AVALIAÇÃO DA ESCOLHA MODAL PARA O TRANSPORTE FERROVIÁRIO DE PASSAGEIROS NA REGIÃO SUDESTE ATRAVÉS DE ÁRVORES DE DECISÃO

Cassiano Augusto Isler
Cira Souza Pitombo
Universidade de São Paulo
Escola de Engenharia de São Carlos
Departamento de Engenharia de Transportes

RESUMO
O investimento na infraestrutura de transporte ainda é um desafio para melhoria da eficiência e qualidade dos deslocamentos da população brasileira. Mesmo com atuais propostas de projetos de sistemas de Trens de Alto Desempenho e Trens de Alta Velocidade para o transporte intermunicipal de passageiros na Região Sudeste, existe uma deficiência na representação adequada da escolha modal para novos sistemas dessa natureza. Assim, o objetivo deste artigo é avaliar a escolha modal em cenários de novos serviços ferroviários utilizando informações de uma pesquisa de preferência declarada e algoritmos de Árvores de Decisão. Os resultados da modelagem indicam que o algoritmo *CART* provê o maior índice de acertos em relação às respostas da pesquisa e que, nas condições avaliadas, a inclusão de variáveis socioeconômicas não aumenta esses acertos. As árvores de decisão em diferentes cenários são geradas em função dos atributos das viagens e dos modos de transporte.

ABSTRACT
The investment in the transport infrastructure is still a challenge to increase the efficiency and quality of the mobility in Brazil. Despite the current proposals of new High Performance Rail and High Speed Rail projects for passenger transport, there is a deficiency to represent the modal split of such new services adequately. In this paper we evaluate the modal split of new railway passenger scenarios in the Southeastern Region of Brazil based on the information of a stated preference survey and Decision Trees techniques. The modeling results indicate that the CART method provides the highest accuracy compared to the answers of the survey and that the inclusion of socioeconomic variables does not increase the matching rate. The decision trees of different scenarios are available based on the attributes of the trips and the travel mode.

1. INTRODUÇÃO

Apesar da estabilidade econômica do Brasil nos últimos 20 anos, o investimento sistemático na modernização da infraestrutura existente e construção de novos sistemas de transportes ainda é um desafio para aumento da eficiência e qualidade dos deslocamentos da população brasileira. Recentemente o Governo Federal apresentou um projeto de Trem de Alta Velocidade para conexão das cidades de Campinas, São Paulo e Rio de Janeiro com veículos caracterizados por velocidade média 300 km/h (TAV Brasil, 2014). Entretanto, o projeto não apresenta perspectivas de operação em um futuro próximo dadas, entre outras causas, as dúvidas a respeito da real demanda que possa utilizar o sistema.

Nessa mesma conjuntura, sob a representação da Companhia Paulista de Trens Metropolitanos, o Governo do Estado de São Paulo vem apresentando propostas de conexões da Região Metropolitana de São Paulo às cidades de Campinas, Sorocaba, São José dos Campos e Santos (CPTM, 2014) através de trens caracterizados por velocidade média de 150 km/h, doravante denominados neste artigo como Trens de Alto Desempenho.

Diferentemente de muitos países europeus em que a malha ferroviária de trens convencionais está consolidada e existe uma propensão à melhoria infraestrutura pela operação de trens de alto desempenho - e instalação e expansão das malhas de alta velocidade -, a inexistência de serviços competitivos dessa natureza no Brasil dificulta a estimativa da divisão modal no caso de implantação de novos sistemas. Assim, no contexto de incertezas a respeito da divisão modal nos cenários de reativação do transporte ferroviário de passageiros, o objetivo deste

artigo é avaliar a propensão à escolha modal por Trens de Alto Desempenho ou de Alta Velocidade a partir de dados de uma pesquisa de preferência declarada conduzida na Região Sudeste e resultados da aplicação de diferentes algoritmos de Árvores de Decisão (AD).

A maioria dos modelos de análise de escolha modal é baseada nos princípios de maximização da utilidade. Em geral, os autores investigam fatores que influenciam a escolha modal através de técnicas de coleta de dados como pesquisas de preferência declarada e revelada, e calibração de modelos como *Logit* e *Probit* (Ahern e Tapley, 2008; Grange *et al*., 2013). No entanto, a modelagem da divisão modal pode ser definida como um problema de reconhecimento de padrões comportamentais e definição de probabilidades de escolhas entre diferentes alternativas (Xie *et al*., 2007). Assim, o uso de técnicas de mineração de dados (como algoritmos de Árvore de Decisão) para avaliar o comportamento individual relativo a viagens pode ser uma alternativa às abordagens econométricas clássicas.

2. ALGORITMOS DE ÁRVORES DE DECISÃO

O desenvolvimento de mineradores de dados emergiu a partir da década de 1990 com o objetivo de apresentar informações úteis a partir de um grande conjunto de dados através de regras que caracterizam padrões de ocorrência (Mannila, 1997). A Árvore de Decisão é uma técnica não paramétrica possível para avaliação da escolha modal, sendo considerada uma forma simples de classificação das relações entre atributos do conjunto de dados. Uma AD permite classificar uma base de dados em um número finito de classes através de regras hierárquicas e da sua divisão em grupos, organizando os dados de maneira compacta e permitindo uma visão geral da natureza do processo (Quinlan, 1983).

A hierarquia resultante da classificação é denominada árvore e cada segmento é denominado nó. O segmento original (nó raiz) contém o conjunto dos dados que podem ser subdivididos em outros sub-nós (nós filhos) e em um nó terminal ou folha, quando estes não podem ser mais subdivididos com base em um critério preestabelecido. O algoritmo usado para dividir os dados nos modelos de árvore identificam as variáveis independentes que maximizam a segregação dos dados segundo a variável dependente. Alguns dos algoritmos existentes são o *C4.5* (Quilan, 1993), *CHAID* (Kass, 1980) e *CART* (Breiman *et al*., 1984) e *QUEST* (Loh e Shih, 1997).

De modo geral, o algoritmo de AD torna os subconjuntos resultantes cada vez mais homogêneos em relação à variável resposta mediante sucessivas criações de subconjuntos. A cada passo no crescimento da árvore, o particionamento dos dados se faz a partir da minimização do desvio em relação à variável dependente nas divisões dos nós da árvore, buscando-se a diminuição da aleatoriedade na previsão de uma variável resposta – (Breiman *et al.*, 1984). Neste artigo, três algoritmos para geração de AD (*CHAID*, *CART* e *QUEST*) são comparados para avaliação da escolha modal de potenciais novos serviços de transporte ferroviário de passageiros na Região Sudeste.

3. PESQUISA DE PREFERÊNCIA DECLARADA

Ortúzar and Willumsen (2011) definem três classes de pesquisa de preferência declarada: *Contingent Valuation* (*CV*), *Conjoint Analysis* (*CA*) and *Stated Choice* (*SC*). Na *CV*, os respondentes são questionados sobre o quanto pagariam por um serviço ou produto diferente daquele que já é ofertado (Mitchell e Carson, 1989). A *CA* sugere que os participantes ordenem uma série de alternativas quanto à sua preferência e a *SC* propõe que o respondente

escolha uma alternativa dentre novos serviços (ou produtos) e aqueles existentes segundo diferentes restrições estabelecidas pelo analista.

Nesse artigo descreve-se uma *SC* em que os participantes devem indicar a escolha modal em situações hipotéticas de operação de novos serviços ferroviários intermunicipais de transporte público de passageiros. A pesquisa destina-se à identificação da propensão à escolha modal em viagens de longa distância (entre 100 km e 1000 km) envolvendo os modos Automóvel, Ônibus, Trem (Alto Desempenho ou Alta Velocidade) e Avião. Para facilidade de representação algumas denominações são previamente definidas no contexto da língua portuguesa e representadas por siglas que remetem à sua tradução em inglês. No caso dos modos de transporte são atribuídas as respectivas denominações *CAR*, *BUS*, *HPR* (para *High Performance Rail*), *HSR* (para *High Speed Rail*) e *AIR*.

Inicialmente, é preponderante considerar que em algumas viagens de longa distância o modo aéreo não está disponível pela ausência de infraestrutura aeroportuária. Ainda, dado que o motivo da viagem é um fator importante na escolha modal e que afeta os resultados de possíveis aplicações da pesquisa, optou-se pela segmentação dos respondentes que realizaram viagens por motivo trabalho (*Work,* representado por "*W*") ou lazer, estudos e outros (genericamente denominado como *Leisure* e representado por "*L*"). Assim, considerando a oferta de serviços ferroviários (*HPR* ou *HSR*), os motivos de viagem e a disponibilidade do modo aéreo, são definidos oito cenários para formulação da pesquisa de preferência declarada: *HPR without AIR, HPR with AIR, HSR without AIR* e *HSR with AIR* pelo motivo *W*, e *HPR without AIR, HPR with AIR, HSR without AIR* e *HSR with AIR* pelo motivo *L*.

3.1 Características dos modos

3.1.1 Atributos

Para evitar o foco em apenas alguns atributos que caracterizam as viagens pelos diferentes modos de transporte (Carson *et al.*, 1994) e respostas aleatórias pelo excesso de informações (Sælensminde, 1999), o número máximo de atributos em cada um deles foi limitado a três, de forma a viabilizar o questionário em termos de número de perguntas apresentadas aos respondentes e adequação à pouca experiência da população brasileira no uso de transporte ferroviário intermunicipal.

O modo *CAR* foi caracterizado pelo Tempo de Viagem (*Travel Time* - TT_{CAR}), custo de combustível (*Petrol* - PE_{CAR}) e custo de pedágio (*Toll* - TO_{CAR}) e os modos *BUS*, *HPR*, *HSR* e *AIR* foram caracterizados pelo tempo de viagem (*Travel Time* - TT_{BUS}, TT_{HPR}, TT_{HSR}, TT_{AIR}), frequência em termos de intervalo entre serviços (*Frequency* - FR_{BUS}, FR_{HPR}, FR_{HSR}, FR_{AIR}) e preço da passagem (*Fare* – FA_{BUS}, FA_{HPR}, FA_{HSR}, FA_{AIR}). Apesar de nem todas as rodovias da Região Sudeste serem operadas pela iniciativa privada, os custos dos automóveis foram divididos em combustível e pedágio pela tendência de expansão dos modelos de concessões.

3.1.2. Níveis

A cada um dos atributos dos respectivos modos de transporte foram atribuídos três níveis de valores numéricos (Baixo, Médio e Alto) de modo a captar a variabilidade do atributo em função da decisão dos respondentes. O tempo de viagem foi definido com base na distância (dependendo das conexões entre cidades definidas na próxima seção) e velocidade média dos veículos conforme a Tabela 1, imputando nisso as características dos veículos, os limites de velocidade impostos pela lei e os atrasos eventuais devido a congestionamentos, condições

climáticas adversas e restrições operacionais. A frequência dos serviços de ônibus, trens e avião também foi definida em três níveis em função do intervalo entre partidas sucessivas, com valores "Baixa" para partidas a cada 12 horas, "Média" a cada 6 horas e "Alta" a cada 3 horas.

Tabela 1: Níveis de velocidade média para definição do tempo de viagem

Nível	*CAR*	*BUS*	*HPR*	*HSR*	*AIR*
Baixo	70	60	100	200	400
Médio	90	75	150	250	500
Alto	110	90	200	300	600

Por fim, os custos da viagem foram estabelecidos em função da distância de viagem e de custos médios para os atributos de cada modo, conforme a Tabela 2. Para os automóveis, os níveis de custo de combustível (PE_{CAR}) foram estabelecidos com base em valores médios informados pela Agência Nacional de Petróleo (ANP, 2013) e os níveis de pedágio (TO_{CAR}) com base nos preços médios praticados nas rodovias atualmente concessionadas pela iniciativa privada, fornecidos pela Associação Brasileira de Concessionárias de Rodovias (ABCR, 2013). Os preços das passagens de ônibus foram definidos pelas tarifas médias autorizadas pela Agência Nacional de Transporte Terrestre (ANTT, 2013) e os das passagens de avião com base nas regulações da Agência Nacional de Aviação Civil (ANAC, 2013).

Há uma dificuldade na definição do preço da passagem dos trens intermunicipais uma vez que esses serviços são praticamente inexistentes no país. Assim, os níveis desses atributos foram adotados a partir de estudos de *benchmarking* das tarifas praticadas nos sistemas ferroviários dos países europeus para trens convencionais e de alto desempenho (MVAConsultancy, 2013) e trens de alta velocidade (Prodan, 2011). Para evitar a dominância de um modo quanto ao custo de viagem, os preços dessas passagens foram estabelecidos no intervalo entre o custo total das viagens por automóvel e o preço da passagem de avião, uma vez que os tempos de viagem calculados com base na Tabela 1 diminuem na ordem decrescente desses modos.

Tabela 2: Cálculo dos níveis dos atributos de custo em função da distância de viagem (R$)

NÍVEL	*CAR*[1]	*BUS*[2]	*HPR*[2]	*HSR*[2]	*AIR*[2]
BAIXO	[*DI*/10] · 2,5 + *DI* · 0.125	*DI* · 0,1814	*DI* · 0,620	*DI* · 0,676	*DI* · 1,084
MÉDIO	[*DI*/10] · 3,0 + *DI* · 0,166	*DI* · 0,2002	*DI* · 0,804	*DI* · 0,794	*DI* · 1,285
ALTO	[*DI*/10] · 3,5 + *DI* · 0,220	*DI* · 0,2918	*DI* · 0,919	*DI* · 0,984	*DI* · 1,527

[1][DISTÂNCIA (*DI*-km)/DESEMPENHO AUTOMÓVEL (DM-km/litro)]. Combustível (R$/km) + DISTÂNCIA (*DI*-km). Pedágio (R$/km); [2]DISTÂNCIA (*DI*-km). Passagem (R$/km).

3.2 Planejamento do Experimento

Inicialmente foram definidas as cidades de origem e destino com maior volume de passageiros em voos domésticos regulares (ANAC, 2012): Bauru (BAU), Belo Horizonte (Pampulha - PLU), Campinas (VCP), Campos dos Goytacazes (CAW), Ipatinga (IPN), Juiz de Fora (JDF), Macaé (MEA), Presidente Prudente (PPB), Ribeirão Preto (RAO), Rio de Janeiro (Santos Dumont - SDU), São José do Rio Preto (SJP), São Paulo (Congonhas – CGH), Uberlândia (UDI) e Vitória (VIX). As 38 conexões mais representativas (Tabela 3) em termos de população das cidades e número de passageiros transportados foram escolhidas para garantir a diversificação das conexões avaliadas (*i.e.*, a variabilidade dos níveis dos atributos), considerando as distâncias aproximadas pelas distâncias rodoviárias entre cidades.

O planejamento do experimento visa definir a combinação de níveis dos atributos dos modos de transporte que são apresentadas aos respondentes em cada situação hipotética de uma pesquisa (ou linha do experimento). A abordagem de fatorial completo (*Full Factorial*) contempla todas as possibilidades de combinações de níveis dos atributos e apresenta crescimento exponencial do número de linhas do experimento (questões apresentadas aos respondentes) na inserção de novos atributos e níveis. Por outro lado, o planejamento fatorial fracionado (*Fractional Factorial*) é definido como um subconjunto das combinações do fatorial completo, podendo ser escolhidas aleatoriamente ou dividindo-se essas combinações em conjuntos e atribuindo cada um a diferentes respondentes (Ortúzar e Willumsen, 2011).

Tabela 3: Cidades de origem, destino, distâncias rodoviárias (*DR*) e distâncias estabelecidas (*DI*) para definição da pesquisa de preferência declarada

Origem	Destino	Distância		Origem	Destino	Distância		Origem	Destino	Distância	
		DR	*DI*			*DR*	*DI*			*DR*	*DI*
VCP	CGH	103	100	SDU	PLU	437	440	VCP	PLU	579	580
CGH	VCP	103	100	VCP	SDU	430	440	PLU	SJP	671	670
CAW	MEA	105	100	IPN	VIX	430	440	RAO	SDU	715	715
SDU	JDF	178	235	PLU	SDU	437	440	SDU	RAO	715	715
IPN	PLU	235	235	RAO	PLU	516	520	CAW	CGH	712	715
PLU	IPN	235	235	VIX	SDU	516	520	CGH	IPN	801	800
CAW	VIX	243	235	PLU	VIX	527	520	IPN	CGH	802	800
VCP	BAU	261	235	SDU	VIX	516	520	VIX	CGH	941	940
CAW	SDU	275	320	VIX	PLU	527	520	CGH	VIX	941	940
RAO	CGH	318	320	CAW	PLU	525	520	PLU	PPB	956	955
CGH	RAO	318	320	IPN	SDU	563	565	SDU	UDI	982	980
SDU	CGH	430	440	CGH	PLU	584	580	VCP	VIX	999	1000
CGH	SDU	430	440	PLU	CGH	584	580				

O *software* SPSS® (SPSS, 2012) foi utilizado no planejamento do experimento devido à disponibilidade de licença e sua adequação aos propósitos da pesquisa na aplicação da abordagem fatorial fracionado ortogonal. Assim, para cada par OD da Tabela 3, a execução do planejamento ortogonal pelo referido *software* resultou em 27 linhas (situações hipotéticas para escolha modal) para os cenários *HPR with AIR*, *HPR without AIR*, *HSR with AIR* e *HSR without AIR*. Apesar de todas as cidades possuírem aeroportos, o primeiro e o terceiro cenário foram estabelecidos sem a existência do modo aéreo para que fosse possível a modelagem da divisão modal entre cidades que não possuem este modo disponível.

Considerando que 27 questões apresentadas para um mesmo respondente são excessivas, possivelmente comprometendo a eficiência da pesquisa, cada cenário foi dividido em blocos de nove questões como exemplificado na Tabela 4. Assim, uma replicação do experimento implica na identificação de 24 participantes (três blocos para os quatro cenários pelos dois motivos de viagem) e resulta em 216 escolhas modais.

A plataforma *on-line* de formatação de pesquisas *Qualtrics* (*qualtrics.com*) foi utilizada para diagramação do experimento dada a vasta disponibilidade de recursos visuais e o acesso a uma licença comercial. Na execução da pesquisa foi de interesse identificar participantes que tivessem realizado pelo menos uma viagem entre uma das cidades de origem e destino definidas anteriormente, pois a relevância das respostas é diretamente relacionada ao conhecimento do usuário sobre as características das viagens.

Tabela 4: Exemplo de bloco para um cenário *HSR WITH AIR* de distância de 580 km (*e.g.*, entre Campinas-SP e Belo Horizonte-MG)

TT_{CAR}	TT_{BUS}	TT_{HSR}	TT_{AIR}	FR_{BUS}	FR_{HSR}	FR_{AIR}	FA_{BUS}	FA_{HSR}	FA_{AIR}	PE_{CAR}	TO_{CAR}
6h50	8h15	2h55	1h10	12	6	6	170	570	750	150	70
6h50	7h15	2h20	1h0	12	12	3	170	460	750	200	100
6h30	7h45	2h20	1h10	3	12	6	110	390	750	150	100
6h5	7h15	1h55	1h25	3	6	3	170	570	890	150	100
6h50	8h15	2h55	1h25	6	12	3	110	390	890	150	130
6h50	7h45	1h55	1h10	3	12	3	120	570	630	170	70
6h5	7h15	1h55	1h0	12	12	12	110	390	630	150	70
6h30	8h15	1h55	1h10	12	6	12	120	390	890	200	100
6h50	7h15	2h20	1h25	3	6	6	120	390	630	200	130

3.3 Aplicação do Questionário

Após uma explicação do propósito da pesquisa e a sequência do questionário, o participante é orientado a escolher a cidade de onde está respondendo a pesquisa dentre aquelas origens disponibilizadas na Tabela 3 e, caso não esteja em nenhuma das localidades, é direcionado para o encerramento da pesquisa. No caso de prosseguimento, o respondente é induzido a escolher uma cidade para onde já tenha realizado pelo menos uma viagem para um dos destinos e, novamente, caso nunca tenha viajado para nenhum dos destinos possíveis o questionário é encerrado (Figura 1).

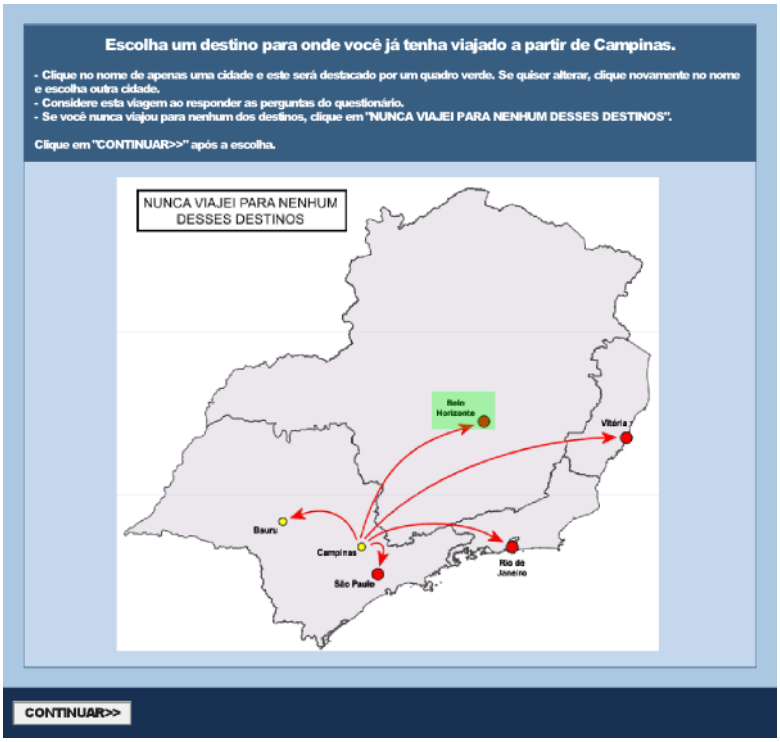

Figura 1: Exemplo de escolha da origem e destino de uma viagem realizada pelo participante

O participante que continua a pesquisa deve escolher o modo de transporte que utilizou na viagem realizada, sendo orientado que essa seja tomada como referência para as respostas às perguntas subsequentes. Em seguida é questionado sobre as características da viagem (motivo, tempo de acesso e egresso aos terminais no caso de uso do modo ônibus ou avião) e uma estimativa dos tempos de acesso e egresso aos terminais para os modos alternativos àquele escolhido (inclusive para as estações ferroviárias dado que todas as cidades de origem e destino contam com essa infraestrutura, mesmo que inoperantes).

Ao final dessa etapa, os modos de transporte são caracterizados com base na escolha aleatória do cenário a que o respondente é submetido e nove situações hipotéticas são estabelecidas para a escolha modal de fato, conforme exemplificado na Figura 2. Na etapa final são apresentadas questões de caráter socioeconômico como idade, vínculo empregatício, renda média familiar, número de pessoas na residência e posse de Carteira Nacional de Habilitação.

O tamanho da amostra (número de participantes) é um item importante para a modelagem das respostas de uma pesquisa de preferência declarada. No caso deste artigo, o fator determinante na definição desse parâmetro foi o orçamento disponível para identificação dos participantes e coleta das respostas. A execução da versão final do questionário entre os dias 24 de Outubro e 5 de Novembro de 2013 contou com a participação de 580 respondentes, dos quais 279 (48%) mostraram-se aptos a responder o questionário por terem viajado pelo menos uma vez entre um par de cidades por motivo de trabalho e 301 (52%) por motivo de lazer ou outros.

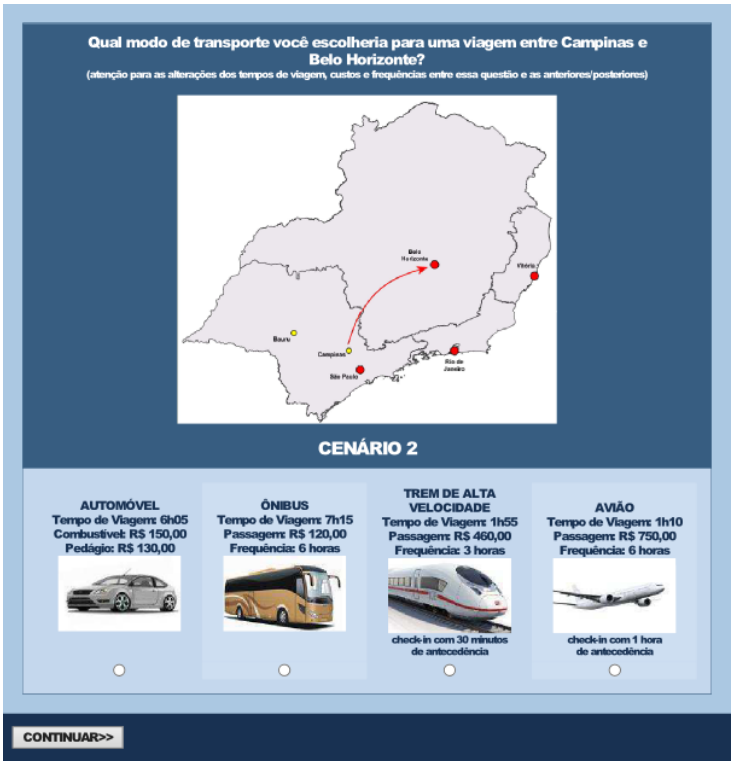

Figura 2: Situação hipotética para escolha modal em um cenário específico

A Tabela 5 apresenta o número de respostas obtidas em cada cenário por motivo de viagem, sendo que cada bloco de nove respostas corresponde a um participante no cenário que lhe foi apresentado.

Tabela 5: Número de respostas obtidas em cada cenário da versão final da pesquisa e porcentagem de escolhas de cada modo.

CENÁRIO		***HPR without AIR***		***HPR with AIR***		***HSR without AIR***		***HSR with AIR***	
MOTIVO		***W***	***L***	***W***	***L***	***W***	***L***	***W***	***L***
RESPOSTAS		729	549	603	639	648	612	729	711
	CAR (%)	31,3	23,1	23,5	17,4	18,7	19,9	28,4	17,4
PROPORÇÃO	**BUS (%)**	39,1	36,1	32,8	23,6	33,0	25,5	21,9	19,1
DAS	**HPR (%)**	29,6	40,8	23,7	27,1	0,0	0,0	0,0	0,0
ESCOLHAS	**HSR (%)**	0,0	0,0	0,0	0,0	48,3	54,6	29,4	40,7
	AIR (%)	0,0	0,0	19,9	31,9	0,0	0,0	20,3	22,8

Cabe ressaltar a proporcionalidade das escolhas dos modos para os cenários de *HPR* conforme apresentado na tabela anterior, indicando a eficácia na definição dos níveis de seus atributos ao evitar a dominância de um modo de transporte entre as alternativas disponíveis. Entretanto, para os cenários de *HSR* observa-se a tendência de maior escolha desse modo em detrimento dos demais - eventualmente devido ao maior apelo sobre as possibilidades de construção de uma infraestrutura dessa natureza no país - o que não necessariamente indica a dominância em relação aos demais modos nos respectivos cenários quanto aos níveis dos seus atributos.

4. ÁRVORES DE DECISÃO APLICADAS AOS RESULTADOS DA PESQUISA

Os algoritmos dos métodos *CHAID*, *CART* e *QUEST* foram executados no *software* SPSS® a partir dos dados obtidos pela aplicação do questionário considerando 70% das respostas da pesquisa para treinamento e 30% para testes (validação) dos modelos.

As variáveis admitidas para geração das árvores de decisão sem as características socioeconômicas dos respondentes foram: distância entre as cidades de origem e destino (*DI*), tempo de viagem para todos os modos (*TT*), custos de combustível (*PE*) e pedágio (*TO*) para o modo automóvel, e custo da passagem (*FA*) e frequência (*FR*) para os modos ônibus, trem e avião. No caso da geração das árvores contemplando as variáveis socioeconômicas, além das definidas anteriormente, foram consideradas a idade (*AGE*), vínculo empregatício (*EMP*), renda média mensal (*INC*) e posse de carteira de habilitação (*LIC*). A Tabela 6 apresenta a porcentagem de acertos para a amostra de teste dos algoritmos de árvore de decisão (comparação entre observados e estimados), para cada cenário estabelecido previamente com e sem a inserção das variáveis socioeconômicas.

Tabela 6: Porcentagem de acertos da amostra de validação (30% das respostas).

CENÁRIO/MÉTODO	CHAID[1]	CART[1]	QUEST[1]
HPR with AIR (L)	44.4 / 39.0 [-12.2%]	37.7 / 37.7 [0%]	41.6 / 29.4 [-29.3%]
HPR with AIR (W)	35.2 / 35.2 [0%]	43.5 / 41.1 [-5.5%]	38.4 / 38.4 [0%]
HPR without AIR (L)	47.6 / 43.8 [-8%]	56.8 / 47.5 [-16.4%]	44.3 / 44.3 [0%]
HPR without AIR (W)	43.7 / 40.8 [-6.6%]	42.5 / 41.6 [-2.1%]	37.1 / 37.1 [0%]
HSR with AIR (L)	35.3 / 45.1 [27.8%]	43.5 / 43.5 [0%]	44.7 / 44.7 [0%]
HSR with AIR (W)	33.8 / 36.7 [8.6%]	42.0 / 38.1 [-9.3%]	36.3 / 32.6 [-10.2%]
HSR without AIR (L)	53.2 / 55.5 [4.3%]	59.4 / 53.6 [-9.8%]	51.1 / 51.1 [0%]
HSR without AIR (W)	42.6 / 43.6 [2.3%]	42.6 / 42.6 [0%]	47.1 / 47.1 [0%]
MÉDIA	42.0 / 42.5 [1.2%]	46.0 / 43.2 [-6.1%]	42.6 / 40.6 [-4.7%]

[1]Com variáveis socioeconômicas (%) / Sem variáveis socioeconômicas (%) [Diferença percentual = (% acertos sem variáveis socioeconômicas – % acertos com variáveis socioeconômicas)/(% acertos com variáveis socioeconômicas)].

Na comparação entre os valores verifica-se que o método *CART* apresenta o maior índice médio de acertos, com 46,0% para o modelo com variáveis socioeconômicas e 43,2 % sem essas variáveis. Ainda, pode-se verificar que a inclusão de variáveis socioeconômicas nos modelos de AD não influencia fortemente a quantidade de acertos, pois o percentual médio de acertos aumenta em apenas 1,2% para o caso do algoritmo *CHAID* e diminui nos demais algoritmos. Isso demostra que a importância maior para análise da escolha modal para novos serviços ferroviários na Região Sudeste está nas características de viagens. As figuras seguintes representam as classes das partições dos dados resultantes do algoritmo *CART* para os diferentes cenários (com ou sem *AIR* e pelos motivos *W* e *L*) sem as variáveis socioeconômicas.

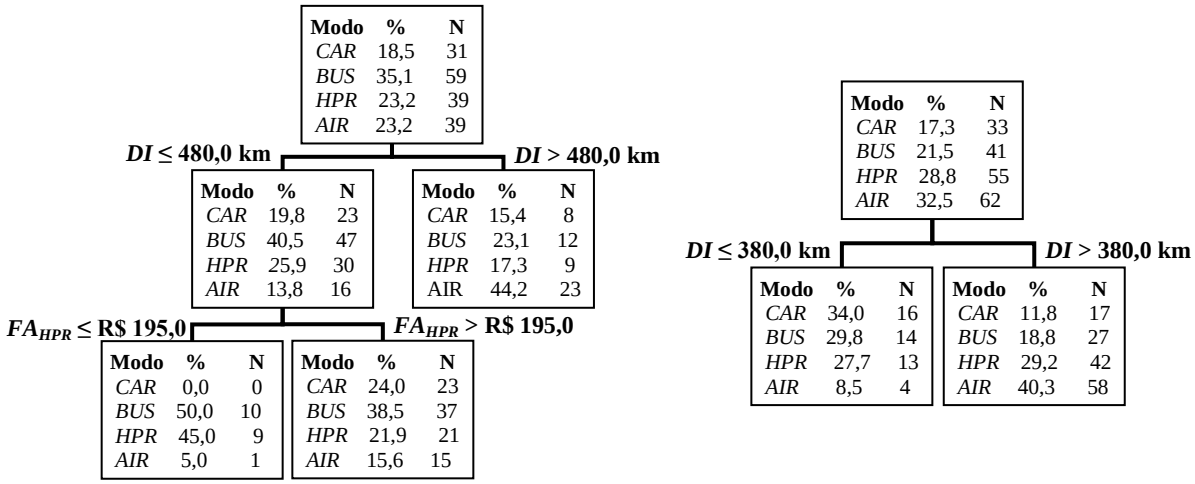

Figura 3: Árvores de Decisão do cenário *HPR with AIR* por motivo *W* (esq.) e *L* (dir.) do método *CART*

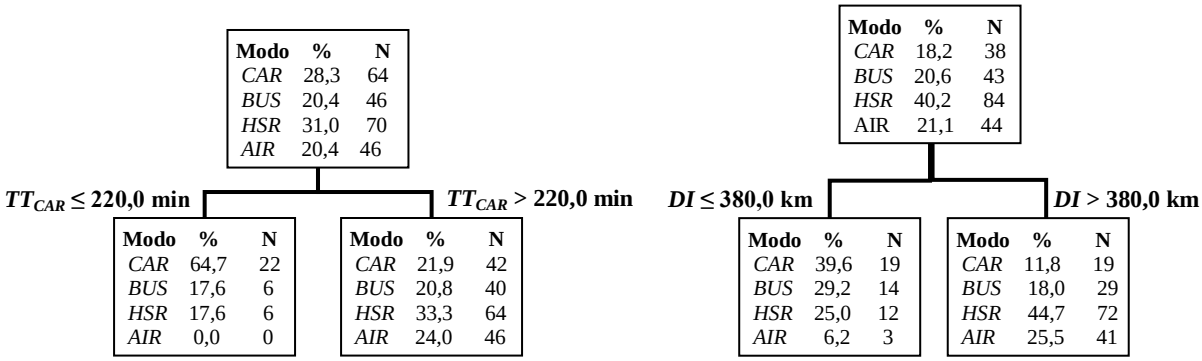

Figura 4: Árvores de Decisão do cenário *HSR with AIR* por motivo *W* (esq.) e *L* (dir.) do método *CART*

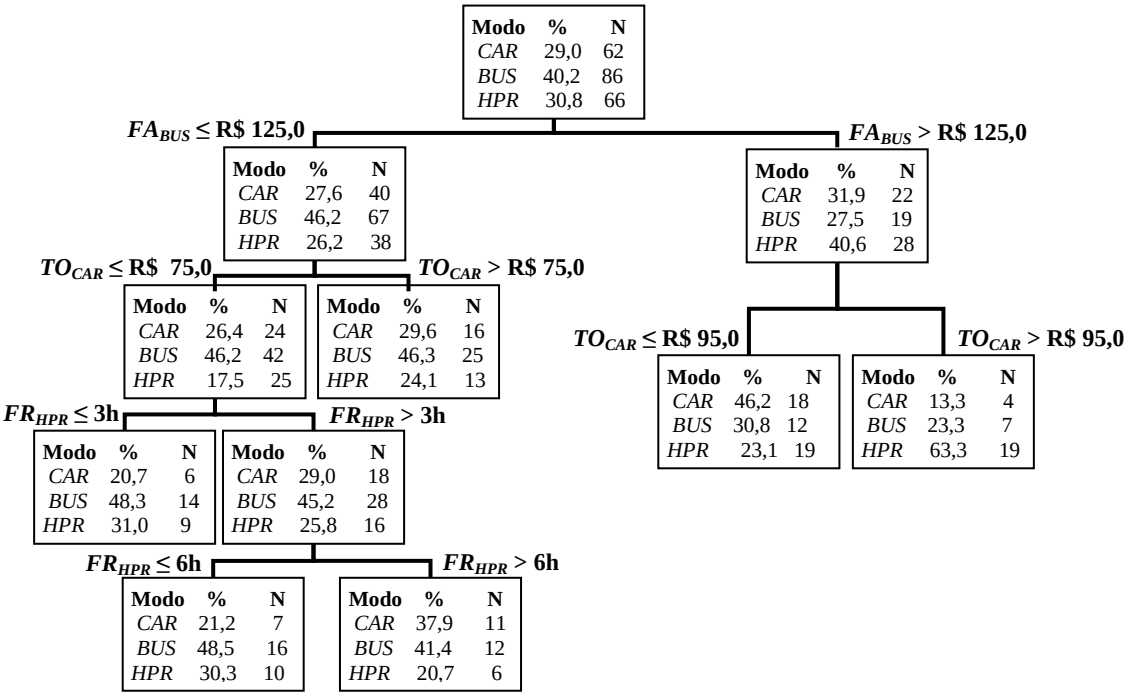

Figura 5: Árvores de Decisão do cenário *HPR without AIR* por motivo *W* do método *CART*

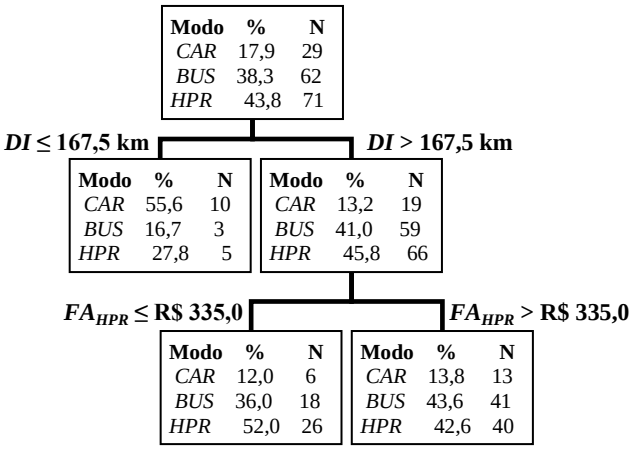

Figura 6: Árvores de Decisão do cenário *HPR without AIR* por motivo *L* do método *CART*

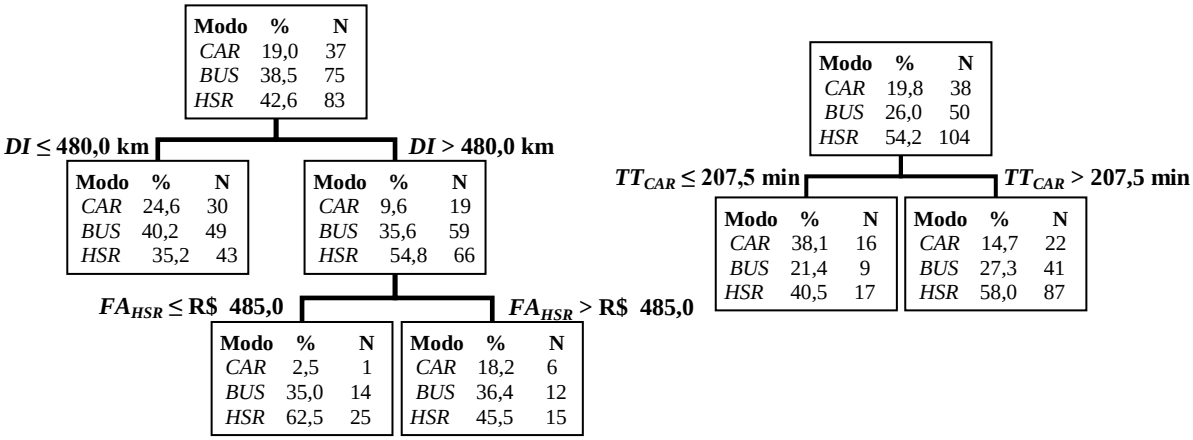

Figura 7: Árvores de Decisão do cenário *HSR without AIR* por motivo *W* (esq.) e *L* (dir.) do método *CART*

Como proposto, foram obtidos os nós terminais, ou folhas, que caracterizam as classes de indivíduos segundo características de viagens relativas às alternativas e suas respectivas probabilidades de escolha. Entre as variáveis consideradas, as seguintes foram selecionadas nos respectivos cenários: FA_{BUS}, TO_{CAR} e FR_{HPR} (*HPR without AIR* por motivo *W*); *DI* e FA_{HPR} (*HPR without AIR* por motivo *L* e *HPR with AIR* por motivo *W*); *DI* e FA_{HSR} (*HSR without AIR* por motivo *W*); TT_{CAR} (*HSR without AIR* por motivo *L* e *HSR with AIR* por motivo *W*); e *DI* (*HPR with AIR* por motivo *L* e *HSR with AIR* por motivo *L*). Assim, verifica-se que as variáveis *DI* (distância em km), FA_{BUS} (tarifa do ônibus em reais) e TT_{CAR} (tempo de viagem por automóvel em minutos) são as mais importantes para segmentação dos dados.

Além do percentual de acertos observados para as amostras de validação (Tabela 6), foi realizado o teste Qui-quadrado de avaliação do grau de associação entre as categorias dos valores observados e estimados pelo algoritmo *CART*. A Tabela 7 apresenta os resultados desses testes, demonstrando a acurácia e precisão dos modelos de árvore. Os valores da estatística Qui-quadrado e das probabilidades associadas a esse valores (α bicaudal) demonstram que, a um nível de significância de 5%, não há indícios de que os valores resultantes das árvores de decisão pela aplicação do algoritmo *CART* não estão associados às respostas dadas pelos participantes.

Tabela 7: Resultados do teste Qui-quadrado para o *CRT* sem variáveis socioeconômicas.

CENÁRIO	Estatística Qui-quadrado	Graus de Liberdade	*α* (bicaudal)
***HPR without AIR* (*L*)**	70,9	4	0,000
***HPR without AIR* (*W*)**	51,1	4	0,000
***HPR with AIR* (*L*)**	59,8	3	0,000
***HPR with AIR* (*W*)**	101,4	6	0,000
***HSR without AIR* (*L*)**	54,4	2	0,000
***HSR without AIR* (*W*)[1]**	-	-	-
***HSR with AIR* (*L*)**	76,4	3	0,000
***HSR with AIR* (*W*)**	60,6	3	0,000

[1]Nenhuma estatística foi calculada pois as respostas do modelo são constantes.

5. CONCLUSÕES E PESQUISAS FUTURAS

Esse artigo descreve uma pesquisa de preferência declarada para avaliação da propensão à escolha modal de novos serviços de trens de alto desempenho e de alta velocidade na Região Sudeste do Brasil utilizando técnicas de árvores de decisão. Apesar de a aplicação de técnicas de árvore de decisão não permitir a obtenção da significância das variáveis, tal como nas técnicas confirmatórias usuais, estas são de fácil aplicação e interpretação dos resultados.

Além disso, as técnicas admitem o uso de qualquer tipo de variável (contínua, discreta e categórica) e avaliação dos padrões das respostas da pesquisa de preferência declarada em consonância com os objetivos desse artigo. O poder preditivo dos modelos de árvore pode ser, de certa forma, mensurado pela quantidade de acertos e estatística qui-quadrado, tal como realizado e apresentado neste artigo.

A partir dos dados coletados verificou-se que, pela aplicação de diferentes algoritmos de AD, as variáveis socioeconômicas podem não afetar significativamente a escolha modal sob as condições estabelecidas na pesquisa. Essa conclusão é importante como referência na aplicação de outras abordagens para caracterização da escolha dos modos de transporte pelos usuários, como a calibração de modelos *Logit Multinomial* ou *Nested Logit Multinomial.*

Como perspectivas de pesquisas futuras, os dados obtidos a respeito das viagens já realizadas pelos participantes (modo de transporte utilizado, tempo estimado de acesso e egresso à infraestrutura de transporte público existente) permitem - além da estimativa de parâmetros de modelos *Logit* - a aplicação de modelagem proeminente de associação de dados de pesquisa de preferência declarada e de preferência revelada (Ortúzar e Willumsen, 2011).

Ainda do ponto de vista metodológico, pode ser realizada uma avaliação da abordagem de planejamento eficiente de experimentos (Rose *et al*., 2008) em detrimento do planejamento ortogonal, uma vez que os coeficientes dos parâmetros das funções utilidade que caracterizam os diferentes modos de transporte podem ser determinados pelo método *Logit Multinomial*, os quais são requeridos para aplicação daquele método de delineamento de experimentos.

AGRADECIMENTOS

A pesquisa de preferência declarada foi executada com financiamento concedido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq).

REFERÊNCIAS BIBLIOGRÁFICAS

ABCR (2013) Tarifas de Pedágio. Disponível em:< http://www.abcr.org.br/TarifasPedagio/TarifaPedagio.aspx>. Acesso em: 25/08/2013.

Ahern, A. e N. Tapley (2008) The use of stated preference techniques to model modal choices on interurban trips in Ireland. *Transportation Research Part A: Policy and Practice*, v. 42, n. 1, p. 15–27.

ANAC (2012). Anuário Estatístico do Transporte Aéreo 2012. Disponível em: <http://www2.anac.gov.br/arquivos/zip/Anuario2012.zip>. Acesso em: 12/02/2014.

ANAC (2013) Tarifas Aéreas Domésticas. (26ª ed.). Disponível em:<http://www2.anac.gov.br/estatistica/tarifasaereas/>. Acesso em: 20/08/2013.

ANP (2013) Boletim Anual de Preços de Combustível. Disponível em:<http://www.anp.gov.br/?pg=65870&m=&t1=&t2=&t3=&t4=&ar=&ps=&cachebust=1369405304193>. Acesso em:17/08/2013.

ANTT (2013) Coeficientes Tarifários de Serviços de Ônibus Intermunicipais. Disponível em: <https://www.antt.gov.br/sgp/src.br.gov.antt/apresentacao/consultas>. Acesso em: 15/08/2013.

Breiman, L.; J. Friedman; R. A. Olshen e J. Stone (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, USA.

Carson, R.T.; J. J. Louviere; D. A. Anderson; P. Arabie; D. S. Bunch; D. A. Hensher; R. M. Johnson; W. F. Kuhfeld; D. Steinberg; J. Swait; H. Timmermans e J. B.Wiley (1994) Experimental analysis of choice. *Marketing Letters*, v. 5, p. 351–368.

CPTM (2014) Companhia Paulista de Trens Metropolitanos – Programa Trens Regionais. Disponível em:<http://www.antt.gov.br/html/objects/_downloadblob.php?cod_blob=12456>.Acesso em: 15/03/2014.

Grange, L.; F. González; I. Vargas e J. C. Muñoz (2013) A polarized logit model. *Transportation Research Part A: Policy and Practice*, v. 53, p. 1–9.

Hosmer, D. W. e S. Lemeshow (2000) *Applied Logistic Regression*. Wiley, New York, USA.

Hsieh, F. Y. (1989) Sample size tables for logistic regression. *Statistics in Medicine*, v. 8, p.795–802.

Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* , v. 29, p. 119–127.

Loh, W. Y. e Y. S. Shih (1997) Split selection methods for classification trees. Statistic Sinic, v. 7, p. 815–840.

Mannila, H. (1997) *Methods and Problems in Data Mining*. Database Theory. p. 41-55. Springer, Berlin.

Mitchell, R.C. e Carson, R.T. (1989) *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources for the Future, Washington, DC.

Orme, B. (1998) *Sample Size Issues for Conjoint Analysis Studies*. Technical Paper. Disponível em: <http://www.sawtoothsoftware.com/download/techpap/samplesz.pdf>. Acesso em 20/11/2013.

Ortúzar, J. D. e L. G. Willumsen (2011) *Modelling Transport* (4th ed.). Wiley, New York, USA.

Peduzzi P; J. Concato; E. Kemper; T. R. Holford e A. R. Feinstein (1996) A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, v. 49, p. 1373–1379.

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

Sælensminde, K. (1999) *Valuation of Non-Market Goods for Use in Cost-Benefit Analyses: Methodological Issues*. PhD Thesis, Department of Economics and Social Sciences, Agricultural University of Norway.

TAV Brasil (2014) Trem de Alta Velocidade no Brasil – Agência Nacional de Transportes Terrestres. Disponível em: < http://www.antt.gov.br/index.php/content/view/5448.html>. Acesso em: 22/04/2014.

Whittemore, A. (1981) Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, v. 76, p. 27–32.

Xie, C.; L. Jinyang e E. Parkany (2007) Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks. *Transportation Research Record*, v. 1854, p. 50–61.

Rose, J. M.; M. C. J. Bliemer; D. A. Hensher e A. T. Collins (2008) Designing efficient stated choice experiments in the presence of reference alternatives. *Transportation Research Part B:Methodological*, v. 42, n. 4, p. 396–406.

MVAConsultancy (2013) *Comparison of International Rail fares and ticketing*. Disponível em:< http://www.passengerfocus.org.uk/research/publications>. Acesso em: 12/09/2013.

Prodan A. (2011) *Infrastructure Pricing Models for New High-Speed Railway Corridors in Europe*.. PhD Thesis, Complex Transport Infrastrucutre Systems, MIT Portugal Programm.

Cassiano Augusto Isler (cassiano.isler@usp.br)
Cira Souza Pitombo (cirapitombo@gmail.com)
Departamento de Transportes, Escola de Engenharia de São Carlos, Universidade de São Paulo
Av. Trabalhador São-Carlense, 400 – São Carlos, SP, Brasil