COMMUTE MODE CHOICE IN CITY OF SÃO PAULO: AN EMPIRICAL ANALYSIS

Resumo: O presente trabalho trata do processo de escolha do modo de transporte no deslocamento de casa para o trabalho. O objetivo é analisar como o tempo e o custo de deslocamento, assim como características do trabalhador, estão associados com a escolha de um determinado modal. Trabalha-se com hipótese de que tempo e custo de deslocamento influenciam negativamente a escolha dos modos de transporte. O método utilizado é o Logit Condicional com Variáveis Específicas das Alternativas. Nossa principal fonte de dados são as Pesquisas Origem Destino de 2007 e 2012 realizadas pela Companhia do Metropolitano de São Paulo. Os resultados corroboram nossa hipótese inicial de que o custo e o tempo de deslocamento estão negativamente correlacionados com a escolha modal, além disso, o modelo revelou um Valor do Tempo de Deslocamento de 1,78 e 1,09 (R$ nominais) para 2007 e 2012.

Palavras-Chave: Escolha do modo de transporte, tempos de deslocamento casa para o trabalho, modelo de escolha discreta

Código JEL: R41, R48, C13.

Abstract: The present works deals with the commute mode choice process. Our aim is to analyze how travel time and travel cost, as well as the commuter characteristics, are associated with the probability of choice of a certain mode of transport. Our main hypothesis is that both time and cost are negatively associated with the choice of any mode. Our method is the Alternative-Specific Conditional Logit.. Our main data sources are from the 2007 and 2012 Origin Destination Surveys conducted by the Companhia do Metropolitana de São Paulo. The results corroborate our initial hypothesis that travel time and cost are negatively correlated with mode choice, besides that, our model revealed a Value of Travel Time of 1.78 and 1.09 (nominal R$) for 2007 and 2012.

Keywords: Mode choice, commute travel times, discrete choice model.

JEL code: R41, R48, C13.

1. Introduction

The present work aims to explore the relationship between travel cost and travel time and the probability of choosing a certain mode of transport in the city of São Paulo. Our hypothesis, as well as the theoretical and empirical evidence points that both correlations are negative: an increase in the cost/time of a given mode of transport is associated with a lower probability of choosing it and a higher probability of choosing competitor modes. Perhaps the true question is how strong this negative association is.

The literature whose geographic interest is also the city of São Paulo - comprised by Lucinda, Meyer and Ledo (2013), Barcellos (2014), Pacheco and Chagas (2015) – usually apply a mixed or random coefficients logit framework (in this model some variables have fixed coefficients and others might assume a normal distribution), using data from the 2007 Metrô Survey. There are several possible improvements in these works, for example, the

Metrô has already published a more recent survey in 2012 which updates all previous information brought by the previous survey. Also, these works have some methodology issues, for example, they employed the same model for work and education trips; also they failed to show how exactly they handled and constructed their final database. Besides these studies, we are only aware of another one for the city of São Paulo: Swait and Ben-Akiva (1987) which used data of 1985 Metrô's survey to calibrate a Parametrized Logit Captivity (PLC) mode, which is a generalization of the *dogit* model (GAUDRY, DAGENAIS, 1979).

International literature is increasingly focusing in the Generalized Extreme-Value (GEV) logit models, especially the Cross-Nested specification, leaving behind models which rely on the Independence of the Irrelevant Alternative (IIA) assumption, as the Multinomial and Conditional Logit models. Nesting is the practice of grouping similar alternatives in nests, which allow for some degree of dependency of alternatives, whereas cross-nesting is the practice of allowing the same alternative to belong to different nests. Small (1987) proposed an Ordered GEV model departure time, which is a continuous variable usually model as discrete intervals; however, he accounted for both the order and the dependency of the outcomes. Vovsha (1997) proposed a cross-nested logit model to study is Tel Aviv, Israel, by using both Stated Preference (SP) and Revealed Preference (RP) data to model automobile, bus and rail commute choices and their respective access and leg modes. Bierlaire, Axhausen and Abbay (2001) explore the stability of different model approaches to the SwissMetro transport data: multinomial logit, MNL with non-linear utility function, nested logit and cross-nested logit. The previous works also have some areas of improvement, for example, they failed to control for individual characteristics which can be highly correlated with the mode choice; besides that, they could better explain the structure of their database and how exactly they constructed their counterfactual trips (if it was necessary)..

Our study makes a few contributions to the mode choice literature. First, we use both publicly available transport surveys in the city of São Paulo: the 2007 Pesquisa Origem Destino and the 2012 Pesquisa de Mobilidade. Most studies use only one cross-section, which makes comparisons between two points in time impossible; and also, having at least two cross-sections allow us to cross-check the results by comparing them. Second, we conduct a thorough description and processing on our dataset, explaining how and why every single change, exclusion and creation was made, increasing the transparency of the present work and the reproducibility of the results. Third, we created a new method for calculating the travel cost for the private mode of transport which involves first estimating the yearly mileage from sources outside the Metrô data and also a new insight for estimating the counterfactual travel times by first finding each mode of transport speed. Lastly, we took advantage of the full potential of our dataset including, sociodemographic characteristics of the commuter, which is not so common in practice, see for instance the works of Bierlaire, Axhausen and Abbay (2001), Dissanayake and Morikawa (2010) and Washbrook, Haider and Jaccard (2006).

2. Methodology

The following discussion of the random utility maximization (RUM) model is based on Koppelman and Bhat (2006). The utility, U, is a function of the attributes of the alternatives and the characteristics of the individuals. If one alternative is chosen, it means that it has higher utility than others in the choice set. In other words, if alternative j is chosen

if and only if the utility of alternative j is greater than or equal to the utility of all other alternatives k, in the choice set, C:

$$If\ U(X_j,\ S_i) \geq U(X_k,\ S_i)\ \ \forall j \rightarrow j > k\ \ \ \forall k \in C \tag{1}$$

Given that, we can say that alternative j is preferred over other alternatives k, in the choice set C. The random utility states that the utility function of individual i and alternative j, U_{ij}, has both a deterministic or observable portion, V_{ij}, and a stochastic or random portion, ε_{ij}, such that:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{2}$$

Also, we can say that the probability P_{ij} that individual i chooses j is equal to the probability of U_{ij} being the largest of all $U_{ij}, \ldots, U_{iJ}$. With $y_i \in \{1 \ldots J\}$ denoting the alternative that decision maker i chooses, this probability is:

$$P_{ij} = Pr(y_i = j) = Pr(U_{ij} > U_{ik}\ \forall\ k = 1 \ldots J\ : k \neq j) = Pr(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - \tag{3}$$

Given the deterministic part of the function, the probability will depend upon the distribution of the stochastic errors. Equation (12) shows two interesting features of RUM probabilities: i) it is based on utility differences only, that is, the addition of a constant doesn't change the outcome probabilities; and ii) the scale of the utility is not identified, multiplying each utility by a constant doesn't change the probabilities, so RUM models must normalize the utilities.

The systematic (or deterministic) utility can also be divided into three parts: $V(S_i)$ associated with the characteristics of the individual i; $V(X_j)$ associated with the attributes of the alternative j; and $V(S_i, X_j)$ which results from an interaction between the attributes of alternative j and the characteristics of individual i:

$$V_{it} = V(S_i) + V(X_j) + V(S_i, X_j) \tag{4}$$

Considering the Conditional Logit (CL) – but the assumptions also applies to the Multinomial Logit (MNL) – it assumes that the error terms are independent and identically distributed (i.i.d.) as extreme value type I with a variance $\sigma^2 = \frac{\pi^2}{6}$. McFadden (1974) demonstrated that the P_{ij} that individual i chooses j is:

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} \tag{5}$$

In the CL, the errors are assumed i.i.d. and they capture all unobserved determinants. If two alternatives are similar, errors might be positively correlated, for example, in the case of a mode choice model, bus and train are both transit alternatives. If the assumption of i.i.d. errors is violated, it means that CL and MNL parameters are biased. Therefore, the CL model relies on a strong assumption: the Independence of Irrelevant Alternative (IIA).

3. Empirical Strategy and Data

The database used in the present research comes from various sources, but its core is composed by the microdata from the transport surveys done by the Companhia do Metropolitano de São Paulo[1] (Metrô) in 2007 and 2012. Besides that, we also collected monthly fuel prices from the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis[2]

[1] Company of the Metropolitan of São Paulo. Website: http://www.metro.sp.gov.br/
[2] National Agency of Petroleum, Natural Gas and Biofuels. Website: http://www.anp.gov.br/

(ANP) and yearly vehicle mileage from the Instituto Nacional de Metrologia, Qualidade e Tecnologia[3] (INMETRO). In this section, we will explain in detail each dataset, especially those from Metrô.

Both the 2007 and 2012 Metrô Surveys are very similar in their three-block structure. The first block of information is about the household, its durable goods and socio-economic variables of the residents. The second block informs the characteristics and localization of the school, first and second jobs. The third has information about individual trips: which weekday; coordinates and traffic zone of origin, first, second and third transferences and destination; purpose on origin/destination; departure and arrival time (hour and minute), mode choice, time walking at origin/destination, length and distance. Until four trips are recorded per person. It is answered by everyone who took at least one trip the previous day.

We collected monthly gasoline information for each Brazilian State from the ANP (ANP, 2012). We merged this information into the Metrô's dataset according to the year and month variables presented in both sources. We also collected data about the energetic efficiency of light motor vehicles from the INMETRO. There is information about the vehicle category, brand/company, model, version, engine, speed transmission, air conditioning, assisted steering, fuel, mileage (divided between urban and rural cycle and hydrous ethanol and gasoline) and the Programa Brasileiro de Etiquetagem (PBE) fuel efficiency classification.

After pre-processing the data (removing missing values and logical inconsistencies), our initial task was estimate the vehicle's mileage gasoline consumption. Both 2007 and 2012 Metrô's Surveys present the number of automobiles *per* household as well as the year of manufacture of three of them. We could use the INMETRO's average mileage to merge into the Metrô's surveys; however, there are many automobiles in the Metrô's dataset which were manufactured prior 2009, as we can see from the following table:

Table 1: Year of manufacture of the first automobile in 2007 and 2012 Metrô's Surveys

Year	Frequency	Percent	Cumulated
Prior 2009	17,321	90.89	90.89
2009	264	1.39	92.28
2010	493	2.59	94.86
2011	480	2.52	97.38
2012	445	2.34	99.72
2013	54	0.28	100
Total	19,057	100	

Source: elaborated by the author

Our method to overcome this data gap (mileage prior 2009) relies on one reasonable assumption: each passing year, vehicles become more efficient in terms of mileage (or for our purpose we can think backwards: each previous year automobiles were less efficient in terms of mileage). This is assumption seems to hold for the INMETRO dataset as we can see from the following bar graphic showing average mileage for each year. We can argue that this assumption might hold in the "real world" if we think that as time passes, technology improves, and as technology improves, new fuel efficient engines are created, car body

[3] National Institute of Metrology, Quality and Technology. Website: http://www.inmetro.gov.br/

becomes lighter and other aspects are improved, implicitly "time is making mileage increases".

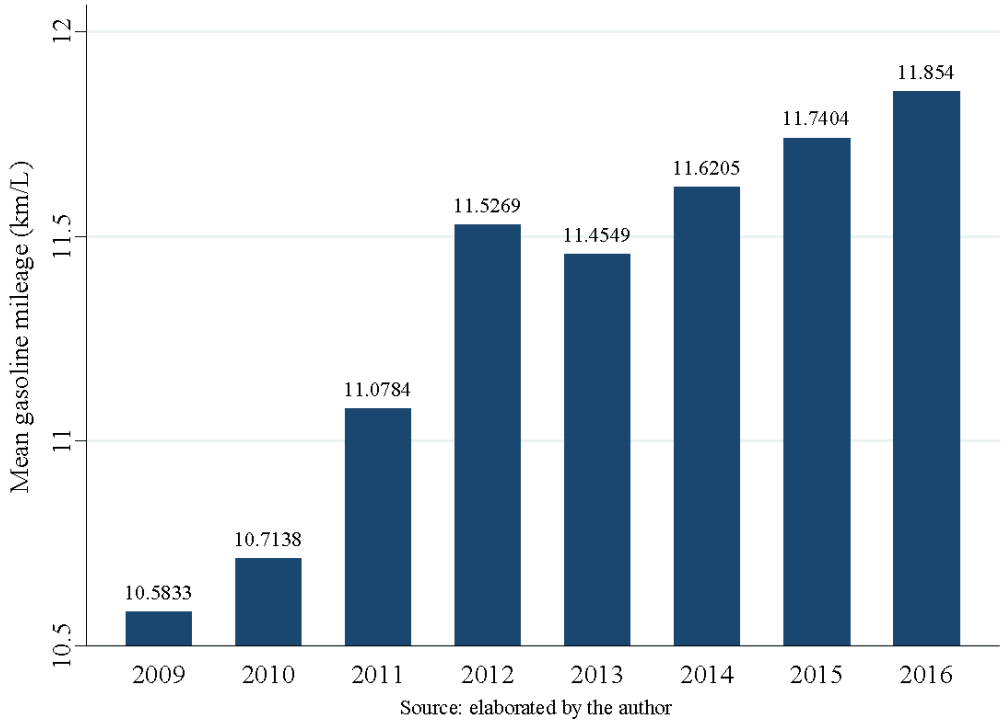


Figure 1: Mean gasoline mileage over year (2009-2016), sub-compact and compact vehicle categories

Therefore, our objective is to estimate if and how much time affects automobile mileage. To fulfill our objective, we ran a dummy variable fixed effects model with a deterministic trend. We also included two other variables besides these those. According to the INMETRO (2009-2016) methodology, the reference values of models' mileage[4] were tested with the optional air conditioning and assisted steering, as indicated. So, we included in the model a dummy variable for the presence of air conditioner (with value equal to "1" indicating the presence of air conditioner and "0" otherwise) and dummy variables for each type of assisted steering (electric, electrohydraulic, hydraulic and mechanic). We grouped the individual variable by brand/company and model (totalizing 61 automobiles/individuals). We are primarily interested in the coefficient of the deterministic trend which tells us how much the mileage increases with each passing year.

Table 2: Estimating the effect of time on vehicular gasoline consumption

VARIABLES	Gasoline mileage (city)

[4] According to the INMETRO methodology (INMETRO, 2017), the reference values are obtained from consumption measures performed in laboratory according to the NBR 7024 standards. To approximate the its values to those perceived by actual drivers, the INMETRO adopted the same adjustment factors as the United States Environment Protection Agency.

Deterministic trend	0.1367***
	(0.01566)
Air conditioning	-0.8113***
	(0.08282)
Assisted steering:	
electrohydraulic	0.6406**
	(0.2769)
hydraulic	-0.6359***
	(0.2089)
mechanic	-0.2623
	(0.1866)
Constant	-263.55***
	(31.577)
Observations	822
R-squared	0.710
Fixed effects	Yes

Assisted steering base category: electric

t statistics in parentheses
Source: elaborated by the author
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

As we can see from Table 2, the variables were significant (except one of the assisted steering dummies). In particular, the deterministic trend was significant at the 1% level and we can interpret it as time increases by one year, gasoline mileage increases, in average, by 0.1367 km/l, holding other variables constant. Using this estimate, we can now fill in the missing values for the mileage of cars manufactured prior 2009. For the manufacture years prior 1970, we treated mileage as being equal to the year of 1970, which is 5.5658 km/l, otherwise values would be nearly zero for the oldest manufactured automobiles (1920 is the oldest automobile in the dataset). Also, there were 94 cases where the year of the first automobile is missing, even though the household had at least one item, so we replaced those missing observations by the sample median of the year of the manufacture of the first automobile, which was 2002 for the 2007 Survey and 2007 for the 2012 Survey. We also used the automobile median to generate the year of manufacture of the motorcycles and twice the respective auto mileage.

Our next task is to estimate the counterfactual travel times of the dataset. The structure of the dataset that we currently have is one observation (one row) for each individual, that is, we have information about the observed or actual choice; however, we need to have or to estimate the counterfactual scenarios, that is, what would have happen had the individual chosen a different mode to commute.

Our assumption is that people remember more accurately their origin and destination points (since during the interview, interviewees only need to answer the address of each location and can cite references, like gas stations, supermarkets, based on that information the data process stage will fill the latitude and longitude coordinates, which decreases recalling error) than at what time they departure and arrived at such locations. Given that, instead of

trying to estimate directly the travel times, our strategy is to estimate the average speed - the ratio between distance and time – regressing each mode's speed on individual characteristics.

There isn't any literature regarding specifically the prediction of transit speeds and there is no demographic variable that can explain it *per se*. However, we can think about the transit speed as the sum of two speeds: the speed of the individual while walking to/from a station and the speed of the vehicle itself. For the "first" speed, we included as explanatory variables age and sex variable as in the non-motorized speed regression (latter explained). For the "second" speed, we included variables that could indicate the effect of time and space over the vehicle speed: the day of the week (Weekday dummy variables, equal to "3" if the trip occurs on a Tuesday, "4" if the trip occurs on a Wednesday and so on, the base category is Monday), the departure hour (ranging from "0" to "23", since early departure should avoid the rush hour and supply of vehicles might also change depending on the time of the day) and dummy variables for each origin TAZ. Since the latter variable is different for the 2007 and the 2012 Surveys, we estimated a different regression for each survey. Table 3 displays the regression of transit speed on explanatory variables for the 2007 and 2012 Surveys.

Table 3: Estimating average transit speed for each Metrô Survey

	2007 Survey	2012 Survey
Age	-0.0060	-0.0095
	(0.0047)	(0.0066)
Sex	0.1784	0.3315*
	(0.1164)	(0.1578)
Monday	0.0000	0.0000
	(.)	(.)
Tuesday	0.1569	-0.0759
	(0.2324)	(0.3311)
Wednesday	0.1584	0.1623
	(0.2319)	(0.3352)
Thursday	0.1149	0.2017
	(0.2178)	(0.3175)
Friday	0.2761	0.1197
	(0.1942)	(0.2794)
Departure hour	0.0853***	0.0458
	(0.0191)	(0.0257)
_cons	5.9745***	6.8457***
	(0.8624)	(0.5633)
N	8,209	2,586
R^2	0.2012	0.1299
TAZ_effect	Yes	Yes

Standard errors in parentheses

Source: elaborated by the author

$^{*}\ p < 0.05,\ ^{**}\ p < 0.01,\ ^{***}\ p < 0.001$

There also isn't any literature regarding specifically the prediction of private motorized speeds and there is no demographic variable that can explain it *per se*. There are, however, some papers on driving behavior and individual characteristics, especially sex and gender. For example, Rhodes and Pivik (2011) applied a phone survey in the state of Alabama on 504 teens and 409 adults showed a riskier driving behavior for teen and male than for adult and female drivers. Özkan and Lajunen (2006) investigated the differences between sex and gender on driving skills and accident involvement among young drivers. Reasoning as we did on the transit speed regression, the private motorized speed can be described as the sum of two factors: the behavior of the individual driver and speed of the vehicle itself. Regarding the "first" speed, we included socio-demographic characteristics: age, sex, number of family dwellers, family income (nominal R$) and employment relationship (base category is formal contract). For the "second" speed", we used the day of the week, the departure hour and dummy variable for each origin TAZ. Table 4 displays the results of the regression of transit speed on the previous explanatory variables for the 2007 and 2012 Surveys. All variables were individually significant and overall goodness of fit was higher than expected.

Non-motorized speed has a more robust literature body, being that the medical research area has special interest in the field. For example, Murray *et al.* (1966) employed a photographic method for recording simultaneously the displacements which occur in walking, studying the gait pattern of sixty men. The patterns for fast speed walking had a mean of 7.848 km/h and a standard deviation of 0.9, while free speed had a mean of 5.436 km/h and a standard deviation of 0.72. Also, younger men (20-25 years old) walked significantly faster than the older men, taking the longest strides in the shortest time, while older men (60-65 years old) tended to take the shortest strides. Height also influenced the stride length, showing the greatest magnitude for the tall subjects and the least for the short subjects. All of the displacement patterns, except the stride width and the foot angles, were notably similar for repeated walking trials of the individual subjects. Murray, Kory and Clarkson (1969) continue the previous study by extending the upper limits of the age range to 87 to test whether a "presenile" walking pattern is consistent and progressive with advanced age. Sixty-four "normal" men (normal strength and range of motion) in age groups from 20 to 87 years were recorded by interrupted-light photography. Free speed mean was 5.004 km/h with a standard deviation of 0.828 whereas fast speed mean was 7.02 km/h with a standard deviation of 1.44. Again, walking speed of the men in the three oldest age groups was significantly lower than that of the younger men for both their free and fast speed walking trials.

Table 4: Estimating average private speed for each Metrô Survey

	2007 Survey	2012 Survey
Age	-0.0363***	-0.0031
	(0.0086)	(0.0134)
Sex	0.8846***	1.1292***
	(0.2139)	(0.3366)
Family_dwellers	-0.0999	0.0395
	(0.0803)	(0.1248)
Family_income	0.0000	-0.0000
	(0.0000)	(0.0000)
Formal contract	0.0000	0.0000

	(.)	(.)
Informal contract	-0.6185	-0.5244
	(0.4645)	(0.7479)
Public agent	0.2705	-1.1646*
	(0.3981)	(0.5677)
Self employed	-0.2930	-0.8554
	(0.2910)	(0.4613)
Employer	-0.8903*	-0.0429
	(0.3614)	(0.6776)
Independent professional	-0.3116	-0.9619
	(0.3532)	(0.5900)
Family business employer	-1.5244**	-0.4360
	(0.4797)	(0.8819)
Family business employee	-1.4148	2.2747
	(0.8117)	(2.1843)
Monday	0.0000	0.0000
	(.)	(.)
Tuesday	0.0308	1.7123**
	(0.3995)	(0.6237)
Wednesday	0.1668	0.8811
	(0.3713)	(0.6422)
Thursday	-0.0606	-0.5484
	(0.3739)	(0.5711)
Friday	0.2158	0.9009
	(0.3008)	(0.5103)
Departure_hour	-0.0974**	-0.0450
	(0.0340)	(0.0498)
Constant	6.4358***	9.4912***
	(1.0280)	(1.1275)
Observations	8,394	1,873
R^2	0.1408	0.0862
TAZ_effect	Yes	Yes

Standard errors in parentheses

Source: elaborated by the author

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Given the previous works, age, sex and weight are the standard variables for explaining walking speed. Since we did not reject the hypothesis that cycling and walking had different mean speeds and cycling holds a small mode share, we treated cycling and walking as the same mode. Since we do not observe the individual's weight, we regressed the average non-motorized speed on age (years) and sex ("1" for male and "0" for females). We found a negative relationship between speed and age and males had a higher speed than females. Both individual t-tests rejected the hypothesis that the coefficients are equal to zero at 1% significance, although the coefficient of determination (the r-squared) resulted smaller than we expected.

Table 5: Estimating average non-motorized speed for both Surveys

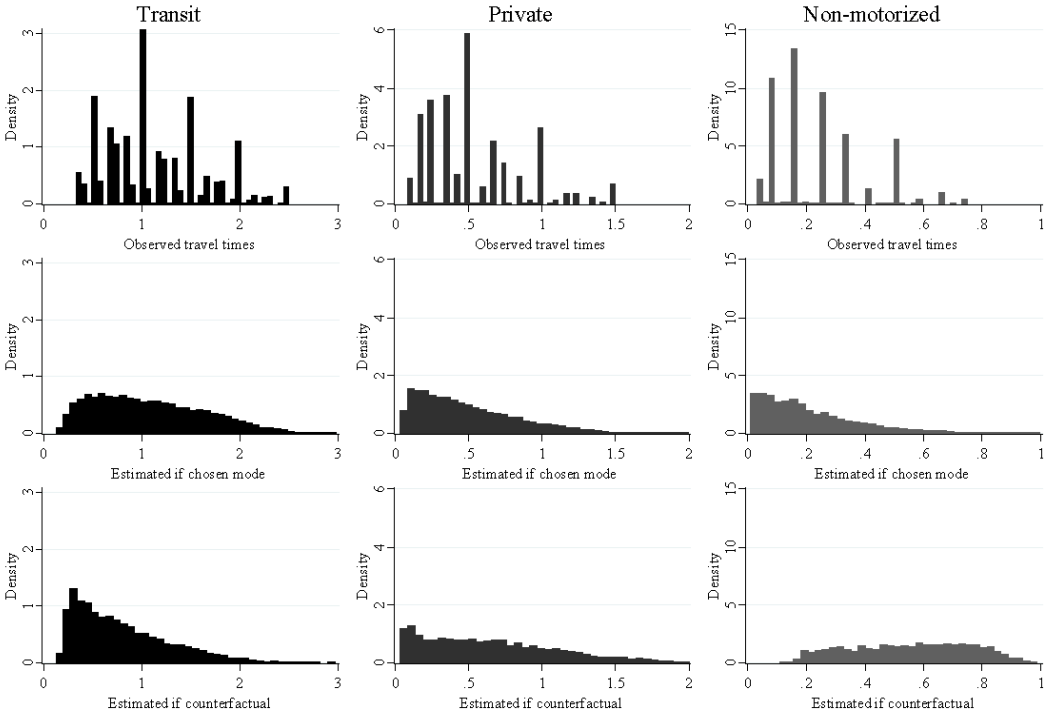
	Speed
Age	-0.0112**
	(0.0037)
Sex	0.3063**
	(0.1045)
Constant	3.4854***
	(0.1554)
Observations	5,736
R^2	0.0073

Standard errors in parentheses

Source: elaborated by the author

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Figure 2 displays the observed and estimated travel times: the first column shows transit travel times; the second, private; and the third, the non-motorized. The first row shows the observed travel times (what we had in our "original" dataset); the second row shows the travel times for those same modes as they were estimated by our models; and the third row shows the travel times for the counterfactuals, that is, the modes that weren't actually chosen by the commuters and that we also estimated. We can perceive in the first row how observed travel times show large spikes around sharp times and has small mass between those spikes. The second column shows more even distributions, however, they all share a trend to be right/positive skewed, maybe that is a sign that our models overestimated the speeds, resulting in lower travel times. This trend is even more evident in the third row for transit and private modes, which were more right-skewed: this may be due to the fact that those low travel times are the counterfactuals of non-motorized trips, which tend to be shorter, and since transit and private display higher speeds, this (dividing a shorter distance by a higher speed) result in lower travel times. Non-motorized column in the third row, on the contrary, seem to be a little left-skewed: since it is the counterfactual of the other modes and it has a lower speed, it results in higher travel times. We will continue to deal with our counterfactual choice set generating as we move on to the cost estimating.



Source: elaborated by the author

Figure 2: Observed and estimated sample distribution of travel times by mode

Our next step was creating the travel cost variable. We did it based on the inherent characteristics of some modes as well as some information given by the Metrô dataset and external sources. The next table presents in the first column the mode of transportation, then the next two columns our calculation for each one of the Metrô's Surveys and the sources.

Table 6: Travel cost calculation by mode of transportation for 2007/2008 and 2012 (R$)

Mode	**2007/2008**	**2012**
Bus from the São Paulo municipality	2.3 (SPTRANS, 2017)	3 (SPTRANS, 2017)
Chartered bus	2.3	3
Driving automobile	Equation (6)	Equation (6)
Passenger of automobile	Equation (6)	Equation (6)
Taxi	3.5+(2.1*Distance) (SÃO PAULO, 2006)	4.1+(2.5*Distance) (SÃO PAULO, 2010)
Microbus from the São Paulo municipality	2.3	3
Subway (tube)	2.3 (SPTRANS, 2017)	3 (SPTRANS, 2017)
Train (rail)	2.3 (SPTRANS, 2017)	3 (SPTRANS, 2017)
Motorcycle	[Equation (6)]/2	[Equation (6)]/2
Bicycle	0	0
Walking	0	0

Source: elaborated by the author

$$Auto\ cost = \left(\frac{distance(km)}{mileage(km/l)}\right) \times p_g(R\$/l) \quad (6)$$

Next two tables provide the summary statistics for cost and time variables for the 2007 and 2012 Surveys, respectively. Sample mean of both cost and time is higher for transit than private; and higher for private than cost, for both Surveys. Contrary to most works, where there is a clear trade-off between cost and time: faster commutes usually result in higher costs, in our dataset transit is both more time consuming and more expensive than private modes. There are some possible explanations for this phenomenon: i) a direct effect is the government subsidy for the gasoline through its monopolistic company Petróleo Brasileiro S. A. (Petrobras); ii) and an indirect effect being the Downs-Thomson paradox taken place in the practice, as the private vehicle real prices are actually declining due to heavy subsidizing, it shifts people from public transportation towards private transportation, increasing the fare as the system cost is beared by less users; and iii) the way we created the private cost, we only took into consideration the gasoline costs, not the other various direct and indirect cost of both owning and using a private motorized mode. We can also note that, despite non-motorized modes being in practice slower (in terms of speed) than both transit and private, their travel times is, on average, lower than the other two modes. This can only be explained by the fact the distance commuted by non-motorized modes are smaller, that is, people live near their work or work near where they live. As most works, the standard deviation of transit travel times was higher than private travel times for both Surveys: a simple interpretation might be that transit is a less reliable mode than private and non-motorized, which has a practical explanation, since the transit user does not control the schedule and the route of this mode. Lastly, we can note that, although we created a sample around 15-20% the size of the original dataset for each Survey, the sample still account for almost 6 million trips for the 2007 Survey and 6.7 million trips for 2012.

Table 7: Summary statistics of Cost (R$) and Time (h) variables for the 2007 Survey

	Summary of Cost		**Summary of Time**			
Mode	Mean	Std. Dev.	Mean	Std. Dev.	Freq.	Weighted freq.
Transit	2.16	0.38	0.94	0.55	15,592	1,769,072
Private	1.79	2.92	0.55	0.40	13,658	1,091,313
Non-motorized	0.00	0.00	0.36	0.25	7,830	717,377
Total	1.57	1.97	0.68	0.50	37,080	3,577,762

Source: elaborated by the author

Table 8: Summary statistics of Cost (R$) and Time (h) variables for the 2012 Survey

	Summary of Cost		**Summary of Time**			
Mode	Mean	Std. Dev.	Mean	Std. Dev.	Freq.	Weighted freq.
Transit	2.82	0.48	1.03	0.59	4,196	1,997,714
Private	2.23	3.67	0.64	0.47	3,462	1,224,060
Non-motorized	0.00	0.00	0.34	0.25	1,911	837,060
Total	2.04	2.47	0.68	0.50	9,569	4,058,834

Source: elaborated by the author

First of all, let us define our dependent variable, the choice of mode of transportation. The respondents in the Metrô's Surveys may choose among an exhaustive choice set of seventeen options considered by the questionnaire: bus from the São Paulo municipality, bus from other municipalities, intercity bus, chartered bus, school vehicle, driving a car, passenger of a car, taxi (cab), microbus from the São Paulo municipality, microbus from other municipality, intercity microbus, subway (tube), train (rail), motorcycle, bicycle, walking and others, as shown in the first column of the next table. Our mode aggregation considers three outcomes: the first one we shall call "transit", the second "private motorized" (or simply private) and the third one "non-motorized"; from the sample point of view, it divides in roughly "40-40-20" percent.

Table 9 specifies the independent variables that compose the deterministic portion of the utility function. As already mentioned, travel time and travel cost are the two most consecrated variables that determine home-work mode choice. Since there might be other unobserved variables that can affect mode choice and might be correlated with travel time and travel cost, we also included socio-economic characteristics of the individual as control variables: age, gender, whether he studies and what is the employment relationship. A great difference, and also a possible disadvantage, of our work compared to the existing literature is that we did not include alternative specific dummies, not because of some methodological motive, but because of practical reasons: the dataset structure does not allow for this kind of variable since there is "only" information about the chosen outcome and nothing about the unselected alternatives.

Table 9: Explanatory variables - code, variable, description and unit of measure

Code	Variable	Description	Unit of measure
TT	Travel time	Total travel time between origin and destination	Hours
TC	Travel cost	Total travel (variable) cost between origin and destination	R$
Age	Age	Age of the individual	Years
Sex	Sex	Sex of the individual	1-male, 0-female
Study	Study	Whether the person is currently studying	1-yes, 0-otherwise
Employ	Employment relationship	Dummy variable for each employment relationship (employee with formal contract; employee without formal contract; public servant; self employed; employer; liberal professional; family business owner; and family worker)	1-if person works as *eth* employment relationship; 0-otherwise
Degree	Education degree	Dummy variable for each education degree (illiterate/incomplete elementary school, complete elementary/incomplete secondary school, complete secondary/incomplete high school, complete high/incomplete college, complete college	1-if person works as *dth* education degree; 0-otherwise

Source: elaborated by the author

4. Results

The next table presents the Alternative-specific Conditional Logit model – which is also known as mixed logit - for the 2007 and 2012 Survey. In the "usual" Conditional Logit, the explanatory variables are the same for all outcomes (there's only one coefficient for all alternatives), the Alternative-specific Conditional Logit is a sub-specification of this mode, for, in practice, the case variables are interacted with J-1 outcome dummy-variables and the J*th* outcome is the base outcome, as in the Multinomial Logit. The alternative-specific variables are travel cost and travel time and the characteristics of the choice maker used in the previous model are the case variables. The first column shows the explanatory variables; the next three columns show the estimates for the 2007 Survey; and the last three columns show the estimates for the 2012 Survey.

The alternative-specific variables are also called "generic" variables, since they are associated with all outcomes equally, that is, there is one estimate for all alternatives. The ratio between the cost and time coefficients is known in the literature as the Value of Travel Time savings (VoTT), which is the amount of money people are willing to spend to save one hour of their commute or it is simply a measure of the value time. According to Cameron and Trivedi (2011), a negative sign of an alternative-specific variable means its own-effect is negative and the cross-effect is positive, that is, a negative coefficient means that an increase on that variable decreases the probability of choosing that alternative and increases the probability of choosing the other alternatives. Both travel time and cost variables have a negative sign, which is in accordance with our expectations and also is the usual in the literature. Other variables should be understood as in the previous Multinomial Logit case. The generic variables were all individually significant, except the travel time in 2012, other variables were significant in one Survey and not in the other or significant for transit and not for private mode, for example, sex was significant in all cases, except for transit in the 2012 Survey. One final note is that some cases were dropped in each Survey, because these cases had only one alternative in their choice set: the observed outcome and their counterfactuals were excluded because they weren't feasible choices.

Table 10: Alternative-specific Conditional Logit for the 2007 and 2012 Surveys

	2007 Survey			2012 Survey		
	Generic	Transit	Private	Generic	Transit	Private
Cost	-0.483***			-0.325***		
	(-7.43)			(-4.11)		
Time	-0.861***			-0.354		
	(-5.00)			(-1.13)		
Age		0.010	0.028***		-0.018*	0.001
		(1.79)	(5.28)		(-2.39)	(0.09)
Sex		-0.350**	0.618***		0.064	0.974**
						*
		(-2.79)	(5.05)		(0.35)	(5.49)
Study		-0.268	0.034		-0.456	-0.270
		(-1.29)	(0.18)		(-1.46)	(-0.96)
Informal contract		-0.341	0.078		0.004	0.180
		(-1.89)	(0.46)		(0.01)	(0.64)

Retired		-0.153	0.154	0.423	0.664
		(-0.61)	(0.63)	(1.20)	(1.93)
Sick leave		-0.590***	0.668***	-0.405	0.832**
		(-3.29)	(3.87)	(-1.61)	(3.55)
No work		-1.060**	1.392***	-1.850*	0.675
		(-2.91)	(4.90)	(-2.66)	(1.74)
Never worked		0.124	0.980**	0.912	1.363*
		(0.34)	(2.93)	(1.51)	(2.34)
Housewife		-0.328	1.646***	-0.091	2.095*
		(-0.74)	(4.52)	(-0.10)	(2.43)
Student		-0.299	2.113***	-3.056*	-0.050
		(-0.41)	(3.57)	(-2.33)	(-0.05)
Elementary school		-0.031	0.521	-0.190	-0.076
		(-0.09)	(1.49)	(-0.37)	(-0.15)
Secondary school		-0.235	0.549	-0.593	-0.158
		(-0.68)	(1.60)	(-1.17)	(-0.32)
High school		0.267	1.106***	-0.196	0.451
		(0.82)	(3.41)	(-0.40)	(0.96)
College		0.191	1.965***	-0.540	1.056*
		(0.57)	(5.94)	(-1.08)	(2.16)
_cons		0.772	-2.189***	2.409**	-0.319
		(1.93)	(-5.64)	(3.88)	(-0.55)
Observations	30,095		7,545		
Cases	14,202		3,587		
Cases dropped	6,985		2,024		
VoTT	1.78		1.09		

t statistics in parentheses

Source: elaborated by the author

Employment relationship base category: Formal contract

Education degree base category: Incomplete elementary school

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We can see that the magnitude of both the cost and time variables decreased from 2007 to 2012, especially the latter. This change also decreased the Value of Travel Time Savings from R$ 1.78 to R$ 1.09. This decrease means that commuters were willing to pay in 2012 nearly half as much as commuters in 2007 for faster travels. The 2007 figure is in accordance with other works, such as Barcellos (2014), but is underestimated compared to Lucinda, Meyer and Ledo (2013) and Pacheco and Chagas (2016), which found a VoTT of R$ 8.58 and R$ 6.88, respectively. The discrepancy in the figures in the three previously cited works is surprising since all of them have almost identical methodologies. Besides our own assumption and model framework errors, we could imagine one major reason why the VoTT

of São Paulo is so low: given that São Paulo's transit is simultaneously slower, costlier and more inflexible (both in schedule and in departure time) compared to the private modes, it makes sense that the São Paulo's citizens value so little their travel time savings, especially in view of its high average hourly-wage compared to the rest of Brazil. Usually, what happens in most cities is a clear trade-off between travel cost and time: slower modes of transport are cheaper (e.g. transit) and faster ones are also more expensive (e.g. automobile), then commuters are willing to pay more for faster commutes. The VoTT is also a strong indicator for the public policy: a low number reduces extremely the return over infrastructure investment, since even a major speed gain resulted from a road expansion, for example, would be lowly valued by the commuters (at the extreme a VoTT equal to 0 would mean that any investment cost would result in no benefit).

As we did in the Multinomial Logit, after the regression procedure we estimated the Average Marginal Effect for each Survey. However, this time, the table looks a little different: we didn't show the cases (person or commuters) variables, only the alternative-specific variables cost and time. The first column shows two blocks, one for each variable, cost and time, and each block is divided by mode of transportation, that is, the first line refers to the AME of the transit cost. The next three columns refer to the 2007 Survey and the last three, to the 2012 Survey. Each Survey has three columns, one for the probability of choosing each one of the commute modes. For example, the first coefficient (first row second column), -0.091, shows the AME of the transit cost on the probability of choosing transit as a commute mode; the next coefficient on the side (first line third column), 0.034, shows the AME of the transit cost on the probability of choosing private as a commute mode, both for the 2007 Survey. As we said in the previous paragraph, the negative sign of both cost and time coefficients implied that its own-effect is negative and the cross-effect is positive. For example, in the first line, a one-unit (R$ 1) increase in the transit cost is associated with a decrease on the probability of choosing transit by 9.1 percentage points (pp.), an increase on the probability of choosing private by 3.4 pp. and an increase on the probability of choosing a non-motorized mode by 5.7 pp. All average marginal effects were individually statistically significant, except travel time for the 2012 Survey.

Table 11: Alternative-specific Conditional Logit models' average marginal effects

	2007 Survey			2012 Survey		
Cost	Transit	Private	Non-motorized	Transit	Private	Non-motorized
Transit	-0.091***	0.034***	0.057***	-0.073***	0.040***	0.032**
	(-10.15)	(11.61)	(6.43)	(-4.70)	(7.23)	(2.86)
Private	0.034***	-0.097***	0.063***	0.040***	-0.075***	0.034**
	(11.61)	(-11.14)	(7.19)	(7.23)	(-4.70)	(2.88)
Non-motorized	0.057***	0.063	-0.120***	0.032**	0.034**	-0.067**
	(6.43)	(7.19)	(-6.85)	(2.86)	(2.88)	(-2.88)
Time						
Transit	-0.163***	0.061***	0.101***	-0.079	0.044	0.035
	(-6.48)	(11.72)	(4.62)	(-1.18)	(1.36)	(1.01)
Private	0.061***	-0.174***	0.112***	0.044	-0.081	0.037
	(11.72)	(-6.48)	(4.78)	(1.36)	(-1.17)	(1.00)
Non-motorized	0.101***	0.112***	-0.214***	0.035	0.037	-0.073

(4.62)	(4.78)	(-4.72)	(1.01)	(1.00)	(-1.00)

z statistics in parentheses

Source: elaborated by the author

$^{*}\ p < 0.05,\ ^{**}\ p < 0.01,\ ^{***}\ p < 0.001$

From this table, we can infer how sensitive commute choice is to cost and time changes and how policy interventions would impact the commute mode shares. For example, an increase in the transit cost would favor the choice of non-motorized over private in 2007 and private over non-motorized in 2012. In 2007, non-motorized cost and time own-effects are higher than the other modes, that is, it is more sensitive; on the other hand, in 2012, the three modes are more or less equally sensitive in its own-effect, with non-motorized having a slightly lower own-effect. A policy to incentive transit ridership when subsidizing the transit fare by R$ 1 would increase the probability of transit ridership by 7.3 pp. and decrease the private commute by 4 pp and non-motorized by 3.2 pp. in 2012. A policy to inhibit private commute when taxing gasoline by R$ 1 would decrease the probability of choosing private by 9.7 pp. and increase the probability of transit ridership by 3.4 pp. and non-motorized by 6.3 pp. in 2007. A policy to decrease transit travel times, for example, the creation of exclusive bus lanes, by 6 min (or 0.1 hours in our measure) would increase the probability of transit ridership by 0.79 pp. and decrease the private probability by 0.44 pp. and non-motorized by 0.35 pp. Another possible policy could be to incentive bicycle ridership to work. Like what France and United Stated do, Brazil could incentive bicycle and walk commuting by subsidizing it by the amount of the transit fare, R$ 3.2 in 2012, which would, on average, increase the probability of non-motorizing commuting by 21.44 pp. and decrease transit by 10.24 and private by 10.88 pp.

5. Discussion and final remarks

This research attempted to measure the correlation between economic and demographics variables and the probability of choosing a certain transport commute mode. Our main objective was to measure the correlation between travel time and cost and the probability of choosing a certain mode of transport. Our main hypothesis is that this correlation is negative for all modes of transport. The core dataset is composed of two of the most recent transportation surveys carried out by the Metrô company, but we also collected gasoline prices from the ANP and yearly vehicle mileage from the INMETRO. We begin our strategy by showing the estimation of the counterfactual travel times, that is, the travel times of the modes not taken (not observed in the dataset). Then, we estimated the observed and counterfactual travel costs. Finally, we made explicit our model specification for both the dependent and the independent variables.

We further specify a Conditional Logit model using socio-economicand also generic variables for travel time and cost, as expected they turned out to be negatively correlated with the probability of choosing any commute mode. The Value of Travel Time Savings (VoTT), that is the ratio between the time and cost coefficients, we found was fairly low compared to other studies, this means that, among other things, public investment in transportation infrastructure would lead to low benefit to commuters. We also estimated the average marginal effects (AVE) for the cost and time variables, with this information we can precisely

know how this variables are correlated with the probability of choosing a certain commute mode. For example, were we interested in discouraging private commuting by raising a gasoline tax by R$ 1/L, we would expect that in 2012 the probability of choosing private mode decrease by 7.5 p.p. and increase in 4 p.p. and 3.4 p.p. for transit and non-motorized, respectively.

In view of the previous results, we fully accept our initial hypothesis that travel cost and time is negatively associated with the probability of choosing a certain commute mode, and also that its own-effect is negative and the cross-effect is positive, for both variables. We also showed that the demographic characteristics – age, sex, study, employment relationship and education degree - were also highly and significantly associated with the probability of choosing a certain commute mode.

The present work can serve as basis for further research in the transport mode choice, especially in the Brazilian context. However, we are fully aware that the present work can be improved in many aspects. For example, instead of estimating travel time counterfactuals using econometric techniques, there are more appropriate methodologies specifically for this purpose, micro-simulation is one such case, as implemented by softwares like TransCAD, Vissim, MATSim, Quadstone Paramics, TRANUS and others. Also, past data could be gathered for São Paulo to see changes in time, because the Metrô's survey was first made in 1960; however, public data is only available for 2007 and 2012 at the moment. We could expand the mode choice analysis for other Brazilian cities that had at least one origin-destination survey, for example, one such city that available publicly and freely its data is Belo Horizonte (for 2002 and 2012).

References

ANP. (2012). Série histórica do levantamento de preços e de margens de comercialização de combustíveis - Série histórica mensal. Available at: <http://www.anp.gov.br/wwwanp/Precos/Mensal2001-2012/Estados.xlsx>. Accessed in: 12/12/2016

BARCELLOS, T. M. (2014). Não são só 20 centavos: efeitos sobre o tráfego da Região Metropolitana de São Paulo devido a redução na tarifa de ônibus financiada pelo aumento da CIDE nos combustíveis da cidade de São Paulo. 73f. Dissertação (Mestrado em Economia Aplicada) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, Ribeirão Preto.

CAMERON, A. C.; TRIVEDI, P. K. (2009). Microeconometrics using Stata. Stata Press.

DISSANAYAKE, D.; MORIKAWA, T. (2010). Investigating household vehicle ownership, mode choice and trip sharing decisions using a combined revealed preference/stated preference Nested Logit model: case study in Bangkok Metropolitan Region. Journal of Transport Geography. Volume 18 (3). Pages 402-410.

GAUDRY, M. J. I.; DAGENAIS, M. G. (1979). The dogit model. Transportation Research. Volume 13 (2). Pages 105-111.

INMETRO. (2017). Metodologia para divulgação de dados de consumo veicular. Available at: <www.inmetro.gov.br/consumidor/pbe/Metodologia_Consumo_Veicular.pdf>. Accessed in: 28/06/2017.

INMETRO. (2009). Tabelas PBE veicular – veículos leves 2009-2016. Available at: <http://www.inmetro.gov.br/consumidor/pbe>. Accessed in: 19/12/2016.

KOPPELMAN, F. M.; BHAT, C. (2016). A self instructing course in mode choice modelling: multinomial and nested logit models. Available at: <http://www.caee.utexas.edu/prof/bhat/COURSES/LM_Draft_060131Final-060630.pdf>. Accessed in: 17/06/2016.

LUCINDA, C. R.; MEYER, L. G.; LEDO, B. A. (2013). Urban road tax in a large emerging market: some Brazilian evidence. In: Encontro Brasileiro de Econometria, 35. Foz do Iguaçu, Paraná, Brasil. Anais… Foz do Iguaçu: Sociedade Brasileira de Econometria.

MCFADDEN, D. (1974). Conditional logit analysis of qualitative choice behavior. In: Frontiers in Econometrics, Zarembka, P. (ed.) (NY: Academic Press): 105.

METRÔ. (2012a). Pesquisa de Mobilidade da Região Metropolitana de São Paulo. Available at: <http://www.metro.sp.gov.br/metro/arquivos/mobilidade-2012/banco-de-dados/Dbase.zip>. Accessed in: 05/12/2015.

METRÔ. (2012b). Pesquisa de Mobilidade da Região Metropolitana de São Paulo – manual da pesquisa domiciliar. Available at: <http://www.metro.sp.gov.br/metro/arquivos/mobilidade-2012/manuais/manual codificador domiciliar_2012.pdf>. Accessed in: 05/12/2015.

METRÔ. (2012c). Zoneamento da Pesquisa de Mobilidade 2012 (em formato jpg). Available at: <http://www.metro.sp.gov.br/metro/arquivos/mobilidade-2012/mapas/ZonasMobilidade2012.JPG>. Accessed in: 05/12/2015.

METRÔ. Pesquisa Origem e Destino. (2007a). Available at: http://www.metro.sp.gov.br/metro/arquivos/OD2007/dbase.zip. Accessed in: 05/12/2015.

METRÔ. (2007b) Pesquisa Origem e Destino 2007 – manual da pesquisa domiciliar. 2007b. Available at: <http://www.metro.sp.gov.br/metro/arquivos/OD2007/manual-domiciliar-2007.pdf>. Accessed in: 05/12/2015.

MURRAY, M. P.; *et al*. (1966). Comparison of free and fast speed walking patterns of normal men. American Journal of Physical Medicine and Rehabilitation. Volume 45 (1). Pages 8-24.

MURRAY, M. P.; KORY, R. C.; CLARKSON, B. H. (1969). Walking patterns in healthy old men. Journal of Gerontology. Volume 24 (2). Pages 169-178.

ÖZKAN, T.; LAJUNEN, T. (2006). What causes the differences in driving between young men and women? The effects of gender roles and sex on young drivers' driving behaviour and self-assessment of skills. Transportation Research Part F: Traffic Psychology and Behaviour. Volume 9 (4). Pages 269-277.

PACHECO, T. S.; CHAGAS, A. L. S. (2015). Demanda por transporte na Região Metropolitana de São Paulo e política de pedágio urbano para redução de congestionamento. In: Encontro Nacional de Economia, 43, Florianópolis. Anais... Florianópolis: Associação Nacional dos Centros de Pós-Graduação em Economia.

RHODES, N.; PIVIK, K. (2011). Age and gender differences in risky driving: The roles of positive affect and risk perception. Accident Analysis and Prevention. Volume 43 (3). Pages 923-931.

SMALL, K. (1987). A discrete choice model for ordered alternatives. Econometrica. Volume 55 (2). Pages. 409-424.

SPTRANS. (2017). A SPTRANS. Available at: <http://www.sptrans.com.br/>. Accessed in: 19/01/2017.

STATACORP. (2009a). Stata Statistical Software: Release 11. StataCorp LP.

STATACORP. (2009b). Stata 11 Base Reference Manual. Stata Press.
SWAIT, J.; BEN-AKIVA, M. (1987). Empirical test of a constrained choice discrete model: mode choice in Sao Paulo, Brazil. Transportation Research Part B: Methodological. Volume 21 (2). Pages 103-115.
VOVSHA, P. (1997). Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area. Transportation Research Record: Journal of the Transportation Research Board. Volume 1607. Pages 6-15.
WASHBROOK, K.; HAIDER, W.; JACCARD, M. (2006). Estimating commuter mode choice: a discrete choice analysis of the impact of road pricing and parking charges. Transportation. Volume 33 (6). Pages 621-639.