Analysis of Drivers’ Route Choice Behavior

Considering Probability Choice Sets

Gang Xu
Graduate School of Environmental Studies, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
(e-mail) xu.gang @g.mbox.nagoya-u.ac.jp
Tel: +81-52-789-3730
Fax: +81-52-789-5728

Tomio Miwa
EcoTopia Science Institute &
Green Mobility Collaborative Research Center,
Nagoya University, Japan
(e-mail) miwa@civil.nagoya-u.ac.jp

Takayuki Morikawa
Graduate School of Environmental Studies &
Green Mobility Collaborative Research Center,
Nagoya University, Japan
(e-mail) morikawa@nagoya-u.jp
Abstract

In this work, 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.
1. Introduction

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 technological developments, 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 analysis of how travelers select possible routes for a trip and which route they actually choose to take. 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.

1.2 Past Research

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. 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 from which the driver ultimately makes a choice (Kaplan, 2012). Figure 1.1 illustrates the relationships among these sets.

![Fig 1.1: Route choice set from drivers’ perspective.](image)

Early researchers (McFadden, 1975) 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, 2007)

### 1.3 Research Objectives

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.

### 1.4 Paper Structure

The remainder of this paper is structured as follows. Section 2 explains the models used in the investigation. Section 3 describes the data source and the data processing used. Section 4 presents the estimation results and includes some discussion of the results. In Section 5, the conclusions are given and a direction for future research is proposed.
2. Models

The various approaches lead to different models of route choice behavior; Figure 2.1 illustrates the situation (Frejinger, 2007):

![Route choice modeling diagram]

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 $P(i|C)$ 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.
2.1 Multinomial Logit Model

The most widely used models in route choice analysis would be the multinomial logit (MNL) (Ben-akiva et al., 1985), Probit (Sheffi, 1985) and modified Probit and Logit models, including C-Logit (Cascetta, et al., 1996), Path-Size (Ben-akiva and Bierlaire, 1999) and the Mixed Logit model (Train, 2003).

In the MNL model, choice makers are assumed to be rational decision makers who choose the alternative with the greatest utility. Generally, the utility function comprises two additive parts as follows:

\[ U_{in} = V_{in} + \epsilon_{in} \]

Where

- \( V_{in} \): the systematic component of the utility of alternative \( i \) to individual \( n \);

- \( \epsilon_{in} \): disturbances.

Usually, the systematic component is a linear function of the parameter:

\[ V_{in} = \beta^* X_{in} \]

Where

- \( \beta \): the vector of unknown parameters to be estimated;

- \( X_{in} \): the explainable variables of alternative \( i \) to individual \( n \).

If disturbances are assumed to follow a Gumbel distribution, the probability of choosing alternative \( i \) in the MNL model would be as follows:

\[ P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in M} e^{V_{jn}}} \]

Where:

- \( P_n(i) \): the probability of individual \( n \) choosing alternative \( i \);

- \( M \): the choice set including all of the alternatives.
2.2 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:

$$P_n(i) = \sum_{C \in G} P_n(i|C) * Q_n(C|G)$$

Where

- $P_n(i)$: probability of individual $n$ choosing route $i$ from master set $M$;
- $P_n(i|C)$: probability of individual $n$ choosing route $i$ from given choice set $C$;
- $G$: set of all non-empty subsets of $M$;
- $Q_n(C|G)$: probability of individual $n$’s choice set being $C$.

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:

$$q_n(i) = \prod_{k=1}^{K} q_{kn}(i)$$  \hspace{1cm} (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:

$$E_{kn}(i) = \alpha_k * w_{in} - \delta_m$$  \hspace{1cm} (6)

Where:

- $\alpha_k$: vector of unknown parameters to be estimated;
- $w_{in}$: vector of variables affecting the constraints;
- $\delta_m$: disturbance.

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

$$q_{kn}(i) = \text{Prob}(E_{kn}(i) \geq \mu_k) = \text{Prob}(\alpha_k * w_{in} - \delta_m \geq \mu_k) = \text{Prob}(\delta_m \leq \alpha_k * w_{in} - \mu_k)$$  \hspace{1cm} (7)
In this equation, $\mu_k$ is the threshold value of the $k$-th constraint. If we assume $\delta_n$ follows a logical distribution, Equation (7) and Equation (5) can be rewritten as:

$$q_{in}^i(i) = \frac{1}{1 + e^{-(q_i + \delta_n - \mu_k)}}$$

$$q_n(i) = \prod_{k=1}^{K} \frac{1}{1 + e^{-(q_i + \delta_n - \mu_k)}}$$

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

$$Q_n(C|G) = \frac{1}{1 - Q_n(\emptyset)} \times \prod_{j=1}^{M} q_n(j)^{d_c} \{1 - q_n(i)\}^{1-d_c}$$

Where:

- $Q_n(\emptyset)$: $\prod_{m=1}^{M} 1 - q_n(m)$, the probability of individual $n$'s choice set being empty;

- $d_c$: 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 (4) and (10), we obtain the following:

$$P_n(i) = \sum_{C \in G} P_n(i|C) \ast Q_n(C|G)$$

$$= \frac{1}{1 - Q_n(\emptyset)} \times \sum_{C \in G} \left\{ \sum_{k=1}^{K} e^{\lambda_n} \times \prod_{j=1}^{M} q_n(j)^{d_c} \{1 - q_n(i)\}^{1-d_c} \right\}$$

Equation (11) 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 (11) 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):

$$
P_n(i) = \frac{1}{1 - Q_n(\emptyset)} \times \text{Prob}(i \in C_n) \times \text{Prob} \left[ \left\{ (1 \in C_n) \cap (U_{in} \geq U_{jn}) \right\} \cup \left\{ 1 \notin C_n \right\} \right]
$$

and

$$
P_n(i) = \frac{1}{1 - Q_n(\emptyset)} \times q_n(i) \times \text{Prob} \left[ \bigcap_{j \in M, j \neq i} \left\{ j \in C_n \right\} \cap (V_{in} + \varepsilon_{in} - V_{jn} + \varepsilon_{jn} \geq 0) \right] \cup \left\{ j \notin C_n \right\}
$$

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 $n$;

$\varepsilon_{in}$: disturbance of utility of alternative $i$ for individual $n$.

Taking the conditional probability on the disturbance component of utility $\varepsilon_{in}$, Equation (12) can be written as:

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

Where:

$f(\bullet)$: probabilistic density function of $\varepsilon$;

$F(\bullet)$: cumulative density function of $\varepsilon$.

If the $\varepsilon$ is assumed to follow a Gumbel distribution, then
\[ P_n(i) = \frac{q_n(i)}{\sum_{m \in M} q_n(m)} \cdot e^{-\sum_{j \in M, j \neq i} \left( q_n(i) e^{v_{n,i} - v_{n,j}} + (1 - q_n(j)) \right)} \]

The calculation of the log-likelihood value would be complex due to the one-dimensional integral in Equation (14), so here we use Simpson’s rule to calculate it:

\[ \int_{a}^{b} f(x) \, dx \approx \frac{b-a}{6} \left[ f(a) + 4f\left( \frac{a+b}{2} \right) + f(b) \right] \]

At this point, Equation (14) can theoretically be applied to calculate the likelihood no matter how many alternatives there are in the master set. However, computation time would become a problem for bigger master sets. Furthermore, using this equation, the model can be estimated without knowledge about the decision makers’ choice set, but with knowledge of only the decision makers’ choice.
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 3.1 shows the data structure.

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

Figure 3.1 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 3.1 as the research target area as this is where probe vehicles passed most frequently.

The rectangular target area is shown in detail in Figure 3.2. 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 3.3 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.

Fig. 3.1: Accumulated number of probe vehicle passes on the road network of Toyota city

Fig 3.2: 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:

- **Gender:** the driver gender
- **Age:** the driver age
- **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

Figures 3.4 and 3.6 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 3.5 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.

![Driver characteristics for en-route OD-1](image1)

Fig 3.4: Driver characteristics for en-route OD-1

![Total trip distance distribution](image2)

Fig 3.5: 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 3.7, 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.
Fig 3.6: Driver characteristics for en-route OD-2

Fig 3.7: Total trip distance distribution for vehicles passing through the target area by en-route OD-2
4. Results and Discussion

In this research, the data for both en-route OD pairs are used together to estimate the models. The set of observed routes for each en-route OD is treated as the master set for the drivers.

4.1 MNL Model

In the MNL model, the systematic component of utility function is as follows, where values $\beta$ are the parameters to be estimated:

$$V_i = \beta_1 * \text{Distance} + \beta_2 * \text{ARR} + \beta_3 * \text{Turns}$$

Table 4.1: Estimation result of MNL model for en-route OD-1 (305 samples) and en-route OD-2 (144 samples)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Parameter</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (km)</td>
<td>-10.732</td>
<td>-3.8</td>
</tr>
<tr>
<td>ARRT</td>
<td>0.653</td>
<td>2.8</td>
</tr>
<tr>
<td>Turns</td>
<td>-1.475</td>
<td>-20.0</td>
</tr>
<tr>
<td>Log-likelihood at zero</td>
<td>-1177.212</td>
<td></td>
</tr>
<tr>
<td>Log-likelihood at estimates</td>
<td>-749.349</td>
<td></td>
</tr>
<tr>
<td>McFadden adjusted rho-bar squared</td>
<td>0.361</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>1504.7</td>
<td></td>
</tr>
</tbody>
</table>

The estimation result obtained with the MNL model is shown in Table 4.1. 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.

4.2 PCS Model

For the PCS model, the random utility model is applied in the choice making stage and the constraint-based choice set generation model is used in the choice set generation stage. The implementation of the choice making stage is simply as for the MNL model discussed above; therefore, the crucial part of the PCS model is the choice set generation stage.

This is a constraint-based generation model, so it is important to define the constraints at this stage. In many route choice modeling contexts, drivers are always assumed to have perfect information about the network, which is impossible. For a given OD pair, the differences in distance between all possible routes might be small, while perception of travel time changes according to traffic congestion. However, it is possible for a driver to know the number of turns entailed in a particular candidate route in advance, especially if the he or she knows the road network well. Then, as a result of his or her particular network knowledge, driving habits and characteristics, a driver will generate an individual route choice set according to a single constraint, which is the number of turns that have to be made.

In this research, the number of turns along each route is taken as the only constraint on the generation of the individual choice set. Under this assumption, Equation (8) can be rewritten as

$$q_n(i) = \frac{1}{1 + e^{-\beta \text{turns} - \text{Threshold}}}$$

Table 4.2: Estimation result of PCS model for en-route OD-1 (305 samples) and en-route OD-2 (144 samples)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Choice Making Stage – MNL</th>
<th>Choice Set Generation Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (km)</td>
<td>-26.831</td>
<td>-1.573</td>
</tr>
<tr>
<td>ARR</td>
<td>0.879</td>
<td>-1.774</td>
</tr>
<tr>
<td>β</td>
<td>-1.573</td>
<td>-19.4</td>
</tr>
<tr>
<td>Threshold</td>
<td>-1.774</td>
<td>-7.0</td>
</tr>
</tbody>
</table>
Table 4.2 shows the estimation result obtained with the PCS model. It should be noted that the number of turns 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, so 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 an en-route variable that ranges from 0 to 1 according to how much of the trip has been completed. The influence of DO/TD 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 4.1, driver T’s house is located at point A and his destination is point B, a
shop in the city centre. 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.

![Example route from T’s house to a city center destination](image1)

**Fig. 4.1:** Example route from T’s house to a city center destination

![Relationship between number of turns in target area and trip progress ratio](image2)

**Fig. 4.2:** Relationship between number of turns in target area and trip progress ratio

Figure 4.2 shows the relationship between number of turns in the research area and the vehicles’ DO/TD 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.

We can rewrite $q_n(i)$ as follows after introducing the en-route variable:

$$q_n(i) = \frac{1}{1 + e^{-\left(\beta \text{Turns} + \alpha \frac{DO}{TD} + \gamma \frac{DO}{TD} - \text{Threshold}\right)}}$$

The estimation result for this updated model is shown in Table 4.3. After introducing one variable, the explanatory power of the model is improved. Further, the $\chi^2$ statistic for the likelihood of the joint hypothesis that $\alpha = 0$ and $\gamma = 0$, $-2*(-737.644 + 730.9) = 13.4$, 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 4.2, which explains how the en-route variable 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.

Table 4.3: Estimation result (2) for PCS model with en-route OD-1 (305 samples) and en-route OD-2 (144 samples)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>t-Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choice making stage – MNL</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>-26.564</td>
</tr>
<tr>
<td>Highway ratio</td>
<td>7.32</td>
</tr>
<tr>
<td>Choice set generation stage</td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>-1.396</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>17.066</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-8.723</td>
</tr>
<tr>
<td>Threshold</td>
<td>-3.454</td>
</tr>
</tbody>
</table>
Fig. 4.3: Influence of en-route variable on $q_{n}(i)$
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 driver’s awareness 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.

An en-route variable 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.

This work points to several possible further research directions. One is to consider the heterogeneity of both the choice set generation and choice making stages. A second is to try different models in the choice set generation stage to compare the results.
References