Centre for Transport Studies
STOCKHOLM

An empirical study of aggregation of alternatives and its influence on prediction: case study of car type choice in Sweden

Shiva Habibi[a] Emma Frejinger[b]
Marcus Sundberg[a]

[a]*Centre for Transport Studies, Teknikringen 10, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden E-mail addresses: shiva.habibi@abe.kth.se, marcus.sundberg@abe.kth.se*

[b] *University of Montreal, Department of Computer Science and Operations Research, Pav. PAVILLON ANDRE-AISENSTADT, CP 6128 Succursale Centre-Ville, Montréal QC H3C 3J.*
emma.frejinger@umontreal.ca

CTS Working Paper 2015:3

Abstract
In the car type choice models, alternatives are usually grouped into categories by some of their main characteristics such as make, model, vintage, body type and/or fuel type. Each of these categories contains different versions of the cars that are usually not recognized in the applied literature. In this study we empirically investigate whether including the heterogeneity of these versions in the modeling do matter in estimation and prediction or not. We have detailed data on alternatives available on the market down to the versions level of each model which enables us to account for heterogeneity in the model. We also have Swedish car registry data as demand. We estimate different discrete choice models with different methods of correction for alternative aggregation including nesting structure. We estimate these models on based on year 2006 Swedish registry data for new cars, predict for 2007 and compare the results. The results show that including heterogeneity of cars' versions in the model improves model fitness but it does not necessarily improve prediction results.

Keywords: Aggregate alternatives, prediction, car type choice, discrete choice modeling, clean vehicles

Centre for Transport Studies
SE-100 44 Stockholm
Sweden
www.cts.kth.se

An empirical study of aggregation of alternatives and its influence on prediction: case study of car type choice in Sweden

Shiva Habibi[1], Emma Frejinger [2], and Marcus Sundberg [1]

[1] *KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. , Emails: shiva.habibi@kth.se, marcus.sundberg@abe.kth.se*

[2] *University of Montreal, Department of Computer Science and Operations Research, Pav. PAVILLON ANDRE-AISENSTADT, CP 6128 Succursale Centre-Ville, Montral QC H3C 3J. Email: emma.frejinger@umontreal.ca*

Abstract

In the car type choice models, alternatives are usually grouped into categories by some of their main characteristics such as make, model ,vintage, body type and/or fuel type. Each of these categories contains different versions of the cars that are usually not recognized in the applied literature. In this study we empirically investigate whether including the heterogeneity of these versions in the modeling do matter in estimation and prediction or not. We have detailed data on alternatives available on the market down to the versions level of each model which enables us to account for heterogeneity in the model. We also have Swedish car registry data as demand. We estimate different discrete choice models with different methods of correction for alternative aggregation including nesting structure. We estimate these models on based on year 2006 Swedish registry data for new cars, predict for 2007 and compare the results. The results show that including heterogeneity of cars' versions in the model improves model fitness but it does not necessarily improve prediction results.[1].

Keywords: Aggregate alternatives, prediction, car type choice, discrete choice modeling, clean vehicles

[1] The results of this paper have been presented in IATBR 2012

1 Introduction

The field of car type choice modeling has been extensively studied during previous decades (For the extensive reviews of car type choice literature see e.g. De Jong et al., 2004; Potoglou and Kanaroglou, 2008). These models are of interest to policy makers due to the high contribution of cars in energy consumptions and green house emissions. These models are employed to evaluate possible policies aiming at influencing the composition of the car fleet towards more energy and emission efficient fleet. The studies of Hugosson and Algers, 2012,Hensher and Plastrier, 1985,Mannering, 1983 and Page et al., 2000 are examples of estimation and application of theses models to analyze different policies in Sweden, Australia, US and UK, respectively.

One of the challenges in car type choice models is the great number of cars (alternatives) available to choose among. The problem is either lack of detailed data of available car choices or computational space capacity. The latter is not a determining factor anymore considering the recent advances in computer technology. Choo and Mokhtarian, 2004 summarize car type choice studies according to their sample size, model type, number of available car alternatives and final number of aggregated (grouped) alternatives. The common practice to deal with this problem has been grouping alternatives into categories by some of their main characteristics such as make, model ,vintage, body type and/or fuel type. Each of these categories contains different types of the cars with the same main characteristics *e.g.* make, model and vintage but different other characteristics such as engine size, power, weight and consequently different price. Each of these categories are represented by an alternative which characteristics are averaged over characteristics of the cars within that category. The study of Lave and Train (1979) was the first to employ MNL model to estimate a car type choice model with the purpose to evaluating transportation energy consumption policies. They estimate a multinomial logit model (MNL) in which available car alternatives are grouped into 10 size/price categories for the newly purchased cars. They take sale-weighted average of characteristics of cars in each group to represent aggregate alternatives. Ben-Akiva and Lerman (1985) introduce an approach for including aggregated alternatives in discrete choice models. However, very few studies have implemented this method in practice. Among

them, The work of Mabit, 2011 can be mentioned where he estimates a multinomial logit (MNL) model on Danish registry data to analyze the effect of Danish 2007 differentiated vehicle tax reform on new vehicle market. He groups cars into 424 categories by make/model/body-type/fuel-type of cars, each defining an aggregate alternative. He includes the 'log' of the number of alternatives within each category (*i.e.* sub-alternatives) for each aggregate alternative as an explanatory variable in his model. He also includes alternative specific constants for each alternative in his model. The parameter for this variable becomes positive and very significant explaining that considering the fact that each aggregate alternative represents certain sub-alternatives is very important in car type choices models. He concludes that the results indicate the influence of supply in car type choice which is usually not considered in literature.

Having access to the rich datasets of Swedish registry data as well as detailed data on alternatives available on the Swedish market, denoted by supply, enables us not only to include size of sub-alternatives (versions of car models) as Mabit, 2011 does, but also include heterogeneity of sub-alternatives, explicitly. The objective of this study is to investigate empirically whether or not the sub-alternatives are heterogeneous, also, whether considering the heterogeneity of versions in modeling improves explanatory power and/or goodness of fit or not. In our study, in order to observe chosen alternatives from registry data in supply we need to aggregate (group) alternatives based on make/model/fuel-type. We employ Ben-Akiva and Lerman's (1985) approach for aggregating alternatives in which the measure of size and heterogeneity of sub-alternatives are included in the model. To do so, we estimate different discrete choice models based on 2006 Swedish registry data. We employ these models to predict for 2007 considering the 10,000 SEK[2] purchase policy for clean cars implemented in 2007 and compare the predicted results of different models as well as with actual outcomes. To the best of our knowledge none of the existing literature in car type choice field, compare predictions to actual outcomes as we do here. The paper is organized as follows: next, in section 2, we describe the data used in this study. Section 3 presents modeling methodology and different model specifications. Estimation and prediction results are discussed in section 4, where we also compare the prediction results of different models and finally

[2] SEK equals to 0.15 US Dollars in April 2013

we propose future work and research directions in section 5.

2 Data

For the results presented in this paper we merge two different data sources for the two years of interest, namely 2006 and 2007. The first data source is the car register that contains all passenger cars in the Swedish fleet and some characteristics of each car. The second data source contains very detailed information about all car models, including price, that were available on the Swedish market these years. In this section we describe each of these data sources and finally how the two are merged.

2.1 Demand

The car register contains all passenger cars that are owned privately or by a company. The cars that are owned by companies can either be used for the company's activities or privately by employees who in this case pay a benefit tax. Since company cars are defined in an ambiguous way in the registry data, we focus on this segment in this paper. There were approximately 3.3 million privately owned passenger cars in traffic 2006 and 2007.

In addition to information specific to the registration of the car (e.g. first registration date and date for last status change), some main car characteristics are stored in the register such as brand, model name, vehicle year, fuel type, weight, power and body type. The age, gender and home municipality of the owner are also given in the register. The vehicle year is defined based on a combination of three attributes; model year, production year and first registration date because all three attributes are not available for all observations. Vehicle year is equal to model year if it is available, otherwise, the production year of the car and if this is not available either then it is equal to the year of first registration date.

Since we are interested in new cars these observations need to be selected. For this purpose cars that are registered for the first time a given year but that are actually older should be excluded. We consider that a car has been bought new in 2006 if the first registration date is equal to 2006 and the vehicle year is equal to 2006 or 2007. We

define new cars for 2007 in the same way. Imported cars are not included in any case. With this information there are 107,717 observations in 2006 and 116,566 in 2007. This definition of a newly bought car is slightly different from the one used in the official statistics that also counts older cars in. We choose to exclude these so that we can have a more accurate idea about the price paid for the car.

Table 1 reports the number and share of new cars sold by fuel type in 2006 and 2007. The share of sold petrol cars decreased with 20% mainly in favor of diesel cars but also ethanol cars. One can also note an increase in the share of clean petrol and diesel cars. The share of electic-hybrid cars and gas cars remain almost the same. It should be mentioned here that in Sweden, petrol, diesel and electric-hybrid cars are considered clean when their emission is less than 120gr/km and they should meet the Euro 4 (2005) standard requirements, furthermore, diesel cars should contain filters for particles.

Fuel Type	**2006**		**2007**	
	Number	Share	Number	Share
Petrol	83416	77.4	67011	57.5
clean petrol	2044	1.9	4959	4.3
Diesel	18650	17.3	38118	32.7
clean diesel	76	0.1	1508	1.3
El-hybrid	475	0.4	586	0.5
clean El	314	0.3	421	0.4
Ethanol	5107	4.7	10739	9.2
Gas	69	0.1	112	0.1

Table 1: Observations by fuel type in 2006 and 2007

2.2 Supply

Some interesting attributes of the chosen cars such as price, fuel consumption and CO2 emission are missing in the car register. In order to impute this information as well as defining the choice set we use an additional data source provided by a consultant company, Ynnor, containing detailed information about all cars available on the Swedish market on the make/model/version level of detail. For 2006 and 2007 there are 2320 and 2679 cars available, respectively, corresponding to 45 different makes. Table 2 shows the share of available cars by fuel type in 2006 and 2007. For petrol, diesel and electric-

hybrid the number and share corresponding to the clean car definition are also reported (these are also included in the total for each fuel). Since there is an increase in the number of cars available on the market from 2006 to 2007 one can note that the number of petrol cars increase but the share decrease in favor of diesel cars. Moreover, there are 14 more clean diesel cars introduced in the market and 28 ethanol cars.

Fuel Type	**2006**		**2007**	
	Number	Share	Number	Share
Petrol	1579	68.0	1748	65.2
clean petrol	24	1.9	31	1.1
Diesel	703	30.3	863	32.7
clean diesel	1	0.04	15	0.6
El-hybrid	11	0.5	13	0.5
clean El	5	0.2	6	0.2
Ethanol	16	0.7	44	1.6
Gas	11	0.5	11	0.4

Table 2: Cars available in the market by fuel type for 2006 and 2007

2.3 Data matching

As described above we have on one hand the demand data from the car register where the characteristics of the chosen cars are crudely defined. On the other hand, the alternatives in the supply data are defined at a very detailed level. When matching these two data sources to impute missing information several alternatives may correspond to the same observation. Therefore we aggregate (or group) alternatives available in supply based on vehicle-year/make/model/fuel-type since these characteristics can also be observed from demand data. Figure 1(a) shows as an example that an observed choice from demand (*i.e.* Volvo-S40-diesel) can correspond to different versions of Volvo-S40 running on diesel. Therefore, these versions are grouped as an aggregate alternative to match with observed choice from demand. The resulting data set contains 103,155 & 112964 observations [3] and 398 & 485 aggregated alternatives in 2006 and 2007, respectively.

[3]The difference in numbers here with that of registered new cars owned privately, presented in Section 2.2 is due to the ambiguous way of encoding of make and model in registry data also different way of coding of fuel type in two sources of available data which do not always match.

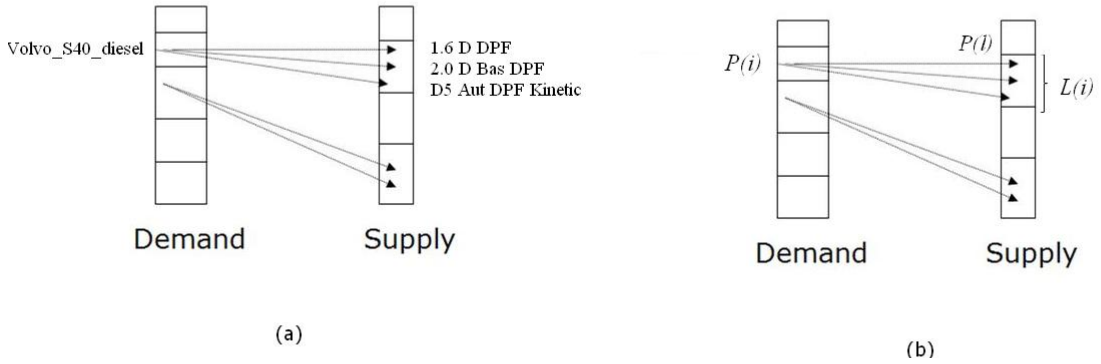

Figure 1: Matching demand with supply

Since price is highly variant in the supply, its coefficient of variation (CV), defined as the ratio of standard deviation to mean, is presented in Figure 2. CV shows the dispersion of the price of disaggregate alternatives corresponding to an aggregate alternative. $CV = 0$ indicates the homogeneity of price over disaggregate alternatives in the respective set. As it can be seen, only 15% of sets (aggregate alternatives) have only $CV = 0$ and those are corresponding to the sets including only one version(disaggregate-alternative) and for the rest the prices for the versions are variant. This graph shows that the assumption that versions (sub-alternatives) are homogenous is not plausible.

3 Theory and model specification

In this section we present three different logit models that all have a linear-in-parameters specification of the deterministic utility function at the disaggregate level. We denote an aggregate alternative by i and a disaggregate by l. P_i, the probability of an observation, is formulated as follows:

$$P_i = \sum_{l \in L_i} P_l$$

where L_i is the set of disaggregate alternatives corresponding to aggregate alternative i and P_l is the probability of a disaggregate alternative. To calculate deterministic utility of aggregate alternative, V_i, we use the approach presented by Ben-Akiva and

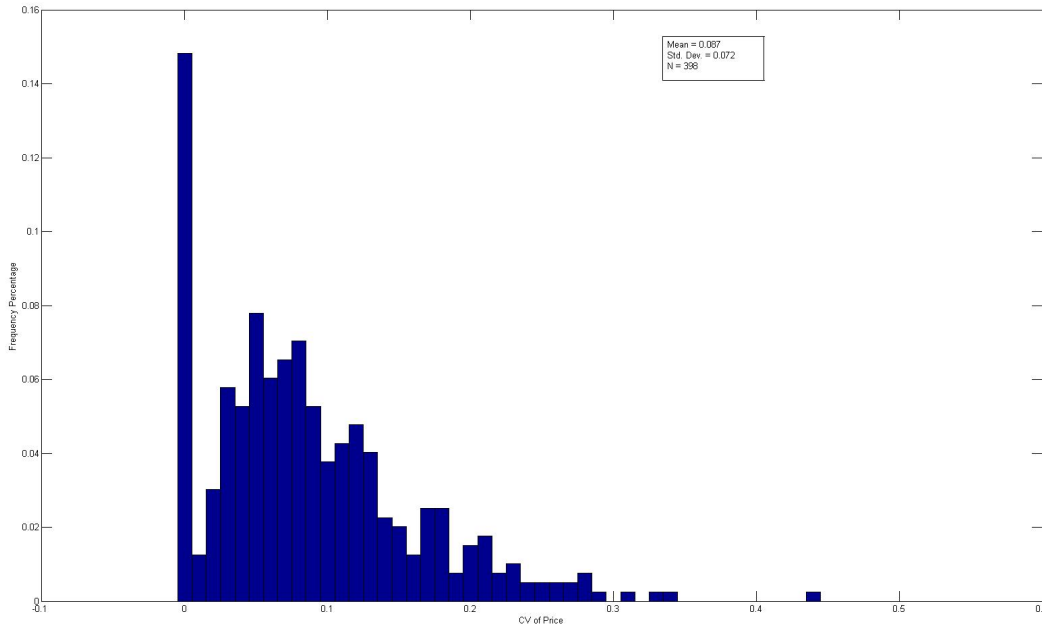

Figure 2: Coefficient of variation of price for disaggregate alternatives corresponding to an aggregate alternative in supply 2006

Lerman, 1985 to aggregate alternatives:

$$V_i = \bar{V}_i + \mu \ln m_i + \mu \ln[\frac{1}{m_i} \sum_{l \in L_i} \exp^{\frac{(V_l - \bar{V}_i)}{\mu}}] \tag{1}$$

where,

$\bar{V}_i = \frac{1}{m_i} \sum_{l \in L_i} V_l$, average of disaggregate alternatives' deterministic utilities,
m_i, number of disaggregate alternatives in the L_i, and,
μ, nesting parameter.
The second term of the formula is the measure for the size and the third term is the measure for the heterogeneity. In the literature, usually the first term is included which the attributes are averaged over the disaggregate alternatives without any correction for the aggregation. Mabit (2011) includes the the size in the model formulation and adds alternative specific constants which can capture the effect of heterogeneity. The first

model is multinomial logit (MNL) in which we only include the measure for the size without considering heterogeneity:

$$V_i = \bar{V}_i + \mu \ln m_i \tag{2}$$

In the second model we include the measure for the heterogeneity as well as size. It should be noted that the models is MNL, hence $\mu = 1$:

$$V_i = \bar{V}_i + \ln m_i + \ln[\sum_{l \in L_i} \exp^{(V_l - \bar{V}_i)}] \tag{3}$$

In the third model we consider the nesting structure in which each aggregate alternative of make/model/fuel-type will be a nest, as shown in Figure 3 and we estimate μ and the equation will be the same as equation 1. Table 3 shows the car attributes used for the modeling.

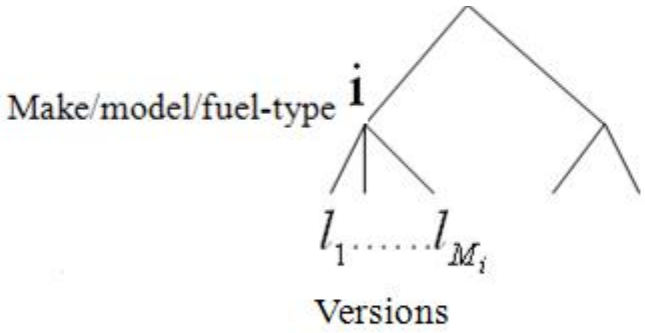

Figure 3: Nesting structure

We introduce two series of models; the first series include brand specific constants and the second ones include origin specific constants which are specified based the origin country of the cars' manufacturers. We are aware of the issue with brands changing owners and hence origin country. The aim has been to categorize each brand with the country it is in general associated with. The results are presented in following section.

4 Estimation results and forecasts

Due to the large amount of data, a tailor made MATLAB code is used for the estimation, the results of the four models are presented in Tables 4 and 5.

Table 3: Description of cars attributes

Attribute	**Description**
Brand	Dummies for brands
Origin	Dummies for origins
Cabriolet	Dummy for cabriolets
Copue	Dummy for Coupes
Hatch-backs	Dummy for Hatch-backs
Minibuss	Dummy for minibuss
Minivan	Dummy for minivan
MPV	Dummy for MPVs (multi purpose vehicles)
Sedan	Dummy for Hatch -backs
SUV	Dummy for Hatch -backs
GAS	Dummy for gas cars
E85	Dummy for ethanol-hybrid cars
El	Dummy for electrical-hybrid cars
Diesel	Dummy for diesel cars
AFV	Dummy for alternative fuel cars
Price	Purchase Price in 1,000,000 SEK
Tax	Vehicle circulation tax in 1000 SEK[1]
Tank-volume	in liter
Weight/power	kg/kw/10
Lux	Dummy for luxury car (purchase price over 800,000 SEK)
Clean	Dummy for clean cars
m	Number of elemental alternatives

[1] vehicle circulation tax= base tax(360 SEK) + CO2 component (20 SEK/gr of CO2 emission for conventional, 10 SEK/gr of CO2 emission for alternative fuels. For diesel cars, tax of conventional car tax is multiplied by 3.15. 1 USD is approx 7.2 SEK in June, 2012.

without heterogeneity-MNL			with heterogeneity-MNL		with heterogeneity-NL	
Est. Parameter	**value**	**t-value**	**value**	**t-value**	**value**	**t-value**
AMERICAN	-2.16	(-136)	-2.21	(-63.9)	-2.11	(-124)
BRITISH	-2.84	(-51.7)	-2.85	(-51.9)	-2.9	(-52.7)
CZECH	-0.863	(-51.7)	-0.86	(-30.1)	-0.744	(-43.2)
FRENCH	-1.37	(-104)	-1.42	(-48.7)	-1.38	(-104)
GERMAN	-1.38	(-122)	-1.42	(-106)	-1.39	(-126)
ITALIAN	-3.67	(-63.5)	-3.62	(-55.9)	-3.65	(-63.1)
JAPANESE	-1.33	(-112)	-1.39	(-60.1)	-1.38	(-111)
KOREAN	-1.63	(-82)	-1.56	(-44.1)	-1.65	(-81.3)
SPANISH	-2.92	(-78.8)	-2.86	(-52.2)	-2.83	(-75.4)
SWEDISH	0	fixed	0	fixed	0	fixed
Cabriolet	-1.06	(-33.6)	-1.13	(-25.8)	-1.07	(-31.1)
Coupe	-1.24	(-32.8)	-1.56	(-31.7)	-1.5	(-31.2)
Hatch	-0.0659	(-5.07)	-0.0763	(-4.35)	-0.0615	(-4.83)
Minibuss	-2.79	(-29.3)	-2.43	(-23.1)	-2.53	(-29.5)
Minivan	-1.1	(-43.9)	-1.18	(-40.4)	-1.27	(-43.6)
MPV	-1.99	(-31.9)	-1.9	(-28)	-2.02	(-32.4)
Sedan	-1.35	(-93.1)	-1.41	(-85.2)	-1.47	(-86.1)
SUV	-0.571	(-31.6)	-0.482	(-25.7)	-0.444	(-24.8)
Gas	0.399	(2.81)	0.482	(0.691)	0.472	(2.48)
E85	3.53	(40)	3.64	(11.3)	3.56	(31.3)
El	0.588	(7.21)	0.636	(1.55)	0.587	(4.96)
Diesel	-2.1	(-73.8)	-2.15	(-59.7)	-2.22	(-76.8)
Price	-0.798	(-7.65)	-0.595	(-1.24)	-0.777	(-5.96)
Price*Clean	-5.02	(-14.1)	-5.91	(-6.17)	-5.39	(-13.7)
Tax	-1.19	(-57.2)	-1.26	(-25.4)	-1.36	(-58.9)
Tax*Diesel	1	(64)	1.07	(33.6)	1.13	(65.5)
Tax*AFV	-2.13	(-30.2)	-2.19	(-11.3)	-2.27	(-27)
Tank	3.32	(39.6)	3.29	(31.4)	3.78	(46.5)
Weight/Power	-0.767	(-25.9)	-0.623	(-7.67)	-0.773	(-21)
Lux	0.125	(1.47)	0.361	(1.31)	0.354	(3.89)
Clean	0.643	(12.8)	0.887	(7.06)	0.734	(14)
Log(m)	0.939	(193)	-	-	-	-
μ	-	-	1	fixed	0.859	(165)
Final Log-likelihood	-516,107.43		-515,882.75		-515,518.00	
Null Log-likelihood	-617532.45		-617532.45		-617532.45	
$\bar{\rho}^2$	0.164		0.164		0.165	
AIC	1,032,276.86		1,031,825.55		1,031,098.00	

Table 4: Estimation results with origin specific constants

without heterogeneity-MNL			with heterogeneity-MNL		with heterogeneity-NL	
Est. Parameter	**value**	**t-value**	**value**	**t-value**	**value**	**t-value**
ALFAROMEO	-3.46	(-34.1)	-3.63	(-35.5)	-3.66	(-35.9)
AUDI	-1.53	(-83.7)	-1.61	(-86.2)	-1.6	(-87.1)
BENTELY	0.453	(1.23)	0.981	(3.08)	0.869	(2.92)
BMW	-1.17	(-57.7)	-1.27	(-64.7)	-1.26	(-64.3)
CADILLAC	-2.94	(-15.7)	-2.78	(-15.1)	-2.85	(-15.6)
CHEVROLEET	-2.72	(-56.7)	-2.72	(-56.7)	-2.81	(-59.2)
CHRYSLER	-1.42	(-30)	-1.44	(-30.2)	-1.48	(-32)
CITROEN	-1.92	(-84.5)	-2.07	(-91)	-2.04	(-93)
DODDGE	-3.4	(-27.7)	-3.38	(-27.3)	-3.46	(-28.3)
FERRARI	0.189	(0.597)	0.615	(2.01)	0.514	(1.83)
FIAT	-4.07	(-54.3)	-4.18	(-55.4)	-4.15	(-55.4)
FORD	-2.36	(-126)	-2.52	(-145)	-2.37	(-128)
HONDA	-1.2	(-48.8)	-1.23	(-50.9)	-1.31	(-53.8)
HYUNDAI	-1.48	(-63.9)	-1.52	(-65.9)	-1.57	(-70)
JAGUAR	-3	(-32.1)	-2.94	(-31.2)	-2.98	(-32)
JEEP	-0.948	(-12.3)	-0.851	(-10.7)	-0.961	(-12.6)
KIA	-2.59	(-70.9)	-2.65	(-72.7)	-2.68	(-74.4)
LAMBORGINI	-0.0609	(-0.185)	0.488	(1.93)	0.374	(1.48)
LAND ROVER	-2.98	(-25.1)	-3.1	(-26)	-3.05	(-25.7)
LEXUS	-1.76	(-25.6)	-1.7	(-23.9)	-1.81	(-25.2)
LOTOUS	-4.94	(-8.55)	-4.6	(-7.52)	-4.75	(-8.6)
MASERATI	-0.885	(-1.95)	-0.53	(-2.18)	-0.611	(-1.89)
MAZDA	-1.81	(-73.3)	-1.95	(-79.2)	-1.9	(-78.2)
MERCEDES	-1.75	(-75.1)	-1.84	(-79.5)	-1.84	(-79.5)
MINI	-2.93	(-34.1)	-3	(-34.6)	-2.97	(-34.7)
MITSUBISHI	-1.72	(-65.7)	-1.79	(-68.1)	-1.78	(-69.7)
MORGAN	-6.09	(-12.2)	-5.74	(-11.2)	-5.89	(-12)
NISSAN	-2.65	(-93.6)	-2.79	(-98.5)	-2.72	(-97.8)
OPEL	-1.72	(-83)	-1.86	(-92.4)	-1.79	(-90.8)
PEUGEOT	-1.29	(-69.2)	-1.51	(-88.4)	-1.39	(-82)
PORSCHE	-1.74	(-24.9)	-1.52	(-21)	-1.51	(-21.3)
RENAULT	-1.88	(-95.3)	-1.99	(-105)	-1.9	(-101)
SAAB	-0.499	(-29)	-0.671	(-41.2)	-0.615	(-37.4)
SEAT	-3.05	(-80.2)	-3.1	(-81.1)	-3.06	(-81.6)
SKODA	-1.06	(-57.5)	-1.15	(-66.4)	-1.03	(-58.2)
SMART	-7.12	(-23.5)	-7.29	(-24)	-7.06	(-23.3)
SSANGYU	-4.57	(-7.91)	-4.64	(-8.36)	-4.63	(-8.45)
SUBARU	-1.13	(-38.5)	-1.25	(-42)	-1.29	(-44.5)
SUZUKI	-2.85	(-67.4)	-2.91	(-68.6)	-2.91	(-69.7)
TOYOTA	-0.93	(-57.9)	-1.13	(-73.2)	-1.05	(-68)
VOLKSWAGEN	-1.26	(-71.1)	-1.47	(-82.6)	-1.4	(-79.9)
VOLVO	0	FIXED	0	FIXED	0	FIXED
Cabriolet	-0.212	(-6.79)	-0.36	(-10.4)	-0.486	(-13.7)
Coupe	-1.19	(-28)	-1.07	(-21)	-1.12	(-21.5)
Hatch	-0.237	(-15.9)	-0.214	(-14.7)	-0.19	(-13.2)
Minibuss	-2.64	(-29.7)	-2.44	(-28.2)	-2.54	(-29.3)
Minivan	-1.16	(-43.4)	-1.17	(-40)	-1.24	(-42.1)
MPV	-2	(-31.8)	-1.95	(-31.1)	-2.03	(-32.5)
Sedan	-1.34	(-90.4)	-1.44	(-85.9)	-1.48	(-86.8)
SUV	-0.253	(-12.7)	-0.222	(-11.2)	-0.201	(-10.3)
Gas	-0.836	(-5.68)	-0.758	(-5.73)	-0.745	(-4.97)
E85	1.72	(18)	1.69	(18.6)	1.74	(18.5)
El	-0.0599	(-0.696)	-0.0599	(-0.747)	-0.0147	(-0.18)
Diesel	-2.41	(-78)	-2.42	(-77)	-2.45	(-79.2)
Price	-2.02	(-13.9)	-2.21	(-17.7)	-2.04	(-16.8)
Price*Clean	-6.48	(-15.8)	-7.45	(-17.4)	-7.33	(-17.3)
Tax	-1.45	(-55.3)	-1.51	(-56.6)	-1.55	(-61.9)
Tax*Diesel	1.24	(63.8)	1.29	(64.1)	1.31	(68.8)
Tax*AFV	-0.584	(-7.12)	-0.491	(-5.89)	-0.566	(-6.89)
Tank	3.78	(38.5)	4.01	(45.4)	4.1	(47.3)
Weight/Power	-1.07	(-30.6)	-0.863	(-24.6)	-1.02	(-27.8)
Lux	1.47	(14)	1.48	(18.4)	1.33	(17.8)
Clean	0.805	(15)	1.01	(18.5)	0.85	(15.7)
Log(m)	0.954	(167)	-	-	-	-
μ			1	FIXED	0.881	(142)
Final Log-likelihood	-506,799.20		-506,484.93		-506,302.05	
Null Log-likelihood	-617532.45		-617532.45		-617532.45	
$\bar{\rho}^2$	0.179		0.179		0.180	
AIC	1,013,724.4		1,013,093.86		1,012,730.10	

Table 5: Estimation results for the MNL model with brand specific constants

Volvo station-wagon petrol is defined as the base case and respective dummies have coefficients fixed to zero as well as Swedish as origin constant, correspondingly. All other parameters associated with brand, origin and body type are negative showing the preference over Volvo station-wagon, all other attributes equal to zero. All the estimated parameters have their expected signs and are highly significant; except constants related to some luxury brands such as Ferrari and Lamborghini, dummy for electric-hybrid cars and luxury cars in some models. Price is insignificant only in MNL with heterogeneity and origin specific constant. The price parameter is negative as expected, both for clean cars and non-clean ones. The estimated parameters show a larger price sensitivity for clean cars. The tax parameter is always negative and significant. However, the combined effect of tax with diesel dummy is positive. Yet, the sum of this parameter is negative for diesel cars but the value is very small showing individuals are less sensitive to tax, when considering diesel cars. The interacted effect of tax for alternative fuel vehicles (AFV) is negative showing even more sensitivity to tax when considering AFV cars compared to conventional ones. This explains the fact that lower taxes for AFV cars might be good incentives to buying these cars. The attempts to include fuel cost in the model, did not lead to a better model fit. Clean cars parameter is also always positive and significant, showing preference over these cars in Sweden. Tank volume and weight/power ratio parameters have their expected positive and negative signs, respectively and are significant. The parameter for the luxury cars are also positive and significant in most models.

ρ^2 is the likelihood ratio index, which is defined as: $\rho^2 = 1 - \frac{LL}{LL_0}$, where LL is final log-likelihood and LL_0 is null log-likelihood. ρ^2 provides a relative value between 0 to 1 indicating the improvement from the null model. Adjusted ρ^2, $\bar{\rho}^2$, is used to penalize the addition of variables to the model compared to the number of observations. Generally, as more independent variables added to the model, ρ^2 will increase, therefore, $\bar{\rho}^2$ should be corrected for the number of variables in the model. However, in our study due to the large number of observation is very large, there is not a negligible difference between ρ^2 and $\bar{\rho}^2$. As can be seen from Tables 4 and 5, across the same constants (brand or origin), $\bar{\rho}^2$ is the same form model 1 (MNL without heterogeneity) to model 3 (NL). However, comparing models over the constants, models with brand specific constants

having higher value for $\bar{\rho}^2$ and therefore are better estimated models. We, also compare these models based on other model selection methods such as AIC or BIC[4]. AIC and BIC, also, penalize extra parameters, and BIC does penalize more heavily while taking number of observations into account. The lower AIC and BIC indicate better models. As mentioned earlier, due to the large number of observations, in this study, BIC can not be a relevant criterion. Considering value of AIC, it can be seen that in both two series of models moving from model 1 to model 3 result in a better model. Also, AIC decreases moving from the models with origin specific constants to the models with brand specific ones, showing better explanatory power of these models.

The parameter for log(m) is highly significant stating that the fact that the aggregate alternatives contain sub-disaggregate alternatives is really important. Yet, this model is not accounting for heterogeneity as in the other two models. μ in NL model is highly significant and significantly different from one ($t_{(\mu=1)} = 19.19$ at the level of 0.05) indicating that nesting structure is important and heterogeneity of sub-alternatives matters in modeling.

We continue this section by presenting prediction results obtained by applying the six estimated models to the actual market 2007 considering the clean car purchase subsidy of 10,000 S. Table 6 reports the share of different brands and Table 7 the share of ethanol cars. It would be more interesting to see the share of clean cars but spotting actual choice of clean cars from the demand data is ambiguous since emissions and fuel consumptions are missing from this database, therefore we present the prediction of ethanol cars to avoid uncertainty in results that is caused by clean cars.

Referring to the Table 6, all models under-predict share of Volvo and over-predict shares of Saab and BMW. However, Saab and Volvo have less difference in models with brand specific constants suggesting that Volvo and Saab constants should be differentiated in models with origin specific constants as well. BMW's share is highly over-predicted in models with the brand specific constants, it could be due to the fact the number of versions introduced to market has increased 133% from 2006 to 2007. However, BMW share is lower in models with origin specific constants that have more

[4]The AIC is $-2LL + 2K$, where LL is the value of log-likelihood and K is the number of parameters. The BIC is $-2LL + \log(N)K$, where N is sample size

Brand	**Actual market**	**Predicted market**					
		Origin without heterogeneity MNL	Origin with heterogeneity MNL	Origin with heterogeneity NL	Brand without heterogeneity MNL	Brand with heterogeneity MNL	Brand with heterogeneity NL
ALFAROMEO	0.03	0.06	0.05	0.06	0.09	0.08	0.08
AUDI	3.76	3.81	3.49	3.65	3.63	3.58	3.64
BENTLEY	0	0	0	0	0.01	0.01	0.01
BMW	3.33	5.87	6.51	6.11	8.38	9.23	8.7
CADILLAC	0.02	0.11	0.08	0.09	0.05	0.05	0.05
CHEVROLET	0.21	0.48	0.4	0.52	0.36	0.35	0.35
CHRYSLER	0.24	0.2	0.18	0.21	0.44	0.43	0.44
CITROEN	5.81	5.84	5.62	6	3.73	3.69	3.7
DODGE	0.18	0.22	0.19	0.24	0.1	0.1	0.1
FERRARI	0.02	0	0	0	0.01	0.01	0.01
FIAT	0.71	0.18	0.18	0.19	0.14	0.14	0.14
FORD	5.87	5.06	5.26	5.1	4.79	4.87	4.93
HONDA	3.12	1.81	1.71	1.79	2.58	2.66	2.52
HUMMER	0	2.59	2.57	2.53	3.7	3.75	3.68
HYUNDAI	2.2	0.09	0.09	0.09	0.09	0.09	0.09
JAGUAR	0.06	0.19	0.16	0.2	0.84	0.88	0.88
JEEP	0.12	2.46	2.54	2.46	1.17	1.19	1.16
KIA	3	0	0	0	0.01	0.01	0.01
LANDROVER	0.18	0.14	0.14	0.11	0.12	0.11	0.1
LEXUS	0.22	1.07	0.91	0.99	0.34	0.32	0.33
LOTUS	0	0.01	0.01	0.01	0	0	0
MASERATI	0.01	0	0	0	0.04	0.04	0.04
MAZDA	1.77	3.14	3.13	3.05	2.44	2.39	2.4
MERCEDES	1.92	3.06	3.02	3.07	2.33	2.29	2.32
MINI	0.29	0.3	0.34	0.31	0.42	0.43	0.41
MITSUBISHI	1.12	1.07	1.01	1.1	1.14	1.11	1.12
MORGAN	0	0.04	0.03	0.04	0	0	0
NISSAN	1.79	5.19	5.21	5.18	1.86	1.83	1.86
OPEL	3.73	4.34	4.2	4.19	3.68	3.59	3.66
PEUGEOT	8.13	6.76	6.75	6.69	8.95	8.54	8.8
PORSCHE	0.07	0.17	0.14	0.15	0.2	0.2	0.21
RENAULT	3.31	5.24	4.86	5.13	3.93	3.77	4
SAAB	4.68	7.81	9.82	9.04	7.12	7.37	6.97
SEAT	0.54	0.68	0.68	0.7	0.67	0.67	0.69
SKODA	5.59	5.88	6.2	6.08	5.68	6.08	6.01
SMART	0.05	1.11	1.22	1.18	0.01	0.01	0.01
SSANGYONG	0.08	0.22	0.23	0.23	0.02	0.02	0.02
SUBARU	1.11	1.44	1.44	1.5	2.05	2.07	2.03
SUZUKI	1.07	1.28	1.16	1.23	0.38	0.38	0.39
TOYOTA	8.24	4.11	4.16	4.13	7.29	6.99	7.2
VOLKSWAGEN	7.26	3.94	3.76	3.82	5.06	4.54	4.68
VOLVO	19.45	14.01	12.55	12.86	16.17	16.16	16.28
RMSE		1.72	1.97	1.89	1.43	1.54	1.46

Table 6: Prediction results vs. actual for different brands

aggregate constants. These results show very high sensitivity of the prediction of these models to the number of different versions introduced to the market, i.e. supply side. Unlike estimation, root mean square error (RMSE) is improving (getting smaller) when heterogeneity does not included in the models. The results also indicate that the models with brand specific constants show smaller error in prediction of brand shares.

Brand	**Actual market (E85 share)**	**Predicted market (E85 share)**					
		Origin without heterogeneity	Origin with heterogeneity MNL	Origin with heterogeneity NL	Brand without heterogeneity	Brand with heterogeneity MNL	Brand with heterogeneity NL
CADILLAC	0	0.02	0.01	0.01	0.01	0.01	0.01
CITROEN	0.09	1.35	1.34	1.45	0.66	0.68	0.69
FORD	3	1.79	1.88	1.72	1.55	1.57	1.5
PEUGEOT	0.28	1.17	1.18	1.17	1.49	1.44	1.48
RENAULT	0.4	1.08	1.05	1.13	0.72	0.76	0.81
SAAB	3	3.29	4.5	4.08	3.74	4	3.66
SEAT	0	0.06	0.06	0.07	0.06	0.07	0.07
VOLVO	2.4	2.74	2.36	2.7	3.4	3.45	3.66
All	9.23	11.5	12.38	12.33	11.63	11.98	11.88
RMSE	-	0.67	0.79	0.78	0.74	0.77	0.79

Table 7: Prediction results vs. actual for E85 cars

	Actual market (E85 share)	**Predicted market (E85 share)**					
		Origin without heterogeneity	Origin with heterogeneity MNL	Origin with heterogeneity NL	Brand without heterogeneity	Brand with heterogeneity MNL	Brand with heterogeneity NL
All	9.23	10.94	11.72	11.70	10.82	11.04	10.97

Table 8: Prediction for E85 market share without policy

As it can be seen in Table 7, all the models over-predict the share of the ethanol cars in 2007. Similar to the brand share prediction, presented in Table 6 root mean square error (RMSE) is improving (getting smaller) when heterogeneity is not included in the models. However, here, the model with origin specific constants give the smaller error. One can question whether the reason to this over-prediction of ethanol cars could be that fact that models usually predict the long-run effect after learning the new policies

which in our case is purchase subsidy for clean cars. Therefore, we provide prediction without considering implemented policy in 2007 and the results are presented in 8. As can be seen all the models still over-predict the share of ethanol cars while, but the values are smaller compared to that of prediction with policy. This low sensitivity to purchase subsidy can be explained by very small value for coefficient of price in all models. Hence, the problem of over-prediction refers to the structure of the models. It should be noted that MNL model without heterogeneity and origin specific constant gives the closest predicted value for the share of total ethanol cars considering purchase policy in 2007 while this value is smallest in the same model but with brand specific constant when no policy is considered.

5 Conclusion and future work

As it can be seen from the results, number of considering disaggregate alternatives forming the aggregate alternatives and their heterogeneity will significantly improve the fitness of the models. Also, they show that aggregation level and nesting structure are defined appropriately. However, this will affect prediction adversely. These findings are in line with Mabit, 2011's study about importance of considering supply in the modeling. Comparing the estimation results with prediction ones, with or without policy, raise the question that whether or not are the 'best' model for estimation necessarily the 'best' ones for prediction as well? This question plays an important role in car fleet modeling since the main objective of these models is to build a decision support tool for predicting the changes under implemented. Therefore, what is more important here is the prediction accuracy of models while what is done in literature is to find the best estimated model and use it to project results in the future.

We also, have compared MNL with a specific nesting structure which puts each aggregated alternative in one nest. It should be mentioned here that there are several assumptions and restrictions existing in the definition of MNL models including independence of irrelevance alternatives (IIA). Here, having nesting parameter, μ between 0 and 1, indicates the importance of nesting structure on this data and violation of IIA assumption. Generally, to avoid the restrictions imposed by MNL models (specifically

IIA), nested structure is employed in the literature such as influential studies of Berkovec and Rust, 1985 and Hensher and Plastrier, 1985. Also, the car type choice model that developed by (see Transek, 2006) for Sweden has a nesting structure. Therefore, in order to estimate a better model and achieve likely more accurate prediction results ,nesting structure should be tested on this data as well. However, the different nesting structures should be investigated for the data used in this study, and other ones might be more relevant here.

Acknowledgments

We are grateful to Yu Shen who started the work on merging the two data sources during his Master's thesis. We would also like to thank Anders Karlström and Staffan Algers for their valuable comments and also to thank Mohammad-Reza Yahya for assisting us with the registry data, and we are thankful to YNNOR for providing supply data.

References

Ben-Akiva, M. and Lerman, S. (1985). *Discrete choice analysis: theory and application to travel demand*, Vol. 9, MIT press.

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

Choo, S. and Mokhtarian, P. L. (2004). What type of vehicle do peaple drive? the role of attitude and lifestyle in influencing vehicle type choice, *Transportation Research Part A* **38**(3): 201–222.

De Jong, G., Fox, J., Pieters, M., Daly, A. J. and Smit, R. (2004). A comparison of car ownership models, *Transport Reviews* **24**(4): 379–408.

Hensher, D. A. and Plastrier, V. L. (1985). Towards a dynamic discrete-choice model of household automobile fleet size and composition, *Transportation Research Part B: Methodological* **19**(6): 481–495.

Hugosson, M. and Algers, S. (2012). Accelerated Introduction of Clean Cars in Sweden, *in* T. I. Zachariadis (ed.), *Cars and Carbon SE - 11*, Springer Netherlands, pp. 247–268.

Lave, C. A. and Train, K. (1979). A disaggregate model of auto-type choice, *Transportation Research Part A: General* **13**(1): 1–9.

Mabit, S. L. (2011). Vehicle type choice and differentiated registration taxes, *European Transport Conference* .

Mannering, F. L. (1983). An econometric analysis of vihicle use in multivehicle households, *Transportation Research Part A: General* **17**(3): 183–189.

Page, M., Whelan, G. and Daly, A. (2000). Modelling The Factors which Influence New Car Purchasing, *European Transport Conference 2001 proceedings*, Association for European Transport, Homerton College, Cambridge.

Potoglou, D. and Kanaroglou, P. S. (2008). Disaggregate Demand Analyses for Conventional and Alternative Fueled Automobiles: A Review, *International Journal of Sustainable Transportation* **2**(4): 234–259.

Transek (2006). Bilparksmodell (car fleet model), *Technical report*, Transek AB, Sweden.