A discrete continuous model of vehicle ownership and use in Flanders

Franckx, Laurent and Michiels, Hans and Mayeres, Inge

Flemish Institute for Technological Research (VITO)

25 August 2014

A DISCRETE CONTINUOUS MODEL OF VEHICLE OWNERSHIP AND USE IN FLANDERS

Laurent Franckx[1], Hans Michiels and Inge Mayeres, Flemish Institute for Technological Research (VITO)

ABSTRACT[2]

We estimate a discrete-continuous model of vehicle demand and use for the Belgian region of Flanders, combining the results of the official regional travel survey with a detailed database of vehicle characteristics. The overall predictive value of the submodel predicting the number of vehicles owned by each household is satisfying, and in line with expectations. However, existing data turn out to be relatively poor predictors of the vehicle class owned by households and of the annual mileage per vehicle. We argue that the current travel survey focuses on determinants of travel in the peak periods. In order to predict overall travel demand, future versions of the travel survey should identify indicators with a higher predictive value for travel behavior for other than commuting purposes.

1. INTRODUCTION

Understanding the determinants of the type of cars people own, how many cars they own and how much they drive remains a highly topical research area. From a policy point of view, the environmental impacts (such as local air quality, traffic noise and greenhouse gas emissions) of vehicle choice and use speak for themselves. However, both the acquisition cost and the variable costs of a car are also affected by variables that are set by governments: acquisition taxes, annual circulation taxes, fuel taxes, etc. Reform of these variables will therefore not just affect the environmental performance of the transport system, but also tax revenues and possibly income distribution.

In this paper, we have jointly estimated these three decisions using a joint discrete/continuous model of vehicle choice and use – we refer to De Jong et al. (2004) for a comprehensive survey of other approaches to car ownership, choice and use modelling.

The literature on discrete/continuous modelling of consumer demand using a unified theoretical framework has originated with Haneman's (1984) seminal work. Train (1986) provides a comprehensive application to vehicle ownership and use. De Jong (1991, 1997) has used similar models for the joint modelling of the number of vehicles owned and distance traveled, but without considering the choice of car type.

In this paper, we will follow closely the approach pioneered by Train.

This is not to say that no relevant further developments have taken place since Train's work. One important alternative to the approach used here are Multiple Discrete-Continuous Choice Models - Bhat and Pinjari (forthcoming) provide a comprehensive discussion of the relevant literature. These models allow for the fact that consumers may choose multiple alternatives that are imperfect substitutes for each other. In the case of car demand models, this approach reflects that

[1] Corresponding author: Boeretang 200, 2400 Mol, laurent.franckx@vito.be

[2] This research has benefited from a research grant from the Flemish Policy Research Centre on Fiscal Policy. We would like to thank the Scientific Advisory Committee and Stef Proost for stimulating and insightful comments on previous drafts. All remaining errors are ours.

households may own a mix of vehicle types to satisfy different functional or variety-seeking needs (Bhat and Sen 2006). As discussed by Bhat and Sen, a drawback of this approach, however, is that it is based on the assumption that the total household annual mileage across all personal motorized vehicle types is known *a priori*. Therefore, such models cannot capture the effect of variables that are likely to affect total vehicle use, such as the fuel cost. Moreover, as pointed out by Fang (2008), such model do not allow for the possibility that households may own two vehicles belonging to the same class. We stick to the approach used by Train, and take the annual mileage of both vehicles owned by the households as endogenous.

For reasons to be clarified below, we will also follow Train's approach (1986) in modelling the choice of vehicle class, and not of individual models. It is common to use average values of vehicle characteristics when modelling the choice of a class of vehicles. However, it is not always recognized that consistent estimation of such models requires that one also incorporates information on the variance of vehicle characteristics within each vehicle class. In our model, we have stuck rigorously to the approach that McFadden (1978) has developed for approximating the Inclusive Value of underlying vehicle models in each class (see further for more details).

The focus on car ownership[3] (rather than on car purchase) also implies that we are actually modelling two choices in a single model: (a) the initial choice of a specific car and (b) the recurrent annual decision to keep the current car. Ideally, we would use a fully dynamic model[4]. However, as our data are limited to a cross-section sample, it was not possible to capture dynamics explicitly.

A possible alternative would be to capture dynamics (partially) with the use of transaction cost dummies, representing the search costs of acquiring a new vehicle. For instance, Berkovec and Rust (1985) use a transaction costs dummy that is 1 if the currently held vehicle was owned in the previous year and 0 otherwise. Unfortunately, adding this dummy to our model prevented it from converging. Train (1986) uses a transaction cost dummy which is equal to one for vehicles in class that a household did not own in the previous year, and zero for all others. Data limitations have prevented us from implementing this in our model. Therefore, we had to maintain a strictly static approach.

We shall proceed as follows. We first provide an overview of the structure of the model, which consists of four sub-models: an (implicit) model of vehicle model choice, a discrete choice model of vehicle class choice, a discrete model of vehicle quantity choice and a continuous model of annual mileage per vehicle (conditional on the number of vehicles owned). We then proceed with a description of the data we have used, a discussion of the most important preliminary data manipulations, and the key descriptive statistics. The main part of the paper consists in a discussion of the main results of each sub-model. We finish with a summary of the results, and a general discussion of the general predictive performance of the model.

2. STRUCTURE OF THE MODEL

For households owning two or less cars, we have addressed the following questions:

- How do the socio-economic characteristics of households and the characteristics of cars affect the choice of the car(s) owned by the household? (the "vehicle choice" model)

[3] Note that, in parallel with research, a model of vehicle purchases has been developed (see Mayeres and Vanhulsel 2014).

[4] A classical reference in this field is Hensher et al. (1992).

- How do the socio-economic characteristics of households and the characteristics of cars affect the numbers of car(s) owned by the household? (the "vehicle quantity" model)
- How do the socio-economic characteristics of households and the characteristics of cars affect the annual distance driven per car owned by the household? (the "distance" model)

A separate model has been estimated for each of these choices. Although these models are discussed sequentially, they are not logically independent. The interdependence of the different models is represented in Figure 1.

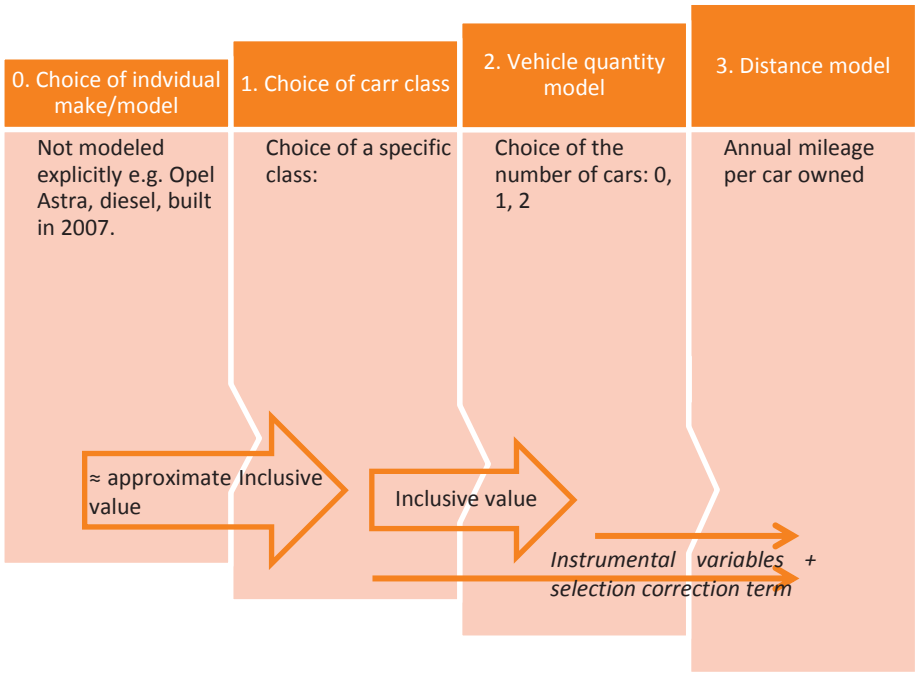

Figure 1 Structure of the model

Submodel 0 corresponds to the choice of a specific make/model e.g. an Opel Astra diesel built in 2007. We follow Train (1993, p. 142) in not modelling this choice explicitly.

Indeed, if we model car choice at the level of individual makes and models, then we can only perform forecasts related to the models that have been observed in the current sample. However, the number of different car models that are observed in the sample used for estimation is likely to be small compared to the total number of models that are available on the market[5]. Therefore, the predictive value of such a model would be limited.

Therefore, the first explicit choice modelling takes place at the level of **submodel 1**, the vehicle choice model. In this submodel, we model the probability that a household chooses a specific car class (if the household chooses one car) or a combination of two classes (if the household chooses two cars). Each car class corresponds to a body type, the fuel used (diesel or gasoline) and the

[5] The Febiac database with technical data we have used (see further) distinguishes between 33403 different car models. The travel survey with the household choices contains just 466 different models.

vintage[6]– see Table 1. *All* models available in the FEBIAC database are then allocated to a given class. We consider the choice between thirty possible classes[7].

Table 1: Definition of the body types used in the vehicle choice model

Body type
Urban car+ Compact car
Family car + executive
convertible + coupé
SUV
Mini-van

It is easy to see the analogy between this approach and the structure of a Nested Logit (NL) model where each "car class" would correspond to a nest. Interestingly, McFadden (1978) has indeed shown that submodel1 and submodel2 can be linked using an approximation to a nested logit model. The basic insight of McFadden (1978) is that the utility of a given "car class" does not just depend on the average characteristics of the models contained in this class, but also on their variances and the covariances – we refer to the annex for a more formal treatment.

Thus, all the terms in the covariance matrix of the explanatory variables are added to the model, and the probability that a household chooses a given car class is then modelled with a standard discrete choice model. We have tested both Multinomial Logit (MNL) and Nested Logit (NL) model[8].

The results of submodel 1 are then used to estimate the choice of the "vehicle quantity" with an MNL (**submodel 2**). Indeed, the probability that a household chooses a given number of cars is affected, not just by the characteristics of the households, but also by the characteristics of the existing cars, which are represented by the Inclusive Value of owning one or two cars: this represents the expected value of the maximum utility of owning one or two cars.

Finally, in **submodel** 3, we address the distances travelled with each car owned by a household. As distance travelled is a continuous variable, we have used linear regression in this submodel.

However, we have good reasons to expect that Ordinary Least Squares (OLS) will lead to inconsistent estimates.

First, several variables that affect distance travelled are also likely to affect the choice of the car class. For instance, a household that anticipates a high annual mileage will attach a different value to some vehicle characteristics (such as fuel consumption) than if anticipates a low annual mileage. Fuel consumption is thus an endogenous variable, and we need thus to test whether an Instrumental Variable estimation is required.

[6] We consider three vintage classes: < 2001, 2001-2005, > 2005.

[7] We have also estimated the probability that a one-car household chooses a given class out of 36 possible classes. However, for a two-car household, 36 possible classes imply $\left(\binom{36}{2}\right) = \frac{37*36}{2} = 666$ possible choices, while our econometric software imposes a maximum of 500 possible choices. The results for the one-car household model with 36 classes are close to the results for the one-car household model with 30 classes, and are available on request.

[8] See Hensher et al. (2005) or Train (2009) for extensive discussions of MNL and NL models.

Second, it also seems reasonable that a given parameter will affect annual mileage for each car differently depending on the number of cars owned by the family. Therefore, we estimate a separate distance model for one car households than from two car households. However, some variables that affect annual mileage also affect the number of cars chosen. Indeed, a household will evaluate a certain number of parameters (such as the distance to work, the number and the activities of the household members, the availability of public transport, the distance to close family and friends) to determine whether or not it needs a second car. Hence, anticipated annual mileage driven and the number of cars per family are not independent of each other. As a result, the variables which influence the probability of choosing a given number of cars will now be correlated with the error term of the distance model. Thus, we need to correct our estimates for sample selection bias.

3. DATA DESCRIPTION

In this section, we describe the data we have used and the most important data manipulations that we have undertaken before estimating the models.

We have used two main sources of information.

First, we have used the Flemish transport survey (Onderzoek Verplaatsingsgedrag) (henceforth "OVG"). 5046 households participated in the version we have used (version 4.1 to 4.3). The survey contains information pertaining to the main household characteristics (income, family size, employment status, age of the head of the household etc), the vehicles used by the household (make, model, fuel, construction, year) and a sample of trips performed by household members.

For some variables, there were significant gaps in the answers. Some of these missing values were probably due to the reluctance of the respondent to report information that was considered sensitive (such as the household income), even if these data remain confidential. In other cases[9], the question may have been due to the difficulty of the question, or the respondent may have filled in a blank instead of a "zero" value. This has led to difficult trade-offs between including additional explanatory variables and keeping a sufficiently large sample size. Moreover, several continuous variables responses were collected in categorical form, which results in the loss of some information. For instance, household income was reported as belonging to one out of six income classes.

Also, for several models, there exist variants to the base model, but households can only report the model, not the variant. We have therefore assumed that, in these cases, the household chooses the cheapest variant. We do not know how this choice has affected the overall performance of the model.

Second, we have purchased the proprietary database of FEBIAC, the Belgian car and motorcycle federation. This database contains the technical data and prices of all cars that have been sold on the Belgian first-hand market since 1990. The detail of the data pertaining to the cars goes much deeper than in the OVG survey.

[9] For instance, the extent to which the car is used for professional purposes, or participation in a car-pool system.

However, there are also significant gaps. For instance, the only indicator of variable costs per kilometer driven is the fuel cost per kilometer – we have no data on depreciation or maintenance cost per kilometer. Also, while the volume of a car is probably an important criterion in the purchase decision, it is not included in the database, and we had to use an approximate value (length*width*height), which seriously overestimates the volume of sedans compared to station wagons. However, the data do not distinguish between sedans and station wagons, and it is not possible to correct for this bias.

Finally, the purchase price in the Febiac database is the price for new cars, and there are no reliable estimates of the values on the second had market. Therefore, in order to estimate the opportunity cost of the car owned by the household, we had to rely on a depreciation schedule based on expert judgment. The vehicle purchase tax[10] and (when applicable) the subsidy for cars with low CO_2 emissions were added to the price to obtain the total acquisition cost.

In order to estimate the model, we have performed some additional operations on both datasets.

We first grouped the vehicles in the Febiac database according to their class. Each vehicle model was associated with the average values and the covariance matrix of the characteristics of the cars belonging to the same class. The average values and the covariance matrix were then joined with the OVG database according to the class of the vehicle(s) owned by the household. For the purposes of the two car household model, we have performed similar operations for each possible pair of models.

We have not considered households owning three or more cars. Indeed, as will become clear below, the methods we use here become intractable once we have to model the ownership of more than two cars.

Moreover, we have not considered families with one or more company cars. Indeed, people who can use their company car for private purposes are not confronted with the full costs of acquiring and using a car. One can therefore expect that the type of cars they use, and how much they drive, differs significantly from the households who own their cars. Therefore, we have estimated a separate model for the households owning their cars and kept the modelling of company car choice and utilization as a topic for further research.

4. KEY DESCRIPTIVE STATISTICS

Before we proceed with the results of the econometric analysis, we present some key descriptive statistics in Table 2, split according to the number of cars owned by the family.

First, all quartiles and the mean of the annual distance per car are lower for families who own one car. This suggests that the mobility needs in one-car families are different from those in two-car families, and that separate distance models will be needed for both family types[11]. A comparison between the quartiles and the mean value also indicates that the distribution of the distances per car is heavily skewed positively.

[10] Which, in Belgium, also applies to the purchase of second hand cars.

[11] Keep in mind that we do not consider families who use one or more company cars.

Table 2: key descriptive statistics

Number of cars per family	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Annual distances per car						
1	5	6000	10000	12540	17000	120000
2	10	7000	11000	14050	20000	150000
Age of the head of the family						
1	22	43	54	55.65	68	93
2	25	43	49	50.16	57	86
Distance to work						
1	0	0	0	6.708	7.5	160
2	0	0	6.5	12.59	16	160

Second, both the median and the mean of the age of the head of the family are higher in families with two cars than in families with one car. However, the third quartile is much higher in one-car families. This suggests that the age profiles are different for both types of households, probably because the age of the household head is indicative of the presence of dependent children in the household.

Third, the median, the mean and the third quartile of the distance to work is higher in two car families than in one car families. Both distributions are heavily skewed to the right: 75% of the one car families face a maximum commuting distance of less than 7.5 km, and for 50% of these families, the distance is even negligibly small. For two car families, these quartiles are higher, but not spectacularly so (and much lower than the maximal distances).

Other key descriptive variables (distance per income class, distance as a function of the number of family members, distance as a function of the type of municipality[12], distance as a function of the intensity of use of alternative modes) are available on request.

5. THE VEHICLE CHOICE MODEL FOR ONE-CAR FAMILIES

For one car families, our preferred model is a 2 level Nested Logit where the nests are defined according to the vintage of the car classes: class 1-12, 13-24 and 25-36 correspond to nests 1, 2 and 3 respectively[13].

In Table 3, we represent the estimation results for the significant variables in the preferred model. The definition of the variables is given in Table 4 (leaving out the estimates for the Alternative Specific Constants). This model has been estimated with 1740 observations.

We can summarize the main findings as follows.

[12] The OVG distinguishes 8 types of municipalities, ranging from “large” urban areas to “rural” areas. This variable has been included in both the vehicle quantity and the distance model. This is in line with the hypothesis that people living in dense urban areas travel shorter distances, and are more likely to use public transport – for a recent review, see Boussauw 2011

[13] We have also estimated a MNL model, and alternative nesting structures based on body types and/or fuels. Full results are available on request.

PR11YLOW and **PR11YHIG** are interaction variables between the acquisition cost of a car and effects coded dummy variables corresponding to the "low" and "high" income classes. For "medium" income families (the reference level for the "income class" dummy), the implied coefficient is -0.03879E-04. Thus, for low and medium income families, an increase in the average acquisition cost of a car class reduces the utility that a family obtains from choosing a car in this class, which is the expected effect. The opposite is true for high income families, suggesting that, for higher income classes, there is a "snob effect": all other things being equal, they prefer more expensive car classes.

VOLMEDGZ and **OLGRGZ** are interaction variables between the car's volume and the family size. As expected, for "large" families (5 or more members) an increase in the average volume of a car class increases the utility that a family chooses a car from this class. For both "small" (the reference level) and "medium" sized families the opposite holds true: a decrease in the average volume of a car class increases the utility that a family chooses a car from this class. One possible explanation for this counterintuitive result is that the "volume" variable is an overestimation of the true volume of the car, especially for sedan cars. If small families are more likely to own sedan cars than station wagons[14], then our data systematically overestimate the volumes of cars owned by small households.

VOLYLOW shows that, for low income families, an increase in the average volume of a car class lowers the utility that a family obtains from choosing a car from this class. For medium income families (the reference level) the opposite holds (implied coefficient of +0.08156). For high income families, there is no significant effect.

TRAFFTAX shows that the coefficient for the annual traffic tax has the expected sign but is not highly significant. The low significance of this specific effect can be explained as follows. In most cases, the variance of the traffic tax within each class is higher than the variance of traffic tax between the different classes. The opposite holds true for the acquisition cost[15], which is generally much higher than the expected present value of all annual traffic taxes paid over a vehicle's life cycle. It is therefore not surprising that the impact of the traffic tax is relatively unimportant compared to the impact of the acquisition cost.

FUELCOST is significant with the expected sign.

LOGRC (= $\log(N_i/N)$), links each vehicle class with the implicit underlying nest of models contained in the class. As discussed above, McFadden (1978) has shown that this variable needs to be added to the explanatory variables as one of the proxy variables to the real Inclusive Value of the underlying nest. The estimated confidence interval for the parameter implies that we cannot reject the hypothesis that θ lies in the [0,1] interval, which is required for consistency with utility maximization (Hensher et al., 2005).

The ninth to eleventh variables are elements from the covariance matrix between the characteristics of the individual models contained in each vehicle class. As discussed above, McFadden (1978) has shown that this covariance matrix needs to be added to the explanatory variables as one of the proxy variables to the real Inclusive Value of the underlying nest. Only three covariance terms turned out to be significant.

[14] Keep in mind that the available data do not allow to distinguish between sedans and station wagons.
[15] Detailed results are available on request.

OMTTXGG is the covariance between, on the one hand, the interaction term between volume and family size (for large families), and, on the other hand, the annual traffic tax. The negative sign of the coefficient implies that if (within a given vehicle class), there is a high covariance between volume and traffic tax, this reduces the expected utility of this vehicle class compared to a class with a high low covariance between volume and traffic tax (at least in the case of large families).

Remarkably, **OMTTXGHY** and **OMTTXGHO** are the covariance between the annual traffic tax and a variable which is not individually significant in the NL model: an interaction term between the power of cars in the vehicle class and the age of the head of the family. This result shows that, if the head of the family is "young", a high covariance between the traffic tax and the power of a car leads to a lower expected utility. The opposite effect holds true for "old" family heads (older than 65). In the interpretation of this result, one has to take into account that the cubic capacity of the engine is a key determinant of the annual traffic tax, and that there is a high correlation between this cubic capacity and the engine power[16].

It is noteworthy that the three significant terms from the covariance matrix all included the annual traffic tax, showing that the effect of the traffic tax goes beyond the effect of the "average" tax within each vehicle class.

IVOLD, **IVMED** and **IVYNG** are the coefficients of the Inclusive Values of the "old", "medium" and "young" nests, respectively. All coefficients are highly significant. Moreover, their 95% confidence intervals lie between 0 and 1, and are thus compatible with utility maximization.

Finally, in the NL model, none of the performance indicators (such as the average maximum speed, the average cubic capacity and the average engine power in each class) turned out to be significant.

Table 3 Nested logit model for 1 car household

Explanatory variable		Coefficient	95% confidence interval	
1.	**PR11YLOW**	-0.23423E-04***	-0.34493E-04	-0.12353E-04
2.	**PR11YHIG**	0.25876E-04***	0.10062E-04	0.41690E-04
3.	**VOLMEDGZ**	-0.02404*	-0.04824	0.00015
4.	**VOLGRGZ**	0.11426***	0.04421	0.18431
5.	**VOLYLOW**	-0.03086**	-0.05845	-0.00327
6.	**TRAFFTAX**	-0.00118*	-0.00239	0.00003
7.	**FUELCOST**	-0.17428**	-0.31230	-0.03625
8.	**LOGRC**	0.69615**	0.15656	1.23575
9.	**OMTTXGG**	-0.00045**	-0.00089	-0.00002
10.	**OMTTXGHY**	-0.17303E-04**	-0.31733E-04	-0.28721E-05
11.	**OMTTXGHO**	0.99640E-05**	0.10889E-05	0.18839E-04
12.	**IVOLD**	0.45917***	0.15296	0.76538
13.	**IVMED**	0.46530***	0.16832	0.76227
14.	**IVYNG**	0.55531***	0.23055	0.88006

Table 4: variables used in the NL model for one car household

Explanatory variable	Definition
1. **PR11YLOW**	Interaction variable between the total acquisition cost (depreciated purchase price, circulation tax and CO2 subsidy) and an effects coded dummy for low income families (net family income ≤2000 EUR per month)

[16] R^2= 0.86 for the sample of 1740 households that was used in the analysis.

2.	**PR11YHIG**	Interaction variable between the total acquisition cost (depreciated purchase price, circulation tax and CO2 subsidy) and an effects coded dummy for high income families (net family income > 4000 EUR per month)
3.	**VOLMEDGZ**	Interaction variable between approximate volume (length x width x height) and an effects coded dummy for medium sized families (3 or 4 family members)
4.	**VOLGRGZ**	Interaction variable between approximate volume (length x width x height) and an effects coded dummy for large sized families (> 4 family members)
5.	**VOLYLOW**	Interaction variable between approximate volume (length x width x height) and an effects coded dummy for low income families (net family income ≤2000 EUR per month)
6.	**TRAFFTAX**	Annual traffic tax in 2011
7.	**FUELCOST**	Fuel cost (EUR/100km)
8.	**LOGRC**	The log of the ratio between the number of models in this class and the total number of available models
9.	**OMTTXGG**	Covariance between VOLGRGZ and TRAFFTAX
10.	**KWGHFDYG**	Interaction variable between the power of the car and effects coded dummy for "young" head of the family
11.	**KWGHFDOD**	Interaction variable between the power of the car and effects coded dummy for "old" head of the family
12.	**OMTTXGHY**	Covariance between KWGHFDYG and TRAFFTAX
13.	**OMTTXGHO**	Covariance between KWGHFDOD and TRAFFTAX
14.	**IVOLD**	**Inclusive value** of the nest '**OLD**', i.e. cars with vintage <2001 (car class 1-12)
15.	**IVMED**	**Inclusive value** of the nest '**MED**', i.e. cars with vintage 2001-2005 (car class 13-24)
16.	**IVYNG**	**Inclusive value** of the nest '**YNG**', i.e. cars with vintage >2005 (car class 25-36)

Despite the high number of highly significant covariates, the overall predictive power of the model turned out to be limited (pseudo-R^2[17] of 0.0192).

6. THE VEHICLE CHOICE MODEL FOR TWO-CAR FAMILIES

In the case of two car families, the dependent variable is the probability that a household chooses a *pair* of cars from the vehicle classes. It is not clear how such pairs can be grouped meaningfully in nests, and we have therefore limited ourselves to a MNL.

For most explanatory variables, we expect that it is the *sum* of the average characteristics of the classes to which the chosen cars belong that matter in the choice of the vehicle pair. Thus, we construct the sum of the average acquisition costs, the sum of the annual traffic taxes etc.

In some specific cases (such as the car's volume), we reckon that the expected absolute value of the difference between the two classes also matters[18]: for a given total volume, we would expected families to prefer one large and one small car, rather than two medium cars.

If volume is normally distributed among the models in each class, then the expected absolute value of this difference is given by (see Train (1986)):

$$E_{ij\,\epsilon c^1 \times c^2}\left|V^i - V^j\right| = \left(V^i - V^j\right)\left(\Phi\left(\frac{V^i - V^j}{\sigma}\right) - \Phi\left(\frac{V^j - V^i}{\sigma}\right)\right) + 2\,\sigma\Phi\left(\frac{V^i - V^j}{\sigma}\right)$$

Where:

- V^i is the volume of model *i*;

[17] Calculated compared to a "base" model with equal market shares.

[18] As discussed before, Multiple Discrete-Continuous Choice Models explicitly account for preferences for diversity, but face limitations of their own.

- c^1 and c^2 are the sets of models in class 1 and 2 respectively
- Φ is the cumulative standard normal distribution
- ϕ is the standard normal probability density
- σ is the variance of $V^{i-} - V^{j}$

Due to the significant increase in the number of possible choices, we have not estimated any Alternative Specific Constant (ASC) or coefficient of the covariance matrix in order to keep the number of covariates within the constraints imposed by the econometric software.

The estimation results are summarized in Table 5 (limited to the significant effects), and the definition of the variables is given in Table 6.

Table 5 Estimation result for the two car household model

Explanatory variables		Estimated coefficient	t-statistic	95% confidence interval	
1.	**PR11TYLOW**	-0.29852E-04***	4.91	-0.41779E-04	-0.17924E-04
2.	**PR11TYHIG**	0.27980E-04***	7.50	0.20671E-04	0.35288E-04
3.	**VOLTGRGZ**	0.05347***	2.62	0.01354	0.09340
4.	**VOLTYLOW**	0.08116***	3.75	0.03877	0.12355
5.	**KWTGHFDYG**	-0.00675***	5.44	-0.00918	-0.00432
6.	**TRAFFTAXT**	-0.00372***	14.80	-0.00421	-0.00323
7.	**FUELCOSTT**	-0.03434***	2.80	-0.05834	-0.01034
8.	**LOGRCT**	0.09933***	4.03	0.05107	0.14759
9.	**EXP_VOLD**	0.71508***	4.70	0.41670	1.01346

*MNL-model estimated with 465 alternatives (pairs of car classes). Number of observation used: 866 out of the 1145 families (279 families with unknown values for one or more explanatory variables were not witheld). log likelihood-value of -8422. A 1%, 5% or 10%- significance level is represented with ***, ** or *.*

Table 6 Variables used in the MNL model for two car households

Explanatory variables		Definition
1.	**PR11TYLOW**	Interaction variable between the sum of the total acquisition costs (depreciated purchase price, circulation tax and CO2 subsidy) and an effects coded dummy for low income families (net family income ≤2000 EUR per month)
2.	**PR11TYHIG**	Interaction variable between the sum of the total acquisition costs (depreciated purchase price, circulation tax and CO2 subsidy) and an effects coded dummy for high income families (net family income > 3000 EUR per month)
3.	**VOLTGRGZ**	Interaction variable between the sum of the approximate volumes (length x width x height) and an effects coded dummy for large families (> 4 family members)
4.	**VOLTYLOW**	Interaction variable between the sum of the approximate volumes (length x width x height) and an effects coded dummy for low income families (net family income ≤2000 EUR per month)
5.	**KWTGHFDYG**	Interaction variable between the sum of the engine powers and an effects coded dummy for young head of the family (<40 year)
6.	**TRAFFTAXT**	Sum of the annual traffic taxes in 2011
7.	**FUELCOSTT**	Sum of the fuel costs (EUR/100km)
8.	**LOGRCT**	LOGRCT = LOG(number of possible combinations of models within the pair of vehicle classes /total number of possible combinations of models over all vehicle classes), this is the link with the implicit nest of individual models
9.	**EXP_VOLD**	Expected absolute difference between the approximate volumes of both cars

The discussion of the results is largely analogous to the discussion for one car families.

For instance, the interaction terms between the acquisition costs and the family income confirm the existence of a “snob” effect for high income households, while “low” and “medium” income

households react as expected. Also, we cannot reject the hypothesis that θ lies in the [0 ; 1], and thus that the observed choices are consistent with utility maximization.

The impact of the traffic tax is highly significant, contrary to what we obtained for the one car households.

The impact of the sum of the volumes is also as expected. Interestingly, the expected absolute difference between the volumes is also highly significant: the higher the expected difference, the more likely that a given pair will be chosen (for a given sum of volumes). This confirms that households prefer cars that are complementary.

The interpretation of the other coefficients is straightforward.

Note that it was not possible to calculate a pseudo-R^2 in this model: the estimation of the base model with observed market shares would require the estimation of 464 alternative-specific constants, which exceeds the limits of our econometric software.

7. THE VEHICLE QUANTITY MODEL

In this submodel, we estimate the probability that a family owns 0, 1 or 2 cars. We thus exclude families with 3 or more cars, or families who use one or more company cars.

This submodel can be interpreted as the highest level of a nested logit model, where each nest corresponds to the number of cars chosen. Thus, in the case of families with 1 or 2 cars, the expected value of the maximal utility that can be derived from owning 1 or 2 cars has to be added as explanatory variable.

Table 8 and Table 9 describe the explanatory variables for the expected utility of owning one or two cars, respectively. For the utility of owning zero cars, the only explanatory variable used is the ASC.

Compared to a base model using the observed market shares, we obtain a pseudo-R^2 of 0.2627. The vehicle quantity model has thus a much higher predictive power than the vehicle choice model for one car families.

Table 7 Estimation results for the quantity model

Explanatory variable	Estimated coefficient	t-statistic	95% confidence interval	
1. LOGSUM2	0.21965***	5.22	0.13718	0.32011
2. LOGSUM1	0.35784**	2.41	0.06685	0.64884
3. GHFD_VR2	-0.18368*	1.68	-0.39797	0.03061
4. GHFD_VR1	-0.23966***	2.88	-0.40255	-0.07678
5. SINGLE2	-2.20145***	5.97	-2.92470	-1.47819
6. SINGLE1	-0.79457***	7.75	-0.99544	-0.59371
7. LEDENA1	-0.19888**	2.16	-0.37902	-0.01873
8. Y_LOW2	-1.86867***	12.33	-2.16566	-1.57169
9. Y_LOW1	-0.39457***	3.16	-0.63910	-0.15005
10. Y_HIGH2	1.02555***	7.90	0.77103	1.28006
11. DIPL_L2	-0.76937***	5.70	-1.03382	-0.50491
12. DIPL_L1	-0.60847***	5.50	-0.82530	-0.39165
13. DIPL_H2	0.57568***	3.48	0.25169	0.89968

14.	**DIPL_H1**	0.44842***	3.02	0.15779	0.73904
15.	**GEMTH_GR2**	-0.39526***	3.78	-0.60002	-0.19050
16.	**GEMTH_GR1**	-0.22657***	2.60	-0.39722	-0.05593
17.	**BUSWEK2**	-1.08041***	9.11	-1.31287	-0.84795
18.	**BUSWEK1**	-0.68349***	7.76	-0.85608	-0.51089
19.	**FIETSWEK1**	0.14135***	3.05	0.05062	0.23208
20.	**TREINWEK2**	-0.57566***	2.92	-0.96167	-0.18965
21.	**TREINWEK1**	-0.49402***	2.89	-0.82939	-0.15865
22.	**LFTGHFD2**	0.20841***	5.09	0.12815	0.28868
23.	**LFTGHFD1**	0.16489***	5.92	0.11027	0.21952
24.	**LFTGHFDE2_2**	-0.00189***	5.16	-0.00261	-0.00117
25.	**LFTGHFDE2_1**	-0.00145***	6.08	-0.00191	-0.00098
26.	**WNWRK2**	0.06027***	5.01	0.03668	0.08387
27.	**WNWRK1**	0.03116***	3.07	0.01126	0.05106
28.	**WNWRKE2_2**	-0.00023***	2.86	-0.00039	-0.00007

*MNL-model estimated with 3 alternatives (0 car / 1 car / 2 cars). Observations used for the estimation: 2407. log likelihood= -1672. A 1%, 5% or 10%- significance level is represented with ***, ** or *.*

Table 8 Explanatory variables used in the quantity model (utility of 1 car alternative)

Verklarende variabele		Definitie
0.	**LOGSUM1**	The expected value of the maximum utility that can be obtained from choosing one car.
1.	**GHFD_VR1**	Effects coded dummy for a female head of family
2.	**SINGLE1**	Effects coded dummy for single member household
3.	**LEDENA1**	The number of household member
4.	**Y_LOW1**	Effects-coded dummy for low income household (net family income ≤2000 EUR/month)
5.	**DIPL_L1**	Effects code dummy if the highest level of qualification of the head of family is primary education or less
6.	**DIPL_H1**	Effects code dummy if the head of family has obtained a higher education
7.	**GEMTH_GR1**	Effects coded dummy if the household has its home address in a large or medium sized city.
8.	**BUSWEK1**	Effects coded dummy for families who use the bus at least once per week
9.	**FIETSWEK1**	Effects coded dummy for families who use the bicycle at least once per week
10.	**TREINWEK1**	Effects coded dummy for families who use the train at least once per week
11.	**LFTGHFD1**	The age of the head of the family
12.	**LFTGHFDE2_1**	The square of the age of the head of the family
13.	**WNWRK1**	Distance from the home address to the place of work

Table 9 Explanatory variables used in the quantity model (utility of 2 car alternative)

Verklarende variabele		Definitie
0.	**LOGSUM2**	The log sum for the 2 car alternative: the expected value of the maximum utility that can be obtained from choosing two cars.
1.	**GHFD_VR2**	Effects coded dummy for a female head of family
2.	**SINGLE2**	Effects coded dummy for single member household
3.	**Y_LOW2**	Effects-coded dummy for low income household (net family income ≤2000 EUR/month)
4.	**Y_HIGH2**	Effects-coded dummy for high income household (net family income >4000 EUR/month)
5.	**DIPL_L2**	Effects code dummy if the highest level of qualification of the head of family is primary education or less
6.	**DIPL_H2**	Effects code dummy if the head of family has obtained a higher education
7.	**GEMTH_GR2**	Effects coded dummy if the household has its home address in a large or medium sized city.
8.	**BUSWEK2**	Effects coded dummy for families who use the bus at least once per week
9.	**TREINWEK2**	Effects coded dummy for families who use the train at least once per week
10.	**LFTGHFD2**	The age of the head of the family
11.	**LFTGHFDE2_2**	The square of the age of the head of the family
12.	**WNWRK2**	Distance from the home address to the place of work

Verklarende variabele	Definitie
13. **WNWRKE2_2**	Square of the distance from the home address to the place of work

Table 7 summarizes the estimation results. All coefficients are significant at the 1% level, except LOGSUM1, LEDENA1 (both significant at the 5%-level) and GHFD_VR1 (at the 10%-level).

The coefficients for **LOGSUM2** and **LOGSUM1** confirm that the expected value of the maximal utility that can be obtained from owning one or two cars has a significant impact on the utility of choosing one or two cars. This means that the probability that a family chooses one or two cars is affected by the characteristics of the available models, but also by how the socio-economic characteristics of households influence the chosen vehicle class.

From **GHFD_VR2** , **GHFD_VR1**, **SINGLE2** and **SINGLE1**, we infer that having a female head of family or living alone reduces the utility of owning cars compared to owning no cars.

The sign of **LEDENA1** shows that an increase in the number of household members leads to a decrease in the utility that a household will own one car. However, the coefficient for **LEDENA2** was not significant, and it is thus not clear what should be inferred from this result.

The coefficients for **Y_LOW2**, **Y_LOW1** and **Y_HIGH2** show that having a low income reduces the utility of owning at least one car, and that a high income increases the utility of owning two cars[19].

DIPL_L2, **DIPL_L1**, **DIPL_H2** and **DIPL_H1** reflect the impact of the educational qualifications of the head of the family. Having a head of family who has a primary education or less reduces significantly the utility of owning at least car, while having a head of family who has followed a higher education has the opposite effect.

Living in a large or medium sized city (**GEMTH_GR2** and **GEMTH_GR1**) also significantly reduces the utility of owning at least one car, possibly because of better accessibility of most destinations or due to a higher supply of alternative transport modes in cities.

The impact of the use of alternative modes is also captured directly in **BUSWEK2**, **BUSWEK1**, **FIETSWEK1**, **TREINWEK2** and **TREINWEK1**. Families who use the bus or the train on at least a weekly basis have a lower utility of owning at least one car than families who don't. There was however no significant impact of bicycle use on the utility of owning two cars, although it has a positive impact on the utility of owning one car. The effects of metro and tram use were not found to be significant. However, one should be careful in the interpretation of these coefficients. Indeed, the actual use of alternative modes (as opposed to the supply of alternative modes) is also an endogenous variable: the use of different transport modes (including of cars) is determined simultaneously. These coefficients indicate that there are some common "deep" variables affecting bus and train use on the one hand, and car ownership on the other hand, but do not imply causality in one direction or the other.

The sign of the coefficients of the age related variables **LFTGHFD2**, **LFTGHFD1**, **LFTGHFDE2_2** and **LFTGHFDE2_1** confirm that the utility of owning one or two cars is a concave quadratic function of the age of the household head. Thus, when the head of the household becomes older, the utility of owning one or two cars first increases, and then decreases. The maximum utility, all other things being equal, of owning two cars is attained at the age of 55, while it is 57 for the utility of owning

[19] The effect on the probability of owning one car was not significant.

one car. This suggests that the utility of owning more than one car is strongly linked to certain phases in life, such as having children combined with a professional activity.

Finally, the variables linked to the home-work distance **WNWRK2**, **WNWRK1** and **WNWRKE2_2** show that the utility of owning two cars is a concave quadratic function of this distance. This is what we would have expected. Indeed, the best alternatives to private cars are available for either short (walking, cycling) or long distances (train). It therefore seems logical that the highest utility of owning two cars is obtained for intermediate distances, and that households are more likely to avoid the expense of a second car for the extreme cases. The utility of owning one car is an increasing linear function of the home-work distance, probably reflecting that owning at least one car offers significant benefits besides the use for commuting purposes.

8. DISTANCE MODEL FOR ONE CAR HOUSEHOLDS

For the distance model, we have used the technical and cost characteristics of the models owned by the households, rather than the averages of the classes to which these models belong as we did in the vehicle choice and quantity choice model.

In the case of one car households, the number of family members had no significant impact on the annual mileage. We have also tested the hypothesis whether the frequency of use of alternative modes has an impact on distance travelled. This has only been confirmed for the motorcycle. "Composite" indicators of public transport use have not been found to be significant either. As economic theory predicts, the fixed costs of car ownership have no impact on annual mileage.

The following variables have not been included due to the large number of missing values: participation in a carpool system, the use of the car for professional reasons and the gender of the head of family.

Table 10 is the correlation matrix for the continuous variables (where we have only represented the correlations exceeding 0.5). We see that the collinearity between the key explanatory variables is limited. The tables also suggests that the cylinder displacement, the power of the engine and the weight of the car are reasonable candidates for acting as instruments for the fuel cost per kilometer and the volume of the car, to the extent that they are themselves uncorrelated with the error term of the distance function.

Table 10: correlation matrix for the one car distance model

	totpr11	trafftax	Vastkm	ghfdgb	fuel_con	ledena	vol	cyl	kw	gewleeg
totpr11	1.00	.	.	.	.	.	.	.	0.58	0.51
Trafftax	.	1.00	.	.	.	.	.	0.91	0.72	0.64
Vastkm	.	.	1.00	.	.	.	.	.	.	.
Ghfdgb	.	.	.	1.00	.	.	.	.	.	.
fuel_con	.	.	.	.	1.00	.	.	.	0.56	.
Ledena	.	.	.	.	.	1.00	.	.	.	.
Vol	.	.	.	.	.	.	1.00	0.59	.	0.86
Cyl	.	0.91	.	.	.	.	0.59	1.00	0.81	0.78
Kw	0.58	0.72	.	.	0.56	.	.	0.81	1.00	0.65
Gewleeg	0.51	0.64	.	.	.	.	0.86	0.78	0.65	1.00

The results of the preferred models are summarized in Table 12 and the variables used are defined in Table 11. We have estimated three models: an OLS model, an IV model to correct for the endogeneity of the car's volume and fuel cost, and a model with correction for self-selection. For clarity, we have only represented the estimates of the regression coefficients and their respective significance levels[20]. The model is based on 1345 observations. The test statistics that evaluate the overall performance of each model are represented in separate tables.

Table 11: definition of the explanatory variables

Code	Definition
(Intercept)	Intercept term
vastkm	Distance to work
I(2011- ghfdgb)	Age of the head of the family in 2011
totink*i*	Income class
gemthuistypei	Municipality type
vol	Volume of the car
gmotori	Frequency of use of the motorcycle
fuel_con	Fuel cost in EUR per 100 km

Table 12: distance model for one car families

Covariates	OLS estimates	IV estimates	Correction for self selection
(Intercept)	2081.9916	180.8785	4595.6871
vastkm	109.6693 (**)	123.7998 (***)	102.1987 (**)
I((vastkm)^2)	-0.8082 (*)	-0.985 (*)	-0.7724 (.)
I(2011- ghfdgb)	189.0042 (.)	152.5368	106.4316
I((2011-ghfdgb)^2)	-2.8828 (**)	-2.4466 (*)	-2.1262 (*)
totink1	-810.7889	-510.03	-16.5944
totink2	-1423.5554 (*)	-1283.04 (*)	-941.172
totink3	215.8619	191.6528	-37.0803
totink4	-510.1942	-718.002	-870.329
totink5	105.478	-185.227	-196.6466
gemthuistype1	1801.972	1911.566 (.)	1987.8361 (.)
gemthuistype2	-2152.3556 (**)	-2057.51 (**)	-1958.189 (**)
gemthuistype3	-1002.4367	-857.799	-913.4155
gemthuistype4	-1224.3277	-1362.58	-1223.7164
gemthuistype5	-466.9044	-574.638	-564.4079
gemthuistype6	-7.4463	-39.9051	-55.7532
gemthuistype7	1784.5079 (**)	1671.663 (**)	1628.9089 (*)
vol	882.1992 (***)	1483.259 (***)	852.1694

[20] The significance codes are: 0 '***', 0.001 '**, 0.01 '*', 0.05 '.'

			(***)
gmotor1	2601.7336 (*)	2649.057 (***)	2607.1482 (*)
gmotor2	-3417.6329	-3396.63	-3228.9473
gmotor3	0.6413	125.7557	-129.8802
gmotor4	-5195.043	-5891.47 (*)	-5182.8357 (.)
fuel_con	-250.6279 (**)	-622.787 (*)	-249.9021 (**)
corr_cars_nu			-1409.2041
corr_cars_0			1901.9068 (.)

Table 13: Overall performance of the OLS model

Residual standard error:	**8486**
Multiple R-squared:	0.1788
Adjusted R-squared:	0.16
F-statistic:	13.08
p-value:	< 2.2e-16

Let us first consider the overall performance of the OLS model. We see that the variance of the explanatory variables explain barely 17.88% of the variance of the dependent variable. As we have considered close to all variables included in the OVG and the Febiac database in our regression, we can conclude that the existing data are relatively poor predictors of the annual mileage – we shall come back to this point later.

Using the Breusch-Pagan test, we cannot reject the null hypothesis of homoscedasticty.

Table 14: Breusch-Pagan test for the OLS model

BP = 20.9048
df = 22
p-value = 0.5266

How does the IV model compare to the OLS model?

As discussed above, one can reasonably expect that families who anticipate a high annual mileage, will prefer cars with (all other things being equal) lower fuel costs and a larger volume. We have therefore estimated a IV model with the following instruments: the acquisition cost of the car, the annual circulation tax, the number of members in the family, the number of bicycles in the possession of the household, the use of alternative modi, and an indicator of the flexibility of the household's head working regime.

Table 15 summarizes the main performance indicators of this alternative model.

Instrumental Variables must fulfil two requirements.

On the one hand, the chosen instruments need to be correlated with the endogenous variables: otherwise the instruments cannot accurately predict these variables. If the test statistic "weak instruments" is smaller than 10 (see Green (2012) p 250 for the details), one can conclude that the instruments are too weak – this is not the case here.

On the other hand, consistency requires that the instruments cannot be correlated with the error term of the regression. Under the null hypothesis that the instruments and the residues are not correlated, the test statistic of the Sargan test follows a χ^2 distribution whose degrees of freedom equal the difference between the number of instruments used and the number of endogenous variables. In this case, we cannot reject the null hypothesis.

Thus, we can conclude that the instruments we have used fulfill two criteria: sufficiently high correlation with the endogenous variables and independence from the error terms.

However, we also need to assess whether the IV estimates are an improvement compared to the OLS estimates. The null hypothesis is that both the OLS and the IV estimators are consistent. Under this null-hypothesis, the Wu-Hausman statistic follows a χ^2 distribution with degrees of freedom equal to the number of endogenous variables (see Greene (2012) p 237 for details). In this case, the null hypothesis cannot be rejected, and we prefer to use the OLS estimates, because they are more accurate[21].

Table 15: Overall performance of the IV model

Tests IV					
Diagnostic	tests:				
	df1	df2	statistic	p-value	
Weak instruments	29	1295	18.309	<2e-16	***
Sargan	27	NA	24.433		
Wu-Hausman	2	1320	2.148	0.117	
Residual standard error:	8574	on 1322 degrees of freedom			
Multiple R-squared:	0.1615				
Adjusted R-squared:	0.1475				
Wald test	11.95	on 22 and 1322 DF,			
p-value:	< 2.2e-16				

In the third version of the model, we have added two terms to correct for self-selection. In order to understand how these terms have been calculated, we first define some additional variables:

- σ is the standard deviation of the OLS model
- *J=1,…,M* are the available discrete choices
- P_j is the probability that alternative *j* is chosen
- r_j is the correlation coefficient between the error term in the linear model and the unobservable variables for each alternative in the choice model

In our case, the *M* alternatives are: (1) the household chooses two cars, (2) the household chooses one car (3) the family has no car.

[21] To be more precise, their covariance matrix is asymptotically smaller.

Dubin en McFadden (1984) have shown that, if the probability that a given number of cars is chosen are logit, and the OLS error terms follow a normal distribution, then we need to add the following correction term to the independent variables:

$$\sigma \frac{\sqrt{6}}{\pi} \sum_{j=2,\cdots,M} r_j \left(\frac{P_j ln(P_j)}{1-P_j} + ln(P_1) \right)$$

As the choice probabilities have been estimated in the vehicle quantity model, it is possible, for each *j*, to estimate σr_j jointly with the distance function using OLS.

In our model, *corr_cars_nu* is the Dubin-McFadden correction term for the "two cars" alternative, and *corr_cars_0* is the Dubin-McFadden correction term for the "zero cars" alternative. None of these terms is significantly different from zero at the 5 % level, and *corr_cars_0* is only significant at the 10 level. Nevertheless, we can observe that the introduction of these correction terms has led to a decrease in the significance levels of several variables (the age of the head of the household, the family income and the quadratic term for the distance to work) who also affect the vehicle quantity model.

Anyhow, we cannot reject the null hypothesis that the correlation between the error terms of the distance model and the error term of the quantity model is zero. We therefore stick to the OLS model without correction term.

We can conclude that, contrary to our prior expectations, OLS is an appropriate approach for the distance model for "one car" households. We will therefore focus our interpretation of the model on the OLS results.

As expected, distance to work (*vastkm*) affects the annual mileage. Both the linear and the quadratic term are statistically significant. The lower significance of the quadratic term can be understood in the light of the distribution's skewness (see **Error! Reference source not found.**): the vast majority of observed values of *vastkm* are in the [0,1] interval, and estimates outside this interval are thus subject to very high uncertainty.

The marginal influence of vastkm is thus represented by $y = g(x) = -0.81 * x^2 + 109.67 * \mathrm{x}$. This function is maximized when x = 68 km. This probably corresponds to the point where the distance to work has become so large, that a household probably gains by buying a second car, or where commuting by train becomes more attractive than commuting by car[22]. However, modal choice falls outside the scope of this study, and we will not explicitly test this hypothesis.

We also see that the age of the head of the family has a quadratic impact on the annual mileage, with marginal effect: $y = f(x) = -2{,}88 * x^2 + 189 * \mathrm{x}$. This function is maximized when x = 33. One possible explanation for the quadratic term is that, when people reach a certain age, they enter a phase in their life where the annual total mileage of the family increases significantly[23]. Households who previously owned just one car may then decide to purchase a second one. This high annual mileage will then be spread over two cars- remember that we have shown that the age of the head of household has indeed a significant effect on the number of cars owned.

[22] In the vehicle quantity model, we have indeed shown that distance to work has an impact on the number of cars owned by the household.

[23] The typical explanation being dependent children who need to brought to school or to leisure activities.

On prior grounds, one would expect a household's income to be an important determinant of annual mileage, especially for leisure purposes: indeed, a higher income household cannot just afford to drive longer discretionary miles, but it can also afford to pay for the leisure activities that are offered at the destinations of travel. However, contrary to our expectation, a household's income does not seem to be a good predictor of its annual mileage (given ownership): it is only for the second income class[24] that the income dummy is significant.

Similarly, the characteristics of the household's municipality of residence are insignificant for most types of municipalities that are distinguished in the OVG. For type 2 ("central municipality in large urban areas"), the effect is negative, as expected. For type 7 ("small urban area") the effect is positive, also as expected.

The volume of the car is highly significant and positive. Although we cannot reject the hypothesis that the OLS and the IV model are both consistent, the effect is almost twice as large in the IV model as in the OLS model.

As already discussed above, the only alternative mode which seems to affect annual mileage is the motorcycle, and even then the effect is limited to one level of the variable: people who drive weekly with their motorcycle (level 1) drive significantly less.

Finally, the fuel cost per kilometer has the expected negative sign. Although we cannot reject the hypothesis that the OLS and the IV model are both consistent, the effect is almost twice as large (in absolute value) in the IV model as in the OLS model.

9. DISTANCE MODEL FOR TWO CAR HOUSEHOLDS

In this section, we provide estimates for the annual mileage for each car owned by the two-car households.

We have maintained all variables that we had also used for the one car family model. Moreover, we have taken into account that the use profile of both cars may be different. It is for instance likely that the largest car will mainly be used for travelling long distances for professional or commuting purposes, while the smaller car will mainly be used by the adult member of the household who faces shorter commuting distances. The larger car will probably also be used for longer leisure travel. We have therefore created three additional dummy variables[25], for the oldest car (*older_car*) , the cheaper car (*cheap_car*) and the smaller car (*small_car*).

Table 16 summarizes the distance model for two car household, based on a sample with 1188 observations (594 households with two cars). We have again put the three estimation results in a single table.

[24] Income between 1000 and 2000 EUR per month.

[25] As in the vehicle choice model for two car families, this is again a way to implicitly account for the preferences for diversity that are considered explicitly in the Multiple Discrete-Continuous Choice Models.

Table 16 Distance model for two car households

Covariates	OLS estimates	IV estimates	Correction for self selection
(Intercept)	2290.3014	87.7391	6975.6456
Vastkm	137.2902 (***)	141.7094 (***)	98.435 (*)
I((vastkm)^2)	-0.7404 (.)	-0.7836 (.)	-0.4098
totink1	-3499.859	-3664.76	-2354.4666
totink2	-2564.2338 (*)	-2512.02 (*)	-1272.8479
totink3	595.7108	717.0636	1149.6901
totink4	57.3232	152.8517	-331.2264
totink5	1973.8203 (.)	1962.748 (.)	688.3815
gemthuistype1	-978.9672	-921.424	-806.4346
gemthuistype2	-2777.3173 (.)	-2728.21 (.)	-2667.8379 (.)
gemthuistype3	155.4574	111.5158	298.0022
gemthuistype4	2088.6777	2044.433	2071.3801
gemthuistype5	859.8023	882.5155	766.1516
gemthuistype6	-372.0287	-420.623	-417.4241
gemthuistype7	1273.8538	1312.605	1109.9191
vol	1082.7289 (***)	1121.908 (.)	1051.8994 (***)
gmotor1	658.2269	659.0196	641.479
gmotor2	5984.3335 (*)	5933.616	5917.4667 (*)
gmotor3	-2683.2124	-2702.71	-2555.2844
gmotor4	-1650.8558	-1663.69	-1727.7983
fuel_con	23.0891	158.9388	36.0056
older_car	-1472.7303 (*)	-1507.76 (*)	-1479.0506 (*)
cheap_car	-399.2922	-291.399	-399.7472
small_car	-2379.8382 (**)	-2159.67	-2434.6618 (**)
gtram1	-2094.4403 (.)	-2112.22 (.)	-1960.7652
gtram2	-2557.3713 (*)	-2566.19 (*)	-2381.7721 (.)
gtram3	-1643.4318	-1622.49	-1679.8462
gtram4	-2239.9742	-2202.59	-2044.5486
gtrein1	2041.7605 (**)	2058.722 (**)	1892.0577 (*)
gtrein2	2037.5634 (*)	2062.681 (**)	1962.1992 (*)
gtrein3	1843.6011	1816.456	1921.1299
gtrein4	-1842.0633	-1898.33	-1936.7262
corr_cars_nu			4847.5088
corr_cars_0			-3701.5331

Let us first discuss the key statistics of each model.

Although the model contains several highly significant regressors, we also observe that the overall fit of the OLS model is low. As in the one car model, we cannot reject the null hypothesis of homoscedasticity.

Table 17: Overall fit of the OLS model for two cars

Residual standard error:	10880	on 1156 degrees of freedom
Multiple R-squared:	0.1329	
Adjusted R-squared:	0.1097	
F-statistic:	5.718	on 31 and 1156 DF
p-value:	< 2.2e-16	

Table 18: Breusch-Pagan test the OLS estimates for the two car distance model

BP = 24.481
df = 22
p-value = 0.3225

For the IV model, we have used all the instruments that we had used in the one car model, but we have also added the age of the head of the family (which was not significant as a regressor). As in the one car model, the "weak instruments" instruments test shows that the correlation between the instruments and the endogenous variables is sufficiently high, and that we cannot reject the hypothesis that the instruments are not correlated with the error terms. However, the "Wu - Hausman" test also shows that we cannot reject the hypothesis that the OLS estimators are consistent if the IV estimators are consistent. We therefore continue to use the OLS model, which is more accurate.

Table 19 diagnostics tests for the use of IV in the two car model

	df1	df2	statistic	p-value	
Weak instruments	18	1140	28.944	<2e-16	***
Sargan	16	NA	14.182	0.585	
Wu-Hausman	2	1154	0.465	0.628	

In the third version of the model, we have added two correction terms for self-selection: *corr_cars_nu* is the DubinMcFadden correction term for the "one car" alternative, and *corr_cars_0* is the DubinMcFadden correction term for the "1 auto" alternative. None of these correction terms is significantly different from zero. Nevertheless, we observe a decrease in the significance levels for distance to work, the income class and the type of municipality, three variables that also affect the number of cars owned.

Therefore, we will again focus our discussion of the individual regressors on the OLS estimates, with an emphasis on the differences with the one car model.

First, the age of the head of the household is no longer significant. One possible explanation is that households only purchase a second car if several household members need a car for commuting to

work or school, or for their leisure activities. For these families, once the decision has been taken to buy a second car, the discretion to modify the annual mileage is limited. As discussed above, these reasons to buy a second car are to some extent age-related, and the quantity model has confirmed that the age of the household head has a significant impact on the number of cars owned.

Second, we also see that significance of the quadratic term for the distance to work is now much lower, and the maximum mileage is obtained for a distance to work of 93 km. As discussed above, one possible explanation is that families who do not purchase a second car have better access to public transport. Testing this hypothesis would need additional data, and is a possible avenue for further research.

Third, the influence of the family income remains weak. One possible explanation is that most families have limited overall discretion in their annual mileage, which is mainly determined by the distance between their place of residence and the places where they perform activities. Therefore, a household which expects to drive a lot will rather save money by buying a cheaper car than by reducing distance driven. The vehicle choice model has indeed confirmed that a household income's has an impact on how the average acquisition cost of a car class affects the utility of choosing a car from this class.

Fourth, the impact of the place of residence is even lower than in the one car model. Here as well, one plausible explanation is that the impact of the place of residence is mainly felt through the number of cars owned by the household, rather than on the distances traveled per car.

Fifth, the car's volume continues to have a highly significant effect, but the fuel cost is no longer significant.

Sixth, the influence of motorcycle use remains limited.

Seventh, as expected, the oldest car is used significantly less than the new car. The same holds true for the smaller car. This confirms that, in two car households, these cars perform different functions.

Eight, tram and train use have a limited significant influence in the two car model, while they had no significant impact in the one car models. However, the sign of some of the dummies for tram use are counterintuitive, as they imply that people who use the tram less than once a month travel less by car than people who use the tram on a daily basis. The results for train use, however, are in line with intuition, as they imply that frequent train users drive less.

10. CONCLUSIONS

In this paper, we have estimated a discrete-continuous model of vehicle demand and use for the Belgian region of Flanders, combining the results of the official regional travel survey with a detailed database of vehicle characteristics.

In a first step, we have modelled the choice of the class of car(s) owned by the household. For one car households, we have used a Nested Logit model, where the nests are defined by the average construction year of the cars in each class (see Table 3). For two car households, we have used a Multinomial Logit model (see Table 5) to model the choice of a pair of vehicle classes. Several

variables have a highly significant impact in line with the theoretical expectations. Nevertheless, the overall predictive value of these models is low.

In a second step we have modelled the number of cars owned by the household for households with two or less cars – see Table 7. We see that the number of cars is affected, not just by the socio-demographic characteristics of the household, but also by the economic and technical characteristics of the car models that are available on the market. For this model, the overall predictive value is satisfactory (pseudo-R^2 equals 0.26).

Finally, we have modelled the annual mileage of each car owned by the household, conditionally on the number of cars owned. Formal statistical tests have shown that there was no need to correct for endogeneity or self-selection bias. The preferred model for one car households is summarized in Table 12 and for two car households in Table 16. As in the vehicle choice model, we combine several highly significant individual regressors with a low overall predictive value.

We can thus conclude that the only variable for which we have obtained a satisfactory overall predictive value is the number of cars owned. It is important to understand the underlying reasons for these results.

A first point is that many responses in the OVG are incomplete. We have already mentioned above that several potentially important variables were not included in the model because the response rate for these variables was too low. Moreover, many responses were inaccurate. We have introduced several filters in our preliminary data analysis to eliminate obviously wrong answers but there is no waterproof way to filter inaccurate but credible answers.

Second, there are also important gaps in the information with respect to the car models. As explained before, these gaps were, where possible, filled using our expert judgment, but it is clear that more detailed observed characteristics would be preferable.

Third, there are several, potentially key, determinants of mobility behavior that are not included in the OVG. As is generally acknowledged, transport demand is a derived demand. Travel surveys should therefore include information on the activities in which people participate. Distance to work is for instance an indicator for the activity “commuting”. However, for other activities (such as visiting friends and relatives) we only have very gross proxies, such as the size of the household. It is therefore highly desirable that future versions of this survey identify indicators with a higher predictive value for travel behavior for other than commuting purposes.

This is especially relevant in the light of important future challenges. Indeed, if the objective of transport modelling is to plan for capacity, then understanding peak behavior suffices. However, there is an increasing interest in the environmental impacts of transport, and these are related to total travel, not just peak travel.

Fourth, we can expect that the supply of alternative modes has an impact on the number of cars owned and the yearly mileage. In the current paper, we have used the frequency of use of these alternative modes as a proxy, although this is also an endogenous variable in an overall model of travel behavior. The type of municipality where a household resides could also act as a proxy for the availability of public transport, but our results show that this is enough. However, developing reliable indicators for public transport availability is not obvious: the proximity of a bus stop does for instance not contain any information on the frequency of the offer, or on the quality of the connections.

Fifth, we have limited ourselves to families who effectively own their cars. We have argued above why households with company cars are likely to behave differently from families who own their cars. Households with company cars can however also be expected to have different socio-economic characteristics. They are more likely to also drive more kilometers for professional reasons, and are also more likely to have higher incomes[26]. As a result, our estimates for high income families are probably not representative for the total population of high income families. Therefore, developing a vehicle choice and use model for company cars is an important subject for further research – a crucial element will be how to estimate the actual cost faced by the households.

A final point is related to the definitions of the car classes. In the vehicle type model, we have grouped the make and models in different classes, mainly based on the body type. However, we have seen that the available data do not always contain all the information that is needed for a meaningful classification of individual models. For instance, it is not possible to identify station wagons. It is thus possible that, for some non-observed elements, the variation in a class is larger than the variation between classes. The results of the choice model suggest that the criteria that were used for this classification may not be the criteria that households use in the choice of a specific car class. Finding more relevant classification criteria is a possible subject for further research.

REFERENCES

Berkovec J. and Rust J. (1985), A nested logit model of automobile holdings for one vehicle households, Transportation Research – B, Vol 19B, N° 4, pp 275-285

Bhat, C.R., and A. Pinjari, "Multiple Discrete-Continuous Choice Models: A Reflective Analysis and a Prospective View," forthcoming, Handbook of Choice Modelling edited by S. Hess and A. Daly, Edward Elgar Publishing Ltd.

Bhat, C.R., and S. Sen (2006), "Household Vehicle Type Holdings and Usage: An Application of the Multiple Discrete-Continuous Extreme Value (MDCEV) Model," Transportation Research Part B, Vol. 40, No. 1, pp. 35-53.

Boussauw K. (2011). Ruimte, regio en mobiliteit. Aspecten van ruimtelijke nabijheid en duurzaam verplaatsingsgedrag in Vlaanderen. ("Space, region and mobility. Topics in spatial proximity and sustainable mobility behavior in Flanders.")) Garant Uitgevers

De Jong, G. (1991), An indirect utility model of car ownership and use. European Economic Review, 34, 971-985

De Jong, G. (1997). A micro-economic model of the joint decision on car ownership and car use in: P. Stopher and M Lee-Gosselin (Eds): Understanding travel behavior in an era of change. Pergamon, Oxford.

de Jong G., Fox J., Pieters M., Daly A. & Smith R. (2004). *A comparison of car ownership models*, Transport Reviews 24(4): 379-408, White Rose University Consortium.

[26] In Belgium, the private use of company cars is a widespread technique to avoid income taxes and social security contributions.

de Jong G. & Gunn H. (2001). *Recent Evidence on Car Cost and Time Elasticities of Travel Demand in Europe*, Journal of Transport Economics and Policy 35(2): 137-160, White Rose University Consortium.

de Jong G., Kouwenhoven M., Geurs K., Bucci P. & Tuinenga J. (2009). *The Impact of Fixed and Variable Costs on Household Car Ownership*, Journal of Choice Modelling 2(2): 179-199.

Dubin, J. A & McFadden, D. L, 1984. "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," Econometrica, Econometric Society, vol. 52(2), pages 345-62, March.

Econometric Software, Inc., *Frequently Asked Questions - Questions about R-squareds*, consulted on 3/2/2014 via
http://www.limdep.com/support/faq/

Fang, H.A. (2008), A discrete-continuous model of households' vehicle choice and usage, with an application to the effects of residential density, Transportation Research part B 42, pp 736-758

Hanemann (1984), Discrete/Continuous Models of Consumer Demand, Econometrican Vol 52, N°3, pp 541-562

Hensher D., Smith N.C., Milthrope F.W. and Barnard P.O. (1992), *Dimensions of Automobile Demand. A Longitudinal Study of Household Automobile Onwership and Use*. Elsevier Science Publishers. Amsterdam

Hensher D., Rose J. & Greene W. (2005). *Applied Choice Analysis - A Primer*, ISBN 978-0-521-60577-9, 717 pp, uitgegeven door Cambridge University Press, Cambridge.

Hoetker G. (2007). *The Use of Logit and Probit Models in Strategic Management Research: Critical Issues*, **Strategic Management Journal 28**: 331-343, uitgegeven door John Wiley & Sons, Ltd.

IDRE - Institute for Digital Research and Education, *FAQ: What are pseudo R-squareds?*, consulted on 3/2/2014 via
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

Kaufman J., *Can pseudo-R-squareds from logistic regressions be compared and used as a measure of fit?*, consulted on 3/2/2014 via
http://andrewgelman.com/2009/11/03/can_pseudo-r-sq/

Louviere, J., Hensher, D., & Swait, J. 2000. *Stated Choice Methods - Analysis and Applications* Cambridge, Cambridge University Press.

Manski C.F. and Sherman L. (1980), An empirical analysis of houeshold choice among motor vehicles, Transportation Research A, Vol 14A, pp 349-366

Mayeres, I. & M. Vanhulsel (2014), Simulatiemodel voor de hervorming van de verkeersbelastingen, Rapport Steunpunt Fiscaliteit en Begroting II ("Simulation model for the reform of traffic taxes. Report for the Policy Research Centre – Fiscal Policy).

McFadden D. (1978). *Spatial Interaction Theory and Planning Models - [25] Modelling the Choice of Residential Location*, ISBN 0886-0416, Karlqvist A., Lundqvist L., Snickars F. & Weibull J. (eds.), 388 pp, North Holland Publishing Company, Amsterdam.

Train K. (1986). *Qualitative Choice Analysis - Theory Econometrics, and an Application to Automobile Demand*, ISBN 0-262-20055-4, 247 pp, The MIT Press, Cambridge, Massachusetts.

Train, K. (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press, 2nd edition

Vlaamse Overheid - Departement Mobiliteit en Openbare Werken (MOW). Onderzoek Verplaatsingsgedrag Vlaanderen. Mobiel Vlaanderen ("Flemish Government – Department of Mobility and Public Works. Travel survey Flanders ") Consulted on 3/2/2014 via http://www.mobielvlaanderen.be/ovg/ovg04.php?a=19&nav=11

ANNEX: MCFADDEN'S APPROACH TO APPROXIMATING INCLUSIVE VALUES

Let $c = 1,\ldots,C$ be the vehicle classes with $m_c = 1,\ldots, M_c$ the individual models in class c. x_{cm} is the vector of observed attributes of the individual models. If it is appropriate to model the choice between individual models with a nested logit structure, using the classes as nests, then the probability of choosing model m_c in class c is given by (where 1 -σ is the degree of independence of the random terms for the models within a given class[27]:

$$P_{mc} = \frac{e^{(1-\sigma)I_c}}{\sum_{b=1}^{C} e^{(1-\sigma)I_b}}$$

where I_b is the Inclusive Value of class b - this is the expected value of the maximum utility that can be obtained from choosing a model in class b:

$$I_c = \log \sum_{m=1}^{N_c} e^{\frac{\beta' x_{cm}}{1-\sigma}}$$

where β are the parameters to be estimated.

McFadden has shown that, if the x_{cm} are identically normally distributed with mean x_c^* , then:

$$I_c \underset{a.s.}{\longrightarrow} \frac{\beta' x_c^*}{1-\sigma} + \log N_c + \frac{1}{2}\left(\frac{\omega_c}{1-\sigma}\right)^2$$

Where N_c is the number of models in class c and $\omega_c^2 = var(\beta' x_{cm})$.

The intuition behind this result is that, if the covariates follow a multivariate normal distribution, then the information contained in the expected values, the variances and covariances is sufficient to approximate the expected value of maximal utility. All other things being equal, households prefer a class with more models, and a higher variance in the underlying characteristics of the individual models: this reflects that this class offers a wider range of potential choices.

For estimation purposes, we assume that $\omega_c^2 = \beta' \Omega_c \beta$ where Ω_c is the covariance matrix of x_{cm}.

A full maximum likelihood estimation would require to estimate $\frac{\beta' x_c^*}{1-\sigma} + \log N_c + \frac{1}{2}\left(\frac{\beta' \Omega_c \beta}{1-\sigma}\right)^2$, taking into account the non-linear constraints. However, McFadden has shown that consistent estimators can be obtained by writing out the terms in the quadratic form $\beta' \Omega_c \beta$ as independent parameters and ignoring the non-linear constraints. This is the approach we have chosen here.

[27] We follow the notation used by McFadden, but leave out the terms that are not directly relevant for the present analysis.