Prediction-driven approaches to discrete choice models with application to forecasting car type demand

SHIVA HABIBI

Doctoral Thesis in Transport Science
Stockholm, Sweden 2016

KTH Royal Institute of Technology
School of Architecture and the Built Environment
Department of Transport Science
Division of Systems Analysis and Economics
SE-100 44 Stockholm
SWEDEN

Prediction-driven approaches to discrete choice models with application to forecasting car type demand

TRITA-TSC-PHD 16-002
ISBN 978-91-87353-82-6

KTH Royal Institute of Technology
School of Architecture and the Built Environment
Department of Transport Science
Division of Systems Analysis and Economics
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i transportvetenskap onsdagen den 3 februari 2016, klockan 9:00 i Kollegiesalen Brinellvägen 8.

Tryck: Universitetsservice US-AB

تقدیم به پدر و مادرم

To my parents
who did not raise me as a “girl”

"Still, it is an error to argue in front of your data. You can find yourself insensibly twisting them round to suit your theories"

— Sherlock Holmes Quote, *A Scandal in Bohemia*

Abstract

Models that can predict consumer choices are essential technical support for decision makers in many contexts. The focus of this thesis is to address prediction problems in discrete choice models and to develop methods to increase the predictive power of these models with application to car type choice. In this thesis we challenge the common practice of prediction that is using statistical inference to estimate and select the 'best' model and project the results to a future situation. We show that while the inference approaches are powerful explanatory tools in validating the existing theories, their restrictive theory-driven assumptions make them not tailor-made for predictions. We further explore how modeling considerations for inference and prediction are different.

Different papers of this thesis present various aspects of the prediction problem and suggest approaches and solutions to each of them.

In paper 1, the problem of aggregation over alternatives, and its effects on both estimation and prediction, is discussed. The focus of paper 2 is the model selection for the purpose of improving the predictive power of discrete choice models. In paper 3, the problem of consistency when using disaggregate logit models for an aggregate prediction question is discussed, and a model combination is proposed as tool. In paper 4, an updated version of the Swedish car fleet model is applied to assess a Bonus-Malus policy package. Finally, in the last paper, we present the real world applications of the Swedish car fleet model where the sensitivity of logit models to the specification of choice set affects prediction accuracy.

Sammanfattning

Modeller som kan användas för att förutsäga konsumenters val på en marknad är viktiga verktyg för beslutsfattare i olika sammanhang. Denna avhandling fokuserar på metoder för att förbättra diskret valmodeller i dess användning som prognosverktyg med tillämpning på för konsumentens val av biltyp. I avhandlingen utmanar vi det gängse bruket att använda statistisk inferens för val av bästa modell när syftet är att använda modellen för prognostisering. Vi visar att inferens-drivna ansatser är kraftfulla verktyg för validera teorier, men att de samtidigt sådana ansatser inte ger de bästa skräddarsydda metoderna för prognoser. Vi påvisar därför hur avvägningarna skiljer sig beroende på om syftet är inferens eller prognos.

Avhandlingens olika uppsatser behandlar olika aspekter av prognosproblement och påvisar ansatser och lösningar för att adressera dessa.

I det första papperet diskuterar vi aggregering över alternativ, och dess effekt såväl på estimering som prognostisering. I det andra papperet fokuserar vi på val av bästa diskret valmodell, med syfte att hitta den modell som har bäst prognosvärde. I det tredje papperet diskuterar vi konsistens av disaggregerade logitmodeller som används för aggregerade prognoser, och vi föreslår modelkombination (model combination) som ett verktyg i detta sammanhang. I papper 4 analyserar vi hur specifikationen av valalternativen påverkar prognosförmågan hos en logitmodell för svenska bilflottan. I papper 4 använder vi en uppdaterad version av en modell för den svenska bilflottan för att analysera ett Bonus-Malus-system. Slutligen, i det sista papperet analyserar vi hur specifikationen av valalternativen påverkar prognosförmågan hos en logitmodell för svenska bilflottan.

Acknowledgment

I would like to extend my appreciation to the dozens of people whose help and support contributed to this thesis in many different ways.

Special mention goes to my supervisors Anders Karlström, Marcus Sundberg, Muriel Hugosson, Staffan Algers, and Emma Frejinger. I had an opportunity to learn from diverse expertise. They each taught me how to think about problems from different perspectives.

Profound gratitude goes to Oded Cats and Karin Brundell-Freij for their comments on the earlier version of the thesis, licentiate thesis, and to Joel Franklin for the final review. Towards the end, Oded also played the role of a shadow supervisor in giving advice and support, and not to mention, he has always played the role of a dear friend. To Stef Proost, not only because of his comments on parts of the work but also his care, his advice and support, especially at the final stages when I needed it the most. To Haris N. Koutsopoulos, my master thesis supervisor, who even during his short visits to Stockholm made sure that I will see "the light at the end of the tunnel". And to David Brownstone and Andrew Daly, for the helpful and inspiring discussions and their genuine intentions to help Ph.D. students.

Grateful acknowledgment goes to the funding source that made my Ph.D. work possible. I was funded by the Centre for Transport Studies, Stockholm. Many appreciations to its former and current directors, Jonas Eliasson and Maria Börjesson.

Many thanks to all the folks at the Department of Transport Science for inspiring discussions and supporting company. Especially to Lars-Göran Mattsson, my very first contact person in Sweden, to the best officemates ever, Christer Persson and Maria Nordström, to Yusak Susilo, Eva Pettersson, Susanne Jarl, Bibbi Nissan, Wilco Burghout, and, Emma Engström. Special mention goes to the "Sex rummet", for all intense and vibrant discussions. Sitting with you guys enriched my life in many ways. Thanks, Jake, Masoud, Oskar, Per, Siamak, and Vivi. Thanks for the all the fun. Special thanks to Nina, My Swedish teacher, who not only made me take great leaps in my Swedish studies but for a while played a role of a mentor.

The endless gratitude goes to the friends who always incented me to strive towards my goals, without whom Ph.D. life or even life in Stockholm would not be imaginable. For all the laughter and stupid fights, for all the wine and stimulating discussions that we shared, THANK YOU. Thanks Alicja, Anne, Beata, Burcu, Cihan, Faria and Ramy. Thanks to my Iranian gang: Roya and Davood. Thanks to Tahmineh and Fati, my first hosts in Stockholm.

The wordless gratitude goes to my family. I am grateful to my parents for being wonderful role models for me. I am also grateful to my sister, Shole and my brothers, Reza and Amir, for the love and encouragement that will always exist, no matter what.

List of papers

I Habibi, Sh., Frejinger E.R. and Sundberg M. (2012) Aggregation of alternatives and its influence on prediction, *to be submitted.*

Extended version of paper presented at The 13th International Conference on Travel Behavior Research, Toronto, Ontario, Canada, July 15-20, 2012.

II Habibi, Sh., Sundberg M. and Karlström A. (2015) Prediction-driven approach to model selection using feature selection and nonrandom hold-out validation, *submitted to Journal of choice modeling.*

Extended version of the paper presented at 2nd Symposium of the European Association for Research in Transportation (hEART), Stockholm, 4-6th 2013, Stockholm, Sweden, and,
poster presented at The Transportation Research Board (TRB) 93rd Annual Meeting , The Transportation Research Board (TRB), Washington, D.C.

III Habibi, Sh., Sundberg M. and Karlström A. (2015) Model combination for capturing the inconsistency in the aggregate prediction, *to be submitted.*

IV Habibi, Sh., Hugosson, M., Sundbergh P., Algers, S. (2015) Evaluation of Bonus-Malus systems for reducing car fleet CO2 emissions in Sweden. *Submitted to Transport research part D, February, 2015.*

Extended version of the paper presented at 3rd Symposium of the European Association for Research in Transportation (hEART), 10-12th September 2014, Leeds, UK.

V Hugosson, M., Algers, S., Habibi, Sh., Sundbergh P. (2015) The Swedish Car Fleet Model, Evaluation of Recent Applications. *Revised and re-submitted to Transport Policy journal, August, 2015.*

Declaration of contribution

The idea of paper I was proposed by co-authors. Shiva Habibi was responsible for the analysis of results and writing.

The idea of paper II was proposed by Anders Karlström. Shiva Habibi was the main contributor in methodology implementation as well as responsible for the analysis of results and writing

The idea of paper III was proposed by Shiva Habibi. The methodology was developed in discussions between Shiva Habibi and Marcus Sundberg. The first author was the main contributor in methodology implementation and the analysis of results and writing.

Shiva Habibi was the main contributor in the analysis of results and writing of paper IV.

The idea of paper V was jointly proposed by authors. Shiva Habibi contributed in the analysis of results and writing.

Contents

Contents **xvii**

Part I: Introduction **3**

1 Overview and objectives **3**

2 Introduction to car fleet models: A review **5**
Swedish car fleet model: A summary 6

3 Data **9**
Data matching 10

4 Prediction problems **11**

5 Research outlines **15**
Paper I 15
Paper II 16
Paper III 17
Paper IV 18
Paper V 19

6 Conclusions and future work **21**

Bibliography **23**

Part II: Papers **29**

Introduction

1

Overview and objectives

The story of this thesis starts with the application of the Swedish car fleet model in analyzing a clean car promoting policy, in 2007, when the government introduced a subsidy of 1000 Euros for privately bought clean cars to advance the number of clean cars in the fleet. Transek (2006) estimated a model using the 2004 the Swedish register data and stated preference data, and predicted the share of clean cars after the introduction of the subsidy. The prediction severely underestimated the actual share of clean cars for several reasons related to the exogenously defined future alternatives scenario. However, the model predictions were accurate given the correct external inputs.

Based on this background, the objective of this thesis is to identify the prediction problems in the field of choice modeling and develop methods to improve the prediction performance of these models. The application of interest is the Swedish car fleet.

The common approach towards forecasting in the field of car type choice, like other areas in choice modeling, is to estimate the best possible logit model and use it to predict the quantity of interest. The typical assumption is that models with high explanatory power are necessarily own high predictive power. However, there is a distinction between prediction and estimation and different modeling considerations should be examined for either of them.

In the statistical inference view, an unknown true probability distribution exists from which the observed data has been generated, and the objective of inference is to derive the properties of this unknown distribution from the observed data. The inferred conclusions are expected to validate the causal hypotheses about theoretical ideas. In contrast to statistical inference, the objective of prediction is to making probability statements about the distribution of as yet unobserved data; yet, this unobserved data is generated from the same process that has generated the observed one (Geisser and Eddy; 1979). However, current theories do not contain new concepts that might exist in the newer richer datasets. Therefore, the theory-driven restrictions on models may prevent accurate predictions results. In inference, data is a tool to achieve the model as

close as possible to the true model whereas in prediction data is the entity of interest and a model is a tool to generate predictions as close to new data. Although in applied statistics the difference between prediction and estimation is observed, it is not explicitly recognized in statistical methodology, yet. In the statistics literature, discussions about explaining versus predicting are found in the context of model selection. Prediction has also been the focus of machine learning and statistical time series areas, yet as a separate issue but not in contrast to inference (Shmueli; 2010). In statistics and econometric, the area that focuses on prediction is time-series. However, there exists two different disciplines; in one approach model is smoothed on data and therefore, summarizes all the historical data (e.g. vector auto regressive model (VAR)) and the other approach use causal models that explain underlying causal theory in decision making (Shmueli; 2010; Sims; 1986). In the field of choice modeling, there exist few studies that focus on the predictive performance of these models (see e.g. Hensher and Ton; 2000; Huang et al.; 2012; McFadden; 1978; Train; 1979a).

Car fleet models are widely used for analyzing policies regarding environment and energy consumption (for example see the studies of Berkovec; 1985b; Boyd and Mellman; 1980; Goldberg; 1995; Greene; 1986; Hugosson and Algers; 2012; West; 2004). Hence, reliable predictions of effects and costs of different policies are of great importance to policy makers. The study of Lave and Train (1979) was the first to employ the multinomial logit (MNL) model to estimate a car type choice model with the purpose of evaluating transportation energy consumption policies. Since then there have been extensive studies on the employing logit models in the car fleet modeling. However, there exist a few studies that focus on the prediction performance of these studies or evaluate the predicted results of these models with the actual figures. The works of Brownstone et al. (1994) and Mohammadian and Miller (2002), can be mentioned as an examples of those few prediction focused studies. In the following sections, we give a short overview of the works done in the field followed by a summary of the current Swedish car fleet model.

2

Introduction to car fleet models: A review

A car *market* consists of different sectors including new cars, used cars, and scrapping. During each period (e.g. one year or six months) consumers purchase new cars from manufacturers and exchange used cars among themselves or with the scrapping sector. The trading among different sectors forms *demand* for different types of cars. On the other hand, there is car *supply* which consists of different car types with different vintages.

Car prices are adjusted at equilibrium. Equilibrium is reached when supply (production and stock) and demand (consumer and scrapping) are equal for each new and used car in each period. Supply and demand interaction will also determine new cars designs in a long run. However, generally car fleet models assume that car designs and prices are fixed exogenously and also ignore the used-car market. Therefore, these models will tend to over or under predict market shares (Berkovec (1985b)).

Comprehensive car fleet models should be able to model different sectors of the car market as well as their interactions in order to quantify responses of these sectors to different possible policies. To sum up, car fleet models should be able to describe and predict following items (Forsman and Engström (2005)):

- total number of cars in the fleet,
- market shares of different car types and vintages,
- driven distances by different car types and vintages.

The above-mentioned items lead to different components of car fleet models that are described below

Car ownership models quantify the number of cars owned. These models exist at both aggregate and disaggregate level. Aggregate models (see e.g. Button et al.; 1993; Tanner; 1983) are a function of income or GDP. Disaggregate car ownership models relate the effect of car characteristics and socio-demographic and socio-economic to car demand at the level of individuals or households (see e.g. Bhat and Pulugurta; 1998).

Car types models typically proceed car ownership models, since car type choice is conditional on car ownership. However, some studies deal only with car type choice (see e.g. Berkovec; 1985a; Manski and Sherman; 1980)Moreover, some studies investigate the effect of attitudinal and residential location variables (See e.g. Choo and Mokhtarian; 2004; Golob and Hensher; 1998; Tertoolen et al.; 1998, respectively).

Car utilization models Some studies model car ownership and type choice models jointly with car use or distance driven (see e.g. Bhat and Sen; 2006; Mannering and Winston; 1985). They argue that car type choice is not exogenous to car use models, and there may be common unobservable factors that affect both models.

As mentioned earlier, market supply consists of new models introduced by manufacturers, scrapping old cars, export and import and also second-hand cars put on the market by consumers. Each of these components can be modeled through econometric models or other simple models. Among different parts of supply, scrapping models have received the most attention (De Jong et al.; 2001; Kim et al.; 2004; Manski and Goldin; 1983, see e.g.). There exists few studies that address new technology production models (see e.g Berkovec; 1985a,b) or used market (see e.g Schiraldi; 2011). Some of the studies model the whole market including, demand which is car ownership and car type, and supply which is scrapping and new technology production (see e.g Berkovec; 1985a,b; Manski; 1983; Rust; 1985).

In this thesis, the focus is on car type choice and the choice of cars are modeled conditional on the decision to buy a car has already made.

Swedish car fleet model: A summary

Large, heavy and powerful cars have been historically popular in the Swedish car fleet. These cars are characterized by high fuel consumption and CO_2 emissions. In order to have a more fuel efficient and environmentally friendly fleet, many policies targeting both demand and supply have been implemented in recent years, in Sweden. For the extensive review of these policies refer to Hugosson and Algers (2012) and Pädam et al. (2012). These efforts have been successful in increasing the number of clean cars in newly bought cars. However, the Swedish car fleet is still the heaviest fleet in all Europe (Hugosson and Algers; 2012). Figure 2.1 summarizes the policies implemented in Sweden during 2006-2013 and their effects on the market share for new registered private cars.

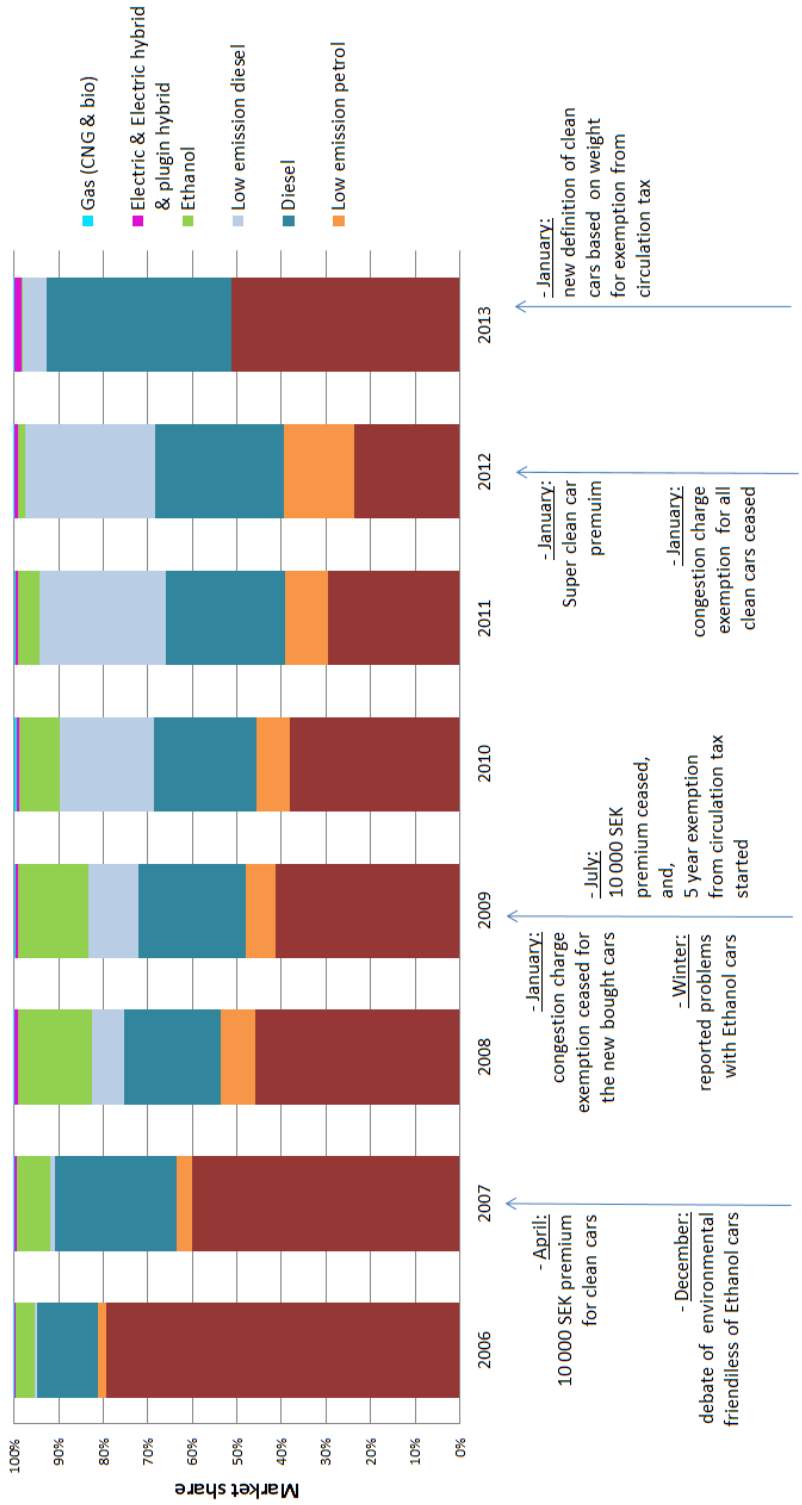

Figure 2.1: Policies implemented each year and the fuel type market share for new registered private cars

It should be noted that the definition of new cars has been changed over time in Sweden, and the current definition is based on the combination of fuel type, fuel consumption, CO_2 emission, weight and environmental class of cars[i].

In 2006, a car fleet model was developed for Swedish road administration (see full description in Hugosson et al. (2014)). This model consists of three sub-models:

- car ownership model that models total fleet size (total number of cars in the fleet), developed by Matstoms (2002)
- car type choice models (for new cars) developed by Transek (2006)
- scrappage model developed by Transek (2006)

Car type choice sub-model is the most sensitive one to the policies. The model has been estimated both on revealed and stated preference data.

[i]for detail information refer to Swedish Transport Administration `http://www.trafikverket.se`

3

Data

To build a car type model, we use data from different sources. On one hand, we have access to the Swedish registry data from 2004-2012. This dataset contains some information specific to the registration of the car (e.g. first registration date and date for the last status change), some main car characteristics such as brand, model name, model year, fuel type, weight, power and body type, and some limited socio-economic information like age, gender and home municipality of the owner. However, this database has some problems as follows:

- there appears to be an ambiguous encoding of the make and model
- some of the essential characteristics of cars are missing such as price.
- lack of data on household

In this thesis, we focus on private cars that have been bought new. Therefore, these observations need to be selected. For this purpose cars that are registered for the first time a given year but are actually older should be excluded. Therefore, we choose to eliminate these older cars so that we can have a more accurate idea about the price paid for a car. A car is considered to have been bought new in year t if the first registration date is equal to the given year t and the "vehicle year" is equal to t or $t+1$. "vehicle year" is defined based on a combination of three attributes; model year, production year and first registration date since all three attributes are not available for all observations. Vehicle year is defined to be equal to model year if it is available, otherwise, the production year of a car and if this is not available either then it is equal to the year of first registration date. Imported cars are not included in the study in any case. This definition of a newly bought car is slightly different from the one used in the official statistics that also counts older cars in.

To gain access to additional attributes of cars, we have another source of data provided by YNNOR[i] company. This dataset includes detailed information on all available cars in Swedish market 1999-2012 up to the versions of cars. This database is called *supply* throughout this thesis.

Data matching

As described above to impute missing information like price, we need to match supply with registry data. Due to the difference in the level of details in two data sources, several alternatives from supply correspond to the same observation from registry data. Therefore, we need to group (aggregate) alternatives available in the supply based on observable characteristics from registry data such as vehicle year, make, model, fuel-type, etc.. Figure 3.1 shows as an example that an observed choice (e.g. Volvo-S40-diesel) can correspond to different versions of respective type (e.g. 1.6 D DPF, 2.0 bas DPF and D5 Aut DPF Kinetic). Therefore, these versions are grouped as an aggregate alternative which is Volvo-S40-diesel to be able to match with the observed choice from registry data.

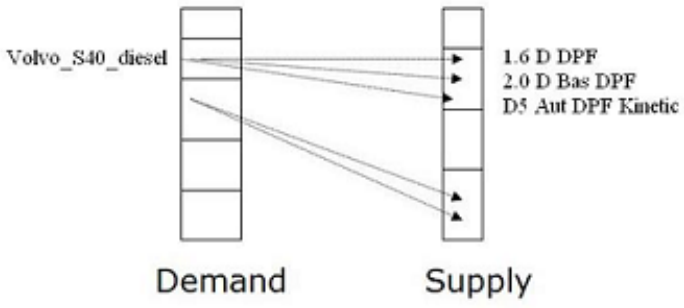

Figure 3.1: Matching demand with supply

Another uncertainty regarding matching is vehicle year. As mentioned earlier, in registry data, the vehicle year is defined based on a combination of model year, production year and first registration data. However, the only time related information exists in supply, year is the market introduction year. Vehicle year and market introduction year are not always equal, and the difference will affect the price as well as the choice set of each year.

[i] `https://www.ynnor.se`

4

Prediction problems

In this chapter, some of the prediction problems that are tackled in this thesis are introduced. The prediction problems dealt with in this thesis include uncertain future alternatives, aggregation, choice set definition and over-fitting. These problems are described below.

Uncertain future alternatives Future alternatives in the field of car fleet modeling can be uncertain in different ways. Uncertainties can be about new technologies introduced to the market, the number of new models or versions that each brand introduces. One way to deal with new technologies in the literature is to use stated preference surveys where consumers report their valuation of different attributes of the hypothetical future alternatives. However, the characteristics of future alternatives are not known with certainty, even for short-term forecasts. The problem of uncertainty about the number and characteristics of the future alternatives make the prediction for car type choice prediction difficult even for short-term prediction.

Aggregation Estimating models on the disaggregate level will capture heterogeneity and interdependencies among individuals as well as making use of all available data that increase the explanatory power of the model. However, when it comes to future prediction, we usually do not have access to the disaggregate data. Therefore, some suggest models estimated at the aggregate level. They also argue that disaggregate models may be misspecified due to the difficulty of capturing the complexity of interactions and interdependencies among decision makers. Moreover, the errors available in data at the disaggregated level might be offset at the aggregate level (see e.g. Aigner and Goldfeld; 1974; Grunfeld and Griliches; 1960). Therefore, the question is whether to estimate a disaggregate model and aggregate individual forecasts to the level of aggregate prediction of interest or to estimate an aggregate model at that level. This question is well-addressed in the literature and initiated by Grunfeld and Griliches (1960) with no particular answer.

According to Winters (1980), there are two types of aggregation problems "numerical" and "functional". Numerical is when both aggregate and disaggregate equations have the same specifications while functional aggregation recognizes the fact that different aggregation level requires different models. The theoretical literature on aggregation primarily deals with numerical aggregation.

Aggregation problem for prediction has several dimensions. Following aspects of aggregation problem are addressed in this thesis:

individuals Discrete choice models are often estimated at the level of individual decision-makers units such as individuals or households but used to predict an aggregate quantity like the market share of clean cars or an average response to a policy change. In discrete choice models, the consistent way of aggregating over individuals is sample enumeration (Train; 2009, p. 31). In the sample enumeration method, the choice probabilities are aggregated over all individuals. However, in practice, due to the lack of access to disaggregate data, the representative individual approach might be used where the attributes of the representative individual is the average over all individuals. Nonetheless, it should be investigated to what extent this approximation might cause bias in prediction.

alternatives Choice sets in logit models can be specified at a disaggregate level for the purpose of estimation such as all available varieties of cars, yet, the models might be used for the prediction of the market share of an aggregate choice, such as the market share of a brand or clean cars. The uncertainty about the aggregate alternatives is often less compared to disaggregate alternatives, in terms of definitions and observations. Therefore, the question will be whether to estimate models on aggregate or disaggregate alternatives.

time The problem of aggregation over time occurs when the data collected is over aggregated time rather than the time it happens actually or when the model is estimated over aggregated time. For example when different types of cars are introduced to the market in different months of a year but the data is available on the yearly basis, or when modeling the car type choice assuming that each individual has access to all alternatives at each time of the year.

Choice set definition Choice set definition problem can be well-defined under the categories of uncertain future alternatives or aggregation. However, we specify a section to it due to its exclusivity to logit models. Choice sets can be defined arbitrarily in many applications, for example for the destination choice; one can consider the exact location or the district or the municipality, or for the choice of the type of the bought car, the choice can be considered as the version of the car with all add-ons or its make and model or other attributes. However, these different choice set specifications will lead to different prediction results. As an example, consider a simple formulation of logit model for the probability of choosing alternative

j (assuming no individual specific variable) which is, $P_j = \frac{e^{(V_j)}}{\sum_j e^{(V_j)}}$. On the aggregate alternative level (clean vs. non-clean), the probability of choosing clean will be $P_{\text{clean}} = \sum_{j=\text{clean}} P_j$. Assuming all clean cars have the same utility and all non-clean as well, then:

$$\frac{P_{\text{clean}}}{P_{\text{non-clean}}} \sim \frac{\text{No clean}}{\text{No non-clean}}$$

Therefore, the probability of choosing a clean car is proportional to the number of clean cars in the supply. Therefore, we suspect that any particular logit model may struggle with out of sample predictions, where the choice sets are uncertain or misspecified.

Over-fitting Over-fitting occurs when a model is too fitted to the available data that loses its generality to be applied to another independent data. There is a trade-off between models with the best fit and the models that will have the highest predictive performance. A "good" in-sample fit of is unlikely, in itself, to give us much confidence in its out-of-sample forecast ability. he question is how to capture this uncertainty through model selection criteria or modeling consideration.

Prediction uncertainty There are so many sources that contribute to uncertainty in prediction. Some of them arise from data, but also from errors arising from the uncertainties in models specifications (Sims; 1986) should also be included. However, the second part of uncertainty is usually ignored (Sims; 1986). To show, any source of uncertainty, the models must provide the probability distribution of forecast, rather than simply make point predictions.

Any model estimated on a large amount of individuals that are assumed to be independent will have a problem in replicating the variance of the aggregate prediction. The reason is the sum of many independent decisions would cause the prediction errors to be offset. Therefore, considering the correlation among individuals is necessary to capture the forecast error between observed data and model aggregate prediction.

Model selection One of the fundamental questions in modeling is to choose the model that explains the data the best. There are different methods for model selection. The conventional methods are classical hypothesis testing (e.g. t-test), maximum likelihood, Akaike information criterion (AIC) that is $-2LL + 2K$, where LL is the value of log-likelihood, and K is the number of parameters and Bayesian information criterion (BIC) that is $-2LL + \log(N)K$, where N is sample size. Another model selection method in the literature is cross-validation (CV). CV splits data, once or several times, part of the data (the training sample) is used for estimation (training), and the remaining part (the validation sample) is used for validating the estimated results. Various ways of splitting data will lead to different methods

of cross-validation. A single data split is called simple validation or hold-out validation. Moreover, the idea of using CV for model selection was discussed by Efron and Morris (1973) and Geisser (1975). CV selects the models that give the smallest error over validation sample. CV has been used widely in different applications due to its simplicity and universality (Arlot and Celisse; 2010).

There are few studies in frequentist approach that address model selection for the purpose of prediction, among them the early work of Geisser and Eddy (1979) can be mentioned. Mosier (1951) criticizes conventional cross-validation methods where validation sample is selected randomly from the training sample by proposing the idea of validation beyond the sample pattern under the name of "validity generalization". Later, Busemeyer and Wang (2000) formally identify the procedures for model selection for more general use in psychology. Although the methods are used widely in psychology, to the best of our knowledge, the only work that in economics that refers to it is the work of Keane and Wolpin (2007) where they deliberately choose a non-random sample of the data in the direction of policy to be evaluated. As mentioned earlier, there exist few studies focusing on prediction. Therefore, many of problems mentioned above have not been addressed in this field, yet. In the following section, we discuss some of these prediction problems in different papers and present methods to tackle them.

5

Research outlines

The thesis put forward here is the effort of a natural work process, and therefore, each work is built on the outcome of the other[i]. In this section, the objectives and main findings of each paper are presented as well as how they are related to each other and to the goal of the thesis that is improving predictive power of logit models.

Paper I

To include a large number of alternatives available in the market is a burdensome task for modeling car type choice. The problems include lack of detailed data on all the available cars in the market or on the observed choice of individuals, or computational limits. However, the latter is not a determining factor anymore considering the recent advances in computer technology. The common practice to deal with this problem is to group alternatives into categories with the same key features such as make, model, vintage, body type and fuel type, but different other characteristics such as engine size, power, weight. The alternatives included in each category are called disaggregate alternatives. Each of these categories is represented by an alternative whose characteristics are averaged over characteristics of the cars within that category. This representative alternative is called aggregate alternative. *The objective is to investigate empirically to what extent the number and heterogeneity of disaggregate alternatives represented by aggregate alternative influences the explanatory power of models and/or their prediction accuracy.*

Method Ben-Akiva and Lerman (1985) introduce an approach for correctly including aggregated alternatives in discrete choice models. they include "measure of size" that

[i]However, the papers are not presented in the chronological order to make the thesis easier to follow for the reader.

is the number of disaggregate alternatives under each aggregate one and "measure of heterogeneity" that is if the disaggregate alternatives are heterogeneous. However, few studies have implemented this method in practice. It could be due to the lack of detailed data. Mabit (2011) includes only the measure of the size in his model. In this paper, we also include the measure of heterogeneity of disaggregate alternatives. Hence, we capture the effects of the supply side in terms of the number and types of added disaggregate alternatives.

The results show that including heterogeneity of disaggregate alternatives in the model improves model explanatory power but does not improve prediction.

Paper II

Based on the results from paper II which shows improvement in model fit does not improve the prediction results, the question raised is, are the 'best' estimated models necessarily the 'best' ones for prediction as well? In other words, what are the criteria for model selection when the purpose is forecasting. *In this paper we analyze the prediction problem and focus on model selection with the objective of improving the predictive ability of discrete choice models. Furthermore, we investigate to what extent different prediction questions lead to different "best" models.*

Method Instead of standard approaches that are typically used for the purpose of inference, we use non-random hold-out validation for model selection in feature (variable) setting. Feature selection is an automatic way of selecting variables for a model such that the criterion is optimized. In general, selecting variables based only on prior knowledge is not likely to be accurate (see e.g. the Train (1979b)). However, this problem will be more severe in car type choice modeling where there exists a large number of car attributes to select among as well as their correlations and interaction (for discussion about the problems of correlations among car attributes see e.g. Train; 1979a). Therefore, feature selection can be a useful method in the car type choice application. We use feature selection as a model selection tool by introducing model selection criterion as its objective function. We introduce two different model selection criteria that are maximum likelihood which is the conventional method of model selection, and root mean squared error of the prediction quantity of interest.

Feature selection is typically used with cross-validation (CV). CV is applied in the nonrandom holdout manner where the holdout sample is the data of a consecutive year. Considering the changes in the supply of each year, using the supply of a given year both for estimation and validation is not likely to provide us with the accurate predicted results since the model might be over-fitted to the supply of that year. *To the best of our knowledge, model selection based on non-random hold-out validation has not been used in economics and choice modeling. Moreover, we use the non-random hold-out validation, in the automatic feature selection formulation.*

The results show that different prediction questions lead to different best models. The results validate the pragmatic approach that alternative models may coexist for

different purposes unlike the absolutist's assumption of the existence of a true model. Models obtained with likelihood as a criterion function have lower predictive performance compared to the ones obtained with the root mean squared error of the prediction question of interest, as their criterion. The reason is that log-likelihood based on individual observation aims at overall highest fit to the data. Therefore, assigns the same weights on all alternatives and observations whereas for prediction we are specifically interested in a sub-section of the data.

Paper III

The problem that is addressed in this paper is what is the appropriate aggregation level for modeling for the purpose of predicting an aggregate question. This question is of great importance in discrete choice modeling since these models are typically estimated at the individual level but often used for the purpose of predicting an aggregate quantity, such as the market share of clean cars. Since neither of the models at the aggregate or disaggregate level is the true model and the data is also not perfect, in this paper, *we propose to tackle the aggregation problem by employing model combination methods to combine aggregate and disaggregate models. To the best of our knowledge, model combination has not been used for tackling the aggregation problem.*

We examine specifically the effect of aggregation on the prediction accuracy of a nested multinomial logit (NMNL) to predict the monthly share of clean cars in the Swedish car fleet. We investigate a situation wherein the large scale models are already estimated, and we are interested in improving their prediction performance in a post-processing manner. The examples of such large-scale models are national travel/demand models or car ownership/car type models. Different aspects of aggregation are covered in this paper; those are aggregation over time, individuals and alternatives.

Method We combine NMNL with a regression tree to capture individual heterogeneity as well as a time-series model to capture dynamics of the market share of clean cars at the aggregate level. Model combination approach recognizes that each model addresses different aspects of data. The model combination has been used in different fields and, in general, has shown improvement in prediction performance (for reviews see e.g. Clemen; 1989; Hoeting et al.; 1999). However, in this paper, model selection is used to overcome the aggregation problem. Models are combined at aggregate and disaggregate level through the latent variable model (Ben-Akiva et al.; 2002; Walker and Ben-Akiva; 2002) and the Bayesian model averaging approach (see e.g. Hoeting et al.; 1999). We do not estimate model combination weights and model parameters simultaneously since we consider a situation that already estimated models are available. Therefore, models and weights are estimated sequentially.

We address the problem of combining models that are estimated at different aggregation levels and propose to use the *aggregate likelihood* to combine them at the aggregate level. The aggregate likelihood is the likelihood of aggregate data given the aggregate point prediction of the model. We suggest using the aggregate likelihood for model com-

bination and selection when the purpose of modeling is aggregate prediction. Moreover, we use aggregate likelihood for sequential estimation of the latent variable model. The latent model estimated with the aggregate likelihood gives the best prediction performance among all combined models.

The prediction results show that all the combined models perform better than any single model. However, the models obtained by the combination at the aggregate level perform better than latent models that are obtained by disaggregate combination. Improvement in the aggregate prediction performance of the latent model combined with a time-series model, show that there is still some aggregate information that are not captured by the disaggregate latent model despite incorporating the aggregate prediction of time-series into the nested logit model. We argue that this aggregation inconsistency is due to not considering correlation among individuals within each month. However, any model estimated on a large amount of individuals that are assumed to be independent will have a problem in replicating the variance of the aggregate prediction. The reason is the sum of many independent decisions would cause the prediction errors to be offset.

Paper IV

In this paper, we demonstrate a real world application of the car fleets. Early 2014, an official Swedish government investigation report released including proposals to promote a Fossil Free Fleet in Sweden by 2050. Unfortunately, these proposals lack quantitative evaluation support. In this final paper, *we have used the Swedish car fleet model to evaluate a Bonus-Malus policy package proposal.* It is quite challenging to evaluate policy packages for the whole fleet using modeling. *We contribute in evaluating policies by considering the whole fleet. Not many studies model the change of the entire car fleet.*

The proposed scenarios address both cars bought privately and by companies and public organizations. These scenarios differ in designs for registration tax, vehicle circulation tax (road usage tax), clean car premiums, company car benefits tax and fuel tax. In addition to updating the car fleet model, we build a simple supply model to predict future supply. The Bonus-Malus is a system that is designed to reward (bonus) car buyers who choose to purchase a car with lower CO_2 emissions and penalize (malus) customers who select a car with higher CO_2 emissions. In such a system those who choose to buy a car with higher CO_2 emissions subsidize the purchase of those who select a car with lower CO_2 emissions. Therefore, the system should pay for itself and not rely on public funding.

The results show it is not likely that the proposal reaches its goals. Since it reduces the number of ethanol and gas cars and the increase rate of electric cars is not enough to reach a fossil independent fleet. However, there is a potential to use Bonus-Malus budget surplus to boost policy efficiency.

Paper V

Since the first application of Swedish car fleet model for the Swedish Environmental Protection Agency (Naturvårdsverket; 2007) to analyze climate policies, the model has been used in several studies, which gives the possibility to evaluate the performance of the model in different situations. These extensive applications give us an excellent opportunity to perform "after and before" study. Therefore, the objective of the first paper is *to compare the predicted effects of applied policies with their actual outcomes.*

The comparison of actual to predicted outcomes is surprisingly few in the choice modeling field (even in economics or marketing (Keane and Wolpin; 2007)). The main finding of this paper is that the car type choice model of the Swedish car fleet model is largely dependent on the new technologies entering the market. Moreover, the uncertainty regarding the time of the year that policies are implemented or new car types introduced to the market plays a significant role. Therefore, improving the car type choice models conditional on these exogenous factors, will not improve the final forecasts, and, the car type choice models should be developed such that the sensitivity to the future scenarios is captured. For the rest of the thesis, we move towards more reliable forecasts by capturing uncertainties mentioned above.

6

Conclusions and future work

The focus of this thesis is on the prediction side of logit models and to investigate the extent the models with the best explanatory power are different from the ones with the best predictive power. While the distinction between prediction and inference is recognized in philosophy of science (Shmueli; 2010), the difference is still not distinguished in statistical literature. Prediction problems are discussed in general and the one specific to logit models. Moreover, we develop methods to overcome these problems. The results of the papers in this thesis support the idea that modeling considerations of modeling to find a model with the best explanatory power are different from those of modeling with the purpose of obtaining the best predictive power.

Car type choice models are typically modeled on the aggregate alternatives nut the actual disaggregate chosen alternative. However, including corrections for the measures for the number and heterogeneity of disaggregate alternatives forming aggregate alternatives in the models, increase the explanatory power of the models or their fitness while these models give less accurate predictions. Moreover, the model selection with the different selecting criteria support the pragmatic view in that, the best models for prediction differ considerably according to the prediction question to answer, and the best models for the purpose of inference are not necessarily the best for that of prediction.

Aggregation is also another problem that rises when estimating models for prediction. Usually, when it comes to forecast, especially long term forecast, the data available is not as rich as the data we have access today, and usually it exists in the more aggregate format. The question is for the purpose of aggregate prediction, whether to estimate on disaggregate or aggregate data. Our result support modeling on the aggregate level or to revise the disaggregate models according to the prediction question of interest. One of the most important points of considerations are aggregate alternatives to the same level of aggregation of prediction question as well as re-estimating logit model with the likelihood function defined at the same level of aggregation. Finally, one less addressed

issue in the literature of logit models are co-relation among individuals which plays an important role in obtaining reliable aggregated results and prediction uncertainty.

Additionally, we present the results of real-world applications where the uncertainty of future new technologies are challenging even for short-term forecast.

As a final note, although, there should be a distinction made between prediction and explanatory variable of the models, in most cases models own some level of both explanatory and predictive power (Shmueli; 2010). However, there is at least one area that model should posses both explanatory and predictive power; policy analysis. For the purpose of policy evaluating, we do need to have accurate prediction about the extent of the effect of the policy that is implemented. On the other hand, the policies are designed based on the theories on how a system behaves and the relevant causal relation. This is not the area that this thesis deals with but has a great potential of further research. This thesis focuses on the prediction problems in the area of discrete choice modeling and within car fleet modeling application and propose different solutions to solve the prediction problems in different papers.

Bibliography

Aigner, D. J. and Goldfeld, S. M. (1974). Estimation and prediction from aggregate data when aggregates are measured more accurately than their components, *Econometrica: Journal of the Econometric Society* pp. 113–134.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection, *Statistics Surveys* **4**: 40–79.

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

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Börsch-Supan, A., Brownstone, D., Bunch, D. S. et al. (2002). Hybrid choice models: progress and challenges, *Marketing Letters* **1**3(3): 163–175.

Berkovec, J. (1985a). Forecasting automobile demand using disaggregate choice models, *Transportation Research Part B* **1**9B(4): 315–329.

Berkovec, J. (1985b). New car sales and used car stocks: a model of automobile market, *Rand Journal of Economics* **1**6(2).

Bhat, C. R. and Pulugurta, V. (1998). A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions, *Transportation Research Part B: Methodological* **3**2(1): 61–75.

Bhat, C. R. and Sen, S. (2006). Household vehicle type holdings and usage: an application of the multiple discerete-continous extreme value (MDCEV) model, *Transportation Research Part B* **4**0: 35–53.

Boyd, J. H. and Mellman, R. E. (1980). The effect of fuel economy standards on the U.S. automotive market: An hedonic demand analysis, *Transportation Research Part A: General* **1**4(5-6): 367–378.

Brownstone, D., Bunch, D. S. and Golob, T. F. (1994). A Demand Forecasting System for Clean-Fuel Vehicles, *Transportation* (221): 15.

Busemeyer, J. R. and Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology, *Journal of Mathematical Psychology* **4**4(1): 171–189.

Button, K., Ngoe, N. and Hine, J. (1993). Modelling vehicle ownership and use in low income countries, *Journal of Transport Economics and Policy* pp. 51–67.

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* **3**8(3): 201–222.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting* **5**(4): 559–583.

De Jong, G., Vellay, C. and James Fox, R. E. N. (2001). Vehicle scrappage: literature and a new stated preference survey, *European Transport Conference 2001 proceedings*, Association for European Transport.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its Competitors An Empirical Bayes Approach, *Journal of the American Statistical Association* **6**8(341): 117–130.

Forsman, A. s. and Engström, I. (2005). The composition and use of the Swedish car fleet - formulation of a forecasting system, *VTI rapport 518A*, VTI, SE-581 95, Link{ö}ping, Sweden.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications, *Journal of the American Statistical Associatioin* **7**0(350): 320–328.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection, *Journal of the American Statistical Association* **7**4(365): 153–160.

Goldberg, P. K. (1995). Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry, *Econometrica* **6**3(4): 891–951.

Golob, T. F. and Hensher, D. A. (1998). Greenhouse gas emissions and Australian commuters' attitudes and behavior concerning abatement policies and personal involvement, *Transportation Research Part D: Transport and Environment* **3**(1): 1–18.

Greene, D. L. (1986). The market share of diesel cars in the USA, 1979-83, *Energy Economics* **8**(1): 13–21.

Grunfeld, Y. and Griliches, Z. (1960). Is aggregation necessarily bad?, *The Review of Economics and Statistics* pp. 1–13.

Hensher, D. A. and Ton, T. T. (2000). A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice, *Transportation Research Part E: Logistics and Transportation Review* **3**6(3): 155–172.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial, *Statistical Science* **1**4(4): pp. 382–401.

Huang, Z., Zhao, H. and Zhu, D. (2012). Two New Prediction-Driven Approaches to Discrete Choice Prediction, *ACM Transactions on Management Information Systems (TMIS)* **3**(2): 9.

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.

Hugosson, M. B., Algers, S., Habibi, S. and Sundbergh, P. (2014). The Swedish car fleet model, S-WoPEc Working Paper No 2014:18. CTS - Centre for Transport Studies Stockholm (KTH and VTI).

Keane, M. P. and Wolpin, K. I. (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior*, *International Economic Review* **4**8(4): 1351–1378.

Kim, H. C., Ross, M. H. and Keoleian, G. A. (2004). Optimal fleet conversion policy from a life cycle perspective, *Transportation Research Part D: Transport and Environment* **9**(3): 229–249.

Lave, C. A. and Train, K. (1979). A disaggregate model of auto-type choice, *Transportation Research Part A: General* **1**3(1): 1–9.

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

Mannering, F. and Winston, C. M. (1985). A Dynamic Empirical Analysis of Household Vehicle Ownership and Utilization, *RAND Journal of Economics* **1**6(2): 215–236.

Manski, C. F. (1983). Analysis of equilibrium automobile holdings in Israel with aggregate discrete choice models, *Transportation Research Part B: Methodological* **1**7(5): 373–389.

Manski, C. F. and Goldin, E. (1983). An Econometric Analysis of Automobile Scrappage, *TRANSPORTATION SCIENCE* **1**7(4): 365–375.

Manski, C. F. and Sherman, L. (1980). An empirical analysis of household choice among motor vehicles, *Transportation Research Part A: General* **1**4(5-6): 349–366.

Matstoms, P. (2002). Modeller och prognoser för regionalt bilinnehav i sverige (models and forecasts for regional car ownership in sweden, *VTI rapport* **4**76.

McFadden, D. L. (1978). Modelling the Choice of Residential Location, *Transportation Research Record* (673): 72–77.

Mohammadian, A. and Miller, E. J. (2002). Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance, *Transportation Research Record: Journal of the Transportation Research Board* **1**807(1): 92–100.

Mosier, C. I. (1951). The need and means of cross validation. i. problems and designs of cross-validation., *Educational and Psychological Measurement* .

Naturvårdsverket (2007). Drivkrafter till bilars minskade koldioxidutsläpp (driving forces to reduce co2 emissions for cars, in swedish with a summary in english), *Report 5755*, Naturvårdsverket, Stockholm.

Pädam, S., Sundbergh, P. and Strömblad, E. (2012). Andelen miljöbilar i nybilsförsäljningen i stockholms län - hur har utvecklingen sett ut och hur kan andelen ökas? (the percentage of green cars in new car sales in stockholm city - how have the developments been in this area and how can this share be increased?), *Technical report*, WSP AB, Sweden.

Rust, J. (1985). Stationary Equilibrium in a Market for Durable Assets, *Econometrica* **5**3(4): pp. 783–805.

Schiraldi, P. (2011). Automobile replacement: a dynamic structural approach, *The RAND journal of economics* **4**2(2): 266–291.

Shmueli, G. (2010). To explain or to predict?, *Statistical science* pp. 289–310.

Sims, C. A. (1986). Are forecasting models usable for policy analysis?, *Federal Reserve Bank of Minneapolis Quarterly Review* **1**0(1): 2–16.

Tanner, J. C. (1983). International comparisons of cars and car usage, *Technical report*.

Tertoolen, G., van Kreveld, D. and Verstraten, B. (1998). Psychological resistance against attempts to reduce private car use, *Transportation Research Part A: Policy and Practice* **3**2(3): 171–181.

Train, K. (1979a). Consumers' responses to fuel-efficient vehicles: a critical review of econometric studies, *Transportation* **8**(3): 237–258.

Train, K. E. (1979b). A comparison of the predictive ability of mode choice models with various levels of complexity, *Transportation Research Part A: General* **1**3(1): 11–16.

Train, K. E. (2009). *Discrete choice methods with simulation*, Cambridge university press.

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

Walker, J. and Ben-Akiva, M. (2002). Generalized random utility model, *Mathematical Social Sciences* **4**3(3): 303–343.

West, S. E. (2004). Distributional effects of alternative vehicle pollution control policies, *Journal of Public Economics* **8**8: 735–757.

Winters, L. (1980). Aggregation in logarithmic models: Some experiments with uk exports*, *Oxford Bulletin of Economics and Statistics* **4**2(1): 36–50.

Papers