# GOOGLE MAPS E ESCOLHA DE MODAL: UMA APLICAÇÃO PARA A REGIÃO METROPOLITANA DE SÃO PAULO

Hector de Moura Luz[1]; Adriana Sbicca Fernandes[2]

**Resumo:** Pesquisas de preferência revelada do tipo Origem e Destino formam bases de dados importantes para análise de políticas de transporte, geralmente embasando modelos de escolha discreta. Um dos desafios no desenvolvimento destas análises é a ausência de atributos de alternativas não escolhidas pelos entrevistados. A realização de pesquisa de preferência declarada e a estimação através de técnicas econométricas são as alternativas mais utilizadas para enfrentá-lo. O presente artigo busca analisar uma alternativa para obtenção destes valores através do Google Maps. O desenvolvimento do trabalho utiliza esse aplicativo e a pesquisa de 2007 da Região Metropolitana de São Paulo para construir uma base de dados análoga àquelas estimadas na literatura por Mínimos Quadrados Ordinários para acrescentar os atributos não observados. Após discutir a dificuldade de comparação dos resultados através de análise de erros, compara-se a significância e adequação dos resultados de escolha discreta, maiores com a utilização dos recursos Google.

## 1. INTRODUÇÃO

À medida que as grandes cidades tornam-se mais complexas, o trânsito ganha mais importância: os impactos deste sobre a vida dos habitantes do município, tais como o excessivo tempo despendido nos deslocamentos, extrapolam tal dimensão e passam a afetar o país como um todo, principalmente através da perda de produtividade (HADDAD et al, 2015).

Para o caso brasileiro, a Região Metropolitana de São Paulo, que quase alcança os 20 milhões de habitantes e detém cerca de 19% do PIB brasileiro (EMPLASA, 2011), desperta especial atenção, uma vez que tais vultosos números refletem sua grande importância econômica.

Buscando melhorar a qualidade de vida da população e melhorar a produtividade da economia, diversas medidas vêm sendo tomadas na região, como a ampliação de trechos das Marginais Tietê e Pinheiros e a implantação de ciclofaixas e faixas exclusivas para ônibus.

Para aumentar a efetividade de medidas como as citadas acima em relação a São Paulo, é necessário compreender a escolha dos indivíduos sobre o modal de transporte. Entre as principais ferramentas utilizadas por economistas para tal fim estão os ditos modelos de escolha discreta, amplamente utilizados internacionalmente (KOPPELMAN; BHAT, 2006; DISSANAYAKE; MORIKAWA, 2010) e cuja crescente utilização no país é evidente nos últimos anos (LUCINDA; MEYER; LEDO, 2013; LUCINDA et al, 2015, PACHECO; CHAGAS, 2016).

1 Doutorando em Economia do Desenvolvimento pela FEA/USP
2 Departamento de Economia, Universidade Federal do Paraná

Pesquisas de preferência revelada do tipo Origem e Destino aparecem como bases de dados importantes para a utilização de tais modelos em estudos sobre transporte. Tais pesquisas, entretanto, apresentam alguns desafios para sua manipulação. Dentre estes, chama a atenção a ausência dos atributos das alternativas não escolhidas pelos entrevistados, necessários para a aplicação da modelagem (HENSHER et al, 2005). O objetivo do presente artigo é apresentar uma alternativa às tradicionais formas de complementação de tais bases de dados: a utilização de informações obtidas com a ferramenta Google Maps. Tal apresentação é realizada com uma aplicação de tal alternativa, buscando complementar a Pesquisa Origem e Destino da Região Metropolitana de São Paulo de 2007. Somada à existência de base de dados para essa região e de pesquisas já desenvolvidas que utilizam essas informações, a Região Metropolitana de São Paulo é utilizada no presente trabalho para a apresentação e discussão a respeito da promissora complementação via Google Maps da Pesquisa Origem-Destino com as alternativas não escolhidas pelas pessoas.

Na seção seguinte, discorrer-se-á acerca da ausência dos atributos das alternativas não escolhidas e as principais formas utilizadas para suprir tal carência; a terceira seção apresentará o Google Maps como alternativa para este fim, assim como as características do banco de dados com ele formado; a quarta apresentará as possibilidades e dificuldades de comparação entre as alternativas, que será realizada na quinta seção. Por fim, conclui-se com uma discussão sobre as vantagens, dificuldades e outras possibilidades apresentadas pela alternativa proposta.

## 2. ESTADO DA ARTE DAS ALTERNATIVAS NÃO-ESCOLHIDAS NAS PESQUISAS DE ESCOLHA DE MODAL

Modelos de demanda por viagens historicamente têm se apoiado em pesquisas de preferência revelada, geralmente consideradas as melhores fontes de informação para tal modelagem (BEN-AKIVA; LERMAN, 1985). Dentre as pesquisas desta categoria, aquelas do tipo Origem e Destino apresentam especial relevância. Para as análises de municípios brasileiros, tal relevância é ainda acentuada pela sua ampla utilização há décadas e em diferentes localidades (SWAIT, 1984; SWAIT; BEN-AKIVA, 1987; LUCINDA; MEYER; LEDO, 2013; LUCINDA et al, 2015, PACHECO; CHAGAS, 2016).

No entanto, tais bases de dados apresentam como característica a disponibilidade de informações apenas referentes à alternativa escolhida. Desta forma, um dos problemas de se utilizar a Pesquisa Origem e Destino como única fonte de dados é a ausência dos atributos das alternativas não escolhidas pelos entrevistados, carência esta que impossibilita a utilização de modelos de escolha discreta conforme desenvolvidos até então. Há quatro principais soluções encontradas na literatura para enfrentar essa deficiência (HENSHER et al., 2005).

Um método envolve a utilização das médias ou medianas dos atributos das alternativas observadas, que são assumidas como sendo os valores para os atributos das alternativas não escolhidas por aquelas pessoas que não as escolheram. A segunda forma envolve uma redistribuição dos níveis dos atributos das alternativas observadas para aquelas não-observadas; tal redistribuição pode ser realizada tanto aleatoriamente quanto seguindo

estratégias de *matching*. A realização de uma pesquisa de preferência declarada, com questionários sobre situações hipotéticas de escolha, é uma maneira de incorporar a uma base de dados alternativas não-escolhidas (DISSANAYAKE; MORIKAWA, 2010). Por fim, a quarta alternativa referenciada é descrita como síntese de dados. Como apontam Hensher et al (2005), a norma desta alternativa é “usar informação conhecida tal como distância de viagem ou outras características sociodemográficas e condicionar os dados sintetizados sobre esses". Exemplificando o uso geral, pode-se citar a utilização de regressões lineares em que a duração de cada modal aparece como variável dependente em cada uma delas, usando dados socioeconômicos e específicos daquela viagem como variáveis independentes[1].

Além das quatro formas referenciadas acima, outras duas vêm sendo apontadas mais recentemente na literatura: a utilização de matrizes *skim*[2], com dados de duração e distância entre zonas, como apontado por Washington et al (2014) e Javanmardi et al (2015), e métodos de imputação múltipla Bayesiana, como realizada por Washington et al (2014).

Embora Hensher et al (2005) recomendem a pesquisa de preferência declarada e a síntese de dados, e as formas mais recentemente utilizadas (e acima referenciadas) venham apresentando bons resultados, as dificuldades (tanto técnicas quantos econômicas) relacionadas à realização de pesquisas de preferência declarada favorece a ampla utilização da alternativa sintética na literatura brasileira. Com relação a esta, destacam-se Lucinda, Meyer e Ledo (2013), Barcellos (2014), Lucinda et al (2015) e Feres (2015), que utilizaram a alternativa sintética para a complementação da Pesquisa Origem e Destino 2007 da Região Metropolitana de São Paulo.

Antes de avançar à discussão de tal síntese, é importante destacar algumas das características de tal pesquisa. Aplicada pela Companhia do Metropolitano de São Paulo, o Metrô, traz 168.582 observações com informações de viagens realizadas na referida região metropolitana. Para tanto, cerca de 30 mil domicílios foram entrevistados e validados, apresentando características sociodemográficas também a nível individual, como renda e idade.

A fim de complementar esta pesquisa quanto à carência dos atributos das alternativas não escolhidas, os referidos autores brasileiros realizaram uma síntese de dados utilizando Mínimos Quadrados Ordinários. Para este caso, foi necessário estimar o custo e o tempo de opções que, embora disponíveis, não tenham sido escolhidas pelo tomador de decisão.

Com exceção do custo de algumas alternativas que pôde ser determinado através de informações externas, como de ônibus, metrô e trem, o custo das demais alternativas e o tempo de todas elas foram estimados a partir de um modelo de regressão. Tal modelo usou as escolhas observadas para cada modal, sendo as variáveis dependentes os logaritmos do custo e da duração da viagem. As variáveis independentes utilizadas foram tanto variáveis *dummy* para a hora de partida e de chegada e para motivação da viagem, como a distância em quilômetros entre as zonas de origem e

[1] Duração é uma variável recorrentemente levantada como essencial à compreensão da escolha de modal.

[2] Matrizes que possuem informações, como distância, pertinentes às zonas da área em questão.

destino[3]. A partir de tal modelo, estimou-se coeficientes que, por sua vez, foram utilizados para a obtenção do tempo e do custo esperados para as alternativas não escolhidas.

É importante ressaltar que este tipo de complementação deve levar em conta a disponibilidade de cada modal para cada tomador de decisão, uma vez que a adequação do conjunto de escolha é essencial para a estimação (SWAIT, 1984). A fim de minimizar tal problema, os autores restringiram a disponibilidade de escolha de trem e metrô apenas para as zonas que tivessem algum respondente escolhendo um destes modais.

Por fim, tal síntese, por sua recente e recorrente aplicação na literatura brasileira, figura como referência para a formação de uma base de dados complementar à Pesquisa Origem e Destino, a ser comparada à alternativa proposta no presente artigo, apresentada na seção seguinte.

## 3. GOOGLE MAPS COMO ALTERNATIVA

Como opção às alternativas propostas por Hensher et al (2005) e em especial à síntese realizada por Lucinda, Meyer e Ledo (2013), Barcellos (2014), Lucinda et al (2015) e Feres (2015), propõe-se, no presente artigo, a utilização da ferramenta de pesquisa e visualização de mapas e rotas da empresa estadunidense Google, o Google Maps. Tal alternativa foi proposta apenas recentemente e em comparação à utilização de matrizes *skim*, pelo trabalho em desenvolvimento de Javanmardi et al (2015).

Dentre sua ampla variedade de recursos, importa à presente pesquisa o fato de tal ferramenta ser capaz de sugerir rotas entre origem e destino através de diversos modais. Sobre tais rotas, provê informações como custo e duração, ficando evidente sua possibilidade de utilização a fim de suprir a referida carência das Pesquisas Origem e Destino.

Em comparação com pesquisas de preferência declarada, a utilização de tal ferramenta apresenta um custo muito menor, além de poder apresentar informações mais precisas em relação aos trajetos que os indivíduos não costumam utilizar. Comparando-se à alternativa sintética, por sua vez, o potencial do Google Maps reside na utilização de maior quantidade e diversidade de informações, que podem ser ilustrados pelos milhões de usuários diários que a ferramenta possui, além da utilização de georreferenciamento e de informações sobre linhas de transporte público. Estas últimas informações, além de serem potencialmente importantes para a obtenção de estimações mais precisas, também permite que se componha o conjunto de escolhas disponível para o tomador de decisão de forma mais adequada.

Com a finalidade de obter tais dados de forma sistemática, foram utilizados outros recursos da própria Google, sendo: Google Maps API, Google Apps Script e Google Sheets. Com a utilização destes recursos, os dados foram então obtidos a partir das coordenadas geográficas fornecidas pela Pesquisa Origem e Destino, de tal forma a obter a duração prevista para viagens a partir de transporte individual motorizado, a pé e das três melhores opções de transporte público de acordo com o ordenamento do próprio

[3] A Pesquisa Origem e Destino da Região Metropolitana de São Paulo de 2007 divide a região em 460 zonas.

Google Maps. As alternativas de transporte público, quando iguais, foram agregadas na de menor duração.

É importante ressaltar que o Google Maps utiliza as definições das linhas de transporte mais recentes para o planejamento da rota. Dessa forma, programou-se para que os resultados que incluíssem linhas de transporte público adicionadas após a Pesquisa Origem e Destino 2007 fossem descartadas.

Enfim, com os dados coletados, foi montada uma base de dados complementar à Pesquisa Origem e Destino 2007 da RMSP de maneira análoga àquela sintetizada pelos referidos autores brasileiros. Devido às restrições quanto a quantidade de dados coletados diariamente, foi utilizado, para o presente artigo, um recorte da Pesquisa Origem e Destino 2007, composta de observações de 25995 indivíduos, referentes às viagens motivadas a trabalho, no horário de pico da manhã (das 6h às 9h59). Tal seleção é análoga à realizada por Swait e Ben-Akiva (1987) e também foi realizada sobre a amostra utilizando a formatação referência para a síntese de dados proposta por Hensher et al (2005).

## 4. DIFICULDADES DE COMPARAÇÃO

A fim de comparar as duas metodologias - a sintética e a coleta via Google Maps -, três formas principais são apresentadas: uma teórica, uma através de análise de erros e, enfim, uma através de verificação de ajuste e significância.

A comparação teórica, no sentido de análise das propriedades dos estimadores, aparece como primeira opção, seguindo a prática comum na literatura econométrica. A dificuldade em se obter o método de estimação da duração utilizado pelo Google Maps - uma espécie de caixa preta - impede, no entanto, tal abordagem.

A segunda alternativa de comparação refere-se à análise de erros, comparando erros médios e erros absolutos médios entre as estimações. No entanto, para obter valores para tais erros, é necessário fazer a comparação da estimação com os dados reais - no caso, com aqueles disponíveis na Pesquisa Origem e Destino. Assim, pode-se comparar apenas os tempos reais e os tempos estimados da mesma observação e, portanto, de um mesmo modal. Em decorrência do método utilizado pelos autores de utilizar a hora de partida e chegada para a estimação, é razoável supor que os erros possíveis de serem avaliados sejam menores do que aqueles referentes às estimações das alternativas não escolhidas, uma vez que não é possível utilizar-se de uma informação tão precisa para este caso.

A fim de exemplificar tal situação, pode-se considerar dois vizinhos que compartilhem o mesmo local de trabalho. O primeiro vizinho se locomove ao trabalho utilizando ônibus e, para que chegue às 8h40, deve sair às 6h40. O segundo vizinho, no entanto, vai de carro e, para chegar no horário do trabalho, deverá sair às 8h10. Desta forma, ainda que a estimação preveja perfeitamente a duração do deslocamento de ônibus para o primeiro vizinho, o fará porque utiliza os "inputs" corretos. Caso realize a previsão a partir dos "inputs" da viagem por carro, provavelmente o fará de maneira errônea, uma vez que, ao invés de utilizar a hora de partida adequada (6h), utilizará aquela

relativa ao carro (8h). Assim, é razoável supor que estime um tempo diferente e, com isso, um erro maior. Tal situação é ilustrada na Tabela 3.

TABELA 1 – EXEMPLO ILUSTRATIVO DA DIFICULDADE DE COMPARAÇÃO POR ANÁLISE DE ERROS

| Vizinhos | Modal Escolhido | Hora de Partida | Hora de Chegada | Duração Modal | Duração Estimada |
|---|---|---|---|---|---|
| Vizinho 1 | Ônibus | 6h | 8h | 2h | 2h |
| Vizinho 2 | Carro | 8h | 8h | 30min | ? 2h |

FONTE: elaboração própria

De fato, tal dificuldade não advém apenas da utilização de tais variáveis, mas é também decorrente de que tal estimação pode apresentar viés de seleção amostral. Conforme apontam Washington et al (2014, tradução nossa): “Imputar ou sintetizar valores de atributos de alternativas não escolhidas é um problema de dados faltantes. Em contraste aos Dados Faltantes Completamente Aleatórios e aos Dados Faltantes Aleatórios, atributos dos modais de transporte faltantes são uma função de seus estado não observado e não-ignorável”.

Desta forma, observa-se que, sendo um caso de Dados Faltantes Não Aleatórios, utilizar-se das informações sobre as viagens através de um modal fornecidas apenas por aqueles que o utilizaram, poderá implicar em viés de seleção quando da estimação para aqueles que não utilizaram tal meio de transporte.

Tal viés de seleção será ilustrado na seção seguinte, que apresentará breve comparação entre as amostras complementares estimadas, assim como comparação entre as estimações dos modelos de escolha discreta utilizando-se das diferentes bases de dados.

## 5. ANÁLISE E COMPARAÇÃO DAS ESTIMAÇÕES

A fim de comparar as alternativas para complementação da base de dados, deve-se, pois, comparar também os resultados das estimações dos modelos de escolha discreta. Dentre os trabalhos referidos que optam por sintetizar os atributos das alternativas não escolhidas, Lucinda, Meyer e Ledo (2013) e Barcellos (2014) apresentam tais informações de forma clara e completa e, desta forma, são aqui utilizados para comparação com as estimações obtidas com uso do Google Maps.

O recorte imposto à base de dados complementada com utilização desta ferramenta a diferenciou das subamostras utilizadas nos referidos artigos. Assim, optou-se por reestimar os valores das variáveis não escolhidas utilizando das mesmas especificações por eles apresentadas - a diferença, final, refere-se apenas a quais foram observações utilizadas para a estimação do modelo de escolha discreta.

Para melhor compreensão dos resultados, esta seção dividir-se-á em três partes, referindo-se primeiramente a uma breve comparação das bases de dados estimadas, seguindo à estimação realizada pelos autores citados e então à dos resultados dos modelos de escolha discreta.

### *5.1. Comparação entre Bases de Dados Estimadas*

A fim de ilustrar indícios de viés de seleção resultante da estimação, através de regressão linear, dos atributos das alternativas não escolhidas (no caso, as durações atraves de tais modais) a partir dos atributos das alternativas escolhidas, duas tabelas são aqui apresentadas.

A primeira delas, Tabela 4, apresenta um apanhado de 3 indivíduos da subamostra utilizada para a estimação dos modelos discretos, com suas respectivas alternativas possíveis, escolhas de modal e durações de viagem, em minutos, tanto da estimação realizada através de regressão, quanto daquela utilizando-se do Google Maps.

TABELA 2 – APANHADO ILUSTRATIVO PARA COMPARAÇÃO ENTRE BASES ESTIMADAS

| **ID do Respondente** | **Alternativa** | **Escolha** | **Duração Regressão** | **Duração Google Maps** |
|---|---|---|---|---|
| 10 | Carro | Não | 32 | 18 |
| 10 | Ônibus | SIm | 90 | 90 |
| 10 | Outros | Não | 22 | 89 |
| 11 | Carro | Não | 36 | 23 |
| 11 | Ônibus | Sim | 90 | 90 |
| 11 | Outros | Não | 25 | 115 |
| 12 | Carro | Não | 40 | 25 |
| 12 | Ônibus | Sim | 105 | 105 |
| 12 | Outros | Não | 27 | 141 |

FONTE: elaboração própria.

Embora outras alternativas de modais estejam disponíveis, foram apresentadas às de maior utilização pela população: carro, ônibus e outros. A partir da observação desses indivíduos, nota-se que as durações obtidas através do Google Maps para viagens de carro são recorrentemente inferiores àquelas estimadas. Isto pode decorrer do fato de que muitos dos indivíduos que não utilizem de automóvel, deixem de usá-lo porque, dada a curta distância do trajeto, optem pelo baixo custo de ir a pé ou de bicicleta (e de outras vantagens que tais modais podem apresentar), apesar de o tempo a ser despendido seja maior do que aquele utilizando-se de carro.

Ainda é importante notar que, ao se analisar os valores estimados para viagens utilizando-se dos modais agrupados em “Outros”, observa-se que os indivíduos apresentam, de acordo com a regressão, viagens com durações menores inclusive do que as viagens de carro. Além disso, apesar de a duração ser tão menor do que a duração de ônibus e não incorrer em custos, a opção dos indivíduos é a mesma por tal transporte coletivo.

Tais características são observadas também através da Tabela 5, que traz as médias de duração em minutos para cada subamostra de interesse.

Observa-se que a regressão estima uma média de tempo utilizando-se de "Outros" inferior àquela estimada para a utilização de automóveis. Ainda que esteja sintonizada com as médias encontradas na Pesquisa Origem e Destino, deve-se relembrar que esta subamostra corresponde apenas às viagens daqueles que optaram por viajar através destes modais. O viés de seleção, neste caso, parece subestimar a duração das viagens.

Enquanto as médias de duração através de automóveis reforçam a interpretação explorada anteriormente, as médias referentes às viagens de ônibus podem também indicar viés de seleção, porém em um sentido contrário - de afastamento entre os valores obtidos entre a subamostra com dados apenas da Pesquisa Origem e Destino e aquela complementada através do uso de regressão. Tal fato pode se dar em decorrência de as pessoas que se utilizam de ônibus o fazerem também pela disponibilidade de infraestrutura, sendo esta possivelmente menor para aqueles que não a utilizam. Desta forma, a regressão poderia obter parâmetros que subestimam o tempo de tal modal.

Por fim, a comparação entre a ordenação e magnitude dos valores imputados através da regressão e do Google Maps aponta que os resultados do último corroboram com o esperado pela intuição: as viagens de carro são as de menor duração, seguidas com alguma margem pelas de ônibus, enquanto as viagens utilizando-se do modal "Outros", como viagens a pé ou de bicicleta apresentam um tempo médio razoavelmente superior aos demais.

TABELA 3 - MÉDIAS DE DURAÇÃO POR SUBAMOSTRA DOS TRÊS MODAIS DE MAIOR UTILIZAÇÃO

| **Subamostra** | **Ônibus** | **Carro** | **Outros** |
|---|---|---|---|
| **Origem e Destino** | 64.6 | 36.3 | 20.4 |
| **Regressão** | 52.1 | 37.0 | 27.5 |
| **Google Maps** | 62.5 | 26.9 | 102.6 |

FONTE: elaboração própria.

### *5.2. Estimação dos Modelos de Escolha Discreta*

Tanto o artigo de Barcellos (2014) quanto de Lucinda et al (2013) utilizam-se dos mesmos modelos de escolha discreta: o Logit Multinomial e o Misto. Antes de prosseguir aos procedimentos práticos concernentes à estimação, é relevante uma breve apresentação de tais modelagens.

Estes modelos compreendem que existem fatores que coletivamente determinam a escolha dos agentes, mas que nem todos são observados pelo pesquisador. O processo de escolha, dessa forma, é expresso por uma função do tipo $y = h(x, \varepsilon)$, em que *x* refere-se aos fatores observados e ***ε*** aos fatores não observados. Contendo termos não observados, pois, não é possível prever *y* com exatidão. Ao invés disso, uma probabilidade de resultado é então derivada a partir da consideração de que os termos não observados são aleatórios com densidade ***f(ε)***. Assim, a probabilidade de um tal resultado, dados os fatores observados, pode ser expressa por: $P(y|x) = Prob(\varepsilon\ s.a.\ h(x, \varepsilon) = y)$.

Utilizando-se do arcabouço de maximização de utilidade, pode-se reexpressar a modelagem acima, assumindo que a *utilidade representativa*, $V_{nj}$, é determinada pelos atributos das $J$ alternativas encontradas pelos $N$ agentes, $x_{nj}$, assim como pelas características destes agentes, $s_n$: $V_{nj} = (x_{nj,}s_n)\ \forall\ j$. A utilidade representativa, no entanto, difere da utilidade real, $U_{nj}$, que possui um componente aleatório: $U_{nj} = V_{nj} + \varepsilon_{nj}$. Desta forma, a probabilidade do indivíduo $n$ escolher a alternativa $i$ em detrimento às outras alternativas é a probabilidade de que a utilidade obtida por tal indivíduo proveniente da escolha de $i$ seja maior do que as demais, tal como $P_{ni} = Prob(U_{ni} > U_{nj}, \forall\ j\ \neq i)$ (TRAIN, 2003).

O modelo Logit Multinomial, o mais simples e amplamente utilizado para este tipo de modelagem, assume que cada $\varepsilon_{nj}$ é um valor extremo independente e identicamente distribuído de acordo com uma distribuição Gumbel. Com esta distribuição de probabilidade, McFadden (1974) obteve a função de probabilidade de escolha Logit Multinomial, $P_{ni} = Prob\big(U_{ni} > U_{nj}, \forall\ j\ \neq i\big) = \frac{e^{V}_{ni}}{\sum_j e^{V}_{nj}}$.

Uma importante limitação deste modelo Logit decorre justamente das hipóteses sobre as características de probabilidades de escolha das quais originalmente Luce (1959) derivou sua formulação. Tais hipóteses constituem a propriedade de independência das alternativas irrelevantes (*independence from irrelevant alternatives* - IIA), segundo a qual, para duas alternativas *i* e *k*, a razão de probabilidade Logit não depende de outras alternativas que não *i* e *k*. Tal propriedade não apenas restringe os padrões de substituição possíveis do modelo, como também implica em erro de previsão sempre que a razão de probabilidades entre duas alternativas mudar com a introdução ou alteração de uma outra alternativa (TRAIN, 2003).

Este Logit básico apresenta duas outras importantes limitações, além desta decorrente da propriedade IIA. Embora possa representar variação de preferência sistemática, é incapaz de fazê-lo para variações aleatórias. Tal modelagem, ainda, não pode ser usada com dados em painel quando fatores observados são correlacionados ao longo do tempo para cada tomador de decisão. (TRAIN, 2003)

Superando tais limitações, foram desenvolvidos os modelos Logit Mistos que, sendo definidos sobre a base da forma funcional de suas probabilidades de escolha, permitem diferentes especificações comportamentais. A fim de ilustrar tal modelagem, assume-se aqui que $U_{nj} = \beta'_n x_{nj} + \epsilon_{nj}$, onde $x_{nj}$ são as variáveis observadas que se relacionam aos tomadores de decisão e às alternativas, $\beta_n$, é um vetor de coeficientes dessas variáveis para a pessoa *n*, e $\varepsilon_{nj}$ é um termo aleatório na população com densidade $f = (\beta)$. Como observado, apresenta, pois, a mesma especificação do Logit Multinomial, com exceção da variabilidade de beta entre os indivíduos. Desta forma, para este modelo, as probabilidades são as integrais de probabilidades Logit Padrão multiplicadas por uma função de densidade dos parâmetros: $P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}\right) f(\beta) d\beta$.

Realizada esta breve apresentação, pode-se prosseguir, então, aos procedimentos diretamente relacionados à estimação. Primeiramente, é importante especificar a forma como os autores agregaram os modais para o

fim da pesquisa - o fizeram constituindo como alternativas as seguintes categorias: automóvel, ônibus, metrô e trem, motocicleta, táxi e outros.

Por fim, quanto às variáveis explicativas, custo e tempo foram utilizadas em todos os modelos como específicas das alternativas. As variáveis dos indivíduos consideradas, por sua vez, foram:

- idade;
- renda individual (em R$1000);
- tamanho da família (medido em número de moradores no domicílio);
- *dummy* para emprego formal;
- *dummy* para gênero feminino;
- *dummy* para estudante.

Enquanto o modelo apresentado por Lucinda et al (2013) utilizou as variáveis apontadas, aquele desenvolvido por Barcellos (2014) incluiu também iterações de custo e tempo com a renda individual. A fim de estabelecer comparações, foi estimado um Logit Multinomial na mesma configuração de Barcellos (2014), e Logit Mistos nas configurações de ambos os trabalhos, utilizando-se da base de dados obtida com o Google Maps. Tais resultados são apresentados a seguir.

### *5.3. Comparação de Resultados dos Modelos de Escolha Discreta*

A primeira comparação é estabelecida entre o modelo Logit Multinomial estimado como Barcellos (2014) e aquele análogo estimado com uso do Google Maps. Tal comparação é apresentada na Tabela 6 e a referência à autora é mantida para fins de facilitar a compreensão, apesar de toda a estimação ter sido replicada no presente trabalho. É importante ressaltar que todas as estimações foram realizadas utilizando a alternativa “Outros” como referência e, portanto, todos os coeficientes apresentados o são em relação a este modal.

TABELA 4 - COMPARAÇÃO DAS ESTIMAÇÕES LOGIT MULTINOMIAL

| | **Google Maps** | **Barcellos (2014)** |
|---|---|---|
| **Variáveis Específicas das Alternativas** | | |
| Custo | -0.1262 *** | -0.1287 *** |
| Tempo | -0.0153 *** | 0.0337 *** |
| Custo X Renda | 0.0098 *** | 0.0143 *** |
| Tempo X Renda | 0.0002 | -0.0051 *** |
| **Ônibus** | | |
| Renda (R$1000) | -0.1360 *** | 0.0663 *** |
| Idade | -0.0032 *** | -0.0120 *** |

| | | |
|---|---|---|
| Mulher | 0.2298 *** | -0.0041 |
| Estudante | 0.2691 *** | 0.1957 *** |
| Tamanho  Família | -0.0219 *** | -0.0634 *** |
| Emprego Formal | 0.7831 *** | 0.4343 *** |
| **Trem e Metrô** | | |
| Renda (R$1000) | -0.0255 | 0.2819 *** |
| Idade | -0.0092 *** | -0.0261 *** |
| Mulher | 0.1457 *** | -0.2894  *** |
| Estudante | 0.2816 *** | 0.0785 * |
| Tamanho  Família | -0.0077 | -0.1201 *** |
| Emprego Formal | 1.1626 *** | 0.3112 *** |
| **Carro** | | |
| Renda (R$1000) | 0.3273 *** | 0.3822 *** |
| Idade | 0.0090 *** | 0.0104 *** |
| Mulher | -1.0953 *** | -1.1802 *** |
| Estudante | -0.2708 *** | 0.2521 *** |
| Tamanho  Família | -0.2169 *** | -0.1656 *** |
| Emprego Formal | 0.0895 ** | 0.4286 *** |
| **Moto** | | |
| Renda (R$1000) | 0.2116 *** | 0.2386 *** |
| Idade | -0.0610 *** | -0.0532 *** |
| Mulher | -2.5687 *** | -2.6767 *** |
| Estudante | -0.2798 *** | -0.1965 *** |
| Tamanho  Família | -0.1390 *** | -0.0861 *** |
| Emprego Formal | 0.2006 ** | 0.7708 *** |
| **Táxi** | | |
| Renda (R$1000) | 0.2773 *** | 0.2160 *** |
| Idade | -0.0134 | -0.0092 |
| Mulher | -0.7819 * | -0.3800 |

| | | |
|---|---|---|
| Estudante | -0.9482 | -0.1456 |
| Tamanho Família | -0.8544 *** | -1.1675 *** |
| Emprego Formal | -0.8846 | -0.2484 |
| Log-Verossimilhança (Convergência) | -26778 | -33630 |
| $R^2$ de McFadden | 0.29044 | 0.10889 |

Fonte: elaboração própria.
p < 0.1: *; p < 0.05: **; p < 0.01 ***

Inicialmente, observa-se um maior valor para o log-verossimilhança de convergência para o modelo imputado utilizando-se do Google Maps. O mesmo ocorre para o $R^2$ de McFadden que, análogo ao $R^2$ tradicional, avalia a significância do modelo estimado. Tal consideração indica uma maior capacidade de previsão e ajustamento do modelo utilizando a base de dados proposta, e vai no mesmo sentido dos resultados obtidos por Javanmarti et al (2015) em comparação à utilização de matrizes *skim*.

Além de tal resultado, é relevante notar que a estimação da forma de Barcellos (2014) gerou um coeficiente com sinal positivo para a variável tempo, enquanto a alternativa proposta apresentou o sinal negativo. O sinal da última corrobora com o esperado intuitivamente: a utilidade dos indivíduos diminui conforme a duração da viagem aumenta. O coeficiente imputado para a iteração entre tempo e renda, no entanto, diminui o problema apresentado no modelo estimado com uso de regressão: com o sinal negativo e estatisticamente significante, indica que, conforme a renda aumenta, a duração passa a ser menos apreciada, até que gere desutilidade.

Este resultado pode ser proveniente da subestimação da duração para o modal “Outros” quando da utilização da regressão linear. Como discutido anteriormente, tal subestimação pode, por sua vez, advir do viés de seleção amostral por utilizar-se apenas das alternativas escolhidas para a estimação dos atributos das alternativas não escolhidas. Corroborando com o apresentado em 5.1, a base de dados imputada com uso do Google Maps parece sofrer menos de tal dificuldade.

Quanto às variáveis relacionadas aos indivíduos, é relevante observar que os coeficientes relacionados aos gênero feminino apresentaram alterações tanto para ônibus, quanto para trem e metrô e táxi. Tais coeficientes passaram a ser estatisticamente significantes com uso da base de dados do Google Maps para os modais ônibus e táxi. O sinal dos coeficientes, no entanto, passaram a ser positivos para ônibus e trem e metrô, e negativos para táxi. Ainda que possam ser contraintuitivos, tal fato pode ser explicado pela tradicional estrutura da família brasileira que, colocando o homem como “chefe de família”, favorece a utilização do homem pelo carro (ou o táxi), restando à mulher os modais alternativos. Diversos estudos, como em Swait (1987), utilizam da variável “chefe de família” para melhor especificar seus modelos. Ainda que esta pudesse ser uma alternativa a ser explorada, optou-se por ser o mais fidedigno possível aos estudos de Barcellos (2014) e Lucinda et al (2013).

Por fim, é importante ressaltar que os coeficientes de renda e tamanho da família relacionados ao modal Trem e Metrô deixaram de ser estatisticamente significantes no modelo que utiliza da base de dados proposta. Por outro lado, os coeficientes de renda da alternativa ônibus e da *dummy* de estudante para a alternativa carro passaram a ter o sinal esperado: negativo.

Finalizando assim a comparação entre os modelos Logit Multinomial, resta analisar os modelos Logit Mistos. Assim, apresenta-se na Tabela 7 a comparação entre as estimações utilizando tal modelagem: a de Barcellos (2014), a de Lucinda et al (2013) e duas estimações utilizando a base de dados do Google Maps, cada uma utilizando uma das especificações de variáveis de cada uma das estimações referidas. Estas estimações estão na coluna central, sendo que se comparam aos modelos apresentados na coluna mais próxima. Para exemplificar, vemos que o coeficiente estimado para renda usando a base de dados e especificação como Barcellos (2014) foi de 0.0164, não sendo estatisticamente significante em níveis abaixo de 10%. Vemos também que, utilizando da base complementar estimada com auxílio do Google Maps, o coeficiente para renda, no modelo de especificação análogo ao de Barcellos (2014) foi de -0.1214, enquanto na especificação análoga à de Lucinda et al (2013) foi de -0.1212 - sendo ambos estatisticamente significantes a 1%. Por fim, o coeficiente para renda com base de dados e especificação como Lucinda et al (2013) foi de -0.0271 e não pode ser considerado estatisticamente significante a níveis menores do que 10%.

TABELA 5 - COMPARAÇÃO ENTRE ESTIMAÇÕES LOGIT MISTAS

| | **Barcellos (2014)** | **Google Maps** | **Lucinda et al (2013)** |
|---|---|---|---|
| **Fixos** | | | |
| **Ônibus** | | | |
| Renda (R$1000) | 0.0164 | -0.1214 *** / -0.1212 *** | -0.0271 |
| Idade | -0.0190 *** | -0.0051 *** / -0.0039 *** | -0.0188 *** |
| Mulher | 0.0292 | 0.2750 *** / 0.2741 *** | 0.0319 |
| Estudante | 0.2152 *** | 0.3080 *** / 0.3108 *** | 0.2178 *** |
| Tamanho Família | -0.0994 *** | -0.0249 *** / -0.0204 *** | -0.0978 *** |
| Emprego Formal | 0.2032 *** | 0.8609 *** / 0.8883 *** | 0.2143 *** |
| **Trem e Metrô** | | | |
| Renda (R$1000) | 0.2484 *** | -0.0282 / 0.0278 | 0.2006 *** |
| Idade | -0.0346 *** | -0.0095 *** / -0.0093 *** | -0.0343 *** |
| Mulher | -0.2660 *** | 0.1926 *** / 0.1918 *** | -0.2622 |

| | | | |
|---|---|---|---|
| Estudante | 0.0614 | 0.2949 *** / 0.3138 *** | 0.0648 |
| Tamanho Família | -0.1653 *** | 0.0035 / 0.0059 | -0.1636 *** |
| Emprego Formal | 0.0816 | 1.3515 *** / 1.3424 *** | 0.0944 * |
| **Carro** | | | |
| Renda (R$1000) | 0.3593 *** | 0.3362 *** / 0.3551 *** | 0.3566 *** |
| Idade | 0.0100 *** | 0.0108 *** / 0.0109 *** | 0.0101 *** |
| Mulher | -1.1891 *** | -1.2124 *** / -1.231 *** | -1.1915 *** |
| Estudante | -0.2906 *** | -0.3346 *** / -0.3405 *** | -0.2923 *** |
| Tamanho Família | -0.1786 *** | -0.2405 *** / -0.2403 *** | -0.0942 *** |
| Emprego Formal | 0.3369 *** | 0.0532 / 0.0544 | 0.3292 *** |
| **Moto** | | | |
| Renda (R$1000) | 0.2318 *** | 0.2702 *** / 0.2239 *** | 0.2342 *** |
| Idade | -0.0528 *** | -0.0733 *** / - 0.0725 *** | -0.0528 *** |
| Mulher | -2.6822 *** | -2.9481 *** / -3.033 *** | -2.6812 *** |
| Estudante | -0.2024 *** | -0.5610 *** / -0.4892 *** | -0.2027 *** |
| Tamanho Família | -0.0939 *** | -0.1311 *** / -0.1335 *** | -0.0942 *** |
| Emprego Formal | 0.7611 *** | 0.3016 *** / 0.2789 *** | 0.7596 *** |
| **Táxi** | | | |
| Renda (R$1000) | 0.3279 *** | 0.9402 *** / 0.9945 *** | 0.5066 *** |
| Idade | -0.0063 | 0.0482 *** / 0.0506 *** | -0.0087 |
| Mulher | -0.7040 | -0.6523 / -0.1605 | -0.7594 * |
| Estudante | -1.3724 | -17.951 *** / -21.475 *** | -1.4547 |
| Tamanho Família | -1.1672 *** | -3.4182 *** / -3.7853 *** | -1.1782 *** |
| Emprego Formal | -1.3238 ** | -13.492 *** / -11.728 *** | -1.4133 *** |
| | | | |
| Custo X Renda | 0.0084 *** | 0.0072 * / - | - |
| Tempo X Renda | -0.0034 *** | -0.0010 *** / - | - |
| Custo | -0.1926 *** | -0.1122 *** / -0.1933 *** | -0.0629 *** |

| | | | |
|---|---|---|---|
| Tempo | -0.0173 *** | -0.0116 *** / -0.0189 *** | 0.0554 *** |
| **Desvios-padrões** | | | |
| Custo | 0.0963 *** | 0.0453 *** / 0.0825 *** | 0.0008 |
| Tempo | 0.0023 | 0.0068 *** / 0.0039 *** | 0.0445 *** |
| Log-Verossimilhança (Convergência) | -32082 | -27377 / -27403 | -32096 |
| $R^2$ de McFadden | 0.14988 | 0.27457 / 0.27389 | 0.14951 |

Fonte: elaboração própria.
p < 0.1: *; p < 0.05: **; p < 0.01 ***

Assim como observado para os modelos anteriores, observa-se de imediato um maior valor para log-verossimilhança de convergência para os dois modelos que utilizaram de bases de dados complementadas com a utilização do Google Maps.

Quanto ao coeficiente de tempo, observa-se que o Logit Misto na formatação de Barcellos (2014) logrou superar a dificuldade anterior, apresentando agora sinal negativo. Aquele que seguiu Lucinda et al (2013), no entanto, manteve o problema. É visto, ainda, que o modelo estimado com a base de dados alternativa e a formatação de Barcellos (2014) passou a apresentar coeficientes estatisticamente significativos e com o sinal esperado para as variáveis iteradas.

Apesar da melhoria apresentada, alguns coeficientes deixaram de ser significantes, como estudante em relação a trem e metrô para ambos os modelos que utilizaram base de dados estimadas com regressão linear. Também deixaram de ser significantes o coeficiente de emprego formal para trem e metrô na estimação como Barcellos (2014) e para carro nas estimações com uso do Google Maps. Ainda, o coeficiente relacionado às mulheres com relação a táxi manteve sua significância apenas no modelo análogo ao de Lucinda et al (2013).

Por fim, é relevante destacar as alterações observadas no modelo de base de dados alternativa quanto ao táxi. Para este modal, os coeficientes de idade e estudante passaram a ser estatisticamente significantes, além de apresentarem sinais razoáveis: quanto maior a idade, maior a utilidade de se utilizar de tal transporte, enquanto ser estudante diminui, e muito, a probabilidade de escolhê-lo.

## 6. CONCLUSÃO

O presente artigo propôs uma alternativa às práticas da literatura quanto à dificuldade de lidar com a ausência dos atributos das alternativas não escolhidas em pesquisas de preferência revelada do tipo Origem e Destino. Ainda que dificuldades tenham sido encontradas para a comparação entre a alternativa proposta, a utilização do Google Maps, e a síntese desenvolvida por Lucinda et al (2013) e Barcellos (2014), a estimação dos modelos de escolha discreta apresentou resultados com melhores propriedades estatísticas com uso da alternativa aqui proposta. Ressalta-se

novamente que os achados corroboram com as semelhantes vantagens percebidas por Javanmarti el al (2015).

É razoável crer que a contribuição da alternativa proposta oscile, a depender das características das informações possuídas pelo Google Maps para cada localidade. Em cidades com menor utilização de ferramentas Google ou relacionadas, como as mais afastadas cidades do interior, a quantidade e qualidade da informação provavelmente serão menos interessantes do que aquelas obtidas em grandes cidades, como São Paulo e Curitiba. No entanto, as poucas barreiras quanto à construção de tal base de dados complementar permite que seja avaliada sua qualidade aplicado-a em novos estudos.

É importante destacar que tal banco de dados possui outras vantagens para além dos aspectos apontados da complementação. Como observado na análise dos resultados, a agregação de metrô e trem em uma mesma alternativa pode representar um problema, pois é razoável crer que as características dos usuários de tais modais sejam bastante diferentes. Tal agregação é justificada pela pequena amostra de usuários de trem, que pode ser resolvida com a amplitude de informações obtidas sobre este modal com o Google Maps - mesmo para aqueles que não o escolhem. Outra vantagem potencial da alternativa proposta reside na melhor especificação da cesta de escolhas disponível para cada indivíduo, decorrente do uso de informações georreferenciadas sobre as linhas de transporte público. Tais vantagens apontadas ilustram, mas não encerram o potencial por ela apresentado.

Deve-se ressaltar, por fim, que outras especificações de modelos poderiam ter sido exploradas, mas a critério de comparação, optou-se por manter adesão às estimações realizadas por Barcellos (2014) e Lucinda et al (2013). Ademais, a base de dados pode ser explorada de outras formas além da estimação direta para modelos de escolha discreta. Seguindo os trabalhos de Washington et al (2014), é razoável a utilização de tal base como distribuição a priori a realizar a imputação Bayesiana pelos autores desenvolvida.

## 7. REFERÊNCIAS


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. Dissertação (Mestrado) - Departamento de Economia, Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, 2014.

BEN-AKIVA, M., LERMAN, S. R. **Discrete Choice Analysis: Theory and Application to Travel Demand.** MIT Press, Cambridge, Massachusetts, 1985.

EMPLASA, **Por Dentro da Região Metropolitana de São Paulo – RMSP**. São Paulo, São Paulo, Junho/2011.

FERES, L. J. *Diferencial de tarifa entre pico e vale como ferramenta de suavização da demanda no sistema de transporte público da cidade de São*

*Paulo*. Dissertação (Mestrado Profissional) - Programa de Mestrado Profissional em Economia, Insper Instituto de Ensino e Pesquisa, São Paulo, 2015.

HADDAD, E. A.; HEWINGS, G. J. D.; PORSSE, A. A.; VAN LEUWEN, E.; VIEIRA, R. S. The Underground Economy: Tracking the Higher-order Economic Impacts of the São Paulo Subway System. **Transportation Research Part A: Policy and Practice**, v.73, p 18-30, 2015.

HENSHER, D. A., ROSE, J. M., GREENE, W. H. **Applied Choice Analysis: A Primer.** Cambridge University Press, New York, 2005.

IBOPE. 9ª Pesquisa sobre Mobilidade Urbana - Rede Nossa São Paulo. *Instituto Brasileiro de Opinião e Estatística, Opinião Pública*. São Paulo, set. de 2015. Disponível em: < http://www.nossasaopaulo.org.br/pesquisas/pesquisaibope2015completa.pdf >. Acessado em: ago. 2016.

JAVANMARDI, M.; LANGERUDI, M. F.; ANBARANI, R. S.; MOHAMMADIAN, A. K. *Mode Choice Modelling Using Personalized Travel Time and Cost Data.* Artigo apresentado em 14th International Conference on Travel Behavior Research, England, July 19-23, 2015.

LUCINDA, C. R.; MEYER, L. G.; LEDO, B. A. *Urban Road Tax in a Large Emerging Market: Some Brazilian Evidence*. Artigo apresentado em 35º Encontro Brasileiro de Econometria. Foz do Iguaçu, 12 dez. 2013. Disponível em: < https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=sbe35&paper_id=11 >. Acesso em: ago 2016.

LUCINDA, C. R.; MOITA, R. M. S; MEYER, L. G.; LEDO, B. A. *The Economics of Sub-optimal Policies for Traffic Congestion*. Working Paper 83, Rede de Economia Aplicada, set. 2015.

METRÔ. **Pesquisa Origem e Destino 2007**. [s.l: s.n.].

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*. TD NEREUS, São Paulo, abr. 2016.

SWAIT, J. **Probabilistic choice set formation in transportation demand models**. Tese (Doutorado) - Departamento de Engenharia Civil, MIT, 1984.

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.21B, p. 103-115, 1987.

WASHINGTON, S.; RAVULAPARTHY, S.; ROSE, J. M.; HENSHER, D.; PENDYALA, R. *Bayesian Imputation of Non-Chosen Attribute Values in*

*Revealed Preference Surveys.* **Journal of Advanced Transportation**. 48:48-65. 2014.

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.** 402-410. 2010.