The Journal of Transport Literature
www.journal-of-transport-literature.org
ITPS International Transport Planning Society

Aplicação conjunta de modelos não paramétricos e paramétricos para previsão de escolha modal

Cira Souza Pitombo[1,+]; Aline Schindler Gomes da Costa[2]
[1] *University of São Paulo, São Carlos, Brazil*
[2] *Federal University of Bahia, Salvador, Brazil*

Article Info

Palavras-chave:
Escolha modal
Árvore de Decisão
Regressão Linear Múltipla

Submitted 6 Jul 2014;
received in revised form 19 Nov 2014; accepted 21 Dec 2014.

Resumo

O presente trabalho visa apresentar uma metodologia realizada em duas etapas, envolvendo aplicação de modelos não paramétricos (Árvore de Decisão - AD) e paramétricos (Regressão Linear Múltipla-RLM) para previsão de escolha modal. A aplicação da AD permite encontrar relações entre variáveis socioeconômicas e escolha modal, bem como discretizar as variáveis numéricas e categóricas para construção dos modelos lineares na etapa posterior. Os dados utilizados para o desenvolvimento deste trabalho são provenientes da entrevista domiciliar da Pesquisa Origem-Destino de 2007/2008, realizada na cidade de São Carlos (SP). O modelo não paramétrico apresentou um total de acertos em torno de 70%, com alta associação entre valores categóricos estimados e observados do modo de transporte. Após transformação de todas as variáveis independentes em binárias através da técnica de AD, foram obtidos modelos lineares pelo método stepwise com bom poder preditivo para as três categorias de modo de transporte consideradas. Além disso, a validação dos modelos lineares com 30% da amostra restante apresentou baixos valores de erro médio e variância dos resíduos. Finalmente, o método proposto pode ser considerado razoável, sendo uma boa alternativa às abordagens tradicionais.

+ Corresponding author. Departamento de Engenharia de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo. Avenida Trabalhador Sãncarlense, 400 - Parque Arnold Schimidt, 13566-590 São Carlos - SP.
E-mail address: cira@sc.usp.br.

Introdução

O presente trabalho visa apresentar uma metodologia realizada em duas etapas, envolvendo aplicação de modelos não paramétricos (Árvore de Decisão - AD) e paramétricos (Regressão Linear Múltipla - RLM) para previsão de escolha modal. A aplicação da AD permite encontrar relações entre variáveis socioeconômicas e escolha modal, bem como discretizar as variáveis contínuas e categóricas para construção dos modelos lineares na etapa posterior. Desta forma, trata-se de uma abordagem exploratória-confirmatória que permite investigar padrões de comportamento dos indivíduos em relação à escolha modal, bem como testar a significância das variáveis em análise através de modelo paramétrico tradicional (RLM). O método proposto pode ser considerado eficiente, sendo uma boa alternativa às abordagens tradicionais, sobretudo em casos onde não há informações relativas a viagens tais como custo e tempo de viagem, imprescindíveis para construção das funções utilidade dos modelos logit.

Tradicionalmente, a escolha modal é influenciada pela distância, pelo custo da viagem e pela disponibilidade ou não de determinado modo de transporte, podendo ser avaliada de acordo com as condições socioeconômicas e a acessibilidade dos habitantes aos modos de transportes (Hutchinson, 1979; Ortúzar e Willumsen, 2011). Por décadas e até os dias atuais, diversos autores têm investigado fatores que influenciam a escolha modal, através de modelos tradicionais, tais como logit e probit, e técnicas de coleta de dados, tais como preferência declarada e revelada (Williams, 1978; Sen et al., 1978.; Ahern e Tapley, 2008; Grange et al., 2013). Caldas e Black (1997) propuseram uma metodologia para estimativa de escolha modal, através de uma amostra composta por variáveis binárias, provenientes de pesquisa de preferência revelada. Brownstone et al. (2000) compararam um modelo logit multinomial e um logit misto, com base em pesquisa domiciliar de preferência relevada e declarada, realizada na Califórnia (EUA).

A modelagem tradicional para estimativas da escolha modal é geralmente baseada nos princípios da maximização da utilidade, provenientes dos modelos econométricos. Adicionalmente às técnicas econométricas utilizadas para análise e modelagem do comportamento relativo a viagens, muitos outros estudos vêm aplicando modelos não paramétricos e exploratórios para investigação da escolha modal, por exemplo. A escolha do modo de transporte pode ser analisada como um problema de reconhecimento de padrões de indivíduos formados por variáveis explicativas e probabilidades de escolhas (Xie et al., 2007). Xie et al. (2007) investigaram o desempenho de duas técnicas não-paramétricas para modelar a escolha modal de viagens com motivo trabalho. As técnicas utilizadas e comparadas pelos autores foram Árvore de Decisão (AD) e Redes Neurais Artificiais. Shmueli et al. (1996) exploraram a aplicação de Redes Neurais Artificiais para comparação de padrões de viagens realizados por homens e mulheres em Israel.

Pitombo et al. (2011) analisaram relações entre variáveis socioeconômicas, de uso do solo, padrões de atividades e padrões de viagens encadeadas através da técnica de Árvore de Decisão (Classification and Regression Tree - CART algoritmo). Pitombo e Costa (2014) propuseram um método para estimação da escolha modal através da mesma técnica exploratória e dados desagregados socioeconômicos e de avaliação do sistema de transporte da cidade de São Carlos (SP).

Arentze and Timmermans (2007) propuseram um modelo baseado em regras para investigar escolha de atividades e viagens. Os autores propuseram uma abordagem híbrida, tal como no atual trabalho, realizando uma análise conjunta de modelos de Árvore de Decisão e logit. Pitombo et al. (2013) também recomendaram uso de técnicas exploratória e confirmatória, conjuntamente, a fim de comparar diferenças nos padrões de viagens de trabalhadores de dois setores diferentes (comércio e indústria), através de dados da Região Metropolitana de São Paulo (SP).

Este trabalho também propõe uma abordagem híbrida, unindo vantagens de técnicas não paramétricas e paramétricas. Este artigo está assim subdividido: Na Seção 1, serão apresentados conceitos básicos relativos às técnicas utilizadas; na Seção 2, é descrito o tratamento do banco de dados, bem como a área de estudo; na Seção 3 e na Seção 4, são descritas as aplicações das duas técnicas e resultados obtidos. Finalmente, as conclusões são apresentadas e, em seguida, as referências bibliográficas utilizadas.

1. Técnicas abordadas

As técnicas abordadas neste trabalho são de Análise Multivariada (AM) de dados. AM pode ser definida como um conjunto de técnicas estatísticas utilizadas com o objetivo de explicar e prever o grau de relações entre diversas variáveis independentes (inclusive entre si) e a variável dependente. Neste trabalho, a estimação da escolha modal será realizada através de aplicação conjunta e sequencial de Árvore de Decisão e Regressão Linear Múltipla, técnicas descritas nas subseções subsequentes.

1.1. Árvore de Decisão (AD)

A primeira técnica (não-paramétrica) utilizada neste trabalho para a análise e estimação da escolha modal é a Árvore de Decisão (AD). Considerada uma forma simples de representação de relação ou de relações existentes em um conjunto de dados. 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.

Para geração do modelo de AD foi utilizado o software SPSS 19.0 ©. A AD contida no SPSS é uma variante do algoritmo do CART (do inglês, Classification and Regression Tree). 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. A cada passo no crescimento da árvore, o particionamento dos dados se faz a partir da minimização do desvio em todas as divisões permitidas nos nós da árvore (Breiman et al., 1984). Essa redução de entropia corresponde à diminuição da aleatoriedade ou dificuldade de previsão de uma variável resposta.

A aplicação da AD neste trabalho tem duas funções: (1) Extrair padrões do banco de dados, formados por grupo de indivíduos com determinadas características (valores de variáveis independentes) e probabilidade de escolha do modo de transporte (variável dependente); (2) Auxiliar no aprimoramento de modelos lineares a partir da discretização das variáveis independentes. Assim, as variáveis contínuas e categóricas foram divididas em classes (binárias) com o intuito de reduzir o efeito da eventual não-linearidade na relação entre variáveis independentes e variável dependente. Cada classe foi associada a uma variável dummy e os valores para escolha das classes das variáveis dummy foram obtidos com a aplicação da AD, descrita mais detalhadamente na Seção 3 deste trabalho.

1.2. Regressão Linear Múltipla (RLM)

A segunda técnica (paramétrica) utilizada neste trabalho foi a Regressão Linear Múltipla (RLM), uma das mais utilizadas e versáteis técnicas tradicionais de AM. É aplicada em uma infinidade de casos, onde se deseja encontrar uma relação entre uma única variável dependente (numérica) e diversas variáveis independentes (numéricas ou dummy), supondo que esta relação seja linear.

A variável dependente é prevista a partir da combinação de todas as variáveis independentes multiplicadas por seus respectivos coeficientes, adicionada a um termo que representa o resíduo (Equação 1). A finalidade é encontrar a combinação linear das variáveis independentes que forneça máxima correlação com a variável dependente.

$$Yi = b_0 + b_1\, x_1 + b_2\, x_2 + \cdots + b_n x_n + \varepsilon_i \quad (1)$$

Em que:

- Yi é a variável dependente;
- b_1 é o coeficiente da primeira variável independente x_1 ;
- b_2 é o coeficiente da segunda variável independente x_2 ;
- b_n é o coeficiente da enésima variável independente x_n ;
- ε é a diferença entre o valor previsto de y e o valor observado considerando o indivíduo/objeto i.

Optou-se, neste trabalho, pelo método de Regressão Linear Múltipla Stepwise. O método Stepwise constrói, iterativamente, uma sequência de modelos de regressão pela adição ou remoção de variáveis em cada etapa, sendo o critério para a seleção ou remoção de variáveis, em qualquer etapa, o teste parcial F.

2. Dados

A área de estudo do presente trabalho é a cidade de São Carlos (São Paulo, Brasil). Com 221.936 habitantes, 96% da população residente na zona urbana e área urbana de aproximadamente 105 km^2 (IBGE, 2010).

Os dados utilizados para o desenvolvimento deste trabalho são provenientes das entrevistas domiciliares da Pesquisa Origem-Destino de 2007/2008, realizada na cidade de São Carlos (SP). Foram utilizados dados relativos a apenas um morador de cada domicílio entrevistado, gerando um banco com 2.791 casos. Isto representava, em 2008, uma amostra de praticamente 1,3% da população da cidade. Foram excluídos alguns registros, por problemas de incoerência nos dados, totalizando uma amostra final composta por 1.216 indivíduos.

3. Aplicação da AD: modelo não paramétrico

Na etapa da aplicação do modelo não paramétrico deste trabalho, foi gerada a AD com a amostra final de 1.216 indivíduos, adotando o mínimo de 50 observações por nó terminal. A variável dependente é formada por três categorias: (1) Modo público; (2) Modo particular motorizado; (3) Modo não motorizado. As variáveis independentes (categóricas e numéricas) são listadas a seguir.

- Socioeconômicas (Quantidade de motocicletas no domicílio; Quantidade de automóveis no domicílio; Idade; Possui Carteira de Habilitação; Sexo; Grau de instrução; Estuda; Condição de atividade; Renda; Ocupação).

Para validação do modelo de AD, a amostra foi dividida aleatoriamente. 70% da amostra foi destinada para calibração do modelo ou treinamento da AD, enquanto que o restante (30%) foi destinado para validação do modelo calibrado ou teste da AD obtida.

A variável de maior importância (que melhor explica a variabilidade dos dados) é "Possui Carteira Nacional de Habilitação (CNH)". A partir da raiz, a árvore se ramifica em dois grupos principais: (1) Não Possui CNH e (2) Possui CNH.

Posteriormente, ocorrem novas segmentações do conjunto de dados. Ao final da segregação dos dados foi encontrado um total de oito folhas. A Figura 1 representa a árvore treinada para investigação da escolha modal. Nas folhas encontram-se ilustradas as categorias da variável dependente e a frequência de cada categoria em cada nó.

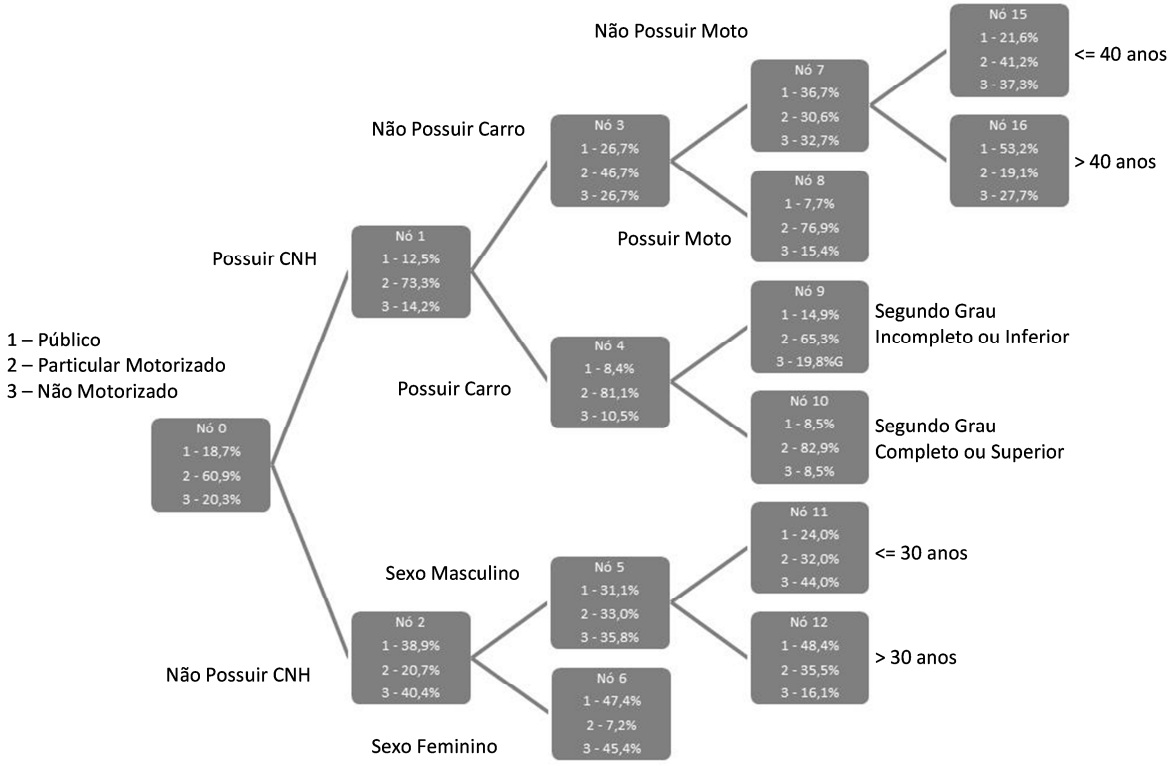

Figura 1 - Árvore treinada (70%) da amostra

Observam-se, através da AD representada na Figura 1, as variáveis selecionadas, bem como sua influência no uso (ou não) de cada um dos três modos de transporte considerados. Ter Carteira Nacional de Habilitação (CNH), por exemplo, influencia positivamente o uso do modo de transporte particular motorizado (Nó 2, 73,3% utiliza automóvel ou motocicleta), assim como grau de instrução superior (Segundo grau completo ou superior – Nó 10, 82,9% utiliza automóvel ou motocicleta). Indivíduos que não possuem CNH e são do sexo feminino são mais propensos ao uso do modo de transporte não motorizado ou Público (Nó 6). Já indivíduos com CNH, mas sem automóveis ou motocicletas no domicílio, utilizam predominantemente o modo de transporte público (Nó 7). Verificou-se também uma boa porcentagem de acertos de estimação para os três modos analisados tanto para a amostra de treinamento quanto para a amostra teste.

Tabela 1 - Variáveis selecionadas pela AD e sua respectiva influência na escolha modal

Variáveis selecionadas	Particular motorizado	Público	Não motorizado
Possuir CNH	Sim	Não	Não
Quantidade de Automóveis	Pelo menos 1 auto	zero auto	zero auto
Quantidade de Motocicletas	Pelo menos 1 moto	zero moto	zero moto
Idade	<= 40 anos	> 40 anos	<= 30 anos
Sexo	Masculino	Feminino	Feminino
Grau de Instrução	2º grau completo/superior	2º grau incompleto/inferior	2º grau incompleto/inferior

Ainda com a finalidade de testar a qualidade do modelo não paramétrico, foi realizado o teste qui-quadrado para testar a significância da associação entre os valores categóricos observados do modo de transporte escolhido e os valores estimados

pelo modelo. Assim, usando um nível usual de significância de 5% e para gl=4, foi encontrado um alto valor da estatística qui-quadrado ($\chi 2$=396) e corroborada a hipótese de associação entre valores estimados pela AD e observados.

3.1. Discretização das variáveis independentes

As variáveis numéricas e categóricas foram divididas em classes. Cada classe foi associada a uma variável dummy e os valores para escolha das classes das variáveis dummy foram obtidos com a aplicação da AD, conforme descrição a seguir: (possuir CNH = 1, não possuir CNH= 0, ter automóvel = 1, não ter automóvel = 0, Homem = 1, Mulher = 0, > 30 anos = 1, <= 30 anos = 0, 2º grau completo ou superior=1 e ter 2º grau incompleto ou inferior = 0).

4. Aplicação da RLM: modelo paramétrico

Visando corroborar as afirmações da etapa não paramétrica e buscando associar inferência estatística às variáveis independentes, a etapa descrita nesta seção consiste na aplicação da Regressão Linear Múltipla (RLM) para estimação de parâmetros que evidenciem a contribuição das variáveis independentes na escolha dos modos de transporte.

A aplicação de RLM deu-se em duas etapas,: (1) aplicação do modelo de RLM com variáveis originais; (2) aplicação do modelo de RLM com todas as variáveis binárias. Como mencionado anteriormente, o processo de discretização foi auxiliado pela aplicação da AD.

4.1. RLM: variáveis originais

A primeira etapa de obtenção de três modelos lineares (para cada modo de transporte considerado) foi realizada a partir das variáveis dependentes e independentes. Para evitar problemas de multicolinearidade, variáveis independentes altamente correlacionadas foram descartadas dos modelos. Desta forma, se duas variáveis independentes tinham alta correlação, era descartada da análise aquela (dentre as duas altamente correlacionadas) que possuía menor correlação com a variável dependente.

A Tabela 2 traz cada um dos três modelos lineares obtidos pelo método stepwise, valores de R2, coeficientes e valores de estatística t associados a cada uma das variáveis. Em todas análises, foi apresentado o melhor modelo obtido pelo método stepwise. As relações encontradas na etapa não paramétrica foram corroboradas pela RLM, considerando os coeficientes obtidos e os valores da estatística t. No entanto, o poder explicativo dos modelos aqui obtidos foi considerado baixo, com coeficiente de determinação variando entre 0,133 (modo não motorizado) e 0,321 (moto particular motorizado).

Tabela 2 - Resultados da aplicação da RLM – etapa 1

Público		**cte**	**AUT**	**MOTO**	**RENDA 0-2SM**	**Mulheres**	**2º grau/superior**	**Nº viagens**	**Não Alfabetizado**	**R^2**
	B	0,157	-0,107	-0,088	0,072	0,102	-0,052	-0,025	0,182	0,139
	est t	3,901	-6,571	-3,972	3,210	3,861	-2,231	-3,319	2,193	
Particular motorizado		**cte**	**AUT**	**MOTO**	**CNH**	**Mulheres**	**2º grau/superior**	**Nº viagens**	**1º a 4º série**	**R^2**
	B	0,886	0,176	0,112	0,250	-0,156	0,141	0,027	-0,320	0,321
	est t	12,582	9,193	4,345	7,505	-4,975	5,500	3,048	-2,074	
Não motorizado		**cte**	**AUT**		**CNH**	**Mulheres**			**1º a 4º série**	**R^2**
	B	-0,119	-0,062		-0,218	0,059			0,438	0,133
	est t	-2,249	-3,550		-7,702	2,013			3,014	

Considerando os resultados dos modelos lineares obtidos com as variáveis originais, observou-se um valor de R2 pequeno para previsão dos três modos de transportes. Contudo, as relações entre as variáveis socioeconômicas e as escolhas modais comprovam resultados obtidos na etapa não paramétrica: (1) A quantidade de automóveis influencia positivamente o uso do modo de transporte particular motorizado e negativamente os demais modos; (2) A quantidade de motocicletas influencia positivamente o uso do modo de transporte particular motorizado e negativamente o modo público; (3) Possuir CNH influencia forte e positivamente o uso do modo de transporte particular motorizado e negativamente o modo não motorizado; (4) Mulheres são mais propensas ao uso do modo de transporte público e não motorizado; (5) Grau de instrução mais alto influencia positivamente o uso do modo particular motorizado e negativamente o modo público.

4.2. RLM: variáveis discretizadas

Com objetivo de melhorar as estimativas obtidas através dos modelos lineares, optou-se pelo uso apenas de variáveis binárias. As variáveis binárias independentes foram obtidas com auxílio da técnica não paramétrica, Árvore de Decisão, A Tabela 3, em seguida, apresenta os principais resultados da RLM. Verifica-se a comprovação das relações obtidas entre variáveis socioeconômicas e escolhas modais. Adicionalmente, os modelos apresentaram uma melhora significativa (R2), apresentando baixos valores de média dos erros e variância dos resíduos na sua validação e realizada com 30% da amostra total.

Tabela 3 - Resultados da aplicação da RLM – etapa 2

Público		**cte**	**CNH**	**GI**	**SEXO**	**IDADE1**	**R^2**
	B	0,370	-0,141	-0,109	-0,101	0,066	0,652
	est t	27,403	-14,772	-13,255	-10,532	5,709	
Particular motorizado		**cte**	**CNH**	**GI**	**SEXO**	**IDADE1**	**R^2**
	B	0,158	0,399	0,215	0,138	0,088	0,829
	est t	8,917	31,961	19,899	10,939	5,794	
Não motorizado		**cte**	**CNH**	**GI**	**SEXO**	**IDADE1**	**R^2**
	B	0,493	-0,271	-0,074	-0,048	-0,034	0,878
	est t	53,247	-42,543	-8,801	-7,613	-4,532	

CNH: 0 – Não; 1 – Sim; GI (Grau de Instrução): 0 – 2º grau incompleto/inferior; 1 – 2º grau completo/superior; SEXO: 0 – Mulheres; 1 – Homens; IDADE1 : 0 - <= 30 anos; 1 - > 30 anos.

Nos três casos, as variáveis dependentes foram explicadas no mínimo em 65,2% (Público) até no máximo de 87,8% (Não motorizado). Observa-se que as variáveis além de serem significativas, comprovam relações encontradas anteriormente.

CNH é a variável mais importante na escolha modal (maiores valores de parâmetros calibrados e estatística t). Possuir CNH influencia forte e positivamente o uso do modo particular motorizado, enquanto que influencia forte e negativamente o uso dos demais modos. O Grau de Instrução - 2º completo ou superior influencia forte e positivamente o uso do automóvel e negativamente o uso dos demais modos; Homens têm maior propensão para o uso do automóvel ou motocicleta, enquanto

que mulheres possuem maior tendência para uso do modo público ou não motorizado; Pessoas com idade inferior a 30 anos são mais propensas para uso do modo não motorizado.

Conclusão

Este trabalho propôs uma metodologia alternativa às abordagens tradicionais para modelagem e estimação de escolha modal. O método é formado por duas etapas principais: modelagem não paramétrica e modelagem paramétrica.

Na etapa de modelagem não paramétrica foi utilizada a técnica de mineração de dados, conhecida como Árvore de Decisão. Foram obtidos oito grupos de indivíduos (nós terminais ou folhas) com diferentes características socioeconômicas e diferentes probabilidades de escolhas modais. Através dos resultados da AD foram encontradas influências (positivas e negativas) nas escolhas de determinado modo de transporte. Possuir CNH, automóvel ou motocicleta no domicílio, Grau de instrução superior ou segundo grau completo e ser do sexo masculino, são características que influenciam o uso do modo particular motorizado, por exemplo. Além disso, a aplicação da AD permitiu a discretização das variáveis numéricas e categóricas para construção dos modelos lineares na etapa seguinte. Vale ressaltar ainda que o modelo não paramétrico apresentou um alto percentual de acertos de escolhas modais.

Visando corroborar as relações inicialmente obtidas e buscando associar inferência estatística às variáveis independentes, foi realizada a modelagem paramétrica. Os modelo lineares obtidos com variáveis independentes categóricas e numéricas (como na Pesquisa OD de São Carlos) não mostraram um bom poder preditivo. No entanto, os modelos obtidos considerando a discretização realizada com auxílio da AD foram considerados razoáveis, com as variáveis independentes significativas, altos valores de coeficiente de determinação, além de baixos valores de erro médio e variância dos resíduos.

Acknowledgements
Ao CNPq - Conselho Nacional de Ensino e Pesquisa – pelo suporte financeiro fornecido à presente pesquisa.

Referências

Ahern, A., & Tapley, N. (2008) The use of stated preference techniques to model modal choices on interurban trips in Ireland. Transportation Research Part A: Policy and Practice, 42(1), 15-27. DOI: 10.1016/j.tra.2007.06.005.

Arentze, T., & Timmermans, H. (2007) Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. Transportation Research Part B: Methodological, 41(7), 772-783. DOI: 10.1016/j.trb.2007.01.001.

Breiman, L., Friedman J.H, Olshen R.A., & Stone C.J. (1984) Classification and Regression Trees. Wadsworth International Group, Califórnia. Brownstone, D.; Bunch, D.; Train, K. (2000) Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research Part B: Methodological, 34(5), 315-338. DOI: 10.1016/S0191-2615(99)00031-4.

Caldas, M. A. F., & Black, I. G. (1997) Formulating a methodology for modelling revealed preference discrete choice data - the selectively replicated logit estimation. Transportation Research Part B: Methodological, 31(6), 463-472. DOI: 10.1016/S0191-2615(97)00008-8.

Grange, L., González, F. , Vargas, I., & Umñoz, J.C. (2013) A polarized logit model. Transportation Research Part A: Policy and Practice, 53, 1-9. DOI: 10.1016/j.tra.2013.06.003.

Hutchinson, B. G. (1979). Princípios de Planejamento dos Sistemas de Transporte Urbano. Rio de Janeiro: Guanabara Dois.

IBGE (2010) Instituto Brasileiro de Geografia e Estatística Cidades. Disponível em: www.ibge.gov.br/cidadesat. Acesso em: 31 abr 2010.

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

Pitombo, C. S., Kawamoto, E., & Sousa, A. J. (2011) . An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. Transport Policy (Oxford), 18, 347-357. DOI: 10.1016/j.tranpol.2010.10.010

Pitombo, C. S., Kawamoto, E., & Sousa, A. J. (2013) Linking activity participation, socioeconomic characteristics, land use and travel patterns: a comparison of industry and commerce sector workers. Journal of Transport Literature, 7, 59-86.

Pitombo, C. S., Costa, A. S. G. (2014) . Decision Tree application for modal choice. In: Panam 2014, 2014, Santander. Panam 2014.

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

Sen, A., Soot, S., & Pagitsas, E. (1978) The logit modal split model: Some theoretical considerations. Transportation Research, Part A, 12(5), 321-324. DOI: 10.1016/0041-1647(78)90006-0.

Shmueli, D., Salomon, I., & Shefer, D. (1996) Neural network analysis of travel behavior: Evaluating tools for prediction. Transportation Research Part C: Emerging Technologies, 4(3), 151-166.

Williams, M. (1978) Factors affecting modal choice decisions in urban travel: Some further evidence. Transportation Research, Part A, 12(2), 91-96. DOI: 10.1016/0041-1647(78)90047-3.

Xie, C., Jinyang, L., & Parkany, E. (2007) Work Travel mode choice modeling with data mining: Decision Trees and Neural Networks. Transportation Research Record: Journal of the Transportation Research Board, 1854, 50-61. DOI: 10.3141/1854-06.

Abstract

This paper presents a two-step methodology, involving application of nonparametric (Decision Tree - DT) and parametric models (Multiple Linear Regression - MLR) to forecast modal choice. The DT application allows finding relationships between socioeconomic variables and modal choice and discretizing the numerical and categorical variables for the construction of linear models in the later stage. The data used are from the Origin-Destination Survey carried out in 2007/2008 in the city of São Carlos (SP). The accuracy of the nonparametric model is around 70%, with high association between estimated and observed categorical values of the travel mode. After discretizing all independent variables using DT, stepwise linear models, with a good accuracy for the three categories of travel mode, were obtained. Moreover, the validation of linear models with 30% of remaining sample showed low average of errors and variance of residuals. Finally, the proposed and reasonable method could be a good alternative to traditional approaches.

Key words: modal choice; Decision Tree; Multiple Linear Regression.