo para estimar um índice potencial de viagens para o planejamento urbano através de dados da Pesquisa O/D e uso de RNAs.

Bocanegra (2002) explorou os procedimentos para tornar mais efetivo o uso das Redes Neurais Artificiais em Planejamento de Transportes e constatou em seus modelos que não haveria melhora somente variando os parâmetros internos (camadas intermediárias e taxa de aprendizado). Propôs então, excluir os dados duvidosos e criou três conjuntos de dados aleatórios para treinar as RNAs. Concluiu que os modelos melhoraram e que as RNAs permitem observar de forma clara, o comportamento dos valores de saída como consequência de variações nos dados de entrada.

No trabalho de Mark, Sadek e Rizzo (2004), os autores utilizaram as RNAs para prever o tempo de viagem entre dois pontos da rede de transportes de uma autoestrada. Já Lin, Zito e Taylor (2005) concluíram que as RNAs têm um potencial maior que outros métodos estatísticos, para estimar e prever o tempo de viagem em redes viárias urbanas.

Na área de Infraestrutura de Transportes, foi encontrado o trabalho de Bosurgi e Trifirò (2005) que utilizaram as Redes Neurais Artificiais e algoritmos genéticos, para definir o uso dos recursos econômicos disponíveis para o recapeamento de pavimentos flexíveis da melhor maneira possível e em um período de curto prazo.

Teodorović et al. (2006) desenvolveram um sistema inteligente baseado em um modelo de RNA e programação dinâmica para aumentar o tempo de verde em tempo real.

Colombaroni e Fusco (2014) propuseram uma RNA para modelar o comportamento dos motoristas de automóveis. Os veículos do experimento foram colocados em fila na via urbana, acoplados com um GPS para a coleta dos dados. Verificaram que as RNAs fornecem uma boa aproximação dos padrões de condução e pode ser adequadamente aplicados em modelos de micro simulação.

Gonçalves, Silva e d'Agosto (2015) utilizaram uma RNA para estimar a matriz origem-destino (OD) de grãos de soja no Brasil destinado à exportação a fim de explicar a variabilidade dos fluxos entre os pares O/D. Os resultados foram comparados com o modelo gravitacional desenvolvido por (SOUZA; D'AGOSTO, 2012). A RNA desenvolvida apresentou uma precisão consideravelmente maior do que os resultados obtidos pelo modelo gravitacional. Portanto, os autores sugerem a RNA como uma nova e confiável opção na distribuição de viagens de carga. Outro trabalho, que utilizou a técnica de RNA para estimativa de distribuição de viagens foi o

desenvolvido por (RASOULI; NIKRAZ, 2013).

O trabalho de Rocha et al. (2015) fez o uso das RNAs para analisar geração de viagens. Avaliaram o potencial das RNAs para estimar viagens produzidas por zona de tráfego. Utilizaram como técnica confirmatória, a regressão linear múltipla. Ambas as técnicas apresentaram modelos com bom poder preditivo. Os autores concluem que a RNA é uma técnica adequada para fins de previsão de geração de viagens, minimizando os erros das estimativas, podendo ser utilizada alternativamente às abordagens tradicionais.

4.3 Estatística e Rede Neural Artificial para pesquisas de transportes: terminologias, diferenças e similaridades.

As técnicas de RNA e Estatística são muito utilizadas na modelagem de dados de transportes. Embora, essas técnicas apresentem os mesmos objetivos, cada técnica apresenta a sua particularidade, vantagens e desvantagens. Por exemplo, nos trabalhos que aplicam técnica de RNA na modelagem de problemas de transportes, alguns apresentados na seção anterior, muitos autores utilizaram a estatística, mais especificamente, os modelos de regressão linear em seus trabalhos, como técnica confirmatória e de análise dos dados e resultados.

Assim sendo, o trabalho de Karlaftis e Vlahogianni (2011) comparou esses dois métodos em pesquisas de transporte, discutiram as diferenças e as similaridades das técnicas e apresentaram um conjunto de ideias para a escolha do método apropriado. As informações que foram consideradas relevantes para este trabalho foram sintetizadas na Tabela 4.3.

Em suma, as terminologias são apresentadas com o intuito de evitar confusões pelos pesquisadores de Engenharia de Transportes. A filosofia das duas técnicas em geral, o processo é o mesmo, independentemente da técnica utilizada; ou seja, para reconhecer e definir o problema, é necessário selecionar um método para resolvê-lo, e, em seguida, interpretar os resultados. Em relação às metas de cada técnica, para a modelagem estatística é necessário conhecer os dados e a estrutura, os elementos do modelo devem ser auto explicativos e esclarecer os fenômenos investigados através da interpretação, o que não é necessário para os modelos de RNA, pois é um mecanismo de dinâmica desconhecido. E por fim, devem ser considerados para o desenvolvimento do modelo o aprendizado, a definição e interpretação dos parâmetros e os pressupostos e limitações.

As redes neurais têm sido amplamente aplicadas a vários problemas de Engenharia de Transportes, especificamente nas áreas: operações de tráfego, gestão de infraestrutura, manutenção e reabilitação, planejamento, meio ambiente e transporte, e segurança e comportamento humano, porque são modelos matemáticos genéricos, precisos e convenientes capazes de simular facilmente os componentes numéricos do modelo. Apesar do amplo uso em pesquisa de transporte, surpreendentemente poucos trabalhos comparam os resultados da estatística com o das RNAs (KARLAFTIS; VLAHOGIANNI, 2011).

Na literatura foram encontrados trabalhos da área de planejamento, que utilizaram a rede perceptron multicamadas e compararam os resultados com os obtidos pelos modelos de regressão. São citados nesta revisão, pois ambos os assuntos tem ampla relação com o tema desta tese e servirá de apoio para o desenvolvimento dos modelos, assim são: Cai, Yin e Xie (2009); Sommer et al. (2008); Longhi et al. (2005); Chang (2005); Jeong e Rilett (2005); Tong e Hung (2002); Al-Deek (2001) e Mozolin, Thill e Usery (2000).

Na Engenharia de Transportes, os pesquisadores tentam encontrar a melhor abordagem de modelagem para os dados, sendo também um dos objetivos desta tese.

Karlaftis e Vlahogianni (2011) recomendam que os pesquisadores se preocupem com a precisão e facilidade de interpretação dos resultados; com o conhecimento prévio sobre o problema a ser estudado; com o tipo de modelo a ser desenvolvido dependo da precisão desejada e com o adequado projeto e avaliação da RNA desenvolvida. Além disso, relatam que ambas as técnicas respondem perguntas sobre os efeitos das variáveis independentes (*inputs*) sobre a variável dependente (*outputs*). Afirmam também, que o aumento no poder computacional torna-se possível resolver problemas mais complexos e não lineares através de técnicas mais avançadas, por exemplo, as RNAs, porém devido algumas limitações, em determinados casos o uso de modelos mais simples podem gerar resultados tão bons quantos os modelos mais complexos.

Tabela 4.3 – Estatística versus RNA (MLP).

DIFERENÇAS	ESTATÍSTICA	RNA (MLP)
TERMINOLOGIAS	variáveis independentes	entrada/saída
	variáveis dependentes	valores de treinamento
	resíduos	erros
	estimação	treinamento, aprendizagem, adaptação e auto organização
	critério de estimação	função erro, função custo
	observações	padrões de treinamento
	estimativas dos parâmetros	pesos sinápticos
	interações	neurônios de ordens elevadas
	transformações	sinais funcionais
	análise de regressão	aprendizagem supervisionada
	redução dos dados	aprendizagem não supervisionada
	análises clusters	aprendizagem competitiva
FILOSOFIAS	não conseguem lidar com dados complexos e lidam com dados lineares	dados complexos e não lineares
	enfatiza a inferência e a determinação e os dados são gerados a partir de processo estocástico	implementação e o processo de geração dos dados é um mecanismo de dinâmica desconhecido
METAS	modelo preditor ou classificador, sendo necessário conhecer os dados e a estrutura, os elementos devem ser auto explicativos e esclarecer os fenômenos investigados através da interpretação	não é necessário um conhecimento prévio dos dados e da estrutura e proporciona uma eficiente representação dos dados (precisão e tempo de desenvolvimento) e boas previsões para o fenômeno em estudo.
MODELO	um único modelo final	independentemente do método utilizado (supervisionado ou não supervisionado, máxima verossimilhança), resulta em mais de um modelo
Aprendizado	modelagem menos flexível pois a forma funcional é assumida a priori	modelagem mais flexível- a curva de aprendizado pode ter vários mínimos locais e pode convergir para várias arquiteturas sequenciais que não são necessariamente aninhadas
	considera importante a interpretação dos dados antes da modelagem	mecanismo de inferência é oculto "caixas pretas" e são desconsideradas as suposiçoes implícitas
Definição e interpretação dos parâmetros	são definidos em termos dos modelos matemáticos usados e das propriedades estatísticas dos resultados	são definidos pela a arquitetura e o algoritmo de aprendizagem
Pressupostos e limitações	é feito uma série de hipóteses e colocadas restrições no desenvolvimento	não é necessário nenhum pressuposto ou restrição
	são especificadas a priori as hipóteses em relação ao termo de erro	os parâmetros são extremamente adaptáveis, poucos pressupostos são feitos em relação ao termo de erro
	problema da multicolinearidade dos dados	não possui esta limitação
	incapaz de lidar com valores extremos, dados ausentes ou ruidosos	capaz de lidar com valores extremos, dados ausentes ou ruidosos
SIMILARIDADES	**ESTATÍSTICA**	**RNA (MLP)**
TOPOLOGIAS	modelos lineares generalizados	rede *feed-forward* sem camada escondida
	regressão linear multivariada	Perceptron linear simples
	regressão logística	Perceptron não linear simples
	regressão por busca de projeção	rede *feed-forward* com uma camada escondida
	regressão Kernel	redes neurais de regressão generalizada
	análise discriminante Kernel	rede probabilística
	k-means clustering	rede de aprendizagem competitiva
	aproximações discretas para principais curvas e superfícies	Mapas de auto-organização de Kohonen
	regressão de componentes principais	redes híbridas (aprendizado supervisionado e não supervisionado)
	análises de componentes principais	aprendizado Hebbiano

Capítulo 5

MATERIAIS E MÉTODO

Nos capítulos 2, 3 e 4 foram discutidos os referenciais teóricos que fundamentam as propostas desta tese. Este capítulo descreve os materiais e método aplicados para o desenvolvimento deste texto. Foram utilizados como banco de dados a Pesquisa Origem e Destino e o Censo Demográfico do IBGE e os pacotes computacionais: o Visual Basic for Application (VBA) e o IBM SPSS 22. Posteriormente aos materiais, é apresentado o detalhamento da proposta elaborada para atingir os objetivos da tese. Está dividida em duas partes: (1) geração da população sintética e (2) modelagem da demanda por transportes.

5.1 Materiais

5.1.1 Banco de dados e área de estudo

Os dados utilizados neste trabalho são referentes a dois levantamentos: (1) a base de dados da Pesquisa Origem e Destino (O/D) realizada na cidade de São Carlos-SP entre os anos de 2007 e 2008 pelo Departamento de Engenharia de Transportes da Escola de Engenharia de São Carlos-SP (EESC) e (2) a base de dados do Censo Demográfico 2010 realizado pelo Instituto de Geografia e Estatística (IBGE). A Pesquisa O/D fornece dados das informações de indivíduos ou domicílios, conhecidos como *dados desagregados.* O Censo Demográfico do IBGE disponibiliza tanto *dados agregados* (informações por setores censitários) quanto *dados desagregados* (microdados 2010 - informações por indivíduos ou domicílios).

A Pesquisa O/D (entrevista domiciliar) foi aplicada a todos os moradores dos domicílios amostrados e coletadas informações geográficas, informações sociodemográficas individuais (idade, gênero, situação do emprego, ocupação), informações dos recursos disponíveis no do-

micílio (propriedade de automóvel, características do domicílio, renda), dados espaciais e não espaciais dos locais de realização das atividades (localização, horário de funcionamento dos estabelecimentos, nível de acessibilidade) e dados referentes à malha viária. Para a efetivação das entrevistas domiciliares, foram selecionados aleatoriamente 5% dos domicílios contidos na base de informações cadastrais cedidas pelo Serviço Autônomo de Água e Esgoto (SAAE) da cidade. Dessa porcentagem selecionada, 98% dos domicílios entrevistados foram considerados válidos (SILVA, 2008).

Deste modo, o banco de dados da Pesquisa O/D é composto originalmente por 10.085 moradores e 3.057 domicílios. Vale ressaltar que os dados dos domicílios foram os dados pertencentes à amostra e não foram expandidos.

O Censo Demográfico do IBGE fornece informações sobre a população brasileira, tais como, as características socioeconômicas dos domicílios ou indivíduos (microdados-dados desagregados) e dos setores censitários (dados agregados). Essas informações são atualizadas a cada dez anos. O IBGE (2010) definiu que microdados consistem no menor nível de desagregação dos dados de uma pesquisa, retratando, sob a forma de códigos numéricos, o conteúdo dos questionários, preservado o sigilo estatístico com vistas a não individualização das informações. Os microdados 2010 estão no formato ASCII, possibilitando aos usuários especializados, com conhecimento em programação, preferencialmente em softwares estatísticos, a leitura dos dados, o cruzamento em diferentes agregações geográficas e a elaboração de múltiplas tabulações segundo sua perspectiva pessoal de interesse. E setor censitário é a unidade territorial de coleta das operações censitárias, com limites físicos identificados, em áreas contínuas e respeitando a divisão político-administrativa do Brasil (IBGE, 2010).

5.1.1.1 A cidade de São Carlos-SP

Localizada no centro do estado de São Paulo e conhecida como um importante polo científico e tecnológico do Brasil com centros de ensino e pesquisa e várias empresas de alta tecnologia. Fundada em 1857 e com uma população de, aproximadamente 240.000 habitantes no ano de 2015, São Carlos-SP é considerada um centro predominantemente urbano, contando apenas com 4% da população residente na área rural (IBGE, 2016). Assim, como outras cidades brasileiras de médio porte, São Carlos-SP apresentou um intenso crescimento nos últimos anos, com uma renda per capita média de R$ 1.086,22 e possuía de um alto nível de IDH igual a 0,805 no ano de 2010, enquanto que no Estado de São Paulo e no país, no mesmo ano, os valores eram 0,783 e 0,813, respectivamente (PNUD, 2016); (IBGE, 2016).

Na época da realização do Censo Demográfico 2010, a cidade de São Carlos-SP contava com uma população aproximada de 212.956 habitantes em 68.833 domicílios distribuídos em 288 setores censitários na área urbanizada da cidade. A Figura 5.1 mostra a localização da cidade de São Carlos-SP e a divisão da cidade em setores censitários. De forma mais detalhada o Anexo A apresenta o mapa em formato (.dbd) dos setores censitários do Censo (2010).

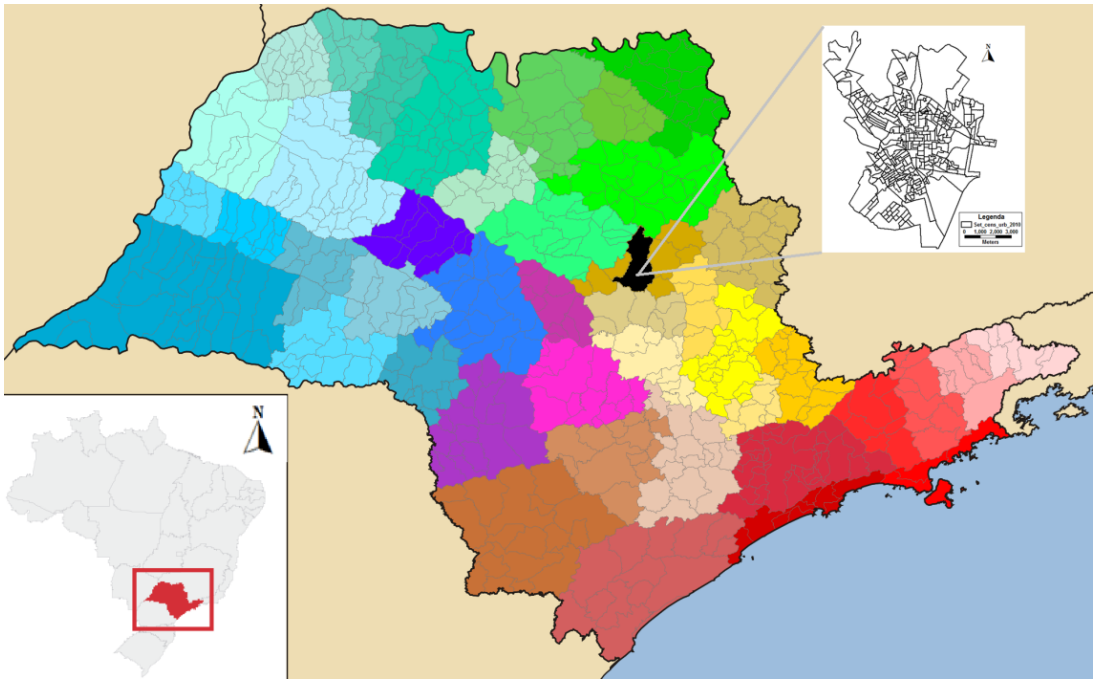


Figura 5.1 – Localização da cidade de São Carlos-SP e os setores censitários.

Fazendo uma caracterização sucinta da cidade de São Carlos-SP, pode-se afirmar que:

- O valor do rendimento nominal médio mensal das pessoas consideradas economicamente ativas é de R$ 1.780,00.

- A cidade de São Carlos -SP apresenta grande parte da população em idade escolar e economicamente ativa, como mostra a pirâmide etária presente na Figura 5.2.

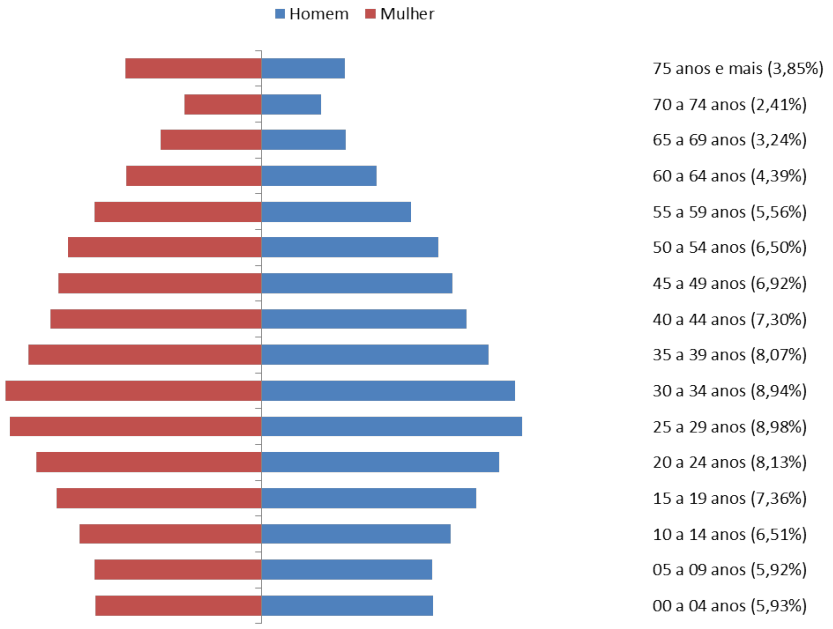


Figura 5.2 – Pirâmide Etária –São Carlos (SP).

Nota-se que a distribuição etária não é uniforme, ou seja, a maioria da população considerada infantil (0-12 anos incompletos) se localiza nas regiões periféricas da cidade e a população idosa (60 anos ou mais) se concentra nas regiões centrais, como mostra a Figura 5.3.

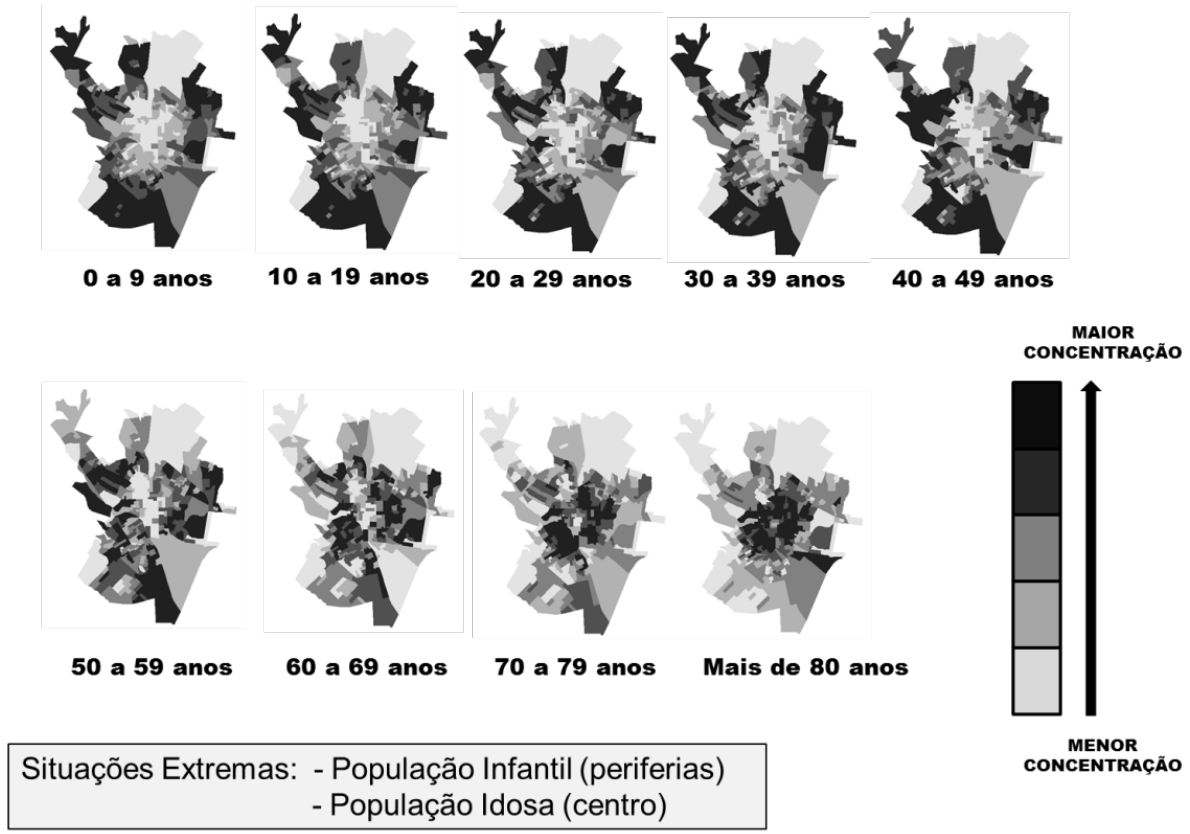


Figura 5.3 – Distribuição da população conforme a faixa etária.

- Quanto ao grau de instrução, 29% do total da amostra possui ensino fundamental incompleto, sendo 28% do total de homens e 30% do total de mulheres e 24% da amostra possui ensino médio completo, sendo 25% do total de homens e 23% do total de mulheres. Este valor do ensino médio, pode estar refletindo o grande número de estudantes universitários da cidade de São Carlos-SP.

- Quanto à condição da atividade, apresentaram um índice de 34% dos entrevistados da Pesquisa O/D indicaram que estavam trabalhando.

- O modo de viagem predominante dos entrevistados da Pesquisa O/D era de automóvel (83%).

- A maior parte dos domicílios entrevistados (59%) possui pelo menos um automóvel, enquanto o número de domicílios sem automóveis corresponde a 41% do total.

-Em relação as informações de viagens dos moradores da cidade de São Carlos-SP, o principal motivo é o retorno para casa (45%), seguido das viagens realizadas por motivo de trabalho, que somam 21% do total de viagens e das viagens por motivo de estudo (17%).

- Quanto ao modo de viagem, 29% das viagens são realizadas a pé e dirigindo automóvel. A porcentagem de viagens realizadas por ônibus (transporte público) segundo a pesquisa por entrevistas domiciliares foi de 16%.

5.1.2 *Software*

Os resultados obtidos neste trabalho valeram-se de um conjunto de ferramentas computacionais, que auxiliaram em diversos campos, tais como: simulação Monte Carlo, análises estatísticas tradicionais e técnicas computacionais avançadas, como a modelagem por Redes Neurais Artificias.

Com relação à simulação Monte Carlo, foi através dela que foi gerada a população sintética proposta nesta tese, codificada em *Visual Basic for Application* (VBA), linguagem de programação do programa Microsoft Excel.

No campo da modelagem de demanda por transportes deste trabalho foram aplicadas duas técnicas: (1) Regressão Linear Múltipla (RLM) e (2) Redes Neurais Artificiais (RNAs). Ambas as técnicas foram realizadas no *software* IBM SPSS *Statistics* 22.

O SPSS foi criado em 1968 e é um dos programas de análise estatística mais usado nas ciências sociais. Foi criado por Norman H. Nie, C. Hadlai (Tex) Hull e Dale H. Bent. Nos anos 70 a Universidade de Chicago ficou responsável pelo desenvolvimento, distribuição e venda do referido *software*. Este *software* é do tipo científico que foi criado para operar o processo analítico inteiro, desde o planejamento e a coleta de dados até a análise, o relatório e a implementação. Foram utilizados nesta tese os módulos de Regressão, Redes Neurais Artificiais e Testes Estatísticos.

De um modo geral, os módulos apresentam diversas funcionalidades que são escolhidas de acordo com o objetivo do trabalho a ser desenvolvido. Para este trabalho, no módulo de regressão foi utilizada a regressão do tipo linear múltipla, no módulo de redes neurais foi utilizado o perceptron multicamadas e no módulo de testes estatísticos foi usado o teste não paramétrico de amostras independentes.

Os aspectos relevantes para o uso do IBM SPSS 22 são justificados pelo fato deste programa estar disponível para pesquisas realizadas pela Universidade de São Paulo, por acessar e analisar rapidamente grandes volumes de dados, além de gerenciar e analisar qualquer tipo de conjunto de dados e principalmente por possuir a ferramenta de RNAs.

5.2 Método

O processo metodológico adotado se baseia nas etapas ilustradas na Figura 5.4 e descritas na sequência.

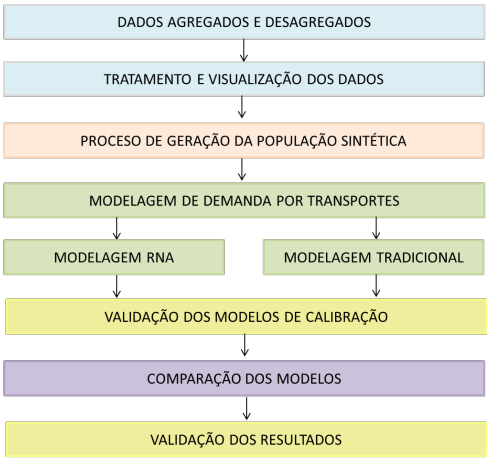


Figura 5.4 – Esquema simplificado do método proposto.

5.2.1 Tratamento e visualização dos dados agregados e desagregados

O propósito desta etapa é apresentar a fonte das variáveis e o procedimento de seleção das mesmas para a construção da base de dados utilizadas no processo de verificação da hipótese deste trabalho. Os procedimentos descritos nesta etapa permitem que outros pesquisadores reapliquem os mesmos procedimentos para a verificação da hipótese desta tese em outro contexto urbano. Ao final do processo de tratamento dos dados foram formadas três bases de dados, conforme apresentado na Figura 5.5.

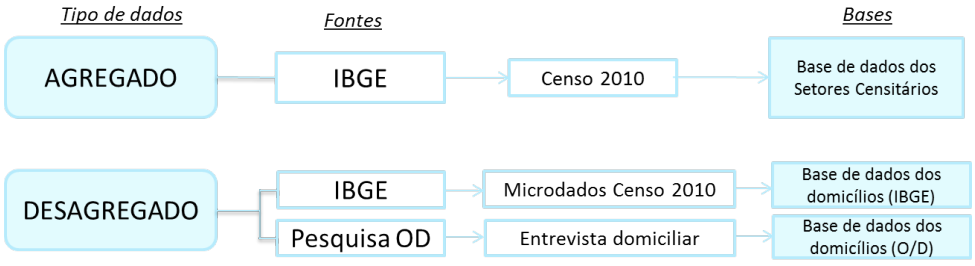


Figura 5.5 – Bases de dados obtidas.

5.2.1.1 Dados agregados

Os dados agregados foram obtidos na base do Censo Demográfico de 2010 relativos a cidade São Carlos-SP e está disponível no site do IBGE. As variáveis contidas no referido censo estão detalhadas no Apêndice A. A fase de tratamento dos dados dos setores censitários iniciou com a eliminação dos setores que não estavam contidos na área urbanizada da cidade e os que não possuíam domicílios e, depois foram selecionados somente os domicílios particulares permanentes contidos nesses setores.

Em se tratando de um estudo de produção de viagens por domicílio, as variáveis que realmente explicam o fenômeno, para o caso de modelagem desagregada são: renda, tamanho da família, posse de automóvel e grau parentesco (ORTÚZAR; WILLUMSEN, 2011). Porém, como o

objetivo principal é obter uma população sintética através de dados agregados de um censo, são selecionadas as informações que estão descritas na Tabela 5.1 para classificar de forma mais detalhada os moradores dos domicílios a serem gerados, por isso as variáveis idade e sexo são selecionadas nesta tese.

Tabela 5.1 – Variáveis disponíveis no censo demográfico 2010 (IBGE).

VARIÁVEIS INDEPENDENTES	NOME	TIPO	TOTAL
Domicílios com 1 morador	x_2	Numérica	9.053
Domicílios com 2 moradores	x_3	Numérica	17.186
Domicílios com 3 moradores	x_4	Numérica	18.094
Domicílios com 4 moradores	x_5	Numérica	14.636
Domicílios com 5 moradores	x_6	Numérica	6.120
Domicílios com 6 moradores	x_7	Numérica	2.218
Domicílios com 7 moradores	x_8	Numérica	849
Domicílios com 8 moradores	x_9	Numérica	339
Domicílios com 9 moradores	x_{10}	Numérica	161
Domicílios com 10 moradores	x_{11}	Numérica	177
NÚMERO DE DOMICÍLIOS			68.833
Número de mulheres	x_{12}	Numérica	108.460
Número de homens	x_{13}	Numérica	103.803
Número de chefes	x_{14}	Numérica	68.922
Número de cônjuges	x_{15}	Numérica	45.357
Número de filhos	x_{16}	Numérica	71.472
Número de outros parentes	x_{17}	Numérica	24.462
Número de agregados	x_{18}	Numérica	324
Número de empregados	x_{19}	Numérica	1.675
Número de visitantes	x_{20}	Numérica	51
NÚMERO DE MORADORES			212.263
Idade até 10 anos	x_{21}	Numérica	28.825 moradores
Idade de 11-20 anos	x_{22}	Numérica	32.198 moradores
Idade de 21-30 anos	x_{23}	Numérica	38.973 moradores
Idade de 31-40 anos	x_{24}	Numérica	33.964moradores
Idade de 41-50 anos	x_{25}	Numérica	29.783 moradores
Idade de 51-60 anos	x_{26}	Numérica	23.448 moradores
Idade 61-70 anos	x_{27}	Numérica	13.700 moradores
Idade de 71-80 anos	x_{28}	Numérica	8.283 moradores
Idade maior de 80 anos	x_{29}	Numérica	3.627 moradores
Sem renda	x_{30}	Numérica	1283 domicílios
Renda 0 - 2 Salários Mínimos	x_{31}	Numérica	47.307 domicílios
Renda 2 - 3 Salários Mínimos	x_{32}	Numérica	8.837 domicílios
Renda 3 - 5 Salários Mínimos	x_{33}	Numérica	6.349 domicílios
Renda 5 - 10 Salários Mínimos	x_{34}	Numérica	3.750 domicílios
Renda 10 ou mais Salários Mínimos	x_{35}	Numérica	1.307 domicílios

De acordo com os dados apresentados na Tabela 5.1 nota-se que a maioria dos domicílios contidos nos setores censitários possuem de dois a quatro moradores e a maioria dos domicílios

(47.307) possuem renda de 0 a 2 salários mínimos. Também, pode-se observar que o total das idades dos moradores apresenta um valor maior que o total de moradores, pois o IBGE disponibiliza as idades das pessoas residentes em domicílios particulares e domicílios coletivos.

A Tabela 5.2 apresenta a caracterização das variáveis quantitativas através das medidas descritivas de média ($\bar{x}$), desvios padrão (σ), mínimo (min.), máximo (max.), quartil 25 (Q_{25}), quartil 50 (Q_{50}) e quartil 75 (Q_{75}).

Tabela 5.2 – Medidas descritivas para as variáveis quantitativas.

variáveis	$\bar{x}$	σ	**Min.**	**Max.**	Q_{25}	Q_{50}	Q_{75}
x_2	31,43	15,69	0,00	91,00	21,00	29,00	37,00
x_3	59,67	23,04	5,00	125,00	43,00	57,00	75,00
x_4	62,83	29,87	1,00	160,00	41,00	60,00	81,00
x_5	50,82	26,50	3,00	147,00	32,00	47,00	68,25
x_6	21,25	13,03	0,00	80,00	12,75	20,00	28,00
x_7	7,70	6,68	0,00	43,00	3,00	6,00	10,25
x_8	2,95	2,98	0,00	19,00	1,00	2,00	4,00
x_9	1,18	1,57	0,00	11,00	0,00	1,00	2,00
x_{10}	0,56	0,95	0,00	6,00	0,00	0,00	1,00
x_{11}	0,62	1,23	0,00	10,00	0,00	0,00	1,00
x_{12}	376,60	163,92	17,00	918,00	257,25	370,00	473,00
x_{13}	360,43	169,34	16,00	973,00	244,75	348,50	457,00
x_{14}	239,31	94,52	11,00	553,00	169,00	239,00	296,00
x_{15}	157,49	73,57	9,00	422,00	107,50	150,00	201,00
x_{16}	248,17	131,78	10,00	757,00	154,00	231,00	323,50
x_{17}	84,94	49,13	3,00	333,00	55,00	78,00	106,25
x_{18}	1,13	1,97	0,00	16,00	0,00	0,00	2,00
x_{19}	5,82	7,78	0,00	64,00	1,00	3,00	7,00
x_{20}	0,18	0,57	0,00	4,00	0,00	0,00	0,00
x_{21}	100,09	72,22	5,00	423,00	50,00	82,50	132,50
x_{22}	111,80	70,31	2,00	396,00	61,00	98,50	146,00
x_{23}	135,32	64,07	2,00	342,00	91,50	131,00	174,25
x_{24}	117,93	68,29	8,00	372,00	67,75	102,00	151,00
x_{25}	103,41	51,00	7,00	286,00	65,75	98,00	134,00
x_{26}	81,42	34,34	4,00	180,00	58,75	79,50	105,00
x_{27}	47,57	21,88	1,00	116,00	31,75	46,00	61,25
x_{28}	28,76	16,75	0,00	76,00	17,00	26,00	38,25
x_{29}	12,59	9,88	0,00	52,00	5,75	10,00	16,25
x_{30}	4,46	6,73	0,00	89,00	1,00	3,00	5,25
x_{31}	164,26	92,67	0,00	500,00	103,75	155,00	213,75
x_{32}	30,68	17,28	0,00	101,00	18,00	28,00	42,00
x_{33}	22,05	18,73	0,00	100,00	8,00	17,00	30,00
x_{34}	13,02	17,03	0,00	101,00	2,00	6,00	17,25
x_{35}	4,54	8,75	0,00	67,00	0,00	1,00	5,00

De acordo com os dados da Tabela 5.2 nota-se que cada setor censitário apresenta na média

63 e no máximo de 160 domicílios com 3 moradores. E a partir, dos domicílios com mais de 5 moradores, a média diminui significativamente, ou seja, o número máximo de domicílios com 7 moradores no setor censitário é 84,8% menor que o número máximo de domicílios com 2 moradores.

Vale ressaltar que, as variáveis selecionadas estão estratificas conforme aparecem disponíveis no censo demográfico, com o objetivo de facilitar para o planejador sem que precise fazer quaisquer modificações nas variáveis de entrada para a aplicação deste método em uma cidade semelhante a São Carlos-SP, justificando o uso dessas variáveis estratificadas na modelagem da produção de viagens por domicílio.

A variável *viagens por domicílio* não constava nos dados agregados do Censo Demográfico de 2010 do IBGE, pois o objetivo do censo é obter dados demográficos da população e não informações sobre as características das viagens. Além disso, esta variável, geralmente é obtida a partir de pesquisas de tráfego, mas devido aos inúmeros fatores que tornam inviáveis a realização deste tipo de pesquisa (custo elevado e longo tempo de preparação e execução), o método proposto nesta tese permitiu a estimação da variável *viagens por domicílio*.

Portanto, esta base de dados agregados do IBGE (Apêndice E.1) foi utilizada no método para obter a população sintética e que a partir desta população sintética, após os modelos propostos serem definidos foi então, obtida a variável *viagens por domicílio*, não apenas de uma amostra da população, mas da população total (população do censo do IBGE-2010).

5.2.1.2 Dados desagregados

Os dados desagregados dos domicílios foram obtidos a partir de dois levantamentos: (1) Pesquisa Origem e Destino (O/D) e (2) microdados do Censo Demográfico IBGE-2010.

(1) Pesquisa Origem e Destino (O/D)

Os dados obtidos pela Pesquisa O/D foram referentes às características socioeconômicas e aos deslocamentos de pessoas na cidade de São Carlos-SP. O detalhamento das informações que foram coletadas na entrevista domiciliar está apresentado no Anexo B.

Os dados desagregados originalmente apresentavam 10.085 registros, em que cada registro representava um indivíduo do domicílio, com suas características socioeconômicas (sexo, idade, grau de instrução, renda, número de veículos, etc.) e de viagens (motivo da viagem, modo principal, quantidade de viagens no domicílio, etc.). Essas informações estão descritas no Apêndice B.

As informações das variáveis da Pesquisa O/D de interesse para esta tese estão descritas na Tabela 5.3, juntamente com a natureza das variáveis e os valores totais por domicílios (desagregados). Contudo, foram selecionadas da Pesquisa O/D as mesmas variáveis que foram escolhidas no banco do Censo Demográfico de 2010 (dados agregados) para futuras comparações, além da variável *viagens por domicílio*, disponível apenas nessa base de dados.

Tabela 5.3 – Informações utilizadas da Pesquisa O/D da cidade de São Carlos-SP.

VARIÁVEIS INDEPENDENTES	NOME	TIPO	TOTAL
Viagens por domicílio	x_1	Numérica	14.702 viagens
Domicílios com 1 morador	x_2	Binária (0-1)	270
Domicílios com 2 moradores	x_3	Binária (0-1)	760
Domicílios com 3 moradores	x_4	Binária (0-1)	756
Domicílios com 4 moradores	x_5	Binária (0-1)	690
Domicílios com 5 moradores	x_6	Binária (0-1)	356
Domicílios com 6 moradores	x_7	Binária (0-1)	140
Domicílios com 7 moradores	x_8	Binária (0-1)	56
Domicílios com 8 moradores	x_9	Binária (0-1)	16
Domicílios com 9 moradores	x_{10}	Binária (0-1)	9
Domicílios com 10 moradores	x_{11}	Binária (0-1)	4
NÚMERO DE DOMICÍLIOS			3.057
Número de mulheres	x_{12}	Numérica	5.226
Número de homens	x_{13}	Numérica	4.859
Número de chefes	x_{14}	Numérica	3.288
Número de cônjuges	x_{15}	Numérica	2.164
Número de filhos	x_{16}	Numérica	3.552
Número de outros parentes	x_{17}	Numérica	988
Número de agregados	x_{18}	Numérica	63
Número de empregados	x_{19}	Numérica	10
Número de visitantes	x_{20}	Numérica	7
NÚMERO DE MORADORES			10.085
Idade com até 10 anos	x_{21}	Numérica	1.309 moradores
Idade de 11-20 anos	x_{22}	Numérica	1.548 moradores
Idade de 21-30 anos	x_{23}	Numérica	1.748 moradores
Idade de 31-40 anos	x_{24}	Numérica	1.334 moradores
Idade de 41-50 anos	x_{25}	Numérica	1.310 moradores
Idade de 51-60 anos	x_{26}	Numérica	1.130 moradores
Idade de 61-70 anos	x_{27}	Numérica	946 moradores
Idade de 71-80 anos	x_{28}	Numérica	547 moradores
Idade maior de 80 anos	x_{29}	Numérica	213 moradores
Sem renda	x_{30}	Numérica	3.527 moradores
Renda 0 - 2 Salários Mínimos	x_{31}	Numérica	2.392 moradores
Renda 2 - 3 Salários Mínimos	x_{32}	Numérica	1.102 moradores
Renda 3 - 5 Salários Mínimos	x_{33}	Numérica	291 moradores
Renda 5 - 10 Salários Mínimos	x_{34}	Numérica	153 moradores
Renda 10 ou mais Salários Mínimos	x_{35}	Numérica	64 moradores

Vale ressaltar que, devido ao objetivo deste estudo, as informações foram agrupadas por domicílios.

De acordo com os dados apresentados na Tabela 5.3, algumas variáveis não foram totalmente obtidas, como a renda e a situação domiciliar. A renda provavelmente foi por motivos de privacidade de alguns entrevistados e em relação a situação domiciliar, pode ter acontecido algum erro na tabulação dos dados.

O método proposto nesta tese foi possível de ser realizado, pois existia uma Pesquisa O/D na cidade de São Carlos – SP. Além disso,os modelos propostos foram calibrados utilizando os dados selecionados desta pesquisa (Apêndice E.2).

Sem as informações de viagens, apenas com a base agregada ou desagregada do censo do IBGE 2010 não seria possível a realização do método proposto neste texto. A Tabela 5.4 apresenta a caracterização das variáveis quantitativas através das medidas descritivas de média ($\bar{x}$), desvios padrão (σ), mínimo (min.), máximo (max.), quartil 25 (Q_{25}), quartil 50 (Q_{50}) e quartil 75 (Q_{75}).

Tabela 5.4 – Medidas descritivas para as variáveis quantitativas.

variáveis	$\bar{x}$	σ	**Min.**	**Max.**	Q_{25}	Q_{50}	Q_{75}
x_1	4,81	4,26	0,00	39,00	2,00	4,00	7,00
x_{12}	1,71	1,01	0,00	9,00	1,00	2,00	2,00
x_{13}	1,59	1,05	0,00	7,00	1,00	1,00	2,00
x_{14}	1,08	0,49	0,00	9,00	1,00	1,00	1,00
x_{15}	0,71	0,48	0,00	4,00	0,00	1,00	1,00
x_{16}	1,16	1,15	0,00	6,00	0,00	1,00	2,00
x_{17}	0,32	0,77	0,00	9,00	0,00	0,00	0,00
x_{18}	0,02	0,22	0,00	6,00	0,00	0,00	0,00
x_{19}	0,00	0,06	0,00	1,00	0,00	0,00	0,00
x_{20}	0,00	0,05	0,00	1,00	0,00	0,00	0,00
x_{21}	0,43	0,77	0,00	7,00	0,00	0,00	1,00
x_{22}	0,51	0,81	0,00	7,00	0,00	0,00	1,00
x_{23}	0,57	0,86	0,00	7,00	0,00	0,00	1,00
x_{24}	0,44	0,70	0,00	4,00	0,00	0,00	1,00
x_{25}	0,43	0,67	0,00	3,00	0,00	0,00	1,00
x_{26}	0,37	0,64	0,00	3,00	0,00	0,00	1,00
x_{27}	0,31	0,61	0,00	3,00	0,00	0,00	0,00
x_{28}	0,18	0,45	0,00	3,00	0,00	0,00	0,00
x_{29}	0,07	0,29	0,00	3,00	0,00	0,00	0,00
x_{30}	1,15	1,28	0,00	9,00	0,00	1,00	2,00
x_{31}	0,78	0,99	0,00	7,00	0,00	0,00	1,00
x_{32}	0,36	0,66	0,00	6,00	0,00	0,00	1,00
x_{33}	0,10	0,34	0,00	3,00	0,00	0,00	0,00
x_{34}	0,05	0,24	0,00	3,00	0,00	0,00	0,00
x_{35}	0,02	0,17	0,00	3,00	0,00	0,00	0,00

Assim, a média de viagens por domicílio da amostra da Pesquisa O/D foi de 4,81 e o número máximo foi de 39 viagens. Em relação ao valor de 39 viagens, provavelmente ocorreu um erro na tabulação dos dados.

(2) Microdados do Censo Demográfico IBGE-2010

No banco de microdados 2010 disponibilizados pelo IBGE, as informações são fornecidas sob a forma de códigos numéricos, devido à privacidade dos dados fornecidos pelos entrevistados. Assim, para a obtenção dos dados desagregados por domicílio (Apêndice E.3), neste trabalho, optou-se por utilizar o *software* IBM SPSS *Statistics* 22 onde foram realizadas as seis etapas apresentadas na Figura 5.6.

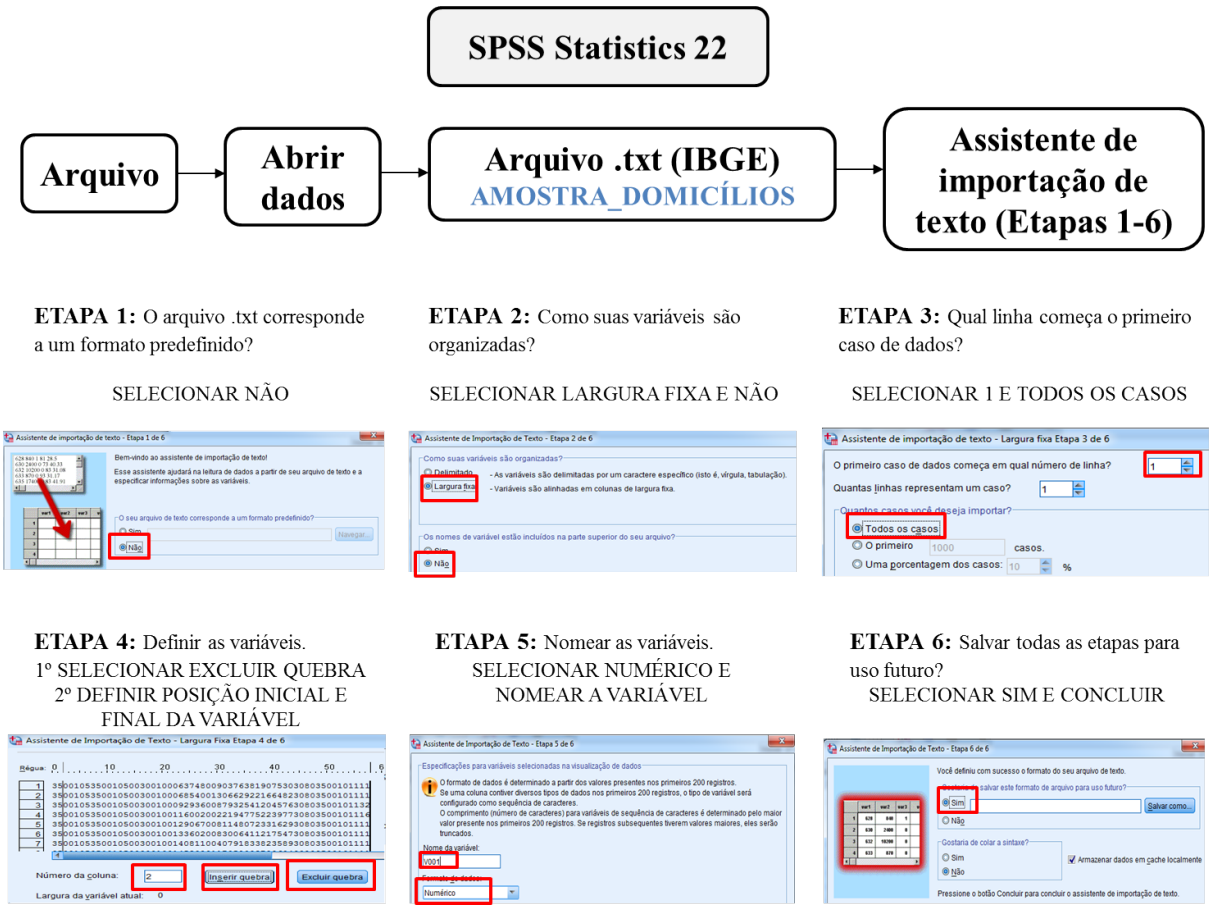


Figura 5.6 – Etapas para obtenção dos microdados 2010 dos domicílios coletados pelo IBGE.

As variáveis contidas na base de microdados 2010 estão detalhadas no Apêndice C e da mesma forma que na base da Pesquisa O/D foram selecionadas as mesmas variáveis utilizadas na base de dados agregada do IBGE.

Lembrando que este trabalho se trata de um estudo de produção de viagens por domicílios, as informações de interesse para esta tese estão descritas na Tabela 5.5, juntamente com a natureza das variáveis e os valores totais por domicílios (desagregados). Da mesma forma, que a variável *viagens por domicílio* não constava no banco de dados agregado do IBGE, também

não constava na base de microdados 2010. Neste caso, a base de dados desagregada do IBGE serviu para validar os dados obtidos pela população sintética.

Tabela 5.5 – Informações utilizadas da base de microdados 2010 do IBGE.

VARIÁVEIS INDEPENDENTES	NOME	TIPO	TOTAL
Domicílios com 1 morador	x_2	Binária (0-1)	864
Domicílios com 2 moradores	x_3	Binária (0-1)	1.695
Domicílios com 3 moradores	x_4	Binária (0-1)	1.804
Domicílios com 4 moradores	x_5	Binária (0-1)	1.459
Domicílios com 5 moradores	x_6	Binária (0-1)	619
Domicílios com 6 moradores	x_7	Binária (0-1)	229
Domicílios com 7 moradores	x_8	Binária (0-1)	83
Domicílios com 8 moradores	x_9	Binária (0-1)	36
Domicílios com 9 moradores	x_{10}	Binária (0-1)	14
Domicílios com 10 moradores	x_{11}	Binária (0-1)	14
NÚMERO DE DOMICÍLIOS			6.817
Número de mulheres	x_{12}	Numérica	10.817
Número de homens	x_{13}	Numérica	10.311
Número de chefes	x_{14}	Numérica	6.817
Número de cônjuges	x_{15}	Numérica	4.532
Número de filhos	x_{16}	Numérica	7.206
Número de outros parentes	x_{17}	Numérica	2.379
Número de agregados	x_{18}	Numérica	35
Número de empregados	x_{19}	Numérica	154
Número de visitantes	x_{20}	Numérica	5
NÚMERO DE MORADORES			21.128
Idade até 10 anos	x_{21}	Numérica	2.890 moradores
Idade de 11-20 anos	x_{22}	Numérica	3.300 moradores
Idade de 21-30 anos	x_{23}	Numérica	3.778 moradores
Idade de 31-40 anos	x_{24}	Numérica	3.378moradores
Idade de 41-50 anos	x_{25}	Numérica	2.916 moradores
Idade de 51-60 anos	x_{26}	Numérica	2.314 moradores
Idade de 61-70 anos	x_{27}	Numérica	1.382 moradores
Idade de 71-80 anos	x_{28}	Numérica	824moradores
Idade maior de 80 anos	x_{29}	Numérica	346 moradores
Sem renda	x_{30}	Binária (0-1)	113 domicílios
Renda 0 - 2 Salários Mínimos	x_{31}	Binária (0-1)	4.422 domicílios
Renda 2 - 3 Salários Mínimos	x_{32}	Binária (0-1)	948 domicílios
Renda 3 - 5 Salários Mínimos	x_{33}	Binária (0-1)	706 domicílios
Renda 5 - 10 Salários Mínimos	x_{34}	Binária (0-1)	441 domicílios
Renda 10 ou mais Salários Mínimos	x_{35}	Binária (0-1)	187 domicílios

A Tabela 5.6 apresenta a caracterização das variáveis quantitativas através das medidas descritivas de média ($\bar{x}$), desvios padrão (σ), mínimo (min.), máximo (max.), quartil 25 (Q_{25}), quartil 50 (Q_{50}) e quartil 75 (Q_{75}).

Tabela 5.6 – Medidas descritivas para as variáveis quantitativas.

variáveis	$\bar{x}$	σ	Min.	Max.	Q_{25}	Q_{50}	Q_{75}
x_{12}	1,59	0,97	0,00	9,00	1,00	0,00	2,00
x_{13}	1,51	1,01	0,00	9,00	1,00	0,00	2,00
x_{14}	1,00	0,00	1,00	1,00	1,00	1,00	1,00
x_{15}	0,67	0,47	0,00	1,00	0,00	0,00	1,00
x_{16}	1,06	1,07	0,00	9,00	0,00	0,00	2,00
x_{17}	0,35	0,85	0,00	10,00	0,00	0,00	0,00
x_{18}	0,01	0,08	0,00	2,00	0,00	0,00	0,00
x_{19}	0,02	0,21	0,00	7,00	0,00	0,00	0,00
x_{20}	0,00	0,03	0,00	1,00	0,00	0,00	0,00
x_{21}	0,42	0,73	0,00	6,00	0,00	0,00	1,00
x_{22}	0,48	0,78	0,00	6,00	0,00	0,00	1,00
x_{23}	0,55	0,78	0,00	6,00	0,00	0,00	1,00
x_{24}	0,50	0,71	0,00	4,00	0,00	0,00	1,00
x_{25}	0,34	0,62	0,00	4,00	0,00	0,00	1,00
x_{26}	0,43	0,67	0,00	3,00	0,00	0,00	1,00
x_{27}	0,20	0,49	0,00	3,00	0,00	0,00	0,00
x_{28}	0,12	0,38	0,00	3,00	0,00	0,00	0,00
x_{29}	0,05	0,24	0,00	2,00	0,00	0,00	0,00

Nota-se na Tabela 5.6 que, em relação a situação domiciliar todo domicílio tem um chefe, um ou nenhum cônjuge e no máximo nove filhos.

Em resumo, foram utilizados os seguintes dados: dados de características socioeconômicas (idade, renda, situação domiciliar e composição do domicílio) e dados referentes às viagens (quantidade de viagens por domicílio). Alguns dados estavam disponíveis agregados (informações médias ou totais dos setores censitários), portanto, neste caso, não se conhecia a renda de um domicílio do setor x; mas o total ou a média de renda de todos os domicílios desse setor. Outros dados estavam disponíveis desagregados por domicílios ou indivíduos, por exemplo, buscar-se-á as viagens totais realizadas no domicílio y, ou até mesmo o número de viagens realizadas pelo indivíduo do domicílio y. Para esta tese, a desagregação dos dados foi por domicílios, devido ao objetivo proposto que é a previsão de viagens produzidas por domicílio.

5.2.2 Processo de geração da população sintética

Uma população sintética representa um conjunto de dados populacionais gerados a partir de informações de uma amostra ((ADIGA et al., 2015); (HAFEZI; HABIB, 2014); e (BARTHELEMY; TOINT, 2013)). Esta população pode ser gerada a partir de informações normalmente disponíveis pelo levantamento censitário feito pelo IBGE (dados agregados do Censo 2010) ou por pesquisas sociodemográficas ligadas ao governo estadual ou municipal, como, por exemplo, o SEADE-SP.

Nesta tese, o processo de geração da população sintética segue os passos de: 1) Definição da

amostra, 2) Método Monte Carlo e 3) Validação.

1) Definição da amostra

Para esta tese, a população sintética foi gerada a partir da amostra dos dados demográficos agregados da população disponibilizados pelo Censo Demográfico do IBGE-2010. Assim, a partir de uma amostra conhecida, é possível expandir os dados para um conjunto maior (população) preservando as características observadas na amostra.

Outro aspecto importante da técnica de gerar população sintética é a desagregação dos dados, isto é, a partir de dados agregados (por exemplo, distribuições de frequência de domicílios, renda, idade, e grau de parentesco) é possível gerar uma população fictícia, detalhando cada domicílio em função da composição familiar de seus moradores (número de moradores, renda, idade e posição familiar). Portanto, foram observados nos moradores do domicílio, por exemplo, quais as relações familiares entre eles, quantos moradores são chefe de família, quais as idades dos moradores e a renda total do domicílio.

É de fundamental importância destacar que para o processo de geração da população sintética desta tese, foi necessário estabelecer dois critérios para a execução do algoritmo proposto: (1) cada domicílio tem obrigatoriamente, pelo menos um chefe e (2) a quantidade de cônjuges, filhos e as outras categorias restringem-se ao número máximo de moradores no domicílio.

2) Método Monte Carlo

Existem vários métodos para geração de população sintética, tais como: técnica de ajuste proporcional iterativo (IPF), probabilidade condicional (Método Monte Carlo-MMC) e otimização combinatória (CO), cada um com seus graus de eficiência. Na elaboração deste conjunto de dados foi utilizado o algoritmo de Monte Carlo, que utiliza uma base estatística para gerar números pseudoaleatórios para resolver este tipo de problema.

Basicamente, o MMC estima a distribuição de uma estatística extraindo amostras aleatórias de uma população e observa o comportamento da estatística sobre as amostras. A geração da população sintética desta tese foi codificada em *Visual Basic for Application* (VBA), linguagem de programação na planilha Microsoft Excel do pacote *Office*. Neste caso, o método Monte Carlo é uma abordagem paramétrica porque a amostra é extraída de uma população com distribuição conhecida.

A construção da população sintética utilizando o MMC resume-se em três etapas, de acordo com a Figura 5.7. Inicia-se com a etapa de seleção do setor censitário, em seguida para cada setor é definida a quantidade de domicílios. Na próxima etapa é sorteado o número de moradores no domicílio e por fim, são definidas as características dos moradores do domicílio. O processo é repetido até que sejam preenchidos na totalidade todos os domicílios que compõem os setores censitários da área em estudo, cujo processo (algoritmo completo) está descrito no Apêndice D.

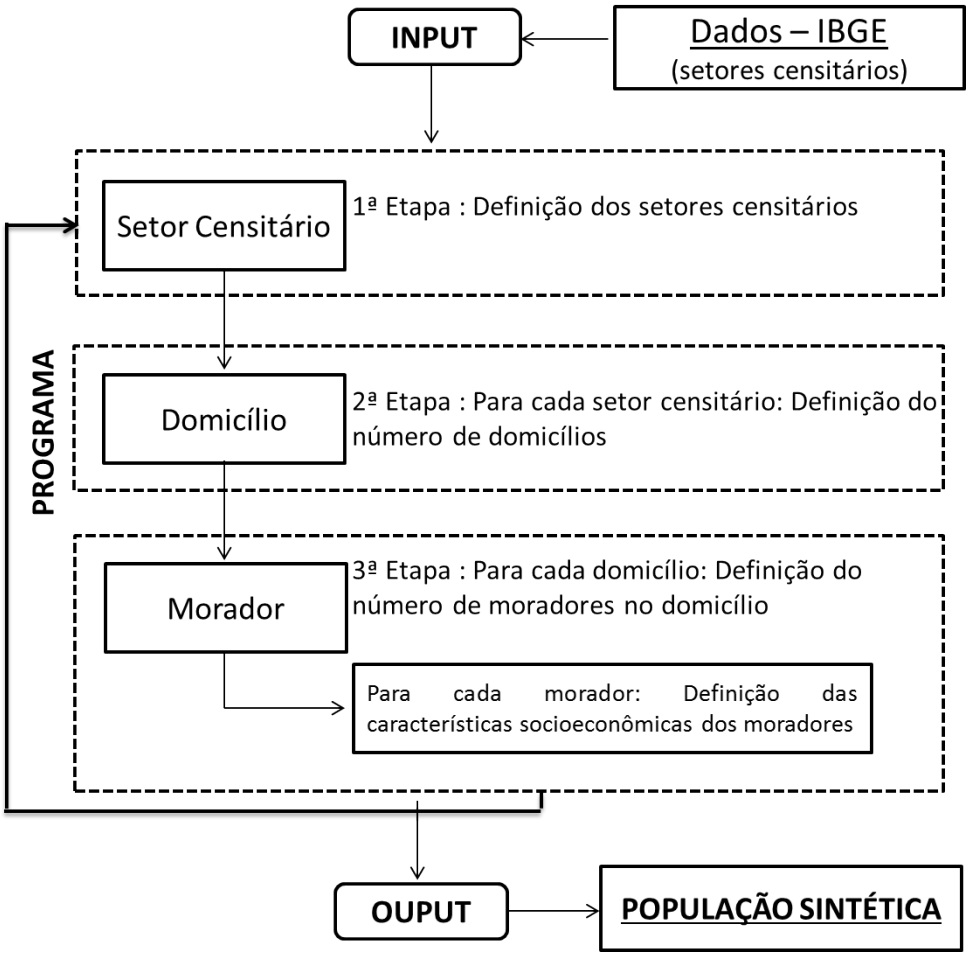


Figura 5.7 – Fluxograma da geração da população sintética.

O pseudocódigo do programa desenvolvido para a geração da população sintética deste trabalho está apresentado na Figura 5.8.

```
Geração de População Sintética
    {
       Para cada Setor censitário [i] faça:
         ndomic = quantidade de domícilios no Setor Censitário [i]
         Para cada domicílio [j] faça:
         nmorad = Gera quantidade de moradores(distribuição de domicílios)
         Para cada morador [k] faça:
         morador[k][j][i] = Gera parentesco(distribuição de parentesco)
                    morador[k][j][i] = Gera idade(distribuição de idade)
                         morador[k][j][i] = Gera sexo(distribuição de sexo)
                    morador[k][j][i] = Gera renda(distribuição de renda)
         Fim-para
       Fim-para
    };
```

Figura 5.8 – Pseudocódigo da geração da população sintética.

As 24 variáveis relativas às características dos moradores utilizadas foram: sexo (2), situação domiciliar (7), idade (9) e renda (6). E foram utilizadas também, variáveis relativas aos

domicílios (10). Foram definidas as categorias das 34 variáveis de acordo com as categorias apresentadas pelo censo de 2010 do IBGE, como mostra a Tabela 5.7.

Tabela 5.7 – Variáveis utilizadas na geração da população sintética.

Domicílios	**Moradores**			
	Sexo	**Sit. domiciliar**	**Idade**	**Renda**
Dom. com 1 morador	Mulheres	Chefes	até 10 anos	Sem renda
Dom. com 2 moradores	Homens	Cônjuges	11-20 anos	0-2 SM
Dom. com 3 moradores		Filhos	21-30 anos	2-3 SM
Dom. com 4 moradores		Outros Parentes	31-40 anos	3-5 SM
Dom. com 5 moradores		Agregados	41-50 anos	5-10 SM
Dom. com 6 moradores		Visitantes	51-60 anos	>10 SM
Dom. com 7 moradores		Empregados	61-70 anos	
Dom. com 8 moradores			71-80 anos	
Dom. com 9 moradores			>80 anos	
Dom. com 10 ou + moradores				

3) Validação

A validação ideal seria através da coleta dos dados reais, ou seja, informações de todos os indivíduos da população e relacioná-los a valores estimados pelo método proposto. Porém, a obtenção através de pesquisa desse conjunto de validação, o custo de realização muitas vezes é elevado, o que é inviável e praticamente impossível. Dessa forma, neste trabalho na etapa de validação da população sintética optou-se pelos testes de significância, também conhecidos como testes de hipóteses, pois são ferramentas que possibilitam a realização de inferência estatística a uma população tomando como base dados experimentais obtidos a partir de amostras desta população.

Os testes de significância são ferramentas importantíssimas para testar a veracidade de dados, ou seja, servem para verificar se os dados fornecem evidências suficientes para aceitar como verdadeira a hipótese da pesquisa (BARBETTA, 2012).

De acordo com Field (2009) os testes estatísticos mais usados para testar a normalidade são os testes de *Kolmogorov-Smirnov* e o de *Shapiro-Wilk*, pois comparam a pontuação da amostra a um conjunto normalmente distribuído com mesma média e desvio padrão, ou seja, ajudam analisar se a distribuição como um todo se desvia de uma distribuição normal.

Assim sendo, o teste estatístico utilizado foi o teste de *Kolmogorov-Smirnov* executado no IBM SPSS *Statistics 22*. Para o caso particular da população sintética proposta, foi utilizado o teste de amostras independentes, em que foram verificados se a distribuição de cada variável estimada era a mesma distribuição das variáveis da amostra de microdados do IBGE, que contém informações de uma parcela de domicílios da cidade de São Carlos-SP.

Como resultado foi apresentado se a variável em análise confirmou ou rejeitou a hipótese nula (a distribuição da variável é a mesma entre as variáveis estimadas e as variáveis da amos-

tra) e foram utilizadas somente as variáveis que passaram no teste estatístico (reteve a hipótese nula) nos modelos finais escolhidos para a previsão de viagens por domicílio.

Entretanto, foi utilizado também no método proposto, o banco completo (variáveis que não passaram no teste estatístico), exclusivamente para verificar se haviam grandes diferenças na modelagem da demanda por transportes quando comparados aos modelos de calibração que utilizaram somente as variáveis que passaram no teste.

5.2.3 Modelagem da demanda por transportes

Posterior à definição e tratamento das variáveis que influenciam a geração de viagens por domicílios, os dados passaram pelas seguintes etapas: (1) geração da população sintética (dados agregados) apresentada anteriormente e (2) modelagem da demanda por transportes (dados desagregados) apresentadas nesta seção, com o intuito de obter o melhor modelo de cada técnica e realizar análises comparativas dos resultados obtidos.

A calibração é uma etapa necessária e importante na modelagem, pois compara dados de desempenho real com os de simulação. O objetivo é corrigir as variáveis de entrada para melhorar a fidelidade do modelo. Neste trabalho, foram chamados de calibração todos os procedimentos de definição e correção dos parâmetros com vistas a obter uma melhor adequação do modelo à realidade.

Os processos de modelagem de produção de viagens por domicílio podem ser observados na Figura 5.9.

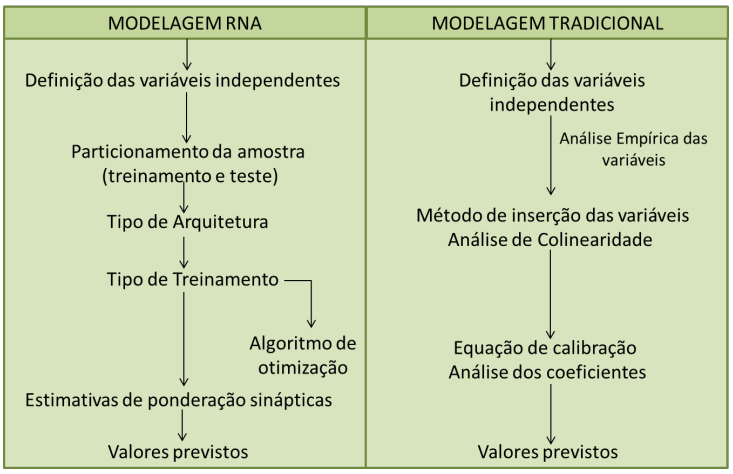


Figura 5.9 – Processo de calibração pela abordagem RNA e tradicional.

O método propõe a calibração de quatro modelos (1, 2, 3 e 4) de produção de viagens por domicílio através das técnicas de Redes Neurais Artificiais e de Regressão Linear Múltipla. Todos os modelos foram calibrados utilizando o *software* SPSS 22 e foram utilizadas as informações desagregadas dos domicílios e das viagens, dados geralmente obtidos por Pesquisas O/D.

Os modelos propostos 1 e 2 utilizaram todas as variáveis selecionadas na etapa de tratamento dos dados, estas foram chamadas de dados O/D e os modelos propostos 3 e 4 utilizaram somente as variáveis que passaram no teste de *Kolmogorov-Smirnov* realizado na etapa de geração da população sintética, chamados de dados O/D filtro, como mostra a Figura 5.10.

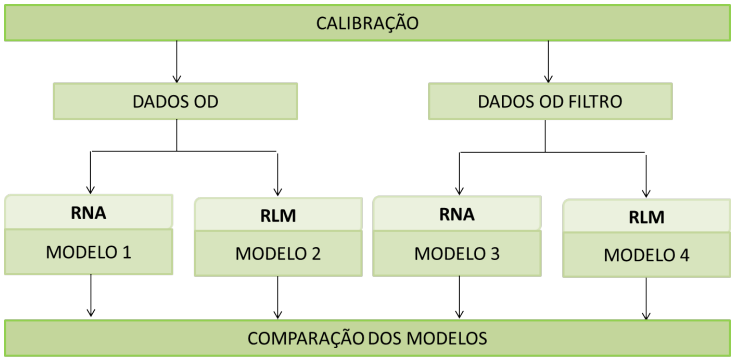


Figura 5.10 – Esquema dos dados utilizados para a calibração dos modelos.

Vale lembrar que os modelos 1 e 2 foram gerados apenas com a finalidade de analisar se apresentariam diferenças significativas nas respostas finais dos modelos quando comparados aos modelos 3 e 4.

Os processos de desenvolvimento dos modelos estão descritos a seguir: (1) Modelos 1 e 3 de Redes Neurais Artificiais (RNAs) e (2) Modelos 2 e 4 de Regressão Linear Múltipla (RLM).

(1) Modelos 1 e 3 de Redes Neurais Artificiais (RNAs):

Esta etapa envolve a calibração dos modelos 1 e 3 através da técnica de RNAs. A modelagem de RNAs segue os passos de: a) Definição das variáveis e particionamento da amostra, b) Tipo de arquitetura e c) Tipo de treinamento.

A escolha da melhor rede neural foi feita através de uma análise de sensibilidade, pois na literatura não existem regras claras a respeito da melhor proporção de amostra de treinamento e teste, nem da mais adequada arquitetura da rede e nem do tipo de treinamento mais eficiente. Assim, a avaliação da sensibilidade foi realizada através das possíveis simulações disponíveis no SPSS.

a) Definição das variáveis e particionamento da amostra

A modelagem de RNAs pode ser realizada por várias opções de redes, porém as redes mais usuais e também, disponíveis no *software* SPSS 22.0 são as redes perceptrons multicamadas (MLP) e as redes de base radial.

A rede MLP é função de variáveis de previsão (variáveis independentes) que minimizam o erro de predição da variável de saída. É composta por uma camada de entrada (variáveis independentes), em que as informações são recebidas; por nenhuma, uma ou mais camadas ocultas e uma camada de saída. A camada de saída fornece a solução do problema.

Neste estudo foi utilizada a rede MLP. Os modelos gerados por RNAs tiveram início com a escolha do tipo de rede a ser treinada, ou seja, a rede perceptrons multicamadas. Em seguida foram definidas as variáveis preditoras (independentes) em variáveis categóricas (fatores) ou numéricas (covariáveis). Para o caso das variáveis numéricas foi estabelecido o tipo de redimensionamento das covariáveis. As medidas disponíveis para o redimensionamento das covariáveis estão contidas na Tabela 5.8.

Tabela 5.8 – Definição das variáveis para a MLP.

VARIÁVEIS		
Dependente	**Independentes**	
numéricas e/ou categóricas	Fatores (categóricas)	Covariáveis (numérica)
Redimensionamento das covariáveis		
Padronizado	*Normalizado*	*Normalizado ajustado*
(x*- média)/desvio padrão)	(x*-mín)/(máx-mín) Valores entre 0 e 1	$[2 \times (x * -min)/(max - min)] - 1$ Valores entre -1 e 1
x*- valor da variável		

Depois de definidas as variáveis da MLP foi realizado o particionamento da amostra. Assim, a amostra foi separada em duas partes: a primeira parte reservada para a calibração dos dados, chamada de amostra de treinamento (dados que treinaram a rede) e a segunda para a validação, chamada de amostra teste usada para avaliar a capacidade da rede neural.

A escolha das proporções da amostra de treinamento e teste foi empírica, pois não há uma proporção padrão para estipular os tamanhos dos conjuntos de calibração e validação. Valores comuns para a razão são: 60/40, 70/30 e 80/20. Contudo, para este trabalho, além das razões comuns foram testadas mais duas razões 50/50 e 90/10. É importante salientar que a divisão deve ser feita de maneira randômica e geralmente é mais eficiente quando a amostra de calibração (treinamento) é maior que a amostra de validação (teste).

b) Tipo de arquitetura

Existem três pontos importantes na concepção da arquitetura de uma RNA, sendo eles (NAGAI, 2006):

- i. Número de camadas escondidas;
- ii. Número de neurônios nas camadas escondidas;
- iii. Função de ativação.

Porém, não existem regras determinadas para a especificação desses pontos e geralmente é utilizado o método de tentativa e erro. O tipo de arquitetura constitui a estrutura da rede neural e os dois primeiros pontos (i e ii) determinam a complexidade do modelo neural.

O especialista pode montar a sua própria arquitetura ou utilizar a arquitetura automática disponíveis em alguns *softwares*. A arquitetura automática calcula o melhor número de unidades na camada oculta (unidade é um nó não observável da rede) e utiliza a função de ativação padrão para as camadas ocultas e de saída. A opção de arquitetura automática do *software* SPSS 22.0 utiliza a rede MLP de uma única camada oculta e permite escolher o número mínimo e máximo de unidades admitidas. As funções de ativação padrão da camada oculta e da camada de saída são a função tangente hiperbólica e a identidade (função que leva valores reais e retorna-os inalterados), respectivamente.

A arquitetura personalizada da rede se dá por meio da definição das camadas ocultas, sendo que cada unidade oculta é a soma ponderada das entradas da rede neural e da camada de saída que contém as variáveis previstas.

Assim, a arquitetura personalizada iniciou-se pela definição da quantidade de camadas ocultas e do número de unidades e pela escolha da função de ativação. Se existir mais de uma camada oculta, por exemplo, na segunda camada cada unidade oculta será a soma ponderada das unidades da primeira camada oculta e não mais das entradas como acontecia na primeira camada oculta, mas a mesma função de ativação é utilizada em ambas as camadas ocultas.

Em seguida, foi determinada a camada de saída. Nessa camada foi escolhida a função de ativação e no caso, da variável dependente ser numérica, o redimensionamento está disponível para modificações se necessárias. A Tabela 5.9 apresenta de forma resumida as topologias mais usuais da rede perceptrons multicamadas (MLP).

Tabela 5.9 – Opções da arquitetura da MLP.

ARQUITETURA MLP		
Camada oculta		
Número de camadas ocultas	1 2	
Número de unidades	automática personalizada	
Função de ativação	tangente hiperbólica	curva sigmoide
Camada de saída	Função de ativação	Redimensionamento
	identidade	padronizado normalizado normalizado ajustado
	tangente hiperbólica	normalizado ajustado
	curva sigmoide	normalizado

Neste trabalho, a MLP foi construída no *software* SPSS 22.0 e por isso, foram utilizadas as opções de arquitetura automática e personalizada. Foram testadas e comparadas as possíveis arquiteturas e em seguida, foi escolhida a arquitetura na qual a rede apresentou melhor desempenho.

c) Tipo de treinamento

Depois de selecionadas as variáveis independentes e dependente, a amostra de treinamento e teste e a arquitetura da rede neural, nesta etapa foi definido o tipo de treinamento da MLP e, consequentemente o algoritmo de otimização da rede neural.

O algoritmo de aprendizado utilizado para o treinamento da RNA foi o *backpropagation*, uma vez ele é o algoritmo padrão do software SPSS 22.0, além de ser normalmente o mais utilizado nas redes MLP. Esse algoritmo é do tipo supervisionado e utiliza pares de valores (entradas, saídas desejadas) que, através de correções de erros, ajusta os pesos da rede.

O tipo de treinamento determina como a rede processa os seus registros e os mais usuais são: o lote (*batch*), o *online* e o mini lote (*mini-batch*) e cada tipo está detalhado a seguir.

- Lote: atualiza os pesos sinápticos após passar todos os registros de dados de treinamento, ou seja, utiliza as informações de todos os registros no conjunto de dados de treinamento. Continua o processo de atualização dos pesos até uma das regras de parada ser atendida. Este tipo de treinamento minimiza diretamente o erro total.

- Online: atualiza os pesos sinápticos após cada registro de dados de treinamento, ou seja, usa a informação de um registro de cada vez. Este treinamento recebe continuamente um registro e só finaliza quando uma das regras de parada seja atendida.

- Mini lote: atualiza os pesos sinápticos depois de passar o grupo, pois divide o conjunto de dados de treinamento em grupos de tamanhos iguais.

Juntamente com a escolha do tipo de treinamento foram estimados os pesos sinápticos e para isso foram utilizados os algoritmos de otimização. Todos os tipos de treinamentos permitem a estimação dos pesos sinápticos pelo algoritmo gradiente descendente, porém apenas o treinamento do tipo lote permite o uso do algoritmo gradiente conjugado em escala, como mostra a Tabela 5.10

Tabela 5.10 – Treinamento e algoritmo de otimização.

TIPO DE TREINAMENTO	ALGORITMO DE OTIMIZAÇÃO
lote	gradiente conjugado em escala e gradiente descendente
online	gradiente descendente
minilote	gradiente descendente

Embora, este trabalho apresente um conjunto de dados grande, foram testados os três tipos de treinamento (lote, online e mini lote) disponíveis no *software* SPSS 22.0. Em consequência disso, também foram utilizados todos os algoritmos de otimização.

Dessa forma, a arquitetura, a função de ativação das camadas ocultas e de saída e o tipo de treinamento da MLP são as características que determinam o desempenho da rede e por isso, foram definidas cuidadosamente.

As opções de combinações que foram testadas estão apresentadas resumidamente na Tabela 5.11.

Tabela 5.11 – Opções de combinações disponíveis pelo SPSS 22.0.

PARTIÇÕES	**ARQUITETURA**	**TREINAMENTO**
Treinamento/Teste	**Automática**	
90% /10%		lote/gradiente conjugado
80% /20%		lote/gradiente descendente
70% /30%		online/gradiente descendente
60% /40%		minilote/gradiente descendente
50% /50%		
Treinamento/Teste	**Personalizada**	
90% /10%	**CAMADAS OCULTAS**	
80% /20%	**nº de camadas**	**Função de ativação**
70% /30%	1	tangente hiperbólica
60% /40%	2	curva sigmoide
50% /50%	**nº de unidades**	
	automática	
	CAMADA DE SAÍDA	**TREINAMENTO**
	Função de ativação	lote/gradiente conjugado
	identidade	lote/gradiente descendente
	tangente hiperbólica	online/gradiente descendente
	curva sigmoide	minilote/gradiente descendente

Ao todo foram testadas 270 combinações da MLP e foi analisado o melhor resultado alcançado, ou seja, o resultado que apresentou o menor erro relativo (combinação ótima) e então, escolhida a rede MLP mais apropriada (o particionamento da amostra, o número de camadas ocultas, a função de ativação das camadas ocultas e de saída, o número de unidades e o tipo de treinamento). O erro relativo (ER) foi calculado pelo *software* SPSS 22.0 através da Equação 5.1.

$$ER = \frac{(\sum(\text{viagens observadas-viagens previstas})^2)}{(\sum(\text{viagens observadas-média das viagens observadas})^2)} \tag{5.1}$$

Em vista disso, após a rede mais adequada ter sido escolhida, o modelo 1 foi calibrado com os dados da Pesquisa O/D e assim, foram estimadas as viagens por domicílio. A combinação ótima da MLP escolhida para o treinamento do modelo 1, foi repetida para o modelo 3.

A diferença do treinamento do modelo 1 para o modelo 3, foi em relação ao número de variáveis de entrada. No modelo 1 empregou-se todas as variáveis (dados O/D) e no modelo

3 somente as variáveis que passaram no teste estatístico de *Kolmogorov-Smirnov* (dados O/D filtro).

(1) Modelos 2 e 4 por Regressão Linear Múltipla (RLM):

O objetivo da RLM é encontrar a combinação linear das variáveis independentes que forneça máxima correlação com a variável dependente. Apesar de ser, obviamente, irreal, ela trata o número de viagens estimadas como variável contínua com suposição de distribuição normal (podendo assumir inclusive valores negativos). Mesmo assim, ainda é a técnica mais indicada e utilizada pelos pesquisadores nas análises de problemas de transportes.

O processo de modelagem tradicional segue os passos de: a) Definição e análise empírica das variáveis, b) Método de inserção das variáveis e análise de colinearidade e c) Equação de calibração e análise dos coeficientes.

a) Definição e análise empírica das variáveis

A modelagem tradicional teve início com a definição das variáveis independentes e dependente. Em seguida, foi realizada uma análise empírica das variáveis e eleitas as variáveis que realmente explicavam o fenômeno da produção de viagens por domicílio, pois as variáveis explicativas devem ser selecionadas não somente pela correlação forte com a variável dependente, mas também por explicar adequadamente o fenômeno.

b) Método de inserção das variáveis e análise de colinearidade

Depois de escolhidas as variáveis que realmente explicavam o fenômeno de geração de viagens domiciliares foi escolhido o método de inserção e estimação. O *software* SPSS 22.0 disponibiliza quatro métodos: *Enter*, *Stepwise*, *Forward Selection* e *Remove*. O modelo escolhido para esta proposta foi o modelo de Regressão Linear Múltipla (RLM) *Stepwise*, pois o procedimento constrói, interativamente, uma sequência de modelos de regressão pela adição ou remoção de variáveis em cada etapa e permite ao pesquisador examinar a contribuição de cada variável independente para o modelo de regressão.

Uma questão importante na interpretação da variável estatística de regressão é a correlação entre as variáveis independentes. O caso ideal para um pesquisador seria ter diversas variáveis independentes altamente correlacionadas com a dependente, mas com pouca correlação entre elas (HAIR et al., 2009). Uma maneira simples de identificar a colinearidade (existência de relação linear entre duas variáveis explicativas) é através da matriz de correlação, utilizando os valores de coeficientes de *Pearson*. Dessa forma, a presença de elevadas correlações entre as variáveis independentes, ou seja, geralmente 0,90 ou maiores é a primeira indicação de multicolinearidade (efeito combinado de duas ou mais variáveis independentes). Porém, a falta de valores elevados de correlação não garante a ausência de colinearidade, por isso, também é

necessário saber o grau que cada variável independente é explicada pelo conjunto de outras variáveis independentes.

Assim, foi estimada a matriz de correlação das variáveis independentes e em seguida, analisada a colinearidade aos pares ou múltiplas através dos fatores de tolerância e de inflação de variância.

A tolerância é uma medida direta de multicolinearidade, definida como a quantia de variabilidade da variável independente selecionada não explicada pelas outras variáveis independentes. Portanto, se o valor de tolerância for alto (próximo de 1,00) significa um pequeno grau de multicolinearidade. O fator de inflação de variância (VIF) é a outra medida de multicolinearidade, que é calculado como o inverso do valor da tolerância. Este não deve ser maior que 2 a fim de evitar problemas de multicolinearidade. Assim, tanto os valores de tolerância quanto de VIF próximos de 1,00 indicam pequeno grau de multicolinearidade.

Por exemplo, valores menores de tolerância e valores maiores para VIF indicam alto nível de multicolinearidade nas variáveis. A raiz quadrada do VIF é o grau em que o erro padrão aumenta devido a multicolinearidade, portanto, se o erro padrão aumenta os intervalos de confiança em torno dos coeficientes estimados tornam-se maiores e fica mais difícil o esclarecimento de que o coeficiente é significativamente diferente de zero (HAIR et al., 2009).

c) Equação de calibração e análise dos coeficientes

A equação de calibração foi obtida pela regressão *Stepwise* considerando como variável dependente a variável *viagens por domicílio* e como variáveis independentes as variáveis domiciliares apresentadas na Tabela 5.3 (página 70).

Depois de obtida a equação de calibração foi realizada uma análise estatística de significância dos parâmetros estimados e do modelo total (estatística t e teste F) e uma análise crítica a respeito das variáveis explicativas selecionadas pelo procedimento *Stepwise*, considerando coerência dos sinais e magnitude dos coeficientes estimados, bem como das variáveis selecionadas.

Finalmente, com todas essas etapas analisadas foram calibrados os modelos 2 e 4 pelo método *Stepwise* e assim, foram obtidas as viagens por domicílio. Diferente da modelagem RNA, em que o modelo 3 apenas repetiu os passos do modelo 1, na modelagem RLM, quando foi modificado o conjunto de dados de entrada do modelo 2 (dados O/D) para (dados O/D filtro) no modelo 4 foi necessário analisar novamente todos os passos da modelagem tradicional realizados para o modelo 2 antes de gerar a nova equação de calibração para o modelo 4.

5.2.4 Validação dos modelos de calibração

O procedimento de validação é fundamental, pois objetiva certificar se a transformação entrada e saída (*input-output*) realizada pelo modelo tem precisão para representar a mesma ocorrência

procedida no sistema real. Primeiramente é necessário rodar o modelo considerando mesmas condições impostas ao sistema e depois comparar os dados gerados pelo modelo e o sistema (BALCI, 2003).

Menner (1995) afirma que a validação é uma maneira para analisar se o comportamento do modelo proposto representa de forma válida o sistema do mundo real que está sendo simulado e pode ser realizada de forma subjetiva ou estatística.

De acordo com Botter (2001) a validação é essencial, pois consiste na confirmação de que o modelo opera da forma que o analista pretendia e que a saída do modelo é confiável e representativa de um sistema real.

O processo de validação dos modelos de demanda por transportes obtidos teve sua eficácia avaliada através do uso de medidas estatísticas de comparação entre os valores previstos (estimativas de viagens por domicílio dos quatro modelos propostos) e os valores observados (viagens por domicílio da Pesquisa O/D). As medidas de desempenho adotadas foram: o Erro médio (Equação 5.2), a Raiz do erro médio (Equação 5.3), o Erro relativo (Equação 5.4) e o Coeficiente de Correlação (Equação 5.5) apresentadas em seguida, respectivamente.

$$EM = \frac{1}{N} \times \sum_{i=1}^{N} (x_i - y_i) \tag{5.2}$$

$$RQEM = \sqrt[2]{\frac{1}{N} \times \sum_{i=1}^{N} (x_i - y_i)} \tag{5.3}$$

$$ER = \frac{\sum_{i=1}^{N} (x_i - y_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{5.4}$$

Sendo: EM=erro médio; x_i=valor observado; y_i=valor previsto; RQEM=raiz do erro médio; ER=erro relativo e $\bar{x}$=valor médio observado.

$$r = \frac{1}{N = r} \sum_{i=1}^{N} \frac{(x_i - \bar{x}) \times (y_i - \bar{y})}{\sigma_x \times \sigma_y} \tag{5.5}$$

Em que: r=coeficiente de correlação; $\bar{x}$=valor médio observado; $\bar{y}$=valor médio estimado; σ_x=desvio padrão dos valores observados e σ_y= desvio padrão dos valores estimados.

5.2.5 Comparação dos modelos de calibração

Finalmente, os resultados das medidas estatísticas da modelagem RNA foram comparados aos da modelagem tradicional RLM. Foram utilizados os valores referentes a 40% da amostra (amostra de teste). Como medidas de comparação foram utilizadas os coeficientes de determinação (R^2) para avaliar o poder de previsão dos modelos 2 e 4, a análise de resíduos para verificar a

qualidade dos modelos e a análise das medidas de desempenho (raiz do erro médio- RSME e correlação-CORREL) para analisar o modelo que apresentou menor erro de previsão.

Deve-se deixar claro que o objetivo não foi checar minuciosamente a qualidade dos modelos lineares, e sim, testar a adequabilidade das RNAs para estimar demanda por viagens (*viagens por domicílio*), que é o objetivo secundário desta pesquisa. Além disso, de acordo com o objetivo principal desta pesquisa, que é gerar uma população sintética para prever viagens por domicílios, a comparação dos resultados valeu-se então, da comparação dos resultados dos modelos (3 e 4) pois, foram os modelos de previsão de viagens por domicílio, os quais utilizaram como dados de entrada somente as variáveis que passaram no teste estatístico de *Kolmogorov-Smirnov*.

Vale ressaltar que os modelos (1 e 2) foram realizados apenas para verificar se existiria grandes diferenças quando foi reduzido o número de variáveis de entrada na calibração dos modelos.

5.2.6 Validação dos resultados

As viagens produzidas por domicílio foram estimadas pelos modelos de calibração 3 e 4 e utilizadas como dados de entrada para os modelos as variáveis independentes da população sintética obtida através de dados agregados.

Todas as medições estão sujeitas a erros e, quanto maior a gravidade dos erros menor a qualidade dos resultados e assim, a análise estatística produzirá resultados diferentes dos desejados. Os erros podem ocorrer devido a diversas circunstâncias, tais como: métodos e técnicas de coleta de dados, como entrevista, observação ou questionário (erros de respostas por desonestidade, confusão e ignorância ou erros de coleta dos dados por falha nos equipamentos e confusão na leitura da resposta), processamento inadequado (utilização de técnica de análise e processamento de dados de pouca confiança ou não apropriados para o problema estudado), armazenamento com pouca confiabilidade (falhas na entrada dos dados para o processamento, ou perda de informações) e outros problemas que podem ocorrer após a coleta de dados.

Quanto à validação das viagens previstas pelos modelos propostos, o ideal seria ter uma amostra real do número de viagens por domicílio da população total, porém só temos uma amostra de 3.057 domicílios coletados pela Pesquisa O/D. Assim, não é possível a validação das viagens por domicílio previstas pelos modelos 3 e 4 utilizando, por exemplo, a análise de resíduos (viagens observadas x viagens previstas). Portanto, a validação dos resultados do método proposto se baseia nas etapas ilustradas na Figura 5.11 e descritas na sequência.

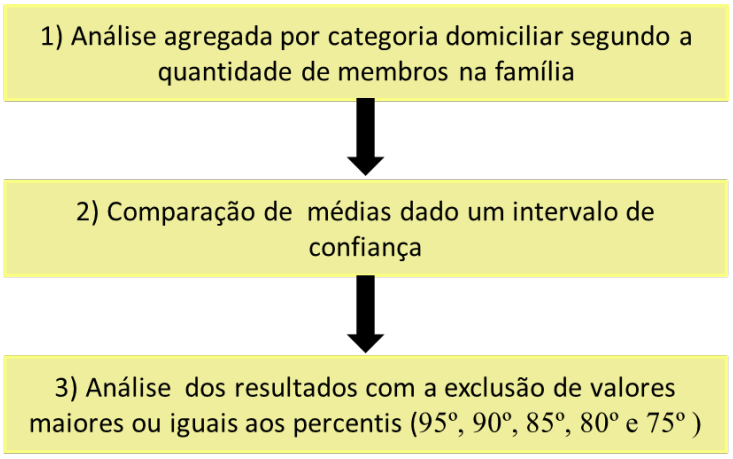


Figura 5.11 – Etapas de validação dos resultados obtidos pelo método proposto.

1) Análise agregada por categoria domiciliar segundo a quantidade de membros na família:

A Figura 5.12 apresenta de forma simplificada a etapa 1.

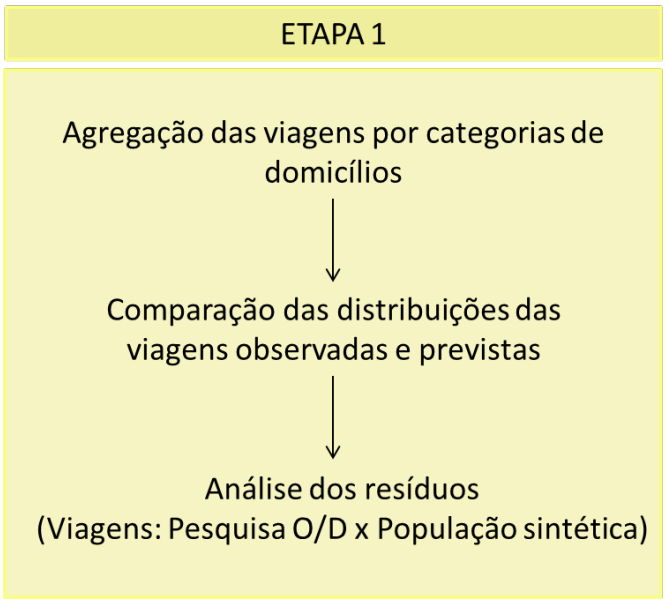


Figura 5.12 – Etapa 1 de validação dos resultados obtidos pelo método proposto.

Desta forma, a validação dos resultados iniciou-se com a etapa 1, que foi realizada uma análise agregada por categoria domiciliar segundo a quantidade de membros na família (total de viagens nos domicílios com 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10 ou mais moradores). Foram utilizadas as porcentagens das viagens por domicílio obtidas por categorias de domicílios e assim, foram comparadas as distribuições das viagens por domicílio observadas (Pesquisa O/D) e as viagens por domicílio da população sintética previstas pelos modelos 3 e 4.

Além disso, foram utilizadas medidas estatísticas de comparação entre os valores observados e os valores previstos. As medidas de desempenho adotadas foram: o erro quadrático médio e o coeficiente de correlação de *Pearson*.

2) Comparação de médias dado um intervalo de confiança:

Outra forma de validação das viagens por domicílio da população sintética previstas foi pela comparação de médias dado um intervalo de confiança.

A etapa 2 segue os passos conforme apresenta a Figura 5.13.

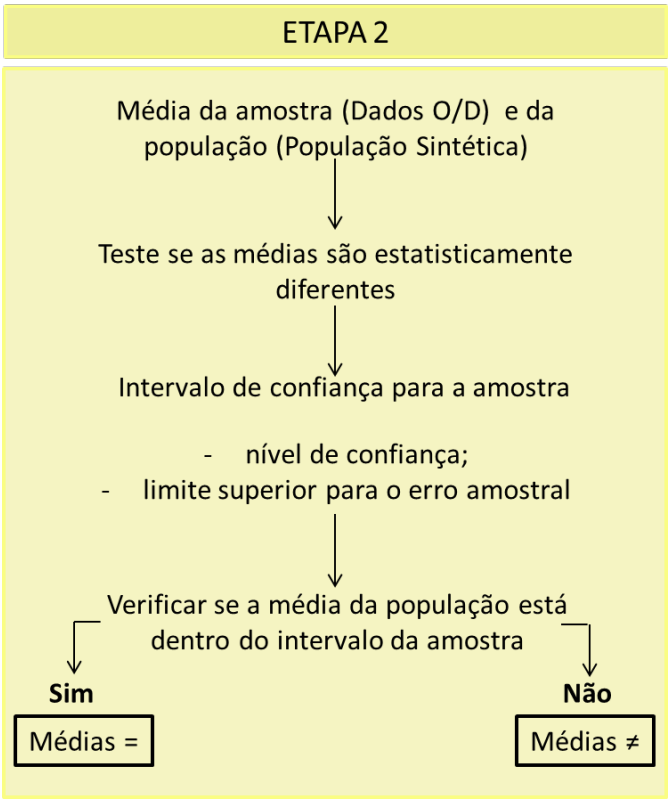


Figura 5.13 – Etapa 2 de validação dos resultados obtidos pelo método proposto.

Primeiramente, foram calculadas as médias da amostra (viagens por domicílio da Pesquisa O/D) e da população (viagens por domicílio da população sintética) por setores censitários. Em seguida, foi testado se essas duas médias eram estatisticamente diferentes. Para descobrir se as médias amostrais calculadas eram estatisticamente diferentes, foi criado um intervalo de confiança para a amostra.

Assim, foi fixado o nível de confiança em 95%, pois é o mais usual na prática e admitindo que a distribuição amostral de $\bar{X}$ é aproximadamente normal, o limite máximo para o erro amostral foi estimado por $1{,}96S_{\bar{X}}$. Sendo o $S_{\bar{X}}$ a divisão do desvio padrão da amostra (S) pela raiz quadrada do tamanho da amostra (n). Portanto, a Equação 5.6 apresenta o intervalo de confiança da amostra.

$$[\bar{X} - 1,96 \times \left(\frac{S}{\sqrt{n}} \right); \bar{X} + 1,96 \times \left(\frac{S}{\sqrt{n}} \right)] \tag{5.6}$$

Depois de obtido o intervalo de confiança para a média da amostra foi verificado se a média calculada para a população estava dentro desse intervalo ou não. Se estiver, significa que

as duas médias são iguais e, então, as viagens por domicílio daquele determinado setor são consideradas válidas. Se não estiver, significa que as médias são diferentes e as viagens por domicílio do setor em questão são descartadas (não válidas).

3) Análise dos resultados com a exclusão de valores maiores ou iguais aos percentis (95, 90, 85, 80 e 75):

A etapa 3 é a última forma de validação proposta e segue os passos da Figura 5.14.

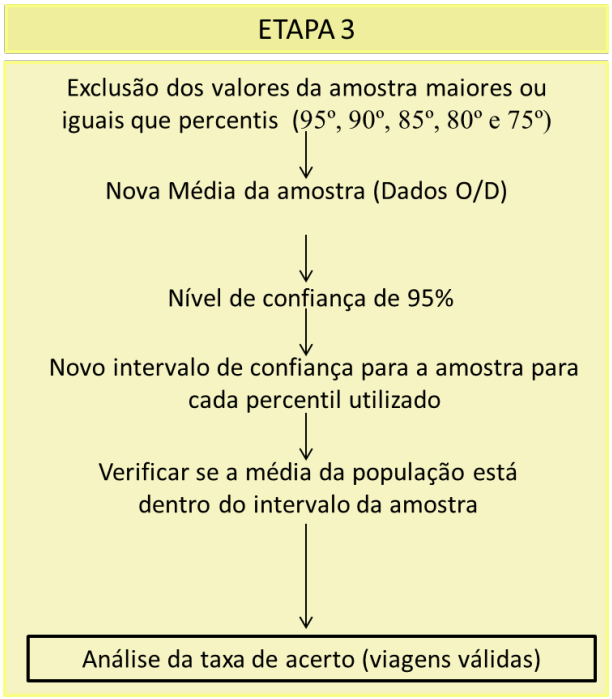


Figura 5.14 – Etapa 3 de validação dos resultados obtidos pelo método proposto.

O objetivo da etapa 3 foi verificar se com a retirada na amostra dos valores maiores ou iguais aos percentis estipulados aumentaria o número de viagens por domicílio da população sintética consideradas válidas.

De tal modo, primeiramente foi realizada na amostra (Pesquisa O/D) a exclusão das viagens por domicílio maiores ou iguais ao percentil estipulado. Em seguida, foi recalculada a média das viagens por domicílio por setores censitários da amostra e, consequentemente foi obtido o novo intervalo de confiança de 95%. Esses passos foram repetidos para cada um dos percentis estipulados para a análise.

Após, foram analisadas se as taxas de acertos aumentaram ou diminuíram com a exclusão das viagens por domicílio maiores ou iguais aos valores referentes a cada percentil.

Este capítulo mostrou o processo para a realização do método proposto neste trabalho, cuja finalidade principal é a obtenção de uma população sintética para modelar número de viagens produzidas por domicílio.

Vale ressaltar a facilidade da utilização do método proposto em outras áreas de estudo semelhantes à cidade de São Carlos-SP, principalmente àquelas que não possuem Pesquisa Origem e Destino. Porém, para utilizar o método proposto em cidades com o porte diferente de São Carlos-SP é necessário ter em mãos pelo menos uma Pesquisa O/D.

Em relação a proposta de testar a adequabilidade das Redes Neurais Artificiais para estimar demanda por viagens, objetivo secundário do presente trabalho foi atingido pela comparação entre os resultados (viagens por domicílio da população sintética) obtidos pelo modelo de RNAs e os resultados do modelo tradicional de modelagem da demanda por transportes, a Regressão Linear Múltipla.

Capítulo 6

RESULTADOS E DISCUSSÕES

Este capítulo descreve a análise e discussão dos resultados obtidos através do método descrito no Capítulo 5. São apresentados resultados principais e validação da população sintética e, também os resultados provenientes da modelagem RNA e modelagem tradicional, bem como comparação das duas técnicas considerando a amostra de validação.

6.1 Tratamento e visualização dos dados

Esta seção envolve a descrição das etapas de tratamento dos dados empregadas neste trabalho. Desta forma, o objetivo do tratamento dos dados foi separar as amostras finais eliminando dados incompletos ou aqueles que não faziam parte dos objetivos da análise. Partiu-se então, da triagem das informações dos dados agregados do Censo 2010 disponibilizado pelo IBGE. Após, foram analisados os dados desagregados, compostos por duas bases: a base da Pesquisa O/D e a base dos microdados do Censo 2010 - IBGE, respectivamente.

6.1.1 Dados agregados

Os dados agregados por setores censitários utilizados neste trabalho foram obtidos após a eliminação dos setores que não estavam contidos na área urbanizada da cidade (32 setores) e os que não possuíam domicílios (2 setores). Além disso, foram selecionados somente os domicílios particulares permanentes contidos nesses setores. Desta forma, foi obtido da amostra original o total de 288 setores censitários na área urbanizada com 68.833 domicílios particulares permanentes. As informações da amostra original e modificada estão apresentadas na Tabela 6.1.

Tabela 6.1 – Amostra dos setores censitários.

Amostra	Setores Censitários	Domicílios	Moradores	Variáveis
Original	322	71.601	221.051	254
Modificada	288	68.833	212.263	34

Assim, das 254 variáveis disponíveis na amostra original, foram selecionadas somente as variáveis de interesse deste trabalho, que foram 34 variáveis socioeconômicas, tais como: sexo, situação domiciliar, características dos domicílios, idade e renda. As características das variáveis selecionadas na base agregada do Censo 2010, bem como as medidas descritivas dessas variáveis foram apresentadas nas Tabelas 5.1 e 5.2, respectivamente (páginas 67 e 68). Vale ressaltar, que essas variáveis foram fundamentais para a geração da população sintética desta tese.

6.1.2 Dados desagregados

(1) Base de dados dos domicílios-Pesquisa O/D

Os dados dos domicílios da Pesquisa O/D selecionados para este trabalho foram obtidos da amostra original de 3.057 domicílios com 10.085 moradores, que realizaram 14.702 viagens. Foram escolhidas as mesmas 34 variáveis socioeconômicas selecionadas na base agregada, tais como: sexo, situação domiciliar, características dos domicílios, idade e renda. Além dessas variáveis, a base de dados da Pesquisa O/D continha a variável *viagens por domicílio*, ou seja, a variável de previsão a ser encontrada para esta tese. As informações da amostra original e modificada estão apresentadas na Tabela 6.2.

Tabela 6.2 – Amostra da Pesquisa O/D.

Amostra	Domicílios	Moradores	Viagens	Variáveis
Original	3.057	10.085	14.702	70
Modificada	3.057	10.085	14.702	35

As características das 35 variáveis selecionadas, bem como as medidas descritivas dessas variáveis foram apresentadas nas Tabelas 5.3 e 5.4, respectivamente (páginas 70 e 71). Logo, essas variáveis foram fundamentais para a calibração dos modelos de demanda por transportes propostos nesta tese, os quais estão apresentados a seguir, neste capítulo.

(2) Base de dados dos domicílios – IBGE

Diferente da base de dados agregada do Censo de 2010, em que as variáveis disponíveis estavam prontas para o uso, a base de microdados apresentava as variáveis na forma de códigos.

Por isso, para a obtenção desse banco de dados foi necessário utilizar a sequência de etapas apresentadas na Figura 5.6 do Capítulo 5: Materiais e Método.

Os dados dos domicílios selecionados da base de microdados 2010 foram obtidos da amostra original de 7.124 domicílios e foram escolhidas as mesmas 34 variáveis socioeconômicas selecionadas na base agregada. As informações da amostra original e modificada estão apresentadas na Tabela 6.3. A base de microdados do IBGE foi de grande importância para o trabalho, pois foi utilizada para a validação da população sintética (domicílios) obtida a partir da amostra agregada do IBGE.

Tabela 6.3 – Amostra dos microdados 2010.

Amostra	Domicílios	Moradores	Variáveis
Original	7.124	22.236	191
Modificada	6.817	21.128	34

6.2 População sintética

6.2.1 Domicílios sintéticos

A população sintética obtida neste trabalho e seus resultados são apresentados a seguir. A distribuição dos dados da população sintética (dados estimados) foi comparada com a distribuição dos dados agregados do Censo de 2010 (dados observados) por meio da representação das frequências relativas. Para ilustrar foram utilizados histogramas de algumas das variáveis, tais como: (a) número de moradores no domicílio, (b) situação domiciliar e (c) faixa etária e estão representadas na Figura 6.1.

A população sintética obtida é composta por 68.833 domicílios sintéticos com 212.263 moradores, ou seja, a mesma quantidade de domicílios e moradores contida na amostra agregada (dados observados). O mesmo aconteceu para a variável sexo (108.460 mulheres e 103.803 homens) e para a variável renda domiciliar, que ambas apresentaram os mesmos valores dos dados observados. As informações da população sintética estão descritas no Apêndice E.4.

Nota-se que os histogramas apresentaram os dados observados e estimados exatamente iguais para as variáveis *número de moradores por domicílio* e *faixa etária das idades*. Isso significa que essas variáveis estimadas pelo método de geração da população sintética (MMC) proposto nesta tese foram precisas, pois apresentaram correlação igual a 1 e erro médio igual a zero.

Ocorreu de maneira diferente para a distribuição dos moradores conforme a situação domiciliar, pois os dados observados não foram exatamente iguais aos estimados. Isso ocorreu, pois foram estabelecidos critérios para a execução do algoritmo Monte Carlo, no qual o pri-

meiro morador é sempre chefe, o segundo morador é preferencialmente, cônjuge e as demais categorias são sorteadas restringindo ao número máximo de moradores no domicílio.

Mesmo com essa diferença, observou-se uma correlação alta entre os dados observados e estimados (0,972), com erro médio zero e erro relativo baixo (0,034), o que também representa uma estimativa precisa dessa variável na população sintética.

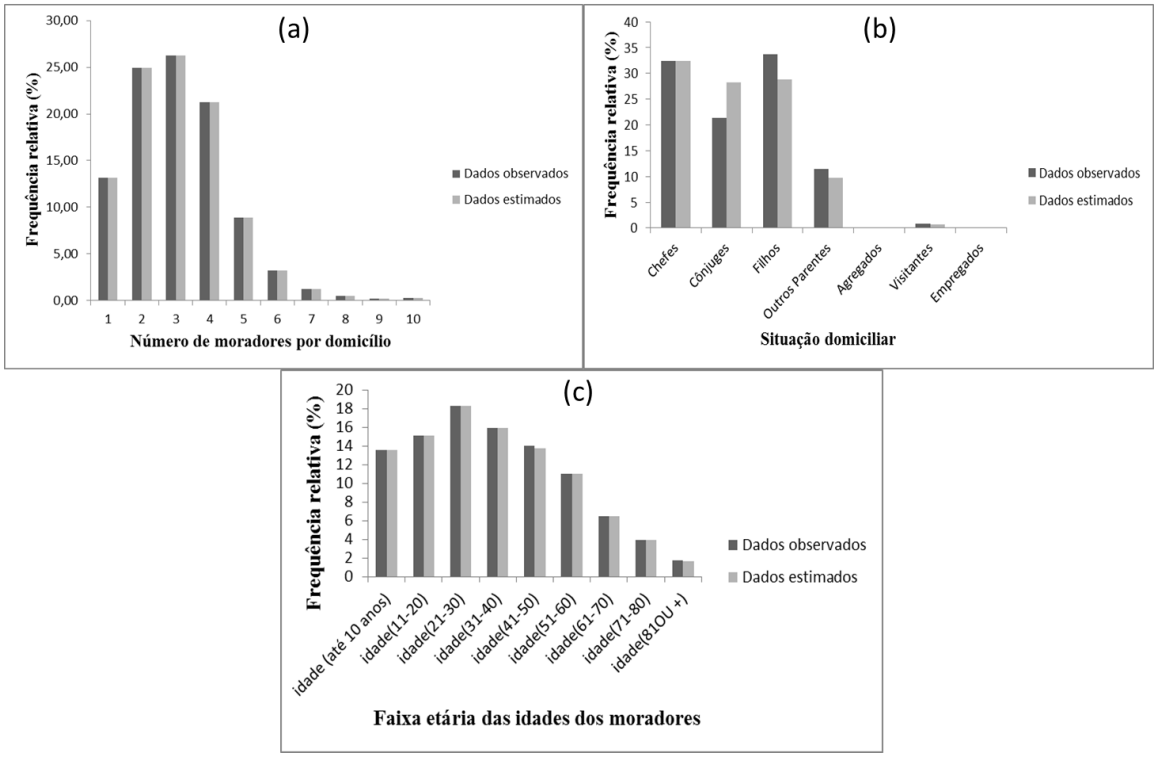


Figura 6.1 – Histogramas das variáveis estimadas x observadas.

6.2.2 Validação da população sintética

Depois de analisar a amostra obtida pela população sintética e comparar com a base agregada do censo 2010, foi realizada a validação dessa amostra utilizando os microdados 2010 do IBGE através do teste estatístico de *Kolmogorov-Smirnov*, realizado no SPSS.

Os resultados obtidos se encontram na Tabela 6.4 e estão apresentadas as decisões de reter ou rejeitar a hipótese nula (distribuição da variável é a mesma entre as categorias da amostra) para cada variável da população sintética.

Neste caso, para o intervalo de confiança de 95%, a hipótese nula foi retida para valores de sig. > 0,05, ou seja, a distribuição é normal e a hipótese nula foi rejeitada para valores de sig. < 0,05, pois a distribuição não é normal. De acordo com os resultados, as variáveis que não passaram no teste foram: sexo (homem, mulher), cônjuges, filhos, outros parentes, idade (até 10 anos, 11-20 anos; 21-30 anos; 31-40 anos; 41-50 anos; 51-60 anos; 61-70 anos) e renda de 0 a 2 salários mínimos, pois apresentaram a distribuição da população (variáveis obtidas pela população sintética) diferente da amostra (variáveis do censo 2010-microdados).

Tabela 6.4 – Variáveis analisadas pelo teste estatístico de Kolmogorov-Smirnov.

Teste de Kolmogorov-Smirnov (Amostras independentes)			
Hipótese nula: A distribuição da variável é a mesma entre as categorias de amostra.			
	Variáveis	**Sig.**	**Decisão (hipótese nula)**
1	Domicílio (1 morador)	0,985	Reter
2	Domicílio (2 moradores)	0,999	Reter
3	Domicílio (3 moradores)	1,000	Reter
4	Domicílio (4 moradores)	1,000	Reter
5	Domicílio (5 moradores)	1,000	Reter
6	Domicílio (6 moradores)	1,000	Reter
7	Domicílio (7 moradores)	1,000	Reter
8	Domicílio (8 moradores)	1,000	Reter
9	Domicílio (9 moradores)	1,000	Reter
10	Domicílio (10 ou mais moradores)	1,000	Reter
11	Mulheres no domicílio	0,000	**Rejeitar**
12	Homens no domicílio	0,000	**Rejeitar**
13	Chefes no domicílio .	1,000	Reter
14	Cônjuges no domicílio	0,000	**Rejeitar**
15	Filhos no domicílio	0,000	**Rejeitar**
16	Outros parentes no domicílio	0,000	**Rejeitar**
17	Agregados no domicílio	1,000	Reter
18	Visitantes no domicílio	1,000	Reter
19	Empregados no domicílio	1,000	Reter
20	Idade até 10 anos	0,001	**Rejeitar**
21	Idade 11-20 anos	0,000	**Rejeitar**
22	Idade 21-30 anos	0,000	**Rejeitar**
23	Idade 31-40 anos	0,000	**Rejeitar**
24	Idade 41-50 anos	0,000	**Rejeitar**
25	Idade 51-60 anos	0,000	**Rejeitar**
26	Idade 61-70 anos	0,035	**Rejeitar**
27	Idade 71-80 anos	0,596	Reter
28	Idade mais que 80 anos	1,000	Reter
29	Renda (sem renda)	1,000	Reter
30	Renda (0-2 sál. mín)	0,000	**Rejeitar**
31	Renda (2-3 sál. mín)	0,479	Reter
32	Renda (3-5 sál. mín)	0,404	Reter
33	Renda (5-10 sál. mín)	0,537	Reter
34	Renda (>10 sál. mín)	0,768	Reter

6.3 Modelagem da demanda por transportes

6.3.1 Modelagem RNA (Modelos 1 e 3)

Conforme apresentado no Capítulo 5, o processo de modelagem RNA através da rede perceptrons multicamadas (MLP) dos modelos 1 e 3 foi através da análise de sensibilidade e consistiu das etapas de: definição das variáveis e particionamento das amostras, tipo de arquitetura e tipo de treinamento.

Iniciou-se com a etapa da escolha da variável dependente (*viagens por domicílio*) e das variáveis independentes, sendo todas variáveis numéricas padronizadas. Em seguida, foram definidas as proporções da amostra de treinamento e teste através do exame das cinco proporções definidas anteriormente (50/50, 60/40, 70/30, 80/20 e 90/10). Desta forma, o produto desta etapa foi a definição da amostra de treinamento e teste mais adequada para a calibração dos quatro modelos propostos no método.

Entretanto, para escolher a proporção da amostra mais adequada foi necessário definir a arquitetura da rede e o tipo de treinamento, pois são duas características com diversas combinações que precisam ser analisadas juntamente com o exame das proporções da amostra de treinamento e teste. A escolha adequada dessas características influenciou no desempenho da rede.

A definição da arquitetura da rede não foi uma tarefa simples, pois foi alcançado por um processo de tentativa e erro considerado um processo de busca exaustiva (elevado custo de tempo). O mesmo ocorreu na definição do tipo de treinamento da MLP, juntamente com a definição do algoritmo de otimização da rede neural.

Vale ressaltar que, foram realizadas inúmeras simulações porém, foram observados com mais detalhes os testes das 270 combinações e foram escolhidas as melhores combinações (30) através dos menores erros relativos, que estão apresentados na Tabela 6.5. Além disso, também são apresentadas nesta tabela, as amostras que apresentaram menores erros relativos de cada uma das partições (amostra de treinamento e teste), que estão destacados em negrito.

As amostras que apresentaram menores erros em todas as proporções estavam constituídas pela seguinte composição:

- arquitetura personalizada composta por duas camadas ocultas (10-8 unidades, respectivamente);
- função de ativação na camada oculta (tangente hiperbólica);
- função de ativação da camada de saída (identidade);
- treinamento do tipo lote;
- algoritmo de otimização gradiente conjugado.

Tabela 6.5 – Melhores combinações obtidas pela MLP.

REDIMENSIONAMENTO DAS COVARIÁVEIS - PADRONIZADO					
Camada ocultas		1 ou 2	**Função de ativação:**	tangente hiperbólica	
Camada de saída		1	**Função de ativação:**	identidade	
	Arquitetura	**Treinamento**		**Parâmetros**	
	Tipo	**Tipo**	**Algoritmo de otimização**	**Erro relativo**	**Erros quadráticos**
Treinamento 90% e Teste 10%					
1	automática	lote	gradiente conjugado	0,495	685,699
2	automática	lote	gradiente descendente	0,542	750,977
3	automática	online	gradiente descendente	0,607	841,236
4	automática	minilote	gradiente descendente	0,631	873,996
5	personalizado	lote	gradiente conjugado	0,489	677,043
6	**personalizado**	**lote**	**gradiente conjugado**	**0,439**	**608,005**
Treinamento 80% e Teste 20%					
7	automática	lote	gradiente conjugado	0,49	602,226
8	automática	lote	gradiente descendente	0,501	615,806
9	automática	online	gradiente descendente	0,658	809,17
10	automática	minilote	gradiente descendente	0,634	779,934
11	personalizado	lote	gradiente conjugado	0,49	603,315
12	**personalizado**	**lote**	**gradiente conjugado**	**0,449**	**552,145**
Treinamento 70% e Teste 30%					
13	automática	lote	gradiente conjugado	0,469	504,885
14	automática	lote	gradiente descendente	0,57	612,954
15	automática	online	gradiente descendente	0,568	610,376
16	automática	minilote	gradiente descendente	0,629	676,136
17	personalizado	lote	gradiente conjugado	0,454	487,982
18	**personalizado**	**lote**	**gradiente conjugado**	**0,444**	**477,561**
Treinamento 60% e Teste 40%					
19	automática	lote	gradiente conjugado	0,444	407,542
20	automática	lote	gradiente descendente	0,446	409,378
21	automática	online	gradiente descendente	0,624	572,775
22	automática	minilote	gradiente descendente	0,592	543,068
23	personalizado	lote	gradiente conjugado	0,457	418,957
24	**personalizado**	**lote**	**gradiente conjugado**	**0,401**	**367,989**
Treinamento 50% e Teste 50%					
25	automática	lote	gradiente conjugado	0,433	331,58
26	automática	lote	gradiente descendente	0,441	338,167
27	automática	online	gradiente descendente	0,57	437,223
28	automática	minilote	gradiente descendente	0,62	475,419
29	personalizado	lote	gradiente conjugado	0,393	301,372
30	**personalizado**	**lote**	**gradiente conjugado**	**0,358**	**274,313**

Em relação aos erros relativos apresentados (dados em negrito), percebe-se que este erro foi menor para a amostra (90%/10%) do que para as amostras (80%/20% e 70%/30%), pois o erro relativo é totalmente dependente do maior ou menor valor da grandeza a ser medida, revelando a precisão da medida feita. No caso da amostra (90%/10%) os valores de viagens foram de no mínimo 0 e no máximo 39 e nas amostras (80%/20% e 70%/30%) foram de no mínimo 0 e no máximo 28 viagens, por isso foi encontrada esta pequena diferença nos valores dos erros relativos. Além disso, o erro relativo foi menor pois apresentou uma correlação nos dados da amostra (90%/10%) maior do que das amostras (80%/20% e 70%/30%).

Outra medida analisada foi a soma dos erros quadráticos, que mede a variabilidade dos dados, ou seja, o desvio das observações em torno da média. Esta medida foi maior para a amostra (90%/10%) e menores para as demais.

Após, definidas as composições da MLP que apresentaram menores erros, os modelos foram treinados com as cinco diferentes partições de amostra (6, 12, 18, 24 e 30) e assim, foram obtidas as viagens produzidas por domicílio para cada partição de amostra.

E por fim, foi escolhida a rede cuja acurácia preditiva se apresentou mais adequada, através das medidas de desempenho dos erros obtidos (erro médio, raiz quadrada do erro médio, erro relativo, correlação e erro relativo), ou seja, a diferença entre a variável observada e a estimada, de acordo com a Tabela 6.6.

Tabela 6.6 – Medidas de desempenho dos erros das cinco partições da amostra.

		ERROS				
Amostra	Teste	EM	RQEM	ER	CORREL	DESVPADA
6 (90%/10%)	10%	-0,123	3,925	0,724	0,555	3,930
12 (80%/20%)	20%	-0,460	4,023	0,714	0,551	4,000
18 (70%/30%)	30%	-0,199	3,820	0,712	0,551	3,817
24 (60%/40%)	40%	-0,106	**3,736**	**0,698**	**0,566**	**3,736**
30 (50%/50%)	50%	-0,172	3,853	0,746	0,537	3,850

A **amostra 24 (60% treinamento e 40% teste)** foi a amostra escolhida, pois apresentou os menores erros e a maior correlação dos dados na amostra de validação (amostra teste). A amostra de treinamento (60%) foi composta por 1.836 domicílios e 100% dos dados foram considerados válidos para a análise. O tempo de treinamento da rede foi de 0:00:01,08 e a função de erro minimizada foi a função soma dos erros quadráticos. Assim sendo, a MLP (60% treinamento e 40% teste) foi utilizada nos Modelos 1 e 3 para prever as viagens por domicílio e os principais resultados obtidos foram: a importância das variáveis independentes e o gráfico de resíduos.

6.3.1.1 Modelo 1

Vale lembrar que os dados de entrada utilizados no Modelo 1 foram as 34 variáveis independentes apresentadas na Tabela 5.3 (página 70).

i) Importância das variáveis independentes

No modelo RNA não é possível dizer o "sentido"da relação entre as variáveis independentes e a variável prevista, somente a importância das variáveis independentes. A importância é uma medida da quantidade de alterações do valor da variável prevista pelo modelo da rede para diferentes valores da variável independente.

O software apresenta as variáveis de maior importância. Tais variáveis provém maior montante informacional ao algoritmo para a correta identificação e diferenciação do número de viagens.

Segundo Ortúzar e Willumsen (2011) as variáveis que têm sido consideradas em vários estudos de geração de viagens por domicílio são: renda, posse de automóvel, estrutura do domicílio e tamanho da família.

A variável que apresentou maior capacidade de produzir viagens por domicílio pela MLP escolhida foi a variável *domicílio com 9 moradores*. Depois, a variável *renda maior que 10 salários mínimos* e *número de chefes no domicílio*, como segunda e terceira variáveis explicativas de maior relevância, respectivamente.

A classificação de cada variável segundo ao grau de importância pode ser visualizada na Figura 6.2.

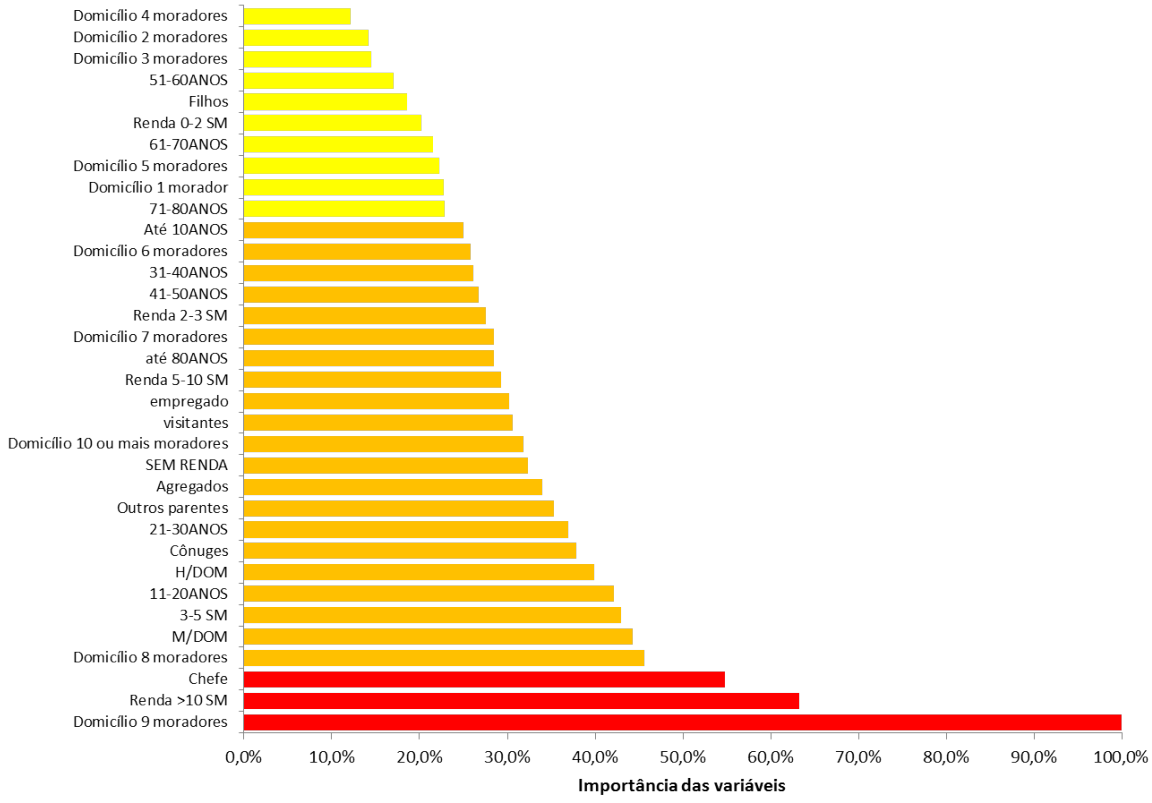


Figura 6.2 – Importância das variáveis analisadas para a produção de viagens por domicílios (Modelo1).

ii) Gráfico de dispersão entre valores observados e previstos.

A validação do Modelo 1 foi através das medidas de desempenho de erros (erro médio, raiz quadrada do erro médio, erro relativo, correlação e erro relativo), dados estes já utilizados anteriormente na escolha da partição da amostra ideal para este trabalho. Os resultados da previsão de viagens por domicílio obtida pelo Modelo 1 foram analisados de uma forma eficiente e visual, a fim de verificar a qualidade do modelo. Isso foi feito através de diagramas de dispersão entre valores observados e previstos (amostra teste-40%). A Figura 6.3 ilustra estes resultados.

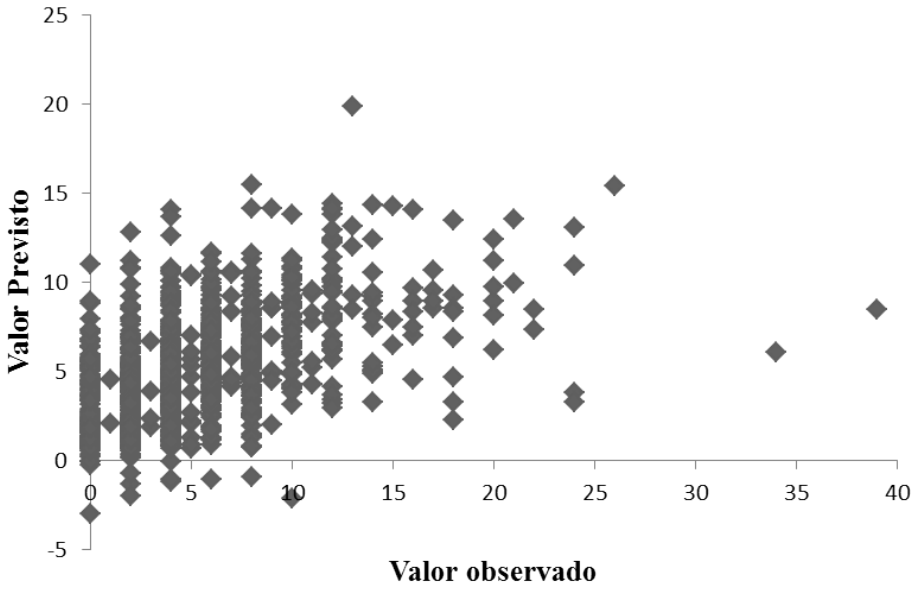


Figura 6.3 – Gráfico de dispersão da variável: viagens por domicílio - Modelo 1:valores observados x previstos.

Verificou-se então, que os resultados obtidos pelo Modelo 1 de geração de viagens por domicílio apresentou uma boa relação entre os dados observados e estimados (Coeficiente de Pearson=0,566).

6.3.1.2 Modelo 3

Os dados de entrada utilizados no Modelo 3, foram os mesmos dados do banco da Pesquisa O/D utilizadas no Modelo 1, porém foram utilizadas somente as variáveis que passaram no teste estatístico de *Kolmogorov-Smirnov* (Tabela 6.4). Portanto, para o Modelo 3 foram utilizadas 21 variáveis independentes.

Foram repetidas as etapas do Modelo 1 e utilizadas as mesmas características da MLP anterior para o treinamento da rede do Modelo 3. A Tabela 6.7 apresenta os principais dados de processamento da rede.

Tabela 6.7 – Dados para processamento da RNA.

Processamento da MLP		
Amostra	Treinamento (60%)	1836
	Teste (40%)	1221
Camada de entrada	Covariáveis	21
Camadas ocultas	Número de camadas	2
	Função de ativação	Tangente Hiperbólica
Camada de saída	Variáveis	1
	Função de ativação	Identidade
	Função de erro	Soma dos Quadrados
Treinamento	Tipo	Lote
	Algoritmo de otimização	Gradiente conjugado

i) Importância das variáveis independentes

A variável que melhor explicou a previsão de viagens por domicílios obtidas no Modelo 3 foi a variável *renda maior que 10 salários mínimos.* Depois, foram as variáveis (*domicílios com 9 moradores; domicílios com 8 moradores; domicílios com 7 moradores; domicílios com 6 moradores e domicílios com 5 moradores)*, ou seja, membros na família.

A sequência de importância das outras variáveis pode ser observada pela Figura 6.4.

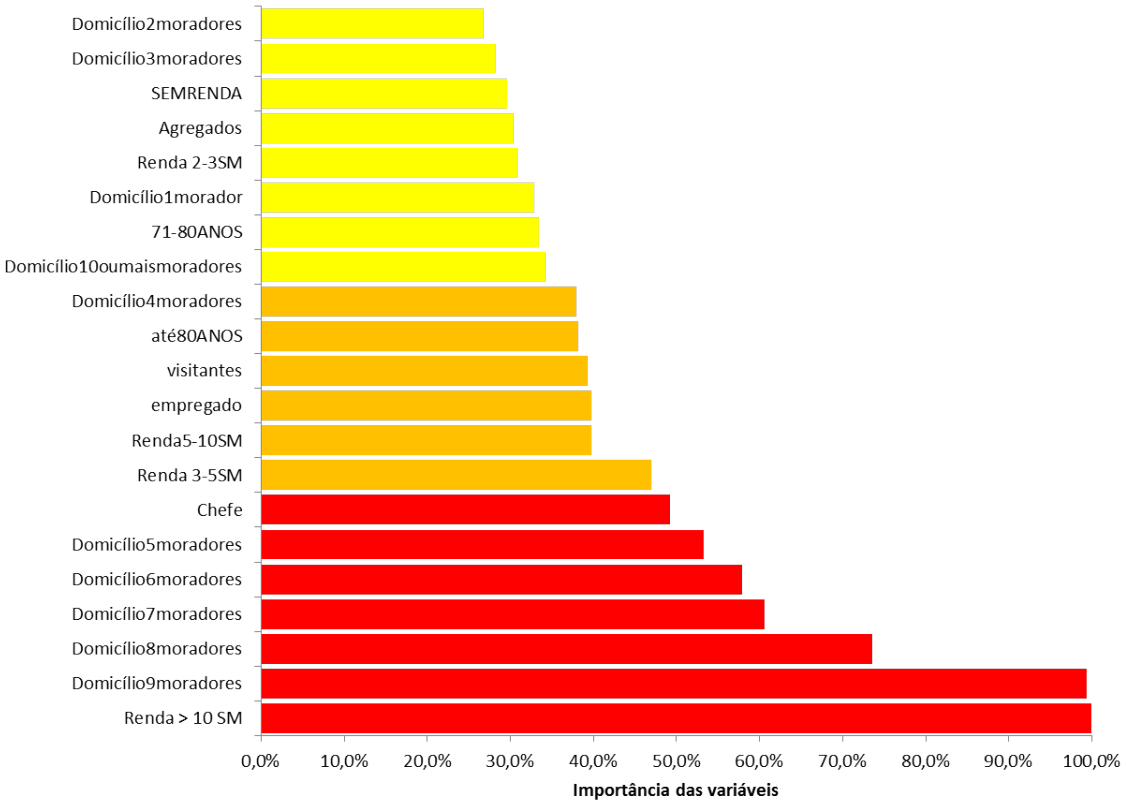


Figura 6.4 – Importância das variáveis analisadas para a produção de viagens por domicílios (Modelo3).

ii) Gráficos de dispersão entre valores observados e previstos.

A validação do Modelo 3 foi examinada através de medidas de desempenho dos erros obtidos da diferença entre a variável observada e a estimada (erro médio, raiz quadrada do erro médio, erro relativo, correlação e erro relativo). Foram também, utilizados os gráficos de dispersão entre valores observados e estimados (amostra teste-40%) para verificar a qualidade do modelo. A Tabela 6.8 e a Figura 6.5 ilustram esses resultados.

Tabela 6.8 – Medidas de desempenho do Modelo 3.

ERROS	TREINO	TESTE
	60%	**40%**
EM	-0,055	0,176
RQEM	2,97	3,64
ER	0,521	0,662
CORREL	0,692	0,583
DESVPADA	2,97	3,638

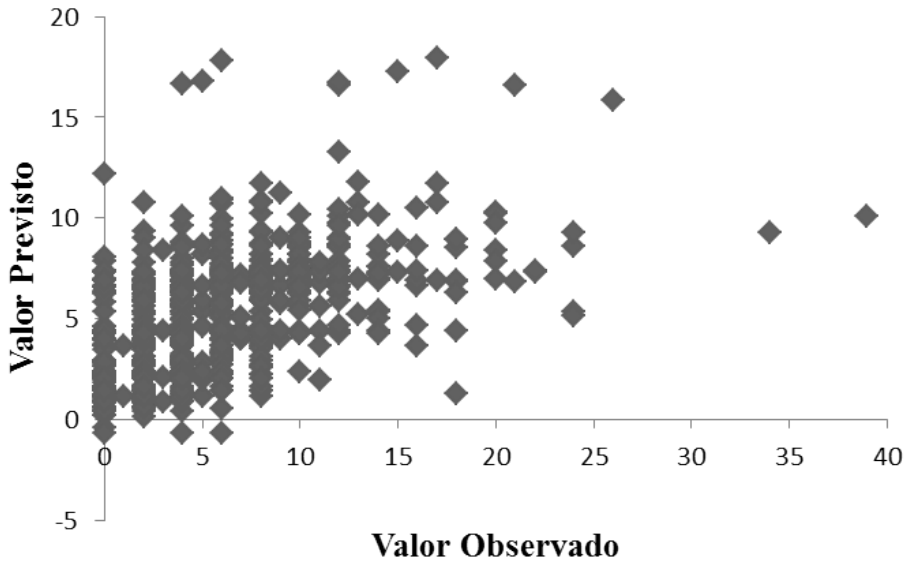


Figura 6.5 – Gráfico de dispersão da variável: viagens por domicílio - Modelo 3:valores observados x estimados.

Constatou-se que os resultados do Modelo 3, assim como o Modelo 1, apresentou uma boa relação entre os dados observados e estimados (Coeficiente de Pearson=0,583). A diferença dos resultados do Modelo 1 e 3 foram sutis.

6.3.2 Modelagem RLM (Modelos 2 e 4)

Apesar de, a Regressão Linear Múltipla não ser a técnica mais indicada para previsão de viagens urbanas, ela foi escolhida por ser a técnica mais utilizada na área de modelagem de demanda por transportes.

O processo de modelagem RLM dos Modelos 2 e 4 consiste das etapas de: definição das variáveis, análise empírica, método de inserção das variáveis, análise de colinearidade, equação de calibração e análise dos coeficientes.

A modelagem tradicional se deu pela análise da variável dependente (*viagens por domicílios*) e das 34 variáveis independentes, estas já definidas anteriormente. Vale ressaltar que, a amostra utilizada na modelagem RLM foi a mesma amostra utilizada para a modelagem RNA (amostra 24-60% treinamento e 40% teste).

A análise empírica das variáveis foi fundamental na modelagem tradicional, pois foram escolhidas as variáveis (23) que realmente explicavam o fenômeno da produção de viagens por domicílio. Segundo (ORTÚZAR; WILLUMSEN, 2011) as variáveis que explicam a produção de viagens por domicílio são: renda, tamanho da família e estrutura domiciliar. Assim, as variáveis utilizadas para a modelagem RLM estão apresentadas na Tabela 6.9.

Tabela 6.9 – Dados de entrada da RLM.

Variáveis-RLM		
Tamanho do domicílio	**Estrutura domiciliar**	**Renda domiciliar**
Dom. com 1 morador	chefe	sem renda
Dom. com 2 moradores	cônjuges	0-2 Sál. Mínimos
Dom. com 3 moradores	filhos	2-3 Sál. Mínimos
Dom. com 4 moradores	outros parentes	3-5 Sál. Mínimos
Dom. com 5 moradores	agregados	5-10 Sál. Mínimos
Dom. com 6 moradores	visitantes	>10 Sál. Mínimos
Dom. com 7 moradores	empregados	
Dom. com 8 moradores		
Dom. com 9 moradores		
Dom. com 10 +moradores		

O método de inserção das variáveis utilizado na modelagem RLM foi o *Stepwise*. Em seguida, antes de obter a equação de calibração foi considerada a correlação entre as variáveis, a fim de evitar problemas de multicolinearidade. A Tabela 6.10 apresenta os valores de Coeficientes de *Pearson* das variáveis selecionadas anteriormente. Não foi observado nenhum valor alto de correlação entre variáveis explicativas que pudesse prejudicar a aplicação da técnica.

O exame da matriz de correlação indica que filhos (X_{15}) tem a mais elevada correlação bivariada com a variável dependente (0,530).

Entretanto, após verificar a matriz de correlação, foi realizada a análise estatística de colinearidade para analisar o impacto de colinearidade nas variáveis independentes da equação de regressão. Para isto, foram utilizados os valores de tolerância e VIF. E por fim, foi realizada uma análise crítica das equações de calibração dos Modelos 2 e 4, em relação as variáveis independentes significativas selecionadas pelo método estatístico RLM. Essas análises estão apresentadas a seguir especificamente para cada um dos modelos.

Tabela 6.10 – Matriz de correlação das variáveis independentes e dependente.

	Viagens	Dom1	Dom 2	Dom 3	Dom4	Dom 5	Dom 6	Dom 7	Dom 8	Dom 9	Dom 10 >	Chefe	Cônj.	Fil.	Out.par.	Agreg.	Empreg.	Visit.	s/Renda	0-2SM	2-3SM	3-5SM	5-10SM	10-20SM
Variável Dependente																								
Viagens por domicilio																								
Variáveis independentes																								
Domicilio1morador	-0,269	1,000																						
Domicilio2moradores	-0,370	-0,182	1,000																					
Domicilio3moradores	-0,066	-0,173	-0,321	1,000																				
Domicilio4moradores	0,242	-0,168	-0,311	-0,296	1,000																			
Domicilio5moradores	0,272	-0,118	-0,218	-0,207	-0,201	1,000																		
Domicilio6moradores	0,199	-0,072	-0,133	-0,127	-0,123	-0,086	1,000																	
Domicilio7moradores	0,142	-0,041	-0,076	-0,072	-0,070	-0,049	-0,030	1,000																
Domicilio8moradores	0,098	-0,024	-0,045	-0,043	-0,042	-0,029	-0,018	-0,010	1,000															
Domicilio9moradores	0,096	-0,021	-0,038	-0,037	-0,035	-0,025	-0,015	-0,009	-0,005	1,000														
Domicilio10oumaismoradores	-0,016	-0,007	-0,014	-0,013	-0,013	-0,009	-0,005	-0,003	-0,002	-0,002	1,000													
Chefe	0,146	-0,062	-0,055	-0,026	0,032	0,015	0,069	0,078	0,070	0,066	-0,004	1,000												
Cônjuges	0,220	-0,461	-0,041	0,034	0,157	0,136	0,067	0,045	0,018	0,041	0,014	-0,158	1,000											
Filhos	**0,530**	-0,316	-0,495	-0,113	0,247	0,396	0,321	0,251	0,115	0,124	0,036	-0,090	0,202	1,000										
Outros parentes	0,113	-0,133	-0,179	-0,066	-0,017	0,120	0,229	0,180	0,242	0,241	0,142	-0,014	-0,119	-0,017	1,000									
Agregados	-0,006	-0,028	-0,046	0,019	0,009	0,039	0,016	-0,012	0,027	-0,006	-0,002	-0,082	-0,068	-0,050	-0,011	1,000								
Empregados	0,001	-0,019	0,045	-0,034	0,009	-0,023	0,026	-0,008	-0,005	-0,004	-0,001	-0,010	-0,055	-0,032	-0,027	0,081	1,000							
Visitantes	-0,015	-0,016	0,018	0,045	-0,028	-0,020	-0,012	-0,007	-0,004	-0,003	-0,001	0,016	-0,078	-0,044	-0,023	0,046	-0,003	1,000						
SEMRENDA	0,410	-0,279	-0,364	-0,074	0,223	0,256	0,199	0,208	0,170	0,119	0,051	0,040	0,264	0,586	0,150	-0,002	-0,042	-0,039	1,000					
0-2SM	0,051	-0,108	-0,059	-0,051	0,052	0,033	0,111	0,080	0,053	0,064	0,122	0,046	0,035	0,102	0,195	0,009	-0,004	0,001	0,013	1,000				
2-3SM	0,151	-0,073	-0,055	-0,017	0,055	0,064	0,000	0,041	0,010	0,037	-0,013	0,093	0,066	0,077	0,017	-0,030	0,018	0,002	0,010	-0,126	1,000			
3-5SM	0,113	-0,029	-0,058	0,008	0,063	0,013	0,022	-0,027	-0,003	-0,020	-0,007	0,038	0,016	0,040	-0,012	-0,005	-0,018	0,043	0,034	-0,138	-0,031	1,000		
5-10SM	0,101	-0,042	-0,032	0,040	0,068	-0,023	-0,027	-0,027	-0,016	-0,014	-0,005	0,068	0,023	0,001	-0,037	-0,019	-0,013	-0,011	0,014	-0,125	-0,041	0,071	1,000	
10-20SM	0,110	-0,017	-0,022	-0,017	0,038	0,010	0,014	0,008	-0,010	-0,008	-0,003	-0,012	0,057	0,015	-0,013	0,035	0,196	-0,007	0,032	-0,055	-0,041	-0,002	-0,013	1,000

6.3.2.1 Modelo 2

As variáveis de entrada utilizadas no Modelo 2 foram as 23 variáveis independentes apresentadas na Tabela 6.9. Juntamente com a matriz de correlação apresentada anteriormente e após a seleção das variáveis consideradas importantes para o modelo pelo Método *Stepwise* foi verificada a multicolinearidade dessas variáveis, de acordo com os valores de tolerância e VIF apresentados na Tabela 6.11.

Tabela 6.11 – Análise de multicolinearidade da RLM (Modelo 2).

Variáveis	Tolerância	VIF
Filhos	**0,586**	1,706
Chefes	0,919	1,088
Cônjuges	0,877	1,141
Outros parentes	0,915	1,093
> 10 Sal. Min	0,992	1,008
5-10 Sal. Min	0,982	1,019
2-3 Sal. Min	0,971	1,03
3-5 Sal. Min	0,987	1,013
Dom. com 4 moradores	0,749	1,335
Dom. com 2 moradores	**0,659**	1,517
Dom. com 5 moradores	0,702	1,425
Agregados	0,965	1,036

Desta forma, apenas duas variáveis apresentaram valores de tolerância menores que 0,700. Isto ressalta que, o impacto da multicolinearidade nos dados é mínimo. Portanto, foi necessário analisá-las (*filhos e domicílios com 2 moradores*) e assim, foi retirada da análise a variável de menor correlação com a dependente, a variável *domicílios com 2 moradores*.

Depois, de finalizadas as etapas de verificação das variáveis significativas, o modelo foi novamente executado pelo método *Stepwise* com a amostra de calibração (60%), porém com um novo conjunto de 22 variáveis independentes. E por fim, foi obtido o Modelo 2 e os principais resultados estão descritos na Tabela 6.12.

A regressão linear múltipla resultou em valores significativos para a maioria dos coeficientes correspondentes a cada variável independente. As exceções foram para as nove variáveis: *domicílios com 1 morador*, *domicílios com 2 moradores*, *domicílios com 6 moradores*, *domicílios com 7 moradores*, *domicílios com 8 moradores*, *domicílios com 9 moradores*, *domicílios com 10 ou mais moradores*, *visitantes e empregados*. Por esta razão, estas variáveis foram suprimidas por não influenciarem os resultados do ajuste do Modelo 2.

Assim, foi realizada uma análise crítica do modelo, em relação às variáveis independentes significativas selecionadas pelo método estatístico e pode-se observar na Tabela 6.12 que todos os valores de coeficientes das variáveis independentes foram positivos e a ordem de grandeza dos parâmetros estimados fez sentido para explicar o fenômeno de previsão de viagens por

domicílio e os valores da estatística t apresentaram-se coerentes com o esperado, significando que as variáveis selecionadas foram significativas. Foi considerada a constante neste modelo, pois se apresentou significativa.

Tabela 6.12 – Principais resultados do modelo linear escolhido (Modelo 2).

Modelo linear: Geração de viagens		
Variáveis significativas	**R^2 =0,405**	
	Coeficientes	**t**
Constante	-1	-3,702
Filhos	1,53	17,171
Chefes	1,804	9,934
Cônjuges	1,076	6,21
Outros parentes	0,697	6,843
> 10 Sal. Min	2,321	5,388
5-10 Sal. Min	1,418	4,521
2-3 Sal. Min	0,551	4,899
3-5 Sal. Min	0,864	4,131
Dom. com 4 moradores	1,075	4,939
Dom. com 5 moradores	0,95	3,357
Agregados	0,763	2,069
Dom. com 3 moradores	0,411	2,099
Sem renda	0,151	2,005

Portanto, pode-se afirmar que o *número de viagens por domicílio* aumenta com o aumento do *número de domicílios com renda maior que 10 salários mínimos*. Da mesma maneira, ocorre para as demais variáveis independentes.

Foram observados os resíduos do modelo linear e verificou-se que os resíduos atendem a suposição de normalidade, como mostra a Figura 6.6, porém não atendem à suposição de homocedasticidade (Figura 6.7).

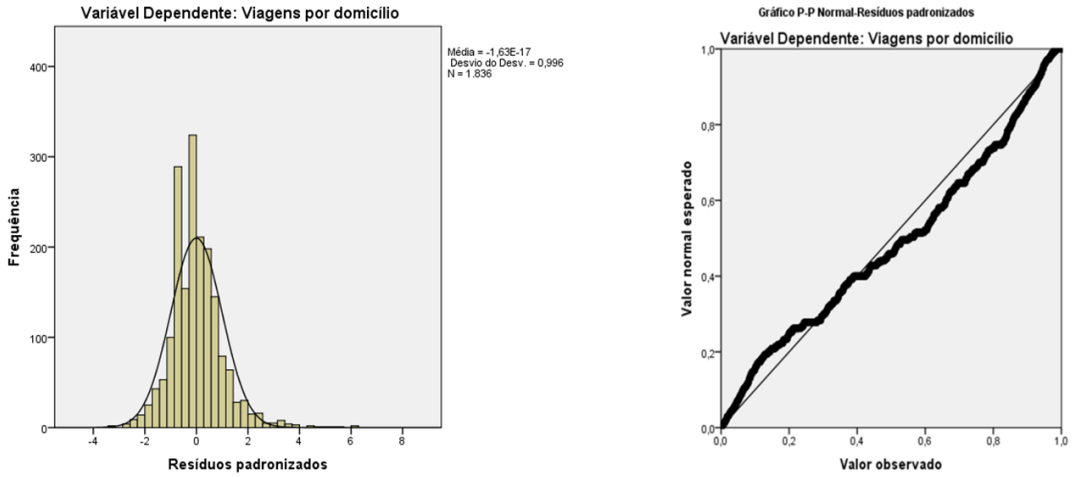


Figura 6.6 – Análise da normalidade (Modelo 2).

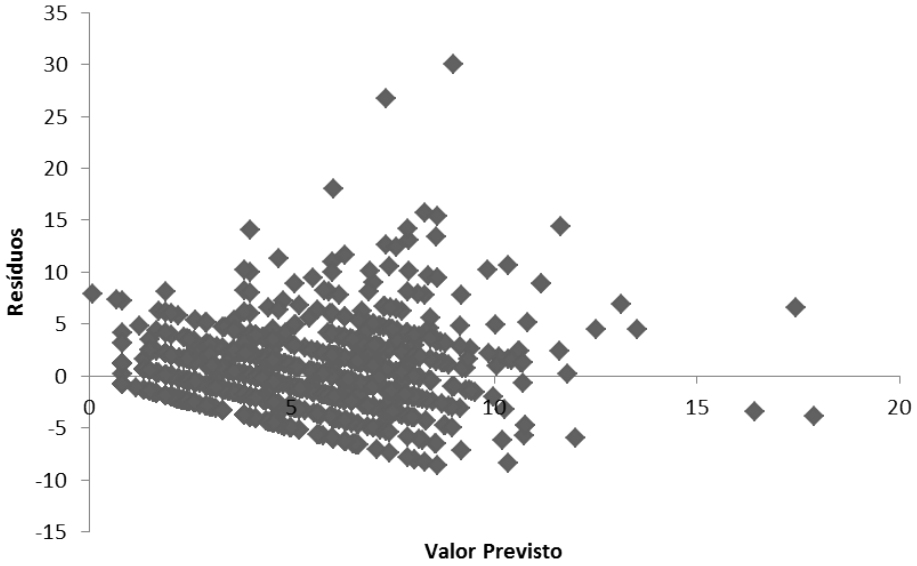


Figura 6.7 – Análise dos resíduos – heteroscedasticidade (Modelo 2).

A qualidade do ajuste do Modelo 2 foi mensurada através do gráfico de dispersão (Figura 6.8, ou seja, previsão das viagens por domicílio entre valores observados e estimados da amostra teste (40%).

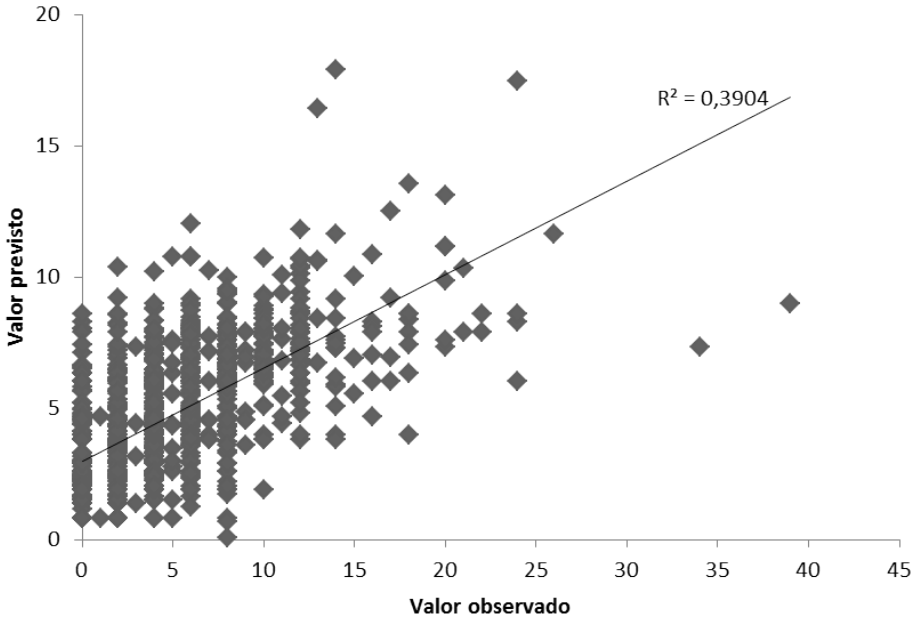


Figura 6.8 – Gráfico de dispersão- Modelo 2: Viagens por domicílio (observada x prevista).

A validação do Modelo 2 foi avaliada através das medidas de desempenho de erros (erro médio, raiz quadrada do erro médio, erro relativo, correlação e erro relativo), de acordo com a Tabela 6.13.

Verificou-se que o Modelo 2 de produção de viagens por domicílio foi considerado signi-

ficativo, pois apresentou um bom poder preditivo (R^2=0,405). Vale lembrar que, por ser um modelo desagregado o R^2 é bem menor que dos modelos agregados.

Tabela 6.13 – Medidas de desempenho de erros (Modelo2).

ERROS	TREINO	TESTE
	60%	40%
EM	0	0,146
RQEM	3,173	3,504
ER	0,595	0,614
CORREL	0,636	0,625
DESVPADA	3,173	3,501

6.3.2.2 Modelo 4

As variáveis independentes utilizadas como dados de entrada para o Modelo 4 foram as mesmas variáveis selecionadas pelo *Stepwise* no Modelo 2 (13 variáveis): *filhos*, *chefes*, *cônjuges*, *outros parentes*, *renda > 10 salários mínimos*, *renda de 5-10 salários mínimos*, *renda de 2-3 salários mínimos*, *renda de 3-5 salários mínimos*, *domicílios com 4 moradores*, *domicílios com 5 moradores*, *agregados*, *domicílio com 3 moradores* e *sem renda*.

Da mesma forma, como realizado na modelagem RNA (Modelo 3), que foram selecionadas somente as variáveis que passaram no teste estatístico de *Kolmogorov-Smirnov*, foi feito o mesmo para a modelagem tradicional (Modelo 4). Das treze variáveis selecionadas no Modelo 2, dez variáveis foram realmente utilizadas como dados de entrada para o Modelo 4, como mostra na Tabela 6.14.

Tabela 6.14 – Dados de entrada da RLM (Modelo 4).

Variáveis-RLM		
Tamanho do domicílio	**Estrutura domiciliar**	**Renda domicliar**
Dom. com 3 moradores	Chefe	sem renda
Dom. com 4 moradores	Agregados	2-3 Sál. Mínimos
Dom. com 5 moradores		3-5 Sál. Mínimos
		5-10 Sál. Mínimos
		>10 Sál. Mínimos

Depois, seguindo as premissas da RLM, antes de executar o método escolhido para estimar viagens, foi realizada juntamente com a matriz de correlação (Tabela 6.10) a análise da multicolinearidade dessas variáveis, de acordo com os valores de tolerância e VIF apresentados na Tabela 6.15.

Nota-se que as variáveis não apresentaram valor de tolerância menor que 0,750. Portanto, o impacto da multicolinearidade nos dados é mínimo.

Tabela 6.15 – Análise de multicolinearidade da RLM (Modelo 4).

Variáveis	Tolerância	VIF
Sem renda	0,847	1,18
Dom. com 5 moradores	0,781	1,28
Dom. com 4 moradores	0,75	1,334
Chefe	0,983	1,017
2-3 Sal. Min	0,976	1,025
> 10 Sal. Min	0,995	1,005
5-10 Sal. Min	0,98	1,02
3-5 Sal. Min	0,987	1,013
Dom. com 3 moradores	0,826	1,21

Logo, com as variáveis significativas do modelo selecionadas, o modelo foi calibrado pelo método *Stepwise*, com 60% da amostra de calibração e assim, foi escolhido o modelo que está descrito na Tabela 6.16.

Tabela 6.16 – Principais resultados do modelo linear escolhido (Modelo 4).

Modelo linear: Geração de viagens		
Variáveis significativas	**R^2 =0,298**	
	Coeficientes	**t**
Constante	1,069	4,457
Sem renda	0,92	13,58
Dom. com 5 moradores	3,128	11,28
Dom. com 4 moradores	2,235	9,983
Chefe	1,026	5,435
2-3 Sal. Min	0,739	6,088
> 10 Sal. Min	2,327	4,988
5-10 Sal. Min	1,336	3,931
3-5 Sal. Min	0,906	3,993
Dom. com 3 moradores	0,748	3,565

Para finalizar, foi realizada uma análise crítica do modelo, em relação às variáveis independentes significativas selecionadas pelo método estatístico e pode-se observar na Tabela 6.16 que todos os valores de coeficientes das variáveis independentes foram positivos, inclusive a constante. Também fizeram sentido a ordem de grandeza dos parâmetros estimados para explicar o fenômeno de previsão de viagens por domicílio e os valores da estatística t apresentaram coerentes com o esperado. Tudo isto, denotou que as variáveis selecionadas neste modelo foram significativas.

Desta forma, é possível afirmar que o *número de viagens por domicílios* aumenta com o aumento do *número de domicílios com 5 moradores*. Da mesma maneira, ocorre para *domicílios com renda maior que 10 salários mínimos* e também, com as demais variáveis independentes. Os resíduos do modelo linear são apresentados na Figura 6.9 e observa-se que os resíduos

atendem a suposição de normalidade, porém não atendem à suposição de homocedasticidade (Figura 6.10).

Com os valores observados das viagens por domicílio obtidas pela Pesquisa O/D, puderam-se realizar através do gráfico de dispersão dos valores observados e estimados (Figura 6.11) uma análise da qualidade do modelo.

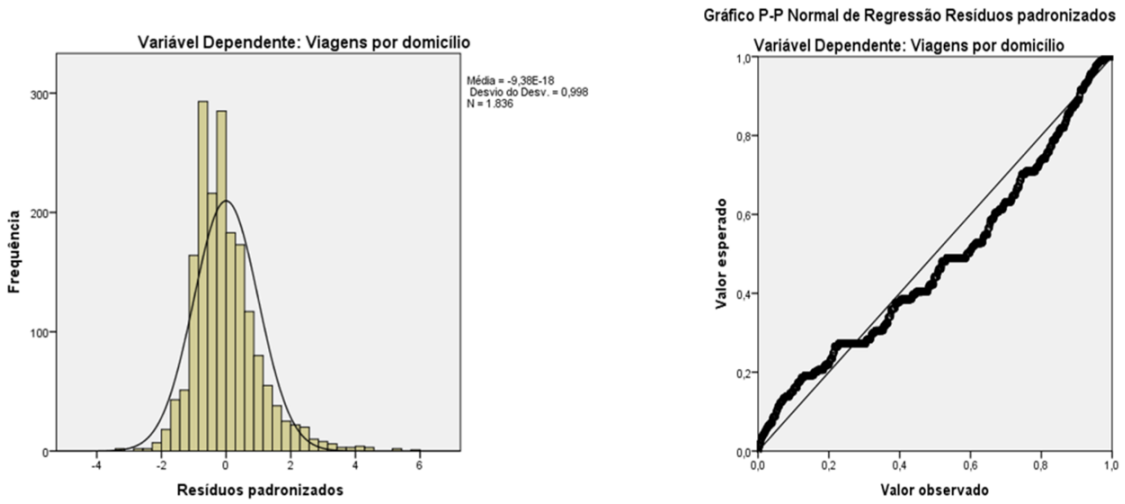


Figura 6.9 – Análise da normalidade (Modelo 4).

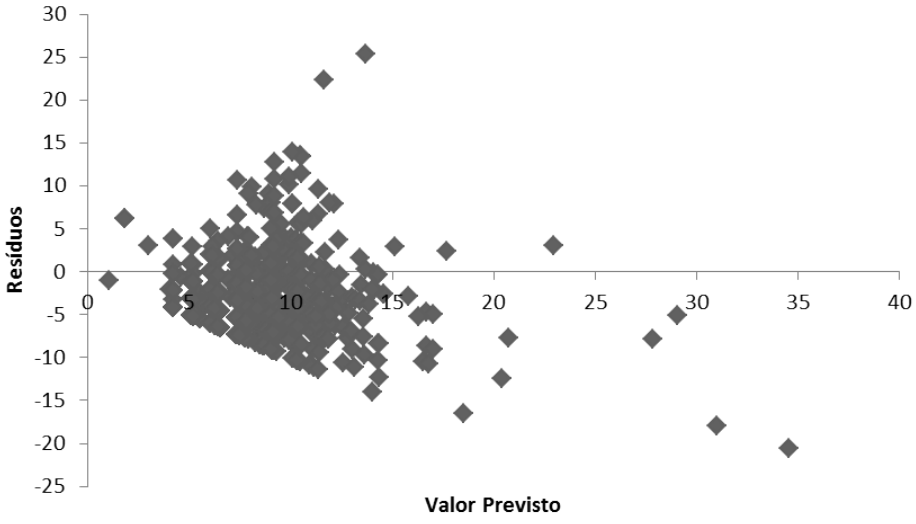


Figura 6.10 – Análise dos resíduos – heteroscedasticidade (Modelo 4).

A validação do Modelo 4 foi mensurada através das medidas de desempenho de erros (erro médio, raiz quadrada do erro médio, erro relativo, correlação e erro relativo) da amostra teste (40%), como mostra a Tabela 6.17. A análise da Tabela 6.17 indica que o Modelo 4 de produção de viagens por domicílio não houve um bom ajuste pela RLM. O valor do coeficiente de determinação foi baixo e os valores dos erros foram maiores que o Modelo 2.

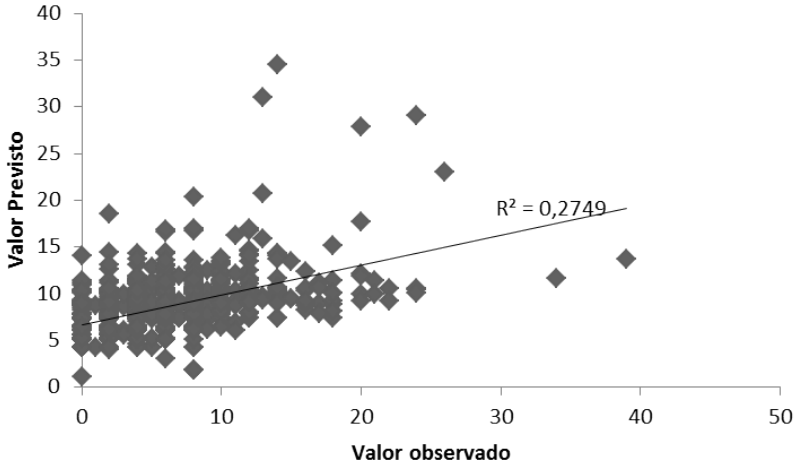


Figura 6.11 – Gráfico de dispersão-Modelo 4:Viagens por domicílio (observada x prevista).

Tabela 6.17 – Medidas de desempenho de erros (Modelo 4).

ERROS	TREINO	TESTE
	60%	**40%**
EM	0	-3,402
RQEM	3,447	5,12
ER	0,702	1,31
CORREL	0,546	0,524
DESVPADA	3,447	3,827

6.4 Comparação dos Modelos

Sintetizando neste capítulo a extração dos resultados para os casos de modelagem RNA e modelagem tradicional, foram promovidas comparações por meio dos valores observados e previstos da amostra teste – 40%. A Tabela 6.18 apresenta os principais resultados dos modelos de calibração (1, 2, 3 e 4).

Em relação a modelagem tradicional (Modelos 2 e 4) foi utilizado para a comparação dos modelos os valores do coeficiente de determinação (R^2), depois de serem analisadas todas as suposições prévias necessárias para a técnica de regressão linear (variáveis todas significativas e os valores dos coeficientes e da estatística t foram coerentes com o esperado). Observou-se um valor de quase 0,4 para o Modelo 2 o que significou um bom poder preditivo para modelagem desagregada.

Apesar do Modelo 2 ter sido mais adequado de acordo com os resultados apresentados (menor erro e maior correlação dos dados) que o Modelo 4, o Modelo 2 poderia ser escolhido para a previsão de viagens por domicílio da população sintética desta tese se todas as variáveis independentes utilizadas para a calibração deste modelo tivessem passado no teste estatístico de *Kolmogorov-Smirnov*.

Tabela 6.18 – Comparação dos modelos de calibração obtidos.

COMPARAÇÃO DOS MODELOS DE CALIBRAÇÃO		
	Modelagem RNA	
Modelo	**1**	**3**
Variáveis	**RNA(OD)**	**RNA(OD-filtro)**
Entrada	34	21
Mais significativa	Domicílios com 9 moradores	Renda > 10 sal. Min.
RQEM	3,736	3,640
CORREL	0,566	0,583
	Modelagem tradicional	
Modelo	**2**	**4**
Variáveis	**RLM(OD)**	**RLM(OD-filtro)**
Entrada	34	13
Mais significativa	Renda > 10 sal. Min.	Domicílios com 5 moradores
R^2	0,390	0,275
RQEM	3,504	5,120
CORREL	0,625	0,524

No âmbito desta tese doutorado, em que o tema é gerar uma população sintética para prever viagens por domicílio, só foram utilizadas as variáveis independentes previstas pela população sintética que passaram no teste estatístico de *Kolmogorov-Smirnov*.

Assim, o Modelo 2 (dados O/D completos) foi utilizado apenas para verificar se apresentariam diferenças significativas em relação ao Modelo 4. Pode-se observar que o poder preditivo do Modelo 4 diminui e o erro aumentou em relação ao Modelo 2. Porém, este modelo será utilizado para a previsão de viagens por domicílio da população sintética apesar de não ter conseguido um bom ajuste pela RLM.

Na modelagem RNA (Modelos 1 e 3), os modelos apresentaram algumas diferenças sutis entre a análise dos erros, porém com a retirada das variáveis que não passaram no teste o Modelo 3 melhorou. O Modelo 3 (RNA) apresentou uma relação mais forte (Coeficiente de *Pearson* = 0,583) entre os valores estimados e observados e o erro foi menor comparado ao Modelo 1.

De maneira análoga, na modelagem tradicional, com a retirada das variáveis que não passaram no teste de *Kolmogorov-Smirnov* o Modelo 4 piorou significativamente, ou seja, o coeficiente de determinação diminuiu em aproximadamente 27%, o erro aumentou quase 32% e a correlação entre os dados diminuiu 16% do valor em relação ao Modelo 2. Desta forma, o Modelo 3 (modelagem RNA) que utiliza técnicas não paramétricas e exploratórias apresentou melhorias na modelagem da demanda por transportes (modelagem desagregada) com dados provenientes de Pesquisa Origem e Destino.

6.5 Validação dos resultados

Finalmente, foi alcançado o objetivo principal desta tese, através do uso dos modelos de demanda por transportes (3 e 4) para previsão de viagens por domicílio da população sintética.

O objetivo deste capítulo é validar as viagens por domicílio da população sintética obtidas pelos modelos propostos 3 e 4 (Apêndice E.5). Devido à falta de dados reais, a validação foi realizada em três etapas:

1. Análise agregada por categoria domiciliar segundo a quantidade de membros na família;
2. Comparação de médias dado um intervalo de confiança, e
3. Análise dos resultados com a exclusão de valores maiores ou iguais aos percentis (95, 90, 85, 80 e 75).

6.5.1 Viagens por domicílio da população sintética obtidas pelo Modelo 3

Foram amostrados 68.833 domicílios com 212.263 moradores que realizaram 295.228 viagens e uma média de 4,3 viagens por domicílio (valores previstos pelo Modelo 3).

1) Análise agregada por categoria domiciliar segundo a quantidade de membros na família:

Para esta primeira análise foi necessário agregar as viagens por categorias de domicílios. Das 295.228 viagens obtidas pelo Modelo 3, verificou-se que 72% das viagens foram realizadas pelos domicílios com 3 a 5 moradores.

Fazendo uma comparação com as viagens observadas (14.702 viagens), apesar das distribuições entre as categorias dos domicílios (domicílio com 3 moradores, domicílio com 4 moradores e domicílio com 5 moradores) apresentarem uma pequena variação entre as porcentagens de viagens observadas e previstas, no total, 72% das viagens por domicílio da Pesquisa O/D também foram realizadas pelos domicílios com 3 a 5 moradores. Isto pode ser observado na Figura 6.12.

Ademais, as viagens por domicílio obtidas pela amostra (Pesquisa O/D) e as viagens por domicílio da população (população sintética) previstas pelo Modelo 3 apresentaram próximo o número médio de viagens por domicílio, 4,80 e 4,30, respectivamente.

Apesar de, existirem algumas diferenças de valores obtidos entre as viagens previstas e observadas, as distribuições das viagens por domicílio agregadas por categorias de domicílios obtida pelo Modelo 3 foi precisa, pois a correlação dos dados foi alta e o erro quadrático médio foi baixo (0,0003). Desta forma, a Figura 6.13 apresenta um diagrama de dispersão entre os valores estimados pelo Modelo 3 e os valores observados pela Pesquisa O/D das viagens por categorias de domicílios, em que o valor do coeficiente de Pearson (medida de intensidade da

associação linear existente entre as variáveis) foi 0,968, corroborando a hipótese de alta relação entre os valores.

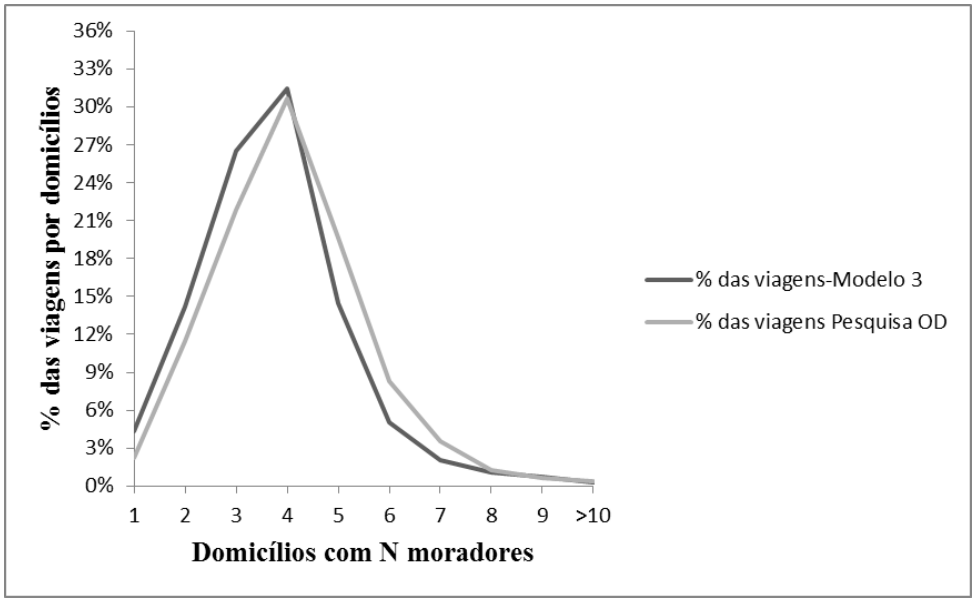


Figura 6.12 – Comparação das % de viagens por domicílio da Pesquisa O/D com a % das viagens por domicílio sintético pelo Modelo 3.

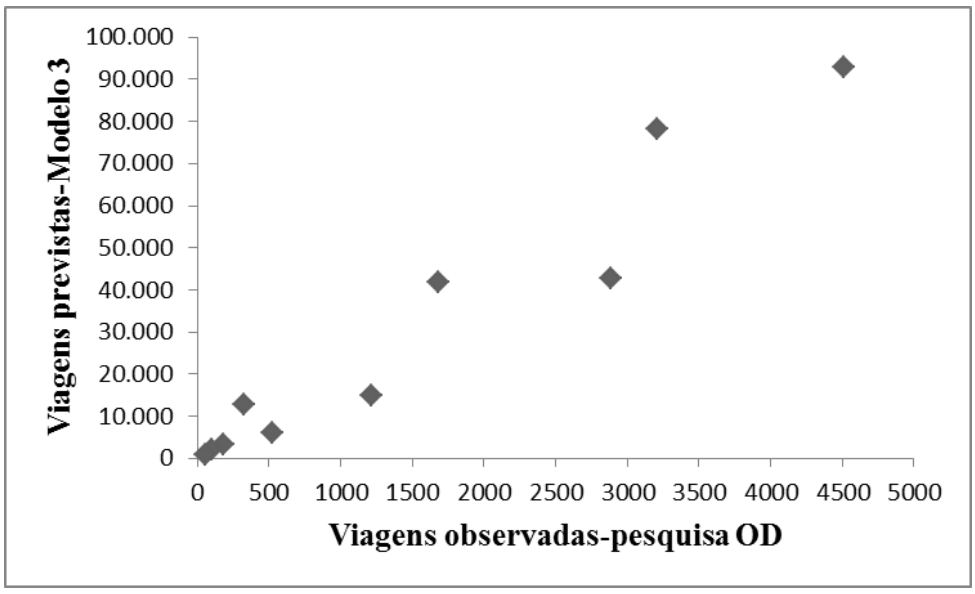


Figura 6.13 – Gráfico de dispersão entre as viagens obtidas pela amostra e as viagens por domicílio previstas pelo Modelo 3.

2) Comparação de médias dado um intervalo de confiança:

Iniciou-se esta etapa com o cálculo das médias da amostra (viagens por domicílio da Pesquisa O/D) e da população (viagens por domicílio da população sintética) por setores censitários.

A Figura 6.14 apresenta os histogramas das médias de viagens nos setores censitários da amostra da Pesquisa O/D (a) e das médias de viagens previstas nos setores censitários pelo

Modelo 3 (b). Observou-se que em ambas as distribuições a média de viagens na maioria dos setores censitários foi de 3 a 6 viagens por domicílio.

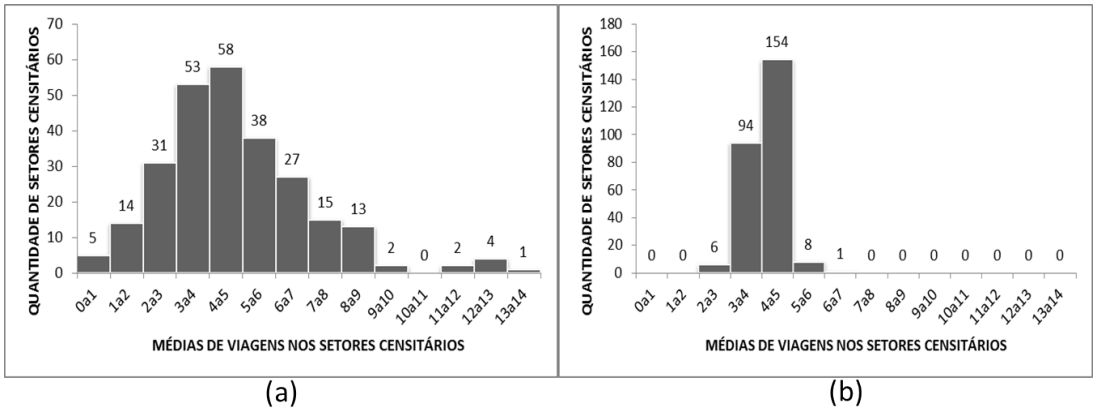


Figura 6.14 – Distribuição das viagens médias por setores. (a) Amostra O/D e (b) Modelo 3 – RNA.

Depois de calculadas e analisadas as médias de viagens por domicílio nos setores censitários foi feito um teste para verificar se as médias de viagens nos setores censitários da amostra e da população eram estatisticamente diferentes. Para isto, foi fixado o nível de confiança de 95% e foram obtidos os intervalos de confiança (limite inferior e superior) para cada setor censitário da amostra (Pesquisa O/D). Essas informações se encontram detalhadas no Apêndice E.6.

Em seguida, foram verificadas se as médias de viagens nos setores censitários obtidas pelo Modelo 3 eram significativamente iguais às médias de viagens por domicílio observadas (Pesquisa O/D), ou seja, estavam dentro dos intervalos de confiança definidos pela amostra.

Vale ressaltar que como a análise das médias foi realizada de forma agregada por setor censitário foram analisados somente os setores censitários contidos na amostra e na população. Desta forma, foram retirados da amostra dois setores censitários e 25 setores censitários da população para que fossem possíveis as comparações. Portanto, concluiu-se que 68,50% dos 263 setores censitários da população possuíam as viagens por domicílio válidas. Assim, a população sintética obtida neste trabalho estimou um total de 270.636 viagens por domicílio nos 63 setores censitários. Após a validação (comparação de médias dado um intervalo de confiança) das viagens, foram consideradas viagens válidas um total de 187.006 viagens por domicílio.

3) Análise dos resultados com a exclusão de valores maiores ou iguais aos percentis (95, 90, 85, 80 e 75):

Após analisadas as médias de viagens dado um intervalo de confiança na etapa 2, foram retirados da amostra (Pesquisa O/D) os valores maiores ou iguais aos percentis 95, 90, 85, 80 e 75 (Apêndice E.7) e, em seguida foi analisado se a taxa de acerto (viagens válidas) das viagens por domicílio da população sintética estimadas pelo Modelo 3 aumentou. A Tabela 6.19 apresenta os valores correspondentes a cada percentil utilizado para a exclusão dos valores na amostra da Pesquisa O/D.

Tabela 6.19 – Valores correspondentes aos percentis.

PERCENTIS	VIAGENS/DOMICÍLIO
95º	Exclusão dos valores >=12
90º	Exclusão dos valores >=10
85º	Exclusão dos valores >=8
80º	Exclusão dos valores >=8
75º	Exclusão dos valores >=7

Depois de retirados os valores maiores ou iguais correspondentes a cada percentil foram recalculadas as médias de viagens por domicílio nos setores censitários da amostra da O/D e obtidos os novos intervalos de confiança, que se encontram detalhados no Apêndice E.7.

Em seguida, foi analisado se as médias de viagens por domicílio da população sintética agregadas nos setores censitários estavam dentro desses novos intervalos de confiança e assim, foram obtidos os setores censitários que possuíam as viagens por domicílio válidas.

Cada percentil foi analisado separadamente e foram obtidos os resultados apresentados na Figura 6.15. Desta forma, a Tabela 6.20 apresenta os principais resultados das viagens estimadas pelo Modelo 3 após a etapa de validação (análise dos percentis).

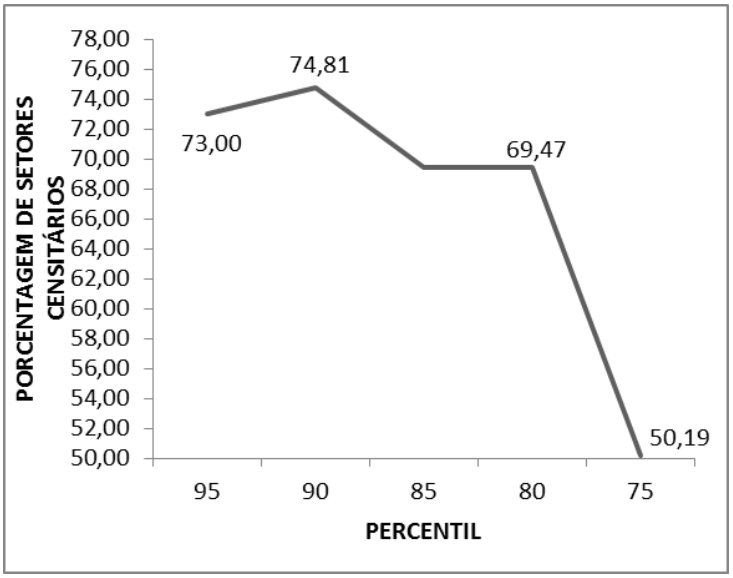


Figura 6.15 – Porcentagem de setores censitários com médias de viagens por domicílio do Modelo 3 que estão dentro dos intervalos da amostra da O/D.

Os resultados da modelagem RNA (Modelo 3) foram muito satisfatórios, pois a validação das viagens por domicílio apresentada anteriormente (etapa 2) obteve taxa de acerto de 69,1% e com a retirada dos valores maiores ou iguais a oito viagens por domicílio (percentis 95, 90, 85, 80) as taxas de acertos aumentaram. De maneira diferente ocorreu na retirada dos valores de viagens por domicílio acima de sete (percentil75) no qual a taxa de acertos diminui quase 25%.

Tabela 6.20 – Resumo dos resultados obtidos.

Intervalos da Amostra O/D	**Viagens válidas**	**Taxa de acertos**
completa	187.006	69,10%
percentil 95º	200.462	74,10%
percentil 90º	204.755	75,70%
percentil 85º	190.295	70,30%
percentil 80º	190.295	70,30%
percentil 75º	123.916	45,80%

A Tabela 6.21 apresenta um resumo das médias das viagens por domicílio da amostra da O/D e da população sintética.

Tabela 6.21 – Análise das médias das viagens na retirada dos valores referentes aos percentis.

População Sintética		**Amostra O/D**	
Média = 4,28		Intervalos de confiança	
	Média	-	+
Completa	4,8	4,65	4,96
Percentil 95º	4,23	4,11	4,35
Percentil 90º	3,93	3,83	4,05
Percentil 85º	3,51	3,41	3,61
Percentil 80º	3,51	3,41	3,61
Percentil 75º	2,9	2,81	2,99

A queda na taxa de acertos com a retirada dos valores maiores ou iguais ao percentil 75 fez com que a média de viagens por domicílio da amostra da O/D diminuísse de 4,80 para 2,90 viagens por domicílio e se justifica, pois quase 60% das viagens por domicílio da amostra são maiores que 7 (Figura 6.16).

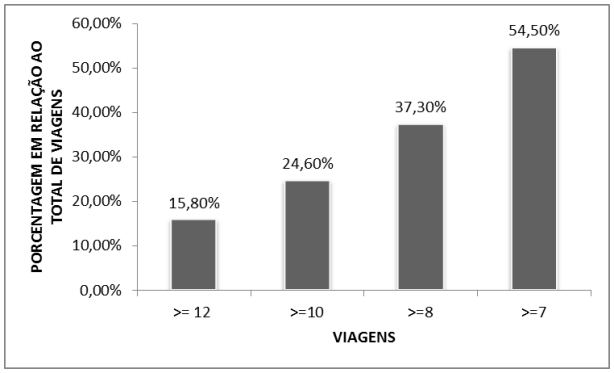


Figura 6.16 – Porcentagem de viagens retiradas da amostra.

Devido a isto, as médias de viagens nos setores censitários da amostra (percentil 75º) também diminuíram. A Figura 6.17 apresenta a distribuição das médias de viagens nos setores censitários da amostra completa e da amostra após a retirada dos valores maiores que sete.

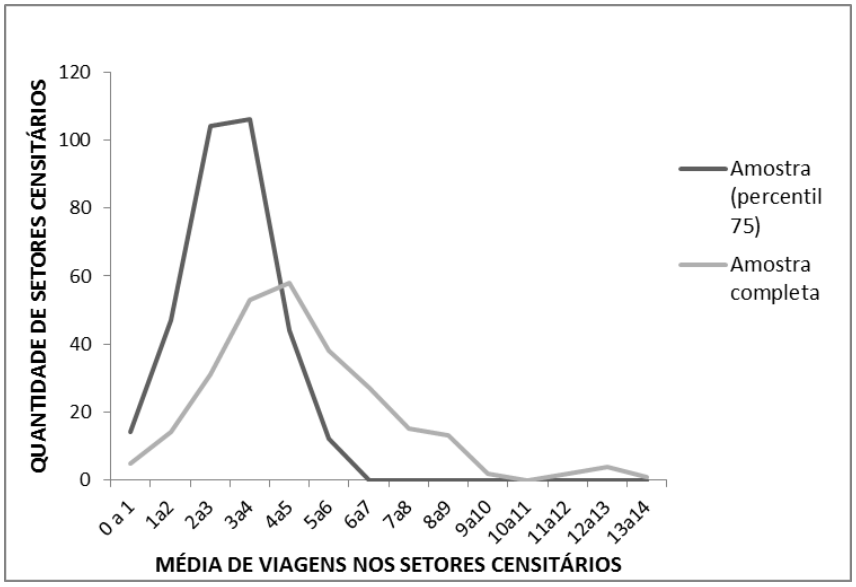


Figura 6.17 – Distribuição das médias de viagens por domicílio nos setores censitários da amostra completa e da amostra após a retirada dos valores maiores ou iguais a sete viagens.

Como a média de viagens da população por setor censitário foi maior que 4 e a maioria dos setores censitários da amostra após a retirada dos valores maiores ou iguais a 7 (percentil 75) apresentaram médias menores que 4, portanto, muitos setores censitários da população não estavam dentro dos intervalos de confiança da amostra, justificando a queda na taxa de acertos.

Contudo, em geral, o uso conjunto da RNA e da população sintética obtida neste trabalho para modelar viagens por domicílio foi considerado satisfatório, pois através da comparação de médias dado um intervalo de confiança, quase 80% das viagens por domicílio da população sintética estimadas pelo Modelo 3 foram consideradas válidas.

6.5.2 Viagens por domicílio da população sintética obtidas pelo Modelo 4:

Foram amostrados 68.833 domicílios com 212.263 moradores que geraram 231.173 viagens por domicílio e uma média de 3,4 viagens por domicílio (valores previstos pelo Modelo 4).

Foram repetidas as três etapas propostas de validação realizadas para o Modelo 3.

1) Análise agregada por categoria domiciliar segundo a quantidade de membros na família:

Das 231.173 viagens por domicílio obtidas pelo Modelo 4 verificou-se que a maioria das viagens (72%) foram realizadas pelos domicílios com 2 a 4 moradores.

Desta forma, a distribuição das viagens agregadas por categoria de domicílios ocorreu de maneira um pouco diferente dos dados observados pela Pesquisa O/D (72% das viagens por domicílio foram realizadas pelos domicílios com 3 a 5 moradores).

De acordo com a Figura 6.18 nota-se que o gráfico das porcentagens das viagens por cate-

gorias de domicílio do Modelo 4 não se encontra muito próximo do gráfico da Pesquisa O/D, admitindo a diferença das distribuições das viagens por categoria de domicílios.

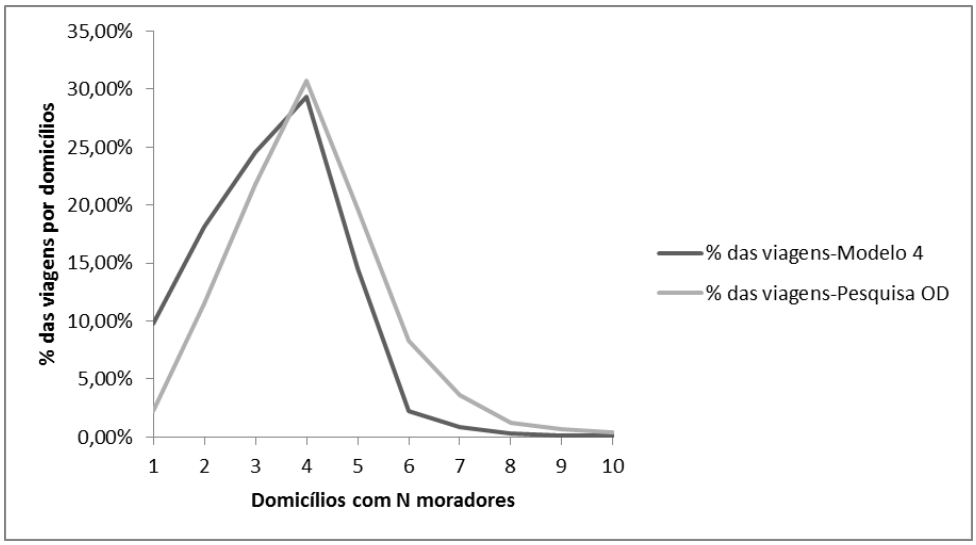


Figura 6.18 – Comparação das % de viagens da Pesquisa O/D com a % das viagens por domicílio sintético pelo Modelo 4 .

Logo, o número médio de viagens por domicílio obtidas pela Pesquisa O/D foi bem maior que o obtido pelo Modelo 4, 4,8 e 3,4 viagens por domicílio, respectivamente. Apesar, das viagens por domicílio agregadas por categorias apresentarem nas distribuições uma pequena diferença entre as viagens estimadas e observadas e uma diferença maior em relação a média total de viagens por domicílio (mais de 40%), no geral a distribuição das viagens por domicílio agregadas por categorias de domicílios obtida pelo Modelo 4 foi precisa, pois a correlação dos dados foi alta e o erro quadrático médio foi baixo (0,0005).

Assim, devido as diferenças apresentadas entre as viagens estimadas pelo Modelo 4 e as viagens observadas pela Pesquisa O/D, observou-se na Figura 6.19 através do diagrama de dispersão que o valor do Coeficiente de Pearson foi um pouco menor que o obtido pelo Modelo 3 (0,914). Porém ainda assim, o Modelo 4 apresentou uma alta relação entre os valores observados e estimados.

2) Comparação de médias dado um intervalo de confiança:

Foram calculadas as médias da amostra (viagens por domicílio da Pesquisa O/D) e da população (viagens por domicílio da população sintética) por setores censitários.

A Figura 6.20 apresenta os histogramas das médias de viagens nos setores censitários da amostra (a) e das médias de viagens da população previstas pelo Modelo 4 (b).

Observou-se que as distribuições são bem diferentes, pois nas viagens da população previstas pelo Modelo 4 quase 100% dos setores censitários apresentaram médias de viagens de 3 a 4 viagens por domicílio e na amostra da O/D, apenas 20,2% dos setores apresentaram essa média.

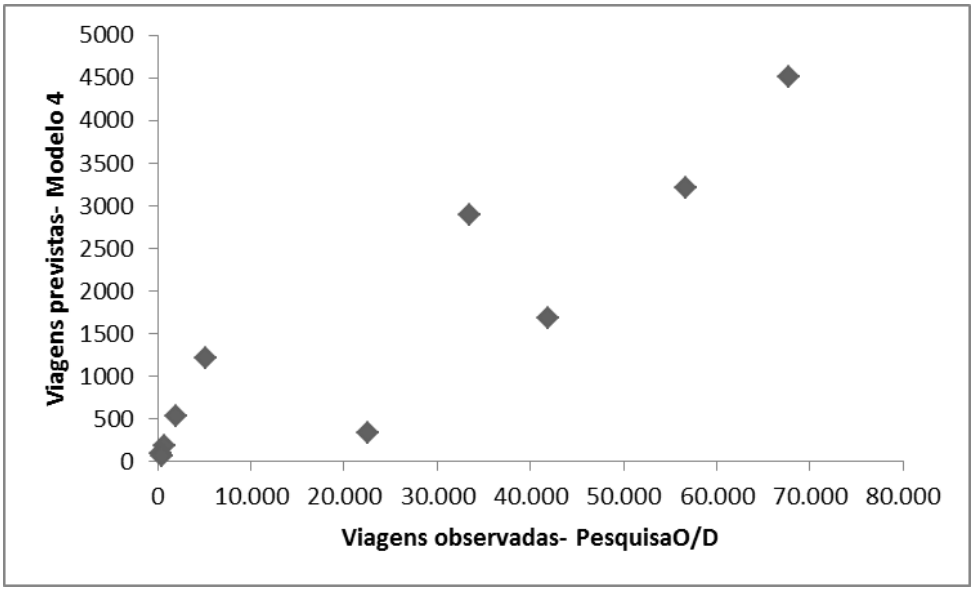


Figura 6.19 – Gráfico de dispersão entre as viagens obtidas pela OD e as viagens obtidas pelo Modelo 4.

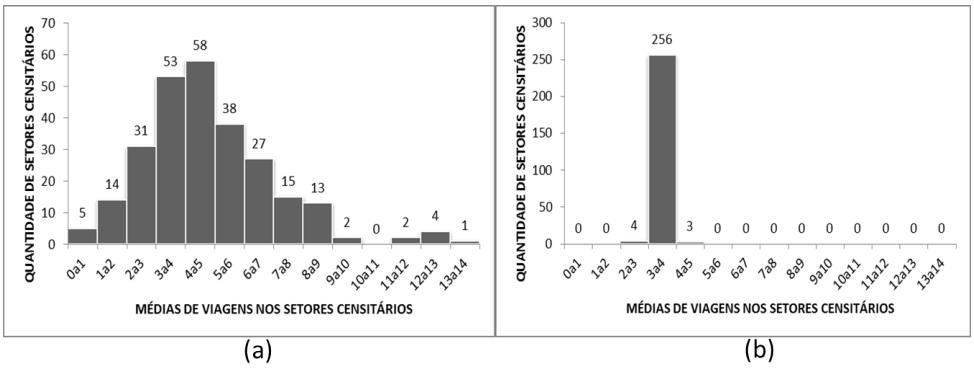


Figura 6.20 – Distribuição das viagens médias por setores. (a) Amostra OD e (b) Modelo 4 – RLM.

Em seguida, foi feito um teste para verificar se as médias da amostra e da população eram estatisticamente diferentes. Para isto, foi fixado o nível de confiança de 95% e foram obtidos os intervalos de confiança (limite inferior e superior) para cada setor censitário da amostra (Pesquisa O/D). Os intervalos de confiança de cada setor censitário se encontram no Apêndice E.6.

Com os intervalos de confiança obtidos, foram verificadas se as médias de viagens nos setores censitários obtidas pelo Modelo 4 eram significativamente iguais às médias de viagens domiciliares observadas (Pesquisa O/D), ou seja, estavam dentro dos intervalos de confiança definidos pela amostra da O/D.

Assim como no Modelo 3, foram analisados somente os setores censitários que continham na amostra e na população. Desta forma, foram retirados da amostra dois setores censitários e 25 setores censitários da população para que fossem possíveis as comparações. Portanto, o Modelo 4 apresentou que 58,90% dos 263 setores censitários possuíam as viagens por domicílio válidas.

A população sintética obtida neste trabalho estimou pelo Modelo 4 um total de 213.327 viagens por domicílios, porém após a validação das viagens nesta etapa, foram consideradas viagens válidas um total de 118.564 viagens por domicílio.

3) Análise dos resultados com a exclusão de valores maiores ou iguais aos percentis (95, 90, 85, 80 e 75):

Da mesma forma que no Modelo 3, foram retirados da amostra (Pesquisa O/D) os valores maiores ou iguais aos percentis (95, 90, 85, 80 e 75) e foi analisado se a retirada dos valores referentes a cada percentil causariam um impacto significativo nos dados das viagens por domicílio da população (Apêndice E.7).

Para isto, foram analisadas se as médias das viagens por domicílio da população sintética agregadas por setores censitários (Modelo 4) estavam dentro desses novos intervalos de confiança e assim, foram obtidos os setores censitários que possuíam as viagens por domicílio válidas. Cada percentil foi analisado separadamente e foram obtidos os resultados apresentados na Figura 6.21.

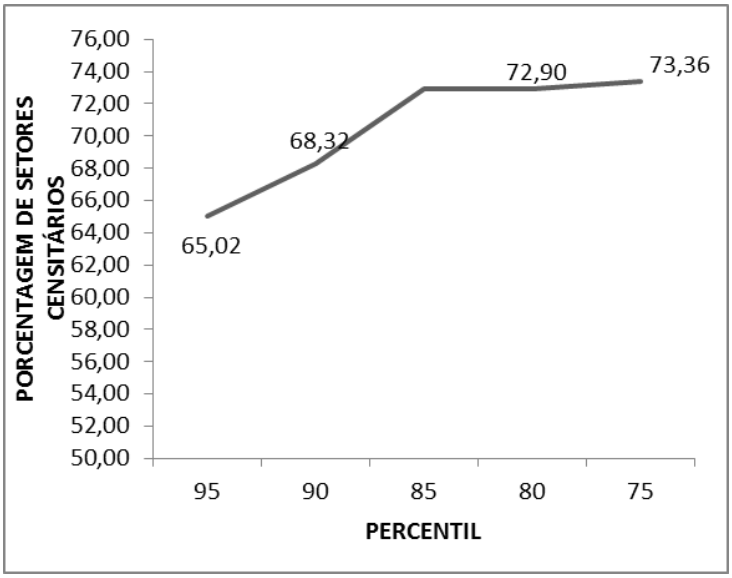


Figura 6.21 – Porcentagem de setores censitários com médias de viagens por domicílio do modelo 4 que estão dentro dos intervalos da amostra da O/D.

Observou-se que a retirada dos valores maiores ou iguais aos percentis, o número de setores censitários da população que possuíam médias de viagens dentro do intervalo de confiança da amostra aumentou em todos os percentis. A Tabela 6.22 apresenta os principais resultados das viagens estimadas pelo Modelo 4 após a etapa de validação (análise dos percentis).

A Tabela 6.23 apresenta um resumo das médias das viagens por domicílio da amostra e da população sintética agregada por setor censitário.

Tabela 6.22 – Resumo dos resultados obtidos.

Intervalos da Amostra O/D	**Viagens válidas**	**Taxa de acertos**
completa	118.564	55,60%
Percentil 95	134.361	63,00%
Percentil 90	141.956	66,60%
Percentil 85	153.414	71,90%
Percentil 80	153.414	71,90%
Percentil 75	157.981	74,00%

Tabela 6.23 – Análise das médias das viagens na retirada dos valores referentes aos percentis.

População Sintética	**Amostra O/D**		
Média = 3,36		Intervalos de confiança	
	Média	-	+
Completa	4,8	4,65	4,96
Percentil 95	4,23	4,11	4,35
Percentil 90	3,93	3,83	4,05
Percentil 85	3,51	3,41	3,61
Percentil 80	3,51	3,41	3,61
Percentil 75	2,9	2,81	2,99

Os resultados da modelagem RLM (Modelo 4) também foram satisfatórios, pois a validação das viagens por domicílio apresentada anteriormente (etapa 2) obteve taxa de acerto de mais de 50% e com a retirada dos valores maiores ou iguais a sete viagens por domicílio (percentis 95, 90, 85, 80 e 75) as taxas de acertos aumentaram em mais de 18%. Isto aconteceu, pois as médias de viagens por domicílio da população sintética (Modelo 4) apresentou quase em 100% dos setores média de viagem entre 3 a 4 viagens por domicílio. Diferentemente do Modelo 3, o Modelo 4 com a diminuição nas médias da amostra devido a retirada dos valores acima de sete foi relevante, pois mais de 70% das viagens estimadas pelo Modelo 4 são menores que 7.

6.5.3 Avaliação geral das viagens por domicílio da população sintética

De acordo com os resultados e as análises apresentadas na seção de validação dos resultados observam-se os seguintes pontos relevantes ao trabalho de pesquisa:

- O Modelo 3 (RNAs) apresentou a distribuição das viagens da população agregadas por categoria domiciliar segundo a quantidade de membros na família muito próxima a distribuição da amostra, ou seja, 72% das viagens por domicílio tanto da amostra quanto da população foram realizadas pelos domicílios com 3 a 5 moradores. Além disso, apresentaram médias totais de viagens por domicílio bem próximas, 4.8 e 4.3 e também uma forte relação entre os dados da amostra e da população (coeficiente de *Pearson* próximo de 1). Em relação as viagens consideradas válidas, o Modelo 3 apresentou na análise mais geral (comparação de médias dado um intervalo de confiança) que 70% das viagens por domicílio da população sintética foram consideradas válidas e na análise mais específica

(análise após a retirada dos valores maiores ou iguais aos percentis 95,90,85,80 e 75) a taxa de acertos (viagens válidas) aumentou para 76%.

- O Modelo 4 (RLM) apresentou a distribuição das viagens da população agregadas por categoria domiciliar segundo a quantidade de membros na família diferente da distribuição da amostra, ou seja, 72% das viagens por domicílio da amostra foram realizadas pelos domicílios com 3 a 5 moradores e da população foram realizadas pelos domicílios com 2 a 4 moradores. Além disso, apresentaram médias totais de viagens por domicílio bem diferentes, 4.8 e 3.4. Porém, a relação entre os dados da amostra e da população foi alta (coeficiente de *Pearson* = 0,914). Em relação as viagens consideradas válidas, o Modelo 4 apresentou na análise mais geral (comparação de médias dado um intervalo de confiança) que 60% das viagens por domicílio da população sintética foram consideradas válidas e na análise mais específica (análise após a retirada dos valores maiores ou iguais aos percentis 95,90,85,80 e 75) a taxa de acertos (viagens válidas) aumentou para 74%.

Um resumo dos resultados da etapa de validação das viagens por domicílio da população sintética encontra-se na Tabela 6.24.

Tabela 6.24 – Resumo das viagens por domicílio da população sintética.

Análises	**Amostra**	**População (RNA)**	**População (RLM)**
Distribuição das viagens agregadas por categoria	72% domicílios 3 a 5 moradores	72% domicílios 3 a 5 moradores	72% domicílios 2 a 4 moradores
Médias totais de viagens	4,8	4,3	3,4
Coeficiente de Pearson		0,968	0,914
Análise geral (viagens válidas)		70%	60%
Análise específica (viagens válidas)		76% (Percentil 90)	74% (Percentil 75)

Apesar de, a técnica de RLM ser a mais utilizada em estudos de transportes, para este trabalho cujo objetivo é modelar viagens por domicílio usando dados sintéticos, o Modelo 3 apresentou resultados melhores que o modelo tradicional (RLM), afirmando que a técnica de RNA é uma alternativa viável para modelar demanda por transportes.

Capítulo 7

Conclusões e Recomendações

Neste capítulo é realizado uma discussão geral sobre o desenvolvimento desta tese. O problema de pesquisa é retomado e respondido com base nos resultados do trabalho. Ao final, são apresentadas sugestões para pesquisas futuras que possam dar continuidade a esta investigação.

7.1 Conclusões

Esta proposta de tese de doutorado teve como propósito o estudo e a aplicação do uso de uma população sintética para prever viagens produzidas por domicílio, além de testar a adequabilidade da aplicação da técnica de Redes Neurais Artificiais na modelagem de transportes. Por meio do desenvolvimento, da aplicação e da análise dos resultados dos modelos propostos foi possível levantar subsídios para responder a pergunta desta pesquisa, a qual foi definida no primeiro capítulo. As duas hipóteses deste trabalho foram feitas de modo que a pergunta tivesse uma resposta afirmativa.

Objetivou-se a análise de dados agregados para geração de população sintética e de dados desagregados para a modelagem de transportes. Assim, a utilização da população sintética propicia a realização de estudos na área de demanda por transportes, pois os dados geralmente necessários são obtidos por pesquisas de tráfego, como por exemplo, a Pesquisa Origem e Destino (O/D). Porém, a realização deste tipo de pesquisa apresenta um custo oneroso e despende de muito tempo para a execução, por isso não é realizada com uma frequência desejada.

Dispondo de um banco de dados agregados e desagregados do Censo Demográfico do IBGE e de um banco de dados desagregados da Pesquisa O/D da cidade de São Carlos-SP, inicialmente

realizou-se a primeira etapa do método proposto: o tratamento e visualização dos dados. Nele foram eliminados possíveis empecilhos como setores censitários que não possuíam domicílios ou que não estavam contidos na área urbanizada da cidade e domicílios que não eram particulares. Neste momento, os dados já estavam aptos ao trabalho, bastando apenas a seleção da variável objeto de estudo, que estava presente somente no banco da Pesquisa O/D, a variável *viagens por domicílio*.

Na sequência, a primeira hipótese da pesquisa é indagada a possibilidade de obter uma população sintética utilizando apenas dados agregados. Por meio da aplicação do Método Monte Carlo, é possível aceitar a primeira hipótese desta proposta de tese, uma vez que as variáveis geradas foram verificadas através do teste estatístico de *Kolmogorov-Smirnov* e 70% das variáveis obtidas na população sintética foram consideradas aptas para o estudo. Dessa forma, foram utilizadas para prever viagens produzidas por domicílio somente as variáveis que passaram no teste estatístico.

Escolhidas as variáveis que tinham relação com o fenômeno de estudo, aplicaram-se a modelagem Redes Neurais Artificiais (RNAs) e a modelagem tradicional para a previsão da variável viagens produzidas por domicílio. Para a modelagem RNA, recorreu-se à rede Perceptron Multicamadas (MLP) e para o caso de modelagem tradicional, recorreu-se a regressão linear múltipla (RLM).

A modelagem RNA teve como primeiro passo a escolha da amostra de particionamento (amostra de treinamento e teste). Em seguida, para todas as proporções de amostras definidas foram testadas as possíveis arquiteturas e os tipos de treinamento.

Após a análise de sensibilidade, a amostra (60% treinamento e 40% teste) foi escolhida para toda a aplicação do método, pois foi a amostra que apresentou o menor erro quadrático médio (3,736) e o maior valor de correlação (0,566) dos dados na amostra de validação (teste).Essa amostra foi definida por uma arquitetura personalizada composta por duas camadas ocultas, função de ativação na camada oculta (tangente hiperbólica), função de ativação da camada de saída (identidade), treinamento do tipo lote e algoritmo de otimização gradiente conjugado.

O processo posterior, a calibração dos Modelos 1 e 3 (RNA), foi realizado para estimar viagens produzidas por domicílio através dos dados desagregados da Pesquisa O/D. Os resultados dos dois modelos apresentaram diferenças sutis e observou-se através de gráficos de dispersão entre valores observados e previstos uma boa relação entre os valores (Coeficiente de *Pearson* = 0,566 e 0,583, respectivamente).

A validação para o método não paramétrico (RNAs) se deu pela análise das medidas descritivas de 40% da amostra (amostra teste). Em geral, os resultados dos dois modelos (1 e 3) foram similares, ambos com mais de 50% de correlação entre os dados observados e previstos.

Em relação a modelagem tradicional, os resultados encontrados nos Modelos 2 e 4, através da técnica mais usual na previsão de geração de viagens por domicílio, a Regressão Linear

Múltipla, foram bem diferentes. Vale lembrar que, foi utilizado o modelo linear, pois é o modelo mais popular na área de demanda por transportes.

Neste estudo,a modelagem tradicional foi utilizada para comparar os resultados obtidos com a modelagem RNA, pois o objetivo não foi checar minuciosamente a qualidade dos modelos lineares,e sim,testar a adequabilidade das RNAs para estimar viagens produzidas por domicílio.

O Modelo 2 (com todas as variáveis) foi o modelo mais adequado para prever viagens por domicílio, pois além do R^2 ser quase 0,4, as variáveis selecionadas pelo Modelo 2 (*Stepwise*) foram todas significativas e os valores dos coeficientes e da estatística t foram coerentes com o esperado.

Pôde-se perceber que o Modelo 4 (variáveis que passaram no teste estatístico) em relação ao Modelo 2 diminui significativamente o poder de preditivo, aumentou o erro e diminuiu a correlação entre os dados observados e previstos.

Contudo, não foi possível utilizar todas as variáveis na calibração dos modelos, pois algumas variáveis não passaram no teste estatístico de *Kolmogorov-Smirnov* e, portanto, os Modelos 1 e 2 não foram utilizados para prever viagens produzidas por domicílio da população sintética. Entretanto, esses modelos foram utilizados apenas para verificar se apresentariam grandes diferenças com a retirada das variáveis que não passaram no teste estatístico de *Kolmogorov-Smirnov*.

Observou-se que o Modelo 1 (RNA) não apresentou diferenças significativas quando comparado ao Modelo 3 (RNA), que utilizou somente as variáveis que passaram no teste estatístico. De maneira análoga, o Modelo 2 (RLM) diminuiu o poder preditivo e aumentou o erro quando comparado ao Modelo 4 (RLM), que também utilizou somente variáveis que passaram no teste estatístico. Dessa forma, foram sugeridos os Modelos 3 e 4 para prever viagens produzidas por domicílio da população sintética.

A segunda hipótese da pesquisa está relacionada ao uso da população sintética obtida nesta tese para estimar número de viagens produzidas por domicílios. Para isto, foram utilizados os Modelos 3 e 4 para estimar as viagens por domicílio. Essa segunda hipótese é confirmada nesta tese no momento em que se aplicam as três etapas de validação dos resultados, pois se observa que a estimativa de viagens produzidas por domicílio obtida pelos Modelos 3 e 4 do método proposto obteve uma boa previsão, ou seja, mais de 70% das viagens produzidas por domicílio da população sintética foram consideradas válidas.

Quanto a execução deste trabalho, destaca-se que para desenvolver um modelo que aborde a produção de viagens de uma população sintética sob uma análise por domicílios, é necessário ter uma base de dados desagregada para permitir os processos de calibração e validação do modelo. Logo, observa-se que o fato de ter disponível uma Pesquisa O/D com elevado nível de desagregação de dados foi decisivo para viabilizar a execução da pesquisa desta tese. Apesar da particularidade dos modelos obtidos nesta tese serem referentes à aplicação na cidade de São

Carlos-SP, é possível que cidades com características semelhantes a de São Carlos-SP utilizem os mesmos modelos.

Além disso, é importante lembrar que, para utilizar o método proposto o planejador precisa ter em mãos um programa que tenha o módulo de redes neurais artificiais ou que saiba programar uma rede neural artificial.

Por fim, o trabalho aqui exposto apresentou um método com resultados considerados satisfatórios para atender a demanda por transportes, e evidencia a eficácia da aplicação da modelagem RNA no planejamento de transportes, especificamente na previsão de viagens produzidas por domicílio.

7.2 Sugestões para próximos trabalhos

Os estudos associados ao uso de uma população sintética para desagregar dados e utilizá-los em estudos de transportes tem muito a ser explorado e utilizado no Brasil, apesar de ser uma técnica que já vem sendo muito utilizada em vários países, como os Estados Unidos. Além disso, outro ponto fundamental deste trabalho foi o uso das Redes Neurais Artificiais como uma alternativa ao uso dos métodos tradicionais de previsão da demanda. É um método também que tem muito a ser explorado com o enfoque de melhoramento na estrutura da rede para o seu uso na modelagem da demanda por transportes. Assim, de forma a dar continuidade a este trabalho são feitas algumas sugestões a seguir.

Neste trabalho, gerou-se a população sintética por meio de dados agregados utilizando o Método Monte Carlo, no entanto recomenda-se que a utilização de outros métodos para a geração da mesma, tais como o método de otimização combinatória (CO-*Combinatorial Optimization*) e o método conhecido como ajuste proporcional iterativo (IPF-*Iterative Proportional Fitting*), fazendo-se em seguida uma comparação entre os dois de forma a verificar a precisão de cada método.

Outro enfoque pertinente para aprimorar o método de geração da população sintética é testar o uso das Redes Neurais Artificiais (RNA) para gerar dados desagregados por meio de dados agregados.

Ainda em relação a população sintética proposta neste trabalho, recomenda-se testar a geração dessa população, que foi por meio de dados agregados do Censo do IBGE-2010, utilizando a base desagregada do Censo do IBGE-2010 disponível no site (base de microdados) e em seguida, compará-la com a população sintética gerada neste trabalho.

Em relação ao uso da RNA na modelagem da demanda por transportes, recomenda-se a utilização de outros *softwares* ou até mesmo o desenvolvimento de uma rede personalizada, utilizando-se, por exemplo o MATLAB. O software utilizado neste trabalho foi o SPSS 22.0 e as opções que estavam disponíveis eram do *default* do programa, o que puderam influenciar no desempenho dos modelos, pois existem outras funções de ativação, bem como outras variações

do algoritmo de aprendizado *backpropagation* que podem ou não melhorar o desempenho da RNA.

Quanto ao modelo proposto, no qual foi desenvolvida apenas a primeira etapa do modelo Quatro Etapas, sugere-se que com a produção de viagens obtidas por este trabalho sejam desenvolvidas as outras três etapas do modelo, ou seja, a distribuição de viagens, a escolha modal e a alocação do tráfego.

Destaca-se também a possibilidade de aplicar a estrutura do modelo proposto em outras áreas urbanas que não possuem nenhuma Pesquisa Origem e Destino aplicar a estrutura do modelo proposto para auxiliar no planejamento de demanda por transportes, alterando apenas os dados de entrada para ajustar o modelo à área de estudo. Porém, vale ressaltar que esta estrutura somente se aplica a cidades que apresentam as características semelhantes as da cidade de São Carlos-SP.

Por fim, ainda em relação a aplicabilidade do método proposto, e de forma a aprimorar e dar continuidade ao trabalho, recomenda-se que outras áreas de estudo que possuem pelo menos uma Pesquisa O/D utilizem o método para que o mesmo se torne mais confiável.

Referências

ADIGA, A. et al. *Generating a synthetic population of the United States*. Virginia Tech, 2015.

AL-DEEK, H. Which method is better for developing freight planning models at seaports—neural networks or multiple regression? *Transportation Research Record: Journal of the Transportation Research Board*, 2001. Transportation Research Board of the National Academies, n. 1763, p. 90–97, 2001.

AMAVI, A. A. et al. Advanced trip generation/attraction models. *Procedia-Social and Behavioral Sciences*, 2014. Elsevier, v. 160, p. 430–439, 2014.

ARENTZE, T.; TIMMERMANS, H.; HOFMAN, F. Creating synthetic household populations: problems and approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2008. Transportation Research Board of the National Academies, 2008.

BADOE, D. A. Forecasting travel demand with alternatively structured models of trip frequency. *Transportation Planning and Technology*, 2007. Taylor & Francis, v. 30, n. 5, p. 455–475, 2007.

BALCI, O. Verification, validation, and certification of modeling and simulation applications. In: WINTER SIMULATION CONFERENCE. *Proceedings of the 35th conference on Winter simulation: driving innovation*. New Jersey, 2003. p. 150–158.

BANKS, J. et al. *Discrete-event system simulation*. Saddle River, NJ: Prentice Hall, 2005.

BARBETTA, P. A. *Estatística aplicada às ciências sociais*. Florianópolis, SC: Ed. UFSC, 2012.

BARROS, E. A. C.; MAZUCHELI, J. *Aplicações de Simulação Monte Carlo e Bootstrap*. 2005. XIV Encontro Anual de Iniciação Científica.

BARTHELEMY, J.; TOINT, P. L. Synthetic population generation without a sample. *Transportation Science*, 2013. Informs, v. 47, n. 2, p. 266–279, 2013.

BATTAILE, C. C. The kinetic monte carlo method: Foundation, implementation, and application. *Computer Methods in Applied Mechanics and Engineering*, 2008. Elsevier, v. 197, n. 41, p. 3386–3398, 2008.

BECKMAN, R. J.; BAGGERLY, K. A.; MCKAY, M. D. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 1996. Elsevier, v. 30, n. 6, p. 415–429, 1996.

BHAT, C. et al. Comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. Transportation Research Board of the National Academies, n. 1894, p. 57–66, 2004.

BILT, K. anne Van de. *Desenvolvimento e validação de um procedimento de projeção desagregada da população associada a um modelo de geração de viagens baseado em análise de segmentação*. Tese (Doutorado) — Universidade de São Paulo, Escola Politécnica da Universidade de São Paulo, 2002.

BIRKIN, M.; CLARKE, M. Synthesis—a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and planning A*, 1988. SAGE Publications, v. 20, n. 12, p. 1645–1671, 1988.

BISHOP, Y. M.; FIENBERG, S. E.; HOLLAND, P. W. *Discrete multivariate analysis: theory and practice*. New York, NY: Springer Science & Business Media, 2007.

BOCANEGRA, C. W. R. *Procedimentos para tornar mais efetivo o uso das redes neurais artificiais em planejamento de transportes*. Dissertação (Mestrado) — Universidade de São Paulo, Escola de Engenharia de São Carlos, 2002. Disponível em: <http://www.teses.usp.br/>. Acesso em: 10.3.2015.

BOSURGI, G.; TRIFIRÒ, F. A model based on artificial neural networks and genetic algorithms for pavement maintenance management. *International Journal of Pavement Engineering*, 2005. Taylor & Francis, v. 6, n. 3, p. 201–209, 2005.

BOTTER, R. *Introdução às técnicas de simulação e ao programa Arena*. 2001. Apostila. Departamento de Engenharia Naval e Oceânica da Universidade de São Paulo.

BOWMAN, J. L. *A comparison of population synthesizers used in microsimulation models of activity and travel demand*. 2004. Unpublished working paper. Disponível em: <http:-//jbowman. net/papers/2004. Bowman. Comparison_ of_PopSyns. pdf>. Acesso em: 20.7.2015.

BRUNDELL-FREIJ, K. Sampling, specification and estimation as sources of inaccuracy in complex transport models-some examples analysed by monte carlo simulation and bootstrap. In: COLLEGE, H. H. (Ed.). *Proceedings of Seminar of the European Transport Conference 2000*. Cambridge, England: [s.n.], 2000. P441.

CAI, M.; YIN, Y.; XIE, M. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transportation Research Part D: Transport and Environment*, 2009. Elsevier, v. 14, n. 1, p. 32–41, 2009.

CARNEIRO, L. G. P. L. *Desenvolvimento de uma Metodologia para Previsão de Demanda de Passageiros para o Transporte Rodoviário Interestadual por ônibus*. Dissertação (Mestrado) — Faculdade de Tecnologia, Universidade de Brasília, 2003. Disponível em: <http://www-.repositorio.unb.br/>. Acesso em: 20.3.2015.

CARVALHO, A. et al. *Inteligência Artificial–uma abordagem de aprendizado de máquina*. 1. ed. Rio de Janeiro: LTC, 2011.

CHANG, J. S. et al. Comparative analysis of trip generation models: results using home-based work trips in the seoul metropolitan area. *Transportation Letters*, 2014. Taylor & Francis, v. 6, n. 2, p. 78–88, 2014.

CHANG, L.-Y. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science*, 2005. Elsevier, v. 43, n. 8, p. 541–557, 2005.

CHWIF, L.; MEDINA, A. *Modelagem e Simulação de Eventos Discretos: Teoria e Aplicações*. 4. ed. São Paulo, Brasil: Elsevier, 2014.

COLOMBARONI, C.; FUSCO, G. Artificial neural network models for car following: experimental analysis and calibration issues. *Journal of Intelligent Transportation Systems*, 2014. Taylor & Francis, v. 18, n. 1, p. 5–16, 2014.

CORRAR, L. J.; THEÓPHILO, C. R. *Pesquisa operacional para decisão em contabilidade e administração: contabilometria*. 2. ed. São Paulo: Atlas, 2004.

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*, 1940. JSTOR, v. 11, n. 4, p. 427–444, 1940.

DONNELLY, R. et al. Advanced practices in travel forecasting. *Transportation Research Board*, 2010. NCHRP Synthesis 406, p. 90, 2010.

DOUGHERTY, M. A review of neural networks applied to transport. *Transportation Research Part C: Emerging Technologies*, 1995. Elsevier, v. 3, n. 4, p. 247–260, 1995.

DUGUAY, G.; JUNG, W.; MCFADDEN, D. *SYNSAM: a methodology for synthesizing household transportation survey data*. Berkeley, C.A.: Urban Travel Demand Forecasting Project, Institute of Transportation Studies, 1976. Working paper.

ESCUDERO, L. F. *La simulatición en la empresa*. Bilbao: Duesto, 1973.

FIELD, A. *Discovering statistics using SPSS*. 3. ed. Washington, DC: SAGE publications, 2009.

GILLESPIE, D. T. Monte carlo simulation of random walks with residence time dependent transition probability rates. *Journal of Computational Physics*, 1978. Elsevier, v. 28, n. 3, p. 395–407, 1978.

GONÇALVES, D. N. S.; SILVA, M. A. d.; D'AGOSTO, M. d. A. Procedimento para uso de redes neurais artificiais no planejamento estratégico de fluxo de carga no brasil. *Journal of Transport Literature*, 2015. SciELO Brasil, v. 9, n. 1, p. 45–49, 2015.

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

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica-5*. Porto Alegre: AMGH Editora, 2011.

GUO, J.; BHAT, C. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2008. Transportation Research Board of the National Academies, 2008.

HAFEZI, M. H.; HABIB, M. A. Synthesizing population for microsimulation-based integrated transport models using atlantic canada micro-data. *Procedia Computer Science*, 2014. Elsevier, v. 37, p. 410–415, 2014.

HAIR, J. F. et al. *Análise multivariada de dados*. Porto Alegre: Bookman, 2009.

HAMMERSLEY, J.; HANDSCOMB, D. *Monte Carlo Methods, Methuen's Monographs on Applied Probability*. New York: Wiley, 1964.

HAYKIN, S. *Neural networks - a comprehensive foundation*. Ontario, Canada: Pretince Hall, 1999.

HEBB, D. O. *The organization of behavior: A neuropsychological approach*. New York, NY: John Wiley & Sons, 1949.

HERTZ, J.; KROGH, A.; PALMER, R. G. *Introduction to the theory of neural computation*. Redwood City, CA: Basic Books, 1991.

HINTON, G. E. How neural networks learn from experience. *Scientific American*, 1992. v. 267, n. 3, p. 145–151, 1992.

HOPFIELD, J. J. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 1984. National Acad Sciences, v. 81, n. 10, p. 3088–3092, 1984.

HUANG, Z.; WILLIAMSON, P. A. *Comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata*. 2002. Working paper. Department of Geography, University of Liverpool.

HUYNH, N. et al. Generating a synthetic population in support of agent-based modeling of transportation in sydney. In: 20TH INTERNATIONAL CONGRESS ON MODELLING AND SIMULATION (MODSIM). *The Modelling and Simulation Society of Australia and New Zealand*. Australia, 2013. p. 1357–1363.

IBGE. *Censo Demográfico do IBGE*. 2010. Disponível em: <http://www.ibge.gov.br/home-/estatistica/ populacao/censo2010>. Acesso em: 21.6.2014.

IBGE. *Instituto Brasileiro de Geografia e Estatística*. 2016. Disponível em: <http://www.ibge-.gov.br>. Acesso em: 21.3.2016.

JEONG, R.; RILETT, L. Prediction model of bus arrival time for real-time applications. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. Transportation Research Board of the National Academies, n. 1927, p. 195–204, 2005.

KARLAFTIS, M.; VLAHOGIANNI, E. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 2011. Elsevier, v. 19, n. 3, p. 387–399, 2011.

KIM, J. et al. Scenario-based approach to analysis of travel time reliability with traffic simulation models. *Transportation Research Record: Journal of the Transportation Research Board*, 2013. Transportation Research Board of the National Academies, n. 2391, p. 56–68, 2013.

KIRKPATRICK, S. et al. Optimization by simmulated annealing. *Science*, 1983. Washington, v. 220, n. 4598, p. 671–680, 1983.

KITAMURA, R. et al. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 2000. Springer, v. 27, n. 1, p. 25–51, 2000.

LIN, H.-E.; ZITO, R.; TAYLOR, M. A review of travel-time prediction in transport and logistics. *The journal of the Eastern Asia Society for transportation studies*, 2005. Asia, v. 5, p. 1433–1448, 2005.

LIPPMANN, R. P. An introduction to computing with neural nets. *ASSP Magazine, IEEE*, 1987. IEEE, v. 4, n. 2, p. 4–22, 1987.

LONGHI, S. et al. Neural network modeling as a tool for forecasting regional employment patterns. *International Regional Science Review*, 2005. Sage Publications, v. 28, n. 3, p. 330–346, 2005.

MA, L. *Generating disaggregate population characteristics for input to travel-demand models*. Tese (Doutorado) — University of Florida, 2011.

MA, L.; SRINIVASAN, S. Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*, 2015. Wiley Online Library, v. 30, n. 2, p. 135–150, 2015.

MARK, C. D.; SADEK, A. W.; RIZZO, D. Predicting experienced travel time with neural networks: a paramics simulation study. In: IEEE (Ed.). *Intelligent Transportation Systems Conference*. Washington, D.C., 2004. p. 906–911.

MENNER, W. Introduction to modeling and simulation. *Johns Hopkins APL Technical Digest*, 1995. JOHNS HOPKINS UNIV APPLIED PHYSICS LABORATORY ATTN: MANAGING EDITOR JOHN HOPKINS RD, BLDG 1-E254, LAUREL, MD 20723-6099, v. 16, n. 1, p. 6–17, 1995.

MILLER, E. J. Microsimulation and activity-based forecasting. In: *Activity-Based Travel Forecasting Conference*. Texas, USA: [s.n.], 1997.

MOECKEL, R.; SPIEKERMANN, K.; WEGENER, M. Creating a synthetic population. In: *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*. Sendai, Japan: [s.n.], 2003. p. 1–18.

MOORE, J. H.; WEATHERFORD, L. R. *Tomada de decisão em administração com planilhas*. Porto Alegre: Bookman, 2005.

MOORE, P. C. Preliminary development of a trip generation manual for texas. *Compendium of Student Papers: 2013 Undergraduate Transportation Scholars Program*, 2013. p. 101, 2013.

MORLOK, E. K. *Introduction to transportation engineering and planning*. New York, USA: McGraw-Hill, 1978.

MOZOLIN, M.; THILL, J.-C.; USERY, E. L. Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological*, 2000. Elsevier, v. 34, n. 1, p. 53–73, 2000.

MÜLLER, K.; AXHAUSEN, K. W. Hierarchical ipf: Generating a synthetic population for switzerland. In: *51 st Congress of the European Regional Science Association*. Barcelona: [s.n.], 2011.

MÜLLER, K.; AXHAUSEN, K. W. Population synthesis for microsimulation: State of the art. In: *90th Annual Meeting of Transportation Research Board*. Washington, DC: [s.n.], 2011.

MÜLLER, K.; AXHAUSEN, K. W. Preparing the swiss public-use sample for generating a synthetic population of switzerland. In: *12th Swiss Transport Research Conference*. Monte Verità, Ascona: [s.n.], 2012.

MÜNNICH, R. et al. Monte carlo simulation study of european surveys. *DACSEIS Deliverables D*, 2003. v. 3, 2003.

NAGAI, E. Y. *Identificação automática de modelos fuzzy inferenciais.* Tese (Doutorado) — Universidade Tecnológica Federal do Paraná, Curitiba, 2006.

NAMAZI-RAD, M. R.; MOKHTARIAN, P.; PEREZ, P. Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. *PloS one*, 2014. Public Library of Science, v. 9, n. 4, p. e94761, 2014.

NASCIMENTO, A.; ZUCCHI, A. *Modelos de simulação*. 1997. Monografia. Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

ORTÚZAR, D. J.; WILLUMSEN, L. G. *Modelling transport*. London: John Wiley & Sons, 2011.

PNUD. *Atlas do Desenvolvimento Humano no Brasil*. 2016. Disponível em: <http://http:/-/www.pnud.org.br/atlas>. Acesso em: 2.4.2016.

RAIA, J. A. A. *Acessibilidade e mobilidade na estimativa de um índice de potencial de viagens utilizando redes neurais artificiais e sistemas de informações geográficas.* Tese (Doutorado) — Universidade de São Paulo, Escola de Engenharia de São Carlos, 2000.

RASOULI, M.; NIKRAZ, H. Trip distribution modelling using neural network. In: *Transport Research Forum 2013*. Brisbane, Australia: [s.n.], 2013.

RASOULI, S.; TIMMERMANS, H. Uncertainty in travel demand forecasting models: literature review and research agenda. *Transportation letters*, 2012. J. Ross Publishing, Inc., v. 4, n. 1, p. 55–73, 2012.

RASSAFI, A. A.; REZAEI, R.; HAJIZAMANI, M. Predicting urban trip generation using a fuzzy expert system. *Iranian Journal of Fuzzy Systems*, 2012. v. 9, n. 3, p. 127–146, 2012.

REIS, S. G. d.; MARTINS, E. Planejamento do balanço bancário: desenvolvimento de um modelo matemático de otimização do retorno econômico ajustado ao risco. *Revista Contabilidade & Finanças*, 2001. SciELO Brasil, v. 12, n. 26, p. 58–80, 2001.

RENDER, B.; STAIR, R. M. J.; HANNA, M. E. *Quantitative Analysis for Management*. India: Prentice Hall, 2000.

RICO, M. T. S.; RODENAS, R. G.; ARANDA, J. L. E. A monte carlo approach to simulate the stochastic demand in a continuous dynamic traffic network loading problem. *Intelligent Transportation Systems, IEEE Transactions on*, 2014. IEEE, v. 15, n. 3, p. 1362–1373, 2014.

ROCHA, S. S. et al. Uso de redes neurais para previsão de produção de viagens: Uma análise agregada. In: *Anais do XXIX Congresso nacional de Pesquisa e Ensino em Transporte*. Ouro Preto, Minas Gerais: Associação Nacional de Pesquisa e Ensino em Transportes (ANPET). 9 a 13 de novembro de 2015, 2015. p. 1995–2006.

ROORDA, M. J. et al. Trip generation of vulnerable populations in three canadian cities: a spatial ordered probit approach. *Transportation*, 2010. Springer, v. 37, n. 3, p. 525–548, 2010.

ROSA, J. L. G. *Fundamentos da inteligência artificial*. Rio de Janeiro: LTC, 2011.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 1958. American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUMELHART, D. E. et al. Parallel distributed processing: Explorations in the microstructure of cognition. *Cambridge, MA*, 1986. v. 1-2, 1986.

RYAN, J.; MAOH, H.; KANAROGLOU, P. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, 2009. Wiley Online Library, v. 41, n. 2, p. 181–203, 2009.

SCHMÖCKER, J.-D. et al. Estimating trip generation of elderly and disabled people: analysis of london data. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. Transportation Research Board of the National Academies, n. 1924, p. 9–18, 2005.

SILVA, A. N. R. d. et al. *SIG: uma plataforma para introdução de técnicas emergentes no planejamento urbano, regional e de transportes: uma ferramenta 3D para análise ambiental urbana, avaliação multi critério, redes neurais artificiais*. São Carlos: EdUFSCAR, 2008.

SILVA, A. N. Rodrigues da. *Pesquisa Origem e Destino da cidade de São Carlos*. 2008. Relatório. Universidade de São Paulo, Escola de Engenharia de São Carlos.

SILVA, C. A. U. *Um método para estimar observáveis GPS usando redes neurais artificiais*. Tese (Doutorado) — Universidade de São Paulo, Escola de Engenharia de São Carlos, 2003.

SOMMER, M. et al. Cognitive and personality determinants of fitness to drive. *Transportation research part F: traffic psychology and behaviour*, 2008. Elsevier, v. 11, n. 5, p. 362–375, 2008.

SOUZA, C. D. R. d.; D'AGOSTO, M. d. A. Modelo de quatro etapas aplicado ao planejamento de transporte de carga. *Journal of Transport Literature*, 2012. v. 7, n. 2, p. 207–234, 2012.

TEODOROVIĆ, D. et al. Dynamic programming—neural network real-time traffic adaptive signal control algorithm. *Annals of Operations Research*, 2006. Springer, v. 143, n. 1, p. 123–131, 2006.

TONG, H.; HUNG, W. Neural network modeling of vehicle discharge headway at signalized intersection: model descriptions and results. *Transportation Research Part A: Policy and Practice*, 2002. Elsevier, v. 36, n. 1, p. 17–40, 2002.

VOAS, D.; WILLIAMSON, P. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, 2000. Wiley Online Library, v. 6, n. 5, p. 349–366, 2000.

VOAS, D.; WILLIAMSON, P. Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, 2001. Taylor & Francis, v. 5, n. 2, p. 177–200, 2001.

WASSERMAN, P. D. *Neural computing*. New York: Van Nostrand Reinhold, 1989.

WATANATADA, T.; BEN-AKIVA, M. Forecasting urban travel demand for quick policy analysis with disaggregate choice models: A monte carlo simulation approach. *Transportation Research Part A: General*, 1979. Elsevier, v. 13, n. 4, p. 241–248, 1979.

WILLIAMS, H.; ORTÚZAR, J. d. D. Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research Part B: Methodological*, 1982. Elsevier, v. 16, n. 3, p. 167–219, 1982.

WILLIAMSON, P.; BIRKIN, M.; REES, P. H. The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, 1998. SAGE Publications, v. 30, n. 5, p. 785–816, 1998.

ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 1998. Elsevier, v. 14, n. 1, p. 35–62, 1998.

ZHANG, T.; XIE, C.; WALLER, S. Integrated equilibrium travel demand model with nested logit structure: Fixed-point formulation and stochastic analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2011. Transportation Research Board of the National Academies, n. 2254, p. 79–96, 2011.

Apêndices

APÊNDICE

A

Sumário das tabelas disponíveis dos dados originais do censo demográfico do IBGE 2010

Tabela A.1 – Planilha_básico.

Situação	Área urbanizada	1
Variáveis	**Descrição**	**Total**
V001	Nº de domicílios particulares permanentes	68833
V002	Nº de moradores em domicílios particulares permanentes	212263
V003	Média do nº de moradores	3,014
V004	Variância do nº de moradores	2,08
V005	Renda média mensal dos chefes com ou sem rendimento	1854,77

Tabela A.2 – Domicílios_02.

Variáveis	**Descrição**	**Total**
V002	Nº moradores em domicílios particulares permanentes	212263
V046	Nº homens em domicílios particulares permanentes	103803
V090	Nº mulheres em domicílios particulares permanentes	108460

Tabela A.3 – Domicílios_01.

Variáveis	Descrição	Total
V002	Domicílios particulares permanentes	68833
V050	Domicílios particulares permanentes com 1 morador	9053
V051	Domicílios particulares permanentes com 2 moradores	17186
V052	Domicílios particulares permanentes com 3 moradores	18094
V053	Domicílios particulares permanentes com 4 moradores	14636
V054	Domicílios particulares permanentes com 5 moradores	6120
V055	Domicílios particulares permanentes com 6 moradores	2218
V056	Domicílios particulares permanentes com 7 moradores	849
V057	Domicílios particulares permanentes com 8 moradores	339
V058	Domicílios particulares permanentes com 9 moradores	161
V059	Domicílios particulares permanentes com 10 ou mais moradores	177
V060	Domicílios particulares permanentes só com mulheres	8085
V061	Domicílios particulares permanentes só com homens	5529
V062	Domicílios particulares permanentes com chefe de família homem (responsável) e mais 1 morador	10348
V063	Domicílios particulares permanentes com chefe de família homem (responsável) e mais 2 moradores	12274
V064	Domicílios particulares permanentes com chefe de família homem (responsável) e mais 3 moradores	10640
V065	Domicílios particulares permanentes com chefe de família homem (responsável) e mais 4 moradores	4223
V066	Domicílios particulares permanentes com chefe de família homem (responsável) e mais 5 moradores	1466
V067	Domicílios particulares permanentes com chefe de família homem (responsável) e mais 6 ou mais moradores	916
V068	Domicílios particulares permanentes com chefe de família homem (responsável) e sem outro morador	4210
V081	Domicílios particulares permanentes com mulher responsável e mais 1 morador	6838
V082	Domicílios particulares permanentes com mulher responsável e mais 2 moradores	5820
V083	Domicílios particulares permanentes com mulher responsável e mais 3 moradores	3996
V084	Domicílios particulares permanentes com mulher responsável e mais 4 moradores	1897
V085	Domicílios particulares permanentes com mulher responsável e mais 5 moradores	752
V086	Domicílios particulares permanentes com mulher responsável e mais 6 ou mais moradores	610
V087	Domicílios particulares permanentes com mulher responsável e sem outro morador	4843

Tabela A.4 – Responsável_Mulher.

Variáveis	Descrição	Total
V001	Pessoas responsáveis, do sexo feminino	24779
V002-V011	Pessoas responsáveis com 10 -19 anos de idade, do sexo feminino	256
V012-V021	Pessoas responsáveis com 20 -29 anos de idade, do sexo feminino	2942
V022-V031	Pessoas responsáveis com 30 -39 anos de idade, do sexo feminino	4734
V032-V041	Pessoas responsáveis com 40 -49 anos de idade, do sexo feminino	5190
V042-V051	Pessoas responsáveis com 50 -59 anos de idade, do sexo feminino	4660
V052-V061	Pessoas responsáveis com 60 -69 anos de idade, do sexo feminino	3391
V062-V092	Pessoas responsáveis com 70 anos ou mais de idade, do sexo feminino	3606
V093	Pessoas alfabetizadas responsáveis, do sexo feminino	23347
V094-V095	Pessoas alfabetizadas responsáveis com 10 a 19 anos de idade, do sexo feminino	256
V096-V097	Pessoas alfabetizadas responsáveis com 20 a 29 anos de idade, do sexo feminino	2931
V098-V099	Pessoas alfabetizadas responsáveis com 30 a 39 anos de idade, do sexo feminino	4665
V100-V101	Pessoas alfabetizadas responsáveis com 40 a 49 anos de idade, do sexo feminino	5043
V102-V103	Pessoas alfabetizadas responsáveis com 50 a 59 anos de idade, do sexo feminino	4435
V104-V105	Pessoas alfabetizadas responsáveis com 60 a 69 anos de idade, do sexo feminino	3052
V106-V108	Pessoas alfabetizadas responsáveis com 70 anos ou mais de idade, do sexo feminino	2965

Tabela A.5 – Responsável_total_homem.

Variáveis	Descrição	Total
V001	Responsáveis totais	68922
V002-V011	Responsáveis totais 10 -19 anos	543
V012-V021	Responsáveis totais 20 -29 anos	8306
V022-V031	Responsáveis totais 30 -39 anos	14824
V032-V041	Responsáveis totais 40 -49 anos	15585
V042-V051	Responsáveis totais 50 -59 anos	13453
V052-V061	Responsáveis totais 60 -69 anos	8743
V062-V092	Responsáveis totais 70 anos ou mais	7468
V093	Alfabetizadas - totais	66292
V094-V095	Responsáveis totais_alfabetizadas 10-19 anos	540
V096-V097	Responsáveis totais_alfabetizadas 20-29 anos	8243
V098-V099	Responsáveis totais_alfabetizadas 30-39 anos	14645
V100-V101	Responsáveis totais_alfabetizadas 40-49 anos	15233
V102-V103	Responsáveis totais_alfabetizadas 50-59 anos	13007
V104-V105	Responsáveis totais_alfabetizadas 60-69 anos	8151
V106-V108	Responsáveis totais_alfabetizadas 70 anos ou mais	6473
V109	Pessoas responsáveis, do sexo masculino	44143
V110-V119	Pessoas responsáveis com 10 -19 anos de idade	287
V120-V129	Pessoas responsáveis com 20 -29 anos de idade	5364
V130-V139	Pessoas responsáveis com 30 -39 anos de idade	10090
V140-V149	Pessoas responsáveis com 40 -49 anos de idade	10395
V150-V159	Pessoas responsáveis com 50 -59 anos de idade	8793
V160-V169	Pessoas responsáveis com 60 -69 anos de idade	5352
V170-V200	Pessoas responsáveis com 70 anos ou mais de idade	3862
V201	Pessoas alfabetizadas responsáveis, do sexo masculino	42945
V202-V203	Pessoas alfabetizadas responsáveis com 10 a 19 anos de idade	284
V204-V205	Pessoas alfabetizadas responsáveis com 20 a 29 anos de idade	5312
V206-V207	Pessoas alfabetizadas responsáveis com 30 a 39 anos de idade	9980
V208-V209	Pessoas alfabetizadas responsáveis com 40 a 49 anos de idade	10190
V210-V211	Pessoas alfabetizadas responsáveis com 50 a 59 anos de idade	8572
V212-V213	Pessoas alfabetizadas responsáveis com 60 a 69 anos de idade	5099
V214-V216	Pessoas alfabetizadas responsáveis com 70 anos ou mais de idade	3508

Tabela A.6 – Grau parentesco_Cônjuges.

Variáveis	Descrição	Total
V001-V010	Cônjuges_tot 10 -19 anos	475
V011-V020	20-29 anos	6940
V021-V030	30-39 anos	11851
V031-V040	40 - 49 anos	10701
V041-V050	50-59 anos	8475
V051-V060	60-69 anos	4394
V061-V071	70 anos ou mais	2521
V072-V081	Cônjuges_HOMEM 10 -19 anos	25
V082-V091	20-29 anos	1255
V092-V101	30-39 anos	2536
V102-V111	40 - 49 anos	2158
V112-V121	50-59 anos	1626
V122-V131	60-69 anos	791
V132-V142	70 anos ou mais	626
V143-V152	Cônjuges_MULHER 10 -19 anos	450
V153-V162	20-29 anos	5685
V163-V172	30-39 anos	9315
V173-V182	40 - 49 anos	8543
V183-V192	50-59 anos	6849
V193-V202	60-69 anos	3603
V203-V213	70 anos ou mais	1895

Tabela A.7 – Grau parentesco_Filhos.

Variáveis	Descrição	Total
V001-V010	Filhos 0-9 anos	17374
V011-V020	10-19 anos	17725
V021-V030	20-29 anos	11948
V031-V040	30-39 anos	3196
V041-V050	40-49 anos	905
V051	50 anos ou mais	218
V103-V 112	ENTEADOS 0-9 anos	685
V113-V122	10-19 anos	1631
V123-V132	20-29 anos	617
V133-V142	30-39 anos	112
V143-V152	40-49 anos	26
V153	50 anos ou mais	17

Tabela A.8 – Grau parentesco_Outros.

Variáveis	Descrição	Total
V001-V002	GENRO 10-19 ANOS	215
V003-V004	20-29 anos	1082
V005-V006	30-39 anos	628
V007-V008	40 - 49 anos	237
V009-V010	50-59 anos	87
V011-V012	60-69 anos	21
V013-V015	70 anos ou mais	13
V046-V047	PAI, PADASTRO, MÃE, MADASTRA 20-29 ANOS	34
V048-V049	30-39 anos	197
V050-V051	40 - 49 anos	469
V052-V053	50-59 anos	688
V054-V055	60-69 anos	633
V056-V058	70 anos ou mais	1644
V085-V086	SOGROS 20-29 ANOS	5
V087-V088	30-39 anos	8
V089-V090	40 - 49 anos	52
V091-V092	50-59 anos	114
V093-V094	60-69 anos	165
V095-V097	70 anos ou mais	664
V124-V125	NETOS 0-9 ANOS	3885
V126-V127	10-19 ANOS	2596
V128-V129	20-29 anos	912
V130-V131	30-39 anos	129
V132-V133	40 - 49 anos	13
V134	50 ANOS OU MAIS	5
V157-V158	BISNETOS 0-9 ANOS	128
V159-V160	10-19 ANOS	36
V161-V162	20-29 anos	4
V163-V164	30-39 anos	0
V165-V166	40 - 49 anos	0
V167	50 ANOS OU MAIS	0
V190-V191	IRMÃOS 0-9 ANOS	90
V192-V193	10-19 ANOS	487
V194-V195	20-29 anos	924
V196-V197	30-39 anos	530
V198-V199	40 - 49 anos	468
V200-V201	50-59 anos	421
V202-V203	60-69 anos	246
V204-V206	70 anos ou mais	205

Tabela A.9 – Pessoa_13_Idade total.

Variáveis	Descrição	Total
V002	Pessoas residentes em domicílios particulares permanentes	212263
V003	Responsáveis pelos domicílios particulares	68922
V004	Cônjuges ou companheiros	45357
V005	Filhos(as) do responsável e do cônjuge em domicílios particulares	51366
V007	Enteados(as) em domicílios particulares	3088
V008	Genros ou noras em domicílios particulares	2283
V009	Pais, mães, padrastos ou madrastas em domicílios particulares	3665
V010	Sogros (as) em domicílios particulares	1008
V011	Netos(as) em domicílios particulares	7540
V012	Bisnetos(as) em domicílios particulares	168
V013	Irmãos ou irmãs em domicílios particulares	3371
V014	Avôs ou avós em domicílios particulares	131
V015	Outros parentes em domicílios particulares	3208
V022-V043	Pessoas 0-9 anos	25742
V044-V053	10-19 anos	31703
V054-V063	20-29 anos	38456
V064-V073	30 - 39 anos	34786
V074-V083	40-49 anos	30222
V084-V093	50-59 anos	24307
V094-V103	60-69 anos	14547
V104-V134	70 anos ou mais	13043

Tabela A.10 – Pessoa_11_Idade homens.

Variáveis	Descrição	Total
V002	Homens residentes em domicílios particulares permanentes	103803
V003	Responsáveis pelos domicílios particulares	44143
V004	Cônjuges ou companheiros	9017
V005	Filhos do responsável e do cônjuge em domicílios particulares	27016
V006	Filhos somente do responsável em domicílios particulares	10769
V007	Enteados em domicílios particulares	1610
V008	Genros ou noras em domicílios particulares	1162
V009	Pais, mães, padrastos ou madrastas em domicílios particulares	998
V010	Sogros em domicílios particulares	217
V011	Netos em domicílios particulares	4028
V012	Bisnetos em domicílios particulares	89
V013	Irmãos ou irmãs em domicílios particulares	1805
V014	Avôs ou avós em domicílios particulares	27
V015	Outros parentes em domicílios particulares, do sexo masculino	1713
V022-V043	HOMENS 0-9 anos	13075
V044-V053	10-19 anos	16237
V054-V063	20-29 anos	19536
V064-V073	30 - 39 anos	17236
V074-V083	40-49 anos	14659
V084-V093	50-59 anos	11502
V094-V103	60-69 anos	6662
V104-V134	70 anos ou mais	5199

Tabela A.11 – Pessoa_12_Idade mulheres.

Variáveis	Descrição	Total
V002	Mulheres residentes em domicílios particulares permanentes	108460
V003	Responsáveis pelos domicílios particulares	24779
V004	Cônjuges ou companheiros	36340
V005	Filhos do responsável e do cônjuge em domicílios particulares	24350
V006	Filhos somente do responsável em domicílios particulares	9337
V007	Enteados em domicílios particulares	1478
V008	Genros ou noras em domicílios particulares	1121
V009	Pais, mães, padrastos ou madrastas em domicílios particulares	2667
V010	Sogros em domicílios particulares	791
V011	Netos em domicílios particulares	3512
V012	Bisnetos em domicílios particulares	79
V013	Irmãos ou irmãs em domicílios particulares	1566
V014	Avôs ou avós em domicílios particulares	104
V015	Outros parentes em domicílios particulares, do sexo masculino	1495
V022-V043	MULHERES 0-9 anos	12667
V044-V053	10-19 anos	15466
V054-V063	20-29 anos	18920
V064-V073	30 - 39 anos	17550
V074-V083	40-49 anos	15563
V084-V093	50-59 anos	12805
V094-V103	60-69 anos	7885
V104-V134	70 anos ou mais	7844

Tabela A.12 – Domicílio_renda.

Variáveis	Descrição	Total
V003	Total do rendimento nominal mensal dos domicílios particulares permanentes	R$ 195.424.628,00
V005-V009	Domicílios particulares com rendimento nominal mensal domiciliar 0-2 salários mínimos	47375
V010-V011	Domicílios particulares com rendimento nominal mensal domiciliar 2,1- 5 salários mínimos	15198
V012	Domicílios particulares com rendimento nominal mensal domiciliar 5,1-10 salários mínimos	3751
V013	Domicílios particulares com rendimento nominal mensal domiciliar 10 ou mais salários mínimos	1309
V014	Domicílios particulares sem rendimento nominal mensal domiciliar per capita	1287

Tabela A.13 – Responsável_renda

Variáveis	Descrição	Total
V067-V069	Pessoas responsáveis moradoras em domicílios particulares permanentes 0 - 2 salários mínimos	29874
V070-V071	Pessoas responsáveis moradoras em domicílios particulares permanentes 2,1- 5 salários mínimos	22820
V072	Pessoas responsáveis moradoras em domicílios particulares permanentes 5,1 - 10 salários mínimos	7437
V073-V075	Pessoas responsáveis moradoras em domicílios particulares permanentes 10 ou mais salários mínimos	3370
V076	Pessoas responsáveis moradoras em domicílios particulares permanentes sem rendimento	5332
V077-V079	Total do rend. mensal das pessoas resp. sem dom. part. per com rendimento nominal mensal 0-2	R$ 21.185.596,00
V080-V081	Total do rend. mensal das pessoas resp. em dom. part. per com rendimento nominal mensal 2,1-5	R$ 37.512.985,00
V082	Total do rend. mensal das pessoas resp. em dom. part. per com rendimento nominal mensal 5,1-10	R$ 27.499.144,00
V083-V085	Total do rend. mensal das pessoas resp. em dom. part. per com rendimento nominal mensal 10 ou mais	R$ 33.508.426,00
V088	Total do rend. mensal das pessoas resp. em dom. particulares permanentes	R$ 119.706.151,00
V089-V091	Homens responsáveis moradoras em domicílios particulares permanentes 0 - 2 salários mínimos	16641
V092-V093	Homens responsáveis moradoras em domicílios particulares permanentes 2,1- 5 salários mínimos	17340
V094	Homens responsáveis moradoras em domicílios particulares permanentes 5,1 - 10 salários mínimos	5713
V095-V097	Homens responsáveis moradoras em domicílios particulares permanentes 10 ou mais salários mínimos	2734
V099-V101	Total do rend. nominal mensal dos Homens resp. em domicílios part. per com rend. nom. mês 0-2	R$ 12.712.043,00
V102-V103	Total do rend. nominal mensal dos Homens resp. em domicílios part. per com rend. nom. mês 2,1-5	R$ 28.534.773,00
V104	Total do rend. nominal mensal dos Homens resp. em domicílios part. per com rend. nom. mês 5,1-10	R$ 21.203.061,00
V105-V107	Total do rend. nominal mensal dos Homens resp. em domicílios part. per com rend. nom. mês 10 ou mais	R$ 27.886.015,00
V110	Total do rend. nominal mensal dos Homens resp. em domicílios particulares permanentes	R$ 90.335.892,00
V111-V113	Mulheres responsáveis moradoras em domicílios particulares permanentes 0 - 2 salários mínimos	13233
V114-V115	Mulheres responsáveis moradoras em domicílios particulares permanentes 2,1- 5 salários mínimos	5480
V116	Mulheres responsáveis moradoras em domicílios particulares permanentes 5,1 - 10 salários mínimos	1724
V117-V119	Mulheres responsáveis moradoras em domicílios particulares permanentes 10 ou mais salários mínimos	636
V121-V123	Total do rend. nominal mensal das Mulheres resp. em domicílios part. per com rend. nom. mês 0-2	R$ 8.473.553,00
V124-V125	Total do rend. nominal mensal das Mulheres resp. em domicílios part. per com rend. nom. mês 2,1-5	R$ 8.978.212,00
V126	Total do rend. nominal mensal das Mulheres resp. em domicílios part. per com rend. nom. mês 5,1-10	R$ 6.296.083,00
V127-V129	Total do rend. nominal mensal das Mulheres resp. em domicílios part. per com rend. nom. mês 10 ou mais	R$ 5.622.411,00
V132	Total do rend. nominal mensal das Mulheres responsáveis em domicílios particulares permanentes	R$ 29.370.259,00

Tabela A.14 – Pessoa_renda

Variáveis	Descrição	Total (R$)
V067-V069	Pessoas de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal de mais de 0 a 2 SM	76706,00
V070-V071	Pessoas de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal de mais de 2 a 5 SM	37469,00
V072	Pessoas de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal de mais de 5 a 10 SM	10667,00
V073-V075	Pessoas de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal de 10 ou mais SM	4450,00
V076	Pessoas de 10 anos ou mais de idade - domicílios part. per sem rendimento nominal mensal	57263,00
V077-V079	Total do rend. nominal mensal das pessoas de 10 anos ou mais de idade com rend. nom. mês > de 0 a 2 SM	51723853,00
V080-V081	Total do rend. nominal mensal das pessoas de 10 anos ou mais de idade com rend. nom. mês > de 2 a 5 SM	60962426,00
V082	Total do rend. nominal mensal das pessoas de 10 anos ou mais de idade com rend. nom. mês > de 5 a 10 SM	39233110,00
V083-V085	Total do rend. nominal mensal das pessoas de 10 anos ou mais de idade com rend. nom. mês de 10 ou + SM	43618173,00
V088	Total do rend. nominal mensal das pessoas de 10 anos ou mais de idade moradoras em dom. part. permanentes	195537562,00
V089-V091	Homens de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal > de 0 a 2 SM	35422,00
V092-V093	Homens de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal > de 2 a 5 SM	24477,00
V094	Homens de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal > de 5 a 10 SM	7196,00
V095-V097	Homens de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal de 10 ou + SM	3325,00
V098	Homens de 10 anos ou mais de idade - domicílios part. per sem rendimento nominal mensal	20324,00
V099-V101	Total do rend. nominal mensal dos Homens de 10 anos ou mais de idade com rend. nom. mês > de 0 a 2 SM	25737396,00
V102-V103	Total do rend. nominal mensal dos Homens de 10 anos ou mais de idade com rend. nom. mês > de 2 a 5 SM	39752539,00
V104	Total do rend. nominal mensal dos Homens de 10 anos ou mais de idade com rend. nom. mês > de 5 a 10 SM	26642406,00
V105-V107	Total do rend. nominal mensal dos Homens de 10 anos ou mais de idade com rend. nom. mês de 10 ou + SM	33892986,00
V110	Total do rend. nominal mensal dos Homens de 10 anos ou mais de idade moradoras em dom. part. permanentes	126025327,00
V111-V113	Mulheres de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal > de 0 a 2 SM	41284,00
V114-V115	Mulheres de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal > de 2 a 5 SM	12992,00
V116	Mulheres de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal > de 5 a 10 SM	3471,00
V117-V119	Mulheres de 10 anos ou mais de idade - domicílios part. per com rendimento nominal mensal de 10 ou mais SM	1125,00
V120	Mulheres de 10 anos ou mais de idade - domicílios part. per sem rendimento nominal mensal	36939,00
V121-V123	Total do rend. nominal mensal das mulheres de 10 anos ou mais de idade com rend. nom. mês > de 0 a 2 SM	25986457,00
V124-V125	Total do rend. nominal mensal das mulheres de 10 anos ou mais de idade com rend. nom. mês > de 2 a 5 SM	21209887,00
V126	Total do rend. nominal mensal das mulheres de 10 anos ou mais de idade com rend. nom. mês > de 5 a 10 SM	12590704,00
V127-V129	Total do rend. nominal mensal das mulheres de 10 anos ou mais de idade com rend. nom. mês de 10 ou + SM	9725187,00
V132	Total do rend. nominal mensal das mulheres de 10 anos ou mais de idade moradoras em dom. part. permanentes	69512235,00

APÊNDICE

B

Variáveis da Pesquisa Origem/Destino (2007/2008) realizada na cidade de São Carlos-SP

Variáveis	Descrição
ID do morador	1-10.085 moradores
Família	Família que o morador pertence
ID do domicílio	Identificação do domicílio
Renda Mensal em salários mínimos	0-15 salários mínimos
Renda Familiar em salários mínimos	0-46 salários mínimos
CNH	1- possui
	2-não possui
Idade	0-110 anos
Sexo	1- Masculino
	2- Feminino
Número de viagens realizadas pelo morador	0-39 viagens
Estuda regularmente	1-Não
	2-Ed. Infantil - Creche/Pré-escola
	3-Ensino Fund. 1ª a 4ª série ou 1º a 4º ano
	4-Ensino Fund. 5ª a 8ª série ou 5º a 9º ano
	5-Ensino Médio
	6-Superior/Universitário
	7-Outros
Grau de Instrução	1-Não alfabetizado
	2-Pré-Escola
	3-Ensino Fundamental Incompleto

	4-Ensino Fundamental Completo
	5-Ensino Médio Incompleto
	6-Ensino Médio Completo
	7-Superior Incompleto
	8-Superior Completo
Situação familiar	1-Chefe
	2-Cônjuge
	3-Filho (a)
	4-Outro parente
	5-Agregado
	6-Empregado residente
	7-Visitante não residente
	8-Residência estudantil (República)
Situação domiciliar	1-Chefe
	2-Cônjuge
	3-Filho (a)
	4-Outro parente
	5-Agregado
	6-Empregado residente
	7-Visitante não residente

Variáveis	Descrição
Condição da atividade	1-Dona de casa
	2-Estudante
	3-Em licença médica
	4-Aposentado/pensionista
	5-Ocupado (tem trabalho)
	6-Ocupado eventualmente
	7-Não ocupado (sem trabalho)
	8-Nunca trabalhou
Modo	1-ônibus
	2-ônibus fretado
	3-transporte escolar
	4-dirigindo automóvel
	5-passageiro de auto
	6-táxi
	7-lotação/perua
	8-moto
	9-bicicleta
	10-à pé
	11-outros
Condição de Renda	1- Não tem renda
	3- 2.1-4.0 SM
	4- 4.1-6.0 SM
	5- 6.1-8.0 SM
	6- 8.1-10 SM
	7-10.1-20 SM
	8-Não respondeu
Setor Censitário 2000	245 Setores-IBGE
Zonas de Tráfego	49 zonas

APÊNDICE C

Tabela microdados 2010 (IBGE) codificada

Variáveis do Registro de Domicílios					
VAR	NOME	POSIÇÃO INICIAL	POSIÇÃO FINAL	INT	DEC
V0001	UNIDADE DA FEDERAÇÃO: 11- Rondônia 12- Acre 13- Amazonas 14- Roraima 15- Pará 16- Amapá 17- Tocantins 21- Maranhão 22- Piauí 23- Ceará 24- Rio Grande do Norte 25- Paraíba 26- Pernambuco 27- Alagoas 28- Sergipe 29- Bahia 31- Minas Gerais 32- Espírito Santo 33- Rio de Janeiro 35- São Paulo 41- Paraná 42- Santa Catarina 43- Rio Grande do Sul 50- Mato Grosso do Sul 51- Mato Grosso 52- Goiás 53- Distrito Federal	1	2	2	

V0002	CÓDIGO DO MUNICÍPIO	3	7	5	
V0011	ÁREA DE PONDERAÇÃO	8	20	13	
V0300	CONTROLE	21	28	8	
V0010	PESO AMOSTRAL	29	44	3	13
V1001	REGIÃO GEOGRÁFICA: 1- Região norte (uf=11 a 17) 2- Região nordeste (uf=21 a 29) 3- Região sudeste (uf=31 a 33 e 35) 4- Região sul (uf=41 a 43) 5- Região centro-oeste (uf=50 a 53)	45	45	1	
V1002	CÓDIGO DA MESORREGIÃO: A relação de códigos encontra-se no arquivo:	46	47	2	
V1003	CÓDIGO DA MICRORREGIÃO: A relação de códigos encontra-se no arquivo:	48	50	3	
V1004	CÓDIGO DA REGIÃO METROPOLITANA: A relação de códigos encontra-se no arquivo:	51	52	2	
V1006	SITUAÇÃO DO DOMICÍLIO: 1- Urbana 2- Rural	53	53	1	
V4001	ESPÉCIE DE UNIDADE VISITADA: 01- Domicílio particular permanente ocupado 02- Domicílio particular permanente ocupado sem entrevista realizada 05- Domicílio particular improvisado ocupado 06- Domicílio coletivo com morador	54	55	2	
V4002	TIPO DE ESPÉCIE: 11- Casa 12- Casa de vila ou em condomínio 13- Apartamento 14- Habitação em: casa de cômodos, cortiço ou cabeça de porco 15- Oca ou maloca 51- Tenda ou barraca 52- Dentro de estabelecimento 53- Outro (vagão, trailer, gruta, etc) 61- Asilo, orfanato e similares com morador 62- Hotel, pensão e similares com morador 63- Alojamento de trabalhadores com morador 64- Penitenciária, presídio ou casa de detenção com morador 65- Outro com morador	56	57	2	
V0201	DOMICÍLIO, CONDIÇÃO DE OCUPAÇÃO: 1- Próprio de algum morador - já pago 2- Próprio de algum morador - ainda pagando 3- Alugado 4- Cedido por empregador 5- Cedido de outra forma 6- Outra condição Branco	58	58	1	
V2011	VALOR DO ALUGUEL (EM REAIS)	59	64	6	
V2012	ALUGUEL EM Nº DE SALÁRIOS MÍNIMOS	65	73	4	5
V0202	MATERIAL PREDOMINANTE, PAREDES EXTERNAS: 1- Alvenaria com revestimento 2- Alvenaria sem revestimento 3- Madeira apropriada para construção (aparelhada) 4- Taipa revestida 5- Taipa não revestida 6- Madeira aproveitada 7- Palha 8- Outro material 9- Sem parede Branco	74	74	1	

V0203	CÔMODOS, NÚMERO: - Branco - 1 a 30	75	76	2	
V6203	DENSIDADE DE MORADOR/CÔMODO	77	79	2	1
V0204	CÔMODOS COMO DORMITÓRIO, NÚMERO: - Branco - 1 a 15	80	81	2	
V6204	DENSIDADE DE MORADOR / DORMITÓRIO	82	84	2	1
V0205	BANHEIROS DE USO EXCLUSIVO, NÚMERO: 0- Zero banheiros 1- Um banheiro 2- Dois banheiros 3- Três banheiros 4- Quatro banheiros 5- Cinco banheiros 6- Seis banheiros 7- Sete banheiros 8- Oito banheiros 9- Nove ou mais banheiros Branco	85	85	1	
V0206	SANITÁRIO OU BURACO PARA DEJEÇÕES, EXISTÊNCIA: 1- Sim 2- Não Branco	86	86	1	
V0207	ESGOTAMENTO SANITÁRIO, TIPO: 1- Rede geral de esgoto ou pluvial 2- Fossa séptica 3- Fossa rudimentar 4- Vala 5- Rio, lago ou mar 6- Outro Branco	87	87	1	

V0208	ABASTECIMENTO DE ÁGUA, FORMA: 01- Rede geral de distribuição 02- Poço ou nascente na propriedade 03- Poço ou nascente fora da propriedade 04- Carro-pipa 05- Água da chuva armazenada em cisterna 06- Água da chuva armazenada de outra forma 07- Rios, açudes, lagos e igarapés 08- Outra 09- Poço ou nascente na aldeia 10- Poço ou nascente fora da aldeia Branco	88	89	2	
V0209	ABASTECIMENTO DE ÁGUA, CANALIZAÇÃO: 1- Sim, em pelo menos um cômodo 2- Sim, só na propriedade ou terreno 3- Não Branco	90	90	1	
V0210	LIXO, DESTINO: 1- Coletado diretamente por serviço de limpeza 2- Colocado em caçamba de serviço de limpeza 3- Queimado (na propriedade) 4- Enterrado (na propriedade) 5- Jogado em terreno baldio ou logradouro 6- Jogado em rio, lago ou mar 7- Tem outro destino Branco	91	91	1	
V0211	ENERGIA ELÉTRICA, EXISTÊNCIA: 1- Sim, de companhia distribuidora 2- Sim, de outras fontes 3- Não existe energia elétrica Branco	92	92	1	

V0212	EXISTÊNCIA DE MEDIDOR OU RELÓGIO, ENERGIA ELÉTRICA, COMPANHIA DISTRIBUIDORA: 1- Sim, de uso exclusivo 2- Sim, de uso comum 3- Não tem medidor ou relógio Branco	93	93	1	
V0213	RÁDIO, EXISTÊNCIA: 1- Sim 2- Não Branco	94	94	1	
V0214	TELEVISÃO, EXISTÊNCIA: 1- Sim 2- Não Branco	95	95	1	
V0215	MÁQUINA DE LAVAR ROUPA, EXISTÊNCIA: 1- Sim 2- Não Branco	96	96	1	
V0216	GELADEIRA, EXISTÊNCIA: 1- Sim 2- Não Branco	97	97	1	
V0217	TELEFONE CELULAR, EXISTÊNCIA: 1- Sim 2- Não Branco	98	98	1	
V0218	TELEFONE FIXO, EXISTÊNCIA: 1- Sim 2- Não Branco	99	99	1	
V0219	MICROCOMPUTADOR, EXISTÊNCIA: 1- Sim 2- Não Branco	100	100	1	

V0220	MICROCOMPUTADOR COM ACESSO À INTERNET, EXISTÊNCIA: 1- Sim 2- Não Branco	101	101	1	
V0221	MOTOCICLETA PARA USO PARTICULAR, EXISTÊNCIA: 1- Sim 2- Não Branco	102	102	1	
V0222	AUTOMÓVEL PARA USO PARTICULAR, EXISTÊNCIA: 1- Sim 2- Não Branco	103	103	1	
V0301	ALGUMA PESSOA QUE MORAVA COM VOCÊ(S) ESTAVA MORANDO EM OUTRO PAÍS EM 31 DE JULHO DE 2010: 1- Sim 2- Não Branco	104	104	1	
V0401	QUANTAS PESSOAS MORAVAM NESTE DOMICÍLIO EM 31 DE JULHO DE 2010	105	106	2	
V0402	A RESPONSABILIDADE PELO DOMICÍLIO É DE: 1- Apenas um morador 2- Mais de um morador 9- Ignorado Branco	107	107	1	
V0701	DE AGOSTO DE 2009 A JULHO DE 2010, FALECEU ALGUMA PESSOA QUE MORAVA COM VOCÊ(S) (INCLUSIVE CRIANÇAS RECÉM-NASCIDAS E IDOSOS): 1- Sim 2- Não Branco	108	108	1	
V6529	RENDIMENTO MENSAL DOMICILIAR EM JULHO DE 2010	109	115	7	
V6530	RENDIMENTO DOMICILIAR, SALÁRIOS MÍNIMOS, EM JULHO DE 2010	116	125	5	5
V6531	RENDIMENTO DOMICILIAR PER CAPITA EM JULHO DE 2010	126	133	6	2
V6532	RENDIMENTO DOMICILIAR PER CAPITA, EM Nº DE SALÁRIOS MÍNIMOS, EM JULHO DE 2010	134	142	4	5

V6600	Espécie da Unidade Doméstica 1- Unipessoal 2- Nuclear 3- Estendida 4- Composta Branco (Domicílio Coletivo)	143	143	1	
V6210	ADEQUAÇÃO DA MORADIA 1- Adequada 2- Semi adequada 3- Inadequada Branco	144	144	1	
M0201	MARCA DE IMPUTAÇÃO NA V0201: 1- Sim 2- Não	145	145	1	
M2011	MARCA DE IMPUTAÇÃO NA V2011: 1- Sim 2- Não	146	146	1	
M0202	MARCA DE IMPUTAÇÃO NA V0202: 1- Sim 2- Não	147	147	1	
M0203	MARCA DE IMPUTAÇÃO NA V0203: 1- Sim 2- Não	148	148	1	
M0204	MARCA DE IMPUTAÇÃO NA V0204: 1- Sim 2- Não	149	149	1	
M0205	MARCA DE IMPUTAÇÃO NA V0205: 1- Sim 2- Não	150	150	1	
M0206	MARCA DE IMPUTAÇÃO NA V0206: 1- Sim 2- Não	151	151	1	
M0207	MARCA DE IMPUTAÇÃO NA V0207: 1- Sim 2- Não	152	152	1	

M0208	MARCA DE IMPUTAÇÃO NA V0208: 1- Sim 2- Não	153	153	1	
M0209	MARCA DE IMPUTAÇÃO NA V0209: 1- Sim 2- Não	154	154	1	
M0210	MARCA DE IMPUTAÇÃO NA V0210: 1- Sim 2- Não	155	155	1	
M0211	MARCA DE IMPUTAÇÃO NA V0211: 1- Sim 2- Não	156	156	1	
M0212	MARCA DE IMPUTAÇÃO NA V0212: 1- Sim 2- Não	157	157	1	
M0213	MARCA DE IMPUTAÇÃO NA V0213: 1- Sim 2- Não	158	158	1	
M0214	MARCA DE IMPUTAÇÃO NA V0214: 1- Sim 2- Não	159	159	1	
M0215	MARCA DE IMPUTAÇÃO NA V0215: 1- Sim 2- Não	160	160	1	
M0216	MARCA DE IMPUTAÇÃO NA V0216: 1- Sim 2- Não	161	161	1	
M0217	MARCA DE IMPUTAÇÃO NA V0217: 1- Sim 2- Não	162	162	1	
M0218	MARCA DE IMPUTAÇÃO NA V0218: 1- Sim 2- Não	163	163	1	
M0219	MARCA DE IMPUTAÇÃO NA V0219: 1- Sim 2- Não	164	164	1	

M0220	MARCA DE IMPUTAÇÃO NA V0220: 1- Sim 2- Não	165	165	1	
M0221	MARCA DE IMPUTAÇÃO NA V0221: 1- Sim 2- Não	166	166	1	
M0222	MARCA DE IMPUTAÇÃO NA V0222: 1- Sim 2- Não	167	167	1	
M0301	MARCA DE IMPUTAÇÃO NA V0301: 1- Sim 2- Não	168	168	1	
M0401	MARCA DE IMPUTAÇÃO NA V0401: 1- Sim 2- Não	169	169	1	
M0402	MARCA DE IMPUTAÇÃO NA V0402: 1- Sim 2- Não	170	170	1	
M0701	MARCA DE IMPUTAÇÃO NA V0701: 1- Sim 2- Não	171	171	1	

APÊNDICE

D

Descrição do algoritmo utilizado para gerar a população sintética

```
Sub GeraPopSintetica()

    ' GERA A POPULAÇÃO SINTÉTICA A PARTIR DOS DADOS GLOBAIS DO IBGE
    ' DISTRIBUIÇÃO DE POPULAÇÃO, PARENTESCO, RENDA E ESCOLARIDADE
    ' POR SETOR CENSITÁRIO

    Dim nplan, i, j As Integer
    Dim ulinha As Long
    Dim noPlan As Boolean
    Dim nomeplan, originplan As String
    nplan = ActiveWorkbook.Worksheets.Count
    noPlan = True
    originplan = ActiveSheet.Name

    '##########################################
    '' VERIFICAÇÕES
    '##########################################
    ' Verifica se os dados estão em porcentagem
    If (ActiveSheet.Range("D2") > 1) Then
       ActiveSheet.Copy  Before:=Sheets(1)
       ActiveSheet.Name = "temp"
       originplan = ActiveSheet.Name
       ActiveSheet.Copy  Before:=Sheets(1)
       ActiveSheet.Name = "Dados_%"
       nomeplan = ActiveSheet.Name
    End If

Call FreqAcum(originplan)

    ' Verifica a existência da planilha "PopSint"
    For i = 1 To nplan
       If (ActiveWorkbook.Sheets(i).Name = "PopSint") Then
          ActiveWorkbook.Sheets(i).Columns("A:XF").Delete
          noPlan = False
          i = nplan
       End If
    Next i
    If (noPlan) Then
       Sheets(1).Select
       Sheets.Add
       ActiveSheet.Name = "PopSint"
    End If
```

```
    Cells.Select
    ' Altera fonte dos dados
    With Selection.Font
        .Name = "Calibri"
        .Size = 11
        .Strikethrough = False
        .Superscript = False
        .Subscript = False
        .OutlineFont = False
        .Shadow = False
        .Underline = xlUnderlineStyleNone
        .TintAndShade = 0
        .ThemeFont = xlThemeFontMinor
    End With

' Determina a população sintética
    Call PopSintetica


End Sub
Sub PopSintetica()

    '###########################################
    ' POPULAÇÃO SINTÉTICA
    '###########################################

    ' Variáveis
    Dim nzonas  As Integer,  ndomic As Integer, nmorad As Integer
    Dim i As Integer, j As Integer, k As Integer
    Dim nomeplan As String
    Dim ulinha As Long
    nomeplan = "temp"

    ' Cabeçalho da planilha
    Sheets("PopSint").Select
    Sheets(nomeplan).Rows("1:1").Copy
    Sheets("PopSint").Rows("1:1").Select
    ActiveSheet.Paste
    Application.CutCopyMode = False
    Sheets("PopSint").Range("B1") = "#Domicílio"
    Cells.EntireColumn.AutoFit
```

```
' Para cada Setor Censitário
  nzonas = Sheets(nomeplan).Range("A1048576").End(xlUp).Row
  ulinha = 2

  Randomize
  For i = 2 To nzonas
    Dim val As Double
    Dim col As Integer

    ' Número de domicílios no Setor Censitário
    ndomic = Sheets(nomeplan).Cells(i, 2)

    ' Verifica se existem domicílios no Setor Censitário
    If (ndomic > 0) Then
        For j = 1 To ndomic

        ' Copia todos os dados
        Sheets(nomeplan).Range("A" & i & ":AT" & i).Copy
        Sheets("PopSint").Range("A" & ulinha & ":AT" & ulinha).Select
        ActiveSheet.Paste
        Application.CutCopyMode = False
        Sheets("PopSint").Range("C" & ulinha & ":AT" & ulinha) = 0

'======================
        ' #ID DOMICÍCLIO
        If (ulinha = 2) Then
          ActiveSheet.Cells(ulinha, 2) = 1
        Else
          ActiveSheet.Cells(ulinha, 2) = ActiveSheet.Cells(ulinha - 1, 2) + 1
        End If

        '======================
        ' #MORADORES DOMICÍLIO
        col = defineParametro(i, "D", "M")
        nmorad = col - 3
        If (nmorad = 10) Then
          nmorad = nmorad + Sheets("temp").Cells(i, 47)
          Sheets("temp").Cells(i, 47) = 0
        End If
        ActiveSheet.Cells(ulinha, 3) = nmorad
```

```
' Acrescenta na Pop. Sintética
        ActiveSheet.Cells(ulinha, col) = ActiveSheet.Cells(ulinha, col) + 1

        '======================
        ' #RENDA DOMICÍLIO
        col = defineParametro(i, "AO", "AT")

        ' Acrescenta na Pop. Sintética
        ActiveSheet.Cells(ulinha, col) = ActiveSheet.Cells(ulinha, col) + 1
        ' Retira da contagem total de domicílios
        'Sheets("temp").Cells(i, 2) = Sheets("temp").Cells(i, 2) – 1

' Gera moradores por domicílio
        For k = 1 To nmorad
          '# SEXO
          col = defineParametro(i, "N", "O")
          ' Acrescenta na Pop. Sintética
          ActiveSheet.Cells(ulinha, col) = ActiveSheet.Cells(ulinha, col) + 1

          '# PARENTESCO
          If (k = 1) Then
            ' O primeiro sempre é CHEFE
            col = Asc("P") - Asc("A") + 1
          ElseIf ((Sheets("temp").Range("Q" & i) > 0) And _
            ((k = 2) Or (Sheets("PopSint").Range("Q" & ulinha) <
Sheets("PopSint").Range("P" & ulinha)))) Then
            ' O segundo morador é, preferencialmente, CONJUGE
            col = Asc("Q") - Asc("A") + 1
          ElseIf ((k > 2) And (Sheets("PopSint").Range("Q" & ulinha) =
Sheets("PopSint").Range("P" & ulinha))) Then
            ' Sorteia categoria menos o CONJUGE
            Dim temp As Integer
            temp = Sheets("temp").Range("Q" & i)
            Sheets("temp").Range("Q" & i) = 0
            col = defineParametro(i, "P", "V")
            Sheets("temp").Range("Q" & i) = temp

          ' %%%%% TROQUEI AQUI
          Else
            col = defineParametro(i, "P", "V")
          ' %%%%% ATE AQUI
```

```
End If
            ' Acrescenta na Pop. Sintética
            ActiveSheet.Cells(ulinha, col) = ActiveSheet.Cells(ulinha, col) + 1

            '# IDADE
            col = defineParametro(i, "W", "AE")
            ' Acrescenta na Pop. Sintética
            ActiveSheet.Cells(ulinha, col) = ActiveSheet.Cells(ulinha, col) + 1
            ' &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
            ' IDADE DE CHEFE > 10 SEGUNDO IBGE
            ' IDADE DE CONJUGE TAMBÉM
            ' &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&

            ' #OBS1: Acrescenta na planilha de contagem de idade,
            ' para obter a escolaridade.
            Sheets("temp").Cells(i, col) = Sheets("temp").Cells(i, col) + 1


            '# ESCOLARIDADE
            val = Rnd
            ' Define a coluna da escolaridade, de acordo com a idade
            col = col + 9
            If (val <= Sheets("Dados_%").Cells(i, col)) Then
               ' Acrescenta na Pop. Sintética
               ActiveSheet.Cells(ulinha, col) = ActiveSheet.Cells(ulinha, col) + 1
               ' Retira da categoria
               Sheets("temp").Cells(i, col) = Sheets("temp").Cells(i, col) - 1
            End If

' #OBS1: Retira a idade acrescentada
            Sheets("temp").Cells(i, col - 9) = Sheets("temp").Cells(i, col - 9) - 1

         Next k
         ulinha = ulinha + 1
      Next j

   End If
Next i
ActiveSheet.Range("a1").Select
Cells.EntireColumn.AutoFit
```

```
 ' Altera fonte dos dados
   With Selection.Font
      .Name = "Calibri"
      .Size = 11
      .Strikethrough = False
      .Superscript = False
      .Subscript = False
      .OutlineFont = False
      .Shadow = False
      .Underline = xlUnderlineStyleNone
      .TintAndShade = 0
      .ThemeFont = xlThemeFontMinor
   End With

   Sheets("Dados_%").Delete
   Sheets("temp").Delete

End Sub
Function defineParametro(lin As Integer, coli As String, colf As String) As Integer
   Dim val As Double
   Dim Rng As Range

   ' Sorteia a categoria
   cond = False
   val = Rnd
   Set Rng = Sheets("Dados_%").Range(coli & lin & ":" & colf & lin)
   For Each cell In Rng
      If (cell > val) Then
         col = cell.Column
         Exit For
      End If
   Next cell

   ' Retira da contagem por categoria
   Sheets("temp").Cells(lin, col) = Sheets("temp").Cells(lin, col) - 1

   defineParametro = col
End Function
Sub FreqAcum(nome As String)
   Dim ulinha As Double
   ulinha = ActiveSheet.Range("A1048576").End(xlUp).Row
```

```
' REFAZ A DISTRIBUIÇÃO DE N. MORADORES POR SETOR CENSITARIO

  Sheets(nome).Range("AV2:BE" & ulinha).FormulaR1C1 = "=ROUND(RC2*(RC[-
44]/SUM(RC4:RC13)),0)"
  Sheets(nome).Range("AV2:BE" & ulinha).Copy
  Sheets(nome).Range("D2:M" & ulinha).PasteSpecial xlValues
  Sheets(nome).Range("AV2:BE" & ulinha).ClearContents

  'Moradores excedentes para domicílios com mais de 10 moradores
  Sheets(nome).Range("AU2").FormulaArray = "=RC[-44]-SUM(RC[-43]:RC[
34]*{1,2,3,4,5,6,7,8,9,10})"
  Sheets(nome).Range("AU2").AutoFill Destination:=Sheets(nome).Range("AU2:AU" &
ulinha)
  Sheets(nome).Range("AU2:AU" & ulinha).Copy
  Sheets(nome).Range("AU2:AU" & ulinha).PasteSpecial xlValues

  ' REFAZ A DISTRIBUIÇÃO DO SEXO

  Sheets(nome).Range("AV2:AW" & ulinha).FormulaR1C1 = "=ROUND(RC3*(RC[-
34]/SUM(RC14:RC15)),0)"
  Sheets(nome).Range("AV2:AW" & ulinha).Copy
  Sheets(nome).Range("N2:O" & ulinha).PasteSpecial xlValues
  Sheets(nome).Range("AX2:AX" & ulinha).FormulaR1C1 = "=RC3-SUM(RC14:RC15)"
  Sheets(nome).Range("AX2:AX" & ulinha).Copy
  Sheets(nome).Range("N2:N" & ulinha).PasteSpecial Paste:=xlValues, Operation:=xlAdd
  Sheets(nome).Range("AV2:AX" & ulinha).ClearContents

' REFAZ A DISTRIBUIÇÃO DO PARENTESCO

  Sheets(nome).Range("AV2:BB" & ulinha).FormulaR1C1 = "=ROUND(RC3*(RC[-
32]/SUM(RC16:RC22)),0)"
  Sheets(nome).Range("AV2:BB" & ulinha).Copy
  Sheets(nome).Range("P2:V" & ulinha).PasteSpecial xlValues
  Sheets(nome).Range("BC2:BC" & ulinha).FormulaR1C1 = "=RC3-SUM(RC16:RC22)"
  Sheets(nome).Range("BC2:BC" & ulinha).Copy
  Sheets(nome).Range("S2:S" & ulinha).PasteSpecial Paste:=xlValues, Operation:=xlAdd
  Sheets(nome).Range("AV2:BC" & ulinha).ClearContents
```

```
' REFAZ A DISTRIBUIÇÃO DA IDADE E ESCOLARIDADE (razão alfabetizados/idade)

   Sheets(nome).Range("AV2:BD" & ulinha).FormulaR1C1 = "=ROUND(RC3*(RC[-
25]/SUM(RC23:RC31)),0)"
   Sheets(nome).Range("BE2:BM" & ulinha).FormulaR1C1 = "=MIN(1,RC[-25]/RC[-34])"
   Sheets(nome).Range("BE2:BM" & ulinha).Copy
   Sheets(nome).Range("BE2:BM" & ulinha).PasteSpecial xlValues

   Sheets(nome).Range("AV2:BD" & ulinha).Copy
   Sheets(nome).Range("W2:AE" & ulinha).PasteSpecial xlValues
   Sheets(nome).Range("AV2:BD" & ulinha).FormulaR1C1 = "=ROUND(RC[9]*RC[-25],0)"
   Sheets(nome).Range("AV2:BD" & ulinha).Copy
   Sheets(nome).Range("AF2:AN" & ulinha).PasteSpecial xlValues
   Sheets(nome).Range("AV2:AV" & ulinha).FormulaR1C1 = "=RC3-SUM(RC[-25]:RC[-17])"
   Sheets(nome).Range("AV2:AV" & ulinha).Copy
   Sheets(nome).Range("AE2:AE" & ulinha).PasteSpecial Paste:=xlValues, Operation:=xlAdd
   Sheets(nome).Range("AV2:BM" & ulinha).ClearContents

   ' REFAZ A DISTRIBUIÇÃO DE RENDA DOMICILIAR

   Sheets(nome).Range("AV2:BA" & ulinha).FormulaR1C1 = "=ROUND(RC2 * (RC[-
7]/SUM(RC41:RC46)),0)"
   Sheets(nome).Range("BB2:BB" & ulinha).FormulaR1C1 = "=RC2 - SUM(RC48:RC53)"
   Sheets(nome).Range("AV2:BB" & ulinha).Copy
   Sheets(nome).Range("AV2:BB" & ulinha).PasteSpecial xlValues
   Sheets(nome).Range("AV2:BA" & ulinha).Copy
   Sheets(nome).Range("AO2:AT" & ulinha).PasteSpecial xlValues
   Sheets(nome).Range("BB2:BB" & ulinha).Copy
   Sheets(nome).Range("AP2:AP" & ulinha).PasteSpecial Paste:=xlPasteValues,
Operation:=xlAdd
   Sheets(nome).Range("AV2:BB" & ulinha).ClearContents

'##########################################
   ' FREQUENCIA ACUMULADA
   '##########################################
   ' Moradores
   ActiveSheet.Range("D2:D" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome &
"!RC4:RC13)"
   ActiveSheet.Range("E2:M" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome &
"!RC4:RC13) + RC[-1]"
```

```
' Sexo

   ActiveSheet.Range("N2:N" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome &
"!RC14:RC15)"
   ActiveSheet.Range("O2:O" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome &
"!RC14:RC15) + RC[-1]"

   ' Parentesco
   ' Retira os chefes de familia

   Sheets("temp").Range("BB2:BB" & ulinha).FormulaR1C1 = "=temp!RC16-temp!RC2"
   Sheets("temp").Range("BB2:BB" & ulinha).Copy
   Sheets("temp").Range("BB2:BB" & ulinha).PasteSpecial xlValues
   Sheets("temp").Range("BB2:BB" & ulinha).Cut Sheets("temp").Range("P2:P" & ulinha)
   ActiveSheet.Range("P2:P" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome &
"!RC16:RC22)"
   ActiveSheet.Range("Q2:V" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome &
"!RC16:RC22) + RC[-1]"

   ' Idade

   ActiveSheet.Range("W2:W" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome
& "!RC23:RC31)"
   ActiveSheet.Range("X2:AE" & ulinha).FormulaR1C1 = "=" & nome & "!RC/SUM(" & nome
& "!RC23:RC31) + RC[-1]"

   ' Alfabetizados
   ' ** Taxa de alfabetizados por idade e por Setor Censitário **

   ActiveSheet.Range("AF2:AN" & ulinha).FormulaR1C1 = "=MIN(1," & nome & "!RC/" &
nome & "!RC[-9])"

   ' Renda
   ' ** Frequencia acumulada por domicílio **

   ActiveSheet.Range("AO2:AO" & ulinha).FormulaR1C1 = "=" & nome & "!RC/(SUM(" &
nome & "!RC41:RC46))"
   ActiveSheet.Range("AP2:AT" & ulinha).FormulaR1C1 = "=MIN(1,(" & nome & "!RC/SUM("
& nome & "!RC41:RC46)) + RC[-1])"

End Sub
```

APÊNDICE NO CD-ROM (em anexo no exemplar impresso)

E

Os apêndices a seguir também podem ser obtidos através da página: https://drive.google.com/open? id=0B-ASewTGCBR1QVRteGZTT2tvcTQ.

E.1 Dados agregados (Censo 2010-IBGE) utilizados para a geração da população sintética

E.2 Dados dos domicílios e número de viagens por domicílio da cidade de São Carlos-SP coletados pela Pesquisa O/D 2007-2008

E.3 Dados desagregados (Microdados Censo 2010-IBGE) utilizados na validação da população sintética

E.4 População Sintética

E.5 Viagens por domicílios da população sintética estimadas pelos Modelos 3 e 4

E.6 Validação das viagens por domicílio da população sintética pelo método do intervalo de confiança a nível de 95%

E.7 Análise dos percentis (95, 90, 85, 80 e 75)

Anexos

ANEXO

A

Mapa dos setores censitários do Censo Demográfico de 2010

O mapa dos setores censitários do Censo Demográfico de 2010 está disponível no CD-ROM (em anexo no exemplar impresso) ou através da página: https://drive.google.com/open?id=0B-ASewTGCBR1TmdPOThHZ3BObmM

ANEXO

B

Formulário para a entrevista domiciliar - Pesquisa O/D (2007/2008) de São Carlos-SP

PESQUISA ORIGEM E DESTINO DE SÃO CARLOS - 2007 - ENTREVISTA DOMICILIAR

Bloco 1

Zona	Domicílio	Tipo	Pesquisador

Visitas ao Domicílio	
Data 1ª visita	Horário
Data 2ª visita	Horário
Data 3ª visita	Horário
Pesquisador	Supervisor
Telefone para contato	

Resultado do Domicilio

1 - Recusa
2 - Fechado
3 - Vago
4 - Incompleto
5 - Completo sem viagem
6 - Completo com viagem

Total de Grupos Familiares	Total de Moradores no Domicílio	Data da Entrevista

Caracterização dos Moradores do Domicílio por Grupo Familiar (perguntas ao chefe ou cônjuge)

1	2	3	4	5	6	7	8	9	10	11	12	13	14
Primeiro Nome da Pessoa (começar pelo chefe)	Nº do Grupo Familiar	Nº da Pessoa (Indiv.)	Sit. Dom.	Sit. Fam.	Idade	Possui Carteira Habilit.	Sexo	Estuda Regular-mente	Grau de Instrução	Cond. de Ativ.	Condição de Renda	Renda Mensal (em R$)	Total Viagens

4 - Situação Domiciliar
1. Chefe
2. Cônjuge
3. Filho (a)
4. Outro Parente
5. Agregado
6. Empregado Residente
7. Visitante Não Residente

5 - Situação Familiar
1. Chefe
2. Cônjuge
3. Filho (a)
4. Outro Parente
5. Agregado
6. Empregado Residente
7. Visitante Não Residente
8. Residência Estudantil (República)

7 - Possui Carteira de Habilitação?
1. Sim
2. Não

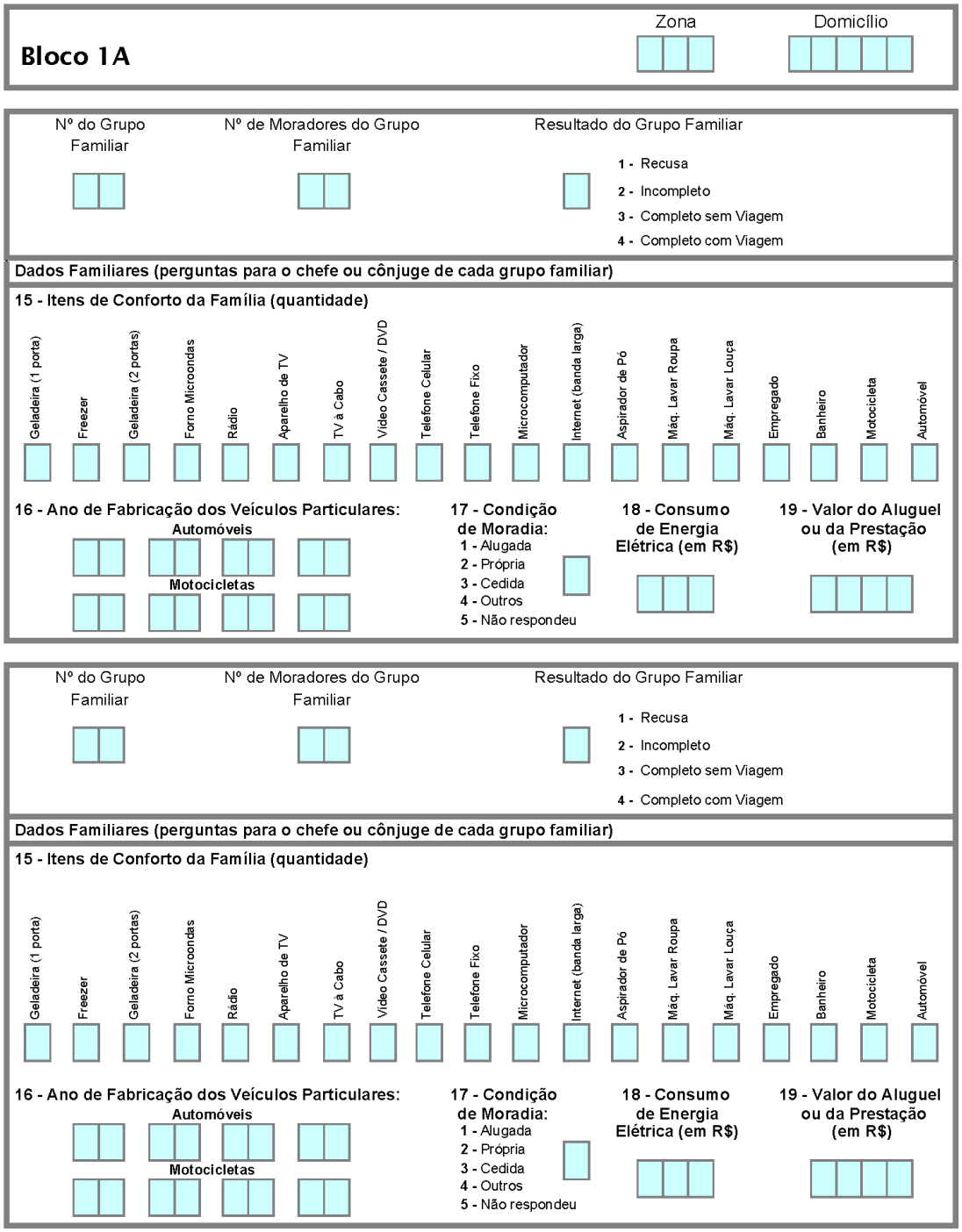
8 - Sexo
1. Masculino
2. Feminino

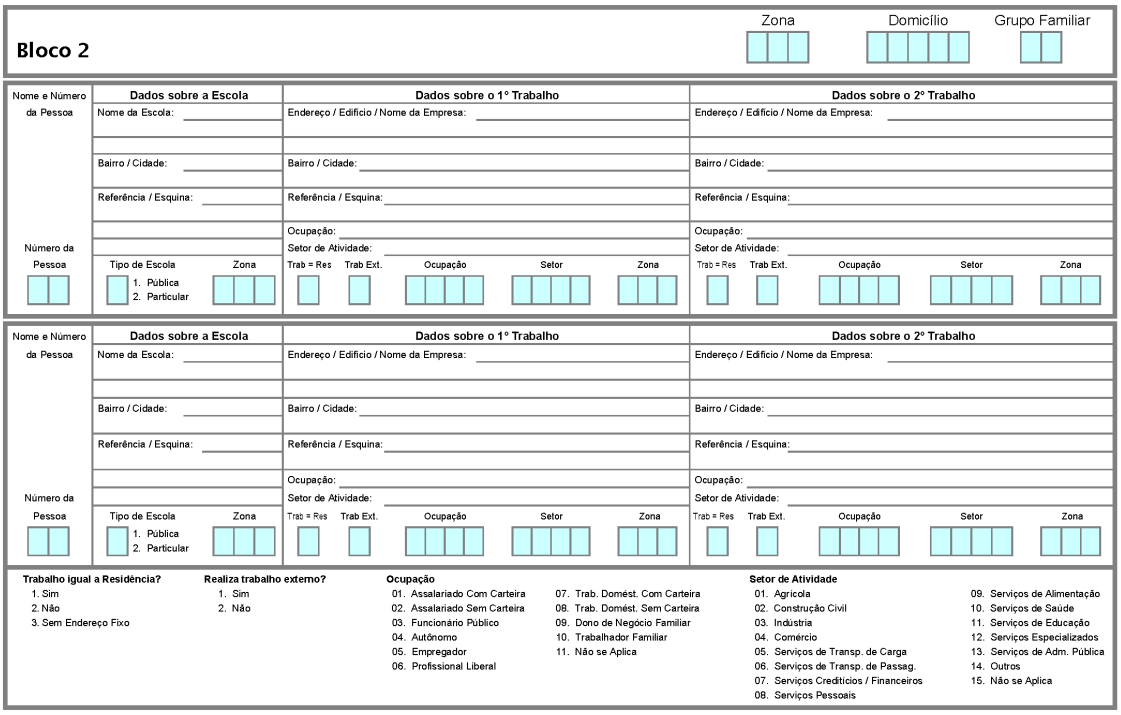
9 - Estuda Regularmente?
1. Não
2. Educação Infantil - Creche / Pré-Escola
3. Ensino Fund. - 1ª a 4ª Série ou 1º ao 4º ano
4. Ensino Fund. - 5ª a 8ª Série ou 5º ao 9º ano
5. Ensino Médio
6. Superior / Universitário
7. Outros

10 - Grau de instrução
1. Não Alfabetizado
2. Pré-Escola
3. Ensino Fundamental Incompleto
4. Ensino Fundamental Completo
5. Ensino Médio Incompleto
6. Ensino Médio Completo
7. Superior Incompleto
8. Superior Completo

11 - Condição de Atividade
1. Dona de Casa
2. Estudante
3. Em Licença Médica
4. Aposentado / Pensionista
5. Ocupado (tem trabalho)
6. Ocupado Eventualmente (faz bico)
7. Não Ocupado (sem trabalho)
8. Nunca Trabalhou

12 - Condição de Renda
1. Não Tem Renda
2. 0,0 - 2,0 SM
3. 2,1 - 4,0 SM
4. 4,1 - 6,0 SM
5. 6,1 - 8,0 SM
6. 8,1 - 10,0 SM
7. 10,1 - 20,0 SM
8. Não Respondeu


Bloco 1A
Zona
Domicílio
Nº do Grupo Familiar
Nº de Moradores do Grupo Familiar
Resultado do Grupo Familiar
1 - Recusa
2 - Incompleto
3 - Completo sem Viagem
4 - Completo com Viagem
Dados Familiares (perguntas para o chefe ou cônjuge de cada grupo familiar)
15 - Itens de Conforto da Família (quantidade)
Geladeira (1 porta)
Freezer
Geladeira (2 portas)
Forno Microondas
Rádio
Aparelho de TV
TV à Cabo
Vídeo Cassete / DVD
Telefone Celular
Telefone Fixo
Microcomputador
Internet (banda larga)
Aspirador de Pó
Máq. Lavar Roupa
Máq. Lavar Louça
Empregado
Banheiro
Motocicleta
Automóvel
16 - Ano de Fabricação dos Veículos Particulares:
Automóveis
Motocicletas
17 - Condição de Moradia:
1 - Alugada
2 - Própria
3 - Cedida
4 - Outros
5 - Não respondeu
18 - Consumo de Energia Elétrica (em R$)
19 - Valor do Aluguel ou da Prestação (em R$)
Nº do Grupo Familiar
Nº de Moradores do Grupo Familiar
Resultado do Grupo Familiar
1 - Recusa
2 - Incompleto
3 - Completo sem Viagem
4 - Completo com Viagem
Dados Familiares (perguntas para o chefe ou cônjuge de cada grupo familiar)
15 - Itens de Conforto da Família (quantidade)
Geladeira (1 porta)
Freezer
Geladeira (2 portas)
Forno Microondas
Rádio
Aparelho de TV
TV à Cabo
Vídeo Cassete / DVD
Telefone Celular
Telefone Fixo
Microcomputador
Internet (banda larga)
Aspirador de Pó
Máq. Lavar Roupa
Máq. Lavar Louça
Empregado
Banheiro
Motocicleta
Automóvel
16 - Ano de Fabricação dos Veículos Particulares:
Automóveis
Motocicletas
17 - Condição de Moradia:
1 - Alugada
2 - Própria
3 - Cedida
4 - Outros
5 - Não respondeu
18 - Consumo de Energia Elétrica (em R$)
19 - Valor do Aluguel ou da Prestação (em R$)



Bloco 2

Zona | Domicílio | Grupo Familiar

Nome e Número da Pessoa	Dados sobre a Escola	Dados sobre o 1° Trabalho	Dados sobre o 2° Trabalho
	Nome da Escola:	Endereço / Edifício / Nome da Empresa:	Endereço / Edifício / Nome da Empresa:
	Bairro / Cidade:	Bairro / Cidade:	Bairro / Cidade:
	Referência / Esquina:	Referência / Esquina:	Referência / Esquina:
		Ocupação:	Ocupação:
Número da Pessoa	Tipo de Escola (1. Pública 2. Particular) / Zona	Setor de Atividade:	Setor de Atividade:
		Trab = Res / Trab Ext. / Ocupação / Setor / Zona	Trab = Res / Trab Ext. / Ocupação / Setor / Zona

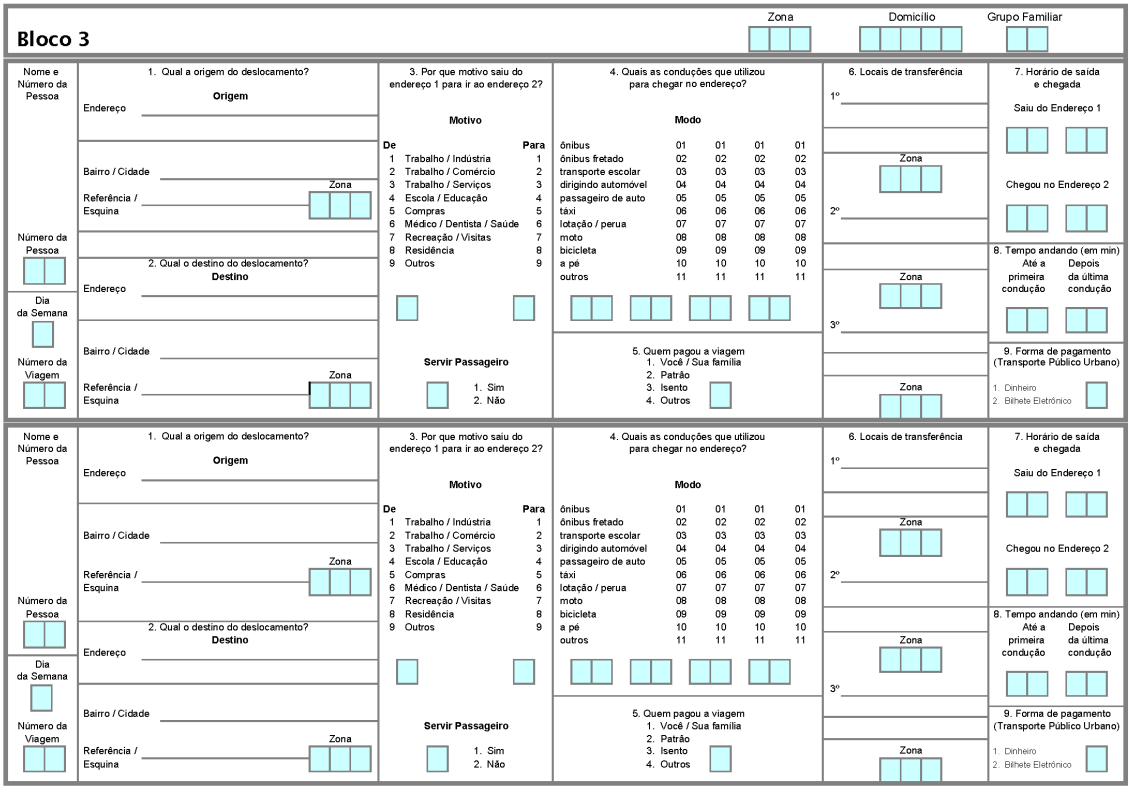
Nome e Número da Pessoa	Dados sobre a Escola	Dados sobre o 1° Trabalho	Dados sobre o 2° Trabalho
	Nome da Escola:	Endereço / Edifício / Nome da Empresa:	Endereço / Edifício / Nome da Empresa:
	Bairro / Cidade:	Bairro / Cidade:	Bairro / Cidade:
	Referência / Esquina:	Referência / Esquina:	Referência / Esquina:
		Ocupação:	Ocupação:
Número da Pessoa	Tipo de Escola (1. Pública 2. Particular) / Zona	Setor de Atividade:	Setor de Atividade:
		Trab = Res / Trab Ext. / Ocupação / Setor / Zona	Trab = Res / Trab Ext. / Ocupação / Setor / Zona

Trabalho igual a Residência?
1. Sim
2. Não
3. Sem Endereço Fixo

Realiza trabalho externo?
1. Sim
2. Não

Ocupação
01. Assalariado Com Carteira
02. Assalariado Sem Carteira
03. Funcionário Público
04. Autônomo
05. Empregador
06. Profissional Liberal
07. Trab. Domést. Com Carteira
08. Trab. Domést. Sem Carteira
09. Dono de Negócio Familiar
10. Trabalhador Familiar
11. Não se Aplica

Setor de Atividade
01. Agrícola
02. Construção Civil
03. Indústria
04. Comércio
05. Serviços de Transp. de Carga
06. Serviços de Transp. de Passag.
07. Serviços Creditícios / Financeiros
08. Serviços Pessoais
09. Serviços de Alimentação
10. Serviços de Saúde
11. Serviços de Educação
12. Serviços Especializados
13. Serviços de Adm. Pública
14. Outros
15. Não se Aplica



Bloco 3

Zona | Domicílio | Grupo Familiar

Nome e Número da Pessoa

1. Qual a origem do deslocamento?
Origem
Endereço
Bairro / Cidade
Referência / Esquina — Zona

2. Qual o destino do deslocamento?
Destino
Endereço
Bairro / Cidade
Referência / Esquina — Zona

Número da Pessoa

Dia da Semana

Número da Viagem

3. Por que motivo saiu do endereço 1 para ir ao endereço 2?
Motivo

De		Para
1	Trabalho / Indústria	1
2	Trabalho / Comércio	2
3	Trabalho / Serviços	3
4	Escola / Educação	4
5	Compras	5
6	Médico / Dentista / Saúde	6
7	Recreação / Visitas	7
8	Residência	8
9	Outros	9

Servir Passageiro
1. Sim
2. Não

4. Quais as conduções que utilizou para chegar no endereço?
Modo

ônibus	01	01	01	01
ônibus fretado	02	02	02	02
transporte escolar	03	03	03	03
dirigindo automóvel	04	04	04	04
passageiro de auto	05	05	05	05
taxi	06	06	06	06
lotação / perua	07	07	07	07
moto	08	08	08	08
bicicleta	09	09	09	09
a pé	10	10	10	10
outros	11	11	11	11

5. Quem pagou a viagem
1. Você / Sua família
2. Patrão
3. Isento
4. Outros

6. Locais de transferência
1°
Zona
2°
Zona
3°
Zona

7. Horário de saída e chegada
Saiu do Endereço 1
Chegou no Endereço 2

8. Tempo andando (em min)
Até a primeira condução
Depois da última condução

9. Forma de pagamento (Transporte Público Urbano)
1. Dinheiro
2. Bilhete Eletrônico

Nome e Número da Pessoa

1. Qual a origem do deslocamento?
Origem
Endereço
Bairro / Cidade
Referência / Esquina — Zona

2. Qual o destino do deslocamento?
Destino
Endereço
Bairro / Cidade
Referência / Esquina — Zona

Número da Pessoa

Dia da Semana

Número da Viagem

3. Por que motivo saiu do endereço 1 para ir ao endereço 2?
Motivo

De		Para
1	Trabalho / Indústria	1
2	Trabalho / Comércio	2
3	Trabalho / Serviços	3
4	Escola / Educação	4
5	Compras	5
6	Médico / Dentista / Saúde	6
7	Recreação / Visitas	7
8	Residência	8
9	Outros	9

Servir Passageiro
1. Sim
2. Não

4. Quais as conduções que utilizou para chegar no endereço?
Modo

ônibus	01	01	01	01
ônibus fretado	02	02	02	02
transporte escolar	03	03	03	03
dirigindo automóvel	04	04	04	04
passageiro de auto	05	05	05	05
taxi	06	06	06	06
lotação / perua	07	07	07	07
moto	08	08	08	08
bicicleta	09	09	09	09
a pé	10	10	10	10
outros	11	11	11	11

5. Quem pagou a viagem
1. Você / Sua família
2. Patrão
3. Isento
4. Outros

6. Locais de transferência
1°
Zona
2°
Zona
3°
Zona

7. Horário de saída e chegada
Saiu do Endereço 1
Chegou no Endereço 2

8. Tempo andando (em min)
Até a primeira condução
Depois da última condução

9. Forma de pagamento (Transporte Público Urbano)
1. Dinheiro
2. Bilhete Eletrônico