Two-Stage Discrete Choice Models for Numerous Alternatives in Travel Decision Contexts

(選択肢が多い交通行動意思決定における二段階離散選択モデルに関する研究)

XU, Gang

(徐　剛)

Doctor of Engineering

Graduate School of Environmental Studies, Nagoya University

(名古屋大学大学院環境学研究科　博士（工学）)

2016

Two-Stage Discrete Choice Models for Numerous Alternatives in Travel Decision Contexts

Doctoral Dissertation

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Engineering

by
XU, Gang

Academic Adviser:
Prof. Takayuki Morikawa

Graduate school of Environmental Studies

Nagoya University

Jan, 2016

Acknowledgements

This dissertation is the summary of the research works which I have done after I came to Nagoya University. It is really an unforgettable experience in my life. I am very fortunate to meet many good friends who I spent pleasant days together, and professors who helped me a lot in my research. Therefore, it is my honor to show my gratitude here to them all.

In the first place, I would like to show my gratitude to Prof. Takayuki Morikawa. He gave me the opportunity to be a research student in Nagoya University, and he not only gave me a lot of help in my research filed, but also very cared about our foreign students' life in Japan. He gave us a very harmonious atmosphere in the laboratory where we could do the work with interests rather than pressure.

I gratefully acknowledge Assoc. Prof Tomio Miwa. He lead me into the research filed from the beginning and patiently taught me everything. He is not only an excellent teacher, but also a trustable friend. He helped me in my whole research, no words can fully express my gratitude.

I am grateful to Prof. Hiroki Tanikawa for his valuable comments and advices as a member of the doctoral examination committee. His advice helped me a lot in improving my dissertation.

Furthermore, I would like to show my gratitude to other members in the laboratory. Prof. Toshiyuki Yamamoto who always pointed out the most important limitations in my research and gave constructive advices. Assistant. Professor Hitomi Sato who helped me with all laboratory routine works.

Finally, I would to give the most sincere thanks to my family, including my parents, my wife, and my son. My parents gave me unreserved supports from the past to the present. For my wife, her love is the reason why I came to Japan and the love between us make our life full of meaning. Finally, I would like to thank my son who make our big family full of happiness since he was born.

Abstract

Choice behavior analysis is an important research area in transportation research filed. It would help researchers to understand how decision-makers makes decisions, the preference on each alternative's attribute. The prediction of future situation would be done by the current choice analysis, and it would be helpful in many areas, such as helping government to design policy to improve the traffic situation.

Discrete choice models are always applied in the choice behavior analysis in the transportation area. During the model estimation process, we assume decision-makers choosing from a choice set. In other word, decision-makers will evaluate all alternatives in the choice set and then make a choice. However, there is an issue in the choice behavior in transportation area is that the number of available alternatives is always huge. For example, in route choice context, the possible routes from an origin to a destination would be hard to be counted if the road network is dense and the distance is long. Therefore, it is not reasonable to make the assumption that decision-makers will evaluate all alternatives in the choice set to make a decision.

In this study, we assume the decision-maker will firstly screen all alternatives in the choice set and include qualified alternatives into a new choice set called the consideration set, then choose the choice alternative from the consideration set. This is call two-stage choice including consideration set formation stage and choice making stage. However,

researchers always do not have the information about decision-makers' consideration set and only have the observed choices information. Therefore, the two-stage choice model should be estimated with only the observed choice information.

The objective of this study is applying the two-stage choice context in the transportation choice analysis. Models with different mechanisms will be assigned to different choice behaviors. The difference between two-stage model and models with only choice-making stage will be compared.

In Chapter 1, we give a brief introduction of the choice behaviors in transportation research area and introduce some basic choice models. The research objective is given and the outline of this dissertation is also presented.

In Chapter 2, a probabilistic choice set (PCS) model is applied to route choice analysis. Route choice behavior is treated as a two-stage process consisting of a choice set generation stage and a choice making stage. In the choice set generation stage, drivers include the routes that satisfy their spatiotemporal constraints into an individual choice set from which an actual route is selected in the following stage. In the choice making stage, drivers choose the route with maximal utility. The data used in this research is 2011 probe vehicle data collected in Toyota city, Japan. This data gives information about drivers' choices in the choice making stage, but lacks any information about the choice set generation stage. In carrying out the computation, models for both stages are estimated simultaneously based on only drivers' choice information. The estimation results demonstrate that the PCS model performs well compared with the multinomial logit

(MNL) model, a result that also indicates the validity of viewing route choice behavior as a two-stage process.

In Chapter 3, we modeled the vehicle purchasing behavior as a two-stage choice process. In the first stage, a household specific consideration set is drawn from the all possible vehicles, and then a discrete choice model is applied to model the final choice based on consideration sets. The hazard-based choice set formation model is utilized in the first stage, and the accepting probability for each vehicle to be considered by each household is decided by the vehicle's price, vehicle's fuel cost and household's socioeconomic attributes. Then, the consideration set for each household is randomly drawn according to each vehicle's accepting probability, and vehicle with higher probability will have bigger chance to be included into the consideration set. Then, a multinomial logit (MNL) model is applied to the final selection step with the consideration sets. In order to investigate the advantage of using consideration set as choice set, MNL models with the universal set which includes all possible vehicles, and a pseudo-random selected choice set for each household are also estimated separately. Estimation results indicate the priority of using consideration set in vehicle purchasing behavior analysis.

In Chapter 4, we try to estimate the two-stage choice model in a Bayes approach with the same data in Chapter 3. The applied Hierarchical Bayes model can avoid the issue of an indifferentiable and irregular likelihood surface caused by thresholds and discontinuities, and the data augmentation and Markov-chain Monte Carlo estimation methods make it possible to estimate two stages simultaneously using only the information about the consumers' actual choices. We tried different screening rules and then compare the

average consideration set size to find out the best one. And then, we also compare choice models in both Chapter 3 and Chapter 4 in constructing the consideration set and prediction accuracy. The result indicate that Bayes model with a good screening rule is better the model in Chapter 2. Finally, we analyze the consumers' behavior under different choice scenarios. If the consumer is intend to buy the first and only vehicle, the compact vehicle, sedan, wagon and SUV are preferred comparing to the consumer who intends to buy an additional vehicles.

In Chapter 5, a conclusion is presented. The two-stage choice model shows the advantage in choice with many possible alternatives in the transportation research area. The biggest difficulty in two-stage choice estimation is in modelling the consideration set formation process, due to the lack of information and the limit of the estimation method. Therefore, in the future, in one hand, it is better to try different estimation methods with the limit information; in the other hand, it is meaningful to compare the consideration set generated by models and the real consideration sets from decision-makers if applicable.

Contents

List of Tables

List of Figures

1. Introduction

1.1 Background

Choice is a basic task which human beings are facing every day and everywhere. In Transportation research area, choice behavior analysis exists in many important topics including travel mode choice, route choice, vehicle purchasing choice, location choice, departure time choice and so on. In every choice situation, decision-makers are assumed that they will choose their final favorite alternative from a choice set. Generally, the choice set includes all available alternatives, such as in the situation of travel mode choice, available alternatives are generally including bus, subway, self-driving, bicycle, taxi and so on. Decision-makers could choose any one of these travel modes as long as they prefer it.

Random utility models are always applied in transportation choice analysis. These models were derived from the utility theory (McFadden, 1974). Under the utility theory, choice makers are assumed to be rationally, and they will choose the goods with the highest utility among the available alternatives. However, we cannot measure the utility to individual directly. Therefore, in random utility models, we suppose that the utility U_{in} of alternative *i* to individual *n* is composed by a deterministic component V_{in} and a stochastic error component ε_{in} , as shown in equation (2.1).

$$U_{in} = V_{in} + \varepsilon_{in} \qquad [1.1]$$

By assuming the error component to follow different distributions, there will be different

random utility models. If the error component follows normal distribution, then the model is called Probit model, and the probability of individual *n* choose alternative *i* in a Probit model (Sheffi, 1985) is given by following equation:

$$P_n(i) = \int_{\varepsilon_{i=-\infty}}^{+\infty} \left(\int_{\varepsilon_{1=-\infty}}^{V_i - V_1 + \varepsilon_i} \int_{\varepsilon_{2=-\infty}}^{V_i - V_2 + \varepsilon_i} \cdots \int_{\varepsilon_{J=-\infty}}^{V_i - V_J + \varepsilon_i} f_\varepsilon(\varepsilon) d\bar{\varepsilon} \right) d\varepsilon_i \qquad [1.2]$$

where $d\bar{\varepsilon} = d\varepsilon_J \ldots d\varepsilon_{j+1} d\varepsilon_{j-1} \ldots d\varepsilon_2 d\varepsilon_1$. Furthermore, if the error component are taken to be independently and identically distributed following the Gumbel distribution, then it will become a Multinomial Logit (MNL) model (Ben-Akiva and Lerman, 1985), and the choice probability will be:

$$P_n(i) = \frac{exp(V_{in})}{\sum_{j \in M} exp(V_{jn})} \qquad [1.3]$$

Generally, when we apply these models in the choice analysis, it is always assumed that decision-makers will consider all alternatives in the choice set and then choose one with the highest utility. However, under some choice situations in transportation area, there is an issue that the number of available alternatives would be numerous. For example, in the vehicle purchasing choice circumstance, there are hundreds of different types of vehicles in the market which are all available for potential customers. Theoretically, for each consumer, he can consider all available vehicles and then choose one from them. However, in fact, this assumption is not reasonable. Consumers will not consider all vehicle types for purchasing but consider only a part of vehicles types which satisfy consumers some constraints on the vehicle. For example, if a consumer's budget is 3 million Yen, then all vehicles which are more expansive than 3 million Yen will be excluded from the consideration, no matter how good they are. Finally, he will only consider vehicles under

3 million Yen and then choose one from them. If the number of all available alternatives is still countable, in the route choice situation in transportation area, the number of available alternatives would beyond imagination in route choice case, especially when the travel distance is long and road network is complex. Firstly, it is impossible for drivers to know all available routes from an origin to destination. Furthermore, under some situations, for example if the driver would like to avoid the toll road, then some available routes which include toll roads would become unavailable for the driver under this situation.

Therefore, we found that it is inappropriate to simply assume decision-makers will evaluate all available alternatives in the choice set and then choose one from them. Sometimes, they will only consider a part of available alternatives and then choose from them. If we call the choice set including all available alternatives is the universal set. Then, the choice set including alternatives which decision-makers would like to evaluate is called a consideration set. The consideration set is a non-empty subset of the universal set which includes the decision-makers' final choices. Shocker et al. (1991) defined ''consideration set'' as ''purposefully constructed and can be viewed as consisting of those goal-satisfying alternatives salient or accessible on a particular occasion.'' Some studies have proved the existence of consideration sets and how they influence choice behaviors (Shocker et al., 1991; Kardes et al., 1993). Parkinson and Reilly (1979) discovered the consideration set formation process using an information processing perspective. Payne (1976) investigated that consumers use consider-then-choose decision processes, and a lot of studies has proved this phenomenon (Shocker et al., 1991; Bronnenberg and Vanhonacker 1996; Brown and Wildt 1992; Hauser and Wernerfelt 1990;

Metha, Rajiv, and Srinivasan, 2003; Paulssen and Bagozzi 2005; Wu and Rangaswamy 2003).

However, it is a hard work to collect the consideration set information directly from the decision-makers. The best method is directly asking decision-makers to describe their consideration sets or strategies about how to construct their own consideration sets. Then researchers would apply such consideration set information into discrete choice models. But, in fact, it is inefficient and cost to ask decision-makers to describe their consideration sets. Therefore, it is better to explore the consideration set formation process with the limited choice information.

Manski (1977) proposed a choice model which introducing the consideration set into it. The probability of choosing alternative *i* in a two-stage choice perspective (consider-then choose) is:

$$P_n(i) = \sum_{C \in G} P_n(i \mid C) * Q_n(C \mid G) \qquad [1.4]$$

where,

- $P_n(i)$: Probability of individual *n* choosing route *i* from master set
- $P_n(i \mid C)$: Probability of individual *n* choosing route *i* from given choice set *C*;
- *G*: set of all non-empty subsets of *M*;
- $Q_n(C \mid G)$: Probability of individual *n*'s choice set being *C*.

Generally, there are three ways to introduce consideration sets into the discrete choice models. The first method is choice-set explosion method where the consideration set is

not deterministically constructed but all possible consideration sets are probabilistically constructed (Morikawa, 1996). In this situation, each possible consideration set has a probability to be the true consideration set. Therefore, the probability of chosen alternative is based on the probability of the potential consideration set to be the true consideration set and the probability of selecting the chosen alternative form the potential consideration set. The Manski' model is describing this situation and suitable to be applied when all possible consideration sets are considered in a probabilistic way. This method is best fitted when the number of alternative in the universal set is small.

The second method is constructing consideration sets in advance and then applying these consideration sets in the discrete choice model (Rashidi et.al, 2011). Then the discrete choice model would be directly applied in the choice making stage. The advantage of this method is that it can be applied with the huge number of alternatives in the universal set. However, the consideration set would be different by the generation method and the bias caused by the consideration generation should be considered in the discrete choice model estimation stage.

The third method is simultaneously estimating the consideration set formation process and choice making process. There are two situations in second method. The first situation is the consideration set is constructed deterministically in the consideration set formation stage and there is only one consideration set for each decision-maker (Gilbride and Allenby, 2004). In this method, although two stages are estimated simultaneously, the consideration set is determined in advance. Then, the choice probability would be calculated by the discrete choice models. This method would be applied with the big

universal set, the consideration set is generated according to decision-maker's thresholds on alternatives' attributes. However, different screening rules with different thresholds combinations would generate different consideration sets, and which screening rule is the best need to be justified. Furthermore, the estimation method is complex and it spends a lot of time for the model estimation.

In previous studies, the simulated data was preferred. Even the real data was used, the universal set size is not big enough or not all available alternatives were included in the universal set. Furthermore, there are rare studies related to comparing the different two-stage choice models with the same data. Different model has the different advantages, therefore, the comparison would tell us which one is suitable for the data. Most of all, in these studies, the changing of alternatives in the consideration set in different choice scenarios was not analyzed. In the consideration set formation stage, consumers' consideration set are formatted by the consumers' thresholds. However, even for the same consumer, the threshold would be changed if the consumer is in different choice scenarios. Therefore, in this dissertation, we would like to apply the real data in all studies. Different two-stage choice models with the same data are compared to find a better one. Finally, we also considered the different consumers' preference in the consideration set formation stage in different choice scenarios.

1.2 Research Objectives

Now, the issue is how to introduce the consideration set into the choice scenario in transportation filed. When there are a mass of available alternatives in the choice scenario, it is appropriate to introduce the consideration set into the choice analysis. If the number

of available alternative is moderately big, we could consider to calculate the probabilities of all possible consideration set to be the true consideration set. However, when the size of the universal set is extremely big, although we still could calculate such probabilities theoretically, the calculation time and precision would reject this way. Constructing a consideration set for each decision-maker in advance and using such consideration set in the following discrete choice analysis would ignore the size of the universal set. If the consideration set was separately constructed, then the common estimation method such as maximum likelihood would be applied in the choice analysis. However, if we want to generate the consideration set during the choice model estimation, the maximum likelihood method is not applicable anymore. The Bayesian approach and machine learning would be the optional method. However, which method is better is not known until we obtain the result from each method.

Therefore, in this studies, we aim to introduce the consideration set into the choice analysis in transportation research, especially when the size of universal set is huge. The following tasks will be proceed in this studies.

1) Applying consideration sets as choice set in transportation choice analysis including route choice and vehicle purchasing choice.

2) Based on the feature of different choice scenarios, choosing different consideration set formation methods and choice model estimation method. The consideration set would be generated in the model estimation process or in advance.

3) Comparing the models' performance which applying the universal set and consideration set separately, Justify the advantage of consideration set in some choice analysis.

4) Comparing different consideration set formation methods. See the differences of consideration sets under different formation methods.

5) Comparing the consideration set under different conditions but in same choice scenario. Finding how the consideration set will be changed under such conditions.

1.3 Research Outline

The outline of this studies would be described as follows. Chapter 1gives the background of this study and shows the research objectives. In chapter 2, we estimate the route choice model with the consideration set. Both the consideration set formation stage and choice making stage are estimated simultaneously with only the observed final choices. In chapter 3, we use hazard-based choice formation model to calculate the probability of each alternative vehicle to be considered, and then randomly select alternative vehicles into the consideration set based on these probabilities, and then apply the MNL model to estimate the vehicle choice behavior. In chapter 4, we applied the same data in chapter 3, but both choice set formation stage and choice making stage would be estimated simultaneously by a hierarchical Bayes model. Then we compare the model performance in chapter 2 and chapter 3. In chapter 5, we give a conclusion and show some future direction to improve the current work.

1.4 Reference

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

2. Bronnenberg, B. J., & Vanhonacker, W. R. (1996). Limited choice sets, local price response and implied measures of price competition. *Journal of Marketing Research*, 163-173.
3. Brown, J. J., & Wildt, A. R. (1992). Consideration set measurement. *Journal of the Academy of Marketing Science*, *20*(3), 235-243.
4. Gilbride, T. J., & Allenby, G. M. (2004). A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules. *Marketing Science*, *23*(3), 391–406. doi:10.2307/30036705
5. Hauser, J. R., & Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of consumer research*, 393-408.
6. Kardes, F. R., Kalyanaram, G., Chandrashekaran, M., & Dornoff, R. J. (1993). Brand retrieval, consideration set composition, consumer choice, and the pioneering advantage. *Journal of Consumer Research*, 62-75.
7. Mehta, N., Rajiv, S., & Srinivasan, K. (2003). Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing science*,*22*(1), 58-84.
8. Morikawa, T. (1996). A Hybrid Probabilistic Choice Set Model with Compensatory and Noncompensatory Choice Rules. Volume 1: Travel Behavior. In *World Transport Research. Proceedings of the 7th World Conference on Transport Research*.
9. Parkinson, T. L., & Reilly, M. (1979). An information processing approach to evoked set formation. *Advances in Consumer Research*, *6*(1), 227-231.
10. Paulssen, M., & Bagozzi, R. P. (2005). A self‐regulatory model of consideration set formation. *Psychology & Marketing*, *22*(10), 785-812.
11. Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human*

performance, *16*(2), 366-387.

12. Rashidi, T. H., J. Auld and A. Mohammadian (2012). "A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection." Transportation Research Part A: Policy and Practice 46(7): 1097-1107.
13. Sheffi, Y. (1985). Urban transportation network. *Equilibrium analysis with mathematical programming methods, Prentice Hall*.
14. Shocker, A. D., Ben-Akiva, M., Boccara, B., & Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing letters*, *2*(3), 181-197.
15. Wu, J., & Rangaswamy, A. (2003). A fuzzy set model of search and consideration with an application to an online market. *Marketing Science*, *22*(3), 411-434.

2. Analysis of driver's route choice behavior considering probability choice sets

2.1 Introduction

2.1.1 Background

Transportation has now become a critical issue in many big cities, with congestion leading to wasted time and fuel, air pollution and other economic losses. City administrators as well as drivers dream of an uncongested road network. Now, thanks to the technological development, advanced tools are available for use in tackling transportation problems, including the Global Position System (GPS), Intelligent Transportation Systems (ITS) and so on. In transportation applications based on these technologies, the concept of route choice plays a very important role.

Route choice is the process of travelers choosing routes. It can be applied to appraise travelers' perceptions of route characteristics, to predict future traffic conditions on transportation networks and to understand travelers' reactions and adaptations to sources of information (Prato, 2009). However, modeling route choice behavior is not as easy as it seems. It is unlike other choice situations in transportation modeling, such as mode choice and destination choice, where the number of alternatives is limited. For example, in mode choice, people might have to choose among bus, taxi, subway, walking, bicycle and driving. However, in a route choice situation, there may be a dense network of roads,

especially in a city, and even if the distance from origin to destination (OD) is only several kilometers, there could exist many thousands of routes for drivers to choose from. Obviously, it is unreasonable to assume that drivers will choose their route from a choice set that includes all routes connecting one OD pair. However, researchers are always lack of travelers' knowledge about the network composition, uncertain about travelers' perceptions of route characteristics and unavailable to exact information about travelers' preferences (Prato, 2009), therefore, how to properly define the choice set for drivers in route choice modeling is always an issue.

2.1.2 Literature Review

As already noted, the universal set of feasible routes between an OD pair might be very numerous; human limitations mean that no driver is likely to know all of them. But a driver may know some of the routes as a result of past driving experience and information from maps and navigation systems. This set of routes is called the awareness set, also called master set. However, a driver might not be choosing from this set of known routes on a particular trip because of certain spatiotemporal constraints that apply, such as a time budget, driving habits and so on. These constraints eliminate the availability of some of the known routes. The remaining known routes constitute a set known as the viable set, also called consideration set from which the driver ultimately makes a choice (Kaplan, 2012).

Early researchers (McFadden, 1981) always assumed that all choice-makers choose from the same choice set. Gaudry and Dagenais (1979) proposed the Dogit model, where an individual is either captive to one alternative or is free to choose from the full choice set.

Manski (1977) proposed the probabilistic choice set (PCS) model in which the choice decision process is divided into two parts: the choice set generation stage and the choice making stage. Some choice set generation models have been proposed in the past and incorporated into the PCS model in some choice modeling investigations (Swait and Ben-akiva, 1986 and Morikawa, 1996), however defining choice set formation in a probabilistic way is complex and has never been done in a full-size application (Frejinger, 2009)

The objective of this research is to model drivers' route choice behavior more accurately. The crucial part of the procedure in this analysis is the step in which the awareness set is reduced to the viable set. Since choice set generation is treated separately, the PCS model is applied. A constraint-based choice set generation model is used to model a driver's choice set generation procedure. Additionally, in the en-route route choice situation, constraints should be varied with the stage of the trip. An en-route variable is introduced to the choice set generation model to see the difference. On the other hand, in the choice making stage, the conventional discrete choice model is used. There is one problem with this approach that should be mentioned: if the number of alternatives is *n*, then the number of all non-empty subsets would be $2^n - 1$. With a large *n*, $2^n - 1$ would increase exponentially to an enormous number. In order to solve the computational problem resulting from this total number of non-empty subsets, a pairwise comparison of alternative methods proposed by Morikawa (1996) is used.

The data used in this research is probe-vehicle data collected in Toyota city, Japan between Feb 2011 and Dec 2011. The two model stages are estimated simultaneously

using information about drivers' actual chosen routes only, as provided by the probe data.

2.2 Models

The various approaches lead to different models of route choice behavior; Figure 2.1 illustrates the situation (Frejinger, 2009):

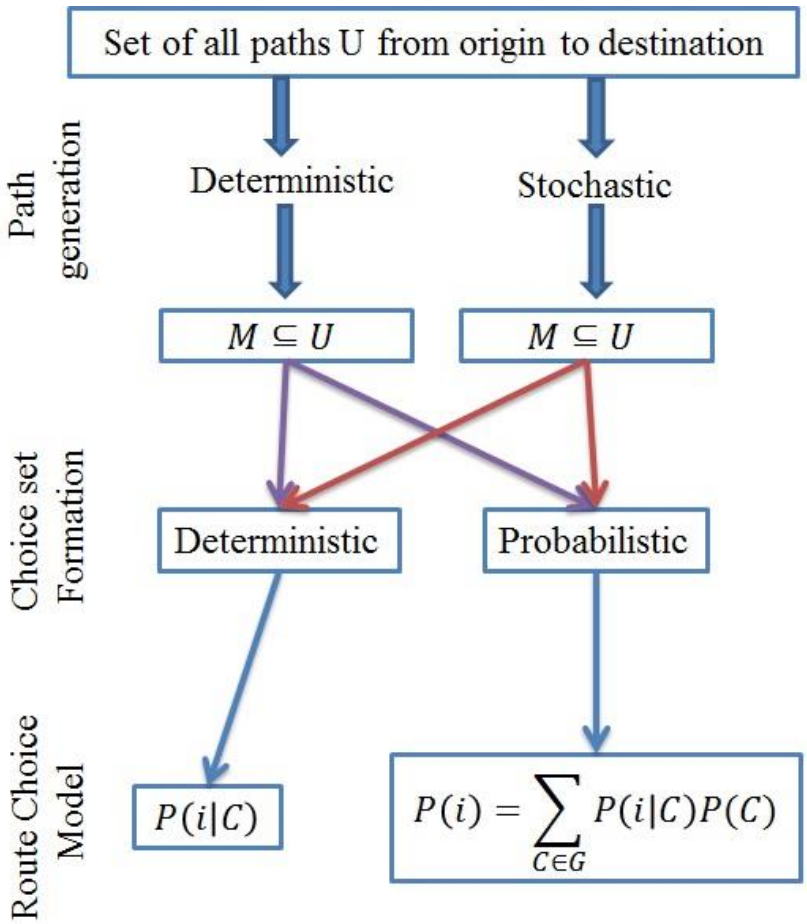

Figure 2-1: Route choice modeling

In this diagram, *U* represents the universal set that includes all possible routes for an OD pair, *M* represents the master set of known routes generated by the researcher using deterministic or stochastic methods in order to approximate a driver's awareness set. And

then, if the choice set generation method is deterministic, the probability of a driver choosing route *i* simply equals $P_n(i|C)$ where C is the individual final viable choice set (and where $C \subseteq M$). On the other hand, if the choice set generation method is a probabilistic approach, the probability would be as shown in the figure, where *G* represents all the non-empty subsets of *M*. A specific explanation of the different models in different situations is given below.

Random utility models are always applied to the route choice modeling, however, there is an issue in route choice should be noted is the similarities between the alternative routes for an OD pair, or so called overlapping problem. Route overlapping should be considered in the route choice model, however, MNL model does not take route overlapping into account. Therefore, some models have been proposed to correct for route overlapping, such as C-Logit model (Cascetta, et al., 1996), Path-size Logit model (Ben-akiva and Bierlaire, 1999). These models are extensions of the MNL model by introducing a factor accounting for the overlapping.

$$CF_i = \ln \sum_{j \in C} \left(\frac{L_{ij}}{\sqrt{L_i L_j}} \right)^{\gamma} \qquad [2.1]$$

Equation (2.1) is the commonality factor (CF) in C-Logit model, where L_{ij} is the length of links common to route *i* and *j*, while L_i and L_j are the lengths of route *i* and *j* respectively. γ is a positive parameter and the summation is extended to all routes belonging to *C* including route *i*. Then the choice probability for C-Logit model is :

$$P_n(i) = \frac{exp\left(V_{in} + \beta_{CF} * CF_i\right)}{\sum_{j \in M} exp\left(V_{jn} + \beta_{CF} * CF_j\right)} \quad [2.2]$$

Path-Size Logit introduce a size variables to account for the overlapping issues:

$$S_i = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C} \delta_{aj} \frac{L_C^*}{L_j}} \quad [2.3]$$

and Γ_i is the set of links in route *i*, l_a and L_i are the length of link *a* and route *i* respectively; δ_{aj} is the link-path incidence variable that is one if link *a* is on route *j* and 0 otherwise; and L_C^* is the length of the shortest route in choice set *C*. And then, the choice probability in Path-Size Logit model is:

$$P_n(i) = \frac{\exp\left(V_{in} + \ln S_i\right)}{\sum_{j \in C} \exp\left(V_{jn} + \ln S_j\right)} \quad [2.4]$$

2.2.1 Probabilistic Choice Set Model

As noted above, there may be a huge number of possible routes for an OD pair and a driver would be unable to consider all of them. This means that it is unreasonable to treat the universal set of routes as the choice set for drivers. Researchers therefore use deterministic or stochastic route generation techniques to create a master set of known routes, *M*, that approximates the driver's awareness set. Such master sets, however, may include many non-viable routes since current route generation techniques do not account for individual spatiotemporal constraints such as existing knowledge of routes, driving habits, route preference and so on (Kaplan, 2012). In this research, the Probabilistic Choice Set (PCS) model proposed by Manski (1977) is introduced as a way to overcome this shortcoming, the model's description is shown in Equation (1.4).

In this model, decision making is taken to be a two-stage process. The first stage is a choice set generation stage; decision makers generate their personal choice sets from the given master set under certain constraints. These constraints are conjunctive, which means that an alternative becomes part of the choice set only if it satisfies all constraints; otherwise it is excluded. The second decision-making stage is a discrete choice model; for this, both the normal and modified MNL models are applied in this research.

Constraint-Based Choice Set Generation Model

Several choice set generation models are available for application in different choice situations. A model in which decision makers face the situation of choosing from a subset with only a single alternative or from the full set was developed by Ben-Akiva (1977). Swait (1984) proposed a choice set generation model in which the choice set has a defined maximum size, reflecting the understanding that the human ability to process information is limited. Richardson (1982) developed a model in which choice set generation is treated as a search process, meaning that a decision maker examines the observed choice results in the choice set only when a final choice is made. In this work, the constraint-based choice set generation model (Swait and Ben-Akiva, 1986) will be applied.

In the route choice context, as already noted, it is impractical to view the universal set as the choice set. Furthermore, it is a reasonable assumption that, before choosing a route, a driver would generate a personal feasible choice set based on the information available to him/her and his/her preferences. This choice set generation stage could be seen as the elimination of alternatives by independent constraints, where the constraints are

conjunctive. The model is as follows:

$$q_n(i) = \prod_{k=1}^{K} q_{kn}(i) \qquad [2.5]$$

where:

- $q_n(i)$: the probability of alternative *i* being included in the choice set of alternative *n*;
- $q_{kn}(i)$: the probability of alternative *i* satisfying the *k-th* constraint for individual *n*.

If we assume that satisfying the constraint means that a latent variable exceeds a certain threshold value, the latent variable has the following structure:

$$E_{kn}(i) = \alpha_k * w_{in} - \delta_{in} \qquad [2.6]$$

where:

- α_k : vector of unknown parameters to be estimated;
- w_{in} : vector of variables affecting the constraints;
- δ_{in} : disturbance.

Then, $q_{kn}(i)$ can be expressed as:

$$\begin{aligned} q_{kn}(i) &= Prob(E_{kn}(i) \geq \mu_k) \\ &= Prob(\alpha_k * w_{in} - \delta_{in} \geq \mu_k) = Prob(\delta_{in} \leq \alpha_k * w_{in} - \mu_k) \end{aligned} \qquad [2.7]$$

In this equation, μ_k is the threshold value of the *k-th* constraint. If we assume δ_{in} follows a logical distribution, Equation (2.6) can be rewritten as:

$$q_{kn}(i) = \frac{1}{1 + e^{-(\alpha_k * w_{in} - \mu_k)}} \qquad [2.8]$$

$$q_n(i) = \prod_{k=1}^{K} \frac{1}{1+e^{-(\alpha_k * w_{in} - \mu_k)}} \qquad [2.9]$$

Within the meaning of $q_n(i)$, $Q_n(C|G)$ in Equation (4) can be written as:

$$Q_n(C|G) = \frac{1}{1-Q_n(\varnothing)} \times \prod_{i \in M} \left[q_n(i)^{d_{iC}} \{1-q_n(i)\}^{1-d_{iC}} \right] \qquad [2.10]$$

Where:

- $Q_n(\varnothing)$: $\prod_{m=1}^{M} 1-q_n(m)$, the probability of individual *n*'s choice set being empty;
- d_{iC}: dummy variable with a value of 1 if the alternative *i* is in individual *n*'s choice set *C*, and otherwise 0.

As mentioned before, the choice making stage in the PCS model is treated as a discrete choice situation, meaning that the MNL model can be applied. Combining Equations (2.5) and (2.11), we obtain the following:

$$\begin{aligned} P_n(i) &= \sum_{C \in G} P_n(i|C) * Q_n(C|G) \\ &= \frac{1}{1-Q_n(\varnothing)} \times \sum_{C \in G} \left\{ \frac{e^{V_{in}}}{\sum_{h \in C} e^{V_{hn}}} \times \prod_{j \in M} \left[q_n(j)^{d_{jC}} \{1-q_n(j)\}^{1-d_{jC}} \right] \right\} \end{aligned} \qquad [2.11]$$

Equation (2.12) expresses the probability of choosing alternative *i* in the PCS model after introducing Swait and Ben-akiva's choice set formation model. This equation can be applied directly when the number of alternatives in the master set is few, such as under five. However, when the number of alternatives increases, the size of the possible choice set, or the number of elements in *G*, increases exponentially and the direct evaluation of

Equation (2.12) becomes virtually impossible (Morikawa, 1996).

Morikawa (1996) proposed a method for solving this exponential problem as follows. Adopting the method of pairwise comparison of alternatives in terms of utility, if individual *n* prefers alternative *i* to alternative *j* in the PCS model, there are two possible scenarios: (1) both *i* and *j* belong to individual *n*'s consideration choice set, and for this individual the utility of alternative *i* is greater than the utility of alternative *j*; or (2) alternative *i* is in individual *n*'s choice set while alternative *j* is not. These two scenarios can be expressed mathematically as follows (Morikawa, 1996):

$$P_n(i) = \frac{1}{1-Q_n(\varnothing)} \times Prob(i \in C_n) \times Prob\left[\begin{array}{c} \{(1 \in C_n) \cap (U_{in} \geq U_{1n})\} \cup \{1 \notin C_n\} \\ and \\ \{(2 \in C_n) \cap (U_{in} \geq U_{2n})\} \cup \{2 \notin C_n\} \\ and \\ \cdots \\ and \\ \{(J \in C_n) \cap (U_{in} \geq U_{Jn})\} \cup \{J \notin C_n\} \end{array}\right] \quad [2.12]$$

$$= \frac{1}{1-Q_n(\varnothing)} \times q_n(i) \times prob\left[\bigcap_{j \in M, j \neq i} \{(j \in C_n) \cap (V_{in} + \varepsilon_{in} - V_{jn} \geq \varepsilon_{jn})\} \cup \{j \notin C_n\}\right]$$

where:

- C_n : latent choice set for individual *n*;
- U_{in} : utility of alternative *i* for individual *n*;
- V_{in} : systematic component of utility of alternative *i* for individual *n*;
- ε_{in} : disturbance of utility of alternative *i* for individual *n*.

taking the conditional probability on the disturbance component of utility ε_{in}, Equation (2.12) can be written as:

$$P_n(i) = \frac{q_n(i)}{1 - Q_n(\varnothing)} \times \int_{-\infty}^{+\infty} f(\varepsilon_{in}) \times \prod_{j \in M, j \neq i} \left\{ q_n(j) \times F(V_{in} - V_{jn} + \varepsilon_{in}) + (1 - q_n(j)) \right\} d\varepsilon_{in} \quad [2.13]$$

where:

- $f(\bullet)$: probabilistic density function of ε;
- $F(\bullet)$: cumulative density function of ε.

if the ε is assumed to follow a Gumbel distribution, then

$$P_n(i) = \frac{q_n(i)}{1 - \prod_{m \in M} \{1 - q_n(m)\}} \times \int_{-\infty}^{+\infty} e^{-\varepsilon_{in}} e^{-e^{-\varepsilon_{in}}} \prod_{j \in M, j \neq i} \left\{ q_n(i) e^{-e^{-V_{in} + V_{jn} - \varepsilon_{in}}} + (1 - q_n(j)) \right\} d\varepsilon_{in} \quad [2.14]$$

then, the log-likelihood of modelling the choice of route *i* from the master set for *N* travelers can be written as:

$$LL = \sum_{n=1}^{N} \log\left(P_n(i)\right) \quad [2.15]$$

Since the model does not have a closed-form expression due to the integral in the choice probability function, here we apply the maximum simulated likelihood with 300 standard Halton draws (Train, 2003).

2.3 Data

The data used in this research is probe vehicle data collected in Toyota city, Japan. The data was collected from private vehicles between February and December 2011. The raw data received from the vehicles was restructured such that each trip was divided into multiple segments representing the links in the trip. Table 2.1 shows the data structure.

Table 2.1: Data structure

Data	Description
User ID	Unique driver ID
Trip ID	Unique trip ID
Trip Sequence	Number Representing link order in trip
DRM Link Mesh Code	Mesh code for the area link belongs to
DRM Node1 ID	The ID of one node of the link
DRM Node2 ID	The ID of the other node of the link
DRM Link Direction Flag	If direction is from Node1 to Node2 = 0, otherwise = 1.
Link Travel Start Time	Travel start time on the link
Link Stay time	Travel time on the link
Day of the week	From 1 to 7 representing Monday to Sunday
Hour of Day	From 1-24

Figure 2.2 shows probe vehicle trips in the area of Toyota city, where the different road link colors represent the accumulated number of probe vehicle passes on these links over one month. (The figure shows data for June 2011.) In order to get enough data to support this research, we select the black rectangle area in Figure 2.2 as the research target area as this is where probe vehicles passed most frequently.

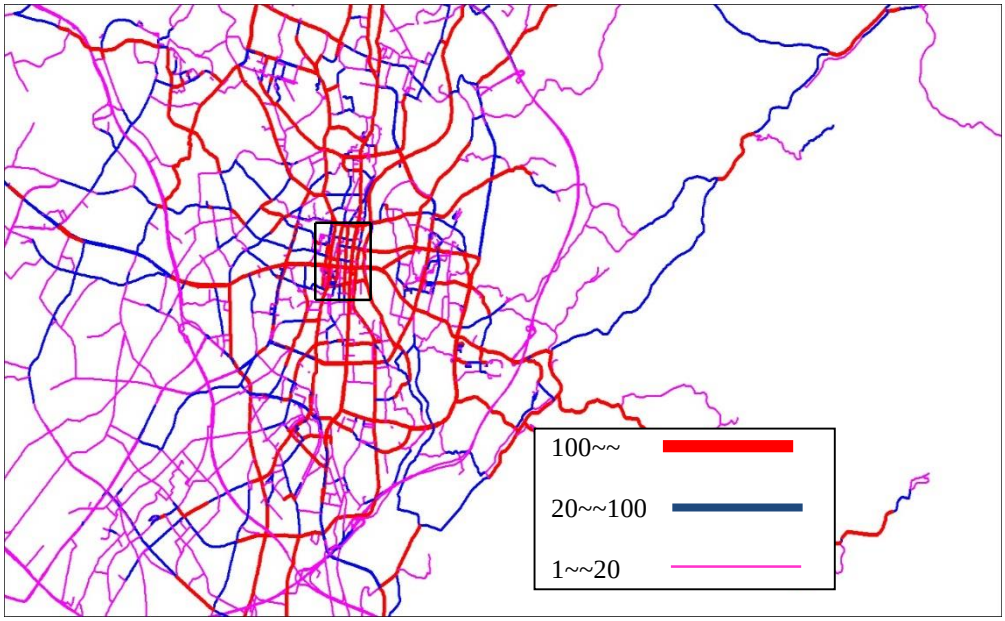

Figure 2-2: Accumulated number of probe vehicle passes on the road network of Toyota city

The rectangular target area is shown in detail in Figure 2.2. Trips are selected for inclusion where the vehicle passes through the target area via nodes on the diagonals, represented by A-C and B-D. It should be noted that vehicles passing through the target area are on longer trips, so the actual origins and destinations of these trips are beyond the target area; therefore, we denote A-C and B-D the en-route origin-destination points to distinguish them from the full trip's origin-destination points. Figure 2.3 shows one typical trip passing through the target area from node A to node C, with the whole trip shown in blue. This research is concerned with only the part of the trip through the target area. Using the data for such trips, we analyze different driving behavior within the target area.

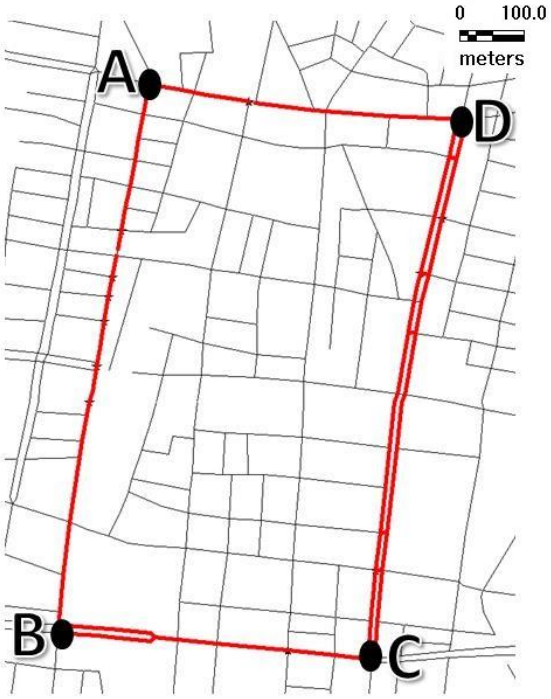

Figure 2-3: Definition of research target area

For these trips, we first analyze driver behavior in the target area. For each trip, we need to obtain the following information:

- Distance (km): the distance traveled on each trip within the target area
- Arterial road ratio (ARR): the ratio of distance traveled on arterial roads within the target area, ranging from 0 to 1
- Turns: the number of turns made within the target area

we then number each of the unique routes within the target area; each trip using the same route would have the same route number. Next, we obtain other information for each trip for use in estimating the models:

- Trip distance (TD): the total travel distance for each trip
- DO: the travel distance from each trip's actual origin to the en-route origin A or B

- Route ID: the number of the route used through the target area for each trip

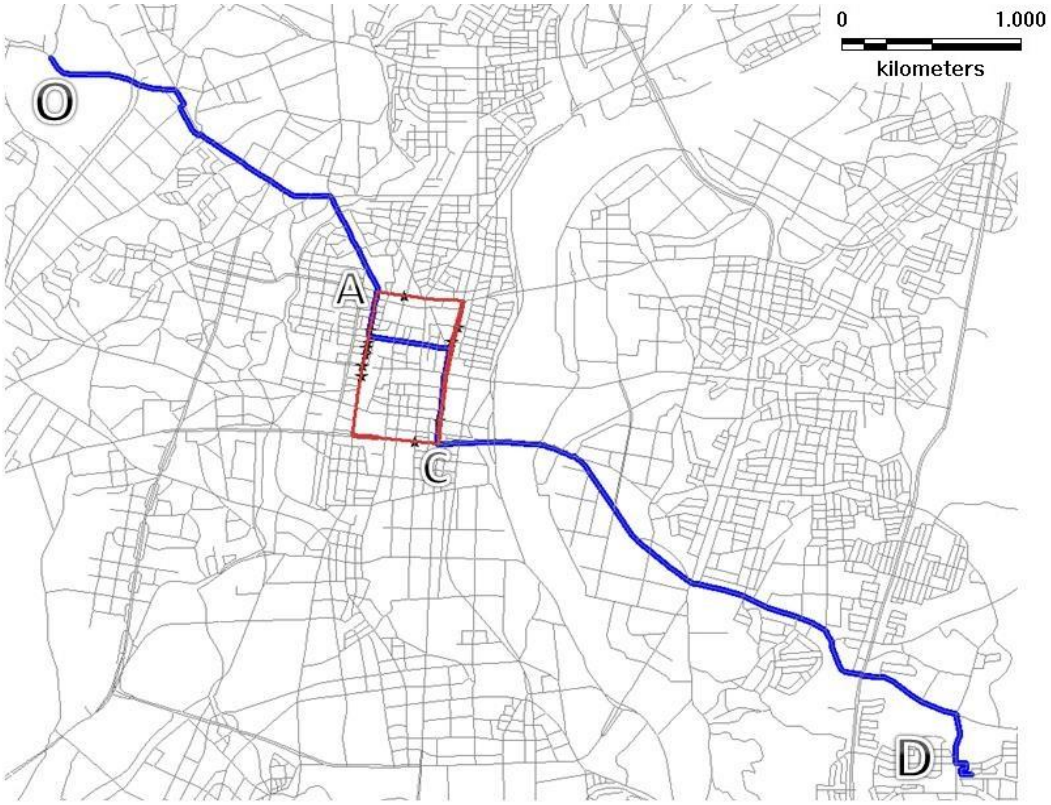

Figure 2-4: A typical trip passing the target area

Figures 2.5 and 2.7 show driver characteristics for the two en-route OD pairs, respectively. For en-route OD-1 (origin A to destination C) there were 305 trips and 16 observed different routes. The number of drivers was 42, of whom 37 were males, and the age range was from 26 to 64. Figure 2.6 shows the distribution of the total trip distance for vehicles passing through the target area by en-route OD1. Most of them are around 5-10 kilometers and few are longer than 25 kilometers.

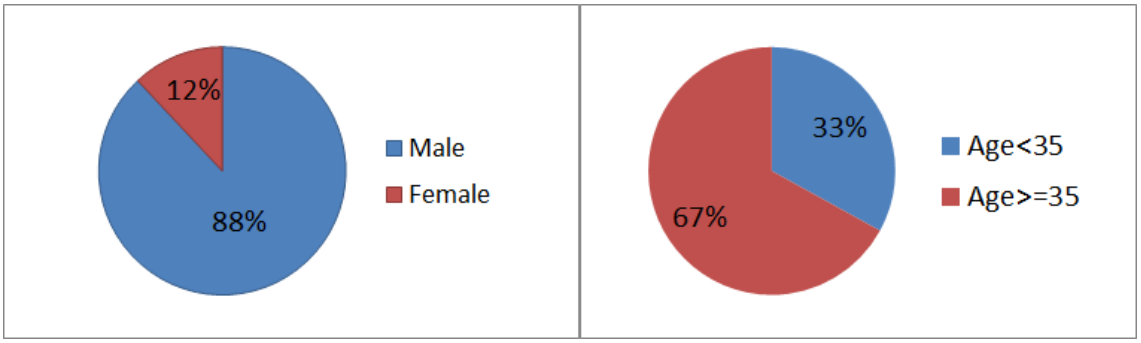

Figure 2-5: Driver characteristics for en-route OD-1

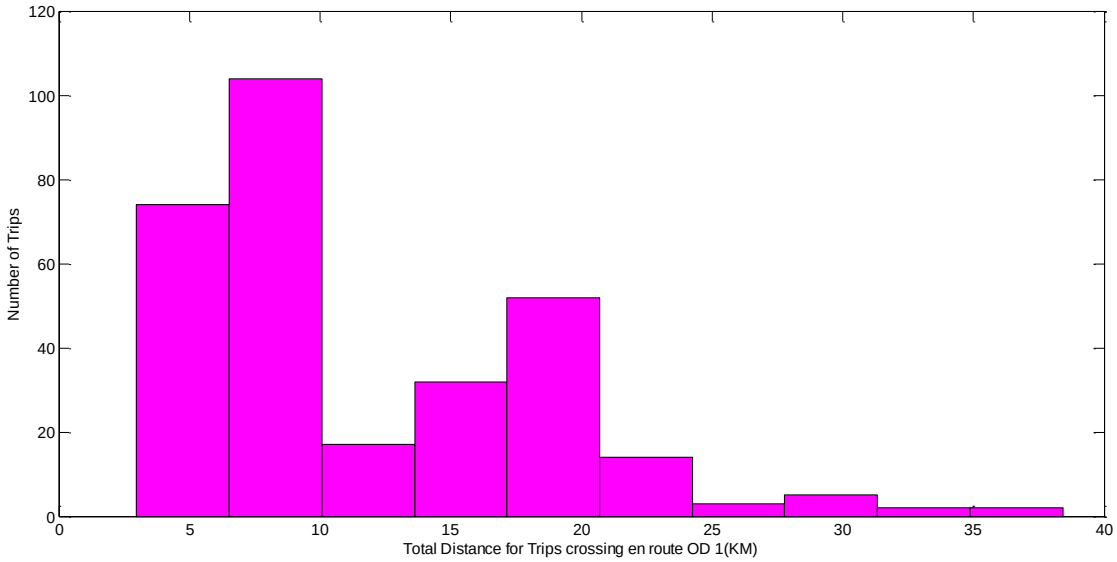

Figure 2-6: Total trip distance distribution for vehicles passing through the target area by en-route OD-1

Similarly, for en-route OD-2 (origin B to destination D) there were 144 trips and the number of observed different routes was 10. These trips were made by 28 drivers, of whom 21 were male, and the age range was from 23 to 58. From the total trip distance distribution in Figure 2.8, it is clear that trips of less than 10 kilometers account for the greater part and there were also many trips in the distance range from 10 to 20 kilometers. There were few trips longer than 35 kilometers.

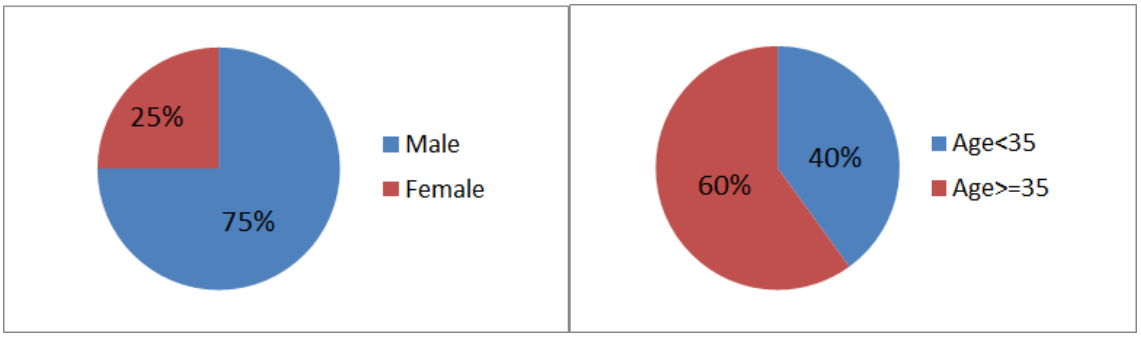

Figure 2-7: Driver characteristics for en-route OD-2

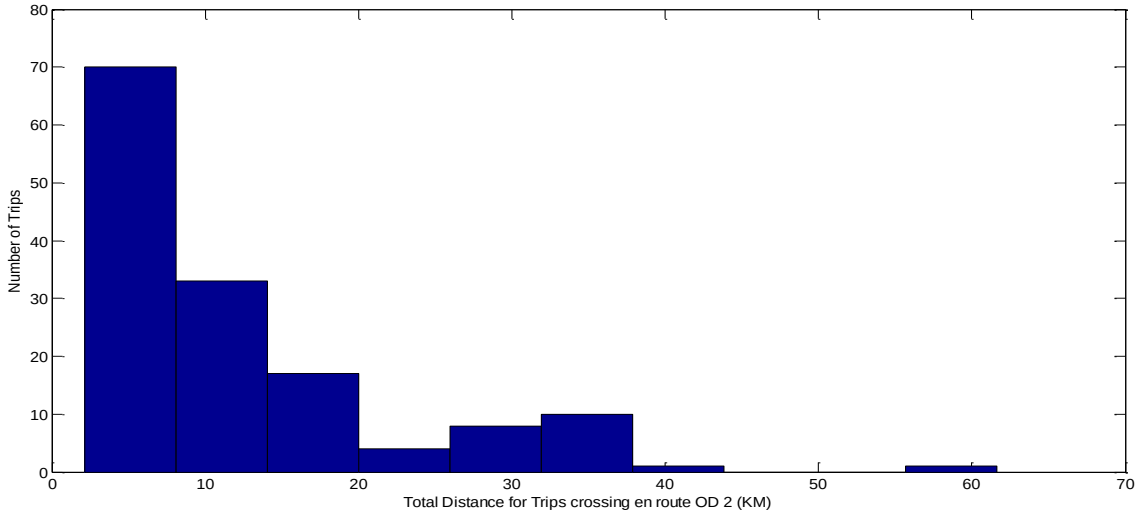

Figure 2-8: Total trip distance distribution for vehicles passing through the target area by en-route OD-2

2.4 Estimation Results and Discussion

In this research, we will apply PCS model to modeling route choice behavior. All the observed routes compose the master sets for two OD pairs respectively, and then the number of turns for each route will be the threshold for drivers to format their viable set before making the choices. Furthermore, as these trips in the study area are part of the

whole trips, the influence of the en-route positions to the viable sets formation will also be analyzed.

Firstly, C-Logit model, the extension of MNL model, will be estimated and the estimation result will be the benchmark for other models. Then, PCS model will be estimated, and the probability for each route to be included in the viable set is:

$$q_n(i) = \frac{1}{1+e^{-(\beta * turns - Threshold)}} \quad [2.16]$$

turns is the number of turns for route *i* in the target area, β and *threshold* are the parameters to be estimated. In the third model, named en-route PCS model, the process indicator, PI=*DO/TD* will be introduced, PI is an indicator which can reflect the trips' rate of progress. And then this PI will be added in the choice set formation stage to see its influence. And the probability for each route to be included in the viable set is:

$$q_n(i) = \frac{1}{1+e^{-\left(\beta * Turns + \alpha * (PI)^2 + \gamma * PI - Threshold\right)}} \quad [2.17]$$

Details about Equation (2.18) will be discussed later. It should be argued that both in PCS and En-route PCS model, *CF* are calculated based on all observed routes for each OD respectively, same as the way in C-Logit model.

Table 2.2: Estimation Results with en-route OD-1 (305 samples) and en-route OD-2 (144 samples)

	`C-Logit	PCS	En-route PCS
	Choice Set Formation Stage		
β	---	-1.383 (-5.2)	-1.193 (-11.3)
α	---	---	13.247 (.4.2)
γ	---	---	-9.289 (-3.7)
Threshold	---	-1.842 (-2.9)	-3.224 (-2.9)
	Choice Making Stage		
Distance(km)	-7.434 (-5.5)	-17.673 (-2.1)	-16.926 (-1.8).
ARR	1.632 (3.3)	1.722 (1.4)	5.182 (2.6)
Turns	-1.68 (-9.4)	---	---
CF	1.134 (2.6)	1.372 (1.2)	1.785 (1.6)
LL0	-1177.212	-1177.212	-1177.212
LL	-745.625	-734.927	-728.431
$\overline{\rho^2}$	0.363	0.372	0.375
AIC	1499.25	1479.85	1470.86

t-statistic value in the parentheses

The first column in Table 2.2 shows the estimation result of C-Logit model. The probability of selecting a certain route from the choice set increases with a decrease in (1) distance and (2) number of turns. The propensity to choose a certain route from the choice set increases with an increase in the ratio of the route that uses an arterial road. This

estimation result is consistent with what we would expect a rational driver to do when making a route choice.

Second column is the estimation result of PCS model. It should be noted that the number of turn variable has been excluded in the choice making stage; this is because it was used as a constraint in the choice set generation stage and the two stages are estimated simultaneously, therefore in order to avoid interference, it is dropped in the choice making stage. In the choice set generation stage, the probability of a route being included in the driver's choice set decreases with increasing number of turns along the route. In the choice making stage, the signs of the two variables are the same as in the MNL model. Here, though, the statistical significance (t-statistics) of the distance parameter is smaller than in the MNL model. This might indicate that, if the choice set is properly considered, the influence of explanatory variables in the choice making stage is reduced. It also means that parameter estimation in a model that does not consider the choice set generation stage may include biases. Furthermore, Akaike's Information Criterion (AIC) value of the PCS model is smaller than that of MNL model, which indicates that the PCS model is preferred in this situation. It also means that taking route choice behavior to be a two-stage process of choice set generation and then choice making is reasonable decision.

Research on probabilistic choice set analysis in route choice behavior usually assumes that the driver chooses a route from a latent choice set that is composed under certain constraints before starting the trip. However, drivers prefer to choose only a partial route, especially when the trip is long and the route network is complex. Furthermore, the constraints used in the choice set generation stage are different for different stages of the

trip. This can be investigated if we define a variable reflecting how much of the trip has been completed: we take the total distance from origin to destination to be TD and the distance from the origin to a particular point en-route to be DO, then define DO/TD as a process indicator (PI) that ranges from 0 to 1 according to how much of the trip has been completed. The influence of PI in the choice set generation stage is not likely to be simply linear or exponential. We might clarify its effects by considering the following example.

In Figure 2.9, driver T's house is located at point A and his destination is point B, a shop in the city center. The purple route is suggested by Google Maps. This route can be treated as comprising three segments, where the first is from T's house to a major road; the second is along major roads until near the destination; and the final segment is from the major road to the destination. Drivers are more sensitive to the number of turns in the second segment than in the first and final segments. That is, driving habits and route preference are not the same throughout the trip.

Figure 2-9: Example route from T's house to a city center destination

Figure 2.10 shows the relationship between number of turns in the research area and the vehicles' PI from all data. It is clear that drivers have different attitudes to the number of turns at different stages of the trip. At the beginning of a trip, a driver accepts that some turns are necessary. Later, acceptance of turns falls and the driver would prefer as few turns as possible. Finally when nearing the destination, acceptance of turns increases. That is, drivers' acceptance of turns during a trip is like a convex function. Therefore, the probability for route to be included in the choice set in en-route PCS model is constructed as Equation (2.18).

The estimation result for this improved model is shown in third column in Table 4.1 After introducing one variable, the explanatory power of the model is improved. Further, the χ^2 statistic for the likelihood of the joint hypothesis that $\alpha = 0$ and $\gamma = 0$, -2*(-734.927 + 728.431) = 13.0, which, with two degrees of freedom, is greater than the critical value of 10.597 at the 0.005 significance level, brings us to the same conclusion.

This can also be seen in Figure 2.11, which explains how the process indicator affects the probability of a route being in the latent choice set. Firstly, at the beginning of a trip, the probability of a route being selected for the driver's latent choice set becomes lower with distance. Then there is a stage of the trip where the probability does not change much, before it rises again continuously to the end. It is clear that a driver is more tolerant of turns the closer he or she is to the destination.

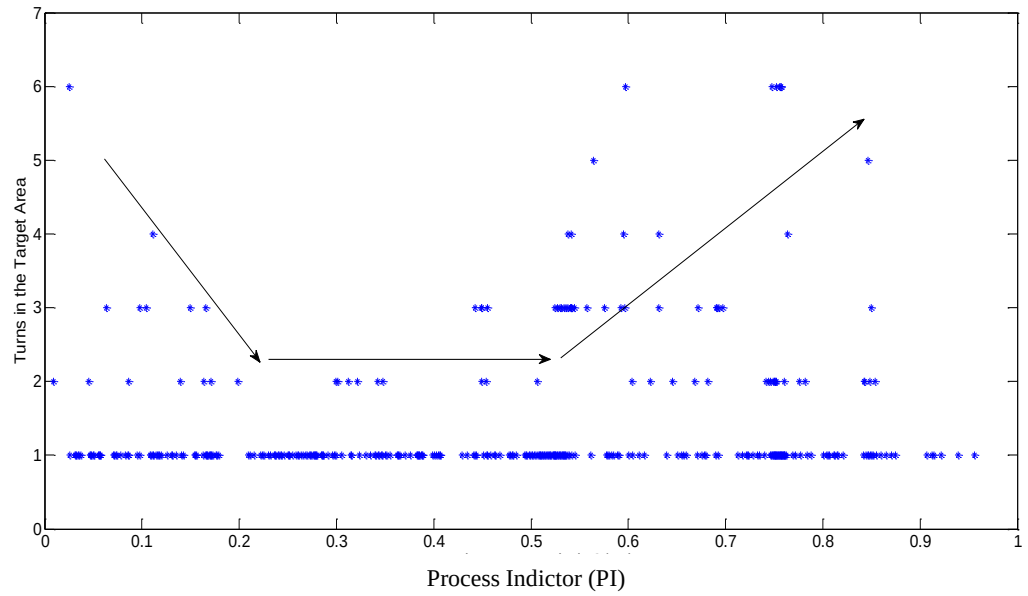

Figure 2-10: Relationship between number of turns in target area and Process Indicator

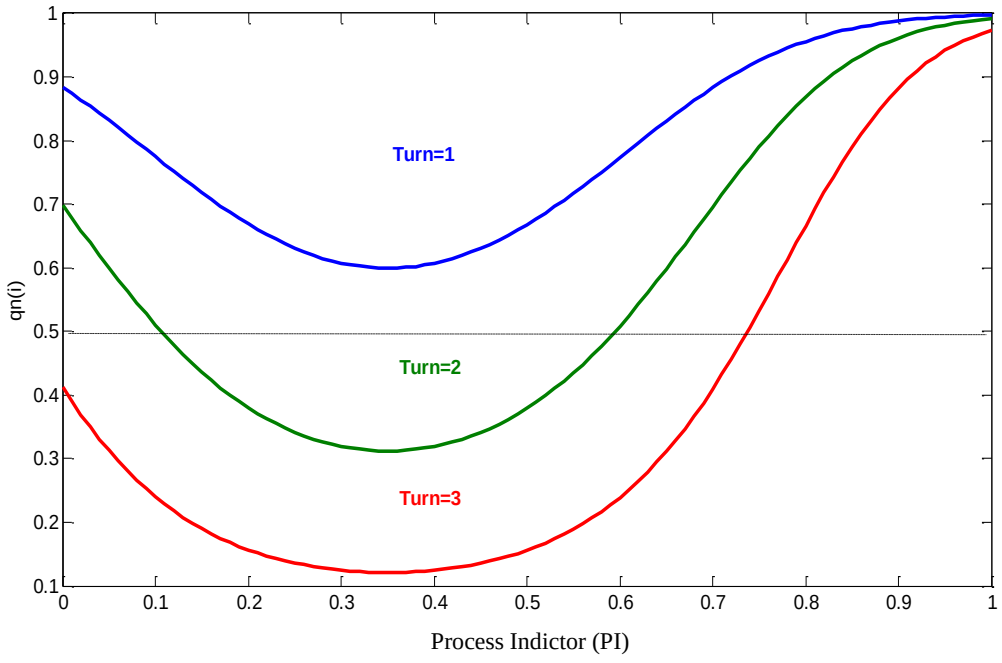

Figure 2-11: Influence of en-route variable on

2.5 Conclusions and further research

Route choice behavior can be treated as a two-stage procedure. The first stage is generation of a feasible choice set from the master set, in which the driver selects routes that fit within his or her spatiotemporal constraints. The second stage is making a choice from among the routes in the feasible choice set. In this work, the two-stage PCS model (Manski, 1977) is used to analyze route choice behavior based on data obtained from probe vehicles.

For the non-compensatory choice set generation stage, the constraint-based choice set generation model (Swait and Ben-Akiva, 1985) is used to calculate the probability of each route being in the driver's feasible choice set, leading to the probability of each subset of the driver's awareness set. Then in the compensatory choice making stage, drivers are assumed to choose the route with the maximal utility, so the random utility model is applied. The models for the two-stages are estimated simultaneously using only information about drivers' actual choices as obtained from the probe vehicle data. The estimation results show that the PCS model offers a significantly better fit to the data than the MNL model.

A process indicator is introduced to represent the proportion of the trip distance already traveled. This is used to analyze changes in the probabilistic choice sets according to how much of the trip has been driven. The estimation results in this case are as expected and demonstrate that the spatiotemporal constraints in the choice set generation stage fluctuate according to the stage of the trip (from origin to a major road; along major roads;

and from major road to destination). This must also influence the probability of an alternative being selected to be in the individual's feasible choice set.

There are many consideration set heuristic rules, including disjunctive, conjunctive, sub-conjunctive and so on. There is no comprehensive research to indicate which rule decision makers will applied in route choice situation. Therefore, in the future, I would like to test different rules and find out which one would give the best result.

2.6 References

1. Ben-Akiva, M., & Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, *12*(1), 9-24.

2. Ben-Akiva, M., & Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science* (pp. 5-33). Springer US.

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

4. Cascetta, E., Nuzzolo, A., Russo, F., & Vitetta, A. (1996, July). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In *Proceedings of the 13th International Symposium on Transportation and Traffic Theory* (pp. 697-711). Oxford, NY, USA: Pergamon.

5. Frejinger, E., Bierlaire, M., & Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*,*43*(10), 984-994.

6. Gaundry, M. J., & Dagenais, M. G. (1979). The dogit model. *Transportation Research Part B: Methodological*, *13*(2), 105-111.

7. Kaplan, S., & Prato, C. G. (2012). Closing the gap between behavior and models in route choice: The role of spatiotemporal constraints and latent traits in choice set formation. *Transportation Research Part F: traffic psychology and behaviour*, *15*(1), 9-24..

8. Manski, C. F. (1977). The structure of random utility models. *Theory and decision*, *8*(3), 229-254.

9. McFadden, D. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, S13-S29.

10. Morikawa, T. (1996). A Hybrid Probabilistic Choice Set Model with Compensatory and Noncompensatory Choice Rules. Volume 1: Travel Behavior. In *World Transport Research. Proceedings of the 7th World Conference on Transport Research*.

11. Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, *2*(1), 65-100.

12. Richardson, A. (1982). Search models and choice set generation. *Transportation Research Part A: General*, *16*(5), 403-419.

13. Sheffi, Y. (1985). Urban transportation network. *Equilibrium analysis with mathematical programming methods*, *Prentice Hall*.

14. Swait Jr, J. D., *Probabilistic choice set generation in transportation demand model.* In **Dept. of Civil Engineering, MIT**. MIT: Cambridge, MA, 1984.

15. Swait Jr, J. D., & Ben-Akiva, M. (1986). *Constraints on individual travel behavior in a Brazilian city* (No. 1085).

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

press.

3. Analyses on choice set appropriateness under vehicle purchasing situation

3.1 Introduction

Vehicle choice behaviors has been an important research topic in transportation area for a long time. Auto makers would like to manufacture a vehicle whose features satisfy consumers' anticipations. Policy makers would like to formulate a policy to lead vehicle consumers' preferences in order to solve some social and economy issues, such as high gasoline price, congestion and so on. Since the introduction of discrete choice theory, multinomial logit model (MNL) were widely used in vehicle choice behavior analysis (Lave and Train, 1979; Manski and Sherman, 1980; Mannering and Winston, 1985; Kitamura et al, 2001; Baltas and Saridakis, 2013; Mabit, 2014). These outstanding past studies analyzed vehicle choice behaviors in different time and different areas, and in these studies, vehicles in choice sets were either categorized by function and size, or randomly selected from a large set, or included all available vehicle types in the market. However, in discrete choice model, the choice set has a significant influence on model's performance (Ben-Akiva and Lerman, 1985), therefore in this study, we would like to investigate a more reasonable choice set formation method in discrete vehicle choice behavior analysis.

Generally, the set including all available alternatives is named the universal set. In vehicle choice analysis, the universal set means the set including all vehicle types in the car market. It is not an easy work to collect such information about all vehicle types,

or even if such information is available, researchers will either directly apply the universal set as the choice set or randomly formulate a subset of the universal set to represent the choice set. These methods would have the following shortcomings. Firstly, there are always hundreds of vehicle types in the market which means the size of universal set will be huge. Actually, due to the limitation of time and ability, when consumers intend to purchase a vehicle, they will not browse all vehicle types but some of them. Secondly, although using randomly selected subset would shorten the model estimation time and provide consistent estimation result (Ben-Akiva and Lerman, 1985), it is still doubtful whether the randomly selected subset would represent the choice set which consumers actually evaluate and choose the final choice from. A subset of the universal set which decision-makers actually screen is named as "consideration set" which "purposefully constructed and can be viewed as consisting of those goal-satisfying alternatives salient or accessible on a particular occasion" (Shocker et al., 1991). Decision-makers would have specific requirements on some alternatives' attributes, and these alternatives which satisfy such requirements will be included into decision-makers' consideration sets. Asking each decision-maker to describe their screening rules is a method to get the information of consideration set, however the work should be processed in the data collection step, and such work is always difficult and costly. Therefore, researcher would like to use the only information of the universal set and decision-makers' final choice to model the consideration set formation process, and then apply it into the discrete choice model.

In this study, we will incorporate consideration set into vehicle purchasing behavior analysis in Japan. Although only the final decision information is available, in the consideration set formation stage, we calculated the probability that the vehicle' price

and fuel cost would satisfy decision-makers' requirements. And then, alternative vehicles were randomly selected to comprise the consideration set, the vehicle with higher satisfying probability will have higher chance to be included into the consideration set. Finally, the consideration set will be treated as the choice set for each decision-maker, and a MNL model incorporating a correction term will be used to analyze the vehicle purchasing behavior in Japan. In order to evaluate the advantage of using consideration set in vehicle purchasing analysis, results of a MNL model estimated with the universal set and a MNL model estimated with the randomly selected alternatives were obtained. Furthermore, we also simulate when the gasoline price is changed, how the average gasoline vehicles' fuel efficiency will be influenced. Simulation results of three estimated models will be provided together.

The rest of this paper is organized as follows. Section 3.2 reviews the past researches about the vehicle choice. Section 3.3 discusses consideration set deeply. Section 3.4 describes the data applied in this study, section 3.5 explains the consideration set formation model in this study and the related estimation result are also presented. Section 3.6 provides estimation results of three MNL models using different choice set. The performance of three models are also compared by some disaggregate validation method. In section 3.7, the simulation result shows the impact of gasoline price on the average gasoline vehicles' fuel efficiency. Section 3.8 concludes this study and points some shortages of this study.

3.2 Literature review

Most published studies of vehicle type choice have examined disaggregate choice models, such as multinomial logit (MNL) and nested logit (NL) models. Lave and Train

(1979) developed an MNL model for estimating the vehicle type choices of 541 new-car buyers using a choice set containing 10 vehicle types. Manski and Sherman (1980) also developed a MNL model to estimate the vehicle possession choices of 1,200 single-vehicle or two-vehicle households in the U.S. In their research, each household's choice set included their choice plus 25 alternatives randomly selected from 600 vehicle types. Manning and Winston (1985) also used an MNL model to estimate household vehicle ownership choices in the U.S. The sample size was 3,842 and the choice set included their choice plus nine alternatives randomly selected from 2,000 vehicle types. Kitamura et al. (2001) used an MNL model and data from 1,898 households in the South Coast metropolitan area (Los Angeles) in 1993 to analyze owned vehicle types; their choice set contained six vehicle types. Baltas and Saridakis (2013) used an MNL model to investigate the impact of behavioral and psychographic consumer characteristics on car preferences. They divided the vehicles into twelve categories based on vehicle size. Hocherman et al. (1983) developed an NL model to estimate the vehicle purchase behaviors of 800 households in Israel. In their NL model, the upper level were buying a first car or replacing an existing car and the lower level comprised the choice plus nineteen randomly selected alternatives from among 950 vehicle types. Berkovec and Rust (1985) developed an NL model of vehicle type choices, where the upper level had three vehicle age groups and the lower level five vehicle classes. The data was a nationwide U.S. sample of 237 single-vehicle households in 1976. Mannering et al. (2002) developed an NL model of vehicle type choice based on data from a survey of 654 U.S. households that had bought a new vehicle between 1993 and 1995. The upper level had two vehicle acquisition types—cash and no-cash—and the lower level comprised the choice plus nine alternatives randomly selected from 175 vehicle types. The literature describes two kinds of vehicle type choice models: the vehicle purchase

model when the chosen vehicle type is a new purchase (Lave and Train, 1979; Hocherman et al., 1983; Mannering et al., 2002), and the owned vehicle model when the vehicle is already owned (Manski and Sherman, 1980; Mannering and Winston, 1985; Berkovec and Rust, 185; Kitamura et al., 2001). The vehicle types in the choice set have been either categorized by functions and size (Lave and Train, 1979; Berkovec and Rust, 1985; Kitamura et. al, 2001) or randomly selected from a large set (Manski and Sherman, 1980; Manning and Winston, 1985; Hocherman et al, 1983; Berkovec and Rust, 1985; Mannering et al., 2002). The number of alternatives used for parameter estimation is small due to the sampling of alternatives from the universal set. However, in the vehicle purchase scenario, for example, the market always offers hundreds of vehicle types for sale and all vehicle types are available to all consumers. Therefore, the dozens of vehicle types in the choice set do not accurately reflect the much larger number of choices available to consumers in the real world. On the other hand, including hundreds of vehicle types in the consumer choice sets complicates the selection process because consumers will not consider some vehicle types and evaluating hundreds of alternatives is a complex task.

3.3 Consideration set formation method

During the consideration set formation process, researchers firstly would like to investigate whether the alternative satisfies the decision-makers' constraints on some or all alternative's attributes. Such information of decision-makers' constraints on attributes is unknown. Therefore, constraints on attributes are estimated by models with individual's attributes, alternative's attributes and the final choice information. In Chapter 2, we introduced the constraint based choice set formation model (Swait and Ben-Akiva, 1987) and applied the method proposed by Morikawa (1996) to estimate

the two-stage choice model simultaneously when the size of universal set is moderately big.

In this chapter, we will apply another method which is that after getting the probability of alternative to be considered, consideration set would be generated by randomly selecting alternatives from the universal set according to their acceptance probability. Higher acceptance probability always means higher probability to be included into the consideration set. The advantage of this method is the size of universal set is no longer an issue in model estimation, however the sampling alternative bias should be corrected in the choice modeling step. Rashidi et al. (2012) applied this framework to model the housing searching behavior where the size of house locations in the universal set is big.

The hazard based choice set formation method is applied to generate the consideration set for each decision-maker in advance, and then the vehicle choice behavior will be estimated by MNL model with consideration sets. The probability of alternative to be considered was calculated firstly, and then the consideration set was generated by randomly selecting alternatives based on their acceptance probabilities.

3.4 Data

An internet-based panel survey about vehicle holding and usage in Japan was used in this study. The data obtained in 2012 was used in this study. We firstly selected households who purchased a new vehicle in the past one year period, and we also found some household socio-demographic attributes including household size, income and so on. For the vehicle's information, we used the vehicle fuel efficiency data in 2012 from Ministry of Land, Infrastructure, Transport and Tourism, Japan. It should be mentioned

that this data included all vehicles from local Japanese manufacturers, however for foreign manufacturers, popular vehicles were included, but rare, expensive or customized vehicles were not included. As this kind of vehicles only took small part in the car market and such vehicles' information was difficult to collect, in this study, we only considered the vehicles in the vehicle fuel efficiency data 2012 to form the universal set.

Finally, the number of qualified households is 1495, where their socio-demographic attributes including number of owned vehicles, household income and number of household members were available from the survey. Furthermore, some dummy variables such as whether this household lived in the three largest cities or not and whether this household owned house or not were also obtained. Table 3.1 shows the household attributes used in the model.

Table 3.1: Sample distribution of household attributes

Parameter	Name	Average	St. dev.
Number of already owned vehicles	NOV	0.971	1.188
Family size (number of household members)	FS	3.231	1.415
Household income (Million Yen)	Income	6.585	3.945
Dummy: 1 for living in the three biggest cities, otherwise 0.	BC	0.227	0.419
Dummy: 1 for owning houses, otherwise 0.	Own_H	0.712	0.453
Dummy: 1 for male driver, otherwise 0.	Gender	0.508	0.500

Vehicles were classified by make, model, car type and fuel type, such as Toyota Prius compact car using hybrid engine. So the classes of vehicles were aggregated into 350. Table 3.2 shows the vehicle attributes and statistics information. The dummy variable "Domestic" is 1 if the vehicle is built by the domestic manufactories in Japan, otherwise it is 0. The fuel cost per kilometer is calculated by the purchased vehicle fuel efficiency and the gasoline price at the time when the vehicle was purchased.

Table 3.2: Descriptive statistics for purchased vehicle attributes

Parameter	Average
Body Type: Kei (Dummy)	0.361
Body Type: Compact (Dummy)	0.314
Body Type: Sedan (Dummy)	0.061
Body Type: Wagon (Dummy)	0.015
Body Type: Van (Dummy)	0.189
Body Type: SUV (Dummy)	0.057
Body Type: Sports (Dummy)	0.003
Domestic (Dummy)	0.965
Fuel cost (Yen/km)	7.313
Price (million Yen)	1.768

The definition of each vehicle body types is as follows:

1. **Kei car**: a Japanese category of small vehicles designed to comply with Japanese government tax and insurance regulations. The maximum size of a Kei car is 3.4 m in length, 1.48 m in width, and 2 m in height. The maximum engine displacement of a Kei car is 660 cc, and the maximum capacity is 4 passengers.

2. **Compact**: In Japan, the maximum size of a compact car generally is 4.2 m in length and 1.7 m in width. Engine displacements range from 1,000 cc to 1,500 cc, but some compact cars have displacements of more than 1500cc. The maximum capacity is 5 passengers.
3. **Sedan**: A passenger car with a three-box configuration and two rows of seats for four or more passengers.
4. **Wagon**: A sedan with a roof that extends to the rear of the vehicle.
5. **Van**: Designed to carry more people, with three rows of seats for 7 or 8 passengers. The car body shape can be comprised of one, one and a half, or two boxes.
6. **SUV**: A sport utility vehicle equipped with four-wheel drive. The shape is similar to that of a wagon but with more ground clearance.
7. **Sports car**: Designed for high performance. The shape is similar to that of a sedan, with two doors and two seats. Some sports cars have two small seats behind the front seats. Some sports cars are very expensive, while others are moderately priced.

3.5 Consideration set formation: methodology and result

In this study, we assumed that consumers will first construct their consideration sets by screening available vehicles with their preferences. And then, they will choose the most favorite vehicle among vehicles in consideration sets. In this section, we will discuss how to construct the consideration set.

When people are buying a goods, usually, the price is one of the most important aspects to be considered. Furthermore, with the rising oil price and public attention to environmental protection, consumers are paying more attention to vehicle's fuel cost.

Therefore, in this study, these two vehicles' attributes including vehicle fuel cost and vehicle price will be utilized in the choice set formation process. In this study, the consideration set formation method is based on the assumption that there exist distributions of acceptable vehicle price and acceptable vehicle fuel cost which are modelled by the hazard-based formulation conditional on the household's socioeconomic attributes (Rashidi et al., 2012; Zolfaghari et al., 2012). In order to find best fit distributions for two vehicle attributes, we do the Kolmogorov-Smirnov test of different distributions to observed purchased vehicle price and fuel cost values, and the result in Table 3.3 indicates that both attributes follow a log-normal distribution.

Table 3.3: Kolmogorov-Smirnov Statistics of Different Distributions

Distribution	**Price**	**Fuel Cost**
	K-S statistic	K-S statistic
Normal	0.177	0.116
Log-Normal	0.091	0.063
Logistic	0.133	0.094
Log-Logistic	0.312	0.646
Weibull	0.175	0.117

According to the definition of hazard function and survival function, the probability density function of accepting a vehicle price and accepting a vehicle fuel cost are as follows:

$$f(p) = \lambda(p) \bullet S(p), \quad f(fc) = \lambda(fc) \bullet S(fc) \qquad [3.1]$$

where $f(p)$ and $f(fc)$ are the probability density functions for accepting a vehicle price and accepting a vehicle fuel cost respectively, $\lambda(p)$ and $\lambda(fc)$ are hazard functions and

$S(p)$ and $S(fc)$ are survival functions. As we use the log-normal distribution here, therefore, the probability density functions (PDF) of vehicle price and fuel cost are as follows:

$$f(p) = \frac{e^{-((\ln(p/m_p))^2/(2\sigma_p^2))}}{p\sigma_p\sqrt{2\pi}}, \quad f(fc) = \frac{e^{-((\ln(fc/m_{fc}))^2/(2\sigma_{fc}^2))}}{fc\sigma_{fc}\sqrt{2\pi}} \qquad [3.2]$$

where σ_p and σ_{fc} are shape parameters for vehicle price and fuel cost respectively, and m_p and m_{fc} are scale parameters. In order to incorporate the socioeconomic attributes into the PDFs, the scale parameter will be parameterized as follows:

$$\log(m) = \theta_0 + \theta X \qquad [3.3]$$

where X are socioeconomic attributes, θ_0 and θ are parameters to be estimated. Then, the PDF will be formulated as follows:

$$f(p) = \frac{e^{-(\ln p - \ln m_p)^2/(2\sigma_p^2))}}{p\sigma_p\sqrt{2\pi}}, \quad f(fc) = \frac{e^{-(\ln fc - \ln m_{fc})^2/(2\sigma_{fc}^2))}}{fc\sigma_{fc}\sqrt{2\pi}} \qquad [3.4]$$

Then, the joint likelihood function would be formulated as follows:

$$L = \prod_{n=1}^{N} f_n^p(p) \bullet f_n^{fc}(fc) \qquad [3.5]$$

where $f_n^p(p)$ and $f_n^{fc}(fc)$ are probability density functions of vehicle price and fuel cost for household n. Estimated parameters includes shape parameters and parameters in Equation (3.4). Maximum likelihood estimation method is applied to estimate the parameters where vehicle price and fuel cost are household's observed quantities.

Table 3.4: Parameter Estimation of Hazard-based Choice set Model

Parameter	Estimate
Acceptable Fuel Cost	
σ (Shape parameter)	0.308**
θ_0 (constant)	1.798**
BC	-0.007
NOV	0.022**
Gender	-0.083**
Income	-0.005**
Own_H	0.062**
FS	-0.040**
Acceptable Vehicle Price	
σ (Shape parameter)	0.381**
θ_0 (constant)	0.168**
BC	0.080**
NOV	-0.032**
Gender	0.188**
Income	0.022**
Own_H	-0.042*
FS	0.034**
Sample size: 1495	
LL(0): -9526.1	
LL: -4647.9	

**: 0.05 significant level; *: 0.1 significant level

Table 3.4 shows the estimation results of the hazard-based choice set formation model. According to the result, for a certain vehicle, the probability to be considered would be higher if the family is not living in the big cities, with higher income and bigger family size. However, if the family had already purchased their house and vehicles, the

probability will be decreased. Furthermore, if the main driver is male, the probability will become higher. The probability for family *n* considering vehicle *i* with price p_i and fuel cost c_i are formulated as:

$$q_n(i) = f_n^{fc}\left(fc_i \mid X_n\right) f_n^{p}\left(p_i \mid X_n\right) \quad [3.6]$$

For each household, all vehicle alternatives will have a probability assigned according to the vehicle's attributes and household attributes. Then we applied weighted random sampling method to select alternatives for consideration sets. Vehicle with higher acceptance probability (q_i) will have higher probability to be selected. The consideration set formation process for each household will be completed until the actually chosen alternative is included into it. The merits of this method are alternatives in the consideration set are close to households' preference, and the finally chosen alternative could be guaranteed in the consideration set. The demerits of this method are that the size of consideration set might become large as the size of the universal set. In this study, after the consideration set formation process, the average size of consideration sets became 105, which is much less than the size of the universal set, 350.

3.6 Vehicle purchase model: methodology and results

After constructing the consideration set for each household, a multinomial logit (MNL) model will be applied to analyze the vehicle purchasing behaviors. However, as the consideration set is a subset of the universal set, in order to keep estimated parameters consistent, an alternative specific correction term for sampling bias will be introduced with the method presented by Ben-Akiva and Lerman (1985). The method was also applied in route choice (Frejinger et al., 2009) and residential location choice (Rashidi et al., 2012). Furthermore, McFadden (1978) also proved that MNL model would be

consistently estimated with a subset of alternatives. The utility function and the probability of household *n* choosing alternative *i* is:

$$\begin{aligned} V_n(i) &= \left(\beta_{Compact}^{constant} + \beta_{Compact}^{FS} * x_{FS}\right) * x_{Compact} \\ &+ \left(\beta_{Compact}^{constant} + \beta_{Sedan}^{FS} * x_{FS}\right) * x_{Sedan} \\ &+ \left(\beta_{Compact}^{constant} + \beta_{Wagon}^{FS} * x_{FS}\right) * x_{Wagon} + \left(\beta_{Compact}^{constant} + \beta_{Van}^{FS} * x_{FS}\right) * x_{Van} \\ &+ \left(\beta_{Compact}^{constant} + \beta_{SUV}^{FS} * x_{FS}\right) * x_{SUV} + \left(\beta_{Compact}^{constant} + \beta_{Sports}^{FS} * x_{FS}\right) * x_{Sports} \\ &+ \left(\beta_{Price}^{constant} + \beta_{Price}^{Income} * x_{income}\right) * x_{Price} + \beta_{FC} * x_{FC} \\ &+ \left(\beta_{Domestic}^{constant} + \beta_{Domestic}^{BC} * x_{BC}\right) * x_{Domestic} \end{aligned} \quad [3.7]$$

$$P_n(i) = \frac{\exp\left(V_n(i) - \ln\left(q_n(i)\right)\right)}{\sum_{j=1}^{J} \exp\left(V_n(j) - \ln\left(q_n(j)\right)\right)} \quad [3.8]$$

where $V_n(i)$ is the deterministic utility, $P_n(i)$ is the probability of household *n* choosing alternative *i*, $q_n(i)$ is the probability that alternative *i* is considered by household *n*. The vehicle with higher utility would have higher probability to be chosen.

Table 3.5: Parameter Estimation of Models with different choice sets

Parameter	MNL with universal set	MNL with Random choice set	MNL with Consideration set
$Constant_{compact}$	0.672**	0.692**	0.859**
$FS_{compact}$	-0.292**	-0.289**	-0.366**
$Constant_{sedan}$	0.913**	0.992**	1.219**
FS_{sedan}	-0.871**	-0.914**	-1.223**
$Constant_{wagon}$	0.381	0.487	0.733*
FS_{wagon}	-1.013**	-1.179**	-1.433**
$Constant_{miniVan}$	0.585**	0.592**	0.718**
$FS_{miniVan}$	0.506**	0.510**	0.520**
$Constant_{suv}$	1.321**	1.436**	1.939**
FS_{suv}	-0.591**	-0.670**	-0.992**
$Constant_{sports}$	-3.072	-3.356	-3.465

FS_{sports}	1.443	1.403	-0.992**
$Constant_{price}$	-2.841**	-2.811**	-3.364**
$Income_{price}$	1.017**	1.005**	1.207**
FuelCost	-0.241**	-0.224**	-0.283**
$Constant_{domestic}$	1.616**	1.634**	1.833**
$BC_{domestic}$	-0.838**	-0.825**	-0.805**
Sample size	1495	1495	1495
LL	-7314.3	-4857.9	-4813.0
APCP	0.010	0.110	0.117

**: 0.05 significant level; *: 0.1 significant level

The estimation results of the MNL model based on the consideration set was shown in Table 3.5. Column 1 shows the estimation result of MNL with universal set. Every household is assumed to choose from all available vehicles in the market. In column 2, vehicles in the choice set for each household are randomly selected from the universal set except that chosen vehicles are automatically included into the set in advance, and the size of choice set is the same as that of consideration set in column 3. The result of MNL model with consideration set is in column 3. According to the result, Japanese families prefer high fuel efficiency vehicles as expected. They also prefer domestic vehicle, however people who are not living in the three biggest cities show more interest to domestic vehicles. Generally, they would like to consider the cheap vehicles, however, for some high income families, they would like to choose the expansive vehicles for higher quality and performance. The number of household members would have big influence on vehicle type choice. Therefore, in the estimation result, we found that the utility of miniVan would be increased within the increasing of the number of family members. For other vehicle types, vehicle utilities will be decreased in different degrees at the same time when the family size become bigger. Estimation results

between model 1 (MNL with universal set) and model 2 (MNL with random set) are almost same, it proved that using the random choice set would get the consistent estimation result with using the universal set. Values of parameters in model 3 are bigger than those in model 1 and model 2, but most of them have the same directions of effect on the utility function.

As the size of choice set for each household is different among these models, the common index such as McFadden value or log-likelihood values at convergence is not appropriate to be used for comparison. Here, we will use another measures of fitness for disaggregate validation (Bhat and Pulugurta, 1998). The measure is called average probability of correct prediction (APCP). The formulation of this measure is as follows:

$$\text{Average Probability of Correct Prediction} = \frac{\sum_n \sum_i y_{ni} p_{ni}}{N} \qquad [3.9]$$

where N is the sample size, y_{ni} is a dummy variable indicating if household n chooses alternative vehicle i, p_{ni} is the predicted probability for household n choosing alternative i which is calculated by the estimated parameters. The APCP values for each model are presented in the last row of Table 3.5. It is obvious that when the size of choice set becomes smaller, the APCP values will be increased significantly. When the sizes of choice sets are same in model 2 and model 3, the APCP value for model using consideration set is slightly higher than the model using random set.

3.7 Conclusion

In this study, we applied hazard-based choice set formation method to generate the consideration set for each household according to sociodemographic and vehicle's attribute. Every available vehicle in the market will be firstly assigned an acceptance

probability according to the hazard model, and vehicle with higher acceptance probability will have higher probability to be included into the consideration set. Then we estimated three MNL models with different choice set. The first one is MNL model with universal set. Households are assumed to choose their favorite vehicles among all available vehicles in the market. The second one is MNL model with random choice set. Vehicles in choice sets are randomly selected from the universal set. The size of random choice set for each household is same as the size of consideration set in model 3. The third one is MNL model with consideration set. According to two introduced indexes, model with consideration set has the best performance and it also shows the best prediction accuracy in the following average gasoline vehicles' fuel efficiency prediction. Finally, in the simulation, it indicated that the increase in gasoline price will also increase the average gasoline vehicle's fuel efficiency as we expected.

This study shows the advantage of applying consideration sets in the MNL model estimation. However, there still are some shortcomings. Firstly, although the consideration set is smaller than the universal set, the average size of consideration sets is not small enough which indicated that the efficiency of consideration set formation process should be improved. Secondly, although consideration set shows the advantage in model estimation, there exists other consideration set formation methods. Therefore, in the future research, we would like to compare different consideration set formation methods, and find which method would generate the most reasonable consideration sets for decision-makers.

3.8 Reference

1. Baltas, G. and C. Saridakis (2013). "An empirical investigation of the impact of behavioural and psychographic consumer characteristics on car preferences: An integrated model of car type choice." Transportation Research Part A: Policy and Practice 54: 92-110.
2. Ben-Akiva, M. and L. S (1985). Discrete choice analysis: theory and application to predict travel demand. Cambridge, MIT Press.
3. Berkovec, J., & Rust, J. (1985). A nested logit model of automobile holdings for one vehicle households. *Transportation Research Part B: Methodological*,*19*(4), 275-285.
4. Bhat, C. R. and V. Pulugurta (1998). A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. 32: 61-75.
5. Bronnenberg, B. J. and W. R. Vanhonacker (1996). "Limited choice sets, local price response and implied measures of price competition." Journal of Marketing Research: 163-173.
6. Brown, J. J. and A. R. Wildt (1992). "Consideration set measurement." Journal of the Academy of Marketing Science 20(3): 235-243.
7. Dzyabura, D. and J. R. Hauser (2011). "Active machine learning for consideration heuristics." Marketing Science 30(5): 801-819.
8. Frejinger, E., M. Bierlaire and M. Ben-Akiva (2009). "Sampling of alternatives for route choice modeling." Transportation Research Part B: Methodological 43(10): 984-994.
9. Hauser, J. R. and B. Wernerfelt (1990). "An evaluation cost model of consideration sets." Journal of consumer research: 393-408.

10. Hocherman, I., Prashker, J. N., & Ben-Akiva, M. (1983). *Estimation and use of dynamic transaction models of automobile ownership* (No. 944).
11. Kitamura, R., T. Akiyama, T. Yamamoto and T. F. Golob (2001). "Accessibility in a metropolis: Toward a better understanding of land use and travel." Transportation Research Record: Journal of the Transportation Research Board 1780(1): 64-75.
12. Lave, C. A. and K. Train (1979). "A disaggregate model of auto-type choice." Transportation Research Part A: General 13(1): 1-9.
13. Mabit, S. L. (2014). "Vehicle type choice under the influence of a tax reform and rising fuel prices." Transportation Research Part A: Policy and Practice 64: 32-42.
14. Mannering, F. and C. Winston (1985). "A dynamic empirical analysis of household vehicle ownership and utilization." The RAND Journal of Economics: 215-236.
15. Mannering, F., Winston, C., & Starkey, W. (2002). An exploratory analysis of automobile leasing by US households. *Journal of Urban Economics*, *52*(1), 154-176.
16. Manski, C. F. (1977). "The Structure of Random UTility Models." Theory and Decision 8: 229-254.
17. Manski, C. F. and L. Sherman (1980). "An empirical analysis of household choice among motor vehicles." Transportation Research Part A: General 14(5-6): 349-366.
18. McFadden, D. (1978). Modelling the choice of residential location, Institute of Transportation Studies, University of California.
19. Mehta, N., S. Rajiv and K. Srinivasan (2003). "Price uncertainty and consumer search: A structural model of consideration set formation." Marketing science 22(1): 58-84.
20. Morikawa, T. (1996). "A hybrid probabilistic choice set model with compensatory and noncompensatory choice rules." In:Proceedings of the 7th World Conference

on Transport Research 1: 317-325.

21. Paulssen, M. and R. P. Bagozzi (2005). "A self‐regulatory model of consideration set formation." Psychology & Marketing 22(10): 785-812.
22. Rashidi, T. H., J. Auld and A. Mohammadian (2012). "A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection." Transportation Research Part A: Policy and Practice 46(7): 1097-1107.
23. Shocker, A. D., M. Ben-Akiva, B. Boccara and P. Nedungadi (1991). "Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions." Marketing Letters 2(3): 181-197.
24. Swait, J. and M. Ben-Akiva (1987). "Empirical test of a constrained choice discrete model: Mode choice in São Paulo, Brazil." Transportation Research Part B: Methodological 21(2): 103-115.
25. Wu, J. and A. Rangaswamy (2003). "A fuzzy set model of search and consideration with an application to an online market." Marketing Science 22(3): 411-434.
26. Xu, G., T. Miwa, T. Morikawa and T. Yamamoto (2015). "Vehicle purchasing behaviors comparison in two-stage choice perspective before and after eco-car promotion policy in Japan." Transportation Research Part D: Transport and Environment 34: 195-207.
27. Zolfaghari, A., a. Sivakumar and J. Polak (2012). "Choice Set Formation in Residential Location Choice Modelling: Empirical Comparison of Alternative Approaches." Transportation Research Board: 1-26.

4. Bayesian approach based vehicle choice model with conjunctive screening rule

4.1 Introduction

In Chapter 3, the hazard-base choice set formation model was applied to calculate the probability of each alternative to be considered by decision-makers, then vehicles were randomly selected into the consideration set where higher consider probability will have high probability to be included, then MNL logit model was estimated with the generated consideration sets. In this chapter, we would like to investigate another method in constructing decision-makers' consideration sets.

Researchers always assume that decision-makers have a threshold value on each alternative's aspect. If we could know or estimate these threshold values, the consideration set would be generated deterministically for each decision-maker, and then the choice model would be estimated based on these consideration sets. Therefore, in this chapter, we will try to discover a proper method to which the consideration set could be deterministically generated.

4.2 Methodology

Gilbride and Allenby (2004) proposed a method that the choice model would be simultaneously estimated with the consideration set by a Bayesian method. First, in the consideration set formation stage, they introduced an indicator function $I(x_i,\gamma)$ to determine whether or not the alternative satisfies the applied screening rules; x_i denotes

a generic argument in which the screening rule is applied to the *i*th alternative, and γ is the relative threshold. When $I(x_i,\gamma)=1$, alternative *i* satisfies the screening rule and can be included in the consideration set, if 0, alternative *i* will be excluded from the consideration set and have zero probability to be chosen.

In a compensatory screening rule, the alternative is considered only when its deterministic portion of the utility V_i exceeds a threshold γ. If it does, then the indicator function is expressed as follows.

$$I\left(x_i,\gamma\right)=I\left(V_i>\gamma\right)=1 \qquad [4.1]$$

If the conjunctive screening rule is applied, then an alternative is considered only when all relevant attributes for this alternative are accepted. Therefore, the indicator functions across the attributes for an alternative should be multiplied using the conjunctive rule.

$$I\left(x_i,\gamma\right)=\prod_m I\left(x_{im}>\gamma_m\right)=1 \qquad [4.2]$$

where x_{im} is the level of attribute *m* for alternative *i* and γ_m is the attribute acceptable level or threshold value. If the threshold value is smaller than the smallest level of the attribute, which means the indicator function for this attribute is always equal to 1, then this attribute is not used for screening alternatives. The disjunctive screening rule, a noncompensatory screening rule, requires acceptance of at least one of the attribute levels. The indicator function is expressed as follows.

$$I\left(x_i,\gamma\right)=\sum_m I\left(x_{im}>\gamma_m\right)\geq 1 \qquad [4.3]$$

Thresholds for continuous variables are parameters to be estimated in both the compensatory and noncompensatory models. However, in the conjunctive and disjunctive screening rules, for discrete variables, the threshold parameter is assumed

to be distributed multinomial. For example, for the vehicle type attribute of compact in Table 3.2, the attribute level is 1 for a compact vehicle and 0 for vehicles with other vehicle types. A grid of possible cutoffs $\{\gamma\}$ is specified in advance:

(1) If the threshold is less than 0, e.g., γ=-0.5, then $I(0>-0.5)=1$ and $I(1>0.5)=1$, which indicates that decision-maker does not use this attribute to screen vehicles. Both compact vehicle and other types of vehicles would be included into the consideration set theoretically.

(2) If the threshold is bigger than 0 and less than 1, e.g., γ=0.5, then $I(0>0.5)=0$ and $I(1>0.5)=1$, which indicates that a compact vehicle will be considered.

(3) If the threshold is bigger than 1, e.g., γ=1.5, then $I(0>1.5)=0$ and $I(1>1.5)=0$, which indicates that a compact vehicle will not be considered.

In this situation, we estimate the mass probabilities of decision-makers applying different possible cut-offs for discrete variables.

In Gilbride and Allenby's model (Gilbride and Allenby, 2004), the choice probability for alternative *i* from the universal set *M* after introducing a consideration set screening rule is expressed as follows.

$$P(i)_k = P\left(V_{ki} + \varepsilon_{ki} > V_{kj} + \varepsilon_{kj} \text{ for all } j \text{ such that } I\left(x_{kj}, \gamma_k\right) = 1\right) \qquad [4.4]$$

where *k* indexes the decision makers, V_{ki} is the deterministic portion of utility of alternative *i* for decision maker *k*, and ε_{ki} is the stochastic portion. The indicator function is expressed according to the screening rule that is applied. If thresholds are assumed as fixed-value or discrete-random variables, then the likelihood surface will become discontinuous and irregular, which leads to indifferentiability, and the gradient-

based estimation method is no longer appropriate. Gilbride and Allenby (2004) applied a Bayesian approach—a data augmentation method (Tanner and Wong, 1987)—to estimate parameters in the model.

If estimating a simple discrete choice model, such as a probit model, then the data augmentation method, $\{V\}$ is augmented with a vector of latent variable z.

$$z_{ki} = V_{ki} + \varepsilon_{ki}, \quad \varepsilon_{ki} \sim \text{Normal}(0,1) \tag{4.5}$$

The choice model can then be written hierarchically.

$$y \mid z \tag{4.6}$$

$$z \mid V \tag{4.7}$$

where y denotes the decision makers' choices. Equation (4.6) shows that an alternative with a maximum value of z will be chosen. Equation (4.7) shows that the latent variables are normally distributed with a mean of V_{ki} and a variance of 1. In probit models, V is usually parameterized as $\boldsymbol{X\beta}$, with $\boldsymbol{\beta}$ being the parameters to be estimated. The model is estimated by iteratively drawing from the following conditional distributions using the Markov Chain Monte Carlo (MCMC) method.

$$\pi(z \mid y, X\beta) = \text{Truncated Normal}(X\beta, \text{IM}) \tag{4.8}$$

$$\pi(\beta \mid z) = \text{Normal}(\beta^0, \sigma^2 \text{IM}) \tag{4.9}$$

where "IM" is the identity matrix. In Equation (4.8), if y_{ki}=1, the distribution of z_{kj} is truncated at the left by 0, and if y_{ki}=0, then the distribution of z_{kj} is truncated at the right by 0. Hyperparameters β^0 and σ^2 are assumed independently.

The choice model with a screening rule can then be written hierarchically.

$$y \mid z, I(x, \gamma) \tag{4.10}$$

$$z \mid V \qquad [4.11]$$

where latent variable *z* is augmented in the same way as in Equation (4.5). Equation (4.10) shows that only alternatives that satisfy the screening rule will be considered and its choice probability is relative to the other alternatives in the consideration set. Then estimation proceeds by drawing iteratively from the conditional distributions.

$$z \mid y, V, I\left(x, \gamma\right) \qquad [4.12]$$

$$\gamma \mid y, z, x \qquad [4.13]$$

$$V \mid z \qquad [4.14]$$

The conditional distribution in Equation (4.12) is a truncated normal distribution for alternatives in the consideration set, and the value of *z* for the chosen alternative should be greater than the other alternatives in the consideration set. For alternatives that are not in the consideration set according to the indicator functions, their latent variables are drawn from a non-truncated distribution.

The conditional distribution for the threshold values in Equation (4.13) is dependent on the choice data *y* and latent variables *z*. Different threshold values place different alternatives in the consideration set. Therefore, permissible threshold values are those that place the observed choice with maximum latent variable *z* in the consideration set. In this model, heterogeneity is also considered, therefore, if attribute *m* is continuously distributed, then it is assumed to follow a normal distribution.

$$\gamma_{km} \sim \text{Normal}\left(\bar{\gamma}_{m,}\sigma_{\gamma}^{2}\right) \qquad [4.15]$$

If the level for attribute *m* is nominally distributed (i.e. 0, 1), a grid of possible threshold values (e.g. -0.5, 0.5, 1.5) is firstly specified. If the attribute has more levels, more possible threshold values will be simultaneously available. The threshold values are

then distributed multinomial by considering the heterogeneity.

$$\gamma_{km} \sim \text{Multinomial}\left(\theta_m\right) \quad [4.16]$$

where θ_m is the vector of multinomial probabilities related to the grid of possible threshold values for the attribute.

The standard normal distribution theory can be used to draw *V* in Equation (4.14). *V* is parameterized as $V = x\beta$ and β are parameters drawn from a normal distribution with a mean and covariance matrix corresponding to the standard OLS model. In order to consider the heterogeneity, β is distributed as follows.

$$\beta \sim \text{MVN}\left(\overline{\beta}, \Sigma_\beta\right) \quad [4.17]$$

After specifying the prior distributions of the hyper parameters, including $\overline{\beta}, \Sigma_\beta$, θ (if there are discrete attributes for the choice models with a conjunctive or disjunctive screening rule), $\overline{\gamma}$, and σ_γ^2, the Markov Chain Monte Carlo estimation for this Hierarchical Bayes model is calculated by drawing iteratively from these conditional distributions. For example, a conjunctive choice model, after the data augmentation, the model can be shown as:

$$y_{kij} = 1 \text{ if } z_{kij} > z_{khj} \text{ for all } h \text{ that } I\left(x_{khj}, \gamma_h\right) \quad [4.18]$$

$$z_{kij} = x_{kij}\beta_k + \varepsilon_{kij}, \quad \varepsilon \sim \text{Normal}\left(0, \sigma_\varepsilon = 1\right) \quad [4.19]$$

$$I\left(x_{kij}, \gamma_h\right) = 1 \text{ if } \prod_m I\left(x_{kijm} > \gamma_{km}\right) = 1 \quad [4.20]$$

where *k* indexes the decision-maker, *j* the choice set, *i* the alternative in the choice set, and *m* the attributes in each alternative. Heterogeneity on the individual parameters could be introduced with the distributions as Equation (4.15), (4.16) and (4.17). Here we assume that there is only one continuous attribute, and other *m*-1 are discrete

attributes. The prior distributions are specified as:

$$\overline{\beta} \sim MVN(0, 100\text{IM}) \quad \Sigma_\beta \sim \text{IW}(v, \Delta)$$

$$\theta_p \sim \text{Dirichlet}(\alpha) \quad \text{for } p=1,\ldots,m\text{-}1 \text{ (discrete attribute)}$$

$$\overline{\gamma} \sim \text{Normal}(0,100) \quad \sigma_\gamma^2 \sim \text{IG}(a,b) \text{ (Continuous attribute)}$$

where v, a and b are positive numbers, $\Delta = v$*IM. θ_p is a vector of dimension n where n means the possible values for γ_p and α is a conforming vector. Then the estimation procedures are as follows:

(1) Generate $z_{kj} \mid y_{kj}, X_{kj}, I\left(X_{kj}, \gamma_k\right), \beta_k, \sigma_\varepsilon = 1$ for $j = 1,..., J$ and $k = 1,..., K$.

Start with i=1,

if $y_{kij} = 1$, then $z_{kij} \sim \text{TN}(x_{kij}\beta_k, \sigma_\varepsilon = 1, z_{kij} >$

z_{khj}, for all h such that $I(x_{khj}, \gamma_k) = 1$),

if $y_{kij} = 0$ and $I(x_{kij}, \gamma_k) = 1$, then $z_{kij} \sim \text{TN}(x_{kij}\beta_k, \sigma_\varepsilon = 1, z_{kij} <$

z_{khj}, where $y_{khj} = 1$).

else $z_{kij} \sim \text{Normal}(x_{kij}\beta_k, \sigma_\varepsilon = 1)$.

Increment i and return to top.

(2) Generate $\beta_k \mid z_k, X_k$ for k=1,...,K.

$$\beta_k \sim \text{MVN}\left(b, \left(X_k^{'} X_k + \Sigma_\beta^{-1}\right)^{-1}\right), \quad b = \left(X_k^{'} X_k + \Sigma_\beta^{-1}\right)^{-1}\left(X_k^{'} z_k + \Sigma_\beta^{-1}\overline{\beta}\right)$$

(3) Generate $\overline{\beta} \mid \{\beta_k\}, \Sigma_\beta$

$$\overline{\beta} \sim \text{MVN}\left(\bar{b}, \left(\left(\Sigma_\beta / H\right)^{-1} + \left(100IM\right)^{-1}\right)^{-1}\right),$$

$$\bar{b} = \left(\left(\Sigma_\beta / H\right)^{-1} + \left(100IM\right)^{-1}\right)^{-1}\left(\Sigma_\beta^{-1}\sum_{k=1}^{\text{K}}\beta_k + \left(100IM\right)^{-1}(0)\right)$$

(4) Generate $\Sigma_\beta \mid \{\beta_k\}, \overline{\beta}$

$$\Sigma_\beta \sim IW\left(v + K, \Delta + \sum_{k=1}^{K}\left(\beta_k - \overline{\beta}\right)^{'}\left(\beta_k - \overline{\beta}\right)\right)$$

(5) Generate $\gamma_{kp} \mid z_k, y_{k,}, X_k$ for p= 1,…m-1 (discrete attributes) and k=1,…,K.

If there is n possible values of γ for each attribute, and let I(γ^a_{kpn})=1 indicate that γ^a_{kpn} is an allowable number.

Start with n=1,

if $y_{kij} = 1$ and $\prod_p I(x_{kijp} > \gamma_{kp})$=0, then I($\gamma^a_{kpn}$)=0,

if $y_{kij} = 0$ and $\prod_p I(x_{kijp} > \gamma_{kp})$=1, then

if $z_{kij} > z_{khj}$, where $y_{khj} = 1$, then I(γ^a_{kpn})=0,

else, I(γ^a_{kpn})=1.

Increment n and return to top.

Then choosing γ_{kp} from γ^a_{kpn} is: $\gamma_{kp} = \gamma^a_{kpn}$ with probability

$I(\gamma^a_{kpn})\theta_{pn} \Big/ \sum_n I(\gamma^a_{kpn})\,\theta_{pn}$.

(6) Generate $\theta_p \mid \{\gamma_{kp}\}$ for p=1,…,m-1 (discrete attribute)

If we define $s_{kpn} = 1$ if $\gamma_{kp} = \gamma_{kpn}$, otherwise $s_{kpn} = 0$. Then,

$$\theta_m \sim \text{Dirichlet}\left(\sum_{k=1}^{K} s_{kp1} + \alpha_1, ..., \sum_{k=1}^{K} s_{kpn} + \alpha_n\right)$$

(7) Generate $\gamma_{km} \mid z_k, y_{k,}, X_k$ for (continuous attribute m) k=1,…,K.

Gilbride and Allenby (2004) applied a traditional random-walk Metropolis-Hasting step here for the generating the possible value for continuous attribute.

It should be mentioned that the value should be checked like in step 2 to see whether it is an allowable value or not.

(8) Generate $\bar{\gamma} \mid \{\gamma_{km}\}, \sigma^2_\gamma$

$$\bar{\gamma} \sim \text{Normal}(c, d), \quad c = \left(\frac{\sigma^2_\gamma / K}{\sigma^2_\gamma / K + 100}(0) + \frac{100}{\sigma^2_\gamma / K + 100}\left[\frac{\sum_{k=1}^{K} \gamma_{km}}{K}\right] \right),$$

$$d = \frac{100\sigma_\gamma^2 / K}{\sigma_\gamma^2 / K + 100}$$

(9) Generate $\sigma_\gamma^2 \mid \{\gamma_{km}\}, \bar{\gamma}$

$$\sigma_\gamma^2 \sim \text{IG}\left(\frac{K}{2} + a, \left\{ \frac{1}{b} + \frac{1}{2} \sum_{k=1}^{K} \left(\gamma_{kp} - \bar{\gamma} \right)^2 \right\}^{-1} \right)$$

The estimation procedure should be iterated sufficient large times to ensure the result convergence. It should be mentioned that changing sequence above will not affect the estimation results.

4.3 Model Specifications and Estimation Results

In this study, we assume that consumers apply the conjunctive screening rule when formulating a consideration set. For other screening rules, the compensatory screening rule still requires decision-makers to first evaluate all alternatives, and an alternative is considered only when its utility exceeds the individual's threshold. When the number of alternatives is large, evaluating all of the alternatives is a complex and impractical task. Under the disjunctive screening rule, the alternative is considered even when only one of its attribute levels is acceptable. If the number of attributes being considered is large or the size of the universal set huge, then the size of the consideration set will be excessively large. Furthermore, using only one attribute level to screen the alternatives is not suitable for a decision as complex and important as purchasing a vehicle for a household. Laroche et al. (2003) also found that conjunctive heuristics is the decision rule that is most often used when forming a consideration set. Therefore, in this study, the conjunctive screening rule is applied during the consideration set formation stage because this rule ensures that the resulting consideration set is rationally sized.

The advantage of this estimation method is that researcher would define various screening rules based on alternative's attributes. For example, in the vehicle choice situations, the available alternative's attributes are including vehicle body types, origin of manufactory, price and fuel efficiency. All these attributes would be applied in the screening rule separately or jointly. In the following, we would like to try different combinations to find out the best screening rule in vehicle purchasing choice.

Firstly, price might be the first important aspect which consumers would like consider before purchasing any kind of goods. Therefore, we would like to use price attribute only to screening the alternative vehicles. The formulation of the screening rule using only price attribute is as follows:

$$I\left(x,\gamma\right)=I\left(-Ln(x_{Price})>\gamma_{Price}\right) \qquad [4.21]$$

Generally, consumers only consider the goods under their purchasing budget, therefore, we transform the value of the price into the –Ln(price), the model using this screening rule is name Bayes Model 1. Then, the utility function in the choice making stage is:

$$\begin{aligned}
z=&\left(\beta_{Compact}^{constant}+\beta_{Compact}^{FS}*x_{FS}\right)*x_{Compact}\\
&+\left(\beta_{Compact}^{constant}+\beta_{Sedan}^{FS}*x_{FS}\right)*x_{Sedan}\\
&+\left(\beta_{Compact}^{constant}+\beta_{Wagon}^{FS}*x_{FS}\right)*x_{Wagon}+\left(\beta_{Compact}^{constant}+\beta_{Van}^{FS}*x_{FS}\right)*x_{Van}\\
&+\left(\beta_{Compact}^{constant}+\beta_{SUV}^{FS}*x_{FS}\right)*x_{SUV}+\left(\beta_{Compact}^{constant}+\beta_{Sports}^{FS}*x_{FS}\right)*x_{Sports}\\
&+\left(\beta_{Price}^{constant}+\beta_{Price}^{Income}*x_{income}\right)*x_{Price}+\beta_{FC}*x_{FC}\\
&+\left(\beta_{Domestic}^{constant}+\beta_{Domestic}^{BC}*x_{BC}\right)*x_{Domestic}+\varepsilon, \qquad \varepsilon\sim\text{Normal}\left(0,1\right)
\end{aligned} \qquad [4.22]$$

This utility function is same as the utility function in Chapter 3. In the estimation, MCMC was run for 3,000 iterations. The first 2,000 iterations were for burn-in, which left 1,000 draws for estimating the posterior distributions. The heterogeneity is

considered using the method described in the methodology section.

Table 4.1: Estimation result of cutoff values in Bayes Model 1

	-Ln(Price)
Mean	-5.284
Variance	0.200

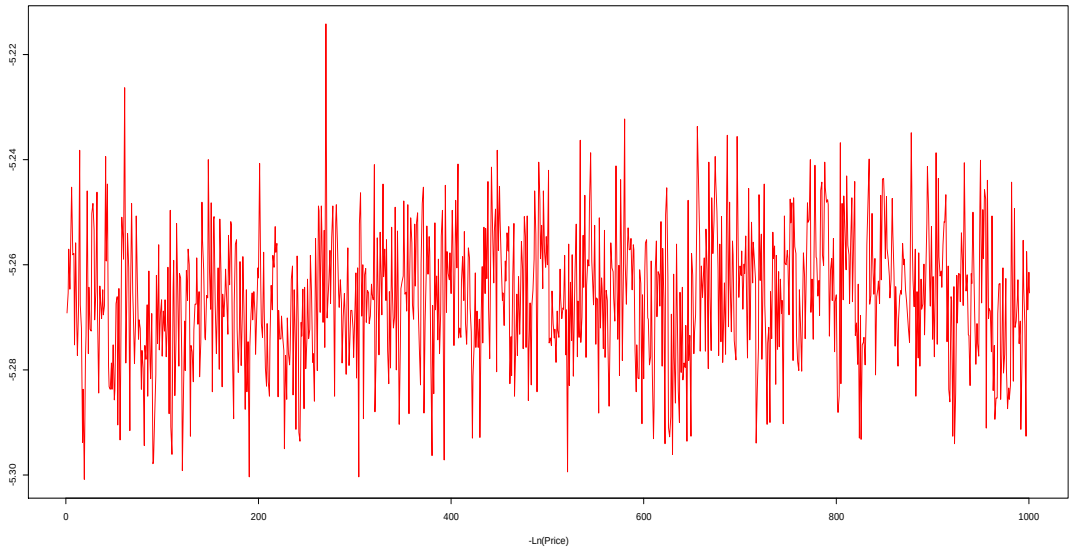

Figure 4-1: MCMC Sequence of -Ln(Price) in Bayes Model 1

According to the estimation result, the average purchasing budget is 1.97 million Yen. And the posterior mean of the consideration set size is 91, which indicate that price is an effective threshold to screen the alternatives. The posterior mean of the parameters in the utility function is shown in the following table:

Table 4.2: Estimates for the parameters in Bayes Model 1

Parameter	Posterior Mean	2.5% Quantiles	97.5% Quantiles
$Constant_{compact}$	-0.021	-0.137	0.083
$FS_{compact}$	-0.270	-0.356	-0.166

$Constant_{sedan}$	0.724	0.624	0.844
FS_{sedan}	-0.623	-0.712	-0.551
$Constant_{wagon}$	-0.195	-0.293	-0.107
FS_{wagon}	-0.598	-0.686	-0.527
$Constant_{van}$	0.118	0.008	0.240
FS_{van}	0.835	0.664	0.978
$Constant_{suv}$	0.909	0.840	0.973
FS_{suv}	-0.358	-0.419	-0.299
$Constant_{sports}$	-0.084	-0.206	-0.084
FS_{sports}	-0.052	-0.145	0.058
$Constant_{price}$	0.437	0.375	0.498
$Income_{price}$	0.023	0.004	0.042
FuelCost	-1.701	-1.838	-1.584
$Constant_{domestic}$	0.743	0.674	0.806
$BC_{Domestic}$	-0.181	-0.293	-0.071

In Bayes model 1, we assume that consumers only consider vehicles when their price under consumers' budget. Therefore, we found that consumers prefer more expansive vehicles in the consideration set, especially for the higher income consumers. The vehicle body type of Van is particularly preferred by the families with many persons. In the consideration set, consumers still prefer the high fuel efficiency vehicles, and domestic manufactured vehicles.

The average of the consideration set size among consumers is 91, which is already smaller than the average consideration set size in the Hazard-based MNL model. In the Bayes model 2, we would like to use both price and fuel efficiency (Km/L) attributes in the screening rule, which are also used in the hazard-based MNL model. Although the mechanism is different between these two models, we would compare the performance of two models. The formulation of the screening is:

$$I(x,\gamma) = I\left(-Ln(x_{Price}) > \gamma_{Price}\right) * I\left(-Ln(x_{FC}) > -\gamma_{FC}\right) \quad [4.23]$$

The estimation result of the cutoff values in Bayes Model 3 is as follows:

Table 4.3: Estimation result of cutoff values in Bayes Model 2

	Threshold	
	-Ln(Price)	-Ln(FC)
Mean	-7.845	-2.019
Variance	0.866	0.116

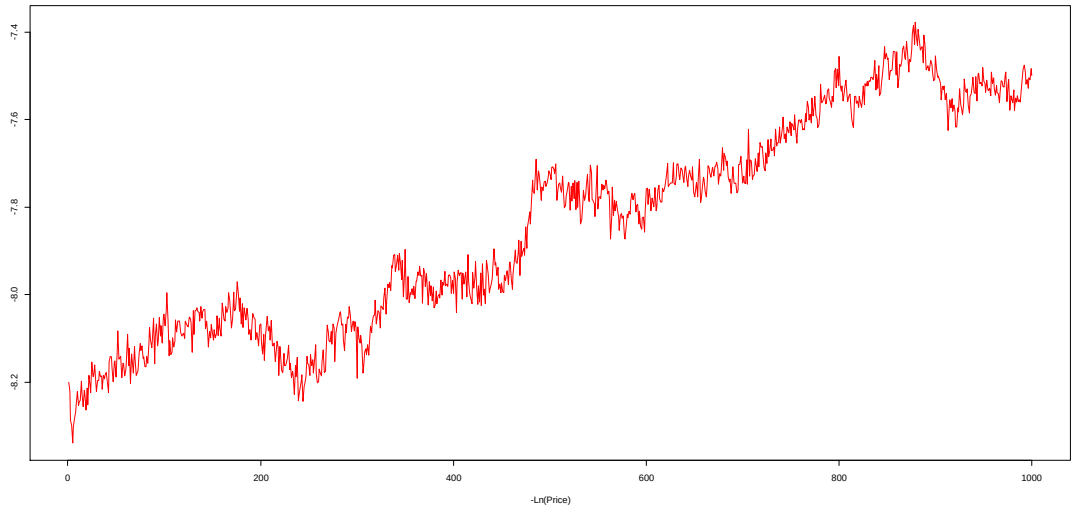

Figure 4-2: MCMC Sequence of -Ln(Price) in Bayes Model 2

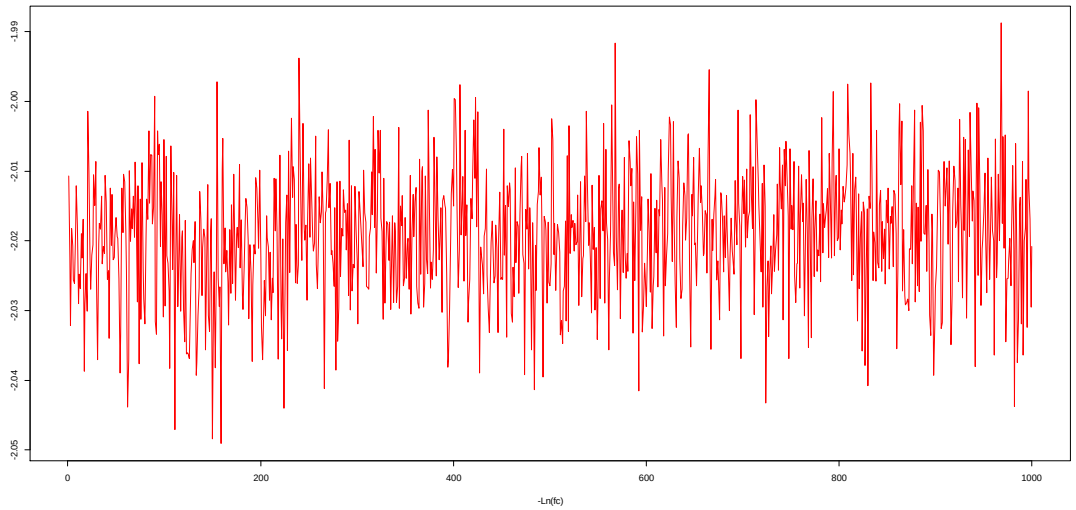

Figure 4-3: MCMC Sequence of -Ln(FC) in Bayes Model 2

According to the estimation result. The posterior mean of the price threshold is 25.5 million Yen which means the price would be useless in the screening rule because almost vehicles are cheaper than this price threshold. According Figure 4.2, it is obvious that the price threshold is not steady in the MCMC sequence. The posterior mean of the fuel cost as a threshold is 7.5 Yen/km. The posterior mean of the consideration set size is 80 which is bigger than the number in Bayes model 2. Therefore, introducing fuel cost attribute into the screening rule would not improve the efficiency in the consideration set formation process.

Table 4.4: Estimates for the parameters in Bayes Model 2

Parameter	Posterior Mean	2.5% Quantiles	97.5% Quantiles
$Constant_{compact}$	-0.418	-0.744	-0.110
$FS_{compact}$	-0.064	-0.348	0.312
$Constant_{sedan}$	0.620	0.262	0.910
FS_{sedan}	-0.444	-0.743	-0.076
$Constant_{wagon}$	-1.00	-1.258	-0.729
FS_{wagon}	-0.474	-0.844	-0.132

$Constant_{van}$	-0.803	-1.057	-0.531
FS_{van}	0.998	0.682	1.292
$Constant_{suv}$	0.751	0.348	1.125
FS_{suv}	-0.876	-1.231	-0.490
$Constant_{sports}$	0.170	-0.069	0.478
FS_{sports}	-0.645	-0.976	-0.417
$Constant_{price}$	-1.667	-1.840	-1.493
$Income_{price}$	0.016	-0.048	0.077
FuelCost	3.724	3.504	4.038
$Constant_{domestic}$	0.890	0.698	1.114
$BC_{domestic}$	-0.126	-0.395	0.112

In Bayes model 2, as the price threshold is high, consumer would like to choose the cheap vehicle in the consideration set. However, as the fuel efficiency is already used in the screening rule, and vehicles in the consideration set will satisfy consumers' requirement on the fuel efficiency. Therefore, they would like to choose low fuel efficiency (high fuel cost) vehicles in the consideration set. Generally speaking, high fuel efficiency is always caused by the new developed technology, and the new technology is always expansive in the beginning. For other attributes, minivan is still preferred by families with more persons. Consumers are still interested to domestic vehicles in the consideration set.

The advantage of this Bayes model is that we can introduce dummy variable into the screening rule which is difficult in the hazard-based model. Therefore, in the next model, we would like to introduce the vehicle body type and origin of manufactory to screen alternatives. The formulation of the screening rule is as follows:

$$I(x,\gamma) = I\left(x_{BodyType} > \gamma_{BodyType}\right) * I\left(x_{Domestic} > \gamma_{Domestic}\right) * I\left(x_{Forgien} > \gamma_{Foreign}\right) \quad [4.24]$$

Here, we consider three levels of cutoff values, (-0.5, 0.5, 1.5), which indicate that (1)

consumers do not use this attribute to screening alternatives; (2) consumers will consider the alternative with this attribute; and (3) consumer will exclude the alternative with this attribute. The model using this screening rule is named as Bayes Model 3. The estimation result of cutoff values is shown in the following table:

Table 4.5: Estimation result of cutoff values in Bayes Model 3

No	Attribute	Probability of each cutoff		
		$\gamma_1 = -0.5$	$\gamma_2 = 0.5$	$\gamma_3 = 1.5$
1	Kei	0.366	0.470	0.164
2	Compact	0.273	0.297	0.430
3	Sedan	0.068	0.067	0.865
4	Wagon	0.043	0.045	0.912
5	Van	0.323	0.327	0.350
6	SUV	0.077	0.065	0.858
7	Sports car	0.017	0.018	0.965
8	Domestic	0.457	0.539	0.004
9	Foreign	0.060	0.070	0.870

According to the result, only 16.4% of consumers will surely exclude the Kei car in the consideration set formation stage. The number of consumers who may consider compact or van in the consideration set formation stage is nearly twice as the number of consumers who purchased these two types of vehicles. For other vehicle types, the number of consumer considering these vehicle types is close to the number of consumers purchased these vehicle types. Almost no consumer will exclude domestic vehicles in the consideration set formation stage, but there are 87% of consumers will

exclude foreign vehicles from their consideration set. In this model, the posterior mean of the consideration set size is 86 which is smaller than the number in Bayes Model 1 and 2 where continuous variables were used in the screening rule. The parameter estimation is shown in the following table:

Table 4.6: Estimates for the parameters in Bayes Model 3

Parameter	Posterior Mean	2.5% Quantiles	97.5% Quantiles
$Constant_{compact}$	0.777	0.586	0.947
$FS_{compact}$	-0.161	-0.257	-0.033
$Constant_{sedan}$	1.501	1.303	1.780
FS_{sedan}	-0.301	-0.403	-0.200
$Constant_{wagon}$	1.120	0.729	1.461
FS_{wagon}	-0.630	-0.855	-0.397
$Constant_{van}$	-1.253	-1.345	-1.151
FS_{van}	1.367	1.252	1.484
$Constant_{suv}$	1.628	1.365	1.927
FS_{suv}	-0.320	-0.448	-0.214
$Constant_{sports}$	1.027	0.732	1.378
FS_{sports}	0.126	0.025	0.236
$Constant_{price}$	-0.146	-0.185	-0.106
$Income_{price}$	0.036	0.017	0.053
FuelCost	-0.549	-0.579	-0.518
$Constant_{domestic}$	0.336	0.232	0.441
$BC_{domestic}$	-0.828	-0.938	-0.725

According to the estimation result, the constant of vehicle body type is bigger if the number of consumer who consider this vehicle body type is close to the number who purchased this type. It means that consumers who purchased these vehicle types would already have the preference in the consideration set formation stage. Big family prefer Van which has a big capacity for passengers. Furthermore, the family size is also has a

positive effect on Sports vehicles. It is because the Sports vehicles is not always purchased as the only vehicles in the family and the size of family with more than one vehicles is always not small. In Bayes model 3, the price is not used in the screening rule. Therefore, we found that the sign of the price constant is negative which means that consumers would like to choose the cheap vehicle from the consideration set. The high fuel efficiency vehicles is also preferred in the consideration set generated by the screening rule in Bayes model 3.

In Bayes Model 3 there are three cutoff levels for the dummy variable including an assumption that consumer do not use this attribute to screen alternatives. However, this assumption sometimes would make the estimation result hard to explain. For example, in Table 4.5. There are 27.3% consumers will not use attribute “compact” to screen alternatives, for these consumers, compact vehicles would appear or disappear in their consideration set. There are 29.7% consumers who included compact vehicles in their consideration set and 31.4% who finally purchased compact vehicles. Therefore, there are at least 1.7% of consumes who did not use compact to screening alternatives but included compact vehicles into their consideration set. However, we cannot determine the accurate value according to the estimation result. Therefore, in Bayes Model 4, we would like to define two levels of cutoff values for dummy variables. We only assume two situations, the first is consumer will accept this attribute, the second is consumer will exclude the alternative with this attribute. The formulation of the screening rule is same as Equation 4.25. The estimation result of cutoff values is shown in the following table:

Table 4.7: Estimation result of cutoff values in Bayes Model 4

No	Attribute	Probabilities of cutoff values	
		γ_1 =0.5	γ_2 = 1.5
1	Kei	0.786	0.214
2	Compact	0.609	0.391
3	Sedan	0.129	0.871
4	Wagon	0.060	0.940
5	Van	0.579	0.421
6	SUV	0.142	0.858
7	Sports car	0.017	0.983
8	Domestic	0.995	0.005
9	Foreign	0.101	0.899

According to the estimation result, Kei, compact and minivan are mostly considered by Japanese consumers. In one hand, it shows the popular of these three vehicle body types in Japan, in the other hand, it also indicates that a significant number of consumers will choose other vehicle body types although they include these kind of vehicles in their consideration sets. Almost all consumers will consider domestic brands and there only 10% consumers will consider foreign brands in purchase vehicle. Finally, the posterior mean of the consideration set size is 81 which is smaller than the number in Bayes model 3. This result shows the two levels of cutoff values for dummy variable is more efficient in consideration set formation process than three levels. The following table shows the estimation of parameters in Bayes Model 4:

Table 4.8: Estimates for the parameters in Bayes Model 4

Parameter	Posterior Mean	2.5% Quantiles	97.5% Quantiles
$Constant_{compact}$	0.333	0.187	0.484
$FS_{compact}$	-0.171	-0.317	-0.041
$Constant_{sedan}$	1.099	0.958	1.252
FS_{sedan}	-0.207	-0.291	-0.120
$Constant_{wagon}$	0.887	0.722	1.012
FS_{wagon}	-0.316	-0.425	-0.167
$Constant_{van}$	-0.958	-1.074	-0.849
FS_{van}	1.101	0.973	1.232
$Constant_{suv}$	0.913	0.779	1.031
FS_{suv}	-0.011	-0.110	0.076
$Constant_{sports}$	1.234	1.066	1.397
FS_{sports}	0.016	-0.059	0.123
$Constant_{price}$	-0.075	-0.110	-0.036
$Income_{price}$	0.035	0.016	0.054
FuelCost	-0.551	-0.586	-0.516
$Constant_{domestic}$	0.077	0.003	0.144
$BC_{domestic}$	-0.197	-0.310	-0.108

In Bayes model 4, all parameters have the same signs as them in Bayes model 3. It means that consideration sets for each consumers would be partly same between two models. However, in Bayes model 4, the estimation result of the cutoff values would be easily understood.

By now, we could found that:

1. Attribute of price is a distinctly efficient cutoff value in the screening rule.
2. Introducing fuel efficiency into the screening rule cannot improve the efficient in consideration set formation process.
3. For the dummy variable, two levels of cutoff values is better than three levels.

Because not only the size of consideration set will be reduced, but also the estimation result would be explained clearly.

Therefore, according to the previous estimation results, we would like to introduce a new screening rule which including two level cutoff values of dummy variables and upper and lower limits of price. The model with this screening is named as Bayes Model 6. The formulation of the screening rule is as follows:

$$I(x,\gamma) = I\left(x_{BodyType} > \gamma_{BodyType}\right) * I\left(x_{Domestic} > \gamma_{Domestic}\right) * I\left(x_{foreign} > \gamma_{foreign}\right) * I\left(-Ln(x_{Price}) > \gamma_{Price}\right) \quad [4.25]$$

According to this new screening rule, only when the alternative satisfy the consumer's vehicle body type constraint and the upper and lower price constraints at the same time, this alternative will be considered to purchase. The estimation result of the cutoff values are shown in the following two tables:

Table 4.9: Estimation result of cutoff values of Price in Bayes Model 5

	-Ln(Price)
Mean	-5.336
Variance	0.206

Table 4.10: Estimation result of cutoff values of vehicle body types in Bayes Model 5

No	Attribute	Probabilities of cutoff values	
		γ_1 =0.5	γ_2 = 1.5
1	Kei	0.547	0.453
2	Compact	0.549	0.451

3	Sedan	0.387	0.613
4	Wagon	0.082	0.918
5	Van	0.839	0.161
6	SUV	0.348	0.652
7	Sports car	0.162	0.838
8	Domestic	0.996	0.004
9	Foreign	0.205	0.795

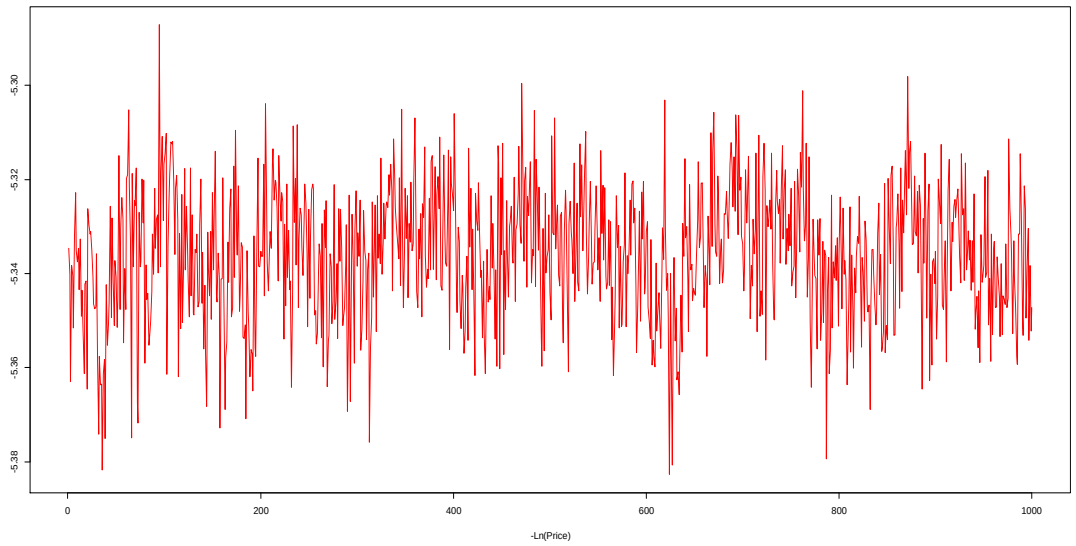

Figure 4-4：MCMC Sequence of -Ln(Price) in Bayes model 5

According to these estimation results, the posterior mean of the price threshold is 2.08 million Yen which is bigger than it in Bayes model 1. The posterior mean of probabilities of cutoff values for vehicle body types are shown in Table 4.10. There are some difference with Bayes model 4. The number of consumers who consider Kei and compact vehicles is slightly decreased but still much bigger than the number of consumers who purchased these kinds of vehicles. For other body types of vehicles, especially for sedan and van, there are more consumers consider them in Bayes model

5. Furthermore, foreign vehicles are also more included into the consideration set in Bayes model 5. However, the mean of consideration size in Bayes model 5 is 46 which is the smallest among these models. It shows the advantage of including both price, vehicle body types and origin of manufactory into the screening rule.

Table 4.11: Estimates for the parameters in Bayes Model 5

Parameter	Posterior Mean	2.5% Quantiles	97.5% Quantiles
$Constant_{compact}$	-1.269	-1.81	-0.754
$FS_{compact}$	-0.081	-0.380	0.228
$Constant_{sedan}$	-1.216	-1.722	-0.597
FS_{sedan}	-0.483	-0.669	-0.271
$Constant_{wagon}$	-0.816	-1.187	-0.473
FS_{wagon}	-0.660	-1.035	-0.304
$Constant_{van}$	-2.613	-3.032	-2.104
FS_{van}	1.256	1.026	1.500
$Constant_{suv}$	-0.847	-1.320	-0.207
FS_{suv}	-0.002	-0.275	0.220
$Constant_{sports}$	-1.569	-2.003	-0.905
FS_{sports}	-0.087	-0.345	0.177
$Constant_{price}$	1.295	0.916	1.208
$Income_{price}$	0.018	-0.039	0.077
FuelCost	-2.231	-2.427	-2.047
$Constant_{domestic}$	-0.745	-0.977	-0.529
$BC_{domestic}$	0.142	-0.263	0.458

According to the estimation result of parameters, parameter relating to vehicle bod types and family size all have the same signs as them in Bayes model 4. That is because they all use the vehicle body types in screening rules. However, the sign of price constant is positive in Bayes model 5. It is same in Bayes model 1 where the price is used to screen alternative vehicles. Based on these estimation results, we could found

that based on the different consideration set generated by different screening rules, the consumers' preference would be different.

4.4 Model performance comparison

In Chapter 2 and Chapter 3, we estimated two choice models with consideration set under the same data, and we tried different screening rules in Bayes models. The estimation method is different, one is maximum likelihood estimation method and the other is a Bayesian approach. The assumption of error component in the utility function is also different which brings one to MNL model and the other to MNP model. Most of all, consideration sets in these models are generated in different ways. Therefore, we would like to use some indexes to find out the difference between these models.

Table 4.12: Means of consideration set size

Model	Mean of the consideration set size
Hazard-based MNL model	105
Bayes Model 1	91
Bayes Model 2	80
Bayes Model 3	86
Bayes Model 4	81
Bayes Model 5	46

In Table 4.12, it shows the average of generated consideration set size in different models. Generally speaking, consideration set sizes in all Bayes models are small than the size in Hazard-based MNL model. In Bayes models, adding fuel efficiency in the

screening rule did not reduce the average consideration set size. Assuming two levels cutoffs for the dummy variable is better than three levels in narrow the consideration set. Finally, we use price and dummy variables in the screening rule in Bayes model 5, and the average consideration set size is the smallest among these models.

One of the aim for analyzing choice behavior is forecasting the future. Therefore, we would like to compare the models' performance in prediction accuracy. Although two models applied the same data, the size and content of consideration set for each decision-makers was different in each model. Therefore, some common indexes are improper to be applied to compare the performance. Here, we would like to check the prediction accuracy of vehicle's fuel efficiency of these models.

Table 4.13: Prediction of Fuel Efficiency (km/L)

	Predicted Fuel Efficiency (km/L)
Hazard-based MNL model	23.5
Bayes Model 1	23.6
Bayes Model 2	29.3
Bayes Model 3	29.5
Bayes Model 4	28.5
Bayes Model 5	22.6

The average fuel efficiency of the purchased vehicles is 22.2 km/L. Hazard-based MNL model and Bayes model 1 have the almost same prediction accuracy on predicted vehicles' fuel efficiency. In Bayes model 2, it is assumed that consumers will include vehicles if the vehicle's fuel efficiency is bigger than consumers' threshold, so the high

fuel efficiency vehicles are always included in the consideration set. Therefore, it is not surprised that the mean of predicted vehicles' fuel efficiency is big. In Bayes model 3 and Bayes model 4, the mean of predicted vehicles' fuel efficiency is even bigger the value in Bayes model 2. In these two model, only vehicle body types and the origin of manufactory are used in the screening rule. If the consumer is prefer one of the vehicle body type, no matter the fuel efficiency and the price of the vehicle, it will be included into the consideration. Therefore, even the average consideration set size of these two models are smaller than other models. The rationality of the consideration set generated only by the vehicle body types as the constraint should be doubted.

4.5 Vehicle choice behavior analysis under different vehicle purchasing scenarios

Consumer's preference in the consideration set formation stage would be influence by various reasons. For example, if the family does not have any vehicles before, their focus on vehicle features could be different with the family who already owned vehicles but want to buy an additional one. Therefore, in this section, we would like to compare the consideration set difference for families with different number of vehicles.

Generally, the vehicle purchase situations would be divided according to the number of consumer owned vehicles. The first situation is that consumers do not own vehicles before they decide to buy a new vehicle, it means that the new purchased vehicle would be their only vehicle. The second situation is that consumers already have one or more vehicles before they purchase the new vehicle, it means that the new purchased vehicle is not the first and only vehicle for these consumers. According to the definition, we divided the data set into two groups. The first group is called single-vehicle group, and

the second group is called multi-vehicle group. Table 4.14 shows the share of different vehicle body types in two groups

Table 4.14 The share of vehicle types in two groups

Vehicle body Type	Single-vehicle group	Multi-vehicle group
Kei	0.299	0.400
Compact	0.339	0.299
Sedan	0.069	0.056
Wagon	0.021	0.011
Van	0.222	0.169
SUV	0.050	0.061
Sports	0.000	0.004

We apply the same screening rule and utility function in Bayes model 5. The estimation results are shown in the following tables.

Table 4.15 Estimation result of cutoff values of vehicle body types

	Single-vehicle group		***Multi-vehicle group***	
	Probabilities of cutoff values			
Attribute	**γ_1 =0.5**	**γ_2 = 1.5**	**γ_1 =0.5**	**γ_2 = 1.5**
Kei	0.462	0.538	0.654	0.346
Compact	0.631	0.369	0.552	0.448
Sedan	0.450	0.550	0.394	0.606
Wagon	0.216	0.784	0.110	0.890
Van	0.826	0.174	0.736	0.264

SUV	0.336	0.664	0.404	0.596
Sports	0.259	0.741	0.209	0.791
Domestic	0.987	0.013	0.992	0.0008
Foreign	0.198	0.802	0.176	0.824

Table 4.16 Estimation result of cutoff values of Price in two groups

	Single-vehicle group	***Multi-vehicle group***
	-Ln(Price)	
Mean	-5.308	-5.360
Variance	0.182	0.236

According to Table 4.14 and 4.15, generally speaking, if the share of vehicle body type is big, it means that there are more consumers consider this kind of vehicle body type. For different vehicle body types in two groups, even the share in two groups only has a little difference, the number of people who consider this vehicle body type would have a significant difference in two groups. Kei car is popular in both groups, but there are more consumers who consider and buy this kind of vehicle. For these four vehicle body types including compact, sedan, wagon and Van, consumers in single-vehicle group have more interests on them. For SUV and sports vehicles, consumers in multi-vehicle groups would like to consider and buy them. This result indicates that in single-vehicle group, because the purchased vehicle is the only vehicle in this family, therefore, the vehicle's practical applicability would be the most important features to be considered. The purchased vehicle should meet the demand of all sorts of use, therefore, vehicle body types such as compact, sedan, wagon and Van take big shares in the single-vehicle group. For consumers who intend to buy an additional vehicle, vehicles with special

features would be more popular. For example, Kei car is very cheap but the fuel efficiency is very high, SUV is a big vehicle and could deal with tough road conditions, sports car always has a powerful engine, dazzling appearance and of course high price. We could found that these special vehicles are preferred in multi-vehicle group. The purchase budget in two groups are: 2.02 million Yen and 2.13 million Yen. Generally, family owing multi-vehicles is richer than family with only one vehicle, therefore, even they are buying an additional vehicle, the budget is higher than the single-vehicle family. The mean of consideration set size in single-vehicle group is 44 and 51 for multi-vehicle group. The high budget would be one of reason for this result.

Table 4.17 Estimates for the parameters in two groups

	Single-Vehicle group	***Multi-Vehicle Group***
Parameter	**Posterior Mean**	**Posterior Mean**
$Constant_{compact}$	-2.042**	-0.631**
$FS_{compact}$	0.185	-0.134
$Constant_{sedan}$	-2.403**	-0.465*
FS_{sedan}	0.688**	-0.579*
$Constant_{wagon}$	-3.332**	0.115
FS_{wagon}	1.077**	-1.322**
$Constant_{van}$	-2.600**	-2.654**
FS_{van}	2.146**	1.712**
$Constant_{suv}$	-1.416**	-0.444
FS_{suv}	0.403*	-0.087
$Constant_{sports}$	-2.847**	-1.791**
FS_{sports}	0.790**	0.754**
$Constant_{price}$	1.324**	0.840**
$Income_{price}$	-0.060	0.003
$FuelCost$	-2.530**	-1.907**
$Constant_{domestic}$	-0.587**	-1.044**
$BC_{domestic}$	-0.522	-0.390

**: 95% posterior mean away from 0
*: 90% posteriors mean away from 0.

Table 4.17 shows the parameter estimation result in two groups. The significant difference is the impact of family size on different vehicle body types. In multi-vehicle group, if the family size is big, the utility of compact, sedan and wagon would be decreased which is different in single-vehicle group.

4.6 Conclusion

This study analyzed the Japanese consumers' vehicle purchasing behavior from a two-stage decision process perspective. When all vehicle models being sold were included in the choice set, consumers only evaluated those models that they would like to consider. In the applied Hierarchal Bayes model, parameters in the consideration set formation stage and choice-making stage were estimated simultaneously, and the estimation process only required minimal information about consumer decisions.

Firstly, we tried different screening rules including only price, price and fuel efficiency, dummy variable (vehicle body type and origin of manufactory) in three levels cutoffs, dummy variable in two level cutoffs, and dummy variables in two level cutoffs and price. By different screening rules, the consumers' preference in the choice making stage would be different. For example, if we assume consumers will only consider vehicles under their purchase budget, then they will prefer expansive vehicles in the consideration set.

Although the consideration set formation process are totally different in these models, the size and the rationality of generated consideration sets are what we are concerned.

The size of the consideration set is much smaller in Bayes models than the size of the universal set. The screening rule including price, vehicle body type and origin of manufactory generated the smallest consideration sets among these models. The rationality of these consideration sets is measured by the difference between the mean of predicted vehicles fuel efficiency and the mean of purchased vehicles fuel efficiency. Although the size of consideration set in Bayes model 1 to model 4 is smaller than the size of consideration set in Hazard-based MNL model, the prediction accuracy of hazard-based MNL model is better than Bayes model 1 to 4. Bayes model 5, which has the smallest size of consideration set, also has the best prediction accuracy among all two-stage choice models. It indicates that the best screening rule is not only narrowing down the size of generated consideration set, but also consistent in consumers' screening logical.

Furthermore, we also compared the choice behaviors under different vehicle purchasing scenarios. For example, in the consideration set formation stage, if the consumer intend to buy the first and only vehicle, the compact vehicle, sedan, wagon and SUV are significantly more popular than the consumer who intends to buy an additional vehicle. Oppositely, Kei car and sports are preferred for the additional vehicle choice. In the choice making stage, the significant result is that when the family size is big, compact vehicle, sedan, wagon and SUV's utility will be decreased, it shows that for the additional vehicle, the consumer would like to choose the high riding capacity vehicle, such as Van, or distinctive vehicles, such as Kei car which is cheap but has a very high fuel efficiency or sports car which has a good appearance and high performance. The research indicate the consumers' difference in both consideration set formation and choice making stage, and the result would give a direction for vehicle manufacture

companies and government to make production plan and vehicle policy.

This study has some limitations. In the consideration set formation stage, we screened vehicles using only several basic vehicle attributes. If there were a greater number of attributes, we could examine their effects on formatting consideration sets. Future studies will analyze the consideration set formation stages in greater detail. Not only a greater number of vehicle attributes but also psychological factors in the consideration set formation stage may be considered. In comparing the vehicle purchasing behavior in different scenarios, if the sample size would be large enough and more consumers' information were available, we would have more specific analysis of consumers' behaviors in both consideration set formation stage and choice making stage. Additionally, other estimation methods, such as machine learning, may be applied to see whether or not they can produce better estimation results

4.7 References

1. Gilbride, T. J., & Allenby, G. M. (2004). A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules. *Marketing Science, 23*(3), 391–406. doi:10.2307/30036705
2. Laroche, M., Kim, C., & Matsui, T. (2003). Which decision heuristics are used in consideration set formation?. *Journal of Consumer Marketing*, *20*(3), 192-209.
3. Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398), 528–540

5. Conclusion

5.1 Conclusions

In this dissertation, we introduced the consideration set into the choice analysis in transportation area. In some transportation research, the number of available alternatives is extremely large, therefore it is not reasonable to assume decision-makers will evaluate all alternatives to make a final decision. Consideration set is a subset of the universal set which includes all available alternatives. Alternatives in the consideration set are these who satisfy decision-makers' constraint on one or more aspects of alternatives. However, the information of each decision-makers' consideration set is barely known by the researchers. Researchers should estimate the consideration set formation and choice behavior with only observed final choice. In these studies, the real data rather than the simulated data were applied in all studies. Different two-stage choice models were compared under the same choice analysis, and the changing of the alternatives in the consideration set were considered in different choice scenarios.

In this dissertation, we introduced different two-stage choice models into choice analysis in transportation area with the real data. Only the observed choice information is available, and there is none of the information about decision-makers' consideration set. Therefore, it required to estimate the consideration set formation stage and choice making stage with the only observed choice information. According to the different situations, the choice-set explosion was applied in the route choice analysis where the number of alternatives in the choice set is moderate. Actually, the number of possible

routes for an OD pair would be numerous, however the number of routes decided by researchers in the choice set would be moderate. Therefore, the choice-set explosion method is best work here. Furthermore, we also considered the change of the consideration set during the driving by introducing a progress indicator. In the vehicle purchasing analysis, all available vehicles in the market were included in the choice set, therefore, the universal set size would be big and the choice-set explosion method is infeasible anymore. Therefore, we applied the hazard-based choice-set formation model to generate the consideration set in advance and then analyzed with the MNL model. Additionally, we also applied a Bayesian method to estimate the consideration set formation stage and choice making stage when the universal set size is big. Various screening rules were tried and the screening rule with vehicle price and vehicle body types gave the best result. In comparing the model's performance, the Bayesian model with the best screening rule had the better prediction accuracy than the MNL model with the hazard-based choice set formation model. Furthermore, we also found that here are significant difference in the consideration set formation stage and choice making stage for consumers whether owned vehicles or not before buying a new vehicle.

5.1.1 Route choice behavior considering probability choice sets.

In this study, we analyzed how drivers would make the following route choice while they are driving on the route, and both consideration set formation stage and choice making stage were simultaneously estimated. The constraint-based choice set formation model was applied in the consideration set formation stage. Drivers were assumed to screen the alternative routes by the number of turns of each route. As all non-empty

subsets of the universal set would be the potential consideration set, therefore when the size of universal is moderately big (more than 5), the difficulties of estimation would be increased exponentially. The method proposed by Morikawa (1996) was applied to solve this issue.

The estimation result indicated that routes with less turns have the higher probabilities to be included into consideration sets. Furthermore, for drivers in different stages of their trips, such as initial stage, intermediate stage or end stage, the probability of the route to be included into the consideration set is not steady but fluctuant.

5.1.2 Vehicle choice analysis with hazard-based choice set formation model.

In this study, we applied MNL model to analyze vehicle choice behavior with the consideration set generated by the hazard-based choice set formation model. We included all available vehicle types in the Japan market into the universal set. In the consideration set formation stage, we used the hazard-based model to calculate the probability of each vehicle to be considered. Then, vehicles were randomly selected according to the probability into the consideration set, the high probability means the higher chance to be considered. Then, the MNL model with a correction term was applied with these consideration set to analyze the vehicle choice behaviors.

In order to measure the performance of the MNL model with the consideration set, we estimated another two models. One is the MNL model with the universal set, where consumers were assumed to evaluate all available vehicles to make a choice. The other is MNL model with randomly generated choice set, where alternatives were randomly

selected from the universal set, and the size of random choice set for each consumer is same as the size of his/her consideration set. Indexes used to measure the model performance were average probability of corrected prediction and hit ratio. Both indexes indicated that the MNL model with consideration set has the best performance among these three models.

5.1.3 Vehicle choice analysis with a hierarchical Bayes model

In this study, we applied a hierarchical Bayes model (Gilbride and Allenby, 2004)to analyze the vehicle choice behavior with same data in Chapter 3. Consumers were assumed to screen alternative vehicles by three variables: vehicle body type, country of origin (Japan or foreign countries), price. A conjunctive screening rule was applied which means only all these three variables of vehicles satisfy consumers' requirements, the vehicle could be included into the consideration set. In the choice making stage, the MNP model was applied. Both two stages were estimated simultaneously by the MCMC method. The estimation result indicated consumers' preference in screening alternative vehicles before purchasing. It should be noted that the estimation result shows that consumers would like to choose the more expansive vehicles in the consideration set while all vehicles in the consideration set are under the purchasing budget.

We applied two different choice model to analyze the vehicle choice behavior with consideration set. Therefore, we would like to compare two models' performance. Firstly, the average size of consideration set is smaller in the hierarchical Bayes model. Secondly, the hit ratio of the model in Chapter 3 is bigger than the model in Chapter 4.

However, the model in Chapter 4 is more accurate in predicating the average vehicles' fuel efficiency. It means that vehicles in the consideration set in Chapter 4 were more similar by the screening rule. Therefore, even the hit ratio is lower than the model in Chapter 3, the prediction ability is better than it. Finally, we could concluded that the model in Chapter 4 would estimate the vehicle choice behavior with the consideration set more efficient and reasonable.

5.2 Recommendations for future work

There are several research directions which would improve the current work. Firstly, if the consideration set information are available, we would directly estimate the choice model with observed consideration sets. Furthermore, we could compare the observed consideration sets with consideration sets generated by the model. Then we would evaluate the performance of these consideration set formation models. Secondly, decision-makers' psychological factors would influence the formation of consideration set, we would like to investigate how these factors would be incorporated into the consideration set formation model. Finally, due to the estimation difficulties caused by the unobserved consideration sets, other estimation methods, such as machine learning, should be applied to see whether these methods would improve the accuracy in estimating the consideration set formation process.

In this dissertation, we only estimated the route choice and vehicle purchasing choice with the consideration set. There are still many other choice situations in the transportation area where the number of available alternative is huge, therefore, it is worth to apply the choice model with consideration set in these research area.

5.3 Reference

1. Gilbride, T. J., & Allenby, G. M. (2004). A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules. *Marketing Science, 23*(3), 391–406. doi:10.2307/30036705
2. Morikawa, T. (1996). A Hybrid Probabilistic Choice Set Model with Compensatory and Noncompensatory Choice Rules. Volume 1: Travel Behavior. In *World Transport Research. Proceedings of the 7th World Conference on Transport Research*