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

Previsão da demanda por viagens domiciliares através de método sequencial baseado em população sintética e redes neurais artificiais

Marcela Navarro Pianucci[1], **Cira Souza Pitombo**[2], **André Luiz Cunha**[3], **Paulo César Lima Segantine**[4]

[1]Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, USP, manavarropg@gmail.com
[2]Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, USP, cirapitombo@usp.br
[3]Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, USP, alcunha@usp.br
[4]Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, USP, seganta@sc.usp.br

Recebido:
20 de junho de 2017
Aceito para publicação:
8 de outubro de 2019
Publicado:
31 de dezembro de 2019
Editor de área:
Helena Beatriz Cybis

Palavras-chaves:
Monte Carlo,
Viagens sintéticas,
Geração de viagens.

Keywords:
Monte Carlo,
Synthetic trips,
Trip generation.

DOI:10.14295/transportes.v27i4.1406

RESUMO
A estimativa de viagens por domicílio é fundamental para a tomada de decisões relativas ao planejamento de transportes. Porém, para obter essa estimativa são necessários dados desagregados dos domicílios, que geralmente são obtidos pela Pesquisa Domiciliar de Origem e Destino. No entanto, a maioria das cidades enfrenta problemas para a aquisição desses dados, uma vez que este tipo de pesquisa é de alto custo de preparação e execução. Desta forma, surge a necessidade de ferramentas que forneçam dados confiáveis e com baixo custo. Assim, o objetivo deste artigo é apresentar um método sequencial, para estimativa de viagens domiciliares, a partir de população sintética e Redes Neurais Artificiais (RNAs). A população sintética foi baseada em dados agregados do censo e simulação Monte Carlo. Os resultados obtidos com as RNAs foram comparados aos resultados de um modelo linear tradicional, mostrando-se melhores e corroborando o potencial do uso de RNAs para modelagem da demanda por transportes. As viagens sintéticas por domicílio foram validadas a partir dos dados desagregados da Pesquisa Origem-Destino (2007) e testes de hipótese para comparação de valores típicos e distribuições populacionais. Em 71% dos setores censitários, as viagens sintéticas foram consideradas similares aos dados reais, confirmando a eficiência do método proposto. Assim, a principal lacuna desta pesquisa, é a apresentação do método sequencial, capaz de tanto minimizar problemas de aquisição de dados quanto atenuar as restrições e suposições matemáticas, inerentes aos modelos tradicionais de demanda por transportes.

ABSTRACT
The estimation of trips per household is essential in the decision-making process related to transportation planning. However, to obtain this estimate, disaggregated data per household is needed, which is usually obtained by an Origin and Destination Survey. Most cities face problems to obtain this data, as this kind of survey needs an amount of time and money to plan and carry it out. Thus, tools for estimation, providing reliable data and low cost, are required. The aim of this paper is to present a sequential method for estimating trips per households using a synthetic population and Artificial Neural Networks (ANNs). The synthetic population was based on aggregated census data and the Monte Carlo Method. The results obtained with ANNs were compared to the results of a traditional linear model and the results were subtly better for ANNs, corroborating their potential in the use of travel demand modeling. The synthetic household trips were validated with the data from the Origin and Destination Survey and hypothesis tests to compare typical values and population distributions. In 71% of the census unit of areas, the synthetic trips were considered similar to the actual data, corroborating the efficiency of the proposed method. Thus, the main research gap is the proposal of the sequential method, capable of minimizing issues of data acquisition and mathematical constraints and assumptions inherent in traditional travel demand forecasting models.

1. INTRODUCÃO

A modelagem tradicional da demanda por transportes pode ser subdividida em quatro etapas: geração de viagens, distribuição de viagens, divisão ou repartição modal e alocação de tráfego (Ortúzar e Willumsen, 2011). Este estudo tratou somente da primeira etapa (geração de viagens), mais especificamente, da estimativa do número de viagens produzidas por domicílio.

Os modelos de demanda por transportes geralmente necessitam de uma base de dados fornecida por Pesquisas de Origem e Destino (OD) para estimativas dos parâmetros envolvidos. Normalmente, são utilizados os dados das pesquisas domiciliares, que fornecem 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 dados socioeconômicos (Silva, 2008; Bruton, 1979; Ortúzar e Willumsen, 2011; Papacostas e Prevedouros, 1993).

Apesar da sua grande importância, a Pesquisa OD normalmente não é realizada com frequência nas cidades, pois é 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. Os microdados do Censo, embora sejam ricos e atendam a várias necessidades de obtenções de dados para análise, não possuem informações relativas aos deslocamentos urbanos (escolha modal; distâncias percorridas; escolhas de destinos ou horários de saída, etc.).

Uma maneira de resolver o problema da falta de dados e da sua periodicidade é a partir de dados sintéticos, geralmente obtidos por meio da geração de uma população sintética – um conjunto de dados referentes a uma população artificial que, em termos estatísticos, representa a população observada. As populações sintéticas podem ser geradas a partir de dados agregados ou desagregados de um censo, por exemplo, tentando obter a maior similaridade possível com a população verdadeira. Mais detalhes sobre métodos de geração de população sintética são encontrados nos trabalhos de Müller e Axhausen (2011); Ma (2011) e Pritchard (2008).

O processo de geração da população sintética preserva a confidencialidade dos indivíduos da amostra quando os dados agregados do censo são utilizados, produzindo atributos e características demográficas realistas (Adiga *et al.,* 2015).

No Brasil, o Censo Demográfico é uma pesquisa realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que disponibiliza, a cada dez anos, uma base de dados que contém informações relativas ao morador e ao uso do solo, agregadas por setores censitários.

A partir de uma base de dados, agregada por unidade de área, disponível do censo, uma população sintética desagregada pode ser gerada por diversas técnicas, sendo que cada uma possui suas limitações. Nos últimos anos, três métodos vêm sendo muito utilizados para a geração deste tipo de população, são eles: (1) Método de ajuste proporcional iterativo (IPF, – Iterative Proportional Fitting), desenvolvido 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 – Combinatorial Optimisation), aplicados nos trabalhos de Williamson *et al*. (1998) e Voas e Williamson (2001).

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 maior 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. O MMC utiliza geradores de números pseudoaleatórios para simular sistemas físicos ou matemáticos, nos quais não se considera o tempo, explicitamente, como uma variável (Chwif e Medina, 2014).

As Redes Neurais Artificiais (RNAs) são aplicáveis em diversas áreas da engenharia, incluindo modelagem de problemas complexos de transportes, haja vista sua grande habilidade de classificar e reconhecer padrões em banco de dados. São modelos matemáticos simples e livres de restrições, resultando em uma ferramenta adaptável, capaz de identificar e aprender padrões nos dados (Adeli, 2001). Dessa forma, uma rede neural pode ser treinada, para aprender a desempenhar uma determinada tarefa. Além disso, as RNAs vêm apresentando um potencial maior que outros métodos estatísticos para estimar e prever dados de demanda por transportes, como apresentado no trabalho de Lin *et al.* (2005) que concluíram que as RNAs têm boa capacidade para estimar tempos de viagem.

O objetivo deste artigo é apresentar um método sequencial para estimativa de viagens domiciliares a partir de uma população sintética e Redes Neurais Artificiais (RNAs).

A população sintética foi obtida através do Método Monte Carlo (MMC), em algoritmo implementado pelos próprios autores. Já o uso das RNAs justifica-se pela ausência de suposições ou restrições matemáticas relativas às variáveis em análise, necessárias para a utilização de Regressão Linear Múltipla (RLM), por exemplo, ferramenta usual de estimativas de viagens domiciliares.

Diante do exposto, este trabalho apresenta uma proposta para estimativa de viagens por domicílios, a partir de uma população sintética, baseada em dados agregados do censo 2010 do IBGE, aplicando o Método Monte Carlo e utilização de Redes Neurais Artificiais (RNAs). Os resultados obtidos com as RNAs foram comparados aos resultados de um modelo linear tradicional.

Em seguida, as viagens domiciliares sintéticas, obtidas a partir do método sequencial proposto, foram validadas utilizando testes de hipótese para comparação de medidas típicas (teste da mediana) e distribuição populacional (Kolmogorov-Smirnov) entre os dados estimados pelo método e dados desagregados da Pesquisa OD. As demais variáveis da população sintética foram validadas com os testes de distribuição populacional (Kolmogorov-Smirnov), fazendo-se comparações pareadas entre microdados do censo e população sintética.

Neste trabalho foram utilizados dados desagregados da Pesquisa OD, realizada em 2007, na cidade de São Carlos, SP, além de dados agregados e microdados provenientes do censo de 2010 (IBGE, 2010).

Assim, a lacuna de pesquisa, associada a este trabalho baseia-se em dois problemas de pesquisa: (1) ausência de dados desagregados devido a fatores econômicos ou necessidade de sigilo e (2) uso de modelos tradicionais para modelagem de demanda por transportes, os quais apresentam várias suposições e restrições matemáticas. Tais suposições normalmente não são atendidas para as variáveis relativas à demanda por transportes. Desta forma, a contribuição deste trabalho baseia-se tanto na capacidade de lidar com ausência de dados desagregados, viabilizando a geração de população sintética, quanto na questão da modelagem de demanda por transportes, através de técnicas de Aprendizagem de Máquinas, considerando os problemas de modelos tradicionais (Regressão Linear, Regressão logística. LogLinear, etc.). No caso da Regressão Linear Múltipla, por exemplo, em que o objetivo principal é encontrar a combinação linear das variáveis independentes que forneça máxima correlação com a variável dependente,

tratar o número de viagens estimadas como variável contínua com suposição de distribuição normal (podendo assumir inclusive valores negativos) é, obviamente, irreal (Schmöcker *et al*., 2005).

Na literatura vigente encontram-se diversos trabalhos que tratam apenas de um dos problemas mencionados anteriormente. Diversos trabalhos, exploram, há anos, diferentes ferramentas para obtenção de dados sintéticos, gerando assim o input a ser utilizado na modelagem (Jain *et al.,* 2015; Moeckel *et al*., 2003; Bowman, 2004; Bowman, 2009; Ye *et al*., 2009; L Ma, 2011).

Outros trabalhos tratam da potencialidade de ferramentas de Aprendizagem de Máquinas, como Redes Neurais Artificiais, na modelagem de demanda por transportes (Pitombo *et al*., 2017; Mozolin *et al.,* 2015; Ichikawa *et al.,* 2002; Xie *et al*., 2003). No entanto, são raros os trabalhos, sobretudo no Brasil, que tratam de ambos os problemas discutidos (Tillema *et al*., 2004).

Em relação a trabalhos que comparam os resultados obtidos com as diferentes técnicas, Redes Neurais Artificiais e Regressões Lineares, especificamente na etapa da geração de viagens, é praticamente unânime, na literatura, o melhor desempenho das RNAs. No entanto, a maior parte dos trabalhos lida com a geração de viagens no nível agregado (Huisken e Coffa, 2000; Openshaw e Openshaw, 1997; Al-Deek, 2001).

As próximas seções deste artigo descrevem as etapas metodológicas, realizadas neste trabalho, bem como descrição teórica das ferramentas utilizadas (MMC e RNAs). A Seção 2 apresenta conceitos e definições teóricas relativas ao método de simulação Monte Carlo. Em seguida, a Seção 3 traz definições teóricas da ferramenta de RNAs. A Seção 4 descreve o banco de dados utilizado e o procedimento metodológico. A simulação da população sintética, a partir de dados agregados do censo por Monte Carlo, é descrita na Seção 5. Já a Seção 6 mostra a aplicação das RNAs para previsão de geração de viagens e na Seção 7 é esboçado o modelo linear calibrado (para fins de validação). A Seção 8 apresenta a validação dos resultados. Finalmente, na Seção 9 são descritas as principais conclusões.

2. MÉTODO MONTE CARLO (MMC)

Observa-se que a simulação por Monte Carlo é amplamente utilizada no campo da Engenharia de Transportes para várias finalidades, entre elas a geração da população sintética. Vários trabalhos ratificam o uso desse método para gerar domicílios sintéticos, dentre eles é importante citar Birkin e Clarke (1988); Beckman, *et al*. (1996), Huang e Williamson (2002); Münnich *et al.* (2003); Namazi-Rad, *et al*. (2014) e Ma e Srinivasan (2015).

O Método Monte Carlo (MMC) é uma parte da matemática experimental que está preocupada em experiências com números pseudoaleatórios e, geralmente, é muito utilizado em modelos complexos, ou não lineares, e uma simulação pode envolver mais de 10.000 replicações do modelo estudado (Hammersley e Handscomb, 1964). Portanto, é uma tarefa difícil e demorada, que no passado só poderia ser realizada por eficientes computadores. 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 (Escudero, 1973).

A execução do MMC acontece de acordo com as seguintes etapas:

1. Primeiramente, são determinadas as distribuições das características de indivíduos da amostra de uma população real;

2. Em seguida, são gerados, aleatoriamente, indivíduos "virtuais", respeitando as distribuições das características da população real, até formar uma população sintética do tamanho da população real;
3. Para a validação da população sintética, uma técnica de amostragem é aplicada, obtendo amostras da população sintética, e calculado o valor da estatística de interesse (para cada amostra) para garantir a aderência das distribuições das características;
4. Todo o processo de geração e amostragem é repetido um número "N" suficiente de vezes, até garantir um intervalo de confiança de 95% na representação das distribuições das características dos indivíduos (Mackay, 1996).

O uso de modelo matemático, para descrever um sistema, pode ser complexo ou até mesmo sem uma solução analítica. Desta forma, o uso da simulação computacional é uma ferramenta de grande valor na obtenção de uma resposta de um problema particular. A simulação por Monte Carlo tem sido utilizada para modelar uma grande variedade de fenômenos, e evoluiu para várias variantes bem estabelecidas após desenvolvimento computacional.

3. REDES NEURAIS ARTIFICIAIS

Existem diversas técnicas de modelagem de dados e produção de informações que buscam simular a inteligência humana para resolver problemas complexos, tais como: Lógica Nebulosa, Sistemas Especialistas, Redes Neurais Artificiais, entre outras. Uma das técnicas de Inteligência Artificial (ou Aprendizagem de Máquinas) mais promissora é a Rede Neural Artificial (RNA), técnica que consegue reproduzir o comportamento de qualquer função matemática, normalmente aquelas não-lineares (Smith, 1996).

As RNAs têm como objetivo compreender o funcionamento do cérebro humano e, de alguma forma, reproduzi-lo, pois ambos compreendem grande número de interconexões das unidades de processamentos não lineares chamados de neurônios, que apresentam como principal função o armazenamento e a disponibilidade de informação. O seu funcionamento é baseado em sistemas de equações, em que o resultado de uma equação é o valor de entrada para várias outras da rede (Dougherty, 1995).

Segundo Haykin (1999) o neurônio forma a base para as RNAs, sendo o neurônio artificial o objeto que simula o comportamento do neurônio biológico, uma unidade de processamento matematicamente simples. Os neurônios se comunicam através de sinapses (região onde dois neurônios entram em contato e através do qual os impulsos nervosos são transmitidos entre eles). Cada neurônio recebe uma ou mais entradas, que correspondem às conexões sinápticas com outras unidades similares a ele, com seus respectivos pesos, transformando em saídas, cujos valores dependem diretamente da somatória ponderada de todas as saídas dos outros neurônios a esse conectados.

A Figura 1 apresenta o modelo de um neurônio, cujos três elementos básicos que formam a base para o projeto de uma Rede Neural Artificial são os pesos sinápticos, a função soma e a função de ativação.

Uma Rede Neural Artificial é composta por várias unidades de processamento, porém o funcionamento é simples, ou seja, as unidades são conectadas por canais de comunicação que estão associados a um peso e fazem operações sobre os dados locais. Em resumo, a operação de uma unidade de processamento acontece com os sinais sendo apresentados à entrada, cada sinal é multiplicado por um número (peso) que indica a sua influência na saída da unidade.

Após, é realizada a soma ponderada dos sinais que produz um nível de atividade (*threshold*) e este nível irá determinar a resposta de saída.

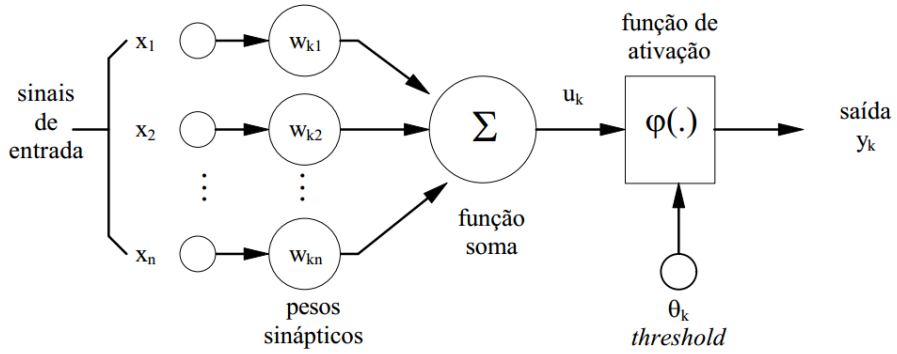


Figura 1. Modelo não linear de um neurônio. Fonte: Adaptado de Haykin (1999)

De acordo com Haykin (1999), o funcionamento da Rede Neural Artificial se realiza através do padrão de conexão entre várias camadas das redes, os números de neurônios em cada camada, a capacidade da aprendizagem e as funções de ativação. Os neurônios de uma rede neural são estruturados de acordo com o algoritmo de aprendizado usado para treinar a rede. Assim, os dois aspectos principais de construção de uma rede neural são: a arquitetura e o aprendizado.

A arquitetura neural é tipicamente organizada em camadas, com unidades que podem estar conectadas às unidades da camada posterior. Essas camadas são classificadas em camada de entrada (padrões são apresentados à rede), camadas intermediárias ou escondidas (realizado a maior parte do processamento) e a camada de saída (resultado final é concluído). Segundo Wasserman (1989) a entrada não é considerada uma camada da rede, pelo fato de apenas distribuir os padrões. A camada com os neurônios, que fornecem a saída da rede, é chamada camada de saída. A arquitetura se apresenta em três diferentes tipos: as redes progressivas de única camada, as redes progressivas de camadas múltiplas e as redes recorrentes. As redes progressivas de camadas múltiplas, *Multilayer Perceptron* (MLP), correspondem a um processador paralelo, constituído de neurônios (unidades de processamento), que são dispostos em uma ou mais camadas interligadas por muitas conexões.

A habilidade de aprender é uma propriedade valiosa das redes neurais. O aprendizado da rede MLP é denominado de treinamento e ocorre através do ajuste dos pesos, utilizando algum algoritmo de treinamento, como por exemplo, o algoritmo *backpropagation* uma técnica de aprendizado supervisionado que utiliza pares (entrada e saída desejada) para, através do cálculo do erro, ajustar os pesos da rede e adquirir conhecimento (Haykin, 1999).

4. MATERIAIS E MÉTODO

Os dados utilizados neste trabalho são referentes à Pesquisa OD, realizada na cidade de São Carlos-SP entre os anos de 2007 e 2008, pelo Departamento de Engenharia de Transportes (EESC-USP), e a pesquisa do Censo Demográfico 2010, realizada pelo Instituto de Geografia e Estatística (IBGE).

A cidade de São Carlos está localizada no centro do estado de São Paulo, 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 é considerada um centro predominantemente urbano, contando apenas com 4% da população residente na área rural (IBGE, 2010). Assim, como outras cidades brasileiras de médio porte, São Carlos apresentou um intenso crescimento econômico nos últimos anos, com uma renda per capita média de R$1.086,22 e um alto nível IDH de 0,805 no ano de 2010. Como exemplo, o estado de São Paulo e o país apresentaram IDH para o mesmo ano de 0,783 e 0,813, respectivamente (PNUD, 2016) e (IBGE, 2010).

Três bases de dados distintas foram utilizadas segundo o procedimento metodológico e apresentam a seguinte amostragem: (i) dados agregados do Censo IBGE - 2010: 288 observações-setores censitários, (ii) dados desagregados da Pesquisa OD: contendo 3.057 observações-domicílios e (iii) microdados do Censo IBGE-2010: 6.817 observações-domicílios.

A sequência metodológica, associada a este trabalho, baseia-se em: (1) Tratamento das bases de dados; (2) Obtenção da população sintética; (3) Calibração do modelo de geração de viagens; (4) Estimativa de viagens domiciliares sintéticas; (5) Validação 1: Validação das variáveis independentes sintéticas e (6) Validação 2: Validação das viagens domiciliares sintéticas. O Tabela 1 1 sintetiza o banco de dados utilizado em cada uma das etapas metodológicas.

Tabela 1 – Síntese do uso das bases de dados das diferentes etapas metodológicas

Etapa metodológica	Base de dados
Etapa 1: Tratamento dos dados	Três bases de dados
Etapa 2: Obtenção da população sintética	Dados agregados do Censo IBGE - 2010: 288 observações-setores censitários
Etapa 3: Calibração dos modelos de geração de viagens	Dados desagregados da Pesquisa OD: contendo 3.057 observações-domicílios
Etapa 4: Estimativa de viagens domiciliares sintéticas	População sintética (variáveis independentes)
Etapa 5: Validação 1: Validação das variáveis independentes sintéticas	Microdados do Censo IBGE-2010: 6.817 observações-domicílios
Etapa 6: Validação 2: Validação das viagens domiciliares sintéticas	Dados desagregados da Pesquisa OD: contendo 3.057 observações-domicílios

A Tabela 2, em seguida, traz uma síntese das descrições das variáveis utilizadas em cada uma das três bases de dados. Vale ressaltar que foram descritas apenas as variáveis aqui utilizadas. Como um dos objetivos deste trabalho é simular uma população sintética por meio de dados agregados do censo, foram escolhidas variáveis que delineavam as características sociodemográficas dessa população. No entanto, por se tratar de um estudo de geração de viagens por domicílio foram selecionadas as variáveis independentes que explicam o fenômeno, segundo literatura conhecida (Novaes, 1986; Bruton, 1979; Kawamoto, 1994; Ortúzar e Willumsen, 2011; Papacostas e Prevedouros, 1993) e a variável dependente *viagens por domicílio*, esta somente obtida pela Pesquisa OD. Vale ressaltar que algumas variáveis da amostra desagregada da Pesquisa OD foram transformadas (numéricas para *dummies*) no intuito de serem compatibilizadas às amostras de microdados e população sintética.

Na etapa de calibração de modelo de geração de viagens por RNAs foi utilizado o banco de dados desagregados da Pesquisa OD e foram utilizadas cinco variáveis independentes: "Tamanho do domicílio", "Renda domiciliar", "Gênero", "Posição na família", "Faixa etária" e a variável dependente "Viagens domiciliares". As variáveis "Tamanho do domicílio" e "Renda Domiciliar" foram utilizadas, na sua forma dicotômica, obedecendo as suas categorias, como por exemplo, a variável "Renda Domiciliar" dividida em 6 categorias (Domicílio sem renda, Domicílio com

renda de 0-2 SM, Domicílio com renda de 2-3 SM, Domicílio com renda de 3-5 SM, Domicílio com renda de 5-10 SM e Domicílio > 10 SM). As variáveis "Gênero" (Quantidade de pessoas no domicílio do gênero feminino e do gênero masculino), "Posição na família" (Quantidade de pessoas no domicílio com a posição de chefes, cônjuges, filhos, etc.) e "Faixa etária" (Quantidade de pessoas no domicílio com idade até 10 anos, de 11 a 20 anos, etc.) foram utilizadas na sua forma numérica. A Tabela 3 sintetiza as categorias das variáveis, tanto em sua forma dicotômica, quanto em sua forma numérica (referindo-se a contagens).

Tabela 2 – Descrição das variáveis utilizadas em cada base de dados

Base de dados	Variáveis/Descrição
Dados agregados do Censo IBGE - 2010	Quantidade de domicílios com X moradores por setor; Quantidade de mulheres e homens nos domicílios por setor; Quantidade de chefes, cônjuges, etc. nos domicílios por setor; Quantidade de moradores com X anos nos domicílios por setor; Quantidade de domicílios com rendimento médio mensal de X SM por setor.
Dados desagregados da Pesquisa OD	Quantidade de viagens produzidas por domicílio (variável dependente); Variáveis independentes: Domicílio com 1 morador, Domicílio com 2 moradores, Domicílio com 3 moradores, etc. (1) SIM e (0) NÃO; Domicílio com renda de 0 a 2 salários mínimos, Domicílio com renda de 2 a 4 salários mínimos, etc: (1) SIM e (0) NÃO; Quantidade de pessoas no domicílio do gênero feminino e do gênero masculino; Quantidade de pessoas no domicílio com a posição de chefes, cônjuges, filhos, etc.; Quantidade de pessoas no domicílio com idade até 10 anos, de 11 a 20 anos, etc.
Microdados do Censo IBGE-2010	Domicílio com 1 morador, Domicílio com 2 moradores, Domicílio com 3 moradores, etc. (1) SIM e (0) NÃO; Domicílio sem renda, Domicílio com renda de 0 a 2 salários mínimos, Domicílio com renda de 2 a 4 salários mínimos, etc.: (1) SIM e (0) NÃO; Quantidade de pessoas no domicílio do gênero feminino e do gênero masculino; Quantidade de pessoas no domicílio com a posição de chefes, cônjuges, filhos, etc.; Quantidade de pessoas no domicílio com idade até 10 anos, de 11 a 20 anos, etc.

Tabela 3 – Descrição, por categorias, das variáveis da Pesquisa OD utilizadas para calibração dos modelos de geração de viagens

Variáveis	Tipo	Descrição das variáveis		Ref.
Tamanho do domicílio	*Dummies*	Domicílios com 1 morador, Domicílios com 2 moradores, ..., Domicílios com 10 moradores	10 categorias	Dom com 1 morador
Renda Domiciliar		Domicílios sem renda, Domicílios com renda de 0 a 2 salários mínimos SM,..., Domicílios com renda acima de 10 SM	6 categorias	Dom sem saída
Gênero	Numéricas discretas	Quantidade de pessoas no domicílio do gênero feminino e do gênero masculino.	2 categorias	
Posição na Família		Quantidade de pessoas no domicílio com a posição de chefes, cônjuges, filhos, empregados, parentes, agregados e visitantes.	7 categorias	Visitantes
Faixa Etária		Quantidade de pessoas no domicílio com idade até 10 anos, de 11 a 20 anos, de 21 a 30 anos,..., Acima de 80 anos	9 categorias	Idade até 10 anos
Viagens Domiciliares		Quantidade de viagens produzidas por domicílio.	-	

A partir dos bancos de dados definidos, a simulação da população sintética foi codificada em *Visual Basic Application* (VBA), linguagem de programação na planilha *Microsoft Excel* do pacote

Office, enquanto as estimativas da variável dependente dos modelos através das ferramentas (Redes Neurais Artificiais e Regressão Linear Múltipla), realizadas no *software* IBM SPSS, versão 24.

A Figura 2 representa o procedimento metodológico realizado no presente trabalho. As etapas do método são sucintamente descritas em seguida. Vale ressaltar que foi realizada uma análise de sensibilidade para a escolha das proporções da amostra de treinamento e teste para o modelo de RNAs, visto que não há um consenso na literatura sobre os tamanhos dos conjuntos de calibração e validação. Os conjuntos testados foram: 60/40, 70/30 e 80/20 para calibração e validação, respectivamente. Após a análise de sensibilidade das proporções, ficou definido 60% da amostra para treinamento da RNA e calibração do modelo linear. Os dados restantes foram utilizados para teste da rede e validação do modelo paramétrico.

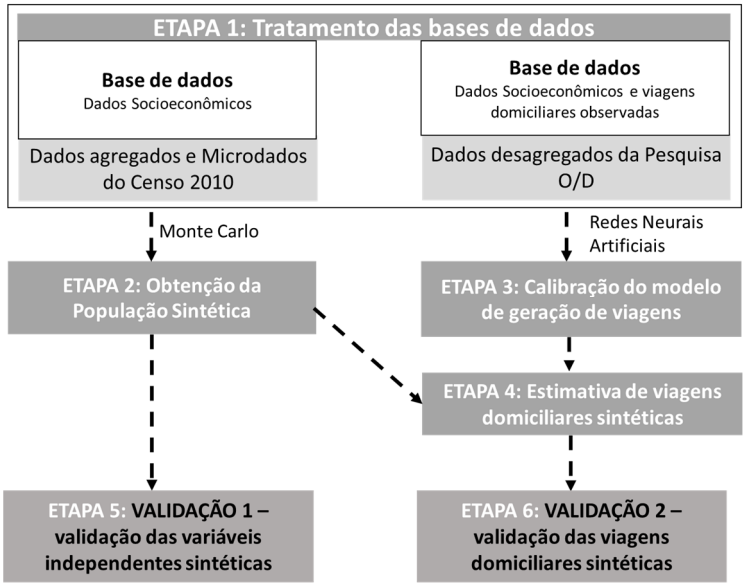


Figura 2. Ilustração das etapas do procedimento metodológico

Etapa 1 - Tratamento dos dados: Inicialmente foi realizado o tratamento da base de dados agregados por setor censitário do IBGE. Foram utilizadas as distribuições populacionais das variáveis da amostra agregada para simulação de dados sintéticos desagregados. Simultaneamente, foi realizado o tratamento de dados da amostra desagregada, por domicílio, dos Microdados e da Pesquisa OD. O banco de dados da Pesquisa OD foi utilizado tanto para calibração do modelo de RNAs, quanto para calibração do modelo linear tradicional.

Etapa 2 - Obtenção da população sintética: Esta etapa metodológica compreende a simulação dos dados sintéticos (descrita detalhadamente na seção seguinte).

Etapa 3 - Calibração dos modelos de geração de viagens: consiste na calibração de modelo de RNAs, com dados da pesquisa OD, para estimativa de viagens domiciliares.

Etapa 4 - Estimativa de viagens domiciliares sintéticas: Com a utilização das variáveis da população sintética, obtida anteriormente pelo MMC (Etapa 2), no modelo de RNAs (Calibrado na Etapa 3), são estimadas as viagens domiciliares sintéticas.

Após realizadas as etapas de 1 a 4, finalizou-se com as etapas de validação, sendo a Etapa 5 a validação 1 (Validação das variáveis independentes sintéticas) e a Etapa 6 a Validação 2 (Validação das viagens domiciliares sintéticas). As etapas de validação, tanto da população sintética

(variáveis explicativas – Validação 1) quanto das viagens estimadas sintéticas (Validação 2), foram realizadas através de testes de hipótese.

Há duas validações e objetivos aparentes, conforme pode-se observar na Figura 3: (1) Validação das variáveis independentes obtidas pela população sintética, com base nas distribuições populacionais dos dados agregados (População sintética); (2) Validação das variáveis de viagens estimadas pelas RNAs e População sintética (Viagens domiciliares sintéticas).

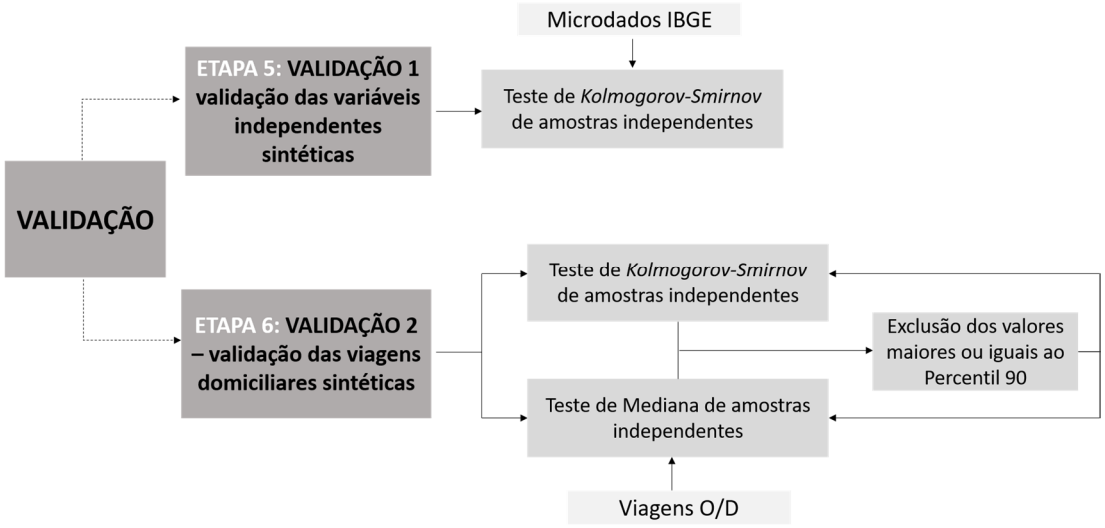


Figura 3. Detalhamento da validação metodológica

Etapa 5 - Validação 1: Validação das variáveis independentes sintéticas

Após a obtenção dos valores desagregados das variáveis independentes para população sintética, é necessário compará-los a uma amostra desagregada observada. Assim a amostra utilizada como base comparativa é proveniente dos Microdados do IBGE.

A ferramenta utilizada é o teste *Kolmogorov-Smirnov* para comparação das distribuições populacionais da amostra sintética e dos microdados do IBGE. Assim, são realizadas comparações pareadas entre as variáveis independentes, obtidas pela população sintética, e as variáveis independentes observadas (Microdados). Vale ressaltar que o teste *Kolmogorov-Smirnov* é realizado na sua forma não paramétrica, com objetivo de comparar duas distribuições populacionais desconhecidas.

Etapa 6 - Validação 2: Viagens domiciliares sintéticas

Como os dados relativos a viagens não são encontrados no censo de 2010 do IBGE, a amostra base para fins comparativos aqui utilizada é a amostra desagregada, proveniente da Pesquisa O/D. Assim, a partir das variáveis independentes, obtidas para a amostra sintética desagregada (Etapa 2) e do modelo de RNAs previamente calibrado (Etapa 3), são estimadas as viagens domiciliares, chamadas neste trabalho de VIAGENS SINTÉTICAS (Etapa 4).

Para validação das viagens estimadas sintéticas, foi feito teste de hipótese *Kolmogorov- Smirnov* para comparação das distribuições populacionais das viagens estimadas sintéticas e das viagens domiciliares da Pesquisa OD.

Também foi realizado o teste da Mediana para comparação da similaridade das medianas entre os mesmos pares de valores (para duas amostras independentes, também de distribuição desconhecida). A Figura 3 ilustra detalhadamente a etapa de validação dos dados. A validação

das viagens é feita para amostras completas e também após a retirada do percentil 90. O intuito foi retirar, da análise de validação, aqueles domicílios atípicos da Pesquisa OD, com a realização de 24 viagens domiciliares, por exemplo.

Vale ressaltar que os testes foram realizados considerando os domicílios contidos em cada um dos setores censitários. Observou-se, posteriormente, o percentual de setores que passaram em ambos os testes de hipóteses, o percentual de setores que passaram em apenas um dos testes, além do percentual de setores que não passaram em nenhum dos testes propostos.

5. SIMULAÇÃO DOS DADOS SINTÉTICOS PELO MÉTODO MONTE CARLO

O algoritmo de Monte Carlo, que utiliza uma base estatística para gerar números pseudoaleatórios, foi utilizado para simular a população sintética (domicílios sintéticos) e, resumiu-se em três etapas, de acordo com a Figura 4. 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. Vale ressaltar que as distribuições utilizadas no método Monte Carlo foram obtidas das informações agregadas do Censo. Ressalta-se que a distribuição populacional, de algumas variáveis agregadas, é Normal, Poisson ou Exponencial. Algumas delas, no entanto, não apresentam nenhuma distribuição conhecida (foram testadas Normal, Poisson, Uniforme e Exponencial).

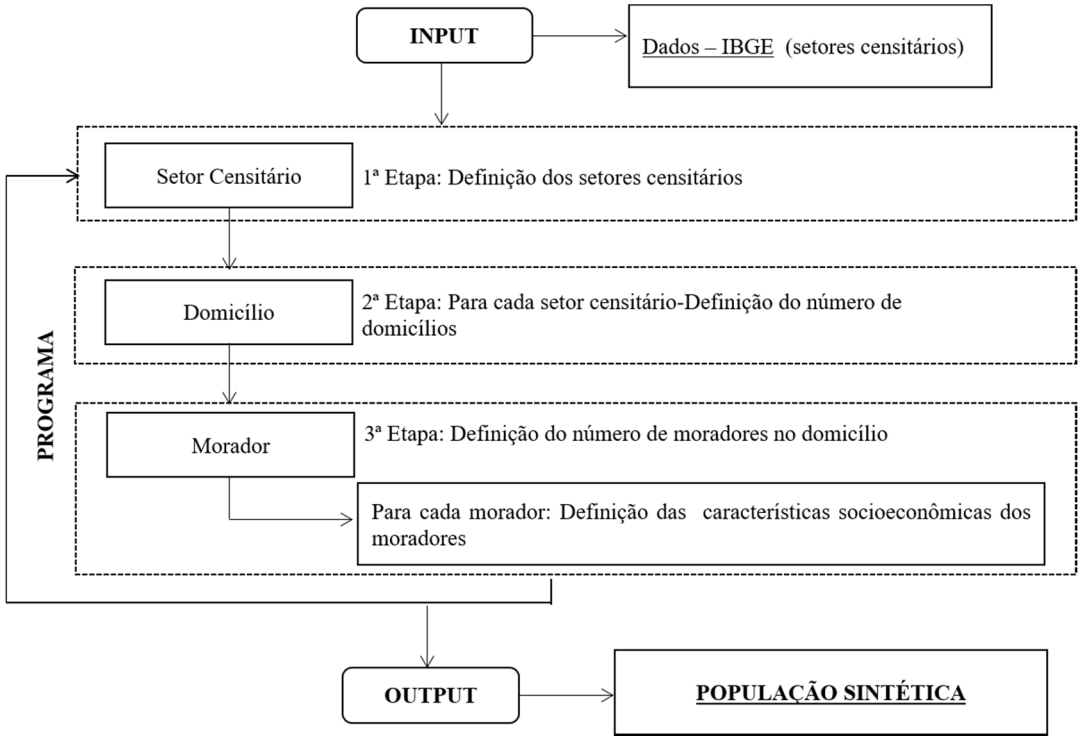


Figura 4. Fluxograma da simulação da população sintética

Com o propósito de avaliar os resultados obtidos pelo Método Monte Carlo, foi realizada a validação dessa população sintética (Etapa 5: Validação 1 – Figura 3) através da amostra de microdados 2010 do IBGE. Optou-se por utilizar o teste estatístico para as amostras indepen-

dentes (na sua forma não paramétrica, ou seja, para variáveis de distribuição populacional desconhecida), o teste de *Kolmogorov - Smirnov*. Da forma que o teste foi realizado neste trabalho, ele compara a distribuição populacional de duas amostras independentes. Assim, foram comparadas, de forma pareada, as distribuições populacionais das variáveis da população sintética e das variáveis originais que compõem os microdados do censo 2010.

Neste teste, a Hipótese Nula seria que as variáveis apresentam distribuições similares e a Hipótese Alternativa seria que as variáveis não apresentam distribuições populacionais similares. Assim, a Hipótese Nula foi retida para as variáveis apresentadas na Tabela 4.

Tabela 4 - Variáveis que passaram no teste de *Kolmogorov-Smirnov*

Variáveis Ho=Reter	p-valor	Variáveis Ho=Reter	p-valor
Domicílios com 1 morador	0,985	Chefes	1,000
Domicílios com 2 moradores	0,999	Agregados	1,000
Domicílios com 3 moradores	1,000	Visitantes	1,000
Domicílios com 4 moradores	1,000	Empregados	1,000
Domicílios com 5 moradores	1,000	Idade 71-80 anos	0,596
Domicílios com 6 moradores	1,000	Idade mais que 80 anos	1,000
Domicílios com 7 moradores	1,000	Renda Domiciliar (sem renda)	1,000
Domicílios com 8 moradores	1,000	Renda Domiciliar (2-3 sal.min.)	0,479
Domicílios com 9 moradores	1,000	Renda Domiciliar (3-5 sal.min.)	0,404
Domicílios com 10 ou mais moradores	1,000	Renda Domiciliar (5-10 sal.min.)	0,537
		Renda Domiciliar (>10 sal.min.)	0,768

6. REDES NEURAIS ARTIFICIAIS PARA PREVER VIAGENS POR DOMICÍLIO

A Rede Neural Artificial (RNA) foi utilizada para prever a geração de viagens por domicílio. O modelo RNA de geração de viagens gerado iniciou-se pela escolha do tipo de rede a ser treinada e optou-se por utilizar a rede *multilayer perceptron* (MLP-Figura 5) que é função de variáveis de previsão (independentes) que minimizam o erro de predição da variável de saída (Haykin, 1999). Em seguida, foram definidas como variáveis de entrada 5 covariáveis - (duas na sua forma dicotômica e 3 na forma numérica, todas distribuídas conforme categorias, como apresentado na Tabela 2 e Tabela 3).

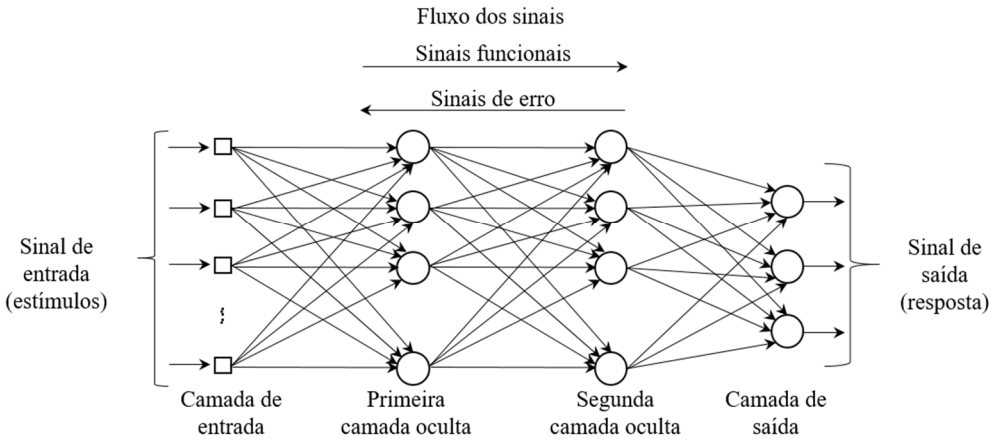


Figura 5. Esboço de uma rede MLP. Fonte: Haykin, (1999)

O processamento da rede para prever a geração de viagens por domicílio foi definido utilizando uma arquitetura personalizada de 2 camadas ocultas, pois foram realizados testes preliminares e não foi observado ganho de desempenho com mais de duas camadas ocultas. Além

disso, o aumento do número de camadas ocultas pode acarretar o aumento da complexidade e do tempo de processamento da rede (Batista, 2012).

A amostra de treinamento da rede foi equivalente a 1.836 domicílios (60%), enquanto a amostra de teste foi de 1.221 domicílios (40%). O processamento em cada neurônio se deu pelas funções de ativação tangente hiperbólica e identidade, pois são funções padrão; além disso, nos testes realizados com as outras funções, estas foram as que apresentaram menor erro. A função de ativação tangente hiperbólica foi adotada com sucesso em outros trabalhos, assim como neste (Aguiar Júnior, 2004; Gonçalves et. al, 2014; Gonçalves et. al, 2015). A função de erro definida para ser minimizada foi a soma dos erros quadráticos.

Turchenko *et al.* (2010) afirmam que as Redes Neurais Artificiais representam uma boa alternativa para a solução de problemas complexos, porém na fase de treinamento requerem grande carga computacional, que pode demorar horas ou até dias para convergir. Assim, para não ter problemas de sobrecarga computacional, os autores sugerem o uso do algoritmo de treinamento por lote, pois a atualização dos pesos sinápticos é realizada após o processamento de todos os padrões de treinamento ao invés de atualizar os pesos a cada padrão apresentado e foi observada uma aceleração positiva. Desta forma, foi selecionado o treinamento do tipo Lote que atualiza os pesos sinápticos aos passar por todos os registros de dados de treinamento, continua o processo de atualização dos pesos e só para quando uma das regras de parada seja atendida.

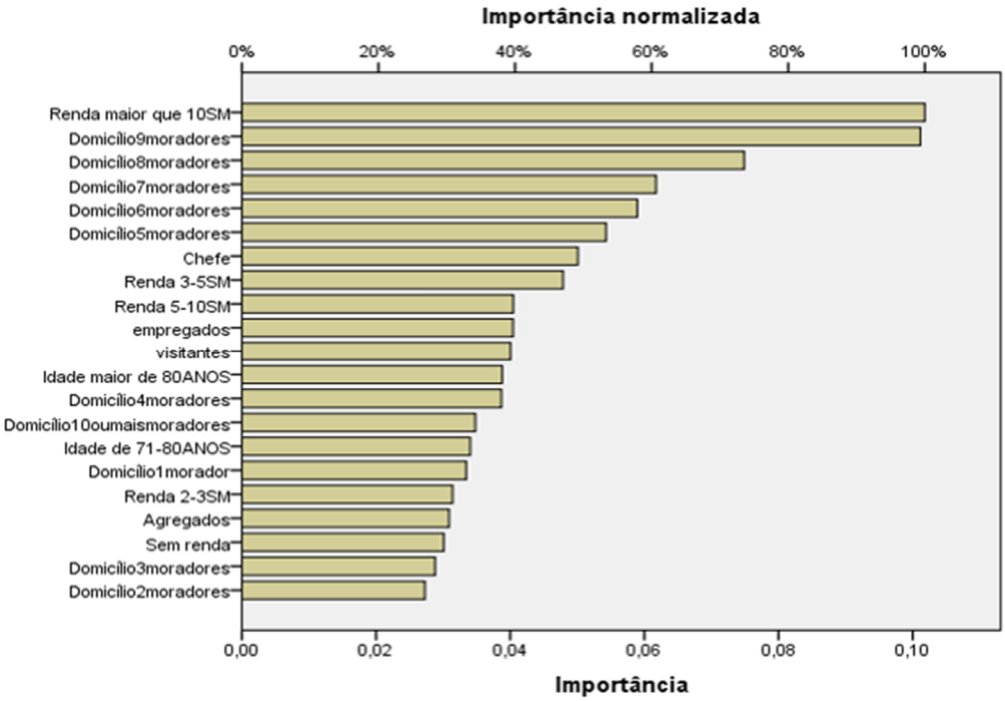


Figura 6. Importância das variáveis analisadas para a geração de viagens por domicílio

A RNA definiu como principais variáveis explicativas da geração de viagens por domicílio as variáveis Domicílio com Renda Domiciliar maior que 10 salários mínimos (dicotômica) e Domicílio com 9 moradores – Dicotômica - (Figura 6). Em seguida, as variáveis Domicílio com 8 moradores – Dicotômica - e Domicílio com 7 moradores – Dicotômica - foram classificadas como

terceira e quarta variáveis explicativas de maior relevância, respectivamente. Observa-se, na literatura tradicional de demanda por transportes, a alta correlação positiva entre o número de indivíduos no domicílio e a quantidade de viagens domiciliares, bem como Renda Domiciliar e quantidade de viagens, sobretudo motorizadas (Ashley, 1978; Atherton e Ben-Akiva, 1976; Bates *et al.,* 1978). Espera-se assim, uma importância maior das variáveis dicotômicas que representam domicílios de renda alta (Domicílio com Renda maior que 10 SM) e domicílios com alto número de membros (Domicílio com 9 moradores, Domicílio com 8 moradores e Domicílio com 7 moradores). Além disso, considerando os resultados da RNA, ressalta-se que a função escolhida tende a minimizar o erro de predição da variável de saída. Por esta razão, encontram-se, nas categorias extremas superiores, os menores erros (valores 1 para categorias extremas superiores, inevitavelmente, produzem maiores valores de viagens domiciliares). Sendo esperado tal resultado para o caso de Redes Neurais, sobretudo para amostra de treinamento, pois são escolhidas aquelas variáveis explicativas que produzem valores de variável dependente com menor variabilidade.

Os resultados tanto do treinamento (60% da amostra) quanto do teste (40% da amostra) foram bons, com erros considerados satisfatórios para o propósito da pesquisa. Com o propósito de avaliar o desempenho dos modelos investigados (RNA *multilayer perceptron* e modelo linear tradicional), foram consideradas algumas medidas de desempenho dos erros obtidos. Tais medidas estão descritas, a seguir, e os resultados obtidos são apresentados na Tabela 5.

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

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

$$ER = \frac{\sum_{i=1}^{N} (x_i - y_i)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{3}$$

Sendo: *EM* = Erro Médio; x_i = valor observado; y_i = valor previsto; *REM* = Raiz do erro médio; *ER* = Erro relativo;

$$r = \frac{1}{N = r} \times \frac{(x_i - \bar{x}) - (y_i - \bar{y})}{\sigma_x \times \sigma_y} \tag{4}$$

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

$$Desvpad\ (x - y) = \sqrt{E(x - y) - \mu^2} \tag{5}$$

Sendo: *D* esvpad (x−y) = desvio padrão dos erros; μ = média dos erros; (x−y) = erro.

Tabela 5 - Medidas de desempenho calculadas para RNA

Amostra	Treino	Teste
Erros	60%	40%
EM	-0,055	0,176
REM	2,970	3,640
ER	0,521	0,662
r	0,692	0,583
DESVPADA	2,970	3,638

Estabelecido o modelo na RNA, foi possível verificar os resultados da previsão da geração de viagens por domicílio. Além da saída com os valores previstos, o IBM SPSS 24.0 apresenta, como resultado, o gráfico de dispersão entre valores observados e previstos da variável dependente. Desta maneira, há uma forma eficiente e visual de verificar a qualidade dos resultados obtidos a partir das RNAs. A Figura 7 ilustra este resultado. Constatou-se que os resultados do modelo de RNAs apresentaram uma boa relação entre os dados observados e estimados (Coeficiente de *Pearson*=0,583).

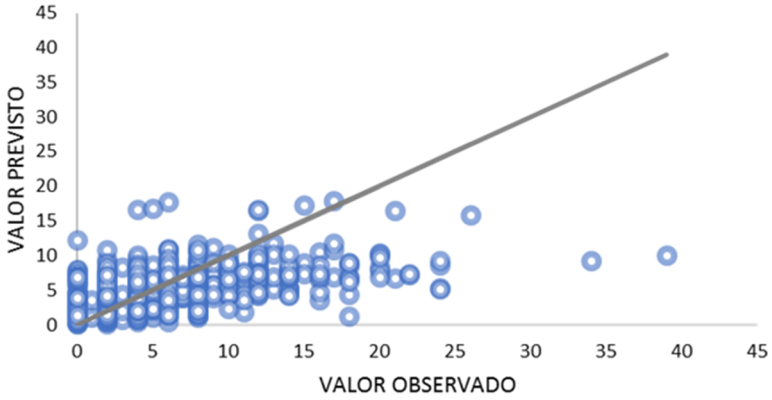


Figura 7. Gráfico de dispersão da variável geração de viagens por domicílio (Observados x estimados por RNAs)

7. MODELO LINEAR CALIBRADO (REGRESSÃO LINEAR MÚLTIPLA)

A Regressão Linear Múltipla (RLM) é aplicada em uma infinidade de casos, buscando uma relação entre uma única variável dependente (numérica) e diversas variáveis independentes (numéricas ou *dummy*), relação esta supostamente linear. A finalidade é encontrar a combinação linear das variáveis independentes que forneça a máxima correlação com a variável dependente (Raymond, 1982). A calibração do modelo de geração de viagens domiciliares (Etapa 3), bem como obtenção das viagens sintéticas, são baseados no algoritmo de RNAs. A calibração através da RLM *Stepwise* foi realizada para efeitos comparativos. Os autores sugerem a calibração de uma técnica comumente empregada para o propósito de testar a potencialidade do uso de RNAs para estimativas de viagens domiciliares.

O procedimento *Stepwise* constrói, iterativamente, uma sequência de modelos de regressão pela adição ou remoção de variáveis em cada etapa, sendo definido pelo teste parcial F (Freedman, 2009). Inicialmente, foi utilizado o banco de dados original com 5 variáveis independentes (34 categorias de variáveis conforme apresentado no Quadro 3) e através do método *Stepwise*, foi selecionado o melhor modelo (maior coeficiente de determinação), com 13 categorias de variáveis independentes selecionadas e 14 parâmetros estimados (incluindo a constante). Em seguida, foi realizada, juntamente com a matriz de correlação das 13 categorias de variáveis selecionadas, a análise de multicolinearidade deste modelo utilizando os valores de tolerância e VIF (Tabela 6).

A tolerância é uma medida direta de multicolinearidade e 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 2,00 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 (Hair *et al.*, 2009).

Tabela 6 - Análise de multicolinearidade

Variáveis	Tolerância	VIF
Dom. com 5 moradores	0,881	1,135
Dom. com 4 moradores	0,865	1,156
Idade -71 a 80 anos	0,958	1,043
Chefe	0,938	1,066
Idade -61 a 70 anos	0,943	1,060
Cônjuges	0,868	1,151
Outros parentes	0,917	1,090
Idade > 80 anos	0,926	1,080
Renda 10-20 SM	0,973	1,028
Renda 3-5 SM	0,985	1,015
Renda 2-3 SM	0,958	1,044
Renda 5-10 SM	0,979	1,021
Agregados	0,981	1,020

Após esta análise, detectou-se multicolinearidade mínima nos dados, pois as variáveis não apresentaram valor de tolerância menor que 0,750 e valores de VIF próximos de 1,000. O modelo escolhido está descrito na Tabela 7.

Tabela 7 - Principais resultados do modelo linear escolhido

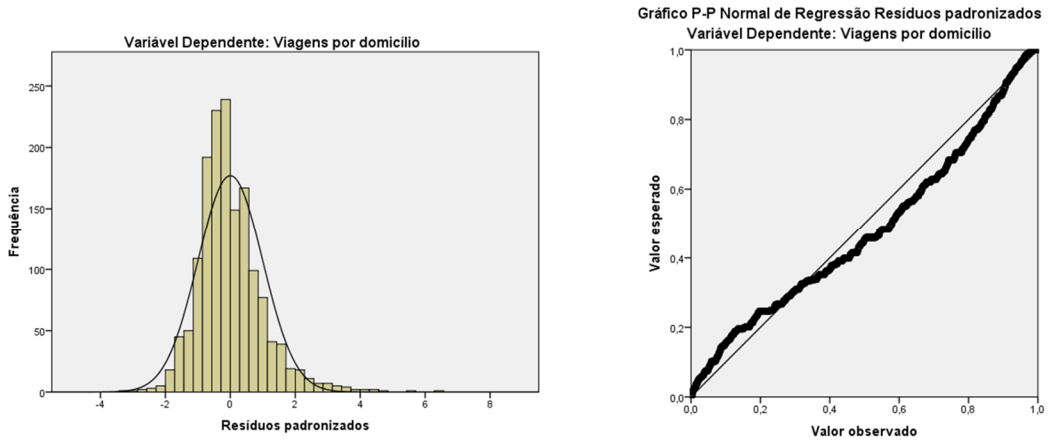
Modelo linear: Geração de viagens por domicílio		
Variáveis significativas	R^2 = 0,317	
	Coeficientes	t
Constante	1,266	4,123
Dom. com 5 moradores	3,082	10,817
Dom. com 4 moradores	1,785	8,151
Idade -71 a 80 anos	-1,576	-7,480
Chefe	1,614	7,644
Idade -61 a 70 anos	-1,322	-8,676
Cônjuges	1,475	7,422
Outros parentes	0,877	7,585
Idade > 80 anos	-1,811	-5,955
Renda 10-20 SM	2,446	5,204
Renda 3-5 SM	1,018	4,473
Renda 2-3 SM	0,591	4,711
Renda 5-10 SM	1,462	4,304
Agregados	1,113	2,093

Por fim, foi realizada uma análise crítica do modelo, em relação às variáveis independentes significativas selecionadas pelo método estatístico e se observa que os valores de coeficientes da maioria das variáveis independentes foram positivos, inclusive a constante. Portanto, é possível afirmar que as viagens aumentam com o aumento do número de moradores no domicílio. Da mesma maneira, quanto maior a renda no domicílio maior o número de viagens realizadas. Novamente, o número de viagens aumenta quanto maior o número de chefes, cônjuges e agregados que o domicílio possui.

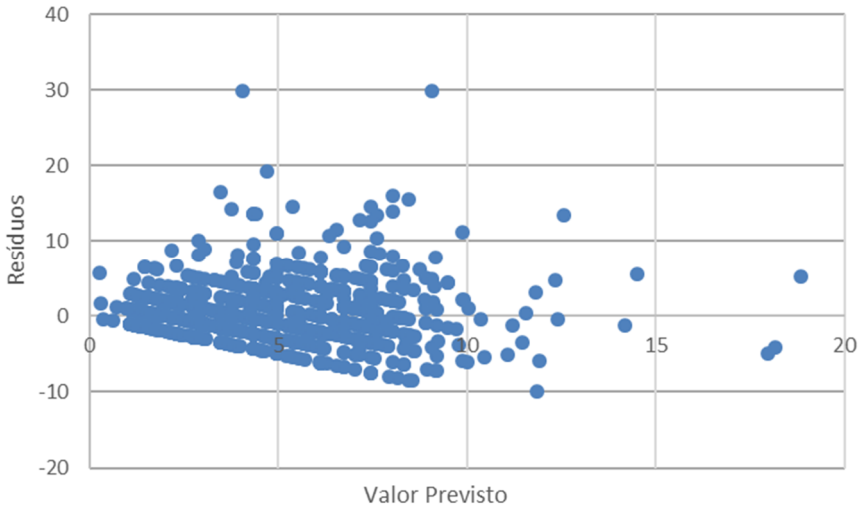
A ordem de grandeza dos parâmetros estimados para explicar o fenômeno de previsão de viagens por domicílio e os valores da estatística t apresentaram valores coerentes com o esperado. Tudo isto denotou que as variáveis selecionadas neste modelo (RLM) foram significativas. As variáveis significativas são coerentes com aquelas apontadas por Ortúzar e Willumsen (2011), que explicam a geração de viagens, através de análise desagregada: renda; posse do carro; tamanho da família e estrutura familiar.

Desta forma, como se tratam de variáveis dicotômicas (exceto a variável "Chefe" que é a quantidade de chefes no domicílio), é possível afirmar que as viagens por domicílio aumentam com o aumento de moradores e também com o aumento da renda familiar, conforme encontrado na literatura tradicional de estimativa desagregada de demanda por transportes (Clifton *et al*, 2012; Ewing *et al*., 1996; Vickerman e Barmby, 1985).

Após análise crítica do modelo obtido, os resíduos do modelo linear foram plotados e ilustrados na Figura 8 e observa-se que atendem à suposição de normalidade (a), porém não atendem à suposição de homocedasticidade (b). Em seguida, os autores apresentam um diagrama de dispersão contendo os valores de viagens observados e estimados pelo modelo de RLM (Figura 9).



(a)



(b)

Figura 8. (a) Análise de normalidade, (b) Análise dos resíduos – heteroscedasticidade

Assim como nas RNAs, para a RLM foram calculadas as medidas de desempenho, apresentadas na seção anterior e os resultados dos erros são apresentados na Tabela 8.

Tabela 8 - Medidas de desempenho de erros

Amostra	Treino	Teste
Erros	60%	40%
EM	0,102	0,160
REM	3,439	3,821
ER	0,691	0,718
r	0,558	0,532
DESVPAD	3,444	3,815

É possível notar que o Modelo RLM de geração de viagens por domicílio apresenta ajuste coerente com resultados encontrados na literatura para dados desagregados (Ashley, 1978; Atherton e Ben-Akiva, 1976). Os valores dos erros, a depender da medida de desempenho adotada, o Erro Médio (EM), a Raiz do Erro Médio (RME), o Erro Relativo (ER), o coeficiente de correlação (r) ou o Desvio Padrão do erro (DESVPAD), foram um pouco maiores que o Modelo por RNA. Desta forma, o desempenho das duas técnicas pode ser classificado como similar, para o presente estudo de caso, com um desempenho sutilmente melhor para o modelo de RNAs.

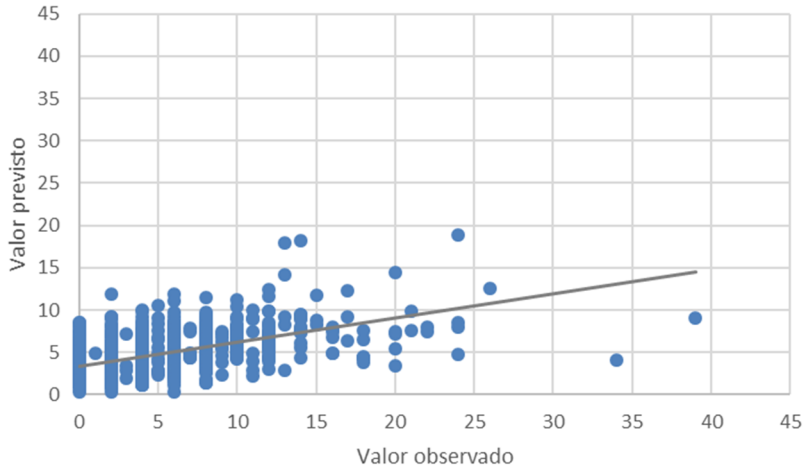


Figura 9. Gráfico de dispersão da variável "viagens domiciliares" (valores observados x valores estimados através da RLM).

Em comparação ao modelo de RNAs, a relação de valores observados e estimados foi um pouco mais forte do que pelo modelo por RLM. Apresentaram coeficientes de *Pearson* de 0,583 e 0,524, respectivamente. No entanto, existe uma vantagem principal em relação ao uso das RNAs, já que esta é uma técnica de Inteligência Artificial, não apresenta as suposições da RLM, sendo livre de restrições, tais como nos modelos lineares tradicionais.

8. VALIDAÇÃO DOS RESULTADOS RELATIVOS À DEMANDA POR TRANSPORTES (ETAPA 6 - VALIDAÇÃO 2)

O objetivo desta etapa de validação foi garantir que o método proposto gerasse informações relativas a viagens confiáveis. Para isto, foram validadas as viagens produzidas sintéticas, por domicílio, estimadas pelo método sequencial proposto (População sintética pelo MMC e RNAs).

Substituindo-se os valores das variáveis independentes da população sintética no modelo de RNAs, previamente calibrado, foram obtidas as viagens domiciliares sintéticas. Para a validação das viagens domiciliares, estimadas a partir do método sequencial proposto, foram utilizados os dados de viagens domiciliares provenientes da Pesquisa OD. A validação proposta foi realizada a partir da comparação de medianas, através do teste de hipótese da mediana e teste *Kolmogorov - Smirnov*, para comparação de distribuições populacionais. A ideia foi verificar se as medianas e distribuição populacional das viagens estimadas pelo método proposto e da Pesquisa OD são similares. O procedimento metodológico compreende a comparação para a amostra completa e para a amostra com a exclusão do percentil 90.

A escolha dos testes, descritos acima, deu-se pela necessidade de incorporar o aspecto não paramétrico. A ideia é comparar tantos valores típicos quanto distribuições para o caso de variáveis que possuam distribuições populacionais desconhecidas, o que não seria o caso do teste t para comparação de médias, por exemplo, que supõe distribuição gaussiana. Este aspecto foi levado em conta considerando que a maior parte das variáveis possuem distribuição populacional desconhecida, embora algumas sigam distribuição Gaussiana, Poisson ou Exponencial.

Vale ressaltar que os testes de hipótese foram realizados considerando os domicílios (viagens sintéticas x viagens observadas da Pesquisa OD) que compõem cada setor censitário. Assim, os testes foram realizados considerando a quantidade de domicílios por setor. A Figura 10 mostra o percentual de setores que passaram em ambos os testes, bem como aqueles que passaram em um dos dois testes propostos e o percentual de setores que não passaram em ambos os testes. A Figura 11 traz os mesmos resultados para a amostra, após a exclusão do percentil 90. Observa-se um padrão espacial, em relação a tais testes estatísticos (Figura 12). Os setores situados em localizações periféricas, predominantemente não passam em ambos os testes. Tal fato pode ser justificado por ausência de amostragem ou poucos domicílios amostrados na Pesquisa OD em tais localizações, haja vista que em 2007 e 2008 (anos de realização da Pesquisa OD em São Carlos) tais setores (zonas de tráfego) tinham pouca população residente.

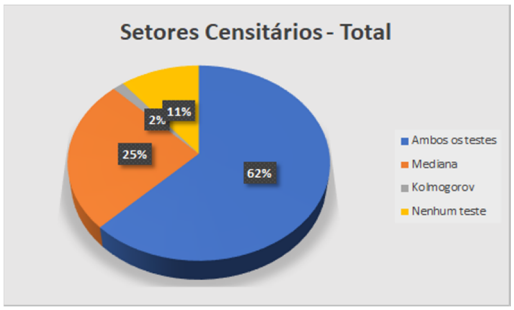


Figura 10. Porcentagem de setores censitários que apresentaram viagens sintéticas válidas

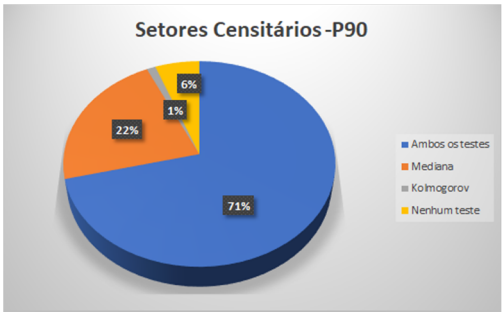


Figura 11. Porcentagem de setores censitários que apresentaram viagens sintéticas válidas após a retirada dos valores do Percentil 90

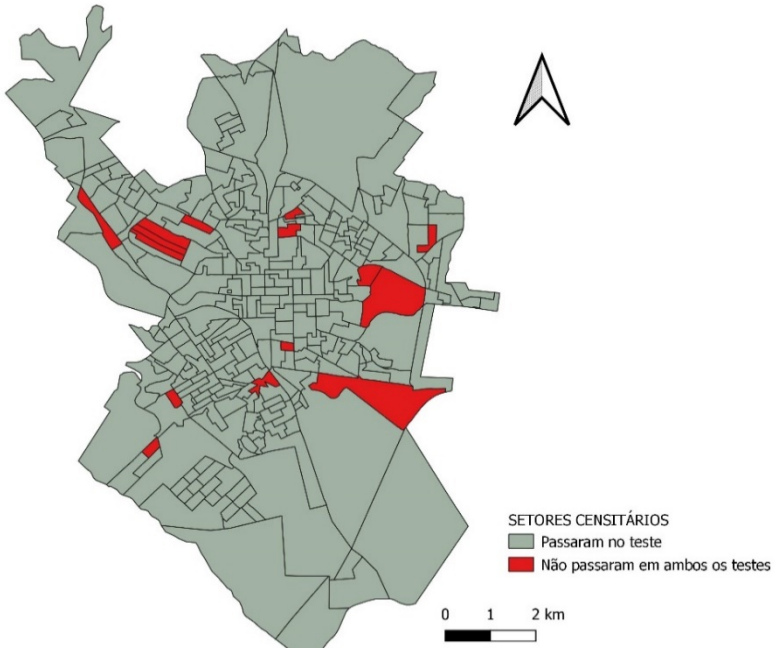


Figura 12. Mapa temático de setores que passaram em ambos os testes estatísticos

9. CONCLUSÕES

A proposta deste trabalho apresentou resultados satisfatórios e evidencia a eficácia do uso conjunto de RNAs e população sintética simulada pelo Método Monte Carlo no planejamento de transportes, especificamente na previsão de geração de viagens domiciliares.

Os resultados encontrados na calibração dos modelos (dados O/D), através da técnica mais usual no planejamento de transportes, a Regressão Linear Múltipla e as RNAs foram bem similares e compatíveis. Vale ressaltar que os resultados foram sutilmente melhores para as RNAs, corroborando o seu potencial no uso de modelagem de demanda por transportes, considerando, principalmente a ausência de restrições e suposições matemáticas associadas à sua aplicação.

O Método Monte Carlo foi eficiente para simular a população sintética, utilizando apenas dados agregados, pois através da validação dessa população, ou seja, da aplicação do teste estatístico de *Kolmogorov-Smirnov* para comparação de distribuições populacionais, observou-se que 62% das variáveis sintéticas foram consideradas aptas, ou seja, de mesma distribuição populacional que os microdados do censo do IBGE.

Importante mencionar que, após a validação das viagens por domicílio da população sintética estimadas pelo método sequencial proposto (MMC seguido de RNAs), constatou-se que em 62% dos setores censitários, as viagens domiciliares foram consideradas similares, segundo distribuições e medida típica (mediana), quando comparadas às viagens observadas pela Pesquisa OD. Com a retirada do percentil 90 da análise (*outliers* da Pesquisa OD), 71% dos setores censitários apresentaram a distribuição e mediana das viagens sintéticas, similares às viagens amostradas pela Pesquisa OD, corroborando a eficiência do método proposto para estimativa de viagens domiciliares.

Assim, através do método sequencial proposto pelos autores, foi possível confrontar dois problemas de pesquisa: (1) ausência de dados desagregados na maioria das cidades brasileiras e (2) restrições e suposições matemáticas associadas à modelagem tradicional de demanda por transportes.

Como principal restrição metodológica, observa-se que a autocorrelação espacial não é, em nenhum momento, considerada para obtenção dos domicílios sintéticos. A dependência espacial é corroborada, ainda, através do mapa temático produzido a partir da validação 2 (Figura 12). Uma solução para tal problema, que seria uma ideia de trabalho futuro, seria aplicação de simulação de dados através de uma ferramenta espacial (Simulação Sequencial Gaussiana). Tal ferramenta pode vir a ser uma boa solução para esta restrição, levando-se em conta variáveis de demanda por Transportes (Linder, 2019; Lindner e Pitombo, 2019).

AGRADECIMENTOS

À Agência de fomento CNPq (Processo 303645/2015-6).

REFERÊNCIAS

Adeli, H. (2001) Neural Network in Civil Engineering: 1989 – 2000. *Computer-Aided Civil and Infrastructure Engineering,* v. 16 (2), p. 126-146. DOI: 10.1111/0885-9507.00219.

Adiga, A., Agashe, A.; Arifuzzaman, S.; Barrett, C.L.; Beckman, R.J.; Bisset, K.R.; Chen, J.; Chungbaek, Y.; Eubank, S.G.; Gupta, S.; Khan, M., Kuhlman; C.J., Lofgren, E.; Lewis, B.L.; Marathe, A.; Marathe, M.V.; Mortveit, H.S.; Nordberg, E.; Rivers, C.; Stretz, P.; Swarup, S.; Wilson, A. e Xie, D. (2015) *Generating a synthetic population of the United States.* Tech. Rep. NDSSL 15-009, Network Dynamics and Simulation Science Laboratory. Disponível em: <https://pdfs.semanticscholar.org/391d/bcde4fda9186f03da174d6b7a4494d0bb6e4.pdf> (Acesso em: 07/08/2019).

Aguiar Júnior, S. R. (2004) *Modelo RAPIDE: uma aplicação de mineração de dados e redes neurais artificiais para a estimativa da demanda por transporte rodoviário interestadual de passageiros no Brasil.* Brasília. Dissertação (Mestrado em Gestão do conhecimento e da Tecnologia da informação), Universidade Católica de Brasília.

Al-Deek, H.M. (2001) Which method is Better for Developing Freight Planning Models at Seaports – neural Networks or Multiple Regression, *Transportation Research Record*, 1763, pp. 90-97, 2001. DOI: 10.3141/1763-14.

Atherton, T.J. and Ben-Akiva, M.E. (1976) Transferability and updating of disaggregate travel demand models. *Transportation Research Record*, 610, 12–18.
Disponível em: <http://onlinepubs.trb.org/Onlinepubs/trr/1976/610/610-003.pdf>. (Acesso em: 07/08/2019).

Ashley, D. J. (1978) The Regional Highway Traffic Model: the home based trip end model. *Proceedings 6th PTRC*. Summer Annual Meeting, University of Warwick, July 1978, England.

Banks, J.; Carson, J.S.; Nelson, B.L. e Nicol, D.M. (2005) *Discrete-Event System Simulation* (4ª ed.). Prentice Hall, Upper Saddle River, NJ.

Bates, J.J., Gunn, H.F. and Roberts, M. (1978) A model of household car ownership. *Traffic Engineering and Control* 19, 486–491, 562–566.

Batista; B. C. F. (2012) *Soluções de Equações Diferenciais usando Redes Neurais de Múltiplas camadas com os métodos da Descida mais íngreme e Levenberg-Marquardt.* Belém. 88p. Dissertação (Mestrado em Matemática), Universidade Federal do Pará.

Beckman, R. J.; Baggerly, K. A. e Mckay, M. D. (1996) Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*. Elsevier, v.30, n.6, p.415–429. DOI: 10.1016/0965-8564(96)00004-3.

Birkin, M.; Clarke, M. (1988) Synthesis—a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and planning A*. SAGE Publications, v.20, n.12, p.1645–1671. DOI: 10.1068/a201645.

Bowman, J. L. (2004) *A comparison of population synthesizers used in microsimulation models of activity and travel demand*. Working paper. Disponível em: <https://pdfs.semanticscholar.org/3d1c/1fda84391752dfa520ccb08426eebf7c5da2.pdf>. Acesso em: 17.3.2019.

Bowman, John L. (2009) *Population Synthesizers, Traffic Engineering and Control,* Vol. 49. No. 9: 342.

Bruton, M. J. (1979) *Introdução ao planejamento dos transportes.* Rio de Janeiro: Interciência.

Chwif, L.; Medina, A. (2014) *Modelagem e Simulação de Eventos Discretos: Teoria e Aplicações* (4ª ed.). Elsevier, São Paulo, Brasil.

Clifton, K. J., K. M. Currans, A. C. Cutter, and R. Schneider (2012) Household travel surveys in a context-based approach for adjusting ITE trip generation rates in urban contexts. *Transportation Research Record* 2307: p.108–119. DOI: 10.3141/2307-12.

Deming, W. E.; Stephan, F. F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, JSTOR, v.11, n.4, p.427–444. DOI: 10.1214/aoms/1177731829.

Dougherty, M. (1995) A review of neural networks applied to transport. *Transportation Research, Part C*, v.3, n.4. p.247-260. DOI: 10.1016/0968-090X(95)00009-8.

Ewing, R., M. DeAnna, and Li, S. C. (1996). Land use impacts on trip generation rates. *Transportation Research Record* 1518: 1–6. DOI: 10.1177/0361198196151800101.

Escudero, L. F. (1973) *La simulatición en la empresa.* Ed. Duesto, Bilbao.

Freedman, D. A. (2009) *Statistical Models: Theory and Practice.* Cambridge University Press.

Gonçalves D. N. S.; Lopes, L. A. S. Da Silva, M. A. V. (2014) Aplicação de Redes Neurais Artificiais para estimar matriz origem-destino de carga. Anais do 28º Congresso Brasileiro de Transporte e Trânsito – ANPET, Curitiba. Disponível em: <http://redpgv.coppe.ufrj.br/index.php/pt-BR/producao-da-rede/artigos-cientificos/2014-1/864-aplicacao-de-redes-neurais-artificiais-para-estimar-matriz-origem-destino-de-carga/file>. (Acesso em: 07/08/2019).

Goncalves, D. N. S.; Silva, M. A.V.; D'Agosto, M. A. Procedimento para uso de Redes Neurais Artificiais no planejamento estratégico de fluxo de carga no Brasil. J. Transp. Lit., Manaus, v. 9, n. 1, p. 45-49, Jan. 2015. DOI: 10.1590/2238-1031.jtl.v9n1a9.

Hair, J. F.; Black, W.C.; Babin, B.J.; Anderson, R.E.; Tatham, R.L. (2009) *Análise multivariada de dados* (6ª ed.). Ed. Bookman, Porto Alegre.

Hammersley, J.; Handscomb, D. (1964) *Monte Carlo Methods, Methuen's Monographs on Applied Probability.*: Wiley, New York.

Haykin, S. (1999) *Neural networks - a comprehensive foundation*. Prentice Hall, Ontario, Canada.

Huang, Z.; Williamson, P. A. (2002) *Comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata*. Working paper. Department of Geography, University of Liverpool. Disponível em: <file:///C:/Users/manav/Desktop/ARTIGOS_Transportes/huang.pdf> (Acesso em: 07/08/2019).

Huisken, G.; Coffa, A. (2000) Neural Networks and Fuzzy Logic to improve Trip Generation Modelling, *Proceedings of the 9th International Association for Travel Behaviour Research Conference*, Institute of Transport Studies, IATBR, Gold Coast, Queensland, Austália, 2000. Disponível em: < https://research.utwente.nl/en/publications/neural-networks-and-fuzzy-logic-to-improve-trip-generation-modell> (Acesso em: 07/08/2019).

IBGE (2010) - *Instituto Brasileiro de Geografia e Estatística*. Disponível em: <http://www.ibge.gov.br>. (Acesso em: 15.3.2017).

Ichikawa, S.M.; Pitombo, C.S.; Kawamoto, E. (2002) Aplicação de Minerador de dados na obtenção de relações entre padrões de viagens encadeadas e características socioeconômicas. *Anais do XVI do Congresso de Pesquisa e Ensino em Transportes,* Anpet, Natal (RN), v. 2, p. 175-186.

Jain, S.; Ronald, N.; Winter, S. (2015) Creating a Synthetic Population: A Comparison of Tools. *In Proceedings of the 3rd Conference Transportation Reserch Group*, Kolkata, India, 17–20 December 2015. Disponível em: <https://www.researchgate.net/publication/291608775_Creating_a_Synthetic_Population_A_Comparison_of_Tools>. (Acesso em:07/08/2019).

Kawamoto, E. (1994) *Análise de sistemas de transporte.* 2ª ed. (Apostila). Departamento de Transportes. Escola de Engenharia de São Carlos, Universidade de São Paulo. São Carlos, Brasil.

Lin, H.; Taylor, M.A.P.; Zito R. (2005) A Review of Travel-Time prediction in Transport and Logistics. *Proceedings of the Eastern Asia Society for Transportation Studies*, v. 5, p. 1433 – 1448. Disponível em: <https://trove.nla.gov.au/work/179520839?q&versionId=223381176+224842764+245312661+253844472>. (Acesso em: 07/08/2019).

Lindner, A. (2019) Métodos heurísticos de desagregação de dados de demanda por transportes através de simulação geoestatística. São Carlos. 118p. Tese (Doutorado) – Programa de Pós-graduação em Engenharia de Transportes e Área de Concentração em Planejamento e Operação de Sistemas de Transportes – Escola de Engenharia de São Carlos da Universidade de São Paulo.

Lindner, A. e Pitombo, C.S; (2019) Sequential Gaussian Simulation as a promising tool in travel demand modeling. Journal of Geovisualization and Spatial Analysis (disponível online). DOI: 10.1007/s41651-019-0038-x.

Ma, L. (2011) *Generating disaggregate population characteristics for input to travel-demand models.* Florida. 124p. Tese (Doutorado) — University of Florida.

Ma, L. e Srinivasan, S. (2015) Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*. Wiley Online Library, v.30, n.2, p.135–150. DOI: 10.1111/mice.12085.

Mackay, D. J. C. (1996) *Introduction to Monte Carlo Methods.* Cambridge (Cambridgeshire).

Moeckel, R.; Spiekermann, K.; Wegener, M.. (2003) Creating a Synthetic Population". *In: Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management* (CUPUM), Sendai, Japan: [s.n.], p. 1-18.

Mozolin, M.; Thill, J.C.; Linn, U.E. (2015) Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological*, v. 34, p. 53-73. DOI: 10.1016/S0191-2615(99)00014-4. Disponível em: <https://ideas.repec.org/a/eee/transb/v34y2000i1p53-73.html>. (Acesso em: 07/08/2019).

Müller, K. e Axhausen, K. W. (2011) Hierarchical IPF: Generating a synthetic population for Switzerland. *In: 51 st Congress of the European Regional Science Association.* Barcelona. Disponível em: < file:///C:/Users/manav/Desktop/ARTIGOS_Transportes/Muller%202011.pdf>. (Acesso em: 08/08/2019).

Münnich, R.; Schürle, J.; Bihler, W.; Boonstra, H. J.; Eckmair, D. e Haslinger, A. (2003) Monte Carlo simulation Study of European Surveys. *DACSEIS Deliverables D*, v.3.

Namazi-rad, M. R.; Mokhtarian, P. e Perez, P. (2014) *Generating a dynamic synthetic population–using an age-structured two-sex model for house hold dynamics.* PloS one. Public Library of Science, v.9, n.4, p. e94761. DOI: 10.1371/journal.pone.0094761.

Novaes, A. G. (1986) *Sistemas de Transportes.* Volume 1: Análise da Demanda. São Paulo: Edgard Blucher.

Openshaw, S.; Openshaw, C. (1997) *Artificial intelligence in Geography*, John Wiley and Sons, Chichester, 1997.

Ortúzar, J.D.; Willumsen, L.G. (2011). *Modelling Transport.* Wiley, 4th Edition.

Papacostas C. S.; Provedouros, P. D. (1993) *Transportation Engineering and Planning*. 2.ed. New Jersey: Prentice Hall.

Pitombo, C. S.; De Souza, A.D.; Lindner, A. (2017) Comparing decision tree algorithms to estimate intercity trip distribution. *Transportation Research Part C* , v. 77, p. 16-32. DOI: 10.1016/j.trc.2017.01.009.

PNUD (2016) - *Atlas do Desenvolvimento Humano no Brasil*. Disponível em: <http://http://www.pnud.org.br/atlas>. Acesso em: 15.3.2017.

Pritchard, D. R. (2008) *Synthesizing Agents and Relationships for Land Use/ transportation Modelling.* 2008. Canadian theses, University of Toronto, Toronto.

Raymond, C. J. (1982) Adapting for Heteroscedasticity in Linear Models. *The Annals of Statistics.* 10 (4): 1224–1233. DOI:10.1214/aos/1176345987. JSTOR 2240725.

Schmöcker, J. D.; Quddus, M. A.; Noland, R. B.; Bell, Michael G.H. (2005) Estimating trip generation of elderly and disabled people: analysis of London data, *Transportation Research Record*, 9–18. DOI: 10.3141/1924-02.

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

Smith, M. (1996) *Neural Networks for Statistical Modeling.* International Thomson Computer Press, Londres, Inglaterra.

Tillema, F., van Zuilekom, K. M.; van Maarseveen, M. F. A. M. (2004) Trip generation: comparison of neural networks and regression models. In C. A. Brebbia, & L. C. Wadhwa (Eds.), *Urban Transport X : Urban transport and the environment in the 21st century* (pp. 121-130). Southampton, UK: WIT Press. Disponível em: <https://research.utwente.nl/en/publications/trip-generation-comparison-of-neural-networks-and-regression-mode>. (Acesso em: 07/08/2019).

Turchenko, V.; Grandinetti, L.; Bosilca, G.; Dongarra, J. (2010). Improvement of parallelization efficiency of batch pattern BP training algorithm using Open MPI. *Procedia CS.* 1. 525-533. DOI: 10.1016/j.procs.2010.04.056.

Vickerman, R. W.; Barmby, T. A. (1985) Household trip generation choice: alternative empirical approaches. *Transportation Research B*, 19(6), 471-479. DOI: 10.1016/0191-2615(85)90042-6.

Voas, D. e Williamson, P. (2001) Evaluating goodness-of-fit measures for synthetic microdata. *Geographic a land Environmental Modelling.* Taylor & Francis, v.5, n.2, p.177–200. DOI: 10.1080/13615930120086078.

Xie, C.; Lu, J.; Parkany, E. (2003) Work travel mode choice modeling with data mining: decision trees and neural networks. *Transportation Research Record: Journal of the Transportation Research Board*, n. 1854, p. 50-61. DOI: 10.3141/1854-06.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana and P. Waddell (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

Wasserman P.D. (1989) *Advanced methods in neural computing*. New York, Van Nostrand Reinhold.

Williamson, P.; Birkin, M.; Rees, P.H. (1998) The estimation of population microdata by using data from small area statistics and samples of anonymized records. *Environment and Planning A,* SAGE Publications, v.30, n.5, p.785–816. DOI: 10.1068/a300785.