JAYME TOLPOLAR ANCHANTE

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

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Economia Aplicada, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
A539c
2017

Anchante, Jayme Tolpolar, 1991-
Commute mode choice in the city of São Paulo : an empirical analysis / Jayme Tolpolar Anchante. – Viçosa, MG, 2017.
xi, 89f. : il. ; 29 cm.

Inclui apêndice.
Orientador: Viviani Silva Lirio.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.82-88.

1. Transporte - Trânsito de passageiros. 2. Transportes - Planejamento. 3. Transportes - Aspectos sociais. I. Universidade Federal de Viçosa. Departamento de Economia Rural. Programa de Pós-graduação em Economia Aplicada. II. Título.

CDD 22 ed. 388.322

JAYME TOLPOLAR ANCHANTE

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

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Economia Aplicada, para obtenção do título de *Magister Scientiae*.

APROVADA: 12 de junho de 2017.

Elaine Aparecida Fernandes

André Luis Squarize Chagas

Janderson Damaceno dos Reis
(Coorientador)

Viviani Silva Lirio
(Orientadora)

À minha família

AGRADECIMENTOS

Gostaria de agradecer primeiramente minha família: minha mãe Miriam, meu padrasto Gerson e minha avó Raquele, por todo seu apoio e carinho durante esta empreitada. Agredeço a minha namorada Jezabel por todo amor e compreensão, assim como toda sua ajuda durante o mestrado.

Agradeço a minha "família" em Viçosa, meus amigos e companheiros de residência Allan, Christopher, Gian, Gustavo, Mário, Matheus, Roberto que me acolheram na república e tanto me ajudaram durante minha estada em Viçosa. Também aos meus amigos Fabiano, Gercione e Rodrigo que me acolheram na minha segunda ida a Viçosa para defender minha dissertação.

Agradeço ao corpo docente do Departamente de Economia Rural pela contribuição a minha formação acadêmica e pessoal, especialmente os professores Viviani Silva Lirio e Janderson Damaceno dos Reis por acreditarem no meu trabalho e na minha capacidade. Agradeço ao corpo discente do Departamento pelas amizades, momentos de estudo e de diversão dentro e fora do DER, especialmente aos colegas que ingressaram no meu período: Adilson, Ana, Ascânio, Carlos, Gessica, Jayne, João, Paulo, Raul, Wallace.

Agradeço ao acolhedor e caloroso povo mineiro e de Viçosa pela hospitalidade. A Universidade Federal de Viçosa por seu maravilhoso campus e incrível infraestrutura. Agradeço especificamente ao Conselho Nacional de Desenvolvimento Científico e Tecnológico e em geral ao povo brasileiro por me possibilitarem uma bolsa para realizar meu aperfeiçoamento na pós-graduação.

BIOGRAFIA

Jayme Tolpolar Anchante, fiilho de Cesar Emilio Anchante Negreiros e Miriam Galbinsky Tolpolar, nasceu em 17 de novembro de 1991, na cidade de Porto Alegre, Rio Grande do Sul.

Iniciou em Março de 2009 o curso de Ciências Econômicas na Universidade Federal do Rio Grande do Sul, graduando-se em julho de 2014.

Em fevereiro de 2015 ingressou no Programa de Pós Graduação em Economia Aplicada da Universidade Federal de Viçosa, concluindo os requisitos para a obtenção do título de *Magister Scientiae* em 2017.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

ANCHANTE, Jayme Tolpolar, M.Sc., Universidade Federal de Viçosa, June, 2017. **Commute mode choice in the city of São Paulo: an empirical analysis**. Advisor: Viviani Silva Lirio. Co-advisor: Janderson Damaceno dos Reis.

The present works deals with the urban mobility in the city of São Paulo, Brazil, specifically 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, a discrete choice econometric model that enables the inclusion of alternative-specific and individual characteristic variables, which has as main assumption the Independence of the Irrelevant Alternative (IIA): the probability ratio of two alternatives depends solely on the characteristics of the these alternatives. We also employed the Nested logit model, which does not relay on the IIA assumption, because it nests similar alternatives. Our main data sources are the microdata from the 2007 and 2012 Origin Destination Surveys conducted by the Companhia do Metropolitana de São Paulo; besides that, we also used monthly gasoline prices from the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis and vehicle mileage from the Instituto Nacional de Metrologia, Qualidade e Tecnologia. The Origin Destination Surveys have information about the household and the individual demographic and economic characteristics; the position of the household, workplace and school of the interviewees; and information about the previous workday travels: weekday, origin/destination zone, purpose at origin/destination, mode of transport, hour and minute of departure/arrival. The dependent variable, the choice of the commute mode of transport, has three alternatives: transit, private motorized transport and non-motorized modes. The independent variables are travel time and cost (alternative-specific characteristics), age, sex, if study, employment relationship and degree of education (characteristics of the individual). The results corroborate our initial hypothesis that travel time and cost are negatively correlated with mode choice, besides that, the Conditional logit revealed a Value of Travel Time of 1.78 and 1.09 (nominal R$) for 2007 and 2012, respectively (the Nested logit revealed a Value of Time between R$ 0.04 and R$ 2.47): this is the value the average citizen of São Paulo is willing to pay

to save one hour of commute travel time. The average marginal effect shows both the own-elasticity and the cross-elasticity of a mode's cost and time, with these information it is possible evaluate a series of transport policies that raise or reduce the mode's cost or time in order to discourage or encourage it. We perform several tests for the IIA assumption and we get mixed results: some reject it and some do not, depending on the model chosen and how we nest the alternatives. We conclude that it is possible to change the behavioral pattern of choosing the commute mode of transport by changing the travel cost and time, grounding possible transport policies that tax or improve the speed flow of certain modes relative to others.

RESUMO

ANCHANTE, Jayme Tolpolar, M.Sc., Universidade Federal de Viçosa, junho de 2017. **Commute mode choice in the city of São Paulo: an empirical analysis**. Orientadora: Viviani Silva Lirio. Coorientador: Janderson Damaceno dos Reis.

O presente trabalho trata do tema da mobilidade urbana da cidade de São Paulo, Brasil, especificamente 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, modelo econométrico de escolha discreta que possibilita a utilização de variáveis específicas das alternativas assim como as do indivíduo, tendo como principal hipótese o pressuposto da Independência da Alternativa Irrelevante (IIA): a razão de probabilidades de duas alternativas depende exclusivamente da característica destas duas alternativas. Também utiliza-se o modelo Logit Aninhado, que possui como grande vantagem o relaxamento da hipótese da IIA ao aninhar alternativas semelhantes. Nossa principal fonte de dados são os micro dados das Pesquisas Origem Destino de 2007 e 2012 realizadas pela Companhia do Metropolitano de São Paulo; além destas pesquisas, também utilizamos dados dos preços mensais da gasolina da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis e a quilometragem de veículos automotores do Instituto Nacional de Metrologia, Qualidade e Tecnologia. As Pesquisas Origem Destino possuem informações sobre as caracterísiticas demográficas e econômicas do domicílio e do indivíduos; localização do domicílio, do trabalho e da escola dos entrevistados; e informações sobre os deslocamentos realizados no dia útil anterior à entrevista, dia da semana, zona de origem/destino, motivo na origem/destino, modo de transporte, hora e minuto de saída e chegada. A variável dependente do modelo, a escolha do meio de transporte para o trabalho, possui três alternativas: transporte público, transporte motorizado privado e modos não motorizados. As variáveis independentes utilizadas são custo e tempo de deslocamento (características das alternativas), idade, sexo, se estuda, vínculo empregatício e grau de instrução (características do indivíduo). 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 Logit condicional revelou um Valor do Tempo de Deslocamento de 1,78 e 1,09 (R$ nominais) para 2007 e 2012, respectivamente (o Logit aninhado revelou um Valor do Tempo entre R$ 0.04 e R$ 2.47): este é o valor que um paulistano médio estaria disposto a pagar para economizar uma hora de viagem para o trabalho. Os efeitos marginais médios mostram tanto as elasticidades próprias de tempo e custo de um modo quanto as elasticidades cruzadas, com essas informações é possível avaliar uma série de políticas de transporte que aumentem ou reduzam o custo e/ou tempo de deslocamento de um modo de transporte com o intuito de (des)incentivá-lo. Vários testes foram realizados para avaliar a validade da hipótese IIA e o resultado foi inconclusivo: alguns a rejeitaram enquanto outros não, dependendo do modelo inicialmente escolhido e da forma como são aninhadas as alternativas. Concluímos ser possível alterar o padrão de comportamento de escolha do modo de transporte para o trabalho alterando o custo e o tempo de deslocamento dos modais, embasando possíveis políticas de transporte que taxem ou melhorem a fluidez de certos modais em relação aos demais.

1. INTRODUCTION

1.1. CONTEXT

Everyday millions of people living in modern urban cities move between residence and work, this phenomenon is known as commute. This task accompanies us for most of our adulthood and requires considerable resources: around 8,820 hours of our lives (around 1 year[1]). Despite the clear motivation (going to work), we don't truly understand it, both from a personal and a social point of view, why do I need to commute, what determines the duration of my commute, why do I choose this specific mode[2] and not a different one? The previous might seem too philosophical, so we, rightfully, stick to more mundane questionings, like will I arrive at my destination on time or should I have taken a taxi (perhaps Uber or Cabify) instead of the bus. The present work is going to address both theoretically and empirically one such issue: why do we prefer a specific mean over the others, this is question is known in the transport economics literature as mode choice.

Ever since the Agriculture Revolution, humans have worked in the same place they lived, conceptually and practically, the work commute almost didn't exist. Farmers planted and harvested their crops; ranchers raised cattle, all in one "property", the division of labor was rare and most of the population was dedicated to the primary sector. This lasted with certain stability up until the XVIII century, when another major change happened in modern history: the Industrial Revolution. In this new urban environment driven by the market, the city center became highly valued land, primarily housing industries of various types, thus concentrating most of the jobs. The workforce was displaced to the outer region of cities because of this appreciation of the land, but their access to the city center – and therefore the industrial jobs – was guaranteed by the incipient transit system, moved first by coal, then fossil fuel (nowadays new sources are appearing, such as the biomass, solar and ethanol). Industrialists and the affluent class also moved to the outskirts of the *urbes*, looking for greener and more spacious areas, away from dust and disease. This was perhaps the foundation of the work commute as a global generalized institution, conceived and driven by the "market forces".

[1] Assuming a one-hour daily commute, 252 work-days in a year and 35 years of work.
[2] We will refer to mode or mean of transport simply as mode or mean.

Later on, other events influenced the way we commute and interact with the urban space. The emergence of the automobile with a gasoline internal-combustion engine in the beginning of the XX century was one such event. It was somewhat already present since the mid-XIX century; however, the introduction of the mass production system started by the Fordism culminated in an accessible, fast and durable vehicle that became dominant as the commute mode. People could live and work wherever they wanted, as long as they had a car, the rest was "assured": the road network was expanding rapidly opening new urban frontiers, gasoline was cheap and highways were fast. Only the 1973 and the 1979 Oil Shocks made us rethink our development model. Gasoline prices skyrocketed "forcing" nations to diverse their energetic and transportation matrix. Today the usual commuter, besides the usual private choices as cars and motorcycles, also has available in its list of modes of transport the public transit in its various forms – buses, trolleybuses, trams (or light rail), trains, subway and ferries. These modes are now part of the urban scenery and proved themselves very efficient in terms of energy, especially in densely populated areas.

Nowadays, in the XXI century, with an ever increasing concern about the future of the planet's environment, the non-motorized – also known as active – modes are returning to vogue, for being non-pollutant, healthy and cheap means of transportation. On the technology side, we are witnessing a major change in individual transportation with the new cellphone apps, such as Uber, Cabify, EasyTaxi and others, that are providing specialized, intelligent and on-demand transportation solutions.

During most of the commute development process, governments played a role of sustaining the fluidity of system constructing and maintaining; organizing it by installing traffic lights, paint and post signaling; and managing it during exceptional times, like accidents, using traffic agents and also during "normal" times, collecting tolls for maintenance, for example. The concept that encompasses the previous interventions is called passive policies, because the government only reacts to changes in the traffic conditions. It is still passive in the sense that it can't make structural changes in the current traffic "equilibrium" - e.g. the current congestion level (DURANTON, TURNER, 2011; HSU, ZHANG, 2014). During the mid-XIX century, governments started to change their behavior towards traffic - one of the milestones was the Smeed Report in the United Kingdom (SMEED, 1964) - planning and managing it keenly, for example, restraining the access of motorized traffic to some places, usually

the City Business District (CBD), and charging users for their congestion impact. These set of policies are called active because, contrary to the passive ones, they have a specific goal, generally improving the social well-being (not private interests of a few) and this goal is usually different of what would have happen had the free forces of the market directed the traffic conditions.

One of the reasons for government intervention is that many private transportation decisions usually lead to sub-optimal social results, as we can see when we deal with underpriced roads from the classical routing paradox of Pigou-Knight-Downs (PIGOU, 1920; KNIGHT, 1924; DOWNS, 1962), unequal competition between two modes as the Downs-Thomson paradox (DOWNS, 1962; THOMSON, 1977) and also when poor network building happens as in the Braess paradox (BRAESS, 1968). Of course, influencing collective decisions can have various outcomes depending on many nuances, for example, the type and objective of the operator of such scheme (private profit maximizing or public welfare maximizing), what is exactly managed (traffic lanes, whole links or region rings) (DE PALMA, LINDSEY, WU, 2008).

Among the active policies, it seems a world pattern that developing cities usually implement some form of road space rationing whereas developed cities ones prefer a congestion charge approach. Some permanent rationing scheme examples are: Athens (1982), Santiago (1986), México City (1989), Metro Manila (1995), São Paulo (1997), Bogotá (1998), La Paz (2003), San José (2005), countrywide in Honduras (2008) and Quito (2010). Other places introduced it temporarily, such as Beijing, Italian cities, Paris and New Delhi. Road pricing schemes have been implemented in Singapore (1975), Stockholm (2008), Helsinki (2011), London (2012) and Milan (2012).

Another emerging policy outside the "usual" instrument set seems to be the incentive to the bicycle commute. For example, France announced a program in partnership with 20 companies (employing a total of 10,000 employees) in the summer of 2014 offering € 0.25/km for people who cycle to work. French Transport Minister Frederic Cuvillier expected that the scheme would boost bike use for commuting by 50% (DE CLERCQ, 2014). However, the results of the six-month trial weren't so impressive: 419 people agreed to participate, but only 19% were shifting from driving, one major obstacle of the scheme is that transit subsidies and free parking are established policies (JAFFE, 2015). Several other countries are implementing such

policies, for example, the United States approved the Emergency Economic Stabilization Act of 2008, which added the section 132(f) (known as the Bicycle Commuter Benefit) to the IRS Code, giving the opportunity of employees to receive a US$ 20/month, tax-free, from their employers under the simple rule of regularly commuting by bicycle, no forms or requirements needed (RUGGIERO, 2016; IRS, 2017). Also, Belgium and Netherlands are offering tax-free payment of € 0.22 and €0.19/km of bike commuting respectively. In Germany commuters biking to work can deduct €0.30/km from their taxable income. In the United Kingdom, more than 10,000 companies use the cycle to work scheme that allows employees to purchase bikes using pre-tax income, thereby saving income tax and other deductions on their pay and the German federal government allows employers to offer "company-paid private bicycles" as a benefit to their employees (BUEHLER, HAMRE, 2014).

A very important aspect, already cited previously, in mode choice is certainly automobile ownership and usage dependency, which was observed in the United States by Downs (1992) and Downs (2004), which was also observed in Brazil by Brinco (2005). The motorization rate (motor vehicles *per* thousand people) is growing rapidly in most of the developing world (GAKENHEIMER, 1999), especially in China (RILEY, 2002). Total vehicle stock will increase from about 812 million in 2002 to over 2.08 billion units in 2030. By this time, 56% of the world's vehicles will be owned by non-OECD countries, compared with 24% in 2002. In particular, China's vehicle stock will increase nearly twenty-fold, to 390 million in 2030 (20.5 million in 2002) and Brazil's stock of vehicles jumped from 1 million in 1960 to 20.8 million units in 2002, with estimates of going up to 83.7 million units in 2030 (DARGAY, GATELY, SOMMER, 2007).

Although the automobile offers extraordinary personal mobility and independence, it also directly effects the air pollution, motor vehicle crashes, pedestrian injuries and fatalities (FRUMKIN, 2002). Cars and trucks account for approximately 30% of emissions of oxides of nitrogen and hydrocarbons in the United States, according to the Environmental Protection Agency (2002). Physical inactivity is a well-documented risk factor for chronic diseases, including coronary heart disease, stroke, some cancers, diabetes, and depression; it is responsible for about 200,000 deaths in the US each year, second only to tobacco, according to the United States Department of Health and Human Services (2010). Although we are not explicitly

saying that the automobile is causing physical inactivity, it is certainly correlated with it, since auto trips (especially daily commuting) made at walkable distances could otherwise combine active transportation (also known as non-motorized, represented mostly, but not totally, by walking and cycling) with an active health lifestyle, collaborating with transportation, urban planning and health fields (SALLIS *et al*., 2004).

On the other side, cycling is related to general health improvement. Hartog *et al.* (2010) and Rojas-Rueda *et al.* (2011) estimated the benefits from a mode shift towards cycling for Netherlands and for Bicing users, the public bicycle sharing initiative in Barcelona, Spain, respectively. Oja *et al.* (2011) made a systematic review on the health benefits of cycling, showing a positive correlation with cardiorespiratory fitness among children/adolescents and an inverse relationship with cardiovascular disease/coronary heart disease mortality, cancer mortality and morbidity among adults. The same goes for walking; Lee and Buchner (2008) reviewed the literature related to walking as a moderate-intensity physical activity. A practitioner would have lower rates of chronic diseases, such as obesity and cardiovascular, less medical expenditures and only a slight increase in activity-related injuries. Mutrie *et al*. (2002) made an intervention in three workplaces in Glasgow, Scotland called "Walk in to work out" to promote walking commute. The intervention group showed a significant improvement in general and in mental health, compared to the control group.

Now we move on to present our geographical region of interest, São Paulo, and its country. Against the policy of most countries and cities that encourage active transportation, Brazil is doing the opposite: auto ownership and usage is subsidized, transit is encumbered. One of the possible reasons is the automotive industry weight in the Gross Domestic Product (GDP) and its chaining power, policy-makers decided to exempt new vehicle sales from the Tax over Industrialized Products (IPI), for example, as a counter-cyclical measure against the world crisis of 2007-2008, and sustained the reduction of the Tax over the following years. According to the Intituto de Pesquisa Econômica Aplicada (IPEA, 2009), the IPI policy represented a renunciation of R$ 1.8 billion in tax revenue, but also made possible to maintain around fifty thousand direct and indirect jobs in the country. Alvarenga *et al.* (2010) estimated that the IPI reduction was responsible for 20.7% of the automobile sales in 2009.

Figure 1 reinforces the arguments made in the previous paragraph, we can see that the IPCA[3]'s sub item Personal vehicle[4], which comprises all costs of owning and using a private motor vehicle, except fuel, is the lowest accumulated inflation in the July, 2006 to December, 2016 period among the selected inflations. In the Panel (a), the Public transportation[5] accumulated inflation is almost always higher than the IPCA (represented by the solid black line), except for a short period between July, 2013 to January, 2015. This is most likely due to the social uprising called the "20 cents manifestation" or "June journeys" whose tipping point that brought the masses to the streets was the transit fare adjustment, besides that, also government dishonesty, underinvestment in education and health and overinvestment in sports events (BARREIRA, 2014). Gasoline[6] accumulated inflation is much lower than the Public transportation or the IPCA during the whole period. Carvalho *et al*. (2013) analyzed the Pesquisa de Orçamentos Familiares data also note that the average household spends 3% of its income with public urban transit. However, this is not the same for all income groups, the 10% poorer committed 15% of their income and the next decile, 11.7%. Also, from the 10% poorer, 30% doesn't undertake any spending with this item, which could be a measure of exclusion, because they lack the pay capacity.

[3] IPCA, short for *Índice de Preços ao Consumidor Amplo* or Broad Consumer Price Index, is the official government target index, which covers families with monthly income ranging from one to forty minimum wages, whatever the source, living in urban areas. During the period shown in the figure, the inflation target was 4.5% (with limit bands of ±2%).From July/2006 to December/2011, the Weighting Structure followed the 02/03 Family Budget Survey, and from January/2012 to December/2016 the 08/09 Family Budget Survey.

[4] Personal vehicle is equal to the group "5102. Veículo próprio" and it comprises: new automobile, license plate and licensing, voluntary vehicle insurance (called DPVAT in Brazil), traffic ticket, lubricating oil, accessories and parts, tire and inner tube, automobile repair, parking, toll, lubrication and washing, used automobile, tow, vehicle paint, vehicle rent and motorcycle.

[5] The Public transportation represents the subgroup created by the author which comprises the sub items "5101001. Ônibus urbano", "5101002. Táxi", "5101004. Trem", "5101006. Ônibus intermunicipal" and "5101011. Metrô".

[6] Gasoline represents the sub item "5104001. Gasolina".

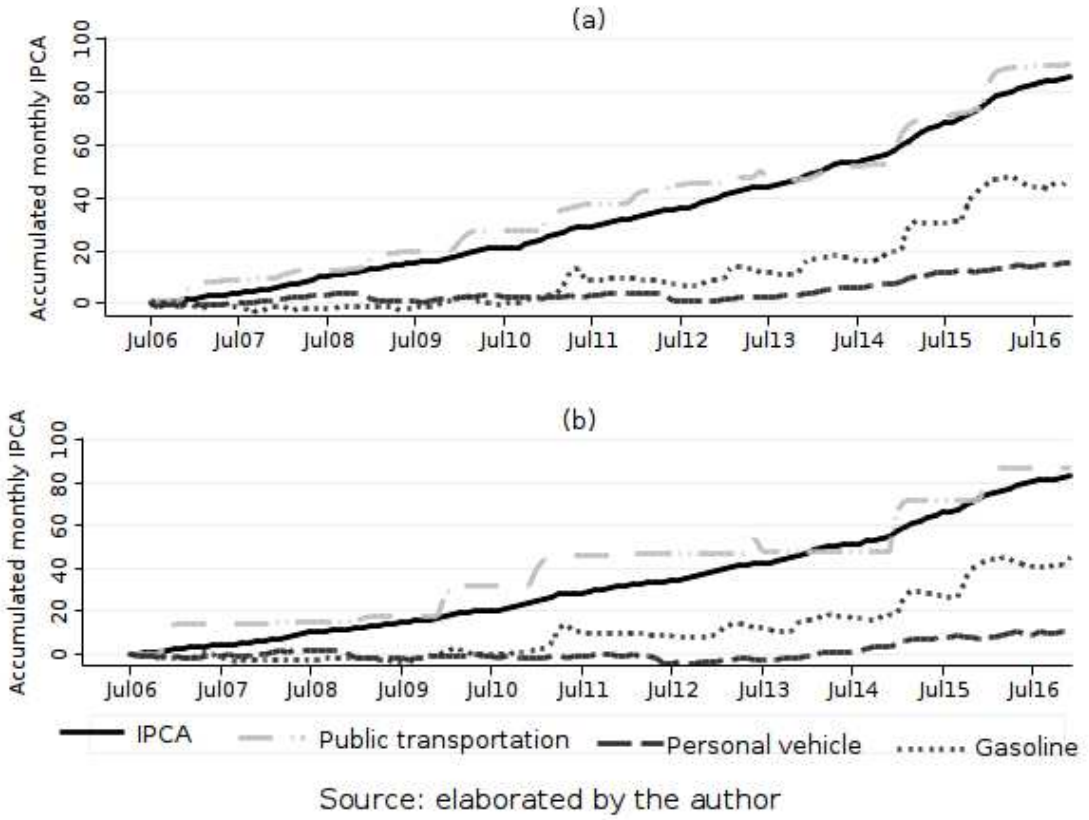


Figure 1: Monthly accumulated Broad Consumer Price Index and selected sub-groups (Jul/2006-Dec/2016). Panel (a) shows the accumulated for Brazil. Panel (b) shows the accumulated for the Metropolitan Region of São Paulo

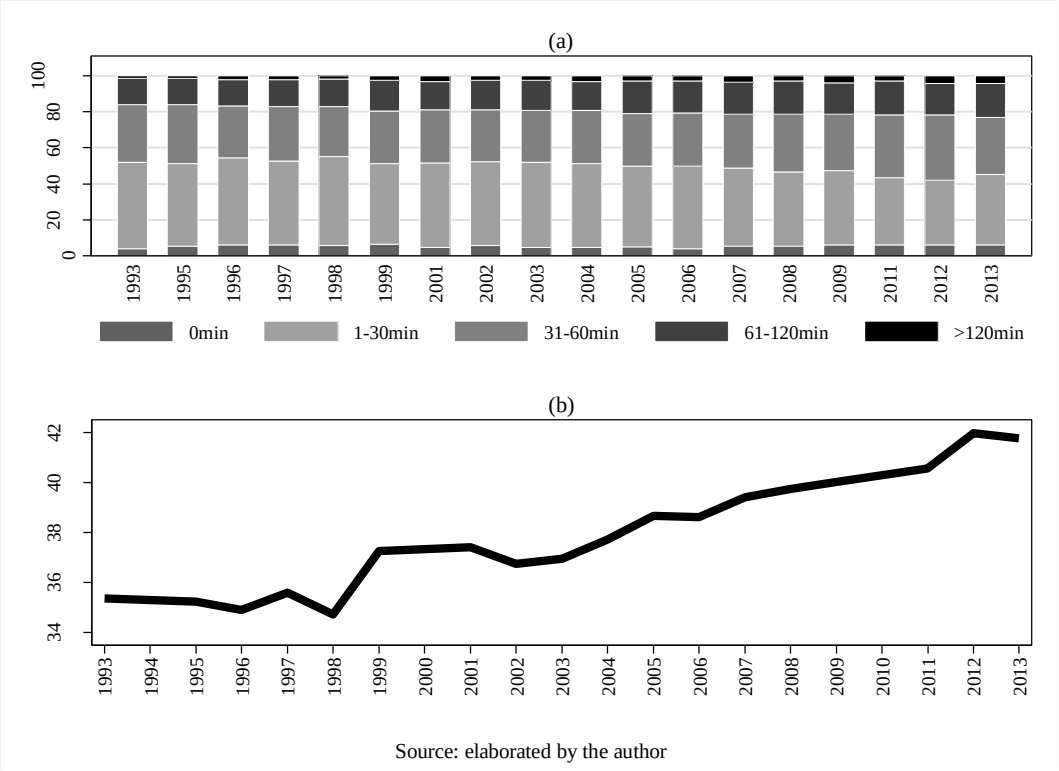
It is a clear fact that the relative prices between private and public transportation has been unfavorable to the latter if we look at Brazil as a whole. In Panel (b), which shows the same accumulated inflations only for the Metropolitan Region of São Paulo, we can make the same points. The relative price is even more disadvantageous to the Public transportation, since Personal vehicle has an accumulated negative inflation (deflation) most of the period up until February, 2014. The period of rise in the Personal vehicle accumulated inflation coincides with the gradual increase in the IPI (2% for vehicles less than 1000cc from 01/01/2012 until 31/12/2013, 3% for the same category until 31/12/2014 and 7% from 01/01/2015, which was the original rate prior the beginning of the reduction).

Given the priority of the automobile over the more sustainable means of transportation, among other factors, apparently traffic conditions are worsening. As

shows the Figure 2, average morning commute travel times are rising steadily over the past twenty years in the Metropolitan Region of São Paulo (MRSP), according to the Instituto Brasileiro de Geografia e Estatística data[7].

A possible explanation for the simplified whole scenario is that as auto trips become more attractive (cost-effective or simply cheaper), more people will use it. This also means that less people will use other modes; in the case of MRSP it means mostly walking, bus and subway. Because of the inherent properties of transit (that we will discuss in the following sections) in regards to the economies of scale, losing passengers means at least one of two things: a higher fare or a decrease in the frequency. In other words, this leads to a lower level of service, so it becomes less attractive and the marginal user will choose auto trips. This process is leveraged by public national policy that incentives car ownership and usage. As time passes by, commuters become more "auto-dependent" and it gets harder to propose policies that try to break rule of Automobile Predominant Cities. Even in developed nations with high transit usage, like Europe and East Asia, road pricing policies face great opposition from citizens, even though it is well established in the academia that the whole society can benefit from such policy.

[7] The calculation methodology can be found in the Appendix A.



Source: elaborated by the author

Figure 2: Morning commute times of the Metropolitan Region of São Paulo from the PNAD data (1993-2013). Panel (a) shows percent sum of categories (%). Panel (b) shows average commute times

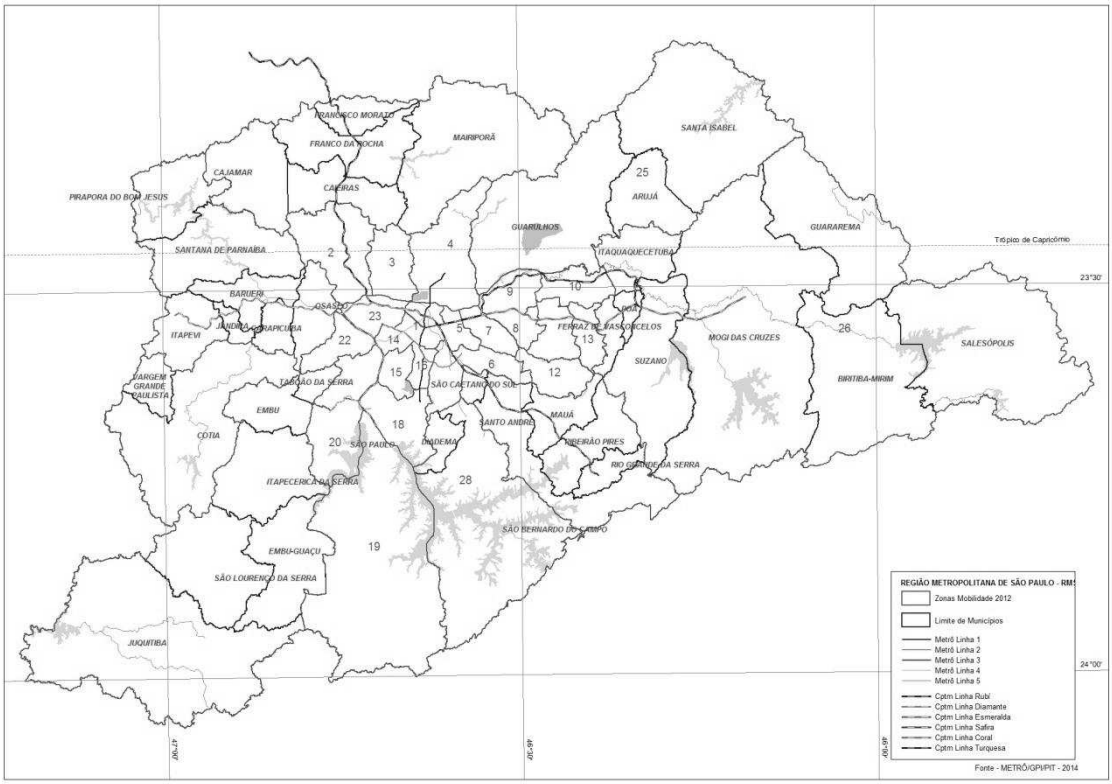
Another important aspect in this process is the public expenditure on transportation. For example, the São Paulo city council spent on urban infrastructure R$ 454,658,604.3 in 2004, R$ 676,466,698.5 in 2007 and 685,571,645.85 in 2012; on road transport R$ 239,960,365.85 in 2004, R$ 443,010,206.2 in 2007 and R$ 633,543,325.79 in 2012; and on other expenditures on the transportation function (mainly transit subsidies) R$ 1,026,263,453.66, R$ 937,682,525.4 in 2007 and R$ 1,708,160,551.41 (all nominal values). This is the huge budget necessary to maintain the São Paulo's transport system functioning: it has the largest values of all Brazilian cities except for the urban infrastructure in 2012, which Rio de Janeiro surpasses it, probably because of the preparations for the 2012 Olympic Games.

Continuing on the MRSP, it implemented some interesting pioneering policies, for example, the Bilhete Mensal, a single fee which gives the right to freely use the city transit system, R$ 140,00 for the bus, and R$ 230,00 for bus, subway and train, with a half discount for students (SÃO, PAULO, 2013). Also, a law guaranteeing a zero-fee

for the metropolitan transit for students of the public education system and low-income of the private system, a measure that could potentially benefit 615,000 students (MACIEL, 2015). São Paulo was also one of the first Brazilian cities to regulate the individual transportation mobile applications in 2016.

With a unique, dynamic and fast grown economy since the XVIII century during the gold exploration (FURTADO, 2007), its strength can also be seen in the transportation indicators. If the MRSP were a country, it would rank among the top 20 countries with a motorization rate (motor vehicles per thousand inhabitants) of 541, according to the Organisation Internationale des Constructeurs d'Automobiles (International Organization of Motor Vehicles Manufacturers or OICA) and the 2013 Departamento Nacional de Tráfego's (DENATRAN) data. The Brazilian average is 198, more than two times less than the MRSP.

Apart from the automobile, the MRSP has a robust transport network and is served by many public modes. Controlled by the State Government of São Paulo, the Companhia Paulista de Trens Metropolitanos (CPTM) and the Companhia do Metropolitano (Metrô) operate the intercity train lines and the subways lines, respectively. The city of São Paulo is also served by 16 private bus consortia under the management of the São Paulo Transportes S.A. (SPTrans), with a total 15 thousand vehicles and more than 1,300 bus routes. The Empresa Metropolitana de Transportes Urbanos de São Paulo (EMTU/SP) is a state-controlled company responsible for the intercity bus system for the five metropolitan regions of the state of São Paulo (including the MRSP itself) (SPTRANS, 2017).



Source: Metrô (2012c)

Figure 3: Map of the Metropolitan Region of São Paulo, city and traffic zone limits, showing the rail transit system (2012)

1.2. RESEARCH QUESTION AND CONTRIBUTION

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. 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).

1.3. HYPOTHESIS

Travel cost and travel time are negatively associated with the probability of choosing a certain commute mode of transport.

The individual characteristics of the commuter have various kinds of association with the probability of choosing a certain commute mode of transport depending on the characteristic: i) sex, males have a higher probability of choosing private modes than females; ii) age, younger people have a higher probability of choosing transit and non-motorized modes than elder people; iii) students have a higher probability of choosing transit than non-students; iv) employment relationship, people with a contract have a higher probability of choosing private and transit modes than other forms of employment relationships; v) education degree, a higher education degree is associated with a higher probability of choosing transit and private modes and a lower probability of choosing non-motorized modes.

1.4. OBJECTIVES

1.4.1. General objective

Analyze how travel cost and travel time, as well as the individual characteristics, are associated with the probability of choosing a certain commute mode of transport in São Paulo in 2007 and in 2012.

1.4.2. Specific objectives

Calculate the Value of Travel Time Savings (VoTT), in other words, how much commuters are willing to pay to save a certain amount of travel time.

Calculate the average marginal effect of a unit increase of a mode's cost on the probability of choosing the same mode (own-effect) and the probability of choosing other modes (cross-effects).

Calculate the average marginal effect of a unit increase of a mode's travel time on the probability of choosing the same mode (own-effect) and probability of choosing other modes (cross-effects).

2. BACKGROUND

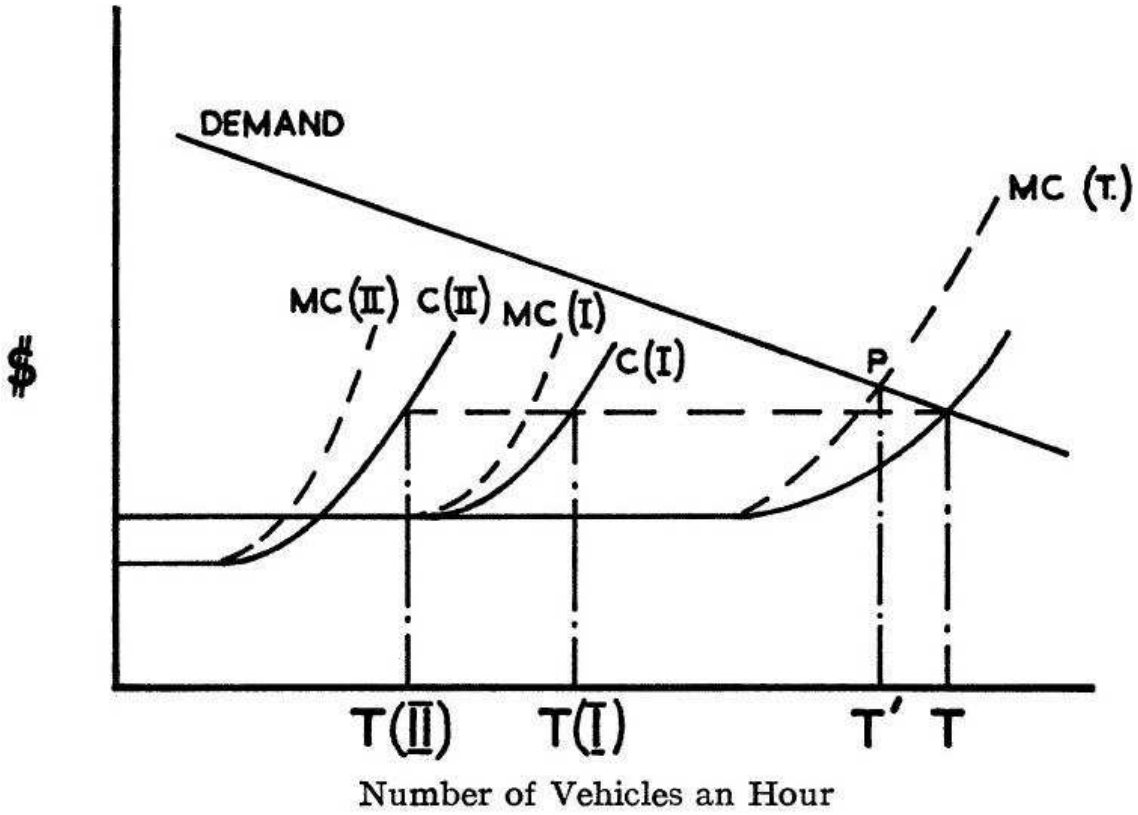
2.1. THEORETICAL FRAMEWORK

2.1.1. Transportation market

The transport market functions like any other market (*e.g.* oranges and labor) it may be represented, from the point of view of neoclassical economics, by supply and demand, equilibrium, price and quantity. Following one of the first formally developed theoretical transport models, we can see these relationships from Walters (1961). Assuming homogeneity (identical vehicles and drivers) and the following traffic conditions:

> Consider two roads connecting A and B and assume that the unit (private) cost of a vehicle using road I is represented by C(I) in Figure 1 [Figure 4 in the present work]. The unit cost is first horizontal but, as the number of vehicles increases, some congestion occurs and unit costs rise. Similar conditions are apparent on road II. The unit cost curve or vehicles using road II is lower than the unit cost on road I for low traffic volumes, but for higher traffic volumes the unit cost on road II rise above the cost on road I. The unit cost curve of road II is shown by C(II). If we assume that the routes are perfect substitutes for one another we can draw one demand curve for the services of the two roads (WALTERS, 1961, p. 677-78).

The main cost associated with the vertical axis is travel time. Therefore, as demand increases, that is, if more vehicles enter a given transport system, the higher will be everyone's travel time. A more severe and prominent case would be the morning commute: as most workers begin at almost the same time (something between 7-9 a.m.) and in the same place (the Center Business District or CBD), a peak-hour demand would make aggregated travel times rise steeply (at an extreme case, it would reach infinity or everyone is simply still).



Source: Walters (1961, p. 678)

Figure 4: Road demand, individual and aggregate private and marginal cost curves in a simplified two-road network with homogeneous drivers

> The unit cost curves C(I) and C(II) in Figure 1 [Figure 3 in the present work] then measure marginal *private* cost. Equilibrium traffic flow can be found by summing the unit cost curves horizontally to give C(T). Then OT is the total traffic flow, OT(I) on highway I and OT(II) on highway II. If we take the total traffic as given at OT the distribution of traffic between roads will be efficient only if at T(I) and T(II) the marginal *social* costs, represented by MC(I) and MC(II), are equal. In general, efficiency requires a reduction in total traffic from T to T' and the distribution between roads such that the marginal social costs are both equal to PT' WALTERS, 1961, p. 678.

Arnott and Small (1994) discusses some extensions of this simple model, which is also known as the Pigou-Knight-Downs paradox, due to the works of Pigou (1920), Knight (1924) and Downs (1962). For example, one can introduce a new mode choice: transit, as in the Downs-Thomson paradox, due to Downs (1962) and Thomson (1977); or a different type of road network can be specified as in the Braess paradox, due to the work of Braess (1968). These and other cases are mathematically formalized in the book of Small and Verhoef (2007), for example, the case of two modes – car and transit – and

two roads - one free and other tolled – under a first-best and a second-best pricing policy.

Other topics and paradigms about the transportation market are highlighted in the following paragraphs. The road network is often cited as an example of a public good. Public goods have both the characteristics of being non-rival and non-excludable. A good is considered non-rival if, for any level of production, the cost of providing it to the marginal individual is zero or if its consumption by one consumer prevents simultaneous consumption by others. A good is non-excludable if non-paying consumers cannot be prevented from accessing it (MYLES, 1995). The simple transportation market representation shows that roads are rival, because, as more people use it, the marginal cost increases. Also, roads are excludable, because it is possible to exclude non-payers from using it, as it is happening in congestion charge schemes previously cited. That is why congestionable goods are better called impure, according to Myles: "Such public goods are termed impure. The utility derived by each household from an impure public good is an increasing function of the level of supply and a decreasing function of its use." (1995, p. 259).

During high congestion periods, the inefficient allocation follows the rule of first capture, as highlighted by Nash (2007): the road is used by those who arrive first and not by those who most need (ambulances, firefighters etc.) or value (high willingness to pay) it. Given a certain congestion level, the roads are used by those who are willing to bear its cost. Therefore, each individual, depending on their marginal cost, will have different incentives to use the roads.

The problem of overuse of the roads can also be seen from the point of view of the Tragedy of the Commons (HARDIN, 1968). Even though society would be better off with less road use (the social equilibrium, for instance), each individual has incentives to deviate and consume more than the "agreed". Again, it occurs because of the difference between the private and the social costs.

2.1.2. Downs-Thomson Paradox

There are many ways our problem could be studied. We will present first a simple model known as Down-Thomson (D-T) Paradox, where the commuter has two route choices: a congestionable facility for auto driving and a transit route, to illustrate how incentives to private modes can be "perverse" to public transportation. Then we will derive a formal mathematical model for the fare level given an underpriced auto alternative.

The D-T Paradox occurs in the sense that every individual traveler suffers from a higher generalized travel cost after the highway capacity expansion due to the direct and indirect effects of the volume shifting. It is named after the seminal papers of Downs (1962) and Thomson (1977).

We will discuss the paradox using the example presented by Arnott and Small (1994). Consider a commuter facing a route decision, whether take: i) a congestionable facility, *e.g.* an expressway with a bridge, for the auto mode; or ii) a transit alternative, *e.g.* rail, which operates with economies of scale, that is, frequency rises (and travel times decreases) as more people use it. Any capacity expansion in the auto route, as well as any policy that benefits the auto usage like the IPI reduction in the Brazilian case, will divert people to the congested road, thus travel times will tend to remain constant.

However, this incentive also imposes an external cost on transit users, as less people use transit, frequency decreases and travel times increases. The paradox only disappears if bridge capacity is expanded above the level where it can accommodate all commuters, which is usually not feasible in practice, due to high investment costs.

We will now turn to the mathematical model presented by Small and Verhoef (2007) about optimal fare level of public transit when there is a competing mode that is underpriced at peak-hour. We will assume that the transit alternative is some form of rail because it doesn't interact with the automobiles on the road, leading to a simpler mathematical formulation.

Let q_A and q_R be the auto and rail trips per unit of time, with vehicle flow rate of $V_A = q_A$. The joint demand for these two types of travel can be derived from a benefit function $B(q_A, q_R)$, which is the area under the inverse demand curve for the case of more than one good. The inverse demand for either mode, given the amount consumed of the other type, is the partial derivative of the benefit function:

$$d_k(q_A, q_R) = \frac{\partial B(q_A, q_R)}{\partial q_k}, k = A, R \quad (1)$$

Let $C_A(\cdot)$ and $C_R(\cdot)$ be the total cost functions for auto and rail, including user cost for auto and both user and agency costs for rail:

$$\begin{gathered} C_A = C_{A-users} = q_A \cdot c_A(q_A) \\ C_R = C_{R-agency} + C_{R-users} = q_R \cdot c_{R-agency}(q_R) + q_R \cdot c_{R-users}(q_R) \end{gathered} \quad (2)$$

As said before, costs on the two modes are independent of each other. The generalized price of mode $k(k = A, R)$ is defined as the average user cost, c_k, plus toll or fare payment, τ_k, and the user equilibrium conditions are such that marginal benefits are equal to this price for each transportation good:

$$\begin{gathered} \frac{\partial B}{\partial q_A} = c_A(q_A) + \tau_A \\ \frac{\partial B}{\partial q_R} = c_{R-users}(q_R) + \tau_R \end{gathered} \quad (3)$$

Social surplus can be defined as benefits minus costs:

$$W = B(q_A, q_R) - C_A(q_A) - C_R(q_R) \quad (4)$$

The first-best solution is the marginal-cost pricing for each mode. First-order conditions for maximizing Eq. (4), after substituting Eq. (3), produce the following first-best policy:

$$\begin{gathered} \tau_A = q_A c'_A \\ \tau_R = q_R q'_{R-users} + q_{R-agency} + q_R c'_{R-agency} \equiv \chi_R \end{gathered} \quad (5)$$

The optimal road tax is the usual Pigouvian toll. The optimal transit fare is the average agency cost with downward adjustments for any scale economies there might be in the user costs (first term) and agency costs (last term). The identity, χ_R, is the short-hand for the three rail-cost terms, known as quasi first-best toll.

Now consider the second-best solution, where auto toll is fixed at zero. We set up the Lagrangian for this case:

$$\begin{gathered} \Lambda = B(q_A, q_R) - C_A(q_A) - C_R(q_R) + \lambda_A \cdot [c_A(q_A) - \partial B/\partial q_A] + \lambda_R \cdot \big[c_{R-agency}(q_R) + \\ \tau_R - \partial B/\partial q_R\big] \end{gathered} \quad (6)$$

We find that $\lambda_R = 0$ and, substituting this and the constraints into the firs-order conditions with respect to q_A and q_R, we obtain the following optimality conditions:

$$\begin{aligned} \tau_R - \chi_R + \lambda_A B''_{AR} &= 0 \\ -q_A c'_A + \lambda_A (c'_A - B''_{AA}) &= 0 \end{aligned} \tag{7}$$

Where B''_{AR} and B''_{AA} are second derivatives of $B(q_A, q_R)$. The second-best transit fare is therefore:

$$\tau_R = \chi_R - q_A c'_A \cdot \frac{-B''_{AR}}{c'_A - B''_{AA}} \tag{8}$$

This second-best optimal fare equals the non-internalized marginal cost of transit, χ_R, less a term that multiplies the marginal congestion externality on the road ($q_A c'_A$) by a weight depending on demand sensitivities and marginal congestion cost. It depends not only on the slope B''_{AA}, but also on the cross-effect B''_{AR}. This last weight is equal the number of new road travelers per rider deterred from the rail.

We can understand Eq. (8) considering some special cases. If modes are perfect substitutes, $B''_{AR} = B''_{AA} < 0$ and both are equal to q', the slope of the combined inverse demand curve for auto and rail trips together. It becomes the same solution as the case for two parallel roads where one is optimally tolled, except for the quasi first-best rail fare, χ_R. When the modes are imperfect substitutes, the fraction remains positive (so the entire term is negative), but it is smaller than the perfect substitutes case. This is because the cross-derivative in the numerator is smaller in absolute value than the second derivative in the denominator, making it less attractive to lower the transit fare in order to reduce auto congestion from an efficiency point of view, because the modes are imperfect substitutes.

When cross-elasticities of demand are zero, the whole fraction becomes zero. The quasi first-best ($\tau_R = \chi_R$) becomes optimal, because auto use can't be affected. When the modes are complements, so that B''_{AR} becomes positive, the fraction becomes negative and the entire term becomes positive: we then raise the rail fare beyond marginal cost in the transit market because doing so reduces auto traffic (complementarity would be a situation where most auto traffic consisted of people traveling to a transit station).

We can also introduce some non-zero τ_A charge for autos replacing $q_A c'_A$ by $q_A c'_A - \tau_A$. We can also rewrite χ_R in terms of average and marginal costs of Eq. (5) and put the result as a per-user transit subsidy σ_R:

$$\sigma_R \equiv c_{R-agency} - \tau_R = (ac_R - mc_R) + (q_A c'_A - \tau_A) \cdot \frac{-B''_{AR}}{c'_A - B''_{AA}} \quad (9)$$

Eq. (9) shows there are two sources of second-best transit subsidies in the model. The first one is scale economies: if average cost exceeds marginal cost, it is desirable to subsidize the difference. The second is auto congestion: insofar as lowering transit price is effective in reducing congestion costs by drawing away automobile users, it is desirable to use subsidies to encourage this result.

The size of the component depends on the degree of on the degree of underpricing on auto traffic. If optimal congestion pricing was in place, the whole term would disappear and optimal subsidy would be much smaller. Therefore, congestion pricing can be viewed as a solution not only to the problem of traffic congestion, but also to the problem of transit finance. We will not turn to the literature review regarding mode choice and pricing policies.

2.1.3. Random utility maximization model

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:

$$If\ U(X_j, S_i) \geq U(X_k, S_i)\ \ \forall j\ \ \rightarrow j > k\ \ \ \forall k \in C \quad (10)$$

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:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \quad (11)$$

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 $U_{ij}, \dots, U_{iJ}$. With $y_i \in \{1 \dots J\}$ denoting the alternative that decision maker i chooses, this probability is:

$$\begin{aligned} P_{ij} = Pr(y_i = j) &= Pr\left(U_{ij} > U_{ik}\ \forall\ k = 1 \dots J : k \neq j\right) \\ &= Pr\left(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik}\ \forall k = 1 \dots J : k \neq j\right) \end{aligned} \tag{12}$$

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:

$$V_{it} = V(S_i) + V(X_j) + V(S_i, X_j) \tag{14}$$

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 $\sigma^2 = \pi^2/6$. McFadden (1974a) demonstrated that the P_{ij} that individual i chooses j is:

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}} \tag{15}$$

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).

2.3.1.1. Independence of Irrelevant Alternative (IIA) tests

The IIA assumption is commonly explained in the context of the blue bus/red bus paradox. Suppose an individual has equal chances (one to one ratio or fifty-fifty per cent) of choosing between its car and the red bus to commute. A new alternative is introduced – the blue bus – and it has the same characteristics as the red bus, except the color. If the IIA holds, the ratio of choosing between the red bus and the car should remain the same as before (one to one ratio), but now it must be 33.33 to 33.33%, because the blue bus also has a 33.33% chance. A more natural scenario would be the car continuing to have a 50% chance and other 50% equally distributed between the red bus and the blue bus.

There are various tests to see if the IIA assumption holds, we will follow the discussing in Cheng and Long (2007), although the author are very reticent about the size and power of the following tests. McFadden, Train and Tye (1981) propose a likelihood ratio test comparing the full model (which includes all possible outcomes for the dependent variable) and a restricted model with $J-1$ alternatives:

$$MTT = -2\left[L_r\left(\widehat{\beta^f}\right) - L_r\left(\widehat{\beta^r}\right)\right] \tag{16}$$

Where L_r is the log-likelihood function for the restricted estimation. Quite simply, the test compares the value of the log-likelihood equation from the restricted estimation to the value obtained by plugging estimates from the full model into the log-likelihood from the restricted model. When IIA holds, MTT is distributed as a chi-square with degrees of freedom equal to the number of parameters in the restricted model.

The Small and Hsiao (1985) showed that MTT is asymptotically biased toward accepting the null hypothesis, so they propose a modified version of the MTT. First, the sample is randomly divided into subsamples A and B of roughly equal size. The full model from equation (1) is estimated on both subsamples, with estimates contained in β_A^f and β_B^f. The weighted average of the coefficients from the two samples is defined as:

$$\widehat{\beta_{A\,B}^f} = \left(\frac{1}{\sqrt{2}}\right)\beta_A^f + \left[1 - \left(\frac{1}{\sqrt{2}}\right)\right]\beta_B^f \tag{17}$$

A restricted subsample is created from subsample B by eliminating all cases with a given value of the dependent variable—in our case, category J. The restricted choice set is estimated using the restricted subsample yielding the estimates β_B^r with the likelihood function L_r. The Small-Hsiao statistic is:

$$SH = -2\left[L_r\left(\widehat{\beta_{A\,B}^{f*}}\right) - L_r\left(\widehat{\beta_B^f}\right)\right] \tag{18}$$

In which SH is asymptotically distributed as chi-square with the degrees of freedom equal to the number of parameters in the restricted choice set.

Hausman and McFadden (1984) proposed a Hausman test (Hausman, 1978) that compares the estimates $\widehat{\beta^f}$, which are consistent and efficient if the null hypothesis is true, to the consistent but inefficient estimates $\widehat{\beta^r}$. The HM test is defined as:

$$HM = \left(\widehat{\beta^r} - \widehat{\beta^f}\right)\left[\widehat{Var}\left(\widehat{\beta^r}\right) - \widehat{Var}\left(\widehat{\beta^f}\right)\right]^{-1}\left(\widehat{\beta^r} - \widehat{\beta^f}\right) \tag{19}$$

Where $\widehat{Var}\left(\widehat{\beta^r}\right)$ and $\widehat{Var}\left(\widehat{\beta^f}\right)$ are the estimated covariance matrices. If IIA holds, HM is asymptotically distributed as chi-square with degrees of freedom equal to the rows in $\widehat{\beta^r}$. Significant values of HM indicate that the IIA assumption has been violated. Hausman and McFadden (1984) note that HM can be negative if $\widehat{Var}\left(\widehat{\beta^r}\right) - \widehat{Var}\left(\widehat{\beta^f}\right)$ is not positive semidefinite, but they conclude that this is evidence that IIA holds.

2.3.1.2. Relaxing the Independence of Irrelevant Alternative assumption

If the IIA assumption is rejected, one could use Nested multinomial logit (NMNL) since it relaxes the i.i.d. errors assumptions by allowing similar alternatives to have correlated errors in the same nest. We follow Heiss (2002) for a derivation of the model consistent with RUM. The researcher can partition the choice set into M subsets (nests) $B_m, m = 1, \dots, M$. The nest which alternative $j = 1, \dots, J$ belongs as $B(j)$:

$$B(j) = \{B_m : j \in B_m,\ m = 1, \dots, M\} \tag{20}$$

The probability of individual i choosing j can be decomposed into two parts, $Pr(y_i = j)$ - for a nesting structure of $M = 2$ – is equal to the product of the probability

to choose some alternative in nest $B(j)$, $Pr\{y_i \epsilon B(j)\}$, and the conditional probability to choose j given that the nest has already been chosen $Pr\{y_i = j|y_i \epsilon B(j)\}$, that is:

$$P_j = \Pr(y = j) = \Pr\{y = j|y \in B(j)\} \cdot \Pr\{y \in B(j)\} \tag{21}$$

In the nested logit, the set of errors $\varepsilon_{i1}, \dots, \varepsilon_{ij}$ are assumed to follow a the generalized extreme-value (GEV) distribution – which is a generalization of the extreme value type I CL assumption - that allows for alternatives within nests to be correlated. Let ρ_m denote the correlation in the nest m, and define the dissimilarity parameter be $\tau_m = (1 - \rho_m)^{1/2}$. The marginal distribution of each error is an extreme value type I.

The inclusive value for the mth nest corresponds to the expected value of the utility that decision maker t obtains by consuming an alternative in nest m. It is a rescaled measure of attractiveness of the nest $B(j)$. Denote this value by IV_m:

$$IV_m = ln \sum_{j \epsilon B_m} exp\left(\frac{V_k}{\tau_m}\right) \tag{22}$$

The marginal choice probability for alternative j, Pr_j, comprises two ratios. The first one shows the conditional probability of choosing j given some alternative in the nest. The second is the probability of choosing some alternative from nest k is a CL probability for the choice between the nests.

$$Pr_j = \frac{e^{\frac{V_j}{\tau(j)}}}{e^{IV(j)}} \cdot \frac{e^{\tau(j)IV(j)}}{\sum_{m=1}^{M} e^{\tau_m IV_m}} \tag{23}$$

This probability is the full information likelihood contribution. The CL model is a special case of the NML where $\tau_m = 1, \forall m = 1, \dots, M$. Also, the NMNL model is consistent with RUM if all τ_m lie in the unit interval.

Another possibility is the more generalized approach of the Cross nested multinomial logit model (CNL), we will follow the explanation of Bierlaire (2006). The probability of choosing alternative i within the choice set C of a given choice maker is:

$$P(i|C) = \frac{y_i \frac{\partial G}{\partial y_i}(y_1, \dots, y_J)}{\mu G(y_1, \dots, y_J)} \tag{24}$$

Where the J is the number of available alternatives, $y_i = e^{V_i}$, V_i is the deterministic part of the utility function associated with alternative i, and G is a non-negative differentiable function defined on R_+^J. A more familiar exposition of the same formula, given homogeneity of G and the Euler's theorem is:

$$P(i|C) = \frac{e^{V_i + lnG_i(\dots)}}{\sum_{j=1}^{J} e^{V_j + lnG_j(\dots)}} \tag{25}$$

Where $G_i = \frac{\partial G}{\partial y_i}$. There are various possible formulations depending on the researcher's objective, for example, Small (1987) studied departure time choice with a model called Ordered GEV and Vovsha (1997) applies the model to a mode choice application, where one of the mode belonged to two nests. Wen and Koppelman (2001) proposed a formulation of a "general" GEV model based on the following generating function:

$$G(y_1, \dots, y_J) = \sum_m \left(\sum_{n' \in N_m} (\alpha_{n'm} y_{n'})^{\frac{1}{\mu_m}} \right)^{\mu_m} \tag{26}$$

Where $\alpha_{n'm} \geq 0$ and $0 < \mu_m \leq 1$. The condition $\sum_m \alpha_{n'm} = 1 \; \forall \, n'$ provides useful interpretation of nest allocation. Also, Wen and Koppelman (2001) provide direct- and cross-elasticities formulae for the model.

2.2. EMPIRICAL STUDIES

This section explores some studies of the vast and increasingly richer mode choice literature. This field "begins" with the seminal paper of McFadden (1974a) as a first attempt to model discrete choice combining statistical analysis and economic behavior theory, since then the author has contributed with a wide range of theoretical and practical models. The empirical part has evolved from the bi-outcome logit, the Multinomial logit (MNL), the Nested logit (NL) to an even broader class of models known as generalized extreme value (GEV) model, from which can be derived all the previous models and even a fairly useful Cross-nested logit (CNL) model, in which the same alternative can belong to more than one nest.

As we said before, McFadden (1974a) gave the microeconomic foundation based on utility theory to the statistical model baptized as Conditional Logit (CL). Among other things, its first axiom of the Independence of Irrelevant Alternative (IIA) states that the odds of y being chosen over x, given the choice set C, depends solely on the characteristics of alternatives y and x, regardless of other alternatives – $z_1, z_2, \dots, z_n$ - in C. We will talk no further here about its properties because we already did it in the previous section.

McFadden (1974b) derives mathematically the commute mode choice based on microeconomic utility maximization. The author starts by stating the general individual decision as a utility maximization given a budget constraint problem, then moves to the population choice behavior or aggregated demand, from which he assumes there is a representative individual with J alternatives described by a vector of x_j attributes, the utility function is made up of a portion of deterministic part and a stochastic term that denotes the individual idiosyncrasies. The ending expression, after some distributional and equation forms, is a conditional logit model. The author uses data on 213 households from the San Francisco Bay Area before and after the introduction of a new rail system called the Bay Area Rapid Transit (BART). First, McFadden ran a binary logit for auto and bus choice and four model specifications (including family income, cost, time, walk time, bus waiting and transfer time and alternative specific constant). After the appropriate discussion of the results, McFadden discusses whether the 'unexplained residual' from the binary logit model was correlated with the independent variables. Regressing the residual estimates on the survey's omitted variables to see if any of them is significant and should be included in the model and also the possibility of simultaneity between mode choice and the independent variables. The author also calculates the BART patronage using the CL, Cascade and Maximum models given certain assumptions.

Ben-Akiva and Lerman (1974) develop the theory and estimation of a joint decision of car ownership and commute mode choice by using a single Multinomial logit (MNL) model. The way they overcome this problem is by simply multiplying each alternative in car ownership choice set by the counterpart of mode choice, generating a joint choice set of the form: (1) own zero autos and take transit; (2) own one auto and use car; (3) own one auto and use transit; (4) own two or more autos and use car; and (5) own two or more autos and use transit. This simplification led to a feasible statistical

estimation, however, there are two major problems with this approach: i) from a practical point of view, implicitly, they are assuming that the both choices are simultaneous, but it is more plausible to think that car ownership is a prior decision; and ii) it also had an econometrical problem regarding the IIA assumption, which can't be sustained under such model. The data consisted of 1968 Home Interview Survey performed in Washington D.C., observations were excluded when no work trips were made, chosen mode was neither transit nor car and intra-zonal work trips. Independent variables were alternative specific constants, number of auto per licensed drivers, household income, travel time, travel cost and a dummy variable for trips whose destination is the CBD. Value of in-vehicle travel time ranged from 8.59 to 11.14US$/h and out-of-vehicle from 39.16 to 58.87US$/h.

McFadden (1978) studies residential location and is the first author to derive the generalized extreme value (GEV) model consistent with economic choice theory, the author is particularly interested in two problems: first, how to model individual (disaggregate) choice among residential location; second, how to estimate such model when the number of elemental alternatives is impractically large. McFadden's insight is that some dwellings are sufficiently similar on their observed characteristics to be treated as different and independent options in the choice, instead dwelling units which are perceived as similar are aggregated, depending on the degree of similarity. At one extreme, the elements of the aggregate will be perceived as independent, and choice will be described by a Multinomial logit with individual dwellings as alternatives. At the other extreme, all dwellings with the same observed attributes will be perceived as virtually the same, and the choice will be described by a Multinomial logit with dwelling types. McFadden moves on to explain the theory of housing location choice, briefly, it depends on the attributes of the alternative, such as quality of public services, neighborhood and dwelling characteristics, as well as the consumer characteristics, such as age, family size, income etc. Then, the author derives the "usual" approach, which is to estimate two separate Multinomial logit models, first estimate the probability of choosing a community (neighborhood) *c*, using a vector of observed attributes which vary only with community, then estimate the probability of choosing some dwelling *n* given the community, using a vector of attributes which vary with both community and dwelling. This approach follows the IIA assumption and requires less computational effort; however, there is a loss of efficiency (relative to estimating both models

simultaneously). Finally, the author proceeds to derive the GEV, whose special cases are the Multinomial (with the inclusive coefficients equal to one) and the nested logit models (where each inclusive value must lie in the unit interval to be consistent with random utility theory). The GEV model allows for general pattern of dependence among unobserved attributes of alternatives and yields an analytically tractable closed form for the choice probabilities.

Train (1980) represents the joint probability of car ownership and commute mode choice as a two-step procedure, being the first is the conditional probability of a worker choosing a certain mode given his car ownership level, the second is the marginal probability of the household choosing a certain car ownership level over all modes. This assumption enabled the estimation processes by two independent MNL models: one for car ownership and another for commute mode choice. They run the model using data of the 1975 BART Survey.

Ben-Akiva and Lerman (1985) textbook is a milestone in the discrete choice literature, since it first compiles very didactically both theory and practice of mode choice analysis. For the non-specialist, is begins with some key statistical concepts and individual choice behavior theory. It moves to Binary and Multinomial logit models, the respective forecasting and testing techniques. The final part addresses the Nested logit and the systems of models, which include not only the mode of transportation decision, but also job and house location, car ownership, destination for non-work trips etc.

Small (1987) might be the first researcher to empirically estimate a GEV model, proposed by McFadden (1978). If there is proximate covariance, that is, a correlation among the unobserved random utility components for alternatives which are closer together on a natural ordering, the IIA might fail. One case is when the dependent variable is a discrete representation of an underlying continuous variable. The model proposed by Small requires only that for a fixed i, the correlation between random utility components for alternatives i and j be a non-increasing function of $|i-j|$. His ordered generalized extreme value (OGEV) model results from the function:

$$G(y_1, \dots, y_J) = \sum_{r=1}^{J+M} \left[\sum_{j \in B_r} w_{r-j} y_j^{1/\rho_r} \right]^{\rho_r} \tag{27}$$

Where M is a positive integer, ρ_r and w_m are constants satisfying $0 < \rho_r \leq 1\ (r = 1, \dots, J+M)$, $w_m \geq 0$, $\sum_{m=0}^{M} w_m = 1\ (m = 0, \dots, M)$ and where $B_r =$

$\{j \in \{1, \dots, J\} | r - M \leq j \leq r\}$. The difference between the OGEV and the NL model is that, in the first one, the subsets B_r overlap, that is, each $(J + M)$ subset contains up to $(M + 1)$ contiguous alternatives. Since each subset can have its own parameter ρ_r, this provides considerable flexibility to the correlation patterns, whose special case is when $\rho_r = 1 \; for \; all \; r$ it reduces to a MNL. The author also discusses an extended ordered GEV (EOGEV) model which further develops the OGEV. Small implements the four models (MNL, NL, OGEV and EOGEV) on two datasets, presented in Small (1982) and Train (1980), respectively; the first one is a sample of 527 commuters who must choose their work arrival time between 12 alternatives (ranging from 40min early to 15min late, with a 5min intervals); the second is a dataset consisting of commuter car ownership and mode choice decisions. The generalized models performed poorly compared to the MNL and NL models, exacerbating errors on misspecification and/or omitted variables.

Available data usually contains information on the actual or observed choice of the commuter, leaving the analyst the task of specifying the remaining not chosen alternatives or counterfactuals. A common practice is just ignoring this problem assuming that all decision-makers have the same choice set. However, Swait and Ben-Akiva (1987) used the MNL model as a benchmark comparing it with a Parametrized Captivity Logit, which incorporates a probabilistic component for generating the alternatives to be included in the choice sets: the model separates the choice makers who are captive to a single mode to the ones that really have a full choice set. The first stage of the estimation process consists of the choice set generation, the second is the actual mode choice. The model was calibrated for São Paulo, Brazil, using the 1977 Origin-Destination Survey's data.

Vovsha (1997) proposes a cross-nested logit model derived from the GEV class more suitable to deal with cross similarities between different pure and combined modes, allowing for differentiated measurement of pairwise similarities among modes. The area of study is Tel Aviv, Israel, which is based primarily on two modes, automobile and bus, however, planes are being made for the development of new, partly competitive and partly complementary, modes: suburban rail, light-rail transit and subway. Vovsha criticizes two problems with standard nesting structure (transit and automobile main nests, and further division of transit modes): (i) unambiguous identification of the main transit leg for each combined trip and (ii) separation of local

feeder modes from trunk. One example where these assumptions are violated are the park-and-ride facilities where commuters can drive and park their cars near (or even inside) trunk transit stations, usually outside the central congested area. The dataset comprises three surveys (1995 and 1996) for the morning peak hour travels in the Tel Aviv metropolitan area: mode composition is about 70% automobile, 29% bus and 1% rail, however, for the Netanya-Tel Aviv corridor rail service is developed, accounting for 10% of these trips. The deterministic portion of utility was comprised of mode-specific constants, travel time and travel cost. Car ownership first segmented the trip markets into three groups (car captives, transit captives and mode choosers), then "main" mode choices were automobile, bus, rail (three-legged, with bus/walking as first and third leg) and park-and-ride (three-legged, with automobile as first leg and bus/walking as third leg). Overall goodness of fit was relatively satisfactory and coefficients were significant, except the cost variable: this can be explained by the narrow range of automobile mode cost because of the absence of parking charges in observations and by the narrow range of transit fares within the Tel Aviv metropolitan area.

Bierlaire, Axhausen and Abbay (2001) explore 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 SwissMetro is a survey which intends to forecast the demand for a magnetic levitation underground system operating at speeds up to 500km/h in partial vacuum connecting major Swiss conurbations. All the models were estimated using the first author's software Biogeme, which can handles any type of GEV model. Contrary to most analysis whose data consists of stated preference (SP), this study is based on revealed preference (RP) sample road (770) and rail (435) users, which can be affected by the new mode. The variables in the utility function are the same for all modes: alternative specific constants, cost, time, age and for transit the frequency, presence of luggage, seat configuration and the possession of annual season ticket. Looking at the log-likelihood value, there was a significant improvement from the MNL to the NL model, and from the latter to the CNL model; the value of time decreased across the MNL-NL-CNL models, from 1.2 to 1.15 to 1.12 sFr/min.

Recent studies have been focusing increasingly in the NL models rather than the MNL ones, because the first model is more flexible regarding the IIA assumption: the

choices in the same "nest" are allowed to be correlated, but not between nests. For example, De Palma and Rochat (2000) estimated a NL model for the commute mode choice in Geneva in which the first level the household chose between one or more than one car (auto ownership) and at the bottom level it chose between private or transit (mode choice).

Dissanayake and Morikawa (2010) had an interesting methodological approach for using both revealed and stated preference data. The stated preference data showed the mode choices of transports already available, on the other side, the revealed preference was used for the Mass Rapid Transit System in Bangkok that would be implemented in the following years. They estimated two models, the first, the benchmark, is a NL with the RP, and the second is a NL with both RP and ST data, both of them had car ownership as the first level and mode choice at the bottom level. The second model performed better, showing more individually significant variables and a scale coefficient smaller than one and significant, according to the expectations.

In regards to pricing policies and mode choice, Washbrook, Haider and Jaccard (2006) applied a questionnaire for people who currently drove alone to their workplaces. They chose between drive alone, carpooling or a hypothetical express bus when choices varied in terms of time and cost attributed. Using this data, they calibrated a Conditional Logit model for the Greater Vancouver suburb to estimate commuters' responsiveness to various policies. They found that the most effective way to disincentive single occupant vehicle choices is raising its cost, in other words, charging users as a road pricing strategy.

Vrtic *et al*. (2010) analyzed a revealed preference survey for four different choices: one for political acceptability of road pricing and the other three (called behavioral experiments) for joint car route and departure time choice (RDC), mode and departure time choice (MDC) and the third with mode and route choice (MRC), all in the presence of road pricing. The questionnaires had 1,005 respondents from a sample of 2,290. They estimated a MNL model based on the joint datasets from the RDC, MDC and MRC experiments showed that, among other things, travelers resent more easily avoidable costs more, such as parking that can be reduced by longer walking times and tolls that can be lessened by different routes and also that the Swiss commuters dislike earlier departure times, however, if choose to drive on a congested non tolled route, then they prefer to depart earlier to make sure that they arrive on time or earlier.

The three following papers are quite similar in various aspects. Besides all the authors being from the São Paulo University, all used the same data source – the 2007 Metrô Survey; they all included work and education trips, running a different model for each purpose; they employed the same empirical methodology, multinomial and mixed logit, and used the same independent variables (cost, time, income, age, sex, household size and two dummy-variables for study and employment status). Also, they calculated the travel cost (unobserved in the original Survey) exactly the same way.

Lucinda, Meyer and Ledo (2013) study mode choice in the Metropolitan Region of São Paulo in order to estimate the welfare and traffic effects of a congestion charge. First, the authors discuss the experience of congestion charge in Singapore, London and Stockholm and the current transport scenario in São Paulo. Then, they discuss the 2007 Metrô Survey, presenting graphics about the departure/arrival time, mode share, mode share for poor families and the importance of the Expanded Center region (São Paulo's CBD). They show their calculation for the unobserved travel cost and also how they estimated the counterfactual trips (what would have been the travel time and cost for the mode snot chosen): they run two separate equations with a dependent variables as the logarithm of the observed travel time and cost on dummy-variables for the departure and arrival times, trip purpose and distance; they only restricted train and subway for some individuals' choice set. In the mixed logit, they found a VoTT of R$ 6.88 for work-trips and also estimated that an expansion of the rotation scheme would lead to a lower welfare loss compared to the congestion charge.

Barcellos (2014) studies mode choice in the Metropolitan Region of São Paulo in order to estimate the effect of a Tax over Fuel (the Fuels' Contribution of Intervention in the Economic Domain or CIDE in Portuguese) on the probability of the travel mode choice. Using the 2007 Metrô data and the mixed logit model results, the simulation considers a tax of R$ 0.1 and R$ 0.5, in the first case, the automobile actually increases its mode share reducing transit's share, whereas in the second case, both transit and automobile increase their share at the expanse of the non-motorized modes.

Pacheco and Chagas (2015) analyzed mode choice of the Metropolitan Region of São Paulo using the Metrô's 2007 Survey. Firstly, the authors discuss the role and characteristics of public roads and the difference between private and social costs. They estimated a multinomial mixed-effects logit model (it contains both a fixed effects and

random effects, also known as random coefficients) with travel time and travel cost as well as socioeconomic characteristics of the individual as explanatory variables (income, age, sex, if the person studies and if the person is employed). They grouped the seventeen mode choices into six categories: non-motorized (walking and bicycle); buses; rail (subway and train); motorcycle; automobile (driver and passenger); and taxi. They estimated the same model (specification) for three different purposes of travel, first using the whole sample, second with work at the origin or destination and thirdly with education at the origin or destination. The authors then move to a discussion of a flat congestion toll and its impact on the demand based on the derivatives of the cost variables.

3. EMPIRICAL STRATEGY AND DATA

3.1. 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[8] (Metrô) in 2007 and 2012. Besides that, we also collected monthly fuel prices from the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis[9] (ANP) and yearly vehicle mileage from the Instituto Nacional de Metrologia, Qualidade e Tecnologia[10] (INMETRO). In this section, we will explain in detail each dataset, especially those from Metrô.

The surveys done by Metrô are traditional sources of transport information about the Metropolitan Region of São Paulo (MRSP), Brazil. It was first implemented in 1967 for the development of studies and projects for the initial subway network. Since then, the survey has been carried out every 10 years (1977, 1987, 1997 and 2007), but since 1997, intermediate surveys (2002 and 2012) are being executed to keep up with fast changing dynamics of the city. Its primary goal is to identify the main daily trips of people, by reason and mode of transportation.

We used the transport data from Metrô for several reasons. It is a complete dataset, with socio-economic and transport specific information, it is well documented and transparent (it contains the Household Research Manual, which explains each variable in detail, the Code Manual, which contains the employment codes, the Zone Matching file, which provides information about how the traffic zone codes changed from previous questionnaires) and it is statistically meaningful, it has a relatively large sample size, statistical weights and it is provided both as a *dBase* and a *SAV* file extension. Specifically, these two surveys, 2007 and 2012, were chosen for two main reasons. They are the most recent ones; therefore, they better portray the present reality of the region. Second, they are public, free-access and readily available for anyone wishing to download it, in Metrô (2007) and Metrô (2012).

[8] Company of the Metropolitan of São Paulo. Website: http://www.metro.sp.gov.br/
[9] National Agency of Petroleum, Natural Gas and Biofuels. Website: http://www.anp.gov.br/
[10] National Institute of Metrology, Quality and Technology. Website: http://www.inmetro.gov.br/

Both 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.

The official name of the 2007 Survey is Pesquisa Origem e Destino 2007 (here we will refer to it simply as *2*007 Survey). It raised information on 30 thousand randomly chosen households and 120 thousand individuals distributed over 460 Traffic Analysis Zones (TAZ), being 320 TAZ in the São Paulo municipality. The official name of the 2012 Survey is Pesquisa de Mobilidade da Região Metropolitana de São Paulo 2012 (here we will refer to it simply as *2*012 Survey). It raised information on 8.1 thousand households and 32.4 thousand people distributed over 31 TAZ.

It defines a series of concepts to standardize empirical analysis. **Household** is the place of living of one or more families, a group or a single individual, it is limited by a wall and covered by a roof; a **dweller** is a family member, people who live in the household most of the week, students and employees; a **non-dweller** is a military, student or worker who doesn't remain most of the week in the household, people who are there only visiting, traveling or doing business; a **qualified respondent** is every dweller in the household older than ten years old and capable of answering the questions (the head of the house might answer for the incapable people older than ten), however, children younger than ten who go alone to school or work can answer; and a **trip** is a movement between two points (an origin and a destination) with a defined reason, using one or more modes of transport, it is considered that the trip took place between 4 a.m. of a given day and 3:59 a.m. of the following day (24h period), the information about the trips are preferably about the previous day of the first visit to the household and must be a workday (Monday to Friday) (METRÔ, 2007b; METRÔ, 2012b).

The 2007 and 2012 surveys consider an exhaustive (the last mode is "others", that is, any mode which doesn't fit in any of the previous categories) mode choice set of

seventeen categories: 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.

The walking mode has some special features. It is only considered as a sole mode when it is the only mode in the whole journey, that is, it never appears with another mode. Trips whose reason is either work or school must always be recorded, regardless of the distance. For other reasons, walking trips are only registered if the distance is over 5 blocks (approximately 500m).

We collected monthly fuel information for each Brazilian State from the ANP, as previously discussed. The fuels surveyed are hydrous ethanol, regular gasoline (we will refer to it simply as gasoline), liquefied/liquid petroleum gas (LPG or commonly referred to as propane or butane), compressed natural gas (CNG), diesel fuel and diesel fuel S10[11]. They provide information about the number of stations surveyed, unit of measurement, average resale price, standard deviation of the resale price, minimum/maximum resale price, average resale margin, and the previous information about the distribution price as well (ANP, 2012). We merged this information into the Metrô's dataset according to the year and month variables presented in both sources.

We 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 (which ranges from A as the lower consumption and E as the higher consumption) from 2009 to 2012. From 2013 to 2016, the INMETRO added the following information: exhaust emissions (non-methane hydrocarbons, carbon monoxide, nitrogen oxides, and carbon dioxide), energetic consumption (in MJ/km), a new PBE 2013 classification and the presence of the Programa Nacional da Racionalização do Uso dos Derivados do Petróleo e Gás Natural (CONPET) Seal of Energetic Efficiency 2013.

[11] Diesel fuel with a 10 mg/kg maximum sulfur limit, as defined by the National Council on the Environment, CONAMA, issued resolution 403/2008.

3.2. EMPIRICAL STRATEGY

This section covers all empirical procedures of the work. Since the data pre-processing, travel time estimation, cost calculation and model specification. All the following data adjustments and statistical analysis were carried out using the software Stata 11.2/SE - StataCorp (2009a) – unless otherwise stated and we shall follow their terminology throughout the work. All Stata code will be fully available at a public repository called "Msc. thesis" in the author's GitHub page (https://github.com/jayme-anchante).

3.2.1. Data pre-processing

Both 2007 and 2012 Metrô databases were appended totalizing 250,203 observations (196,698 from 2007 and 53,505 from 2012). For practical reasons, some changes were made. Some variables that appeared only in 2007 were excluded: "condmora" (housing condition: owned, rented, ceded or other); "qt_micro" (number of microcomputers in the household); "pag_viag" (who paid for the trip: oneself/family, employer, exempted or other); "tp_esauto" (the kind of parking for automobile: didn't park, paid public parking, sponsored, own parking lot, public free parking, one-time parking or monthly parking); "vl_est" (the value paid for the parking); "pe_bici" (why have you traveled on foot or on bicycle: short distance, other modes were expensive, far bus stop, infrequent transit, long trip, crowded transit, physical activity or other reasons); and "tp_esbici" (the kind of parking for bicycle: free bicycle parking, paid bicycle parking, private place, public place or other). Also, the variable "total individual income" can take up to different values depending on the year: six-digit for 2012 and eight-digit for 2007, however, no income higher than 6-digit appears in either survey. A proper day, month and year variables were created based on the "data" variable (the day the interview took place in the household).

For the next paragraphs, we numbers in parenthesis are the total number of observations excluded because of the stated reason. We checked the dataset to find

possible logical inconsistencies and/or typing errors. Some observations (148) were dated prior to 2007, from 1982 and 1999 and some didn't show any date at all (175). The day the trip took place took values from "2" (Monday) to "6" (Friday); however, values "0" were found (457). People could travel from or to outside the researched area (MRSP) and it is foreseen in the Survey Manual (METRÔ, 2007b; METRÔ, 2012b). The code of the workplace municipality should range from 1 to 39, according to the Code Manual, however values "99" appeared as first work (570) and second work (25). Some observations also didn't record information on distances traveled (194).

For methodological reasons, some changes were made. People who had no recorded trips were excluded (33,427). This is due to the tautology of the problem: one cannot study trips that didn't happen. Again, we kept observations of people who went from home to work (home-work based trips) as trip motives, leading us to exclude 170,097 observations. There are several reasons for this narrowing: i) basically all works on the mode choice literature use a dataset consisting of morning commute trips; ii) other-purpose trips have a different set of explanatory variables, such as convenience, destination characteristics, etc; iii) and from the point of view of the public policy, other-purpose trips are not as important as home-work trips, because each purpose accounts for a small portion of traffic and those trips are more flexible, especially the departure/arrival time and location (on the contrary, most jobs start at around 8 a.m. and are located in the Central Business Center). We also kept only the trips that begin and end in the São Paulo municipality (12,853 observations dropped), because it is not feasible to estimate intercity bus fare, which is better explained later on, when we discuss how we created the cost variable.

Other exclusions include: persons which choose the "other" category for mode of transportation (41), since we cannot precise what mode it really is; persons which choose buses from municipalities other than São Paulo (6); persons which choose microbuses from municipalities other than São Paulo (7); persons which choose intercity microbuses (10); households which did not declare their comfort items (29); if the number of automobiles in the household ("qt_auto" is the name of the variable in the dataset) is a missing (481); if the number of bicycles in the household ("qt_bicicle" is the name of the variable in the dataset) is a missing (1); if the Traffic Analysis Zone (TAZ) of the first work is a missing (1260). We excluded extreme values (even though they might represent the behavior of the population), because those values were thought

to be erroneous and/or far from the reality in some form. For example, non-motorized speeds over 200 km/h or general commutes longer than 10h, our arbitrarily chosen measure were values lower than the 5th and higher than 95th percentiles of distance, time and speed variables for each mode, which resulted in 3,708 observations excluded. The final dataset consisted of 26,797 observations, 21,186 belonging to the 2007 Metrô Survey and 5,611 belonging to the 2012 Survey.

3.2.2. Estimating the urban gasoline mileage based on the automobile's year of manufacture

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[12]. There is no other information regarding the category, brand, model, type of engine, transmission and other characteristics of the vehicle which could ultimately lead to an accurate estimation of the vehicle mileage. Moreover, we can't tell what is the year of the automobile used on the trip made by this mode, so we will assume that the trip is always made in the automobile stated in the variable whose dictionary name is "ANO_AUTO1" which is the "year of manufacture – Auto 1" or the vehicle with the latest manufacture year.

Furthermore, we collected data from the INMETRO website which provides official government information on vehicle mileage for gasoline/diesel and hydrous ethanol in rural and urban areas (we are going to use solely gasoline mileage in urban areas). However, information is only available for vehicles manufactured between 2009 and 2016. We downloaded the available archives as Portable Document Format (*.pdf* extension) and using the services of two websites[13] we converted the files to Microsoft Excel Worksheet (*.xlsx* extension), manually appending and correcting the spreadsheets. Since the volume of information was enormous, we selected only vehicles whose category was either Subcompact (defined as passenger vehicles with area until 6.5 m^2, +/- 0.1 m^2) or Compact (defined as passenger vehicles with area from 6.5 to 7 m^2, +/-

12 There is no information about the criteria for defining which are these three vehicles and in which order they are going to be reported, that is, if it is ordered by intensity of use, date of acquisition, date of manufacture, price or other. However, a simple inspection reveals that it follows a decreasing order of the year of manufacture, that is, the first car to be reported is the latest manufacture year.

13 Nitro software online (https://www.pdftoexcelonline.com/en/) for pdfs from 2009 to 2012 and pdftoexcel.com (http://www.pdftoexcel.com/) for pdfs from 2013 to 2016.

0.1 m^2) as proxies for an average commuting vehicle. We could use each year'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 becomes lighter and other aspects are improved, implicitly "time is making mileage increases".

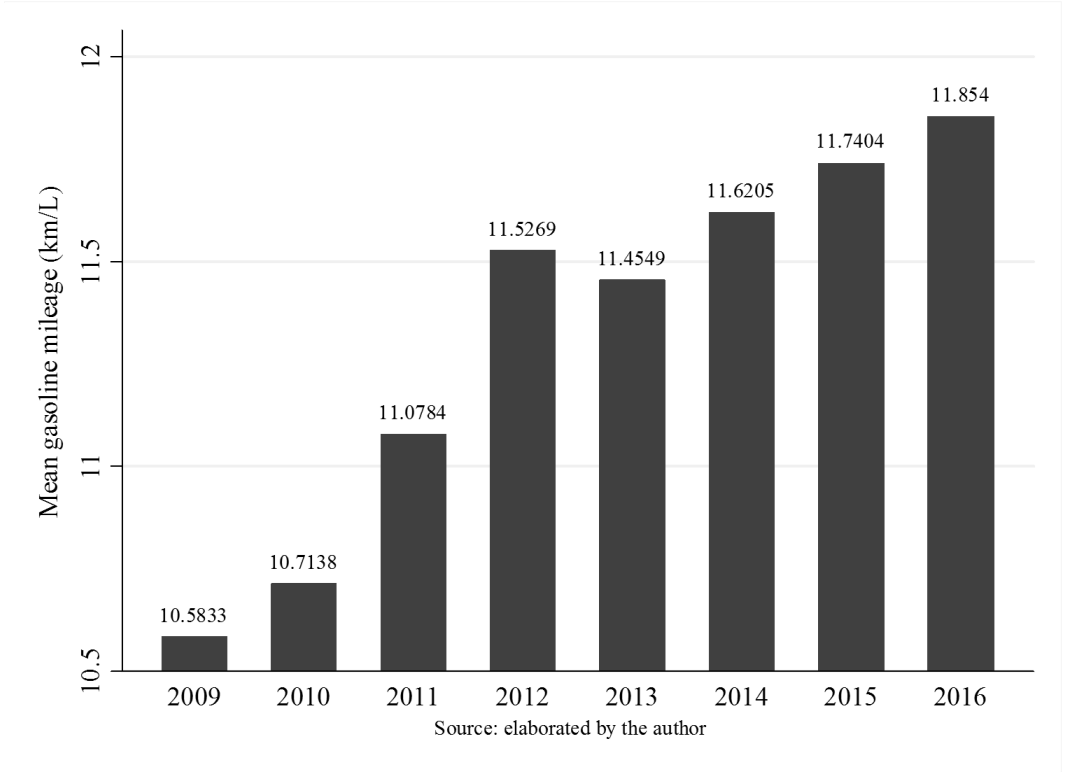


Figure 5: Mean gasoline mileage over year (2009-2016), sub-compact and compact vehicle categories

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[14] 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.

[14] 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.

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

VARIABLES	Gasoline mileage (city)
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 the table above, the variables were significant (except one of the assisted steering dummies). The model's goodness-of-fit represented by the coefficient of determination (R-squared) is relatively high, that is, 71% of the variance of the dependent variable, gasoline mileage, is explained by the model. 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.

3.2.3. Counterfactual travel times estimation

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. The following two tables exemplify what we are explaining. Table 3 displays three individuals, the first one choose private mode, the second transit and the third a non-motorized mode, each one with their own characteristics. However, we should note that the time and cost variables are attributes of the alternatives, whereas sex and age are characteristics of the individuals. Table 4 displays three rows *per* individual, each row shows a different mode choice, noting that each individual chose the same mode as the previous table (given by the Choice column), for the same individual, each mode has a different time and cost, nonetheless, sex and age (as well as other socio-demographic characteristics) remain constant for the same individual.

Table 3: Observed Metrô's data structure

Person_id	Mode	Time	Cost	Sex	Age
1	Private	20	5	Male	30
2	Transit	45	3	Female	25
3	Non-motorized	30	0	Male	20

Source: elaborated by the author

Table 4: Observed Metrô's data structure with mode choice counterfactuals

Person_id	Mode	Choice	Time	Cost	Sex	Age
1	Private	1	20	5	Male	30
1	Transit	0	30	2	Male	30
1	Non-motorized	0	45	0	Male	30
2	Private	0	15	2	Female	25
2	Transit	1	45	3	Female	25
2	Non-motorized	0	60	0	Female	25
3	Private	0	20	4	Male	20
3	Transit	0	25	3	Male	20
3	Non-motorized	1	30	0	Male	20

Source: elaborated by the author

Our next task is "creating" the counterfactual mode choices. 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.

We initially made a visual inspection of histograms for each mode (transit, private and non-motorized), as well as distributions between different modes intra-category, also varying between Surveys and other variables. We also performed various multivariate tests of means to see whether mean speed differed between transit, private and non-motorized options (intra-category differences). For transit options, we rejected the hypothesis that they all have the same sample mean (assuming homogeneity and allowing for heterogeneity), however, since estimating different travel times for each transit option would require to know if the person lives near a bus, a subway and/or a train station, we will assume that all transit options are part of a "unified" system of transit, which have the same mean speed, and all person have access to that system. For private options, we rejected the hypothesis that all sample means are the same; however, if we exclude "motorcycle" from the test, we didn't reject the hypothesis, since it holds a fairly small mode share among private modes, we accepted the hypothesis that the sample mean for private modes are the same. For non-motorized modes, we rejected the

hypothesis that walking and cycling have the same sample mean, but we did not reject the hypothesis that walking/cycling mean speed differ from the two Surveys, since we cannot observe the person's ability of cycling, we are assuming that everyone (has the ability of walking and) chooses walking as non-motorized mode.

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. The following table displays the regression of transit speed on explanatory variables for the 2007 and 2012 Surveys. We can see that age and sex variables had the same sign as the non-motorized speed regression, which is the expected, as people get older they tend to walk slower and also males walk faster than females. For the weekday dummies, all days seem to affect positively the speed compared to Monday and this relationship is even stronger on Friday. The latter the departure hour, the faster the transit speed, contrary to what could be expected (latter departures might face more congestion of the rush hour); however, a possible explanation might be the fact that commutes by transit tend leave home earlier (because transit is less reliable than a motor private mode, for example), and the earlier in the day, the lower the public transportation fleet provision (and higher the time waiting). All variables were individually significant and the R-squared was higher than we expected, since we are estimating transit speed without a proper established theory.

Table 5: 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. The following table displays the results of the

regression of transit speed on the previous explanatory variables for the 2007 and 2012 Surveys. Age had an inverse relationship with private speed whereas males, on average, had a higher speed than females, which is our *a priori* expectation, since elder people might be more conservative than young people in this respect and males might assume a riskier driving behavior than females. The number of family dwellers also showed a negative relationship with private speed; people who live alone might have a riskier behavior than parents with children or couples living with other relatives. Income had a positive relationship with speed, we think about income being positively related to more powerful (engine power) cars and a household located near major (and faster) highways. All employment relationship had a lower speed than the formal contract (base category), it might be the case that formal contractors could have a stiffer arriving time. The weekday dummies had varies relationships with the base category, Monday. The departure time had a negative relationship with speed, this could mean that latter departure might face more congestion. All variables were individually significant and overall goodness of fit was higher than expected.

Table 6: 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

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

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.

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 7: 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

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

Estimating the counterfactuals also implies that we are estimating the individual's choice set and most likely not all alternatives might be available to all individuals. For transit and private counterfactuals, we reasoned that if the distance is too small, the commuter would never choose to go on these modes. We calculated the 1th percentile of the distance for the observed commuters by transit (1.72km) and private (0.63km) modes, and excluded those counterfactual alternatives for individuals whose distances were smaller than those values, which resulted in dropping 7,009 observations for transit counterfactuals and 3,226 observations for private counterfactuals. We used the same reasoning for non-motorized counterfactuals, but thinking that if the distance was too large, the commuter would never choose this mode. We calculated the 99th percentile of the distance for the observed commuters by non-motorized modes (2.78km) and excluded these counterfactual alternatives for individuals whose distances were bigger than this value, which resulted in dropping 17,057 observations for non-motorized counterfactuals.

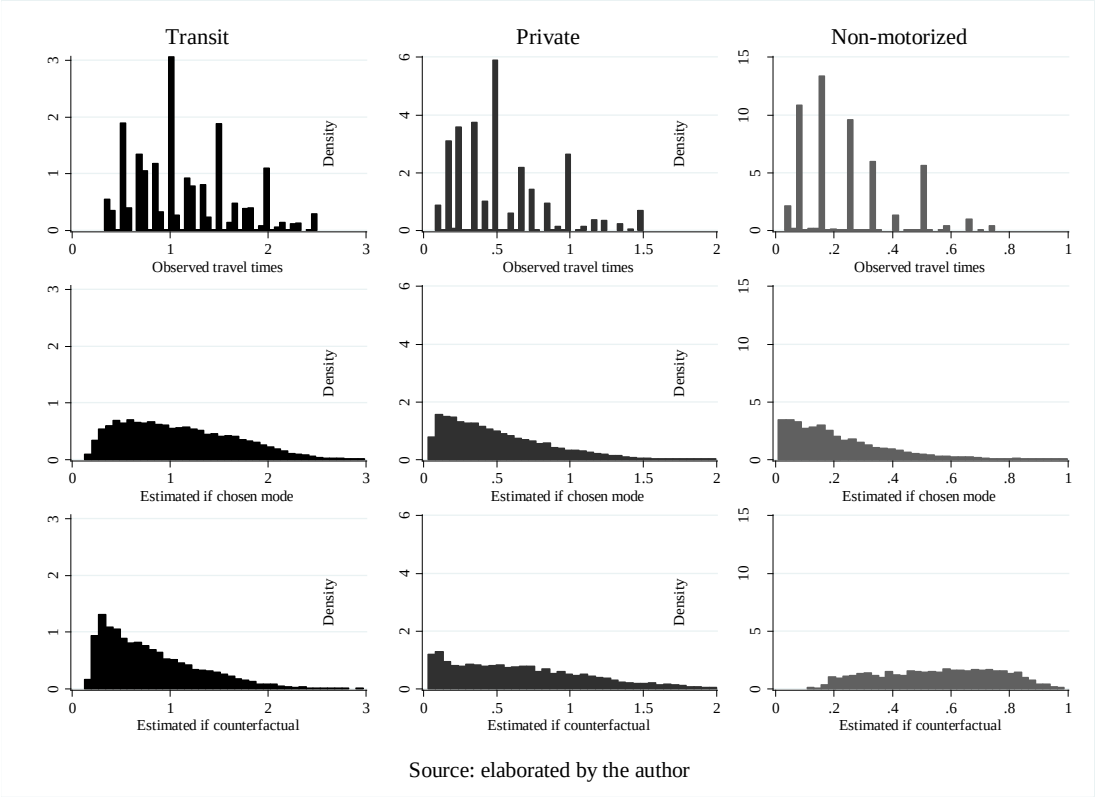
Continuing about the choice set generation; we assumed that every commuter had non-motorized and transit as a counterfactual (after excluding the extreme distance impediments). However, for the private counterfactual we had to look the household's vehicles to see whether a private mode is feasible. We assumed that commuters whose household had at least one automobile could and would commute by auto as a counterfactual. We also assumed that commuters whose household had at least one motorcycle could and would commute by motorcycle, if they haven't already chosen auto as counterfactual. Commuters whose household did not have neither automobiles

nor motorcycles could, in theory, commute by taxi; however, this may not be feasible, since in the original dataset we only observe 148 taxi commutes or 1.44% of the private mode share, which is much less than the 6,653 commuters whose household does not have neither an auto nor a motorcycle. Keeping the private share equal to the observed data, we should have around 192 taxi trips as counterfactuals, since we cannot estimate the probability of choosing taxi over other private modes, we randomly assigned 192 taxi trips for commuters whose household does not have neither an auto nor a motorcycle as draws from a binomial distribution with one trial (*n* equal to one) and probability of success, *p*, equal to 0.03. We excluded the private counterfactuals for commuters whose household did not have neither automobiles nor motorcycles and the ones that weren't assigned the "possibility" of commuting by taxi, resulting in the exclusion of 5,451 counterfactuals.

After estimating each mode's speed (in the case of transit and private categories for different Surveys as well), we divided the observed (reported) distance by the estimated speed for each individual, generating a new variable for the travel time. As already discussed previously, since observed travel times tend to be rounded up in sharp times (because of reported departure/arrival times), we decided to use the estimated travel time not only for the counterfactual travel times, but also for the actually chosen modes, because our estimated turned out to be more evenly distributed around those sharp times, with spikes less accentuated. We did not follow the previous general rule for two cases: the speed differences between automobiles and motorcycles and walking and bicycling. In the observed data, motorcycles had a mean speed 3.89km/h higher than automobiles whereas cycling had a mean speed 2.31km/h higher than walking. We already discussed who commuted by motorcycle as counterfactual, a different and harder task would be who would cycle rather than walk in their non-motorized counterfactual: we can observe whether the commuter has at least one bicycle in their household, however we cannot observe whether he or she has the ability to pedal. In the observed dataset, we had 201 people commuting by bicycle or 3.5% of the non-motorized mode share; an equivalent share in the non-motorized counterfactual would be around 140 cycle commutes. As in the case of taxi, we randomly assigned the possibility of commuters who had at least one bicycle in their households to cycle as draws from a binomial distribution with one trial (*n* equal to one) and probability of success, *p*, equal to 0.08. Motorcyclists and cyclists had an estimated travel time equal

to their distance divided by the estimated speed plus 3.89 km/h for motorcyclists or plus 2.31 km/h for cyclists.

The next figure 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 6: Observed and estimated sample distribution of travel times by mode

3.2.4. Observed and counterfactual travel cost estimation

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 8: 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 (24)	Equation (24)
Passenger of automobile	Equation (24)	Equation (24)
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 (24)]/2	[Equation (24)]/2
Bicycle	0	0
Walking	0	0

Source: elaborated by the author

There were two buses mode choices: bus from the São Paulo municipality and chartered bus. The bus from the São Paulo municipality fare is stipulated in R$ 2.3 for 2007/2008 and R$ 3 for 2012, (SPTRANS, 2017), it should be noted that students only pay half fare (this is valid for all forms of public transportation, subway and train included) as defined by the local law, (SPTRANS, 2017). For buses from other municipalities and inter-city buses we first checked the Empresa Metropolitana de Transportes Urbanos de São Paulo S. A. website (EMTU, 2017) to retrieve the intercity bus fare. There were 6,150 different fares depending on your consortium, type of service, line, section/integration, extension and road toll for the year of 2017 and the Metropolitan Region of São Paulo. Added to this, the fact that there is no exact information about the route or the coordinates of the origin/destination and that there is no fare information prior 2015, it is infeasible to estimate the fare cost for intercity travels. Since this mode accounts for approximately 13% of intercity travels, we decided to keep in the analysis only trips whose origin and destination are the municipality of São Paulo (after doing this, it still remained 6 observations which used buses from other municipalities, we excluded them). We also assumed that the cost for chartered bus is the same as the public buses.

Automobile cost was calculated according to the equation 24. It is composed solely of fuel expenses, which may be an oversimplification, because we did not

account for other important costs like insurance, depreciation, parking, maintenance and taxes, however, drivers only perceive their marginal cost for short term decisions, which is fuel cost, once all other costs are paid (most occur only once at the begin/end of each year). So, auto cost is the ratio between the distance (straight line between origin and destination, in kilometers) and mileage (in kilometers per liter of gasoline), multiplied the price of one liter of gasoline in the month the trip took place (which is an information available at the Metrô's Surveys), prices collected in the ANP (2012). Mileage is not observed in the Metrô dataset, instead we estimated its average by the year of manufacture. Since motorcycle year of manufacture was not observed either, we assumed that it was the median of the year of auto manufacture for each Survey, which is 2002 for the 2007 Survey and 2007 for the 2012 Survey, and that motorcycle mileage is twice the value of auto mileage in the same years.

$$Auto\ cost = \left(\frac{distance(km)}{(mileage(km/l)}\right) \times p_g(R\$/l) \tag{28}$$

Taxi fare is defined by law, as a fixed R$ 3.5 plus "2.1" times the distance traveled, São Paulo (2006), and as a fixed R$ 4.1 plus "2.5" times the distance, São Paulo (2010). Motorcycle cost was also estimated using equation 24, however, it is half the auto cost, because we are assuming the motorcycles are around twice as fuel efficient as automobiles.

For the microbuses, we are also assuming that they follow the same law as the public buses. Subway and train fares are also defined by the same law as buses and they have the same tariffs. Thus, we are assuming that all transit alternatives have the same cost, according to the Survey and if the commuter is a student or not.

Finally, cycling and walking pose no monetary cost to the commuter, so their cost is assumed to be zero. If we think marginally, as we did in the auto cost, this may be true; however, the global cost of walking and cycling is certainly not zero, if walking the person needs special clothing for some kinds of weathers or an umbrella if it is raining and if cycling the person needs a bicycle, which is also may need some kind of maintenance, parking facilities and proper clothing and accessories.

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 9: 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 10: 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

3.2.5. Model specification

First of all, let us define our dependent variable, the choice of mode of transport. 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 Table 11.

We turned the seventeen modes of transport into three broad categories: **Transit** (1), **Private** (2) and **Non-motorized** (3), which is demonstrated in the second column of Table 11. This aggregation followed our subjective judgment based on the modes' inherent characteristics. Transit comprises all public-access mass transportation whose itinerary is not fully under the user control, namely bus from the São Paulo municipality, chartered bus, microbus from the São Paulo municipality, subway and train. Private includes all forms of motorized private means: driving a car, passenger of a car, taxi and motorcycle. Non-motorized, also called active transportation, are all means that require active participation of the traveler: walk and bicycle. Taxi is perhaps the most hard to label, it could be both private or transit. Its price is established by law and it is heavily regulated as other transit options; however, we opted to mark it as private to its cost structure similar to the other private means and also because the user has full control over the itinerary of the travel. The cost structure reason applies generally to all classifications, all transit options share (or were "forced" to share) the

same cost structure as the private and non-motorized. Modes classified in none of the three categories are bus and microbus from other municipalities, intercity bus and microbus for reasons already discussed in the section 3.2.4., school vehicle resulted in no trips because we are study commute trips only and the others category is excluded because we can't specify exactly what mode of transportation it is and because it has a negligible market share.

Table 11: Specification for the dependent variable mode of transportation choice

Modes of transportation	**Mode**
Bus from the São Paulo municipality	(1)
Bus from other municipalities	-
Intercity bus	-
Chartered bus	(1)
School vehicle	-
Driving a car	(2)
Passenger of a car	(2)
Taxi (cab)	(2)
Microbus from the São Paulo municipality	(1)
Microbus from other municipality	-
Intercity microbus	-
Subway (tube)	(1)
Train (rail)	(1)
Motorcycle	(2)
Bicycle	(3)
Walking	(3)
Others	-

Source: elaborated by the author

The following table 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, sex, whether he studies, the employment relationship and the education degree. These variable were chosen both from a theoretic perspective, since they are usually used as control variables in most empirical works, and from a practical point of view, they are the variables we have available in the Metrô microdata that are both exogenous and probably correlated with the mode choice.

Table 12: 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

We didn't include individual income for practical reasons: there were a lot of missings in the data. However, we could include the household income in the model, but we didn't do it for theoretical and econometric reasons. It is well accepted that there is a strong relationship between auto ownership and mode choice; however, one can't directly include the number of autos in the household as an independent variable because of its strong endogeneity. Before the emerge of the generalized extreme value

models, several proposals have been put forward to try to work around this problem, for example Ben-Akiva and Lerman (1974) and Train (1980), including household income as a regressor would have the same problems as auto ownership, as extensively discussed in the previous papers and in the literature. Also, McFadden (1981) formulates a model in which income doesn't influence mode choice since it cancels out when the conditional indirect utility functions of different modes are compared; Small and Rosen (1981) derive the same results by analyzing user benefits within the discrete choice framework. The inclusion of income is empirically advocated by Jara-Díaz and Videla (1989), although they point out that we can't recognize it as an exogenous variable, but rather representing other phenomenon (taste and habits).

The next chapter presents the work's results, its interpretation and discussion with the literature. First, we show the alternative-specific conditional logit models for the 2007 and the 2012 Surveys and the tests on the IIA assumption. We also present two possible nesting structures for the nested multinomial logit model.

5. RESULTS

5.1. ALTERNATIVE-SPECIFIC CONDITIONAL LOGIT MODEL

We first ran two Conditional logit models with case variables (commuter characteristics) for the 2007 and 2012 Surveys. This model is also called Alternative-specific conditional logit in the Stata software and it is also known as McFadden's choice model in honor to its contributions to the logit models. In the "pure" Multinomial logit, all the coefficients are the individual characteristics (known as case variables) and refer to the base alternative (just as when we are interpreting dummy variables). In the other side of the spectrum, in the "pure" Conditional logit, all the coefficients are the alternative characteristics (alternative variables) and the coefficients are the same for all alternatives, it is often called by social scientists as fixed-effect logistic regression because the choices of the same individual are clustered together. The McFadden's choice model is a hybrid because it is composed of both the individual and the alternative variables, in practice one could ran this model within the Conditional logit framework by interacting the case variables with J-1 outcome dummy-variables and setting the J*th* alternative as the base outcome.

Table 15 presents the Alternative-specific conditional logit model for the 2007 and 2012 Survey. The alternative-specific variables are travel cost and travel time and the case variables are sex, age, study, employment relationship and education degree. The first column shows the variable names; the next three columns show the estimates for the 2007 Survey; and the last three columns show the estimates for the 2012 Survey. At the bottom, we have general information about the estimation, the number of observations that we used to run the model, the number of cases or individuals, cases dropped and the value of time, we will detail further the last two. Some cases were dropped or excluded from the model, because they only had one mode in their choice set. Since we are estimating a Conditional logit, the choice needs to be conditional on some choice set, if the choice set possesses a single option, then it is infeasible to include this individual in the model.

In our specification, the ratio between the travel time and cost coefficients is known in the literature as the Value of Travel Time Savings or simply the value of time. This is the amount of money people are willing to spend to save a certain amount of time (in our case, one hour) of their commute. We found a value of time of R$ 1.78/h and R$ 1.09/h, for 2007 and 2012 respectively. The monthly wage (nominal values) of our weighted sample is R$ 1,326.8 and R$ 1,359.18, for 2007 and 2012 respectively; this results in an hourly wage-rate[15] of R$ 7.65 and R$ 7.84. In his paper about cost-benefit analysis, Litman (2008), along with other institutions like the U. S. Department of Transportation, the Australian Bureau of Transport Economics and the UK Department for Transport, among others, suggests that the value of time should be around 35% of the hourly-wage. Our estimates lie between 23% and 14%, for 2007 and 2012 respectively, which is a little lower than expected, especially for 2012, this might be explained by the fact that the travel time coefficient was much higher than that of 2007, it wasn't statistically significant and that the 2012 sample was much smaller than the 2007. Although the next three studies all used the 2007 Metrô survey and almost identical methodologies, they found very different commute values of time: Barcellos (2014) ran two models - a Multinomial logit and a Mixed logit; in the first one she found a value of time of R$ 3.27/h, in the second one, R$ -1.02/h; Lucinda *et al.* (2014) ran a Mixed logit with a value of time of R$ 56.88/h. Pacheco and Chagas found a value of time of R$ 8.58.

Now we will proceed to the analysis of the individual coefficients, first with alternative-specific, then the case variables. 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. These variables were all individually significant, except travel time in 2012, which is unusual in the literature, since travel time is expected to play an important role in the mode choice process. Also, we should note that both time and cost decreased the

[15] Hourly wage-rate given by the ratio between the annual wage (monthly wage times twelve) and the average number of working hours per year (52 - weeks in year - multiplied by 40 – the "usual" working day).

magnitude of their coefficients (they became less negative), especially time, which more than halved its value.

The case variables have many different pattern behaviors depending on the Survey and/or mode of transportation. We will not discuss each variable's sign and magnitude now, only later when we present the average marginal effects. Generally, age was significant only for the 2007 Survey for the private mode, sex was significant overall, study was not significant, the employment relationship dummies were for the private mode in 2007 and not significant in 2012 overall, and the education degree "High school" and "College" dummies were significant for private mode in 2007 and not significant in 2012. The constants, also called alternative-specific constants, were estimated for the sole purpose of accounting for any effects not captured by other variables, a non-significant coefficient is good sign that the model is well specified. In our case, these variables were significant for the private mode in 2007 and for the transit mode in 2012.

Table 13: 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)
Public agent		-0.153	0.154		0.423	0.664
		(-0.61)	(0.63)		(1.20)	(1.93)
Self employed		-0.590***	0.668***		-0.405	0.832***
		(-3.29)	(3.87)		(-1.61)	(3.55)
Employer		-1.060**	1.392***		-1.850**	0.675
		(-2.91)	(4.90)		(-2.66)	(1.74)
Independent professional		0.124	0.980**		0.912	1.363*
		(0.34)	(2.93)		(1.51)	(2.34)
Family business employer		-0.328	1.646***		-0.091	2.095*
		(-0.74)	(4.52)		(-0.10)	(2.43)
Family business employee		-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		
Value of time	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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

There are three variants of the marginal effects: the average marginal effect (AME), marginal effect at a representative value (MER) and marginal effects at the mean (MEM). MER provides an estimate of the marginal effect at a particular value, $x = x^*$, usually the x^* value is the sample mean, thus becoming the marginal effect at the mean (MEM). The AME does a slightly trickier yet elegant computation. We will explain it through an example in our context: suppose we are going to estimate the AME of the sex variable on the probability of choosing a commute mode. The AME computes two scenarios: first, treat each case as if it were female, for example, regardless of what sex the person really is and leaving all other independent variable values as is, then computing the probability of choosing transit as a mode (then private and non-motorized as well). Second, we do the same thing, but treating the person as though it were male. The difference between the two probabilities is the Average Marginal Effect.

Table 14 shows the alternative-specific variables' AME computed for each Survey. 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.

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.

Table 14: Alternative-specific variables' 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

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

Table 15 presents the case variables' average marginal effects for the 2007 Survey. The first column shows the variable names and the next three columns show the AME for each mode of transport. We will comment first the continuous variables, in our case only age, then the categorical variables. Age shows a significant coefficient only for private and non-motorized modes; transit and non-motorized presented a negative sign, this indicates a negative relationship between age and the choice of these modes. The interpretation of its magnitude is as follows: using transit as example, the average marginal effect of a one unit (one year) increase in age is associated with a 0.1 percentage point decrease of choosing transit as a commute mode, adjusted to the sample distributions of all the variables in the model.

Going now to the categorical variables, sex shows a significant coefficient for transit and non-motorized modes. The sign interpretation is slightly different from the continuous variables. A negative sign means a given category has a lower probability of chosen a certain mode of transport compared to the base category. Males have a lower probability of choosing transit and non-motorized than females, and the inverse is true for the private modes. The magnitude interpretation is as follows: being a male raises the probability of choosing private as commute mode by 15 percentage points compared to females. Study was not a significant variable in neither mode of transport, it showed a negative sign for transit and positive for the other modes. For the employment relationship, public agent was not significant; informal contract was significant only for the transit mode; independent professional and family business employee was significant for private only; self employed, employer and family business employer was significant for transit and private modes. All categories were negative for the transit mode compared to the base category, formal contract, positive for the private mode, and negative for non-motorized, except for the informal contract. The higher magnitudes were found for the employer and family business employee categories. For the education degree, elementary and secondary school were significant only for the private mode, high school was significant for private and non-motorized, and college was significant for all modes of transport. A negative sign was found for all the categories for the transit and non-motorized modes, and positive for the private mode. We can see an increase both in magnitude and significance, the higher the education degree.

Table 15: Case variables' average marginal effects for the 2007 Survey

	2007 Survey		
	Transit	Private	Non-motorized
Age	-0.001	0.005***	-0.005***
	(0.22)	(6.20)	(3.75)
Sex	-0.111***	0.150***	-0.039
	(5.63)	(7.24)	(1.32)
Study	-0.053	0.026	0.027
	(1.88)	(0.96)	(0.58)
Informal contract	-0.070**	0.040	0.030
	(2.73)	(1.55)	(0.72)
Public agent	-0.040	0.041	-0.002
	(1.19)	(1.19)	(0.03)
Self employed	-0.159***	0.177***	-0.017
	(5.74)	(6.16)	(0.42)
Employer	-0.301***	0.357***	-0.057
	(4.75)	(6.84)	(0.77)
Independent professional	-0.047	0.189***	-0.143
	(0.92)	(3.93)	(1.71)
Family business employer	-0.180**	0.356***	-0.176
	(2.60)	(6.13)	(1.89)
Family business employee	-0.208	0.449***	-0.241
	(1.82)	(4.74)	(1.60)
Elementary school	-0.043	0.108*	-0.064
	(0.85)	(1.96)	(0.79)
Secondary school	-0.083	0.127*	-0.044
	(1.68)	(2.38)	(0.55)
High school	-0.028	0.205***	-0.176*
	(0.60)	(4.01)	(2.31)
College	-0.105***	0.384***	-0.279***
	(2.03)	(6.90)	(3.59)

z statistics in parentheses
source: elaborated by the author
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16 presents the case variables' average marginal effects for the 2012 Survey, it has the same layout as the previous table. Starting the coefficients interpretation, age shows a significant coefficient for transit and private modes, with a positive sign for private and non-motorized modes. Compared to the previous table, the sign for transit and private remained equal, that means, the older the person, the greater the probability of choosing private and the lower the probability of choosing transit. For the non-motorized, the sign changed from negative to positive; however, it was not statistically significant in 2012, we would expect its sign to be negative and to have a greater magnitude than transit, as it happens in 2007. Sex was significant for all modes

with a positive sign for the private mode only, exactly the same as the previous table: males have a higher probability of choosing private compared to females, and females have a higher probability of choosing transit and non-motorized compared to males. Study was not significant at all, the same as the previous Survey. For the employment relationship, informal contract was not statistically significant; public agent and independent professional were significant only for the private mode; self employed, employer, family business employer and family business employee were significant for transit and private modes. All significant variables were negative for the transit mode, positive for the private mode and not statistically significant for non-motorized mode, this is the same general behavior in terms of significance and sign compared to the previous Survey, although the magnitudes changed slightly according to the variable. For the education degree, college was the only significant variable and only for transit and private modes. All the variables were negative for the transit mode, positive for the private mode, for the non-motorized mode, elementary and secondary school were positive and high school and college were negative, compared to the base category illiterate or incomplete elementary school. For the transit and private modes, we can that the higher the education degree, the greater the coefficient magnitude: more formal education is associated with a lower probability of choosing transit as a commute mode and a higher probability of choosing a private mode. In the non-motorized mode, the effect is somewhat small and neither coefficient is statistically significant. Overall, this is almost the same pattern in terms of significance and sign compared to the previous Survey.

Table 16: Case variables' average marginal effects for the 2012 Survey

	2012 Survey		
	Transit	Private	Non-motorized
Age	-0.004***	0.002*	0.002
	(3.91)	(2.19)	(1.16)
Sex	-0.107***	0.217***	-0.110**
	(3.34)	(8.62)	(2.87)
Study	-0.069	-0.006	0.075
	(1.46)	(0.14)	(1.24)
Informal contract	-0.021	0.041	-0.019
	(0.49)	(0.99)	(0.35)
Public agent	0.012	0.101*	-0.113
	(0.25)	(2.03)	(1.62)
Self employed	-0.195***	0.243***	-0.048
	(4.99)	(7.10)	(0.99)
Employer	-0.501***	0.386***	0.114
	(3.71)	(4.34)	(1.14)
Independent professional	0.036	0.201**	-0.237
	(0.42)	(2.54)	(1.92)
Family business employer	-0.281*	0.496***	-0.215
	(2.21)	(4.02)	(1.23)
Family business employee	-0.682**	0.369*	0.313
	(2.96)	(2.23)	(1.44)
Elementary school	-0.033	0.006	0.027
	(0.42)	(0.07)	(0.28)
Secondary school	-0.114	0.037	0.076
	(1.46)	(0.49)	(0.78)
High school	-0.100	0.129	-0.028
	(1.35)	(1.73)	(0.31)
College	-0.253***	0.312***	-0.058
	(3.29)	(3.99)	(0.61)

z statistics in parentheses
source: elaborated by the author
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 17 provides information about the tests for the IIA assumption on the Conditional logit models for 2007 and 2012, respectively: the Hausman (1978) test and the Hausman and McFadden (1984) test. In the first column we can see the survey year and the mode of transportation. In the next four columns we can see the Hausman statistics and in the next four columns the Hausman-McFadden statistics. The tests were conducted according to the following procedure: one mode of transport is defined as the base alternative, the test statistic is calculated excluding one of the other alternatives, since we have three modes of transportation, there are two possibilities for exclusion,

that is why we have two calculated statistic for each test for each mode of transportation. For example, in the first line, transit was kept as the base outcome, then in the 574.41 chi2, we excluded the private mode, and in the 1076.25 chi2 we excluded the non-motorized mode. Most of the times, we rejected the null hypothesis of the IIA assumption, except in the Hausman test keeping the non-motorized mode as base outcome, when in three cases it didn't meet the asymptotic assumptions (it returned a negative statistic) and in one case it didn't reject the null hypothesis. Also, in the Hausman-McFadden test for the private mode in 2012, the variance matrix was non-symmetric or highly singular; the software returned the coefficients, but not the robust standard errors. Although McFadden (1974a, p. 243) argues that the Multinomial and Conditional logit models should be used in cases where the alternatives "can plausibly be assumed to be distinct and weighted independently in the eyes of each decision maker". This is probably our case, since we have well distinguished alternatives; however, we aggregated the alternatives according to our needs and possibilities, but, possibly, in the eyes of the decision makers they were dependent of each other. Since we rejected the IIA assumption in our tests, we will proceed to further specify a Nested logit model, which doesn't rely on this assumption.

Table 17: Tests for the IIA assumption (2007 and 2012 Surveys)

	Hausman				Hausman-McFadden			
2007	chi2	p	chi2	p	chi2	p	chi2	p
Transit	574.41	0.00	1076.25	0.00	365.82	0.00	1025.18	0.00
Private	443.09	0.00	286.06	0.00	347.28	0.00	231.01	0.00
Non-motorized	-502.88	.*	-206.45	.*	1107.23	0.00	523.51	0.00
2012								
Transit	289.16	0.00	670.04	0.00	156.66	0.00	275.43	0.00
Private	185.83	0.00	0.00	1.00	65981.84	0.00	**	**
Non-motorized	-73.15	.*	5.52	0.99	220.31	0.00	197.97	0.00

H_0: the IIA assumption holds.

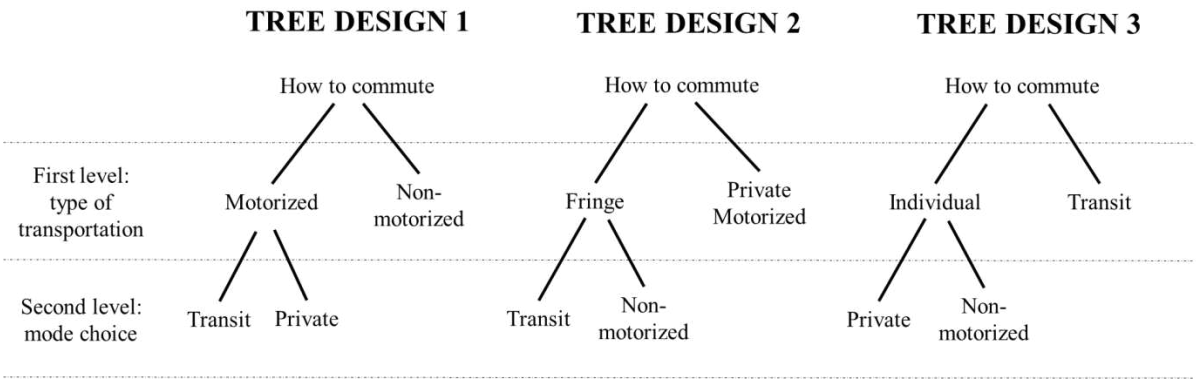
* If chi2<0, the estimated model does not meet asymptotic assumptions.

** The variance matrix is nonsymmetric or highly singular. The software returned the coefficients, but not the robust standard errors.

Since our tests rejected the IIA assumption, we will now proceed to further specify two Nested logit models, which have the special feature of not relying on the IIA assumption. The Nested logit groups similar alternatives into nests, inside a nest we can relax the IIA assumption while between nests, the IIA is still valid.

5.2. NESTED LOGIT MODELS

Although we have a clear gain by relaxing the IIA assumption, the Nested logit also has its disadvantages. The main one is the nesting design. Since we need to group similar alternatives, we have to define *ex ante* which alternatives are similar in a completely arbitrary process. We have only three alternatives, so this is not so much of a problem, since there are only three possible ways which we can group our alternatives. All the possibilities are illustrated in the following figure. Two alternatives will always be grouped together and one alternative will be left alone, with no further branching: this case is known as a degenerate branch, a branch with only one alternative. The tree design number 1 is separates between motorized (transit and private) and non-motorized; tree number 2 separates between what we called "fringe" modes (transit and non-motorized) and private motorized; tree number 3 separates between individual (private and non-motorized) and transit.



Source: elaborated by the author

Figure 7: Tree design possibilities. Design 1: grouping transit and private in the same branch. Design 2: grouping transit and non-motorized in the same branch. Design 3: grouping private and non-motorized in the same branch.

The next three tables show the Nested logit models results, they all have a similar layout. The names of the variables appear in the first column, in the second column we have the coefficients of the 2007 Survey and in the third column we have the coefficients of the 2012 Survey. The first block (first two lines) we have the alternative-specific variables time and cost; in the next block we have the first-level coefficients or the case variables; as second-level coefficients, we only used the alternative-specific constants; then we have the first-level taus or dissimilarity parameters, which measure the degree of correlation of random shocks within each of the types of transportation, dissimilarity parameters greater than one imply that the model is inconsistent with the Random Utility Model (RUM); the Conditional logit is a special case of the Nested logit in which all the dissimilarity parameters are equal to one, at the bottom of the output, we find a likelihood-ratio test of this hypothesis. Equivalently, the property known as the IIA imposed by the Conditional logit model holds if and only if all dissimilarity parameters are equal to one.

Table 18 presents the Nested logit results for the tree design number 1. We will keep the following discussion brief because we already did it extensively in the Conditional logit. The alternative-specific variables were all significant, except time in the 2012 Survey. The implicit Value of Time is R$ 1.13 and R$ 0.26 for 2007 and 2012, respectively, which is a lot lower than expected for 2012. For 2007, age, study, the employment relationship and the highest education degree were significant; for 2012, sex and three categories of the employment relationship were significant only. The alternative-specific constants were significant only for 2007, indicating that there are some unaccounted factors that influence the commute mode choice. The taus are less than or equal to one, that is, they are in the expected interval, one of them is always exactly equal to one, because it is the degenerate branch. The Likelihood Ratio test indicates that we cannot reject the null that all of the log-sum coefficients are 1 and hence we should go back to the Conditional logit model.

Table 18: Nested logit, tree design number 1 (2007 and 2012 Surveys)

	2007 Survey	2012 Survey
Time	-0.283*	-0.0614
	(-2.40)	(-0.29)
Cost	-0.251***	-0.233**
	(-6.00)	(-2.78)
Motorized		
Age	0.0235***	-0.00587
	(7.34)	(-0.96)
Sex	0.0192	0.519***
	(0.25)	(3.38)
Study	0.269*	-0.320
	(2.05)	(-1.22)
Employment relationship:		
Informal contract	-0.114	0.0233
	(-0.98)	(0.09)
Public agent	-0.102	0.380
	(-0.66)	(1.25)
Self employed	0.292**	0.573**
	(2.85)	(2.68)
Employer	0.671***	0.0275
	(3.48)	(0.08)
Independent professional	0.554*	1.064*
	(2.52)	(2.12)
Family business employer	1.634***	1.320*
	(5.52)	(2.41)
Family business employee	1.105**	-0.485
	(2.68)	(-0.52)
Education degree:		
Primary school	0.262	0.159
	(1.20)	(0.31)
Secondary school	0.377	-0.0621
	(1.79)	(-0.12)
High school	0.903***	0.489
	(4.53)	(1.01)
College	1.503***	0.747
	(7.38)	(1.52)
Transit_cons	-0.182***	0.0848
	(-6.77)	(0.96)
Non-motorized_cons	0.986***	-0.486
	(4.11)	(-0.88)
Motorized tau	0.729***	0.711**
	(5.84)	(2.67)
Non-motorized tau	1.000	1.000
	(0.00)	(0.00)
LR test for IIA (tau=1), p-value	0.113	0.575

t statistics in parentheses
source: elaborated by the author
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 19 presents the Nested logit results for the tree design number 2. All the alternative-specific coefficients were significant and have the expected negative sign. The Value of Time is R$ 2.47 and R$ 1.29 for 2007 and 2012, respectively, which is more in accordance with what we expected. For 2007, all first-level coefficients were significant, except study, public agent and primary school; for 2012, age, sex, all the employment relationship categories except informal contract and college were significant. All the alternative-specific constants were significant. The taus were less than or equal to one, making the model consistent with the RUM theory. The Likelihood Ratio test rejected the IIA assumption for both the 2007 and 2012 Surveys.

Table 20 presents the Nested logit results for the tree design number 3. All the alternative-specific variables were significant, except time for 2012. The Value of Time is R$ 1.91 and R$ 0,04 for 2007 and 2012, respectively, the value is much lower than expected for 2012, even though the time coefficient was not significant. For 2007, the case variables sex, age, all the employment relationship categories, except public agent, and college were significant; for 2012, age, sex, all employment relationship categories, except informal contract and public agent, and college were significant. All the alternative-specific constants were significant. The taus were numerically greater than one; however, the Likelihood Ratio test did not reject the IIA assumption for 2007, but it rejects it for 2012.

Overall, the different Nested logit tree designs provided different insights about the commuting behavior of São Paulo. There was some variability in the Value of Time, ranging from R$ 0.04 to R$ 2.47. Also, we have mixed results about in the independence of the alternatives, one design completely rejected the IIA assumption, another did not reject it, and a third one rejected it for one survey and did not reject it for the other. It also means that the rejection or acceptance of the IIA assumption depends not only on the model we estimate to begin with, but also on the way we nest the alternatives.

Table 19: Nested logit, tree design number 2 (2007 and 2012 Surveys)

	2007 Survey	2012 Survey
Time	-0.508***	-0.309*
	(-7.11)	(-2.27)
Cost	-0.206***	-0.239***
	(-15.10)	(-9.68)
Motor private		
Age	0.0191***	0.0135***
	(11.27)	(4.05)
Sex	0.554***	0.744***
	(13.79)	(9.44)
Study	0.0339	0.00457
	(0.53)	(0.03)
Employment relationship:		
Informal contract	0.175*	0.217
	(2.56)	(1.43)
Public agent	0.0351	0.314*
	(0.47)	(2.02)
Self employed	0.831***	1.173***
	(14.66)	(10.43)
Employer	1.714***	1.511***
	(13.10)	(5.64)
Independent professional	0.890***	0.736**
	(8.31)	(3.27)
Family business employer	1.962***	2.109***
	(11.89)	(5.46)
Family business employee	2.008***	1.521*
	(7.37)	(2.50)
Education degree:		
Primary school	0.288	0.0904
	(1.85)	(0.31)
Secondary school	0.467**	0.267
	(3.14)	(0.94)
High school	0.688***	0.517
	(4.92)	(1.92)
College	1.597***	1.342***
	(11.47)	(4.95)
Transit_cons	2.174***	2.135***
	(13.67)	(6.87)
Non-motorized _cons	1.795***	1.440***
	(11.05)	(4.51)
Motor private tau	1.000	1.000
	(0.00)	(0.00)
Fringe tau	0.198***	0.149***
	(8.18)	(3.88)
LR test for IIA (tau=1), p-value	0.000	0.000

t statistics in parentheses
source: elaborated by the author
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 20: Nested logit, tree design number 3 (2007 and 2012 Surveys)

	2007 Survey	2012 Survey
Time	-0.496***	-0.0121
	(-5.44)	(-0.05)
Cost	-0.260***	-0.269***
	(-16.77)	(-9.25)
Individual		
Age	0.0143***	0.0173***
	(7.83)	(4.73)
Sex	0.615***	0.717***
	(14.30)	(8.35)
Study	0.0304	0.1000
	(0.45)	(0.70)
Employment relationship:		
Informal contract	0.255***	0.168
	(3.51)	(0.99)
Public agent	0.0385	0.171
	(0.48)	(1.00)
Self employed	0.827***	1.135***
	(13.35)	(9.16)
Employer	2.109***	2.529***
	(11.77)	(5.41)
Independent professional	0.904***	0.522*
	(7.51)	(2.14)
Family business employer	1.731***	1.938***
	(9.37)	(4.26)
Family business employee	2.104***	2.154**
	(6.21)	(2.65)
Education degree:		
Primary school	0.149	0.120
	(0.92)	(0.39)
Secondary school	0.215	0.317
	(1.39)	(1.05)
High school	0.239	0.440
	(1.66)	(1.53)
College	1.107***	1.310***
	(7.70)	(4.53)
Transit_cons	1.569***	2.155***
	(9.35)	(6.36)
Non-motorized_cons	-1.041***	-1.522***
	(-10.79)	(-7.65)
Individual tau	1.132***	1.579***
	(12.38)	(8.14)
Transit tau	1.000	1.000
	(0.00)	(0.00)
LR test for IIA (tau=1), p-value	0.334	0.004

t statistics in parentheses
source: elaborated by the author
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6. DISCUSSION AND FINAL REMARKS

This research attempted to measure the correlation between economic, demographic and transportation variables and the probability of choosing a certain commute mode of transport. We began by giving a brief history of the work commute, the importance and the relationship of the automobile in the process, and illustrating the commute pattern with the Brazilian and São Paulo average travel times, usage and ownership cost of private and transit modes, and an overview of the city of São Paulo. We proceeded to the contribution of the present work, as well as the statement of the hypothesis and the objectives.

We dived into a more technical explanation of the transportation literature. First, by presenting a simple abstraction of the transportation market, with its supply and demand, showing how congestion deviate the equilibrium from the social optimum to a lower and less efficient private equilibrium, when roads are untolled. We also presented a mathematical formulation of the Down-Thomson paradox, in which every individual traveler suffers from a higher generalized travel cost after the highway capacity expansion due to the direct and indirect effects of the volume shifting. We derived the most important discrete choice statistical models aligned with microeconomic theory based on the Random Utility Maximization, first the Logit, the Multinomial logit (ML) and the Conditional logit (CL), all of which rely on the Independence of Irrelevant Alternative (IIA) assumption, that is, the probability ratio of two alternatives is totally independent of any other alternative in the choice set. Despite the fact that it is heavily a conceptual matter whether the IIA holds or not in a given choice set, we presented three statistical tests to validate or not the IIA assumption. Then, we exposed a recent class of models that do not rely on the IIA assumption, called Generalized Extreme Value (GEV) mode, whose most prominent member are the Nested logit (NL) and Cross-Nested Logit (CNL) models. Finally, we discussed some of the most relevant works in the rich mode choice literature, beginning with works of Daniel McFadden in the decade of 1970's up until the GEV models in the 2000's.

In the third chapter, we explained our data sources as well as our empirical strategy. 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 exactly how we pre-processed our data: dropping missing values, smoothing noisy data, removing outliers and resolving inconsistencies. The second part consisted of estimating the counterfactual travel times, that is, the travel times of the modes not taken (not observed in the dataset). Thirdly, we estimated the observed and counterfactual travel costs. Finally, we made explicit our model specification for both the dependent and the independent variables.

In the fourth chapter, we showed our results from the models estimation for both 2007 and 2012 Surveys. First, we specified a Conditional logit model, which has both alternative-specific and case-specific variables. Cost and travel time, the alternative-specific variables, were negatively correlated with the probability of choosing any commute mode, which was as expected and fully confirmed our initial hypothesis about these coefficients. The Value of Travel Time Savings (VoTT), that is the ratio between the time and cost coefficients, we found was low compared to some studies and high compared to others, in a somewhat middle ground; this is a strong indicator that more studies are necessary to confirm the true magnitude of the São Paulo Value of Time. 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. The AVE for the case variables revealed that males have a higher probability of choosing private modes and a lower probability of choosing transit and non-motorized modes, confirming our initial hypothesis. Older people have a lower probability of choosing transit and non-motorized and a higher probability of choosing private as commute mode, except the non-motorized coefficient for 2012, which was not significant, thus partially confirming our hypothesis. A formal contract as employment relationship leads to a higher probability of choosing transit, in accordance with our initial hypothesis, but is correlated with a lower probability of choosing private as commute mode, contrary to our hypothesis. A higher education degree is associated with a lower probability of choosing transit (contrary to expected), a higher probability of choosing private modes and a lower probability of choosing non-motorized modes, as initially expected. We

performed the Hausman and the Hausman-McFadden tests for the Independence of the Irrelevant Alternative assumption, most of the results pointed to a rejection of the IIA assumption. Our next step was to specify a Nested logit model, since we have three alternatives, there were also three nesting possibilities, which we estimated for each Survey. The significance, sign and magnitude of the coefficients were somewhat similar between the Nested logit models and the Alternative-specific logit models; however, the Value of Time ranged to some degree of variability and, perhaps more importantly, there were some cases were we rejected IIA assumption and others which we did not, based on the results of the Likelihood Ratio test. This indicates the dependence of the IIA assumption on the model specification and the nesting structure, pointing to the need for further investigation in these topics.

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 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

ALVARENGA, G. V.; *et al.* **Políticas anticíclicas na indústria automobilística: uma análise de cointegração dos impactos da redução do IPI sobre as vendas de veículos**. No. 1512. Texto para Discussão, Instituto de Pesquisa Econômica Aplicada (IPEA), 2010.

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

ARNOTT, R.; SMALL, K A. The economics of traffic congestion. **American scientist**, v. 82, n. 5, p. 446-455, 1994.

BARCELLOS, T. M. **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**. 2014. 73f. Dissertação (Mestrado em Economia Aplicada) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, Ribeirão Preto. 2014.

BARREIRA, I. A. F. Ação direta e simbologia das "jornadas de junho": notas para uma sociologia das manifestações. **Contemporânea**, v. 4, n. 1, p. 145-164, 2014.

BEN-AKIVA, M.; LERMAN, S. R. Some estimation results of a simultaneous model of auto ownership and mode choice to work. **Transportation**, v. 3, n. 4, p. 357-376, 1974.

BEN-AKIVA, M. E.; LERMAN, S. R. **Discrete choice analysis: theory and application to travel demand**. Cambridge: MIT Press, 1985.

BIERLAIRE, M. A theoretical analysis of the cross-nested logit model. **Annals of Operations Research**, v. 144, n. 1, p. 287-300, 2006.

BUEHLER, R.; HAMRE, A. **Paying commuters to get on their bikes is not enough**. 2014. Available at: http://theconversation.com/paying-commuters-to-get-on-their-bikes-is-not-enough-28998. Accessed in: 22/03/2017.

BRINCO, R. **Transporte urbano e dependência do automóvel**. Porto Alegre: FEE, 2005.

CAMERON, A. C.; TRIVEDI, P. K. **Microeconometrics using Stata**. College Station, TX: Stata Press, v. 5, 2009.

CARVALHO, C. H. R. de; *et al*. Tarifação e financiamento do transporte público urbano. **Nota Técnica** no. 2 IPEA (2013).

CHENG, S.; LONG, J. S. Testing for IIA in the multinomial logit model. **Sociological Methods and Research**, v. 35, n. 4, p. 583-600, 2007.

DARGAY, J.; GATELY, D.; SOMMER, M. Vehicle ownership and income growth, worldwide: 1960-2030. **The Energy Journal**, v. 28, n. 4, p. 143-170, 2007.

DE CLERCQ, G. **France experiments with paying people to cycle to work**. 2014. Available at: http://www.reuters.com/article/us-france-bicycles-idUSKBN0ED1O120140602. Accessed in: 22/03/2017.

DE PALMA, A.; LINDSEY, R.; WU, F. Private operators and time-of-day tolling on a congested road network. **Journal of Transport Economics and Policy**, v. 42, n. 3, p. 397-433, 2008.

DE PALMA, A.; ROCHAT, D. Mode choices for trips to work in Geneva: an empirical analysis. **Journal of Transport Geography**, v. 8, n. 1, p. 43-51, 2000.

DENATRAN. **Frota de veículos, por tipo e com placa, segundo as Grandes Regiões e Unidades da Federação (2001-2013)**. Available at: www.denatran.gov.br/download/frota/FROTA_2012.zip. Accessed on: 07/04/2016.

DISSANAYAKE, D.; MORIKAWA, T. 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**, v. 18, n. 3, p. 402-410, 2010.

DOWNS, A. The law of peak-hour expressway congestion. **Traffic Quarterly**, v. 16, p. 393-409, 1962.

DOWNS, A. **Stuck in traffic: coping with peak-hour traffic congestion**. Washington: Brookings Institution Press, 1992.

DOWNS, A. **Still stuck in traffic: coping with peak-hour traffic congestion**. Washington: Brookings Institution Press, 2004.

DURANTON, G.; TURNER, M. A. The fundamental law of road congestion: evidence from US cities. **The American Economic Review**, v. 101, n. 6, p. 2616-2652, 2011.

EMTU. **Itinerários e tarifas**. 2017. Available at: http://www.emtu.sp.gov.br/emtu/itinerarios-e-tarifas/tarifas-em-formato-pdf.fss. Accessed on: 19/01/2017

ENVIRONMENTAL PROTECTION AGENCY. **National emission inventory: air pollutant emission trends**. 2002. Available at: www.epa.gov/ttn/chief/trends/index.html. Accessed in: 22/03/2017.

FRUMKIN, H. Urban sprawl and public health. **Public Health Reports**, v. 117, n. 3, p. 201-217, 2002.

FURTADO, C. **Formação econômica do Brasil**. São Paulo: Companhia das Letras, 2007.

GAKENHEIMER, R. Urban mobility in the developing world. **Transportation Research Part A**, v. 33, p. 671-689, 1999.

GAUDRY, M. J. I.; DAGENAIS, M. G. The dogit model. **Transportation Research**, v. 13, n. 2, p. 105-111, 1979.

HARDIN, G. The tragedy of the commons. **Science**, v. 162, p. 1243-1248, 1968.

HARTOG, J. J.; et al. Do the health benefits of cycling outweigh the risks? **Environmental Health Perspectives**, v. 118, n. 8, p. 1109-1116, 2010.

HAUSMAN, J. A. Specification tests in econometrics. **Econometrica: Journal of the Econometric Society**, v. 46, p. 1251-1271, 1978.

HAUSMAN, J. A.; MCFADDEN, D. Specification tests for the multinomial logit model. **Econometrica: Journal of the Econometric Society**, v. 52, p. 1219-1240, 1984.

HEISS, F. Structural choice analysis with nested logit models. **The Stata Journal**, v. 2, n. 3, p. 227-252, 2002.

HSU, W.-T.; ZHANG, H. The fundamental law of highway congestion revisited: evidence from national expressways in Japan. **Journal of Urban Economics**, v. 81, p. 65-76, 2014.

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

INMETRO. **Tabelas PBE veicular – veículos leves 2009**. 2009. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2010**. 2010. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2010.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2011**. 2011. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2011.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2012**. 2012. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2012.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2013**. 2013. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2013.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2014**. 2014. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2014.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2015**. 2015. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2015.pdf. Accessed in: 19/12/2016.

INMETRO. **Tabelas PBE veicular – veículos leves 2016**. 2016. Available at: http://www.inmetro.gov.br/consumidor/pbe/veiculos_leves_2016.pdf. Accessed in: 19/12/2016.

IPEA. **Impactos da redução do imposto sobre produtos industrializados (IPI) de automóveis**. Nota técnica n. 15 (2009). Available at: http://ipea.gov.br/agencia/images/stories/PDFs/2009_nt015_agosto_dimac.pdf. Accessed in: 06/18/2016.

IRS. **Publication 15-B - main content**. 2017. Available at: https://www.irs.gov/publications/p15b/. Accessed in: 22/03/2017.

JAFFE, E. **The problem with paying people to bike to work**. 2015. Available at: https://www.citylab.com/transportation/2015/03/the-problem-with-paying-people-to-bike-to-work/388099/. Accessed in: 22/03/2017.

JARA-DÍAZ, S. R.; VIDELA, J. Detection of income effect in mode choice: theory and application. **Transportation Research Part B: Methodological**, v. 23, n. 6, p. 393-400, 1989.

KOPPELMAN, F. M.; BHAT, C. **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: 06/17/2016.

LEE, I-M.; BUCHNER, D. M. The importance of walking to public health. **Medicine and Science in Sports and Exercise**, v. 40, n. 7, p. S512-S518, 2008.

LITMAN, T. Valuing transit service quality improvements. **Journal of Public Transportation**, v. 11, n. 2, p. 43-63, 2008.

LONG, J. S.; FREESE, J. **Regression models for categorical dependent variables using Stata**. College Station, TX: Stata press, 2006.

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

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

MCFADDEN, D. The measurement of urban travel demand. **Journal of Public Economics**, v. 3, n. 4, p. 303-328, 1974b.

MCFADDEN, D. Modelling the choice of residential location. In: KARLQUIST, A.; *et al*. (eds.). **Spatial Interaction Theory and Residential Location**. Amsterdam: North-Holland, 1978, p. 75–96.

MCFADDEN D. Econometric models of probabilistic choice. In: MANSKI, C. H.; MCFADDEN, D. (eds.). **Structural Analysis of Discrete Data with Econometric Applications**. Cambridge: MIT Press, 1981, p. 198-272.

MCFADDEN, D.; TRAIN, K.; TYE, W. B. An Application of Diagnostic Tests for the Independence From Irrelevant Alternatives Property of the Multinomial Logit Model. **Transportation Research Board Record**, v. 637, p. 39-46, 1981.

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

METRÔ. **Pesquisa de Mobilidade da Região Metropolitana de São Paulo – manual da pesquisa domiciliar**. 2012b. Available at:

http://www.metro.sp.gov.br/metro/arquivos/mobilidade-2012/manuais/manual codificador domiciliar_2012.pdf. Accessed in: 05/12/2015.

METRÔ. **Zoneamento da Pesquisa de Mobilidade 2012 (em formato jpg)**. 2012c. 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. **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.

OJA, P.; *et al*. Health benefits of cycling: a systematic review. **Scandinavian Journal of Medicine and Science in Sports**, v. 21, n. 4, p. 496–509, 2011.

MACIEL, E. Alckmin sanciona lei que garante tarifa zero a estudantes de SP. 2015. Available at: http://g1.globo.com/sao-paulo/noticia/2015/02/alckmin-sanciona-lei-que-da-tarifa-zero-para-estudantes-em-trens-e-metro.html. Accessed in: 23/01/2017.

MYLES, G. D. **Public economics**. Cambridge: Cambridge University Press, 1995.

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

MURRAY, M. P.; KORY, R. C.; CLARKSON, B. H. Walking patterns in healthy old men. **Journal of Gerontology**, v. 24, n. 2, p. 169-178, 1969.

MUTRIE, N.; et al. "Walk in to Work Out": a randomised controlled trial of a self help intervention to promote active commuting. **Journal of Epidemiology and Community Health**, v. 56, p. 407-412, 2002.

NASH, J. R. Economic Efficiency Versus Public Choice: The case of Property Rights in Road Traffic Management. **Boston College Law Review**, v. 49, n. 3, p. 673-739, 2008.

NEWEY, W. K. Efficient estimation of limited dependent variable models with endogenous explanatory variables. **Journal of Econometrics**, v. 36, p. 231–250, 1987.

ÖZKAN, T.; LAJUNEN, T. 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**, v. 9, n. 4, p. 269-277, 2006.

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

PEREIRA, R. H. M.; SCHWANEN, T . **Tempo de Deslocamento Casa - Trabalho no Brasil (1992-2009): diferenças entre regiões metropolitanas, níveis de renda e sexo**. Rio de Janeiro: Ipea, 2013 (Texto para Discussão IPEA, nº. 1813)

PNAD, IBGE. "**Pesquisa nacional por amostra de domicílios**." Rio de Janeiro: IBGE (2001-2013).

RHODES, N.; PIVIK, K. Age and gender differences in risky driving: The roles of positive affect and risk perception. **Accident Analysis and Prevention**, v. 43, n. 3, p. 923-931, 2011.

RILEY, K. Motor vehicles in China: the impact of demographic and economic changes. **Population and Environment**, v. 23, n. 5, p. 479-494, 2002.

ROJAS-RUEDAS, D.; *et al*. The health risks and benefits of cycling in urban environments compared with car use: health impact assessment study. **BMJ**, v. 343, p. 1-8, 2011.

RUGGIERO, A. **Free money? How to get paid biking to work**. 2016. Available at: https://gearjunkie.com/bicycle-commuter-tax-benefit-explained. Accessed in: 22/03/2017.

SALLIS, J. F.; *et al*. Active transportation and physical activity: opportunities for collaboration on transportation and public health research. **Transportation Research Part A: Policy and Practice**, v. 38, p. 249-268, 2004.

SÃO PAULO. **Decreto nº 47.936**, de 30 de novembro de 2006. 2006. Available at: http://www3.prefeitura.sp.gov.br/cadlem/secretarias/negocios_juridicos/cadlem/integra.asp?alt=01122006D%20479360000. Accessed in: 23/01/2017.

SÃO PAULO. **Decreto nº 52.066**, de 30 de dezembro de 2010. 2010. Available at: http://www3.prefeitura.sp.gov.br/cadlem/secretarias/negocios_juridicos/cadlem/integra.asp?alt=27022008D%20520660000. Accessed in: 23/01/2017.

SÃO PAULO. **Decreto nº 54.641**, de 28 de novembro de 2013. Available at: http://www3.prefeitura.sp.gov.br/cadlem/secretarias/negocios_juridicos/cadlem/integra.asp?alt=29112013D%20546410000%20%20%20%20%20%20%20%20%20%20%20%20&secr=&depto=&descr_tipo=DECRETO. Accessed in: 23/01/2017.

SMALL, K. A. The scheduling of consumer activities: work trips. **American Economic Review**, v. 72, p. 467-479, 1982.

SMALL, K. A.; HSIAO, C. Multinomial logit specification tests. **International Economic Review**, v. 26, p. 619-627, 1985.

SMALL, K. A discrete choice model for ordered alternatives. **Econometrica**, v. 55, n. 2, p. 409-424, 1987.

SMALL K.; ROSEN H. Applied welfare economics with discrete choice models. **Econometrica**, v. 49, p. 105-130, 1981.

SMALL, K. A.; VERHOEF, E T. **The economics of urban transportation**. New York: Routledge, 2007.

SMEED, R. J. **Road pricing: the economic and technical possibilities**. London: HMSO, 1964.

SPTRANS. **A SPTRANS**. 2017. Available at: http://www.sptrans.com.br/. Accessed in: 19/01/2017.

STATACORP. **Stata Statistical Software: Release 11**. College Station, TX: StataCorp LP, 2009a.

STATACORP. **Stata 11 Base Reference Manual**. College Station, TX: Stata Press, 2009b.

SWAIT, J.; BEN-AKIVA, M. Empirical test of a constrained choice discrete model: mode choice in Sao Paulo, Brazil. **Transportation Research Part B: Methodological**, v. 21, n. 2, p. 103-115, 1987.

THOMSON, J. M. **Great Cities and their traffic**, Gollancz, London (Published in Peregrine Books, 1977).

TRAIN, K. A structured logit model of auto ownership and mode choice. **Review of**

Economic Studies, v. 47, p. 357-370, 1980.

TUKEY, J. W. Exploratory Data Analysis. Reading, MA: Addison–Wesley, 1977.

UNITED STATES DEPARTMENT OF HEALTH AND HUMAN SERVICES. **Physical activity and health: a report of the Surgeon General**. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, 1996. Available at: https://www.cdc.gov/nccdphp/sgr/pdf/sgrfull.pdf. Accessed in: 04/04/2017.

VOVSHA, P. Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area. **Transportation Research Record: Journal of the Transportation Research Board**, v. 1607, p. 6-15, 1997.

VRTIC, M.; *et al*. The impacts of road pricing on route and mode choice behaviour. **Journal of Choice Modelling**, v. 3, n. 1, p. 109-126, 2010.

WASHBROOK, K.; HAIDER, W.; JACCARD, M. Estimating commuter mode choice: a discrete choice analysis of the impact of road pricing and parking charges. **Transportation**, v. 33, n. 6, p. 621-639, 2006.

WEN, C.-H.; KOPPELMAN, F. S. The generalized nested logit model. **Transportation Research Part B: Methodological**, v. 35, n. 7, p. 627-641, 2001.

WINTER, N. SMHSIAO: Stata module to conduct Small-Hsiao test for IIA in multinomial logit. Boston College Department of Economics: EconPapers, Statistical Software Components, 2000. Available at: http://econpapers.repec.org/software/bocbocode/s410701.htm. Accessed in: 01/03/2017.

Appendix A: Estimating the Metropolitan Region of São Paulo's average commute travel times using data from the Instituto Brasileiro de Geografia e Estatística

The Pesquisa Nacional por Amostra de Domicílios is a Brazilian national survey which began in 1976 and it draws annual information about the demographic and socioeconomic characteristics of the population, like sex, age, education and income, as well as information about the household, and periodically it also gathers information about migration, fertility and nuptiality, among others. Since 1993, it includes three questions about the morning commute travel times.

The first question people are asked is whether they work; if they do, they are asked whether their workplace is the same as their residence (if yes, we account their commute times as 0min); if not, they are asked how much time does it take them to go straight from home to work, according to four categories: less than 30min, between 30 and 60min, between 60 and 120min or more than 120min. We could treat each interval as point data at their mean (e.g. the first category is always equal to a 15min commute, the second is 45min and the third is 90min) and treating the last category as being always equal to a point data of 120min – as in Pereira and Schwanen (2013); however, this does not take into account the true nature of the data: it is interval data and the last category (more than 120min) is right-censored. Therefore, we estimated the average morning commute times by regressing each year's microdata on a constant only, which provides a more accurate proxy for the average metropolitan commute times. The process involves maximizing the following log-likelihood function:

$$lnL = \sum_{j\epsilon R} w_j \log\left\{1 - \Phi\left(\frac{y_{Rj} - x\beta}{\sigma}\right)\right\} + \sum_{j\epsilon I} w_j \log\left\{\Phi\left(\frac{y_{2j} - x\beta}{\sigma}\right) - \Phi\left(\frac{y_{1j} - x\beta}{\sigma}\right)\right\} \quad (29)$$

In which the first sum comprises the observations $j\epsilon$R are right-censored; we know only that the unobserved y_j is greater than or equal to y_{Rj}. The second sum comprises the observations which are intervals $j\epsilon$I, we know only that the unobserved y_j is in the interval $[y_{1j}, y_{2j}]$.