Revista de Engenharia Civil
2019, No. 56, 5-13
http://www.civil.uminho.pt/revista

Uso de árvore de decisão para previsão de geração de viagens como alternativa ao método de classificação cruzada

M.N. Pianucci[a†], C.S. Pitombo[a]

[a] *Universidade de São Paulo, Escola de Engenharia de São Carlos, São Carlos, Brasil*
[†] *Autor para correspondência: marcela.navarro@usp.br*

RESUMO

O objetivo do presente trabalho é apresentar a Árvore de Decisão (AD) para determinação de classes domiciliares, associadas a taxas de viagens como técnica alternativa ao método de Classificação Cruzada. A área de estudo é a cidade de São Carlos (São Paulo, Brasil) e os dados utilizados são provenientes da Pesquisa Origem-Destino (O/D), realizada na cidade em 2007/2008. O método possui sete etapas principais: (1) Escolha das variáveis relevantes ao estudo; (2) Aplicação da AD; (3) Obtenção das classes domiciliares; (4) Aplicação da RLM; (5) Estimativas de geração de viagens obtidas em 2; (6) Estimativas de geração de viagens obtidas em 4; (7) Comparação dos resultados obtidos em 5 e 6. Como resultados, a variável de maior importância (que melhor explica a variabilidade dos dados em relação à geração de viagens por domicílio) foi número de moradores no domicílio e, em ambas as técnicas, os resultados foram muito similares. O uso de AD como ferramenta de auxílio na determinação de estratos para previsão de geração de viagens é adequado, possuindo boa acurácia na previsão do fenômeno de viagens domiciliares e utilizando como critério para determinação das classes medida de impureza (redução da variância).

Cronologia do artigo:

Recebido a 07 junho 2017
Corrigido a 27 setembro 2018
Aceite a 11 abril 2019
Publicado a 27 maio 2019

Palavras-chave:

Regressão linear múltipla
Viagens domiciliares
Modelos desagregados
Árvore de decisão e regressão

1. Introdução

A análise da demanda por transportes tem por objetivo compreender os seus determinantes, as interações que os envolvem e como eles influenciam o volume do tráfego. Os modelos tradicionais, ou modelos quatro etapas, são os mais utilizados nos estudos de demanda e seguem a clássica sequência de geração de viagens, distribuição de viagens, divisão modal e alocação dos fluxos à rede de transportes (Novaes, 1986; Bruton, 1979; Kawamoto, 1994; Ortúzar e Willumsen, 2011; Papacostas e Prevedouros, 1993).

A primeira etapa do modelo sequencial tem como objetivo estimar a produção e atração de viagens em cada zona de tráfego ou viagens geradas por domicílio (Kanafani, 2001; Ortúzar e Willumsen, 2011; Fleet e Sosslau, 1976). Esta fase é essencial ao planejamento de transportes, pois é nela que se analisam os principais fatores que determinam a geração de viagens. Os modelos mais usuais para representação da geração de viagens são Regressão Linear Múltipla (RLM) e Classificação Cruzada (Ortúzar e Willumsen, 2011). Os dois métodos podem ser considerados aceitáveis, até certo ponto, em termos de planejamento de sistemas de transportes. Contudo, alguns problemas críticos podem ser alocados a cada um dos métodos.

No caso da RLM, que o objetivo principal seria 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).

A Classificação Cruzada ou Análise de Categorias é uma técnica que utiliza dados desagregados por categorias de domicílio. As viagens são agrupadas e relacionadas à estrutura familiar e às condições econômicas de cada família. Este modelo considera que as taxas de geração de viagens, independente da categoria da família, são constantes no futuro (Peyrebrune, 1985).

A vantagem desta técnica é que os agrupamentos de classificação cruzada não têm relação com o zoneamento da área de estudo. Não são necessárias suposições prévias sobre a forma da relação entre os dados ou curva de ajuste destes. No entanto, este método apresenta algumas desvantagens, como a necessidade de grande quantidade de dados. Outra desvantagem é a ausência de um método eficaz para

a escolha das variáveis de classificação, ou para escolher melhores agrupamentos a partir das variáveis explicativas (Ortúzar e Willumsen, 2011).

Adicionalmente, a forma arbitrária para escolha das variáveis independentes e, consequentemente, os estratos, pode ser um problema crítico. Além disso, o cálculo de viagens, célula por célula deste método, aumenta a preocupação relativa a estimativas não realistas, particularmente para o caso de amostras pequenas com alta variância (Chang *et al*., 2014).

Desta forma, alguns autores propuseram procedimentos alternativos para determinação de classes domiciliares para previsão de geração de viagens ao longo das últimas décadas. Stopher e MacDonald (1983) propuseram um método alternativo para calibração de modelos de classificação cruzada conhecido como Análise de Múltipla Classificação. O método é baseado na análise de variância (ANOVA), que proporciona um procedimento estruturado para a escolha entre as variáveis independentes e agrupamentos baseados em valores das variáveis independentes escolhidas. Após esta primeira aplicação dos mencionados autores, o método foi replicado diversas vezes para previsão de geração de viagens domiciliares (Ortúzar e Willumsen, 2011; Clark, 1996; TMIP, 2004).

No entanto, tal método pode superestimar o futuro número de viagens. Guevara e Thomas (2007) discutem a necessidade de utilização de formulações mais sofisticadas para modelar geração de viagens baseadas no domicílio.

Desta forma, o objetivo do presente trabalho é apresentar a Árvore de Decisão (AD) para determinação de classes domiciliares, associadas a taxas de viagens, como procedimento alternativo ao método de Classificação Cruzada. Este trabalho apresenta quatro seções, além desta introdução. A seção 2 descreve a técnica de Árvore de Decisão. A descrição dos dados utilizados para a elaboração do estudo se encontra na seção 3 e, em sequência, é apresentado o método seguido neste trabalho. Finalmente, a seção 4 e seção 5 apresentam os principais resultados obtidos, suas discussões e conclusões, respectivamente.

2. Árvore de decisão

Uma forma simples de representação de relação ou de relações existentes em um conjunto de dados é feita pela técnica de Árvore de Decisão (AD). As ADs são estudadas em vários campos de pesquisa como ciências sociais, estatística, engenharia e inteligência artificial. Ela permite classificar uma base de dados em um número finito de classes, com a qual é possível analisar um grande conjunto de dados, através de regras hierárquicas e da sua divisão em grupos, organizando os dados de maneira compacta e obtendo uma visão real da natureza do processo (Quinlan, 1983).

A hierarquia é denominada árvore e cada segmento é denominado nó. O segmento original contém o conjunto completo dos dados, referindo-se ao nó raiz da árvore. Este nó contém dados que podem ser subdivididos dentro de outros sub-nós, chamados de nós filhos. Quando os dados do nó não podem ser mais subdivididos dentro de outro subconjunto ele é considerado um nó terminal ou folha. O algoritmo usado para dividir os dados nos modelos de árvore identificam as variáveis independentes que maximizam a homogeneidade dentro dos subgrupos de dados que compõem cada nó filho, segundo a variável dependente. Alguns dos algoritmos existentes são o C4.5 (Quilan, 1993), CHAID (Kass, 1980), CART (Breiman *et al*., 1984) e QUEST (Loh e Shih, 1997). Neste trabalho optou-se pela utilização do algoritmo CART (*Classification And Regression Tree*), contido no pacote estatístico IBM SPSS 24.0

Quando a variável dependente é qualitativa, as árvores de decisão são usadas para problemas de classificação e são denominadas Árvores de Classificação. Para o caso de variável dependente quantitativa, as árvores de decisão são denominadas Árvores de Regressão, que é o caso deste artigo. A Figura 1 ilustra, esquematicamente, o gráfico acíclico de Árvore de Decisão para divisões binárias (Algoritmo CART).

Nas árvores de decisão e classificação, cada nó terminal ou folha contém um rótulo que indica a classe predita para um determinado conjunto de dados. Para as Árvores de Classificação, os critérios de partição ou medidas de impureza mais conhecidos são baseados na entropia e índice Gini. Já para as Árvores de Regressão, a medida de impureza é chamada de redução da variância a qual representa a redução da variância da variável dependente em cada nó.

A redução da variância, que representa a função de impureza, é apresentado na Equação 1.

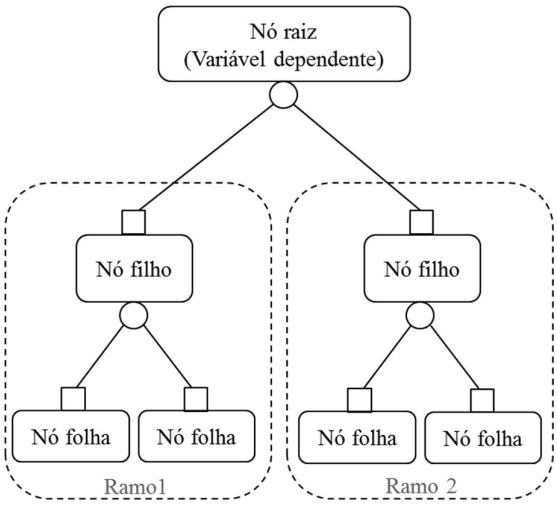

Figura 1 - Esquema de um modelo de AD (Breiman *et al*., 1984).

$$I_V(N) = \frac{1}{|S|^2}\sum_{i\in s}\sum_{j\in s}\frac{1}{2}(x_i - x_j)^2 - \left(\frac{1}{|S_t|^2}\sum_{i\in s_t}\sum_{j\in s_t}\frac{1}{2}(x_i - x_j)^2 + \frac{1}{|S_f|^2}\sum_{i\in s_f}\sum_{j\in s_f}\frac{1}{2}(x_i - x_j)^2\right) \quad (1)$$

Sendo:
Iv (N) = redução da variância no nó N;
S= conjunto da amostra de teste;
St= conjunto da amostra teste do qual o valor da variável explicativa é verdadeiro;
Sf= conjunto da amostra teste do qual o valor da variável explicativa é falso;
xi = valor da variável dependente da amostra teste;
xj = valor da variável dependente da amostra que compõe o nó N.

De um modo geral, o algoritmo da árvore torna os subconjuntos resultantes cada vez mais homogêneos em relação à variável resposta, mediante sucessivas divisões binárias no conjunto de dados. Para o caso deste trabalho, a homogeneidade de cada nó é medida pela pouca variância da variável dependente em cada classe/nó. O valor previsto nos nós terminais corresponde à média da variável dependente do subconjunto de dados que compõe aquela folha, sendo assim, quanto menor a variância dentro do nó terminal, maior a homogeneidade relativa àquela classe de observações (domicílios no caso deste trabalho). A cada passo no crescimento da árvore, o particionamento dos dados se faz a partir da produção da minimização da variância da variável dependente (Breiman *et al*., 1984).

3. Materiais e método

O método utilizado no estudo de geração de viagens por domicílio é composto pelas seguintes etapas: (1) Escolha das variáveis relevantes ao estudo; (2) Geração da AD – amostras de treinamento (70%) e de teste (30%); (3) Obtenção das classes domiciliares (nós terminais da AD); (4) Aplicação de Regressão Linear Múltipla (RLM) com a mesma amostra de treinamento utilizada na AD; (5) Estimativas de geração de viagens obtidas em (2); (6) Estimativas de geração de viagens obtidas em (4) e (7) Comparação dos resultados. Os procedimentos metodológicos são ilustrados na Figura 2. Vale ressaltar que o uso de RLM foi feito apenas para validação do método, através da comparação de resultados obtidos com AD e pelos modelos lineares.

A área de estudo do presente trabalho é a cidade de São Carlos (São Paulo, Brasil). A cidade possui uma área urbana de aproximadamente 105 km^2 e, em 2010 apresentava uma população de 221.950 habitantes (IBGE, 2010).

Os dados utilizados para o desenvolvimento deste trabalho são provenientes da Pesquisa Origem-Destino (O/D), realizada entre os anos de 2007 e 2008 pelo Departamento de Engenharia de Transportes da Escola de Engenharia de São Carlos, Universidade de São Paulo (Rodrigues da Silva, 2008). Neste trabalho, foi utilizado o banco de dados desagregado por domicílios (3057 domicílios no total). A amostra contém características socioeconômicas e de viagens.

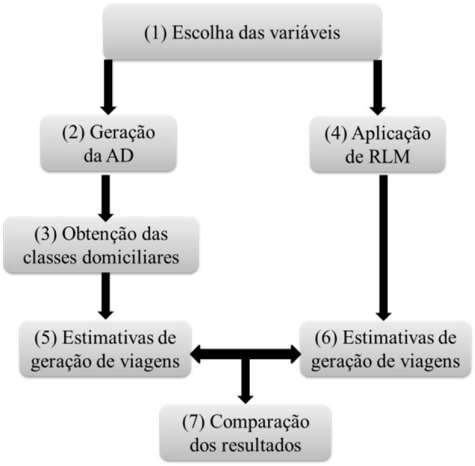

Figura 2 - Diagrama resumido das etapas do método.

3.1. Escolha das variáveis relevantes ao estudo

O primeiro passo do método foi eleger quais as variáveis necessárias para atingir o objetivo do problema. A partir do banco de dados desagregados por domicílio da Pesquisa O/D, as variáveis foram selecionadas (todas numéricas discretas), caracterizando fatores socioeconômicos e de viagens. Após esta primeira seleção, foi realizada uma análise empírica sobre as variáveis que realmente explicam o fenômeno da geração de viagens por domicílio, pois as variáveis explicativas devem ser selecionadas não somente, pela correlação forte com a variável dependente, mas sim por explicar adequadamente o fenômeno. De acordo com Ortúzar e Willumsen (2011), as variáveis mais utilizadas para explicar a geração de viagens, através de análise desagregada, são renda; posse do carro; tamanho da família e estrutura familiar. A Tabela 1 apresenta as variáveis escolhidas e suas medidas descritivas.

Tabela 1 - Descrição e medidas descritivas dos dados originais utilizados no trabalho.

Medidas descritivas							
Variável Dependente	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**	**Quartil 25**	**Mediana**	**Quartil 75**
Viagens *Número de viagens por domicílio*	4,81	4,26	0,00	39,00	2,00	4,00	7,00
Variável Dependente	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**	**Quartil 25**	**Mediana**	**Quartil 75**
Renda *Renda familiar em salários mínimos*	3,03	3,97	0,00	46,00	0,00	2,00	4,00
Moradores *Número de moradores no domicílio*	3,30	1,50	1,00	15,00	2,00	3,00	4,00
Automóveis *Quantidade de automóveis no domicílio*	0,75	0,75	0,00	5,00	0,00	1,00	1,00
Motos *Quantidade de motos no domicílio*	0,19	0,44	0,00	4,00	0,00	0,00	0,00

3.2. Aplicação das Árvores de Decisão (AD)

Nesta etapa, foi utilizada a primeira técnica do método, Árvore de Decisão (AD). As árvores de classificação são usadas quando a variável dependente é categórica (qualitativa) e as árvores de regressão quando a variável dependente é continua (quantitativa). Para o caso deste trabalho, o algoritmo utilizado é denominado Árvore de Regressão, considerando que a variável dependente (número de viagens domiciliares) é numérica discreta. As variáveis independentes foram renda, moradores, automóveis e motos.

Neste estudo, o software utilizado para a geração das árvores foi o IBM SPSS 24.0 que oferece a opção de selecionar a amostra de treinamento e de teste. Foi então selecionada aleatoriamente 70% para treinamento e 30% para teste e geradas as duas árvores, respectivamente.

3.3. Aplicação de Regressão Linear Múltipla (RLM)

Foram definidas as variáveis independentes (variáveis numéricas descritas na Tabela 1) e dependente (geração de viagens domiciliares) a partir do banco de dados da Pesquisa O/D.

Considerou-se, ainda, a correlação entre elas, com valores de coeficientes de *Pearson* das variáveis da amostra a fim de evitar problemas de multicolinearidade, como mostra a Tabela 2.

A presença de elevadas correlações entre as variáveis independentes é a primeira indicação de multicolinearidade (efeito combinado de duas ou mais variáveis independentes) Desta forma, observou-se que não houve nenhum valor alto de correlação entre variáveis explicativas que pudesse prejudicar a aplicação da técnica.

Barbetta (2012) define, genericamente, valores de correlações de *Pearson* altos, moderados e fracos. Valores fracos situam-se próximos ao zero, valores fortes próximos ao 1 ou -1, enquanto valores moderados situam-se entre 0,3 e 0,6 (positivos ou negativos). Vale ressaltar ainda a importância do teste de significância de tais valores sobre R, quando deve-se testar a existência de correlação entre duas variáveis na população.

Tabela 2 - Matriz de correlação entre as variáveis independentes.

Pearson	Viagens	Renda	Moradores	Automóveis	Motos
Viagens		0,247	0,569	0,282	0,159
Renda	0,247		0,135	0,309	0,066
Moradores	0,569	0,135		0,201	0,169
Automóveis	0,282	0,309	0,201		0,086
Motos	0,159	0,066	0,169	0,086	

Após a obtenção da matriz de correlação e análise visual das relações entre variável dependente e variáveis independentes, é obtido o modelo linear para geração de viagens por domicílio. Após análise estatística de significância dos parâmetros estimados e do modelo total (estatística t e teste F), foi realizada 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.

3.4. Comparação dos resultados

Finalmente, foi feita uma comparação dos resultados estimados de viagens domiciliares através da AD, da RLM e dos valores observados pela Pesquisa O/D. Foram utilizadas medidas de desempenho tais como Erro médio, Erro relativo e Coeficiente de Correlação apresentadas em seguida (Equações 2, 3 e 4), respectivamente.

$$EM = \frac{1}{N} . \sum_{i=1}^{N} (x_i - y_i) \tag{2}$$

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

Sendo: EM = Erro Médio; xi = valor observado; yi = valor previsto; ER = Erro relativo; $\bar{x}$ = valor médio observado.

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

Em que: r = Coeficiente de Correlação; $\bar{x}$ = valor médio observado; $\bar{y}$ = valor médio estimado; σx = desvio padrão dos valores observados; σy = desvio padrão dos valores estimados.

4. Resultados e discussões

A Figura 3 e a Figura 4 representam as árvores de treinamento e teste geradas. Ao final da segregação dos dados foi encontrado um total de 12 nós filhos, sendo 7 nós terminais ou folhas.

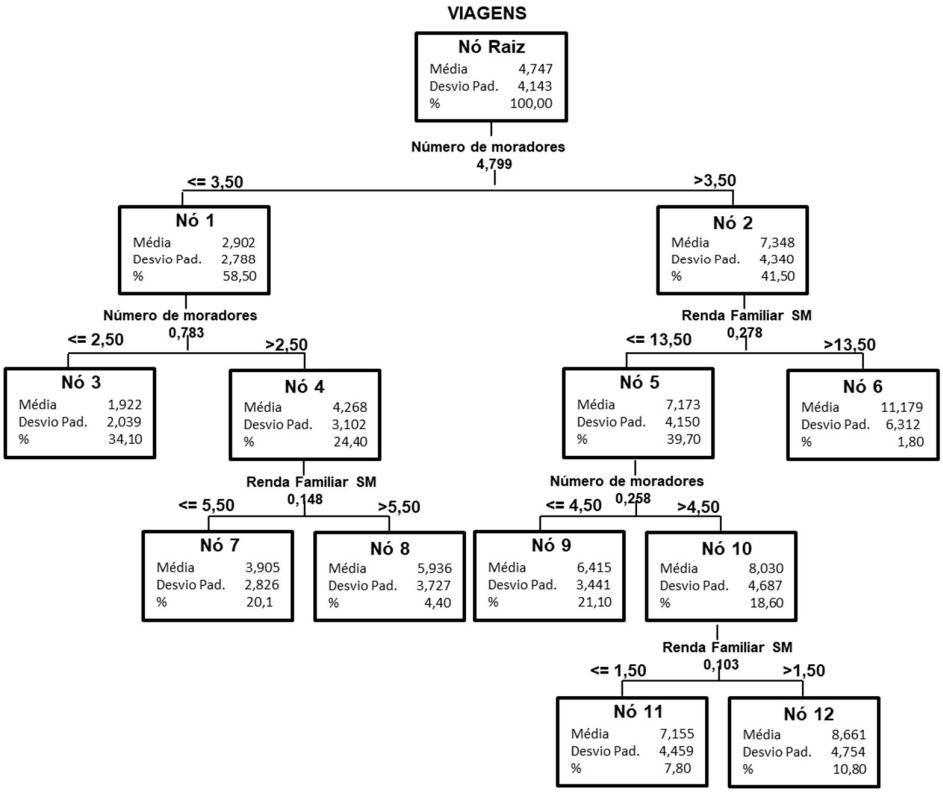

Figura 3 - Árvores de Decisão treinamento (70%).

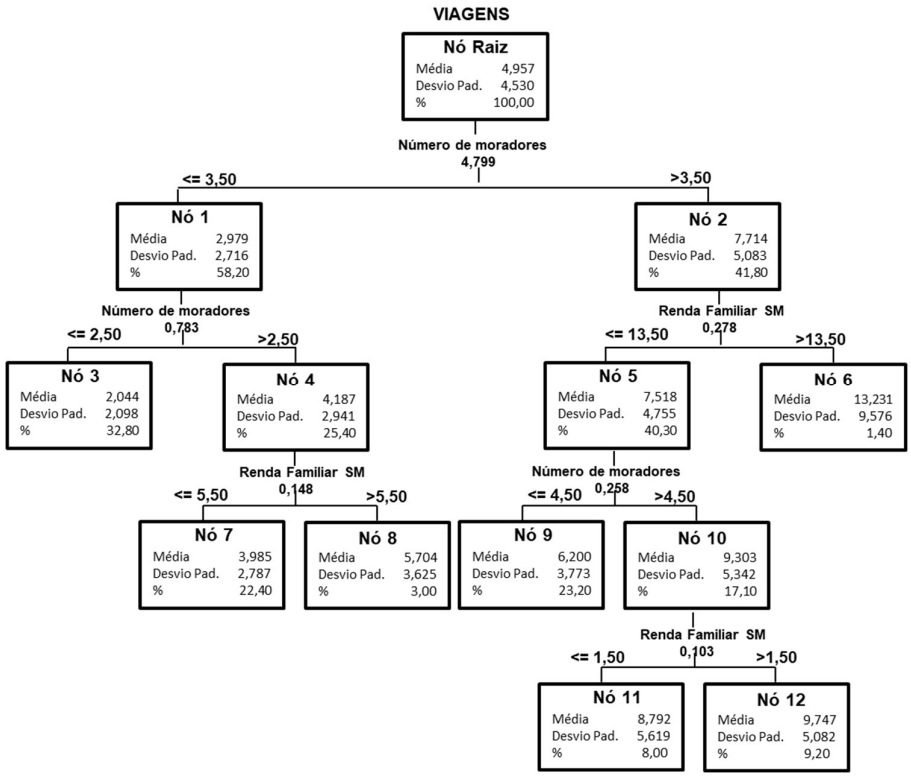

Figura 4 - Árvores de Decisão teste (30%).

Em todos os nós da AD encontram-se ilustrados a média da variável dependente (quantidade média de viagens para os domicílios classificados naquele nó determinado), desvio padrão e a porcentagem (em relação a amostra) de domicílios que compõe cada nó. A variável de maior importância (que melhor explica a variabilidade dos dados em relação à geração de viagens por domicílio) foi número de moradores no domicílio. Ressalta-se que cada nó da AD corresponde a uma classe domiciliar, caracterizada pelas variáveis explicativas e seus respectivos valores de corte, além de valores médios de viagens domiciliares.

Através dos resultados da AD, são apresentados, nos nós terminais da árvore de treinamento, as classes domiciliares associadas à média de viagens estimadas, conforme a Tabela 3.

Tabela 3 - Classes da previsão de viagens por domicílios.

			Viagens		
Classes					**Taxa média de viagens**
1	≤ 3,5 morador	≤ 2,5 morador	-	-	1,922
2	≤ 3,5 morador	> 2,5 morador	≤ 5,5 renda	-	3,905
3	≤ 3,5 morador	> 2,5 morador	> 5,5 renda	-	5,936
4	> 3,5 morador	≤ 13,5 renda	≤ 4,5 morador	-	6,415
5	> 3,5 morador	> 13,5 renda	-	-	11,179
6	> 3,5 morador	≤ 13,5 renda	> 4,5 morador	≤ 1,5 renda	7,155
7	> 3,5 morador	≤ 13,5 renda	> 4,5 morador	> 1,5 renda	8,661

A árvore definiu como principal variável explicativa da produção de viagens domiciliares a variável número de moradores (Figura 5). Em seguida, a variável renda familiar foi selecionada e a terceira variável foi o número de automóveis, seguida pelo número de motos.

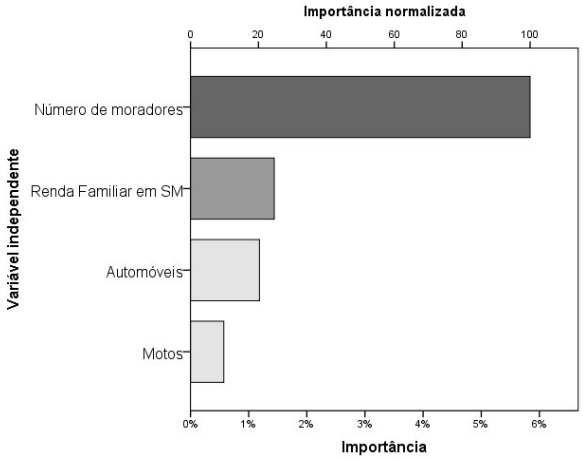

Figura 5 - Importância normalizada das variáveis analisadas para a produção de viagens domiciliares.

O modelo linear, obtido para geração de viagens por domicílio foi considerado significativo, com as seguintes variáveis independentes: moradores, renda, automóveis e motocicletas. A significância do modelo é analisada através da estatística (F) de Fisher e o teste de hipótese associado à mesma. Neste teste, o objetivo seria reter a hipótese alternativa. As hipóteses para o teste da estatística F são as seguintes: Ho: Não existe relação linear entre variável dependente e variáveis independentes; H1: A relação linear entre variável dependente e variáveis independentes é significativa (não é causalidade). De acordo com a Tabela 4, o teste conhecido como teste F do modelo, resultou em F=317,703, com correspondente valor p=0,0000. Como o valor p é extremamente pequeno, o teste estatístico rejeita Ho, indicando que as variáveis Morador, Renda, Automóveis e Motocicletas são significativas para explicar a quantidade de viagens domiciliares.

Além disso, o modelo apresentou um coeficiente de determinação equivalente a 0,372 (R^2=0,372). Este resultado indica que na amostra, 37,2% da variação da variável viagens domiciliares é explicado pelas variáveis independentes. Embora esse valor pareça baixo, para análises desagregadas (indivíduos ou domicílios), espera-se uma explicação menor do que no caso de análises agregadas por unidades de áreas (distritos, zonas de tráfego etc.), onde a variabilidade da variável dependente é menor, pois se

avalia um comportamento médio de residentes daquela unidade de área. Na literatura tradicional de demanda por transportes se encontram valores por volta de 0,3 para coeficientes de determinação para o caso de modelos lineares desagregados (Ashley, 1978; Atherton e Ben-Akiva, 1976; Bates *et al*., 1978). O modelo escolhido para este trabalho está descrito na Tabela 4.

Tabela 4 - Sumário do modelo.

Modelo RLM					
Variável dependente	**Viagens**				
Variáveis significativas	**R^2**	**Coeficiente**	**t**	**F**	**Sig.**
Constante		-1,021	-5,750		0,000
Morador		1,424	29,103		0,000
Renda	0,372	0,137	7,334	317,703	0,000
Automóveis		0,747	7,289		0,003
Motos		0,485	2,943		

Além disso, as variáveis selecionadas no modelo apresentaram valores de coeficientes, estatística t e teste F coerentes como esperado.

O teste t mostra que os parâmetros estimados, relativos às variáveis independentes selecionadas, foram significativos. Corroboram, com um elevado nível de segurança (95%), a hipótese de que os coeficientes das variáveis (morador, renda, automóveis e motocicletas) são significantemente diferentes de zero.

Os coeficientes das variáveis independentes da equação obtida são todos positivos. 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 automóveis e motocicletas que o domicílio possui.

A variável dependente neste modelo (viagens) apresentou correlação moderada com as variáveis: número de moradores, renda e quantidade de automóveis no domicílio (variáveis socioeconômicas).

Para comparação do modelo RLM e da técnica da AD, foram utilizadas as seguintes medidas de desempenho: erro médio, erro relativo e correlação, conforme apresentadas na Tabela 5. Foram observadas, em ambas as técnicas, resultados muito próximos para o procedimento de geração de viagens por domicílio, através das medidas utilizadas para a validação das técnicas abordadas neste trabalho.

Tabela 5 - Validação das técnicas abordadas.

Validação dos dados	**AD**		**RLM**	
	70%	*30%*	*70%*	*30%*
Erro médio	-0,018	0,292	0,000	0,224
Erro relativo	0,632	0,647	0,628	0,597
Correlação	0,607	0,603	0,610	0,643

5. Conclusões

O método tradicional possui algumas desvantagens, relacionadas principalmente à arbitrariedade de escolha das variáveis explicativas e, consequentemente, dos estratos domiciliares obtidos.

Através da técnica de AD, são obtidas as classes domiciliares, através de medida de impureza (Redução da Variância Iv), que possibilita minimizar a heterogeneidade de cada grupo. Neste trabalho, para o estudo de caso da cidade de São Carlos (São Paulo, Brasil), foram obtidos sete estratos considerando número de membros domiciliares, renda, posse de automóveis e motocicletas.

Vale ressaltar que a técnica apresentada pode ser uma ferramenta auxiliar na previsão de demanda por transportes, com bom poder preditivo e estrutura de modelo de fácil interpretação.

Apesar de desempenhos similares entre as técnicas abordadas neste trabalho, são conhecidas todas as desvantagens relacionadas à modelagem de geração de viagens domiciliares através de Regressão Linear Múltipla. Uma delas é que, normalmente, as observações de viagens são disponibilizadas na forma de contagens, o que implica distribuições de probabilidade assimétricas, isto é, distribuições não

normais para a variável resposta. Tendo em vista que as técnicas de regressão mais tradicionais não abrangem, em seu processo de calibração, tais características intrínsecas à demanda por transportes, é interessante o uso de técnicas não-paramétricas, tais como Árvore de Decisão, para o mesmo tipo de previsão. Assim, não são necessários ajustes de funções, restrições vinculadas à independência de erros ou distribuições de probabilidade das variáveis, além de problemas de dados multicolineares.

6. Agradecimentos

À Agência de fomento CNPq.

Referências

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

Atherton, T.J. and Ben-Akiva, M.E., Transferability and updating of disaggregate travel demand models. *Transportation Research Record*, 610, 12–18 (1976).

Barbetta, P. A., Estatística aplicada às ciências sociais. 8. ed., Florianópolis/SC, Editora UFSC (2012).

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

Breiman, L.; Friedman, J.H.; Olshen, R.A. e Stone, C.J., Classification and Regression Trees. *Wadsworth International Group*, Califórnia (1984).

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

Chang, J S, Jung, D., Kim, J. e Kang, T., Comparative analysis of trip generation models: results using home-based work trips in the Seoul metropolitan area. Transportation Letters: *The International Journal of Transportation Research*, 6 (2), 78-88 (2014).

Clark, S., National multi-modal travel forecasts. Literature review: aggregate models. Working Paper 465, Institute of Transportation Studies, University of Leeds (1996).

Fleet, C.; A. Sosslau. Trip generation procedures: An improved design for today's needs. *Institute of Transportation Engineering Journal* (1976).

Guevara C.A. e Thomas A., Multiple classification analysis in trip production models, *Transport Policy*, 14 (6), 514-522, http://dx.doi.org/10.1016/j.tranpol.2007.08.001 (2007).

Instituto Brasileiro de Geografia e Estatística – IBGE, Censo Demográfico Brasileiro (2010). Disponível em < http://www.ibge.gov.br>. Acesso em 03 de abril de 2014.

Kanafani A. Transportation demand analysis. New York, USA (2001).

Kass, G.V., An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29 (2), 119-127 (1980).

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

Loh W.Y. e Shih Y.S., Split selection methods for classification trees. *Statistica Sinica* 7, 815-840 (1997).

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

Ortúzar, J.D. e Willumsen, L.G., Modelling Transport. Londres: Wiley. 4ª ed. 586p (2011).

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

Peyrebrune, J. C., Trip generation: an informational report. Institute of Transportation Engineers, 5ª ed. (1985).

Quinlan, I.R., Learning Efficient Classification Procedures and their Application to Chess end-Games. *Machine Learning: An Artificial Intelligence Approach*, 463-482 (1983).

Rodrigues da Silva, A. N., Pesquisa origem-destino da cidade de São Carlos. Escola de Engenharia de São Carlos - Universidade de São Paulo-USP (2008).

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

Stopher, P.R.; Macdonald K.G., Trip Generation by Cross-Classification: An Alternative Methodology, *Transportation Research Record*, 84-91 (1983).

TMIP, Report on Findings of the Second Peer Review Panel for Southern California Association of Governments (SACG). US DOT Travel Model Improvement Program (2004).