Engenharia Civil UM
2023, No. 63, 18-30

https://revistas.uminho.pt/index.php/ecum

Obtaining characteristics of alternatives from Revealed Preference data using the CART algorithm

Obtenção de características das alternativas a partir de dados de Preferência Revelada e algoritmo CART

V. A. Gomes[a], C. S. Pitombo [a†], L. Assirati[a], C. F. Cerveira[a]

[a] *University of São Paulo, São Carlos School of Engineering,* Transportation *Engineering Department São Carlos, São Paulo, Brazil*
[†] *Associate Professor, corresponding author:cirapitombo@usp.br*

ABSTRACT

In general terms, discrete choice models are calibrated using data obtained from Revealed Preference (RP) and Stated Preference (SP) surveys. In transportation planning, one of the main sources of data is the Origin/Destination (O/D) Survey, which is an RP survey and describes the actual choices and behaviors of individuals. However, it is not possible, through this source, to characterize the alternatives not chosen. This study has two related aims: (1) to propose a criterion to characterize the travel mode alternatives using RP data, and (2) to test the improvement of travel mode choice estimates based on including characteristics of alternatives. First, the CART (Classification and Regression Tree) algorithm was used to characterize the travel times of the travel modes available in the study area (city of São Paulo, Brazil). The trips were classified according to independent variables selected by the algorithm, and average travel time values were obtained for five travel mode alternatives – information not previously available in the RP survey. Finally, the improvement of discrete choice modeling, based on including average travel times, was tested using a validation sample and performance metrics, such as Hit rates and LogLikelihood values. An increase in estimates was observed from including travel duration, and the proposed method is an academic contribution to the modeling based on RP data.

Keywords:
Travel time; Travel mode choice; RP data; Decision Tree

RESUMO

Em termos gerais, a calibração do modelo de escolha discreta se dá através de dados obtidos por pesquisas de Preferência Revelada (PR) e Declarada (PD). No planejamento de transportes, uma das principais fontes de dados é a Pesquisa O/D, que é uma pesquisa de PR e descreve as escolhas e comportamentos reais dos indivíduos. Entretanto, não é possível, através desta fonte, caracterizar as alternativas não escolhidas. Este trabalho possui dois objetivos associados: (1) propor um critério para caracterizar, de forma agregada, as alternativas modais, utilizando dados de PR e (2) testar o aprimoramento de estimativas de escolha modal a partir da inclusão das características agregadas das alternativas. Primeiramente, foi utilizado o algoritmo CART (*Classification And Regression Tree*) para caracterizar os tempos de viagem dos modos de transporte disponíveis na área de estudo (Cidade de São Paulo, Brasil). As viagens foram agrupadas, segundo variáveis independentes selecionadas pelo algoritmo, e foram obtidos valores médios de tempos de viagens para cinco alternativas modais – informação anteriormente não disponível na pesquisa de PR. Finalmente, o aprimoramento da modelagem de escolha discreta, a partir da inclusão dos tempos de viagens médios, é testado através de uma amostra de validação e métricas de desempenho, tais

Palavras-chave:
Tempo de viagem; Escolha do modo de viagem; Dados de PR; Árvores de Decisão.

como Percentual de Acertos e Valor do log da Verossimilhança. Observou-se um incremento das estimativas a partir da inclusão das durações de viagens, sendo o método proposto uma contribuição acadêmica para a modelagem a partir de dados de PR.

1. Discrete choice models and revealed and stated preference Surveys

Discrete choice models were developed to characterize consumer behavior and choices [1,2]. However, the application of these models has been extended to several areas of knowledge, as they combine the economic theory of behavior with an econometric method of dispersion analysis at the individual level.

According to [3], a consumer who is in the process of choosing to purchase a product or service analyzes the available alternatives and chooses the one that provides the greatest satisfaction. The process can be characterized by a system formed by the following elements: (1) decision maker; (2) alternatives; (3) attributes of alternatives; (4) decision rule. The present study focuses on characterizing the attributes of the alternatives, more specifically an attribute of the alternatives of the travel modes available in the study area, the travel time.

Discrete choice model applications provide a good degree of predictability of user behavior and, therefore, have been widely used in the area of transport demand for decades [4,5,6,7,8,9,10,11]. These models make use of data obtained from Revealed Preference (RP) and Stated Preference (SP) techniques, in which the RP data represent choices effectively made by individuals and the SP data refer to choices considering a set of options in which hypothetical scenarios are presented to the consumer for him/her to indicate his/her option [12,13].

For decades, many authors have developed studies combining the two types of surveys [4,14,15,16,17]. However, by using only RP data, individuals and not the alternatives can be characterized, possibly affecting discrete choice modeling.

1.1. Characterization of alternatives from RP data

Origin-Destination (OD) household surveys are important sources of data for studies in the field of transportation engineering and allow future projections of a community's travel needs to be established. For years, these surveys have provided information to urban and transport planning in many cities.

OD surveys, as a Revealed Preference (RP) database, describe the real choices and behaviors of individuals, they have a wide spatial coverage and a large number of interviews. However, the data portray the information of the trips actually made by the interviewee, not presenting, however, information about the other possible alternatives.

To improve discrete choice modeling, it is important to have variables that describe the possible alternatives in the choice set. Some studies have proposed different methods to estimate aggregate characteristics of alternatives, based on Revealed Survey data [18,19,20].

What can be observed is that in these studies, only the time and cost of the trip were used, and the grouping criteria were defined subjectively. In the present article, a set of variables associated with trips is used and a clustering tool is applied with dependency relationships and criteria based on the homogeneity of the groups to estimate, in an aggregated way, the travel time of all alternatives available in the study area.

In this context, this study has two objectives: (1) to propose a criterion to characterize the travel mode alternatives, using RP data and (2) to test the improvement of travel mode choice estimates from the inclusion of aggregate characteristics of the alternatives. First, the CART (Classification and Regression Tree) algorithm was used to characterize the travel times of the travel modes available in the study area (City of São Paulo, Brazil – OD Survey 2007). Then, the Multinomial Logit Model (MNL) was carried out, initially with socioeconomic variables. Finally, utility functions were calibrated including travel times of all travel mode alternatives, previously estimated by the CART algorithms.

Recently, an amount of travel data could be obtnained using cellular networks and app data. Some authors have tried to get trip information using these tools [21,22]. This information

could be important to complete RP surveys for example. However, the location data available could be very sparse on cellular networks. Additionaly, some important information regarding trip purpose were always missing on apps collection. Thus, there were some important discussions considerering how these data can be processed in order to efficiently be used for travel behavior modeling. Finally, despite of all the possibilities of getting travel information, this study purposes a procedure to get travel data easely, associated to importante information such as trip purposes, travel mode use, etc.

2. Description of the tools used

2.1. CART (Classification and Regression Tree)

The CART algorithm successively performs binary partitions of the database, based on inductive rules of the "If...then..." type, to obtain increasingly homogeneous subsets according to dependent variable values. Its structure resembles a tree. The total dataset (root node) is separated by sequential divisions (child nodes). These divisions continue until the terminal nodes (or leaves), when it is no longer possible to obtain any subgroup, considering the stopping rules adopted. For the construction of the tree, three parameters must be defined: a set of rules deliminting data division; a criterion to evaluate the best division to produce the child nodes; and a rule that determines the limit of subdivisions (stop-splitting rule) [23].

For the numerical dependent variable case, as in the work carried out by Pianucci and Pitombo [24], the splitting criterion is called a reduction in variance [25,26] which represents the reduction in variance of the dependent variable within each node. The reduction in variance, which represents the impurity function, is presented in Equation

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

Where:

Iv (N) = a reduction in variance at node N; S = test sample set; St = test sample set for which the explanatory variable value is true; Sf = test sample set for which the explanatory variable value is false; xi = dependent variable value of the test sample; xj = dependent variable value of the sample that comprises node N.

It should be mentioned that the main result, in this case, would be the set of observations of each terminal node, associated with the average dependent variable value. In the case of this study, each set of trips, which comprises each of the obtained terminal nodes, is associated with average values of travel times.

2.2. Discrete choice models

Discrete choice models are based on the microeconomic theory of the consumer, which provides a basis for identifying individual preferences [3]. The principle of discrete choice models is to estimate utility functions. These functions measure the preference for alternatives and are based on a combination of coefficients and variables, which characterize the alternatives and the individuals [1].

In the Multinomial Logit model (MNL), used in this study, utility is treated as a random variable, formed by a component called deterministic or systematic, and another random component, which reflects the "irrationalities" of the individual choice [1]. Thus, the utility for an alternative i for an individual n (U_{in}) can be expressed through Equation 2.

$$U_{in} = V_{in} + \varepsilon_{in} \quad (2)$$

Where U_{in} is the global utility of an alternative i for an individual n, V_{in} is the systematic component of the utility of an alternative i for an individual n and ε_{in}, the random component (which can be a function), represents an unknown portion of the utility function that captures the dispersion of choices and factors not controllable or unknown. The error function is a random deviate, which contains all the unobserved determinants of the utility. One important assumption

of MNL model is the independence of the errors. The logit model is obtained by assuming that each ε_{in} is independently, identically distributed extreme value. The distribution is also called Gumbel and type I extreme value [2]. The most common way of representing the systematic components is linear (Equation 3).

$$V_{in} = \beta_0 + \beta_1 x_{in1} + \beta_2 x_{in2} + \beta_3 x_{in3} + \cdots + \beta_k x_{ink} \quad (3)$$

Where:
k : number of attributes of alternative i for individual n
β_0 :: Constant
β_k : Relative weight of the x_{ink} attribute in the composition of the utility function.
Finally, the probability of choosing an alternative can be calculated by Equation 4.

$$Pr(i/C_n) = Pr\{U_{in} \geq U_{jn} \; \forall \; j \in C_n\} = Pr\{U_{in} = max_{j \in C_n} U_{jn}\} \quad (4)$$

Where Cn corresponds to the set of choices of n individuals.

The Multinominal Logit model is estimated by the maximum likelihood (Equation 5). The maximization of the function is obtained by maximizing the productivity of the probabilities of the alternatives actually chosen by each individual.

$$LL = \sum_{n=1}^{N} \sum_{i \in C_n} P_{ni}{}^{yin} \quad (5)$$

3. Materials and Methods

3.1. Materials

This study used data from the Origin-Destination Survey (OD), carried out in São Paulo Metropolitan Area (SPMA), Brazil in 2007, in which information was collected from 30,000 randomly chosen households. In these households, distributed in the 460 Traffic Analysis Zones (TAZs), approximately 120 thousand people were interviewed. In this study, only the interviews carried out in the city of São Paulo were used (Figure 1).

The latest database available in the period related to this analysis was from 2007 OD survey.

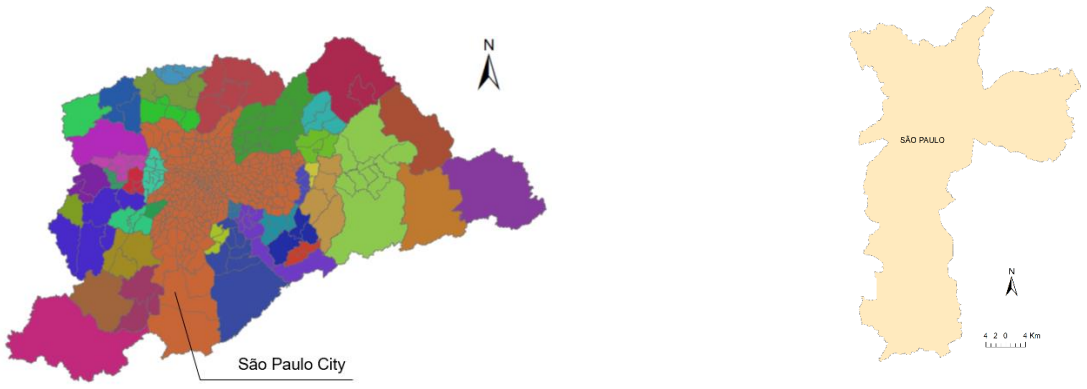

Figure 1 - Origin Destination Survey – SPMA 2007. Source: Metrô (2008)

The survey comprises four databases: aggregated Traffic Analysis Zones: disaggregated trips: disaggregated households: and disaggregated individuals. In this study, disaggregated trip data were used, associated with the individual's identifier, as well as their characteristics and the household variables. Table 1 describes the variables used in this study. The variable “departure trip time” was grouped into six categories, depending on the peak and between peak periods. The table also describes the methodological step in which each of the variables was used.

Table 1 – Sample Variables.

Variables	Nature		Methodological Step	
Level of Education	Ordinal Qualitative			Logit
Number of cars	Quantitative	Discrete		Logit
Family income	Quantitative	Continuous		Logit
Gender	Nominal Qualitative			Logit
Age	Ordinal Qualitative			Logit
Origin trip purpose	Nominal Qualitative		CART	
Destination trip purpose	Nominal Qualitative		CART	
Departure trip time	Nominal Qualitative		CART	
Time walking at the origin	Quantitative	Continuous	CART	
Time walking at the destination	Quantitative	Continuous	CART	
Travel time for used travel mode	Quantitative	Continuous	CART	
Main travel mode	Nominal Qualitative		CART	Logit
Trip Distance	Quantitative	Continuous	CART	

3.2. Method

The proposed method comprises the application of two tools: (1) the CART algorithm, to obtain the travel times, aggregated by groups, of all travel modes available in the study area and (2) the modeling through the Multinomial Logit model in two different steps. The first considering only the socioeconomic characteristics of the interviewees (Model 1) and a second modeling (Model 2) in which the estimated travel times of all travel modes available in the study area were included. Figure 2 illustrates the sequence of the methodological procedure, and the next subsections describe the illustrated steps.

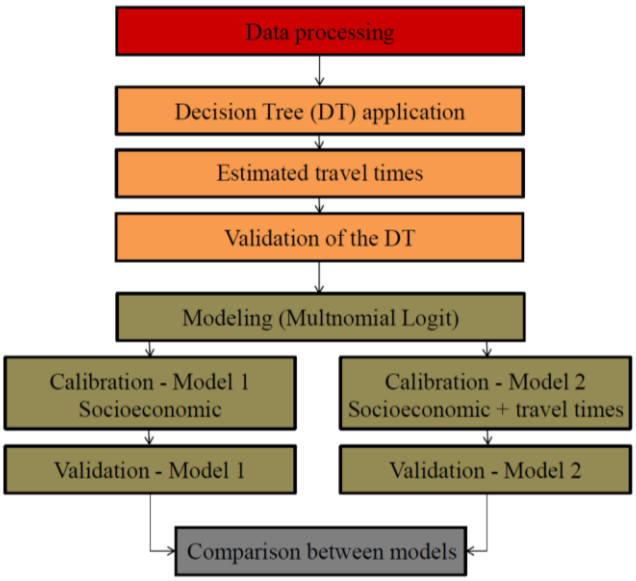

Figure 2 – Methodological sequence

Data processing: Initially, the database was analyzed and the disaggregated final sample (by trips) was obtained with socioeconomic data on the interviewee, household and travel data. The final sample is characterized by the trip carried out, associated with the individual identifier. The data are presented in Table 1.

CART algorithm application and estimation of aggregate travel times: In this step, the CART algorithm was applied to estimate the travel times of all available travel modes. The independent variables used were: trip distance, departure trip time clusters, origin trip purpose, destination trip purpose. The main travel modes were grouped into five categories: private motorized travel mode (1), bus (2), subway or train (3), bicycle (4) and walking (5). This procedure was previously proposed by Gomes *et al.* [27].

Validation of travel time estimates using Decision Tree (DT): In order to validate this step of the method, statistical tests were performed, comparing the estimated travel times (by the travel mode actually used) with the travel duration values actually performed by the interviewee. The sample was randomly split and training (70%) and test (30%) samples were obtained. The error measures were then calculated with the test sample: Mean Square Error, Root Mean Square Error, Mean Absolute Error and Pearson's Correlation.

Multinomial Logit Modeling: In this methodological step, the utility functions of the Multinomial Logit models were defined. Part of the sample was randomly separated for calibration (70%) and validation (30%). In this step, the Biogeme software [28] was used. The first model (Model 1) contained only socioeconomic variables, while the second model (Model 2) incorporated the aggregate variables of travel times.

Model 1: In Model 1, the socioeconomic characteristics of the individuals were considered as independent variables, namely: “level of education”, “age”, “gender”, “number of cars” and “family income”. The dependent variable was the main travel mode (with five categories). Equations 6, 7, 8, 9 and 10 represent utility functions in their literal form. The authors chose to cancel one of the utility functions to reduce the number of parameters to be estimated. The choice of utility 2 as a reference (V2 = 0) occurred, considering the smallest number of observations for category 2 (bus). This modeling structure was proposed to minimize the number of non-significant parameters of the model. Additionally, the authors carried out diverse tests varying the reference utility function. The best model global performance was described as equations 6,7,8,9 and 10.

$$V_1 = ASC_1 + B_1\_GENDER*GENDER + B_1\_AGE*AGE + B_1\_LEVEL\ OF\ EDUCATION*LEVEL\ OF\ EDUCATION + B_1\_FAMILY\ INCOME*FAMILY\ INCOME + B1\_N\ OF\ CARS*N\ OF\ CARS \quad (6)$$

$$V_2 = 0 \quad (7)$$

$$V_3 = ASC_3 + B_3\_GENDER*GENDER + B_3\_AGE*AGE + B_3\_LEVEL\ OF\ EDUCATION*LEVEL\ OF\ EDUCATION + B_3\_FAMILY\ INCOME*FAMILY\ INCOME + B3\_N\ OF\ CARS*N\ OF\ CARS \quad (8)$$

$$V_4 = ASC_4 + B_4\_GENDER*GENDER + B_4\_AGE*AGE + B_4\_LEVEL\ OF\ EDUCATION*LEVEL\ OF\ EDUCATION + B_4\_FAMILY\ INCOME*FAMILY\ INCOME + B4\_N\ OF\ CARS*N\ OF\ CARS \quad (9)$$

$$V_5 = ASC_5 + B_5\_GENDER*GENDER + B_5\_AGE*AGE + B_5\_LEVEL\ OF\ EDUCATION*LEVEL\ OF\ EDUCATION + B_5\_FAMILY\ INCOME*FAMILY\ INCOME + B5\_N\ OF\ CARS*N\ OF\ CARS \quad (10)$$

ASCi: alternative specific constant of alternative i.

Model 2: In Model 2, in addition to the variables listed in Model 1, the travel times estimated in the previous step were considered as independent variables. At this stage, to define if the model was generic or specifc, a simple test (likelihood ratio test) was performed to corroborate the null hypothesis that the coefficients are significantly similar. The null hypothesis is rejected for the following case:

$$-2(L_R - L_u) > \chi^2_{((1-\alpha),gl)} \quad (11)$$

L_R - L_u = difference between the Likelihood of the restricted (generic) and unrestricted model (specific coefficients); $\chi^2_{((1-\alpha),gl)}$ = Chi-Square distribution with for the level of significance and degrees of freedom – gl – equivalent to the difference of parameters estimated by the model with specific coefficients and generic model.

After deciding between calibrating the generic or specific model, its accuracy is evaluated through the adjusted Rho-square, Likelihood value, Loglikelihood and Akaike information criterion. The adjusted rho-quared metric is defined by Equation (12):

$$\rho_*^2 = 1 - \frac{L^* - K}{L_0} \tag{12}$$

L_0 is the likelihood value obtained by assuming all model parameters as zero and L^* is the maximum likelihood value obtained when the parameters correspond to the estimated values. Thus, an ideal model tends to the unit because the ratio L^* (case where the parameters have their optimal values) by L_0 (the case where the parameters are all null), tends to zero because L^* is much smaller than L_0. K is the number of estimated parameters.

The Akaike criterion is defined by Equation 13. The values of K and L* are similar to the previous ones. Established by subtracting the K number of parameters and the logarithm of the maximum likelihood L* value, the Akaike formulation makes the criterion penalize overfitting (the act of adding too many variables to the equations in order to obtain better adjustments, lacking criteria for such addition) and it is for this reason that lower values for this criterion are sought.

$$A = 2K - 2lnL^* \tag{13}$$

Validation, comparison between the validation models and results: The validation and comparison between models were performed with part of the sample, selected at random, and the goodness of fit of models 1 and 2 was measured. The hit rates and Likelihood values were used as parameters for measuring the quality of both models.

4. Results and Discussion

This section presents the results obtained in the aggregate characterization of the alternatives using the CART algorithm (objective 1) and in the modeling stage (objective 2), as well as the comparison of models and evaluation of the tools, using the validation samples.

4.1. The characterization of alternatives using the CART algorithm

For binary partitioning of the data, through the CART algorithm, the stopping criterion was used: minimum of observations in the terminal node = 30 observations. As a result, a total of 57 nodes were obtained, of which 29 were terminal nodes and a depth equal to 5. Each terminal node is characterized according to the cut-off conditions of the independent variables used and the average travel time associated with each travel mode, as shown in Table 2.

Table 2: Cut-off conditions at terminal nodes and travel times for the 5 travel modes.

Node	Cut-off conditions	ATT (min.)				
		Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
16	235 < D <= 512	10.75	26.2	30.45	10.43	12.11
20	1350 < D <= 2445 e Op = 3,2,6,9	23.2	40.66	35.21	20.71	30.25
30	D > 30578	22.19	47.42	48.33	22.5	21.7
31	D <= 76	16.95	36.67	-	-	5.67
32	76 < D <= 235	10.25	24.67	10	10.5	9.1
33	512 < D <= 921 e Dp = 2,4,10,5	11.26	21.3	32.52	14.73	15.38
34	512 < D <= 921 e Dp = 3,8,7,6,1,9	13.82	26.67	32.14	11.8	17.48
35	D > 921 e Dp = 4,7,10,5	12.88	24.69	24.32	17.78	20.95
36	D > 921 e Dp = 3,8,2,6,1,9	16.46	29.48	27.72	18.7	22.5
37	1350 < D <= 2445; Op = 8,4,7,10,5,1 e Dp = 4,7,10	16.94	30.28	28.77	18.47	23.87

38	1350 < D <= 2445; Op = 8,4,7,10,5,1 e Dp = 3,8,2,5,6,1,9	19.2	33.11	31.66	22.62	26.84
39	2445 < D <= 3682; Op = 8,4,7,10,5 e Dp = 4,7,10,5	21.98	34.89	29.93	27.86	25.01
40	2445 < D <= 3682; Op = 8,4,7,10,5 e Dp = 3,8,2,6,1,9	24.81	39.07	34.12	29.77	25.55
41	2445 < D <= 3682; Op = 3,2,6,1,9 e Ch_s <= 5	29.74	47.13	40.19	32	34.68
42	2445 < D <= 3682; Op = 3,2,6,1,9 e Ch_s > 5	24.03	38.39	34.4	28	26.88
43	3682 < D <= 5281; Op = 8,4,7,10,5 e Ch_s <= 5	31.81	46.18	40.2	24.75	31.16
44	3682 < D <= 5281; Op = 8,4,7,10,5 e Ch_s > 5	25.69	39.87	36.32	30	30
45	3682 < D <= 5281; Op = 3,2,6,1,9 e Dp = 3,2,7,10,5	33.22	45.73	40.21	30	20
46	3682 < D <= 5281; Op = 3,2,6,1,9 e Dp = 8,4,6,1	38.15	55.86	48.16	28	44.55
47	5281 < D <= 7669; Op = 3,2,6,1 e Ch_s <= 5	46.04	69.95	56.5	41.25	17.73
48	5281 < D <= 7669; Op = 3,2,6,1 e Ch_s > 5	33.99	54	44.76	-	20
49	5281 < D <= 7669; Op = 8,4,7,10,5,9 e Dp = 3,2,7,6,1	38.14	61.36	49.24	39	12.45
50	5281 < D <= 7669; Op = 8,4,7,10,5,9 e Dp = 8,4,10,5,9	34.81	53.05	47.14	60	13.28
51	7669 < D <= 11134; Op = 8,4,7,10,5 e Ch_s <= 5	45.36	68.49	58.19	-	21
52	7669 < D <= 11134; Op = 8,4,7,10,5 e Ch_s > 5	33.53	65.4	57.07	30	15
53	7669 < D <= 11134; Op = 3,2,6,1,9 e Dp = 8,4,5,6	53.71	80.66	67.77	45	29.55
54	7669 < D <= 11134; Op = 3,2,6,1,9 e Dp = 3,2,7,10,1	46.01	69.4	57.93	-	10.5
55	11134 < D <= 30578 e Op = 4,7,10,5	44.75	77.16	77.21	-	23.73
56	11134 < D <= 30578 e Op = 8,3,2,6,1,9	56.03	93.03	82.22	-	19.59

ATT: Average Travel Time; D: distance (in meters); Op: Origin trip purpose (1, 2, 3 – Work in industry, commerce and services, respectively; 4 – School; 5 – Shopping; 6 – Health, 7 – Leisure; 8 – Household; 9 – Look for employment; 10 – Personal purpose); Dp: Destination trip purpose; Ch_s: *cluster* departure trip time (1: 6 to 9am; 2: 9 am to 12; 3: 12 to 2pm; 4: 2pm to 4pm; 5: 4pm to 8pm; 6: 8pm to 6am). Mode 1: Private Motorized; Mode 2: Bus; Mode 3: Subway and Train; Mode 4: Bicycle; Mode 5: Walking).

Figure 3 illustrates the tree map obtained for the training sample. The 57 nodes obtained are illustrated, as well as the 29 terminal nodes described in Table 2.

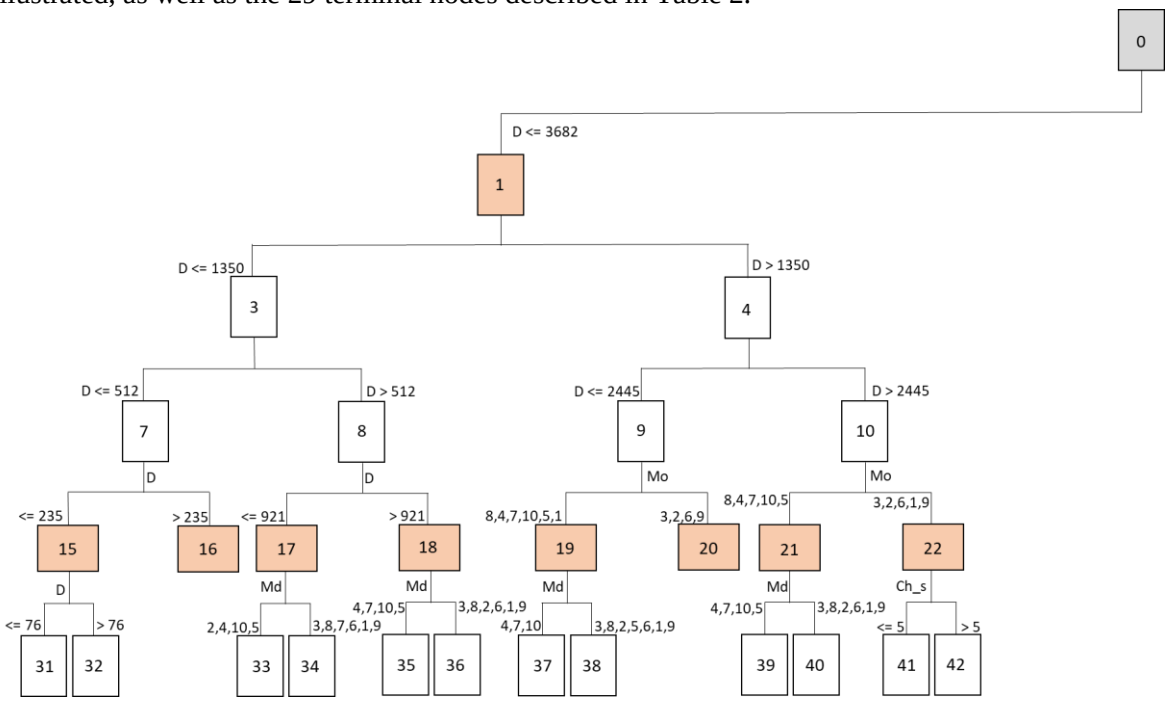

Figure 3 (a) - Map of the CART Algorithm - Branch 1

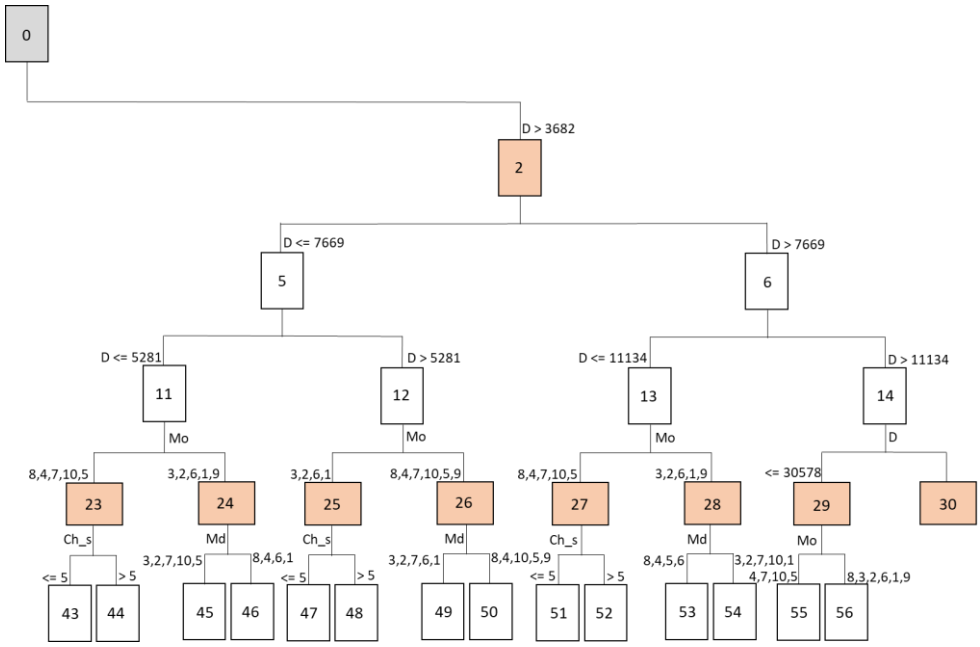

Figure 3(b) - Map of the CART Algorithm - Branch 2

In this tree, the important independent variables were: “D: trip distance”, “Ch_s: cluster trip departure time”, “Op: Origin trip purpose” and “Dp: Destination trip purpose”. Travel times, for all travel mode alternatives, were associated with the terminal nodes obtained at the fourth and fifth levels of the tree. Once the tree was generated, filters were made and the travel times for all five travel modes were identified at each terminal node (1: Private Motorized; 2: Bus; 3: Subway and Train; 4: Bicycle; 5: Walking).

In the validation, the following error measures were obtained: 378.677 for Mean Square Error, 19.46 for Root Mean Square Error, -0.065 for Mean Absolute Error and Pearson's Correlation was 0.638. The calculation of the measurements was performed considering observed and estimated values of travel times of the travel modes actually used from the test sample.

4.2. Modeling the alternatives

4.2.1 Model 1

For the Model 1 calibration sample, the following results were obtained: Rho-square adjusted equal to 0.326 and Akaike Information Criterion of 8.33 x 104. With the validation sample, the following measures were calculated: Hit rates of 60.00%, likelihood value L = 1.47 x 10-42 and log (L) = -96.32.

For disaggregated models for travel mode choice, values and Rho-squared adjusted around 0.20 and 0.40 are the most commonly found in the literature [29,4,30].

Equations 14, 15, 16, 17 and 18 are the calibrated utility functions of the travel modes: (1) private motorized mode, (2) bus, (3) subway or train, (4) bicycle and (5) walking, with the estimated parameters that were significant at a 95% confidence level.

$$V_1 = -1{,}36 - 0.390*\text{GENDER} + 0{,}194*\text{AGE} + 0{,}262*\text{LEVEL OF EDUCATION} + 0{,}0000706*\text{FAMILY INCOME} + 0{,}994*\text{N\_OF CARS} \quad (14)$$

$$V_2 = 0 \quad (15)$$

$$V_3 = -2{,}01 - 0{,}350*\text{GENDER} + 0{,}145*\text{AGE} + 0{,}372*\text{LEVEL OF EDUCATION} + 0{,}000017*\text{FAMILY INCOME} + 0{,}117*\text{N\_OF CARS} \quad (16)$$

$$V_4 = -2{,}35*\text{GENDER} - 0{,}11*\text{LEVEL OF EDUCATION} \quad (17)$$

$$V_5 = 1{,}09 - 0{,}115*\text{GENDER} - 0{,}0598*\text{AGE} - 0{,}0582*\text{LEVEL OF EDUCATION} \quad (18)$$

The analysis of Model 1 brings results already proven in the literature on relationships between socioeconomic variables and travel mode choice [31,32,33]. The modeling shows a greater propensity to use the car (V1) for the male gender (Variable "gender = 0, Men; gender = 1, Women), as well as for older people, higher education level, higher family income and higher number of cars in the household (all parameters associated with such variables are positive). For the use of the subway or train (V3), relationships similar to those found for the use of the car are observed, however such relationships are demonstrably weaker, considering the intensity (in module) of the values of the estimated parameters, except for the case of variable "level of education". The independent term for the utility of the subway also proves a lesser utility of this mode, to the detriment of the automobile. Bicycle use (V4) is associated with males and a lower level of education. In the case of the utility of the walking mode (V5), it presents relationships similar to those found for the use of the bicycle, including the negative influence of the variable "Age".

4.2.2 Model 2

For the modeling that included travel times for the five travel mode options (Model 2), the likelihood ratio test was performed. The null hypothesis of similarity of coefficients (specific/unrestricted and generic/restricted) was refuted and then, a model with specific coefficients was chosen, associated with travel times for each alternative (unrestricted model). The statistically significant parameters, for a confidence level of 95%, are those presented in Equations 19, 21, 22 and 23. In this second model, the adjusted Rho-square equal to 0.454 and the Akaike Information Criterion of 6.17 x 104 were obtained. Concerning the validation sample, the following measures were calculated: Hit rates of 67.47%, with likelihood value L = 6.51 x 10-37 and log (L) = - 83.32. The global improvement of the modeling can be verified, in the calibration stage, by the increase of the adjusted Rho-square metrics and the decrease of the Akaike value. In the validation stage, the improvement of the estimates can be verified through the increase of L and decrease (in module) of Log (L), as well as by the increase in correct answers.

$$V_1 = -1{,}51 - 0{,}402*\text{GENDER} + 0{,}234*\text{AGE} + 0{,}262*\text{LEVEL OF EDUCAYION} + 0{,}0000713*\text{FAMILY INCOME} + 1{,}04*\text{N OF CARS} \quad (19)$$

$$V_2 = 0 \quad (20)$$

$$V_3 = -4{,}05 - 0{,}270*\text{GENDER} + 0{,}159*\text{AGE} + 0{,}316*\text{LEVEL OF EDUCATION} + 0{,}0000286*\text{FAMILY INCOME} + 0{,}126*\text{N OF CARS} + 0{,}0464 * \text{TRAIN TRAVEL TIME} \quad (21)$$

$$V_4 = 1{,}87 - 2{,}55*\text{GENDER} - 0{,}0809* \text{BIKE TRAVEL TIME} \quad (22)$$

$$V_5 = 5{,}68 - 0{,}209*\text{GENDER} + 0{,}031* \text{LEVEL OF EDUCATION} - 0{,}224 * \text{WALKING TRAVEL TIME} \quad (23)$$

Regarding socioeconomic variables, the same relationships between them and the utility of using a particular travel mode (V1; V3; V4 and V5) were found as in the previous modeling (Model 1). All these relationships have been previously proven in the literature. Regarding travel time, it is expected that the increase in travel time of a given travel mode alternative will negatively contribute to its usefulness [4]. This fact is proven in the calibrated equations for bicycle (V4) and walking (V5) modes. For the subway or train (V3), however, a coefficient with low and positive value is found, associated with the duration of the trip by subway/train. This result can be explained by the fact that, as in São Paulo, the fare is fixed (regardless of the travel distance), for very long trips, users opt for public transport, even if they have a car at home.

5. CONCLUSIONS

The present paper aimed to verify the improvement of the travel mode choice estimates, from the inclusion of variables related to the alternatives, obtained through CART algorithms and RP data. Initially, a modeling was performed with socioeconomic variables and Multinomial Logit Model, and later, utility functions were calibrated with the inclusion of travel times of all mode alternatives, previously estimated.

The data showed an improvement in the model from the inclusion of travel times. The CART algorithm, used in this study to estimate travel times, is based on the formation of homogeneous groups, according to the dependent variable, and cluster optimization taking into account the choice of independent variables (as well as cut-off values) that make class divisions meaningful. The procedure makes important contributions taking into account the following factors:

• The OD Survey is traditionally used in many countries. However, it only brings characteristics of the trips actually used, making the proper use of discrete choice modeling unfeasible, due to the lack of data related to unused alternatives.

• Some studies, previously found in the literature, proposed the aggregate characterization of the alternatives based on empirical criteria, according to the choice of variables, as well as cut-off values.

• To characterize the alternatives, through RP, the present article proposed a criterion, based on a non-parametric algorithm, for grouping trips and obtaining average values of variables that characterize alternatives (determined by the terminal nodes).

• The technique is easy to apply, without restrictions related to types of variables or population distributions.

• The same algorithm presented good results according to proposed validations.

• The method can be replicated in the future for any other variable that characterizes the mode alternatives, such as travel cost, for example.

• The modeling increment, through the inclusion of the variable that characterizes the alternative, is observed.

• The methodological sequence proposed here (CART application followed by Multinomial Logit model calibration) can be replicated for other engineering applications that consider choices between alternatives.

Acknowledgments

The authors would like to thank CNPq (304345/2019-9) and CAPES and the Companhia do Metropolitano de São Paulo.

References

[1] M. Ben-Akiva, M. Bierlaire. Discrete choice models with applications to departure time and route choice, In: R.W. Hall (Ed.), Handbook of Transportation Science, second edition, Kluwer Academic, 2003, pp. 7-38.

[2] D. L. McFadden. The Measurement of Urban Travel Demand. J. Public Econ., 4 (1974) 303-328.

[3] M. Ben-Akiva, S. Lerman. Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, Cambridge MA, USA, 1985, pp. 59-99.

[4] M. Ben-Akiva, T. Morikawa. Estimation of switching models from revealed preferences and stated intentions. Transp. Res., 24 (1990) 485–495.

[5] G. Antonini, M. Bierlaire, M. Weber. Discrete choice models of pedestrian walking behavior. Transportation Research Part B, 40 (2006) 667-687. https://doi.org/10.1016/j.trb.2005.09.006.

[6] E. Frejinger. Route choice analysis: data, models, algorithms and applications. Ph.D. Dissertation. École Polytechnique Fédérale de Lausanne, France, 2008.

[7] M. Ruiz-Pérez, J.M. Seguí-Pons. Transport Mode Choice for Residents in a Tourist Destination: The Long Road to Sustainability (the Case of Mallorca, Spain): Sustainability, 22 (2020) 9480. https://doi:10.3390/su12229480.
[8] M. U. C. Caldas, C. S. Pitombo, l. Assirati. Strategy to reduce the number of parameters to be estimated in discrete choice models: an approach to large choice sets. Travel Behav. Soc., 25 (2021) 1-17. https://doi.org/10.1016/j.tbs.2021.05.001.
[9] W.Q. Al-Salih, D. Esztergár-Kiss. Linking Mode Choice with Travel Behavior by Using Logit Model Based on Utility Function: Sustainability, 13 (2021). https://doi.org/10.3390/su13084332.
[10] C.S. Costa, C.S. Pitombo, F.L.U. Souza. Travel behavior before and during the COVID-19 pandemic in Brazil: mobility changes and transport policies for a sustainable transportation system in the post-pandemic period: Sustainability, 14 (2022) 4573. https://doi.org/10.3390/su14084573.
[11] A. Mahdi, J. Hamadneh, D. Esztergár-Kiss. Modeling of Travel Behavior in Budapest: Leisure Travelers. Transp. Res. Proc., 62 (2022) 310-317. https://doi.org/10.1016/j.trpro.2022.02.039.
[12] A.A. Ahern, N. Tapley. The use of stated preference techniques to model modal choices on interurban trips in Ireland. Transp. Res. A., 42 (2008) 15–27. https://doi.org/10.1016/j.tra.2007.06.005.
[13] J.J. Louviere, R.T. Carson, L. Burgess, D. Street, A. Marley. Sequential preference questions factors inuencing completion rates and response times using an online panel: J. Choice Model., 8 (2013) 1-18.
[14] D. Hensher, J. Louviere, J. Swait. Combining sources of preference data. J. Econom., 89 (1999) 197-221.
[15] C. Bhat, S. Castelar. A unified mixed logit framework for modeling revealed and stated preferences: formulation and application to congestion pricing analysis in the San Francisco bay area: Transp. Res. B, 36 (2002), 577-669.
[16] K. Train, W.W. Wilson. Estimation on stated-preference experiments constructed from revealed-preference choices: Transp. Res. B, 42 (2008) 191-203.
[17] Y. Qiao, Y. Huang, F. Yang, M. Zhang, L. Chen. Empirical study of travel mode forecasting improvement for the combined revealed preference/stated preference data-based discrete choice model. Adv. Mech. Eng., 8 (2016). https://doi.org/10.1177/168781401562483.
[18] H.H.H. Souza, F.F.L.M. Sousa, F.M. Oliveira Neto, R.M.C. Freire, C.F.G Loureiro. Estimação do valor do tempo com base em pesquisas domiciliares de origem e destino: desafios teóricos e dificuldades práticas. In: Anais do XXXI Congresso da ANPET, Recife, Brasil, 2017.
[19] C. Fezzi, S. Ferrini, I.J. Bateman. Using revealed preferences to estimate the value of travel time to recreation sites: J. Environ. Econ. Manage, 67 (2014) 58-70. doi:10.1016/j.jeem.2013.10.003.
[20] H. Kato, T. Oda, A. Sakashita. Valuation of travel time saving with revealed preference data in Japan: Further Analysis. In: 13th WCT, CPAPER, Rio de Janeiro, Brasil, 2013.
[21] M. Diao, Y. Zhu, J. Ferreira, C. Ratti. Inferring individual daily activities from mobile phone traces: a Boston example: Environ. Plan. B Plan. Des., 43 (2016) 920–940. https://doi.org/10.1177/0265813515600896.
[22] A. Dypvik Landmark, P. Arnesen, C.-J. Sodersten, O.A. Hjelkrem. Mobile phone data in transportation research: methods for benchmarking against other data sources: Transportation, 48 (2021) 2883–2905. https://doi.org/10.1007/s11116-020-10151-7.
[23] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.
[24] M.N. Pianucci, C.S. Pitombo. Uso de árvore de decisão para previsão de geração de viagens como alternativa ao método de classificação cruzada. Engenharia Civil UM, 56 (2019) 5-13.
[25] P. Geurts. Discretization variance in decision tree induction. Technical report, University of Liège, Dept. of Electrical and Computer Engineering, 2000.
[26] RL. De Mantaras. A distance-based attribute selection measure for decision tree induction: Machine Learning, 6 (1991) 81–92. https://doi.org/10.1023/A:1022694001379.

[27] V. A. Gomes, M. U. C. Caldas, C. S. Pitombo. An investigation of trip-chaining behaviour based on activity participation, socioeconomic variables and aggregated characteristics of modal alternatives: Transportes (Rio de Janeiro), 29 (2021) 173-193.
[28] M. Bierlaire. A short introduction to PandasBiogeme. Technical report TRANSP-OR 200605. Transport and Mobility Laboratory, ENAC, EPFL, 2020.
[29] F. Southworth. Calibration of multinomial logit models of mode and destination choice. Transp. Res. A, 15 (1981) 315-325. https://doi.org/10.1016/0191-2607(81)90013-3.
[30] S. Bekhor, Y. Shiftan. Specification and Estimation of Mode Choice Model Capturing Similarity between Mixed Auto and Transit Alternatives: J. Choice Model., 3 (2010) 29-49.
[31] D.T. Hartgen. Attitudinal and situational variables influencing urban mode choice: Some empirical findings: Transportation, 3 (1974). https://doi.org/10.1007/BF00167967
[32] K. Train, D. McFadden. The Goods/Leisure Tradeoff and Disaggregate Work Trip Mode Choice Models: Transp. Res., 12 (1978) 349-353.
[33] R. Barff, D. Mackay, R.W. Olshavsky. A Selective Review of Travel-Mode Choice Models: J. Consum. Res., 8 (1982) 370–380. https://doi.org/10.1086/208877.
[34] Companhia de Trem Metropolitano de São Paulo. Resultados da Pesquisa Origem-Destino 2007, http://www.metro.sp.gov.br/pesquisa-od/arquivos/OD_2007_Sumario_de_Dados.pdf, 2008. (acesso em 19 junho 2019).

ORCID

V. A. Gomes 0000-0001-7776-7902 (https://orcid.org/0000-0001-7776-7902)
C. S. Pitombo 0000-0001-9864-3175 (https://orcid.org/0000-0001-9864-3175)
L. Assirati 0000-0002-0118-2665 (https://orcid.org/0000-0002-0118-2665)