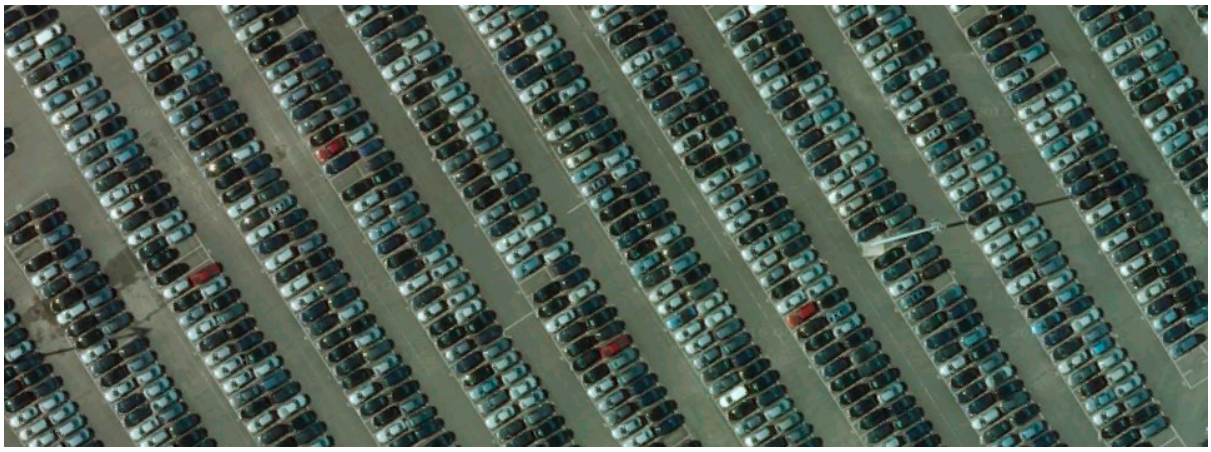


Uncovering substitution patterns in new car sales using a cross nested logit model

Anna Fernández Antolín
Matthieu de Lapparent
Michel Bierlaire

Transportation and Mobility Laboratory, ENAC, EPFL

May 2016

STRC

16th Swiss Transport Research Conference

Monte Verità / Ascona, May 18 – 20, 2016

Transportation and Mobility Laboratory, ENAC, EPFL

Uncovering substitution patterns in new car sales using a cross nested logit model

Anna Fernández Antolín, Matthieu de Lapparent, Michel Bierlaire
Transport and Mobility Laboratory
Ecole Polytechnique Fédérale de Lausanne
Station 18
CH-1015 Lausanne
phone: +41-21-693 24 35
fax: +41-21-693 80 60
{anna.fernandezantolin,matthieu.delapparent,michel.bierlaire}@epfl.ch

May 2016

Abstract

In the context of sales of new cars it is important to understand and model the consumers' substitution patterns as well as their price elasticity towards different types of cars. To do so we develop (i) a multinomial logit model (MNL) and (ii) a cross-nested logit model (CNL) and compare the obtained results. The evolution of the market shares following an increase of the price of one of the alternatives is studied. This is, to the best of our knowledge, the first time that a cross-nested logit model is used to model car-type choice.

For modeling purposes, a choice of car is considered to be a choice of market segment (small, medium, full, luxury, off road or multi-purpose vehicle) and a fuel type (gas, diesel, hybrid or electric). For the imputation of the attributes of the unchosen alternatives bootstrapping techniques are used. The model includes attributes of the car such as price and power and socioeconomic characteristics of the respondent such as gender, age, income level, occupation and education.

Keywords

Car-type choice, Cross nested logit, Substitution patterns

1 Introduction

The automobile sector is of interest for both the public and the private sectors. Governments and other public actors need to understand the car market in order to have valid forecasts of energy consumption, emission levels and even tax revenue. By means of these forecasts they can also derive optimal policy measures to, for instance, incentive the use of electric vehicles to reduce emissions.

It is also interesting for private companies. The interest from automobile firms is obvious, but the car market is linked to many other sectors such as those providing the raw materials – steel, chemicals, textiles – and those working with automobiles – repair and mobility services –. Moreover, according to the European Commission, "the EU is among the world's biggest producers of motor vehicles and the sector represents the largest private investor in research and development (R&D)" [1]

In order to satisfy the needs of these public and private actors it is important to model car ownership, which has many dimensions. Car ownership models can be classified on several criteria according to de Jong *et al.* (2004) such as: **i)** the inclusion of supply and demand, **ii)** the aggregation level, **iii)** dynamic or static, **iv)** long-term or short-term forecasts, **v)** inclusion of car-use and other socioeconomic characteristics, and **vi)** private or business cars among others. In this paper we focus on disaggregate car-type choice model.

The remaining of this paper is structured as follows: the state of the art is presented in Section 2, followed by the methodology used which is discussed in Section 3. In Section 4 we present the case study and Section 5 contains the concluding remarks and future work.

2 State of the art

As mentioned in Section 1, we focus in static disaggregate car-type choice models. There has been a lot of research in this field in the past, but not so much recently. For a complete review of the literature the reader is refered to de Jong *et al.* (2004) and Anowar *et al.* (2014).

In order to estimate a discrete choice model, one of the first things that needs to be defined is the choice set. It is clear that a choice of a private car is a discrete choice, but there is no consensus in the literature in how to define the alternatives. The two main approaches are defined below.

[1] http://ec.europa.eu/growth/sectors/automotive

A first approach is to consider that a combination of make, model, engine and vintage is what defines a car (Birkeland and Jordal-Jorgensen, 2001). Then, for a given year, there are oven 1,000 alternatives. In this case, sampling of alternatives is usually required. However, new results from Mai *et al.* (2015) show that large MEV models can be estimated in relatively low time. They estimate a cross-nested logit model with 500,000 alternatives, 200 nests and 210 parameters in 4.3 hours on an Intel(R) 3.2GHz machine using a non-parallelized code and simulated data. It would therefore be interesting to test this approach with real data. This is left for future research.

The second approach consists in an aggregation of the previous. An example by Page *et al.* (2000) consists in considering that a car is a combination of engine size and fuel type. They have nine alternatives for petrol and seven for diesel. In this case instead of the 1,000 alternatives there are only 16, which simplifies the specification and estimation of the model. This approach is also justifiable from a behavioral point of view.

In terms of the type of discrete choice model , the Multinomial Logit (MNL) is the most used (Wu *et al.,* 1999, Choo and Mokhtarian, 2004). However, MNL models satisfy the Independent of Irrelevant Alternatives (IIA) property, which leads to counterintuitive results when alternatives share unobserved attributes, which might be the case in car-type choice –no matter which of the previous two approaches is chosen–. For this reason, mixed logit models, that overcome the IIA property, have also been used in the literature (Brownstone and Train, 1998, McFadden *et al.,* 2000, Potoglou, 2008), as well as nested logit models (Berkovec and Rust, 1985, McCarthy and Tay, 1998, Mohammadian and Miller, 2002, 2003, Cao *et al.,* 2006).

3 Methodology

In order to model car-type choice we use discrete choice models based on expected maximum utility. This section contains a summary of the models used in Section 4. First, the concept of expected maximum utility is introduced in Section 3.1, followed by the logit model in Section 3.2, the nested logit in Section 3.3 and the cross nested logit in Section 3.4.

3.1 Expected maximum utility

We consider an individual n that faces the choice between different alternatives within his choice set C_n. The choice set contains all the alternatives i that are available to the individual. Each alternative is associated to a random utility that can also depend on the individual U_{in}. U_{in} can

be decomposed in a deterministic part V_{in} and a random part ε_{in} that captures the unobserved components as follows:

$$U_{in} = V_{in} + \varepsilon_{in}. \tag{1}$$

The deterministic part of the utility function, V_{in}, is defined as a function of attributes of alternative i and socioeconomic variables related to individual n. The individual is then assumed to select the alternative associated to the highest utility:

$$\mathrm{P}(i \mid C_n) = Pr(U_{in} \geq U_{jn}, \quad \forall j \in C_n). \tag{2}$$

By substituting Equation (1) in Equation (2) we obtain:

$$\mathrm{P}(i \mid C_n) = Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}, \quad \forall j \in C_n), \tag{3}$$

and by rearranging the terms:

$$\mathrm{P}(i \mid C_n) = Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}, \quad \forall j \in C_n), \tag{4}$$

which is a cumulative distribution function of $\varepsilon_{jn} - \varepsilon_{in}$. Different assumptions about this distribution will lead to different choice probabilities, which will define different models. The logit, nested logit and cross-nested logit are explained below.

3.2 Logit model

The logit model is the most widely used. It is derived when $\varepsilon_{in}, \varepsilon_{jn}$ are assumed to follow an independently and identically distributed Extreme Value. In this case, the expression from Equation (4) becomes:

$$\mathrm{P}(i \mid C_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}, \quad \forall i \in C_n. \tag{5}$$

For the proof of this, we refer the interested reader to Bierlaire *et al.* (2015). For normalization reasons μ is usually set to one.

The logit model has the advantage that it is closed form, and that the likelihood function associated to it is convex –as long as V_{in} is linear in parameters–. However it also has some limitations. One of the most discussed limitations of the logit model is that it satisfies the Independent from Irrelevant Alternatives (IIA) property.

An illustration of the consequences of this property is the *blue bus/red bus* paradox. It can be shown that if we consider a choice set with two alternatives, car and bus, where the only element in the deterministic part of the utility function is travel time, and travel time is equal for both modes as follows:

$$\begin{aligned} U_{\text{car}} &= \beta t + \varepsilon_{\text{car}}, \\ U_{\text{bus}} &= \beta t + \varepsilon_{\text{bus}}. \end{aligned} \tag{6}$$

The choice probabilities are then:

$$\text{P(Bus} \mid \{\text{Car, Bus}\}) = \text{P(Car} \mid \{\text{Car, Bus}\}) = \frac{e^{\beta t}}{e^{\beta t} + e^{\beta t}} = \frac{1}{2} \tag{7}$$

If we now introduce a new alternative, the *red bus* as opposed to the *blue bus* that we had before in the choice set, and define the utilities as:

$$\begin{aligned} U_{\text{car}} &= \beta t + \varepsilon_{\text{car}}, \\ U_{\text{red\_bus}} &= \beta t + \varepsilon_{\text{red\_bus}}, \\ U_{\text{blue\_bus}} &= \beta t + \varepsilon_{\text{blue\_bus}}. \end{aligned} \tag{8}$$

The choice probabilities become:

$$\begin{aligned} &\text{P(red\_bus} \mid \{\text{Car, red\_bus, blue\_bus}\}) = \text{P(blue\_bus} \mid \{\text{Car, red\_bus, blue\_bus}\}) = \\ &\text{P(car} \mid \{\text{Car, red\_bus, blue\_bus}\}) = \frac{e^{\beta t}}{e^{\beta t} + e^{\beta t} + e^{\beta t}} = \frac{1}{3} \end{aligned} \tag{9}$$

The conclusion is that by painting half of the buses of the city, you would increase the share of public transportation from 50% to 66%. This due to the unobserved correlation between two of the alternatives (the blue bus and the red bus), which is not accounted for in the model.

From a formal point of view, Bierlaire *et al.* (2015) describe it as follows, "the IIA property holds that for a specific individual the ratio of the choice probabilities of any two alternatives is entirely unaffected bu the presence (or absence) of any other alternatives in the choice set and by the systematic utilities of any other alternatives." The main assumption that leads to this property is the mutual independence of the error terms.

3.3 Nested logit

The nested logit model allows to overcome the IIA property by changing the assumptions of the error terms. Suppose that the choice set C is partitioned in $\mathcal{M}$ nests, $C_1, C_2, ..., C_M$. The partition of the choice set of individual n, $C_n \subseteq C$ is denoted as $C_{1n}, ..., C_{Mn}$ and defined as $C_{mn} = C_m \cap C_n$.

By defining the following marginal probability:

$$\mathrm{P}(i \mid C_n) = \sum_{m=1}^{M} Pr(i \mid C_{mn}, C_n) Pr(C_{mn} \mid C_n), \tag{10}$$

where $Pr(i \mid C_{mn}, C_n)$ is the probability for individual n to select alternative i within the nest C_m, and $Pr(C_{mn} \mid C_n)$ is the probability to select an alternative in nest C_m, and by making several assumptions on the error terms (see Bierlaire *et al.* (2015)) we obtain the expression:

$$\mathrm{P}(i \mid C_n) = \frac{e^{\mu_m V_{in}}}{\sum_{j \in C_{mn}} e^{\mu_m V_{jn}}} \cdot \frac{\left(\sum_{l \in C_{mn}} e^{\mu_m V_{ln}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{l \in C_{pn}} e^{\mu_p V_{ln}}\right)^{\frac{\mu}{\mu_p}}}. \tag{11}$$

where μ is the scale of the model, and $\mu_i, \quad i = 1, ..., M$ –that are related to the assumptions of the error terms– can be interpreted as the scale of each nest. For details on the normalization of the μ parameters the reader is referred to Bierlaire *et al.* (2015).

3.4 Cross nested logit

The nested logit model can be further generalized by assuming that an alternative can belong to more than one nest. The parameters α_{im} are introduced, and they represent the degree of membership of alternative i to nest C_m. For identification purposes it is bounded between 0 and 1. It is easy to see that the nested logit model is a special instance of the cross-nested logit model, when the α_{im} parameters take value 1 when alternative i belongs to nest m and 0 otherwise The expression of the choice probabilities for a cross-nested logit model are:

$$\mathrm{P}(i \mid C_n) = \sum_{m=1}^{M} \frac{\left(\sum_{j \in C_n} \alpha_{jm}^{\frac{\mu_m}{\mu}} e^{\mu_m V_{jn}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j \in C_n} \alpha_{jp}^{\frac{\mu_p}{\mu}} e^{\mu_p V_{jn}}\right)^{\frac{\mu}{\mu_p}}} \cdot \frac{\alpha_{im}^{\frac{\mu_m}{\mu}} e^{\mu_m V_{in}}}{\sum_{j \in C_n} \alpha_{jm}^{\frac{\mu_m}{\mu}} e^{\mu_m V_{jn}}}. \tag{12}$$

4 Case study

In order to model car-type choice we have data on sales of new cars in France during 2014. The dataset contains over 40,000 purchases. However, after selecting the variables that we use in the models and removing the missing values for any of them, only 20,000 observations are left. It is future work to try to recover some of these missing values.

In order to use discrete choice models we need to define the choice set for each individual, which is discussed in Section 4.1. It is followed by the description of the model specification in Section 4.2. In section 4.3 we describe how the imputation of the attributes of the unchosen alternatives is performed and in Section 4.4 the nesting structure that is used is introduced. Finally Section 4.5 contains the discussion of the results obtained.

4.1 Choice set

In the context of car-type choice, the definition of the choice set is not straight forward, as discussed in Section 2. In our approach we decide to consider a car-type as a combination between a market segment and a fuel type. The market segments we consider are full, luxury, medium, multipurpose vehicles (MPV), off-road and small. The fuel types considered are hybrid, diesel, petrol and electric.

Since for hybrid vehicles there are very few observations and they are always combined with either diesel or petrol, we consider *hybrid* as a market segment rather than as a fuel type. Therefore, we have a total of 15 alternatives summarized in Table 1. Note that for electric vehicles, the only market segment available is *small*. This is due to a data limitation, there are no observations in our dataset for any other market segment.

4.2 Model specification

The attributes and socioeconomic variables considered in our model after trying different specifications are:

- Reported fuel consumption [l/100km]
- Engine power [bhp]
- Reported price after discounts and government schemes [€]
- Reported range [km] (only for electric vehicles)
- Income [€]
- Number of adults in the household
- Number of children in the household
- Residential location: agglomeration vs rural areas
- Education level: university vs. no university

Note that both fuel consumption and price are reported values instead of catalog values. Therefore, these variables contain some measurement errors that might cause endogeneity. Auxiliary

Alternative	Market segment	Fuel type
1	Full	Diesel
2	Luxury	Diesel
3	Medium	Diesel
4	MPV	Diesel
5	Off-road	Diesel
6	Small	Diesel
7	Hybrid	Diesel
8	Full	Petrol
9	Luxury	Petrol
10	Medium	Petrol
11	MPV	Petrol
12	Off-road	Petrol
13	Small	Petrol
14	Hybrid	Petrol
15	Small	Electric

Table 1: List of elements in the choice set

models for the real price and fuel consumption can be estimated and integrated with the choice model to solve this. Moreover, these auxiliary models can also allow to recover the missing variables caused by missing prices and missing reported fuel consumption. This is considered future work.

Tables 2 and 3 show the model specification. Note that both price and fuel consumption are interacted with income. The fuel consumption is also multiplied by the mean fuel price(diesel or petrol) during 2014 in France. The rest of variables appear in a linear form in the model specification. Note also that the specification remains unchanged for the logit, and cross-nested logit that are presented in Section 4.5.

Parameter	1	2	3	4	5	6	7	8
ASC_{full}	1	0	0	0	0	0	0	1
ASC_{luxury}	0	1	0	0	0	0	0	0
ASC_{medium}	0	0	1	0	0	0	0	0
ASC_{MPV}	0	0	0	1	0	0	0	0
$ASC_{offroad}$	0	0	0	0	1	0	0	0
ASC_{petrol}	0	0	0	0	0	0	0	1
$ASC_{electric}$	0	0	0	0	0	0	0	0
β_{inc_full}	$\frac{income}{10000}$	0	0	0	0	0	0	$\frac{income}{10000}$
β_{inc_luxury}	0	$\frac{income}{10000}$	0	0	0	0	0	0
β_{inc_medium}	0	0	$\frac{income}{10000}$	0	0	0	0	0
β_{inc_MPV}	0	0	0	$\frac{income}{10000}$	0	0	0	0
$\beta_{inc\_offroad}$	0	0	0	0	$\frac{income}{10000}$	0	0	0
β_{inc_hybrid}	0	0	0	0	0	0	$\frac{income}{10000}$	0
$\beta_{nbr\_adults\_small}$	0	0	0	0	0	nbr. adults	0	0
$\beta_{nbr\_chilren\_small}$	0	0	0	0	0	nbr. child.	0	0
$\beta_{nbr\_cars\_lux}$	0	nbr. cars	0	0	0	0	0	0
$\beta_{nbr\_cars\_hybrid}$	0	0	0	0	0	0	nbr. cars	0
$\beta_{university}$	0	0	0	0	0	0	1	0
$\beta_{town\_rural\_EV}$	0	0	0	0	0	0	0	0
$\beta_{town\_rural\_hybrid}$	0	0	0	0	0	0	town_rural	0
β_{price_inc}	$\frac{price_1 \cdot 100}{income}$	$\frac{price_2 \cdot 100}{income}$	$\frac{price_3 \cdot 100}{income}$	$\frac{price_4 \cdot 100}{income}$	$\frac{price_5 \cdot 100}{income}$	$\frac{price_6 \cdot 100}{income}$	$\frac{price_7 \cdot 100}{income}$	$\frac{price_8 \cdot 100}{income}$
β_{conso_inc}	$\frac{cons_1 \cdot pd \cdot 100}{income}$	$\frac{cons_2 \cdot pd \cdot 100}{income}$	$\frac{cons_3 \cdot pd \cdot 100}{income}$	$\frac{cons_4 \cdot pd \cdot 100}{income}$	$\frac{cons_5 \cdot pd \cdot 100}{income}$	$\frac{cons_6 \cdot pd \cdot 100}{income}$	$\frac{cons_7 \cdot pd \cdot 100}{income}$	$\frac{cons_8 \cdot pp \cdot 100}{income}$
β_{max_power}	$\frac{max\_power_1}{10}$	$\frac{max\_power_2}{10}$	$\frac{max\_power_3}{10}$	$\frac{max\_power_4}{10}$	$\frac{max\_power_5}{10}$	$\frac{max\_power_6}{10}$	$\frac{max\_power_7}{10}$	$\frac{max\_power_8}{10}$
β_{range_EV}	0	0	0	0	0	0	0	0

Table 2: Model specification (part 1/2)

Parameter	9	10	11	12	13	14	15
ASC_{full}	0	0	0	0	0	0	0
ASC_{luxury}	1	0	0	0	0	0	0
ASC_{medium}	0	1	0	0	0	0	0
ASC_{MPV}	0	0	1	0	0	0	0
$ASC_{offroad}$	0	0	0	1	0	0	0
ASC_{petrol}	1	1	1	1	1	1	0
$ASC_{electric}$	0	0	0	0	0	0	1
β_{inc_full}	0	0	0	0	0	0	0
β_{inc_luxury}	$\frac{income}{10000}$	0	0	0	0	0	0
β_{inc_medium}	0	$\frac{income}{10000}$	0	0	0	0	0
β_{inc_MPV}	0	0	$\frac{income}{10000}$	0	0	0	0
$\beta_{inc\_offroad}$	0	0	0	$\frac{income}{10000}$	0	0	0
β_{inc_hybrid}	0	0	0	0	0	$\frac{income}{10000}$	0
$\beta_{nbr\_adults\_small}$	0	0	0	0	nbr. adults	0	0
$\beta_{nbr\_chilren\_small}$	0	0	0	0	nbr. child.	0	0
$\beta_{nbr\_cars\_lux}$	nbr. cars	0	0	0	0	0	0
$\beta_{nbr\_cars\_hybrid}$	0	0	0	0	0	nbr. cars	0
$\beta_{university}$	0	0	0	0	0	1	1
$\beta_{town\_rural\_EV}$	0	0	0	0	0	0	town_rural
$\beta_{town\_rural\_hybrid}$	0	0	0	0	0	town_rural	0
β_{price_inc}	$\frac{price_9 \cdot 100}{income}$	$\frac{price_{10} \cdot 100}{income}$	$\frac{price_{11} \cdot 100}{income}$	$\frac{price_{12} \cdot 100}{income}$	$\frac{price_{13} \cdot 100}{income}$	$\frac{price_{14} \cdot 100}{income}$	$\frac{price_{15} \cdot 100}{income}$
β_{conso_inc}	$\frac{cons_9 \cdot pp \cdot 100}{income}$	$\frac{cons_{10} \cdot pp \cdot 100}{income}$	$\frac{cons_{11} \cdot pp \cdot 100}{income}$	$\frac{cons_{12} \cdot pp \cdot 100}{income}$	$\frac{cons_{13} \cdot pp \cdot 100}{income}$	$\frac{cons_{14} \cdot pp \cdot 100}{income}$	0
β_{max_power}	$\frac{max\_power_9}{10}$	$\frac{max\_power_{10}}{10}$	$\frac{max\_power_{11}}{10}$	$\frac{max\_power_{12}}{10}$	$\frac{max\_power_{13}}{10}$	$\frac{max\_power_{14}}{10}$	$\frac{max\_power_{15}}{10}$
β_{range_EV}	0	0	0	0	0	0	$\frac{range\_EV}{100}$

Table 3: Model specification (part 2/2)

4.3 Attributes of the unchosen alternatives

Since we aggregate several combinations of make-model-type of vehicles to a single alternative, discussion about the attributes of the unchosen alternatives is required. In other words, if a respondent chose a *small diesel*, we have all the attributes associated to it, but we need to define the attributes for the remaining 14 alternatives. To do so, we consider that we have access to the empirical distribution of all the attributes of the different alternatives. This distribution consists in the observed values of other people's chosen alternatives. This can be done because our sample almost exactly replicates the real market shares of make-models and types.

In a schematic and summarized way, the procedure is the following:

1. Draw a vector of attributes from the empirical distribution for each unchosen alternative for a given respondent.
2. Repeat step 1 for each respondent.
3. Estimate the parameters of the model with this dataset
4. Iterate

Instead of estimating the model one time, we estimate it repeatedly and look at the distribution of the parameters.

4.4 Nesting structures

As explained in Section 3 the nested and cross nested logit models allow to capture correlation between alternatives. Figures 1 and 2 show two very natural nesting structures derived from the definition of alternatives shown in Table 1.

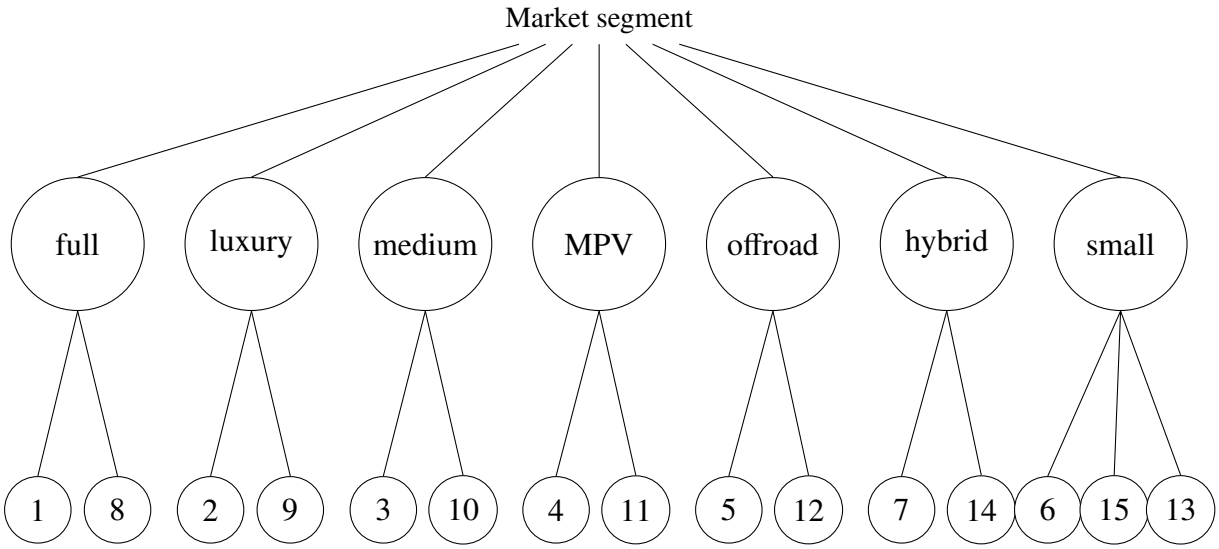


Figure 1: Nested structure by market segment

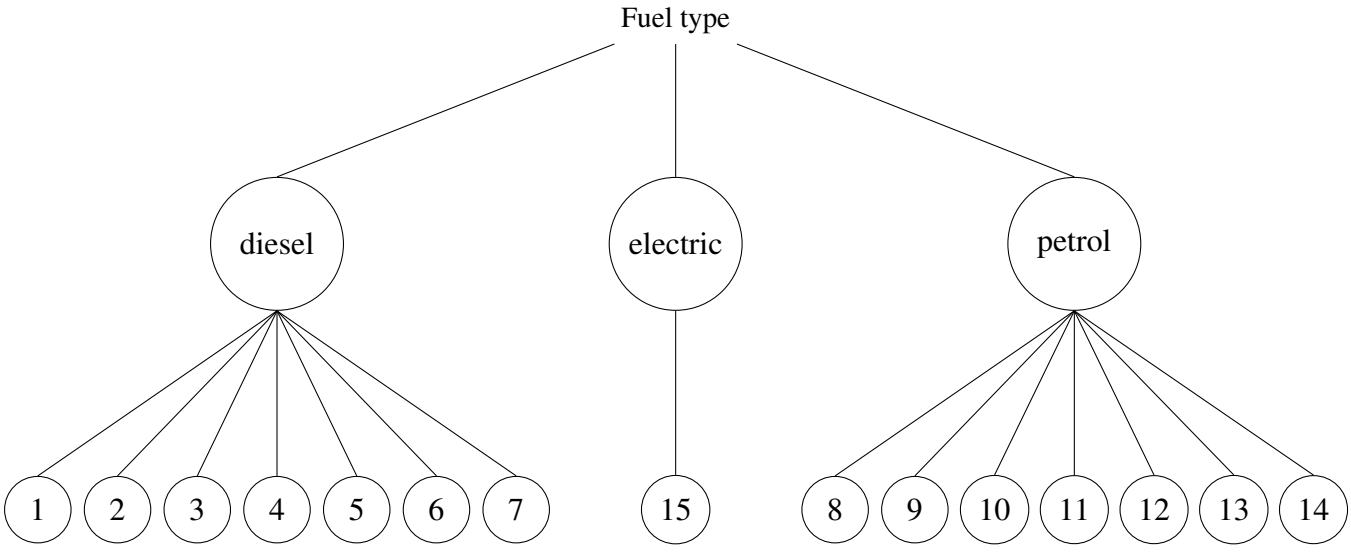


Figure 2: Nested structure by fuel type

These two nesting structures can be combined to a cross-nested structure, that is shown in Figure 3.

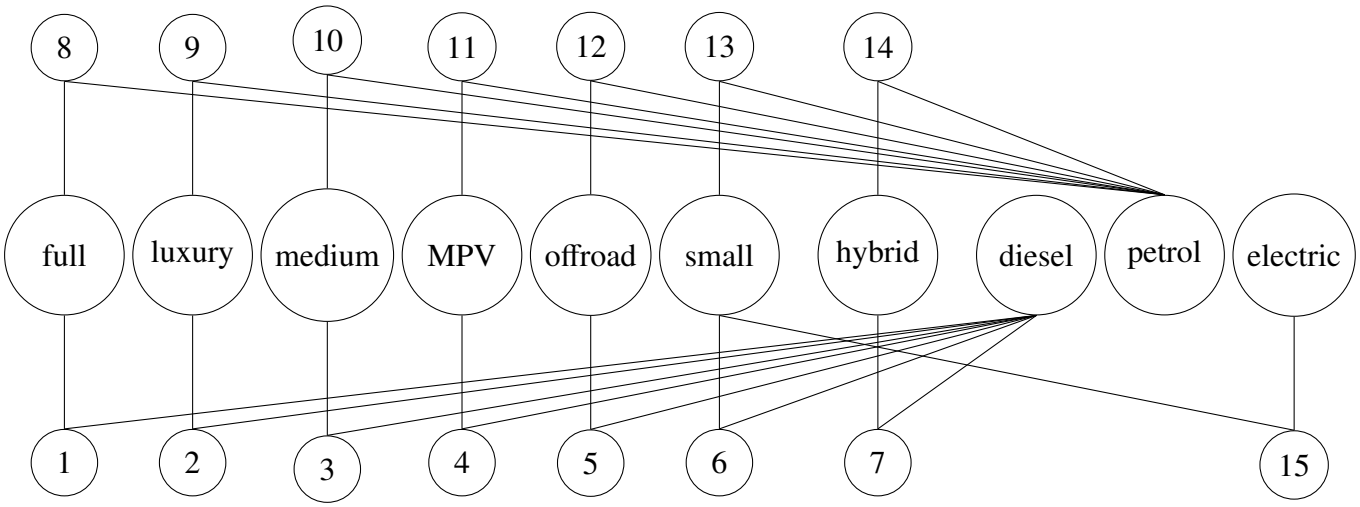


Figure 3: Cross-nested structure

4.5 Estimation results

We show the results from the logit and the cross nested logit models and ommit those of the nested logit model. As explained in Section 4.3 the same specification is estimated for both models, and the estimation is repeated for various draws of the empirical distribution of the vector of attributes of the unchosen alternatives.

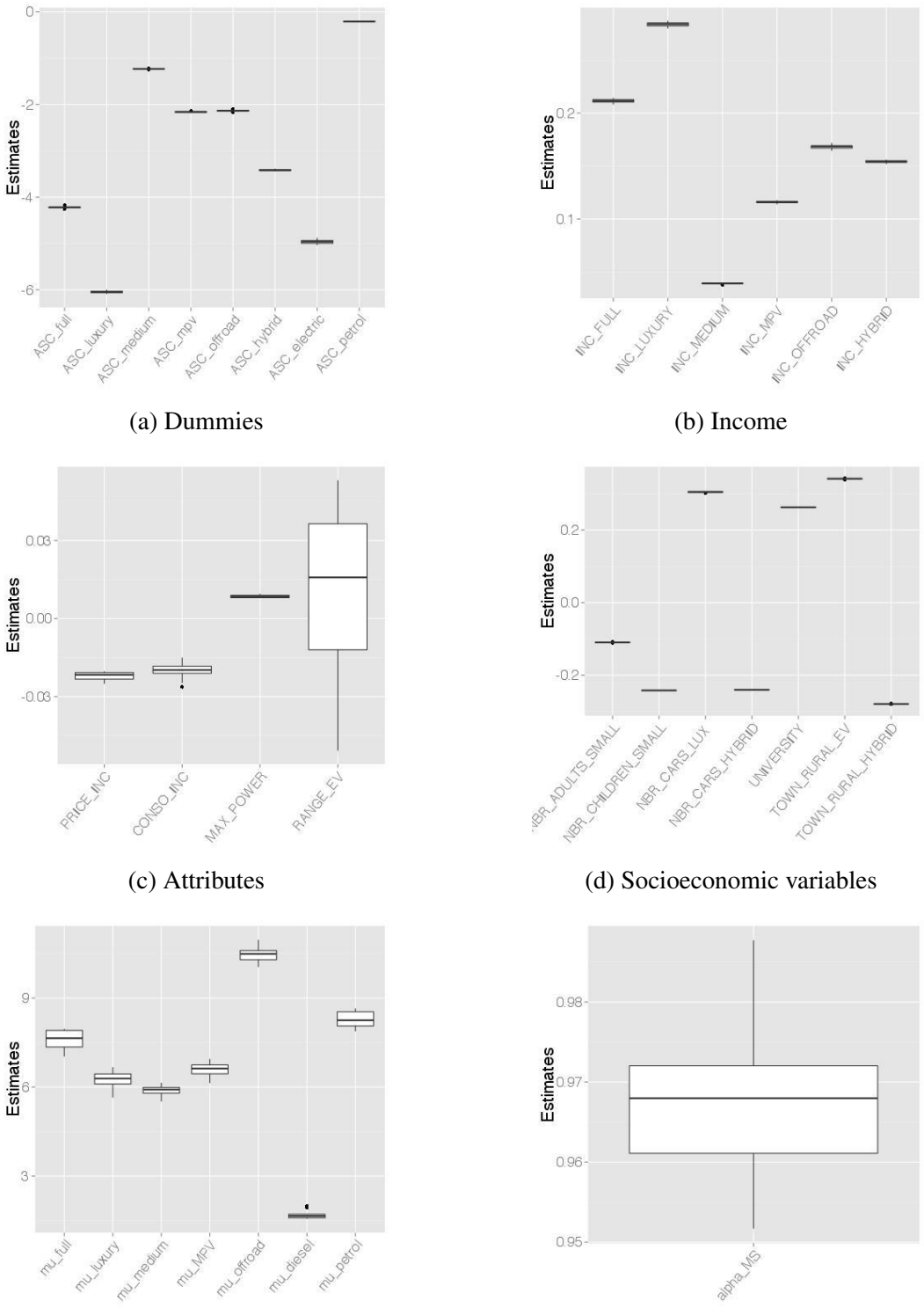
Figure 4 shows the boxplots for the values of the parameters obtained for the cross nested logit model. Due to the long estimation time, the results shown are with 10 draws. However, it is very interesting to see the low variation in the estimates obtained for each draw for all the parameters except for the range, which is discussed later. For the dummy variables (Figure 4(a)) the reference is *small* and *diesel*, and all the rest are negative, meaning that there is an intrinsic preference towards small diesel cars, all else being equal. In Figure 4(b) we can see the effect of income in the utility for different market segments. The reference is *small*. As expected, the largest coefficient corresponds to the *luxury* market segment, and the smallest to *medium*, and they are all positive. The relative magnitude among them is also in line with our expectations.

Figure 4(c) shows the estimation results for the parameters related to car attributes. The signs of all of them are in line with what is expected: the interaction between price and income (β_{price_income}) is negative, as well as the interaction between consumption and income (β_{conso_inc}). The rational behind this interaction is that fuel consumption allows to capture the running costs. The parameter associated to the maximum power of the vehicle (β_{max_power}) is positive as expected. Vehicles with higher maximum power are more attractive, all else being equal.

Interestingly, the parameter associated with the range of electric vehicles (β_{range_EV}) has a very high variability depending on the draw. It takes both positive and negative values, even though the mean is positive. The interpretation of this is that β_{range_EV} is not significantly different from zero and that it does not play a role in the attractiveness of an electric vehicle compared to other vehicle types. To confirm this hypothesis, more draws are required. However a possible explanation –since we only have revealed preference data– is that there is not enough variability in the ranges in the sample. Indeed, $\text{range}_{EV} \in (90, 163)$.

Figure 4(d) shows the effect of several socioeconomic variables. $\beta_{nbr\_adults\_small}$ and $\beta_{nbr\_children\_small}$ are negative meaning that all else being equal, the probability to choose a small car decreases as the number of people in the household increases. $\beta_{nbr\_cars\_lux}$ and$\beta_{nbr_cars_hybrid}$ can be interpreted in the same way, but they have opposite signs. This means that all else being equal, the probability to choose a *luxury* car increases with the number of cars in the household, but the probability to choose a *hybrid* car decreases. $\beta_{university}$ is positive and included in the alternatives related to electric or hybrid vehicles. The interpretation is that all else being equal, highly educated people are more likely to choose more environmental friendly vehicles. This is in line with the literature. Finally, the place of residence is also included. People leaving in towns or rural areas are more likely to choose an electric vehicle than people that live in agglomerations or cities. For hybrid vehicles the result is the opposite. Respondents living in towns or rural areas are less likely to buy a hybrid vehicle compared to those that live in the city or in agglomerations. This needs further research, because it is not in line with the literature. A possibility to explore is to differentiate between towns and urban areas.

Finally, Figures 4(e) and 4(f) show the results associated to the cross nested logit model. The μ parameters associated to *small* and *hybrid* are set to one, because they systematically reach the lower bound during the estimation. The interpretation is that the error terms of those alternatives are not correlated and that the nest does not make sense. The μ parameter associated to *electric* is set to one for normalization reasons (there is only one alternative in the nest). All the other μ parameters are significantly different from one. α_{MS} represents the degree of membership to the market segment (including hybrid as a market segment). The degree of membership associated with the fuel type is therefore $1 - \alpha_{MS}$. Note that by only defining one parameter α we are assuming that the degree of membership to the respective market segment does not depend on the market segment. In other words, a *small diesel* car belongs $(\alpha_{MS} \cdot 100)\%$ to the nest *small* and $(100 - \alpha_{MS} \cdot 100)\%$ to the nest *diesel*, but the same holds for a *medium petrol*, that belongs $(\alpha_{MS} \cdot 100)\%$ to the nest *medium* and $(100 - \alpha_{MS} \cdot 100)\%$ to the nest *petrol*. This needs further investigation, but preliminary results show estimation issues when more α parameters are introduced.



(a) Dummies

(b) Income

(c) Attributes

(d) Socioeconomic variables

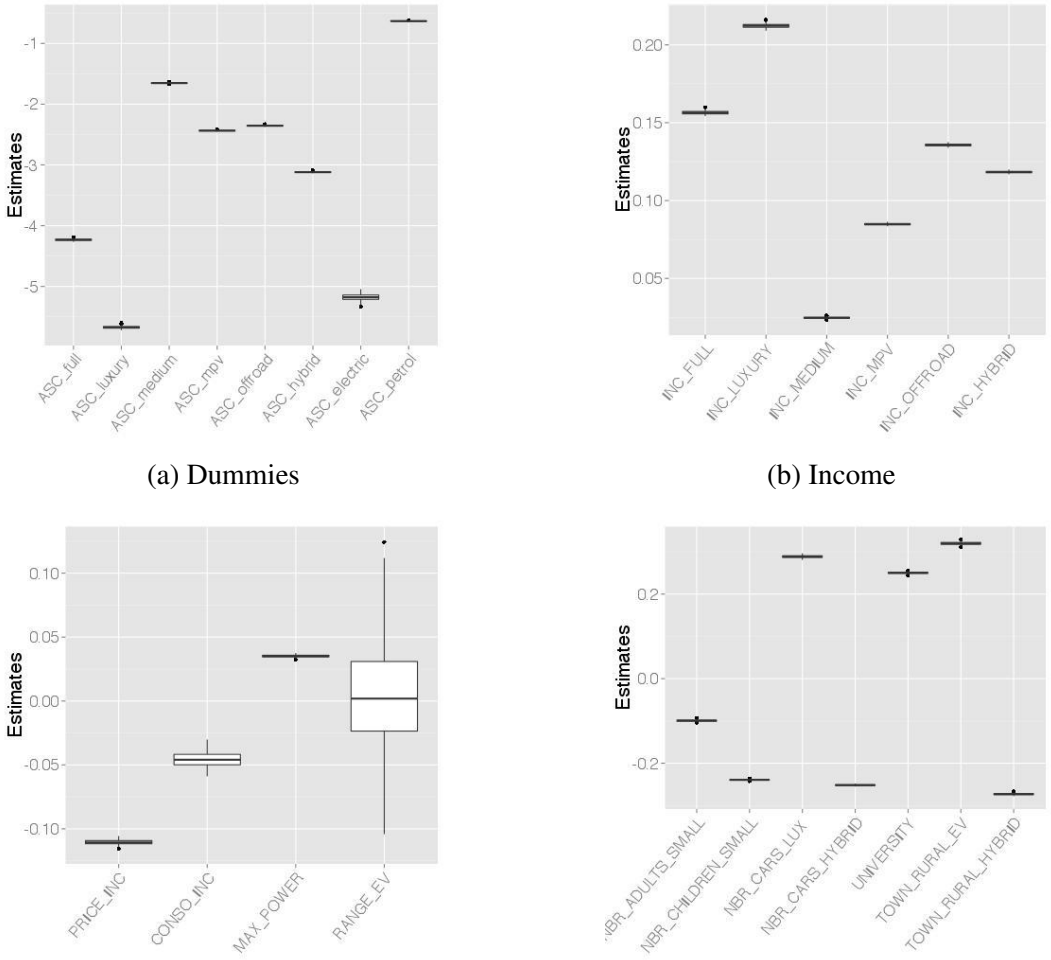
(e) μ parameters

(f) α parameter

Figure 4: Estimation results for a cross-nested logit model with 10 draws

Figure 5 shows the boxplot of the parameter estimates for the logit model. The estimation time is a lot faster, so results are shown with 250 draws. The variability of the parameters is similar to the one observed for the cross nested logit model with only 10 runs and they can all

be interpreted in a similar way.



(a) Dummies

(b) Income

(c) Socioeconomic variables

(d) Attributes

Figure 5: Estimation results for a logit model with 250 draws

Substitution patterns To investigate how the substitution patterns behave, we consider an increase in the price of the *off road diesel* alternative (alternative 5) by 50%. It is not a realistic scenario, but the objective is to see how the market shares change according to the different models used. Table 4 shows the observed market shares from the sample in the first row, followed by the forecasts obtained using the logit model and the cross-nested logit model after the increase in price.

Alternative 5 shares nest with all the alternatives that belong to the nest *diesel*, which are alternatives 1 to 7 and with the only other one that belongs to the nest *off road*, alternative 12. We expect that for the CNL these alternatives will attract a larger portion of the market shares lost by alternative 5 compared to the logit.

Alternative 5's market share decreases from 10.87% to 5.47% and 7.73% for the logit and the CNL respectively. This is expected since the parameter associated to the price is more negative for the logit model than for the CNL. There are some results that might seem counterintuitive. For example, for alternative 6 (*small diesel*) the market share predicted by the logit model is larger than the market share predicted by the CNL even if it belongs to the nest *diesel*. However this is due to the fact that the market share of alternative 5 decreased more for the logit model than for the CNL. If we calculate which proportion of the market share lost by alternative 5 goes to alternative 12 for each model we see that it is $\frac{3.88-1.41}{10.87-7.73} \cdot 100 = 79\%$ for the CNL and only $\frac{4.68-1.41}{10.87-5.47} \cdot 100 = 61\%$ for the logit.

Alternative	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Observed	1.72	0.95	11.84	7.31	10.87	29.45	0.86	0.36	0.29	3.53	1.65	1.41	26.79	2.64	0.35
Logit	1.30	0.81	9.96	6.06	5.47	36.14	2.34	0.83	0.45	5.79	3.13	4.68	21.42	1.24	0.36
CNL	1.48	0.88	10.74	6.48	7.73	30.49	1.90	0.64	0.39	5.69	2.82	3.88	25.01	1.52	0.35

Table 4: Market shares before and after a 50% increase in price of alternative 5 (diesel offroad)

5 Conclusion

This paper is a first step towards modeling complex substitution patterns in the private vehicle market. By using a very rich dataset that contains sales of new cars, we show that our modeling approach – which consists of the definition of the choice set and the way to impute the attributes of the unobserved alternatives – yields mostly intuitive results and that are in line with the literature in terms of the obtained estimates. To the best of our knowledge, this is the first time that a cross-nested logit model is used to model car-type choice. A result that might seem surprising, and for which no evidence has been found in the literature, is that the range of the electric vehicles does not play a role in the car-type choice. As discussed previously, this might be due to a data limitation and not to a behavioral characteristic. In terms of the substitution patterns observed after an increase of price of one of the alternatives, the results are also in line with our expectations, but further analysis is required.

As future work, the first step is to analyze different scenarios to see how the market shares change for the different alternatives, to be able to better understand the difference in substitution patterns between the logit and the cross-nested logit. Moreover, additional models for the real price and fuel consumption will be integrated in order to recover missing variables and to address endogeneity. The final step is to use other choice-probability-generating-function-based models and to compare them with the results obtained both with the logit and the cross nested logit.

As mentioned in Section 2 new results show that it might be possible to estimate the model without aggregating alternatives. It would also be very interesting to try this approach and to compare it with the results obtained by aggregating the alternatives as we do in this paper.

6 References

Anowar, S., N. Eluru and L. F. Miranda-Moreno (2014) Alternative Modeling Approaches Used for Examining Automobile Ownership: A Comprehensive Review, *Transport Reviews*, **34** (4) 441–473, July 2014, ISSN 0144-1647.

Berkovec, J. and J. Rust (1985) A nested logit model of automobile holdings for one vehicle households, *Transportation Research Part B: Methodological*, **19** (4) 275–285, August 1985, ISSN 0191-2615.

Bierlaire, M., M. Ben-Akiva, D. McFadden and J. Walker (2015) *Discrete Choice Analysis*.

Birkeland, M. E. and J. Jordal-Jorgensen (2001) Energy efficiency of passenger cars, Cambridge, UK, ISBN 978-0-86050-339-2.

Brownstone, D. and K. Train (1998) Forecasting new product penetration with flexible substitution patterns, *Journal of Econometrics*, **89** (1–2) 109–129, November 1998, ISSN 0304-4076.

Cao, X., P. L. Mokhtarian and S. L. Handy (2006) Neighborhood design and vehicle type choice: Evidence from Northern California, *Transportation Research Part D: Transport and Environment*, **11** (2) 133–145, March 2006, ISSN 1361-9209.

Choo, S. and P. L. Mokhtarian (2004) What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice, *Transportation Research Part A: Policy and Practice*, **38** (3) 201–222, March 2004, ISSN 0965-8564.

de Jong, G., J. Fox, A. Daly, M. Pieters and R. Smit (2004) Comparison of car ownership models, *Transport Reviews*, **24** (4) 379–408, July 2004, ISSN 0144-1647.

Mai, T., E. Frejinger, M. Fosgerau and F. Bastin (2015) A Dynamic Programming Approach for Quickly Estimating Large MEV Models, *Technical Report*, June 2015.

McCarthy, P. S. and R. S. Tay (1998) New Vehicle Consumption and Fuel Efficiency: A Nested Logit Approach1, *Transportation Research Part E: Logistics and Transportation Review*, **34** (1) 39–51, March 1998, ISSN 1366-5545.

McFadden, D., K. Train *et al.* (2000) Mixed mnl models for discrete response, *Journal of applied Econometrics*, **15** (5) 447–470.

Mohammadian, A. and E. Miller (2002) Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance, *Transportation Research Record: Journal of the Transportation Research Board*, **1807**, 92–100, January 2002, ISSN 0361-1981.

Mohammadian, A. and E. Miller (2003) Empirical Investigation of Household Vehicle Type Choice Decisions, *Transportation Research Record: Journal of the Transportation Research Board*, **1854**, 99–106, January 2003, ISSN 0361-1981.

Page, M., G. Whelan and A. Daly (2000) Modelling the factors which influence new car purchasing.

Potoglou, D. (2008) Vehicle-type choice and neighbourhood characteristics: An empirical study of Hamilton, Canada, *Transportation Research Part D: Transport and Environment*, **13** (3) 177–186, May 2008, ISSN 1361-9209.

Wu, G., T. Yamamoto and R. Kitamura (1999) Vehicle Ownership Model That Incorporates the Causal Structure Underlying Attitudes Toward Vehicle Ownership, *Transportation Research Record: Journal of the Transportation Research Board*, **1676**, 61–67, January 1999, ISSN 0361-1981.