PROPOSTA DE CRITÉRIO PARA CARACTERIZAÇÃO AGREGADA DAS ALTERNATIVAS MODAIS A PARTIR DE DADOS DE PREFERÊNCIA REVELADA

Carolina Fernanda Cerveira
Viviani Antunes Gomes
Cira Souza Pitombo
Universidade de São Paulo
Escola de Engenharia de São Carlos/Departamento de Engenharia de Transportes

RESUMO
O comportamento de escolha individual está associado às características do indivíduo e do conjunto de alternativas. A Pesquisa O/D é uma importante fonte de dados de Preferência Revelada (PR) e descreve as escolhas e comportamentos reais dos indivíduos. Desta forma, através de tal fonte de informação, não é possível caracterizar as alternativas não escolhidas. Este trabalho propõe um critério para caracterizar, de forma agregada, as alternativas modais, utilizando dados de PR. Foram utilizados os algoritmos CHAID (*Chi-squared Automatic Interaction Detection*) e CART (*Classification And Regression Tree)* para estimar o tempo de viagem dos modos de transporte disponíveis na área de estudo (Cidade de São Paulo – Pesquisa OD de 2007). As viagens foram agrupadas, segundo variáveis independentes selecionadas pelos algoritmos, e foram obtidos valores médios de tempos de viagens para as cinco alternativas modais. O critério foi considerado adequado, segundo validações metodológicas propostas, bem como sua facilidade de uso e ausência de restrições matemáticas, inerentes às ferramentas paramétricas.

ABSTRACT
Travel behavior is associated with the characteristics of individuals and a set of alternatives. The O/D Survey is an important source of Revealed Preference (RP) data and describes individuals' real choices and behaviors. Thus, through this source of information, it is not possible to characterize the alternatives not chosen. This paper proposes a criterion to characterize, in an aggregated way, the travel mode alternatives, using RP data. The Chi-squared Automatic Interaction Detection and Classification and Regression Tree algorithms were used to estimate the travel time of the available travel modes in the study area (City of São Paulo, Brazil – OD Survey of 2007). The trips were grouped, according to independent variables selected by algorithms, and the average values of travel time were obtained for the five travel mode alternatives. The criterion was suitable, taking into account the methodological validations, as well as the absence of mathematical constrain, inherent to the parametric tools.

1. INTRODUÇÃO E *BACKGROUND*

Entender os fatores que determinam a tomada de decisão individual relacionada a viagens, como por exemplo, a escolha do modo de transporte, é um aspecto que permite entender o comportamento da população e, consequentemente, a elaborar políticas públicas. As razões de escolha de um modo de transporte em detrimento dos outros são complexas, Ortúzar e Willumsen (2011) apontam alguns fatores que podem interferir nessas escolhas, como: (1) as características do indivíduo (posse do carro, renda, densidade residencial, etc.), (2) as características da viagem (motivo de viagem, período do dia, etc.), e (3) as características das alternativas ou dos modos de viagem (tempo de viagem, custos monetários, conforto, conveniência, etc.).

A aplicação de modelos de escolha discreta tem sido tema de inúmeras pesquisas relativas à modelagem de demanda por transportes (Morikawa, 1989; Ben-Akiva e Morikawa, 1990; Antonini *et al.,* 2006; Frejinger, 2008; Silva, 2015). Sendo assim, as informações referentes às alternativas são essenciais para a previsão de escolhas individuais.

As pesquisas domiciliares, que compõem a Pesquisa Origem-Destino (OD), são importante instrumento para os estudos na área de planejamento de transportes. Possibilitam estabelecer projeções futuras das necessidades de viagens das pessoas. Constituem importante fonte de

dados de Preferência Revelada (PR) e descrevem as escolhas e comportamentos reais dos indivíduos. Têm ampla abrangência espacial e grande número de entrevistas, sendo uma fonte de dados que tem subsidiado o planejamento de transportes em muitas cidades.

No entanto, os dados obtidos através da Pesquisa OD só apresentam as informações das viagens efetivamente realizadas pelo entrevistado, não apresentando, entretanto, informações sobre as outras alternativas possíveis. Para que se consiga aprimorar uma modelagem de escolha discreta é importante ter variáveis que caracterizem as alternativas disponíveis no processo de escolha. Diversos autores (Souza *et al.*, 2017; Fezzi *et al*., 2014; Kato *et al*., 2013) com base em dados de Pesquisa Revelada, propuseram diferentes métodos para estimar características agregadas das demais alternativas (como tempo e custo de viagem, por exemplo).

Souza *et al.* (2017) utilizaram dados de PR para estimar o valor do tempo para diferentes grupos de pessoas. Os autores caracterizaram, de forma agregada, as variáveis relativas a custo e tempo de viagem. No cálculo dos custos médios de viagem, para o modo individual, considerou-se o custo do combustível e estacionamento e para o modo motorizado coletivo o custo foi estimado a partir do valor médio da tarifa. Para a caracterização dos tempos de viagens agregados, os mesmos autores propuseram agrupamentos a partir das distâncias euclidianas das viagens (grupos com amplitude de 500 metros) afim de determinar as médias de tempo de viagens por modo coletivo e individual motorizado. Observa-se que os autores, determinaram, empiricamente, faixas de distâncias para agregação das observações. Não consideram as demais variáveis que compõem a Pesquisa OD e que influenciam na duração das viagens como: horário de saída, tempo de caminhada até o modo principal, zonas de trafego de origem e destino, etc.

Fezzi *et al.* (2014) estimaram o valor do tempo em viagens utilizando rodovias com e sem pedágio através de dados de PR, nas estradas que dão acesso às praias na Itália. As alternativas de rotas foram caracterizadas por diferentes tempos de viagens e custos monetários. Os tempos de viagem foram obtidos utilizando o *website* https://maps.google.com e os custos de viagem foram calculados considerando-se o consumo médio de combustível e o preço do litro. Os autores propuseram agrupamento dos indivíduos, adotando faixas de idade e valor salarial. Além dos dados da rota, foram coletados também dados referentes às viagens, e às características socioeconômicas dos entrevistados.

Kato *et al*. (2013), utilizando os dados do Censo de Tráfego Rodoviário do Japão de 2005 (*Road Traffic Census Survey of Japan*), estimaram o valor da redução do tempo de viagem nas rodovias japonesas. O tempo de viagem foi estimado usando o *website* da pesquisa e o custo da viagem foi estimado com os dados do pedágio, desconsiderando-se o custo de combustível e de manutenção. A agregação, para caracterização das alternativas de rotas, deu-se a partir da escolha, por parte dos autores, de faixas de idade, gênero, tipo de trabalho, intervalos de hora de saída, intervalos de distância de viagem e por motivo.

Nos trabalhos citados utilizou-se apenas o tempo e o custo da viagem, e os critérios de agrupamento foram definidos de forma subjetiva. No presente trabalho, utiliza-se um conjunto de variáveis associadas às viagens e aplica-se uma ferramenta de agrupamento com relações de dependência e critérios baseados na homogeneidade dos grupos.

Neste contexto, este trabalho tem como objetivo propor um critério para caracterização agregada de alternativas modais a partir de dados PR. Os autores utilizaram algoritmos de Árvore de Decisão (CHAID – *Chi-squared Automatic Interaction Detection* e CART - *Classification And Regression Tree*) para estimar o tempo de viagem dos modos de transporte disponíveis na área de estudo (Cidade de São Paulo) através da utilização de dados de PR, obtidos através de pesquisa domiciliar OD de 2007.

2. ALGORITMOS DE ÁRVORE DE DECISAO

Algoritmos de Árvore de Decisão (AD) podem classificar ou prever, de acordo com o tipo de variável dependente. Os modelos são ajustados por sucessivas divisões, a partir de declarações do tipo "Se... então..." com o intuito de se obter subconjuntos cada vez mais homogêneos segundo a variável dependente. São algoritmos não-paramétricos e têm uma estrutura que se assemelha a uma árvore. O conjunto total de dados (nó raiz) é separado por sequenciais divisões (nós filhos) e essas divisões se dão de forma sequencial até os nós terminais (ou folhas), quando não é possível mais nenhum subgrupo. Para a construção da árvore, devem-se definir três parâmetros: um conjunto de regras que demarca a divisão dos dados; um critério que possibilite a melhor divisão para produção dos nós filhos; e uma regra que determine o limite das subdivisões (regra *stop-splitting*) (Breiman *et al.,* 1984).

Quando a variável dependente é categórica, as árvores de decisão são usadas para problemas de classificação e são denominadas Árvores de Classificação. Para o caso de variável dependente quantitativa, que é o caso deste trabalho, as árvores de decisão são denominadas Árvores de Regressão. São muitos os algoritmos de AD, sendo que os principais são: C4.5 (Quinlan, 1983), CHAID (Kass, 1980) e CART (Breiman *et al.,* 1984). Neste trabalho optou-se pelos algoritmos CHAID e CART.

O algoritmo CHAID (*Chi-squared Automatic Interaction Detector)* constrói árvores onde é possível criar mais de duas ramificações ligadas a um único nó e compreende três etapas: fusão, divisão e parada (Magidson, 1994). A variável dependente pode ser categórica nominal, categórica ordinal (árvore baseada no teste quiquadrado) e contínua (árvore de regressão baseada no teste F). Para o caso de variável independente contínua, esta é transformada em uma variável categórica ordinal antes da aplicação do algoritmo.

O procedimento inicial do algoritmo consiste na preparação das variáveis independentes. As variáveis categóricas já possuem as suas classes "naturais". As variáveis independentes contínuas são transformadas em variáveis categóricas dividindo a sua distribuição contínua em categorias com números similares de observações. Após a preparação de variáveis independentes, o algoritmo realiza a etapa de fusão.

A primeira etapa (fusão) corresponde à união das categorias não significativamente diferentes das variáveis independentes, segundo a variável dependente. Para o caso de variável dependente contínua (como neste trabalho que se utilizou duração de viagem), são realizados testes F. Para pares de categorias não significativamente diferentes, os mesmos são unidos. Para os pares significativamente diferentes, então são realizados testes de *Bonferroni* para ajuste de *p-value*. O *p-value* ajustado é calculado para as categorias fundidas, aplicando os ajustes de *Bonferroni*, que é o número de maneiras possíveis pelas quais as categorias podem ser fundidas. Os resultados do teste de *Bonferroni* são utilizados na etapa de divisão.

Na segunda etapa (divisão) é selecionada a variável independente que permite a melhor divisão do nó. A seleção se dá comparando os valores de *p-value* associado a cada variável, obtido na etapa de fusão. A variável independente com o menor *p-value* ajustado (divisão mais significaiva) é selecionada. Caso esse *p-value* ajustado for menor ou igual ao especificado pelo usuário, o nó é dividido usando esta variável. Caso contrário, o nó não é dividido e é considerado como um nó terminal.

A terceira etapa (parada) corresponde às seguintes regras de interrupção: (1) caso todas as observações em um nó tenham os mesmos valores da variável dependente; (2) caso todas as observações, em um nó, tenham os mesmos valores para cada covariável; (3) caso a árvore atinja o limite de profundidade estabelecido; (4) caso o tamanho de um nó seja menor que o valor mínimo estabelecido; (5) caso divisão de um nó resultar em um nó filho de tamanho menor do que o estabelecido pelo usuário (Goodman, 1979; Kass, 1980).

O algoritmo CART (Classification And Regression Tree), desenvolvido por Breiman *et al*. (1984), se dá basicamente, através de divisões binárias dos dados buscando reduzir as impurezas dos nós filhos e maximizar a homogeneidade nos nós terminais.

Para as Árvores de Classificação do algoritmo CART (variável dependente categórica), o critério de partição ou medidas de impureza mais conhecido é o índice Gini. Já para as Árvores de Regressão do CART (variável dependente quantitativa), 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.

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

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.

3. MATERIAIS E MÉTODO

3.1 Materiais

Os dados utilizados neste estudo são referentes à Pesquisa Origem-Destino (OD), realizada na Região Metropolitana de São Paulo (RMSP) em 2007, na qual foram levantadas informações de 30 mil domicílios, escolhidos aleatoriamente. Nestes domicílios, distribuídos nas 460 zonas de tráfego da pesquisa em que foi subdividida a RMSP, foram entrevistadas aproximadamente 120 mil pessoas. Nesse trabalho foram usadas somente as entrevistas realizadas na cidade de São Paulo. A Figura1 ilustra a área onde a pesquisa foi realizada.

A pesquisa é composta por quatro bancos de dados: agregados por zonas de trafego, desagregados por viagens, desagregados por domicílios e desagregados por indivíduos. Nesse trabalho, estão sendo utilizados dados desagregados por viagens. Cada viagem está associada ao identificador do indivíduo, identificador do domicílio de residência, características individuais, domiciliares e de viagens. A Tabela 1 descreve as variáveis utilizadas neste trabalho. A variável "hora de saída" foi agrupada em seis categorias, em função dos períodos pico e entre pico. A Tabela 2, em seguida, traz a caracterização das variáveis numéricas

através das medidas descritivas. Já a Tabela 3 apresenta as frequências das categorias das variáveis nominais. A amostra final foi composta de 54.895 viagens.

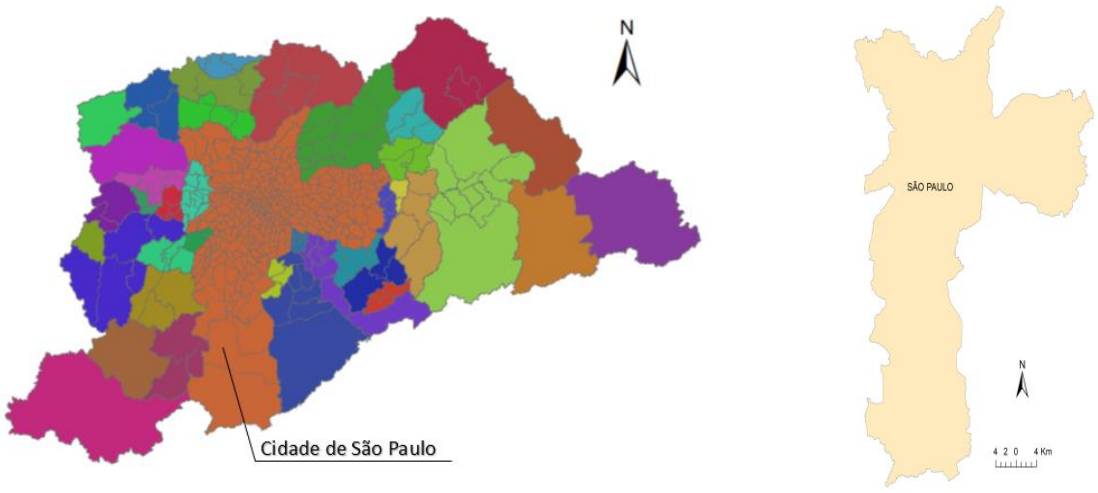

Figura 1 - Pesquisa Origem Destino – RMS 2007. Fonte: Metrô (2008)

Neste trabalho a variável dependente corresponde à duração das viagens. As variáveis independentes, nas duas análises realizadas e descritas na próxima subseção, estão caracterizadas nas tabelas seguintes. Para a estimação dos tempos de viagem utilizou-se os algoritmos CHAID e CART, disponível no pacote IBM SPSS 24.

Tabela 1 – Variáveis da Amostra. Fonte: Metrô (2008)

Variáveis	**Natureza**	
Zona de origem	Qualitativa	
Zona de destino	Qualitativa	
Motivo na origem	Qualitativa	
Motivo no destino	Qualitativa	
Hora de Saída	Quantitativa	Discreta
Duração da viagem (em minutos)	Quantitativa	Continua
Modo principal	Qualitativa	
Distância	Quantitativa	Continua

Tabela 2 - Medidas descritivas das variáveis numéricas.

Dados por domicílio	**Mínimo**	**Máximo**	**Média**	**Desvio Padrão**
Duração (min.)	1	240	29,70	25,27
Distância (km)	1,28	45.876,38	3.600,11	4.192,08

Tabela 3 - Frequência das categorias das variáveis categóricas

Variáveis	Variáveis categóricas	Quantidade na amostra	Percentual (%)
Motivo na origem	1 - Trabalho/indústria	761	1,4
	2 - Trabalho/comercio	2.306	4,2
	3 - Trabalho/serviço	9.645	17,5
	4 - Escola	7.669	13,9
	5 - Compras	2.015	3,6
	6 - Medico/dentista/saúde	1.459	2,6
	7 - Recreação/visitas/lazer	2.455	4,4
	8 - Residência	23.920	43,3
	9 - Procurar emprego	73	0,1
	10 - Assuntos pessoais	4.592	8,3
Motivo no destino	1 - Trabalho/indústria	746	1,4
	2 - Trabalho/comercio	2.393	4,3
	3 - Trabalho/serviço	9.656	17,5
	4 - Escola	7.702	13,9
	5 - Compras	2.021	3,7
	6 - Medico/dentista/saúde	1.475	2,7
	7 - Recreação/visitas/lazer	2.521	4,6
	8 - Residência	23.612	42,8
	9 - Procurar emprego	77	0,1
	10 - Assuntos pessoais	4.692	8,5
Modo principal (AD1)	Modo Coletivo	13.805	25,3
	Modo Individual	41.096	74,7
Modo principal (AD2)	1	25.497	46,4
	2	8.800	15,9
	3	5.005	9,4
	4	292	0,5
	5	15.307	27,8

3.2. Método

A Figura 2 apresenta o método utilizado neste trabalho. Após realizar o tratamento dos dados e gerar a amostra, foram aplicados os algoritmos de CHAID e CART, e estimados os tempos de viagem dos modos de transporte disponíveis através de duas análises diferentes (AD1 e AD2). Na primeira análise (AD1) é gerada a AD a partir de duas variáveis independentes: (1) Distância e Modo de viagem (Particular ou Coletivo). O objetivo desta análise foi fazer uma análise de validação a partir do método sugerido por Souza *et al.* (2017). Na segunda análise (AD2), é gerada a AD com as variáveis explicativas descritas na Tabela 1.

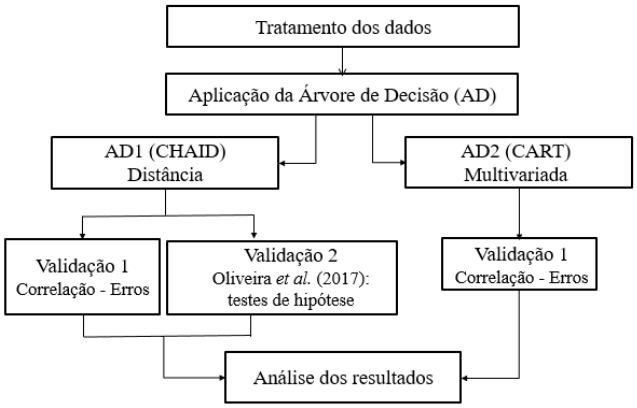

Figura 2 - Método proposto para estimar os tempos de viagem e caracterizar os tempos de viagens das alternativas não utilizadas

Tratamento dos dados: Inicialmente, foi feita a análise do banco de dados e obtenção da amostra final desagregada (por viagens) com dados socioeconômicas do entrevistado, do domicílio e dados das viagens. A amostra final é caracterizada por viagem realizada, associada ao identificador individual. Dados apresentados na Tabela 1.

Aplicação da AD: Nesta etapa, aplicou-se os algoritmos CHAID e CART para estimar os tempos de viagem dos modos de transporte utilizados, além da caracterização dos tempos de viagens dos modos não utilizados (principal objetivo deste trabalho). Foram treinadas duas ADs distintas:

- Primeira AD (AD1) – Nesta primeira árvore as variáveis independentes foram: modo principal e distância de viagem. Esta árvore é composta por dois níveis a partir do nó raiz: o primeiro correspondente aos subgrupos (nós filhos) divididos segundo intervalos de distâncias e o segundo e último nível corresponde aos nós terminais, divididos segundo os modos principais. Nesta árvore os modos principais foram agrupados em apenas duas categorias (Público ou Privado motorizado) com intuito de realizar validação metodológica comparativa ao trabalho de Souza *et al*. (2017). No trabalho de Souza *et al.* (2017), os autores agruparam as viagens segundo classes de distâncias euclidianas (intervalos de 500 metros). O tempo de viagem, por modo, foi equivalente às médias em cada classe. As classes obtidas, pelo trabalho mencionado, foram comparadas às classes obtidas através do CHAID a partir de testes de hipóteses para comparação de medianas (testes de medianas) e distribuições (*Kolmogorov-Sminorv* e *Mann Whitney*).
- Segunda AD (AD2) – Nesta árvore utilizou-se a algoritmo CART e as variáveis independentes foram: distância, clusters da hora de saída, motivo na origem, motivo no destino, zona de origem e zona de destino. Na AD2 a variável modo principal é composta por cinco categorias: modo motorizado privado (1), ônibus (2), metrô ou trem (3), bicicleta (4) e a pé (5).

Validação: Com o intuito de validar o método foram realizadas testes estatísticos, conforme descrito a seguir.

Validação 1: Foi realizada uma primeira validação do método comparando os tempos de viagem estimados (pela modo realmente utilizado) com os valores das durações de viagens efetivamente realizados pelo entrevistado. A amostra foi dividia aleatoriamente e foram obtidas as amostras de treinamento (70%) e teste (30%). Foram então calculadas as medidas de erros com a amostra de teste: erro médio quadrático (Equação 2), raiz do erro médio quadrático (Equação 3), erro médio absoluto (Equação 4) e correlação de *Pearson* (Equação 5).

$$\frac{1}{N}\sum(x_i - y_i)^2 \quad (2) \qquad \frac{1}{N-1}\cdot\sum\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} \quad (3)$$

$$\sqrt{\frac{1\sum(x_i-y_i)^2}{N}} \quad (4) \qquad \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left(\sum(x_i-\bar{x})^2\right)\left(\sum(y_i-\bar{y})^2\right)}} \quad (5)$$

Sendo que, x_i é a medida estimada; y_i é a medida observada; N é o número de medidas, $\bar{x}$ e $\bar{y}$ são as médias das amostras; σ_x e σ_y são os desvios-padrão da amostra.

<u>Validação 2:</u> Em uma segunda validação, os tempos de viagens para duas alternativas modais (Individual motorizado e coletivo), para cada viagem, foram comparados aos resultados obtidos previamente por Souza *et al*. (2017). Foram comparadas distribuições populacionais e medidas típicas (medianas) através de testes estatísticos: (1) teste de mediana (das duas amostras independentes: de treinamento e de teste); (2) *Kolmogorov Sminorv* e (3) *Mann-Whitney*. Os testes foram realizados por classes de distâncias, de forma acumulativa, até que fosse determinada a faixa de distância onde os grupos (Souza *et al*. 2017 e presente trabalho) não fossem considerados similares.

4. RESULTADOS E DISCUSSÕES

4.1. Primeira AD (AD1)

A primeira AD foi gerada através do algoritmo CHAID e os critérios adotados para essa árvore foram: mínimo de observações no nó filho = 30; nível de significância para dividir os nós = 0,05 e para categorias de fusão (teste quiquadrado ou teste F) igual a 0,05. A Figura 3 mostra, esquematicamente, o mapa da árvore obtida (amostra de treinamento). Foram obtidos 60 nós (no total), 40 nós terminais e profundidade igual a 2. A Tabela 4 relaciona os dados dos nós-terminais. Os tempos estimados para as duas alternativas modais (Coletivo ou Individual) são apresentados na Tabela 4. As 20 classes (nós terminais) obtidas pelo CHAID são comparados aos 45 grupos obtidos no trabalho de Souza *et al.* (2017). Observa-se que, para o algoritmo CHAID, categorias de distâncias acima de 12 mil metros (aproximadamente) não foram consideradas significativamente diferentes segundo a variável dependente "tempo de viagem", sendo unidas em uma única categoria na etapa de fusão (Nó 20).

Para caracterização agregada dos tempos de viagens para todas as alternativas modais, os grupos foram associados aos nós filhos obtidos no primeiro nível da AD1 (grupos esses segregados a partir de intervalos de distâncias). Tais grupos são representados na Figura 3 pelos quadrados alaranjados. Cada um desses nós é subdividido em nós terminais segundo alternativas modais. Assim, são obtidos os tempos de viagens para cada uma das duas alternativas. Desta forma, as viagens classificadas no nó 1, por exemplo, terão tempos de viagens agregados correspondentes a 25,97 min por transporte coletivo (nó 21) e 11,58 min por transporte individual (nó 22).

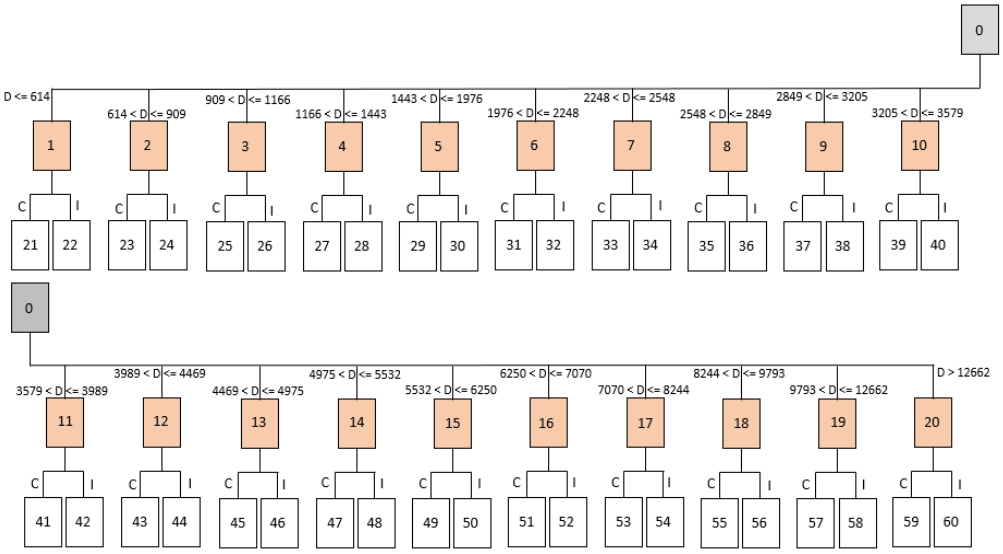

Figura 3 - Mapa da AD1 (Algoritmo CHAID)

Tabela 4: Condições de corte nos nós terminais (AD1: Algoritmo CHAID)

Nó	Condições de corte	TVM (min.)		Nó	Condições de corte	TVM (min.)	
		C (Nó)	I (Nó)			C (Nó)	I (Nó)
1	D <= 614	25,970 (21)	11,588 (22)	11	3579 < D <= 3989	43,229 (41)	29,403 (42)
2	614 < D <= 909	25,186 (23)	12,566 (24)	12	3989 < D <= 4469	45,868 (43)	31,178 (44)
3	909 < D <= 1166	26,733 (25)	14,480 (26)	13	4469 < D <= 4975	46,354 (45)	33,475 (46)
4	1166 < D <= 1443	28,269 (27)	15,697 (28)	14	4975 < D <= 5532	49,452 (47)	35,415 (48)
5	1443 < D <= 1976	33,160 (29)	18,700 (30)	15	5532 < D <= 6250	54,076 (49)	37,369 (50)
6	1976 < D <= 2248	34,714 (31)	20,832 (32)	16	6250 < D <= 7070	58,589 (51)	41,225 (52)
7	2248 < D <= 2548	35,766 (33)	21,970 (34)	17	7070 < D <= 8244	59,795 (53)	42,505 (54)
8	2548 < D <= 2849	36,119 (35)	23,737 (36)	18	8244 < D <= 9793	66,294 (55)	46,328 (56)
9	2849 < D <= 3205	39,023 (37)	25,068 (38)	19	9793< D <= 12662	74,867 (57)	49,235 (58)
10	3205 < D <= 3579	39,821 (39)	27,460 (40)	20	D > 12662	87,488 (59)	52,045 (60)

D: Distância (em metros); TVM: tempo de viagem médio (em min); C: modo coletivo; I: modo individual

A validação 1 foi realizada para essa primeira árvore e foram obtidas as medidas de erro a partir da amostra de teste: 344,86 para o erro médio quadrático, 18,57 para a raiz do erro médio quadrático, 0,046 de erro médio absoluto e a correlação de *Pearson* foi 0,676. O cálculo das medidas foi realizado a partir dos valores de tempos de viagens observados e tempos de viagens estimados pelo modo realmente utilizado.

A validação 2 consistia na realização de testes de hipótese para comparação entre os valores médios de tempos de viagens das classes obtidas no presente trabalho, através do algoritmo CHAID, e das classes propostas por Souza *et al.* (2017). Em tais testes, a hipótese nula afirmava similaridade entre os valores segundo medidas típicas (teste da mediana) e distribuições populacionais (*Mann Whitney* e *Kolmogorov Smirnov*). As comparações foram realizadas de forma acumulativa, segundo aumento de distâncias. A partir da classe 20 (número encontrado neste trabalho), foram acrescentadas classes de distâncias apenas dos resultados obtidos por Souza et al. (2017). A Tabela 5 apresenta os resultados, por grupos de classes, dos testes de hipóteses realizados.

Tabela 5: Resultados referentes à validação 2

	Classes	Classes 1 a 5	Classes 1 a 10	Classes 1 a 15	Classes 1 a 20	Classes 1 a 25	Classes 1 a 29	Classes 1 a 30	Classes 1 a 31
	Souza et al. (2017)	(0 a 2.500m)	(0 a 5.000m)	(0 a 7.500m)	(0 a 10.000m)	(0 a 12.500m)	(0 a 14.500m)	(0 a 15.000m)	(0 a 15.500m)
	CHAID	(0 a 1.976m)	(0 a 3.579m)	(0 a 6.250m)	(0 a 1.2662m) e D>12.662m				
Tempos	Teste da Mediana	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Rejeitar Ho	Rejeitar Ho
Modo	Teste Mann Whitney	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Rejeitar Ho	Rejeitar Ho	Rejeitar Ho	Rejeitar Ho
Coletivo	Teste Kolmogorov Smirnov	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Rejeitar Ho
Tempos	Teste da Mediana	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Rejeitar Ho	Rejeitar Ho
Modo	Teste Mann Whitney	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Rejeitar Ho	Rejeitar Ho	Rejeitar Ho
Privado	Teste Kolmogorov Smirnov	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Reter Ho	Rejeitar Ho

Ho = Há similaridades entre distribuições ou medianas entre tempos de viagens

Observou-se que, para a distância de até 12.000 metros (até classe 24 para Souza et al. (2017) – até classe 20 para CHAID) todas as hipóteses nulas são retidas, ou seja, há similaridades de valores típicos e distribuições populacionais entre os tempos de viagens agregados pelos dois procedimentos. Para 12.500 metros (Até classe 25 para Souza et al. (2017) e até classe 20 para CHAID), os tempos de viagens, para o modo coletivo, não possuem mesma distribuição populacional segundo o teste de *Mann-Whitney*. Até a distância de 14.500 metros o teste de *Mann-Whitney* rejeita similaridades de distribuições de tempos de viagens, tanto para transporte coletivo, quanto para modo de transporte individual. Para distâncias de até 15.000

metros, o teste *Kolmogorov sminorv* é o único que retém a hipótese nula de similaridade entre distribuições populacionais de tempos de viagens. A partir da distância de 15.500 metros, todas as hipóteses nulas são rejeitadas segundo todos os testes estatísticos, corroborando dissimilaridades entre tempos de viagens, levando-se em conta os dois procedimentos de agrupamento.

Para o algoritmo CHAID, não há diferenças significativas, segundo tempos de viagens, a partir da distância de 12.662m. Por esta razão, o algoritmo agrega essas distâncias em uma única categoria (etapa fusão - nó 20). Assim, observa-se que, para distâncias maiores, há dissimilaridades de tempos de viagens entre os procedimentos comparados. Desta forma, pode-se afirmar que o critério de agrupamento de tempos de viagens de transporte coletivo e individual, segundo amplitudes de 500 metros de distâncias euclidianas, proposto por Souza et al. (2017), funciona adequadamente até faixas de distâncias de 12.000 metros. A partir de determinados valores, não há diferenças estatísticas significativas segundo tempos de viagens entre classes de distâncias (comprovadas a partir de teste F do algoritmo CHAID).

4.2. Segunda AD (AD2)

A segunda AD foi gerada através do algoritmo CART e os critérios adotados para parada das divisões foram: mínimo de observações no nó terminal = 30 observações; e valor mínimo de aprimoramento = 0,00001. Como resultado, foram obtidos um total de 57 nós, dos quais 29 foram nós terminais e profundidade igual a 5. A Figura 4 mostra o mapa da árvore (etapa de treinamento), e a Tabela 6 relaciona os dados aos nós-terminais.

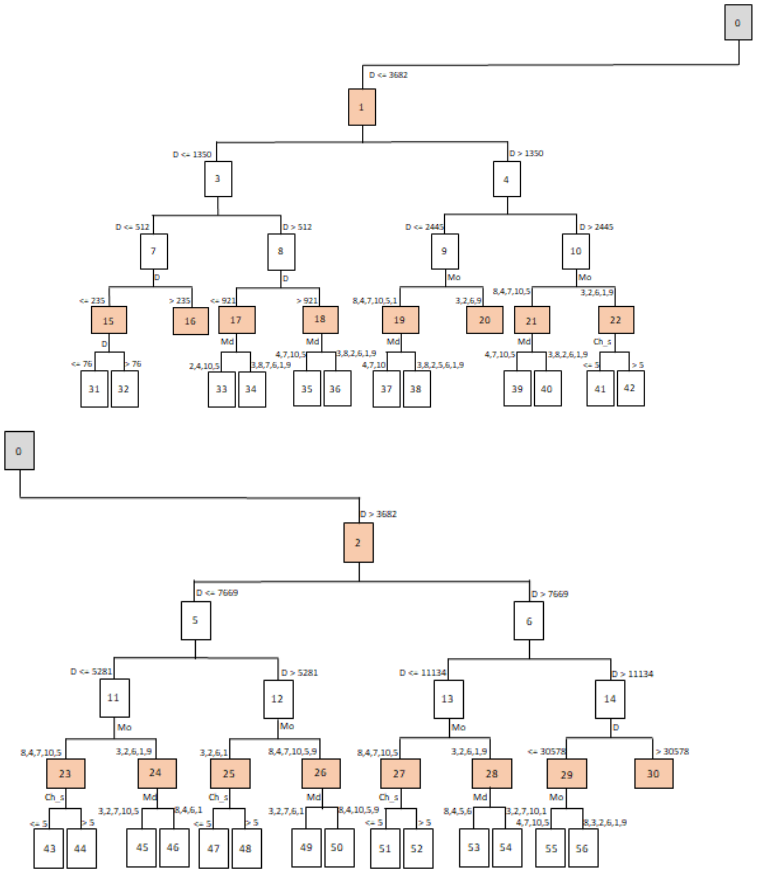

Figura 4 - Mapa da AD2 (Algoritmo CART)

Nesta árvore, as variáveis independentes significativas foram: "distância", "*clusters* da hora de saída", "motivo na origem" e "motivo no destino". Os tempos de viagens, para todas as alternativas modais, foram associados aos nós terminais obtidos no quinto nível da AD2. Gerada a árvore, foram feitos filtros e identificados, em cada nó terminal, os tempos de viagem para todos os cinco modos de viagem (1: Privado Motorizado; 2: Ônibus; 3: Metrô e Trem; 4: Bicicleta; 5: A pé).

Na validação obtiveram-se as seguintes medidas de erro: 378,677 para o erro médio quadrático, 19,46 para a raiz do erro médio quadrático, -0,065 de erro médio absoluto e a correlação de *Pearson* foi 0,638. O cálculo das medidas foi realizado considerando-se valores observados e estimados de tempos de viagens dos modos de transporte realmente utilizados a partir da amostra de teste.

Tabela 6: Condições de corte nos nós terminais (AD2: Algoritmo CHAID)

Nó	Condições de corte	TVM (min.)				
		Modo 1	Modo 2	Modo 3	Modo 4	Modo 5
16	235 < D <= 512	10,75	26,20	30,45	10,43	12,11
20	1350 < D <= 2445 e Mo = 3,2,6,9	23,20	40,66	35,21	20,71	30,25
30	D > 30578	22,19	47,42	48,33	22,50	21,70
31	D <= 76	16,95	36,67	-	-	5,67
32	76 < D <= 235	10,25	24,67	10,00	10,50	9,10
33	512 < D <= 921 e Md = 2,4,10,5	11,26	21,30	32,52	14,73	15,38
34	512 < D <= 921 e Md = 3,8,7,6,1,9	13,82	26,67	32,14	11,80	17,48
35	D > 921 e Md = 4,7,10,5	12,88	24,69	24,32	17,78	20,95
36	D > 921 e Md = 3,8,2,6,1,9	16,46	29,48	27,72	18,70	22,50
37	1350 < D <= 2445; Mo = 8,4,7,10,5,1 e Md = 4,7,10	16,94	30,28	28,77	18,47	23,87
38	1350 < D <= 2445; Mo = 8,4,7,10,5,1 e Md = 3,8,2,5,6,1,9	19,20	33,11	31,66	22,62	26,84
39	2445 < D <= 3682; Mo = 8,4,7,10,5 e Md = 4,7,10,5	21,98	34,89	29,93	27,86	25,01
40	2445 < D <= 3682; Mo = 8,4,7,10,5 e Md = 3,8,2,6,1,9	24,81	39,07	34,12	29,77	25,55
41	2445 < D <= 3682; Mo = 3,2,6,1,9 e Ch_s <= 5	29,74	47,13	40,19	32,00	34,68
42	2445 < D <= 3682; Mo = 3,2,6,1,9 e Ch_s > 5	24,03	38,39	34,40	28,00	26,88
43	3682 < D <= 5281; Mo = 8,4,7,10,5 e Ch_s <= 5	31,81	46,18	40,20	24,75	31,16
44	3682 < D <= 5281; Mo = 8,4,7,10,5 e Ch_s > 5	25,69	39,87	36,32	30,00	30,00
45	3682 < D <= 5281; Mo = 3,2,6,1,9 e Md = 3,2,7,10,5	33,22	45,73	40,21	30,00	20,00
46	3682 < D <= 5281; Mo = 3,2,6,1,9 e Md = 8,4,6,1	38,15	55,86	48,16	28,00	44,55
47	5281 < D <= 7669; Mo = 3,2,6,1 e Ch_s <= 5	46,04	69,95	56,50	41,25	17,73
48	5281 < D <= 7669; Mo = 3,2,6,1 e Ch_s > 5	33,99	54,00	44,76	-	20,00
49	5281 < D <= 7669; Mo = 8,4,7,10,5,9 e Md = 3,2,7,6,1	38,14	61,36	49,24	39,00	12,45
50	5281 < D <= 7669; Mo = 8,4,7,10,5,9 e Md = 8,4,10,5,9	34,81	53,05	47,14	60,00	13,28
51	7669 < D <= 11134; Mo = 8,4,7,10,5 e Ch_s <= 5	45,36	68,49	58,19	-	21,00
52	7669 < D <= 11134; Mo = 8,4,7,10,5 e Ch_s > 5	33,53	65,40	57,07	30,00	15,00
53	7669 < D <= 11134; Mo = 3,2,6,1,9 e Md = 8,4,5,6	53,71	80,66	67,77	45,00	29,55
54	7669 < D <= 11134; Mo = 3,2,6,1,9 e Md = 3,2,7,10,1	46,01	69,40	57,93	-	10,50
55	11134 < D <= 30578 e Mo = 4,7,10,5	44,75	77,16	77,21	-	23,73
56	11134 < D <= 30578 e Mo = 8,3,2,6,1,9	56,03	93,03	82,22	-	19,59

TVM: tempo de viagem médio; D: distância (metros); Mo: motivo na origem (1, 2, 3 – Trabalho na indústria, comércio e serviço, respectivamente; 4 – Escola; 5 – Compras; 6 – Saúde, 7 – Lazer; 8 – Residência; 9 – Procurar emprego; 10 – Assuntos pessoais); Md: motivo no destino; Ch_s: *cluster* hora de saída (1: 6 às 9h; 2: 9 às 12h; 3: 12 às 14h; 4: 14 às 16h; 5: 16 às 20h; 6: 20 às 6h).

5. CONCLUSÕES

O presente trabalho propôs um procedimento para caracterização agregada das alternativas modais com utilização de dados de Preferência Revelada. O procedimento para determinação de classes, a partir de variáveis associadas às viagens, utilizou os algoritmos CHAID e CART.

Os algoritmos baseiam-se na formação de grupos homogêneos, segundo a variável dependente, e otimização de agrupamentos levando-se em conta a escolha de variáveis

independentes (bem como valores de corte) que tornam as divisões de classes significativas. O procedimento apresenta contribuições importantes levando-se em conta os seguintes fatores:

(1) A Pesquisa OD é a mais tradicionalmente utilizada no Brasil para obtenção de dados, no entanto, traz apenas características das viagens realmente utilizadas, inviabilizando o uso adequado de modelagem de escolha discreta, pela ausência de dados relativos às alternativas não utilizadas;
(2) Trabalhos, encontrados na literatura, propuseram a caracterização agregada das alternativas a partir de critérios empíricos, segundo escolha de variáveis, bem como valores de corte;
(3) É proposto um critério, baseado em algoritmo não paramétrico, para agrupamento de viagens e obtenção de valores médios de variáveis que caracterizem alternativas (determinados pelos nós terminais);
(4) A técnica é de fácil aplicação, sem restrições relacionadas a tipos de variáveis ou distribuições populacionais;
(5) O mesmo algoritmo apresentou bons resultados segundo validações propostas;
(6) O método pode ser replicado futuramente para qualquer outra variável que caracterize as alternativas modais, tais como custo de viagem, por exemplo.

Agradecimentos

Os autores agradecem ao CNPq, à FAPESP (Processo 14/06290-3) e à Companhia do Metropolitano de São Paulo.

REFERÊNCIAS

Antonini, G., M. Bierlaire, e M. Weber (2006) Discrete choice models of pedestrian walking behavior. *Transportation Research Part B*, Vol.40, pp.667-687.

Ben-Akiva, M. e T. Morikawa (1990) Estimation of travel demand models from multiple data sources, Koshi, M. *The 11th International Symposium on Transportation and Traffic Theory*, Amsterdam, Netherlands.

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

Fezzi, C., S. Ferrini, e I. J. Bateman, (2014) Using revealed preferences to estimate the value of travel time to recreation sites. *J. Environ. Econ. Manage*., 67(1), 58–70, doi:10.1016/j.jeem.2013.10.003.

Frejinger, E., (2008) *Route choice analysis: data, models, algorithms and applications.* Phd Dissertation. École Polytechnique Fédérale de Lausanne. France.

Goodman, L. A. (1979). Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories. *Journal of the American Statistical Association*, 74, 537-552.

Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. *Appl. Stat*. 29, 119–127. http://dx.doi.org/10.2307/2986296.

Kato, H., T. Oda e Sakashita, A. (2013). Valuation of travel time saving with revealed preference data in Japan: Further Analysis. *In 13th WCT. CPAPER*, Rio de Janeiro, Brasil.

Magidson, J. (1994) The chaid approach to segmentation modeling: chisquared automatic interaction detection. In: Bagozzi, R.P. (Ed.), *Advanced Methods of Marketing Research*. Blackwell, Cambridge, pp. 118–159.

Metro(2008) Companhia de Trem Metropolitano de São Paulo. Resultados da Pesquisa Origem-Destino 2007.

Morikawa, T. (1989) *Incorporating stated preference data in travel demand analysis.* Ph.D. Dissertation, Department of Civil Engineering, MIT.

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

Quinlan, R. (1983) Learning efficient classification procedures and their application to chess end-games. *Machine Learning: An Artificial Intelligence Approach*, Tioga, Palo Alto, pp. 463–482.

Silva, F.G. F (2015) Modelando valor de tempo de viagem para modos concorrentes por diferentes modelos Logit: o que se ganha e o que se perde? Tese de Doutorado. UFC.

Souza, H.H.H., F.F.L.M. Sousa, F.M. Oliveira Neto, R.M.C. Freire e C.F.G. Loureiro (2017). Estimação do valor do tempo com base em pesquisas domiciliares de origem e destino: desafios teóricos e dificuldades práticas. Anais do XXXI Congresso da ANPET – Recife, PE – 2017.