Route choice model for Copenhagen

A data-driven choice set generation approach based on GPS data

Jeppe Rich[1], Stefan L. Mabit, Otto Anker Nielsen

Centre for Traffic and Transport, Technical University of Denmark

Abstract
In order to estimate route choice models based on a random utility maximization approach, data need to be arranged so that utility maximizing agents for each observed route are assigned a relevant choice set. One of the main challenges in this process is the generation of appropriate choice sets. In the literature, several methods have been proposed, including probabilistic methods, constrained enumeration methods, and path search based methods. In this paper we suggest a variant of the constrained enumeration method, which uses real GPS data to build simple connection trees between origin and destinations. The method is based on the assumption that route alternatives can be formed as set of sub-paths, which is defined as unions of sub-paths of GPS-logged routes. In the paper we use the choice set generation approach to estimate a route choice model for Copenhagen with explicit representation of congestion and road charging.

1 INTRODUCTION

The generation of the choice set is a cardinal point when formulating and estimating route choice models based on random utility theory. It is well documented that the size and the composition of the choice set will greatly influence the estimation of route choice model and the subsequent demand prediction (Swait and Ben-Akiva, 1985, 1987). The choice set generation process for route choice models, however, is a complex task compared to the generation for other discrete choice models such as mode choice and destination choice models. Bovy (2006) gives an excellent overview of the special aspects that are related to the choice set for route choice models. One special characteristic is that the choice set may be very large. This requires a filtering or sampling procedure for the model to be tractable and for the model to preserve the discriminatory capacity of choice models (Ruijgrok, 1979). Another issue is that the identification of any route as a genuine alternative is rather complex due to overlapping paths and the space-time domain. Furthermore, routes are affected by the loading pattern (congestion) and the present state of the network (interruptions and road work). This is especially true as more and more cars are

[1]Corresponding author: telephone: +45 45251536, e-mail: jhr@ctt.dtu.dk

equipped with GPS navigators that make use of dynamic traffic information to re-schedule routes as a result of interventions or congestion in the network.

The modeling of the route choice involves a number of complicating factors as outlined by Bovy (2006).

- The size and the composition of the choice set are different. For short routes only a small set of alternatives may be relevant, whereas, for longer routes more alternatives are needed.
- Choice processes may be very different from individual to individual. It may be sequential, simultaneous, or strategic.
- The network used for modeling may be to simplified and outdated compared to the real network. Often demand data such as surveys are several years old, whereas the digital network is upgraded continuously introducing a consistency problem between the two data sources.

1.1 Choice set generation methods

Three classes of choice generation approaches have been proposed in the literature.

- Probabilistic methods
- Constrained enumeration methods
- Path search based methods

All though the class of probabilistic method has been classified as a choice-generation method (Bovy, 2006), it assumes the existence of a master set. As a result it may rather be seen as a tool in which alternatives are ranked according to their relevance (and corrected according to correlation), rather than a true choice set generation procedure. An example is the implicit Availability/Perception (IAP) model proposed by Cascetta and Papola (2001), where a predefined measure of relevance enters the utility function. The path-size logit proposed by Ben-Akiva and Bierlaire (1999) and extended in Ramming (2002)[2] is another example.

Constrained enumeration methods form a master set by constructing connection trees between origin and destinations. The master set may be very large. This requires that the size of the choice set is reduced subsequently by use of filtering or sampling techniques. In principle, all kinds of constraints may be used, including perceptual, cognitive and behavioral preferences.

Path search methods are the most commonly used approaches for route choice generation. They operate by defining utility functions to be used in a shortest-path algorithm. Parameters are then sampled in order to render different shortest-path alternatives, which then enter the choice set. In the sampling, the

[2] This extension has been criticized in Bierlaire and Frejinger (2006).

modeler will try to cover as much of the parameter space as possible to get variability in the choice set. However, the method will predominantly generate the more attractive routes and the choice set will be strongly *selective* in that not all routes will be included (Bovy, 2006).

It is important to realize that a choice set generation usually will use more than one approach. Typically, the first process will be to select a master set, which is then filtered according to various rules. This may be carried out by mean of enumeration methods or shortest-path methods. In a second step, we may make use of probabilistic methods, e.g. by use of a path-size logit.

2 CHOICE SET GENERATION USING GPS DATA

The amount of GPS data available for researchers is growing rapidly. Due to the decrease in the price on GPS technology and the improved quality of the positioning, GPS provide a very cost-effective way of collecting data. The use of advanced map-matching routines, which address the problem of weak coverage in dense city areas, is important in that respect (Zabic and Nielsen, 2006 and Nielsen et al 2007).

Because GPS data include a very detailed description of the space-time domain, typically logged each second, they are a rich source to understand route choice behavior. Moreover, it makes good sense to consider whether this rich pool of data can be used to estimate route choice models, not only by providing the observed route, but also as a source to generate route alternatives. To our knowledge, this approach is widely unexplored in the choice set generation literature, allthough it seems interesting in several respects.

- If the number of routes and individuals is large, the set of sub-paths that might be generated from the observed routes should easily cover all relevant routes.
- On the other hand, since the basis for the route alternatives are only observed routes, irrelevant routes are unlikely to enter the choice set.
- The route choice preferences that are monitored by the GPS reflect elements such as congestion and road-construction work.

On the other hand there are certain issues that make GPS-data less attractive. The programming of an efficient algorithm may seem simple, but is not. The fact that GPS data are likely to be very different in its data representation makes this an even a greater challenge. Another issue is that the outcome of the routine requires that the GPS data are map-matched and divided into unique routes. To the extent this pre-processing will be biased, so will the resulting data be. A solution to this problem has been proposed by Bierlaire and Frejinger (2007), in which route choice models are estimated on the basis of network-free data.

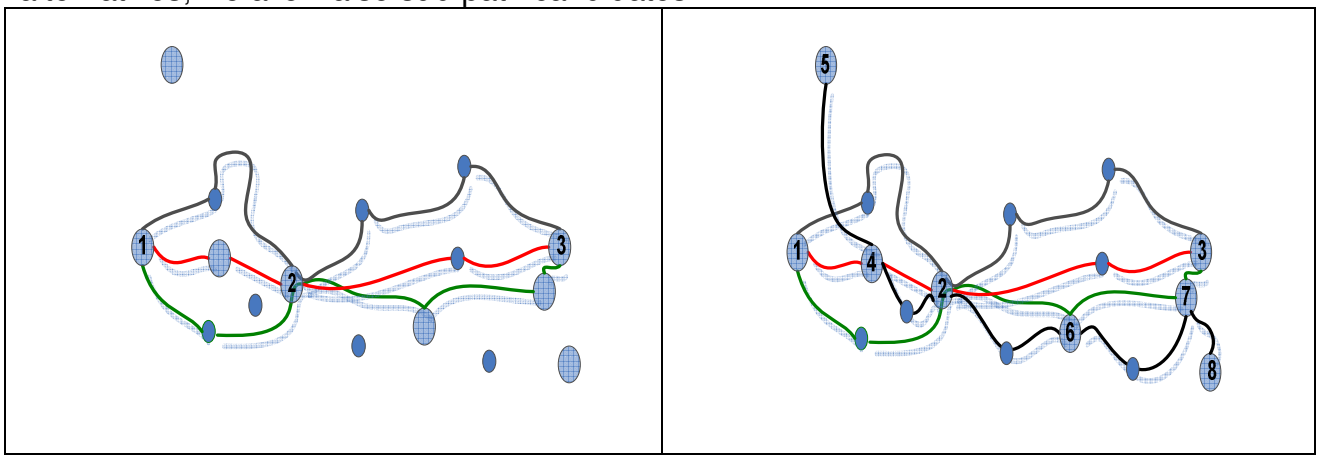
2.1 Background

In this paper we estimate a route choice model based on GPS observations of 500 drivers over several months of driving and under different road charging schemes (the AKTA road pricing experiment, described in Nielsen & Jovicic, 2003). This allows us to estimate, not only the trade-off between travel time and out-of-pocket costs, but more interestingly, between travel time, congestion time and congestion charging. The GPS data include more than 250,000 trips monitored in a road network with more than 400,000 links and 50,000 junctions. The driving behavior has been monitored in various synthetic pricing regimes including kilometer charging and a multi-cordon based charging system (Nielsen, 2004).

2.2 Choice set generation approach

First define $\Omega = \{R_1,\ldots, R_Q\}$ as the GPS route pool, which consist of routes formed as a set of nodes, e.g. $R_q = \{r_1,\ldots, r_n\}$.

Our aim is to generate - artificially – route alternatives between {i,j}. Clearly, the simplest way of generating a choice set would be to select those routes from Ω with common OD combination. However, with more than 50.000 junctions in the network (junctions are the most disaggregate OD representation) there are more than 2.5 billion potential OD pairs. In other words, if we only considered complete routes candidates between unique OD pairs we would get only a few number of competing routes at the OD level and the resulting choice set would be small and with only little variation[3]. To overcome this problem, we suggest using a choice-set generator, which uses a sub-path approach. The idea is that, instead of allowing only complete observed route candidates in the set of choice-set alternatives, we allow also sub-path candidates.



Figur 1: Illustration of sub-paths in a simple network.

[3] This is especially true in the present GPS data because only 500 drivers are monitored, which lead to a sparse coverage in a number of areas.

In figure 1 to the left there are three nodes, linked by 6 sub-paths. If we look at the OD combination {1,3} there are 3 complete connecting routes, which give 3 alternatives. However, if we allow for sub-paths to be joined, 6 additional routes are formed, e.g. 1->2 (red) and 2->3 (green) etc. If in addition we add crossing routes to the network as illustrated in figure 1 to the right, even more sub-paths result. For large networks and GPS pools of data, the number of potential alternatives becomes very large.

For the sake of simplicity, we define that the k'th route alternative between {i,j} given by $R_k(i,j)$ can be represented as only two sub-paths, s_a and s_b.

$R_k(i,j) = s_a \cup s_b$, k=0,…,K(i,j)

Where s_a is a valid sub path with the only requirement that the first node is equal to *i* and the last node of s_b is equal to *j*. Also assume that the observed route is represented by $R_0(i,j)$.

The selection of sub-paths s_a and s_b that make up a choice alternative is the critical point. We have designed the selection procedure in the following way,

1) For a given OD combination {i,j} sample a set of sub-paths S_a and S_b from Ω. Since the number of observed routes is very large, only one sub-path route is sampled for each observed route.
    a. The length of the sub-path - measured in terms of the number of links (e.g. road segments) - is sampled according to draws from a truncated normal distribution. The mean of the normal distribution is set to half of the length of the route from where it is taken and with a 20% variance. This forces variability into the route alternatives in that alternatives which is dominated by either s_a or s_b is minimized. Also, we obtain sub paths that have a higher probability of being matched with a corresponding pair (refer to figure 1). This speeds up the algorithm.

2) Route choice alternatives $R_k(i,j)$ are now formed by joining candidates from S_a and S_b. Only candidates where the last node of s_{ai} equals the first node of s_{bi} are joined (refer to figure 2).

3) The choice-set from 2) can be considered as a master set and may still be very large. As a result we introduce an importance sampling, where the most likely alternatives are sampled. Since we have looped over the entire number of observed routes, there is a tendency that some route structure will be selected many times, whereas others will be selected frequently. The sample probability is determined by the frequency, which has shown to be a good indication of the relevance of the alternative.

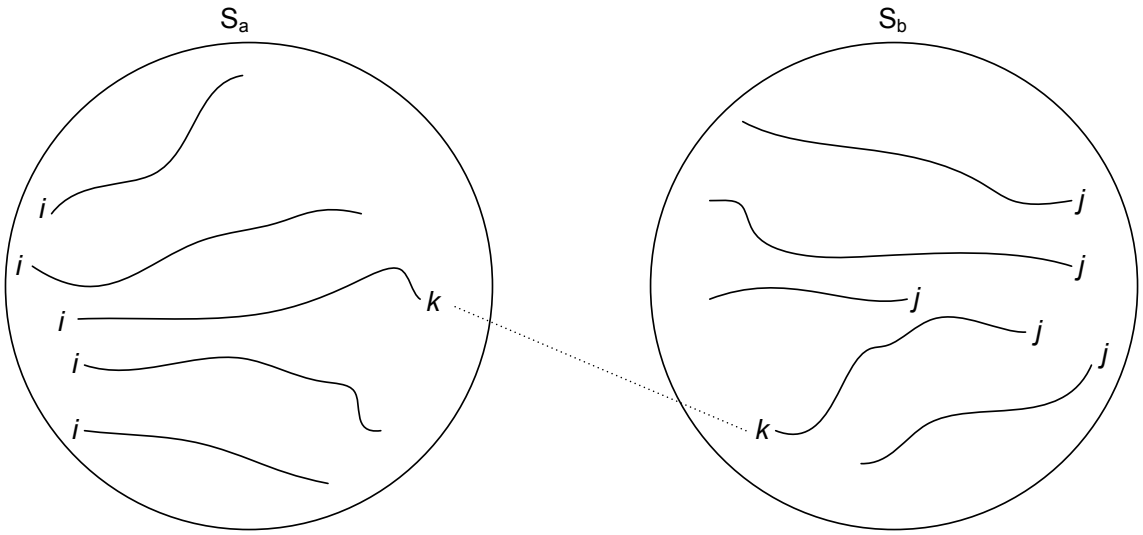


Figure 2: Generation of {*i,j*} route alternatives by joining candidates from S_a and S_b with identical end/start point *k*.

3 MODEL ESTIMATION

Suppose that for observation *n* (may be an individual or the same individual if one person have more than one trip) route *i* is chosen from the choice set D_n, which may have been generated by the approach described in section 2. The probabilistic choice of any given route within D_n can then be described by a random utility maximisation (RUM) model. A general additive RUM model assumes that alternatives $i \in D_n$ have utility functions given by

$$U_{in} = V_{in} + \varepsilon_{in},\ i \in D_n$$

where V_{in} is the deterministic part of utility for alternative *i* and observation *n*.

ε_{in} is the random part of utility.

The most simple route choice model is a multinomial logit (MNL) model with separable cost and time as explanatory variables

$$U_{in} = \beta_c \bullet c_{in} + \beta_t \bullet t_{in} + \varepsilon_{in},$$

where c_{in} is the cost of alternative *i* and observation *n*.

t_{in} is travel time.

ε_{in} is independent identical Gumbel distributed.

The model implicitly assumes independence across the choice set due to the specification of the error term. However, for route choice models, alternatives are likely to be correlated and in the literature, several approaches for dealing with

correlation between routes have been suggested. These include the commonality factor, path size logit and cross nested logit specifications or error components models (for references see, e.g. Cascetta, 2001, or Frejinger and Bierlaire, 2006a). A second way to introduce correlation is through the use of mixed logit models. An example is presented in the next model. Here we distinguish between error components and mixed logit since the mixed logit is connected with the individual making the choice and the error components are based on the network.

A model that has been estimated on the AKTA data using a different choice set generation approach has the following specification (Nielsen, 2004)

$$U_{in} = \beta_l \bullet l_{in} + \beta_c \bullet \beta_l \bullet c_{in} + \beta_{free} \bullet (t_{free,\,in} + \beta_{con} \bullet t_{con,\,in}) + \varepsilon_{in}, \qquad (1)$$

where l_{in} is the lengt of alternative i and observation n.

c_{in} is the road pricing cost.

$t_{free,in}$ is free flow time.

$t_{con,in}$ is congested time.

ε_{in} is independent identical Gumbel distributed

For fixed parameters β_l, β_c, β_{free}, and β_{con} this is a standard MNL model. However, when parameters are allowed to follow different distributions over the population a mixed logit models emerge. In the model β_{free} and β_{con} were assumed to follow lognormal distributions.

In our estimations we will assume an MNL model given by the specification in (1) as our base model. As a second model we will estimate a model corresponding to the earlier application with two lognormal coefficients. The coefficients are allowed to be individual specific while varying in the population, i.e. we use a panel formulation. A third model will allow all four coefficients to follow lognormal distributions.

The mixing of parameters is likely to reduce the problem with correlated alternatives to the extent that correlation can be assumed to be individual specific. However, the correlation is network independent and due to the way data has been generated, it is most likely that partly overlapping routes will share common unobservable characteristics, which will impose a correlation structure among route alternatives.

We will examine how the effect on path size logit can be incorporated into our mixed logit approach.

A further important issue is created by the way we generate our data. Since we join sub paths together it becomes central whether the connection point is a right

turn, left turn or no turn. This might affect the choice behavior and we will therefore allow the number of turns to enter the specification in (1).

4 SUMMARY

The paper examines the possibility of using GPS data as a tool to generate route choice alternatives for the estimation of route choice parameters in random utility models. The idea is to use the GPS data, not only to provide information on the observed route, but also to provide information on alternative routes. To do this, we suggest that sub-paths may be sampled and then joined to form new alternative routes.

However, the paper is at an experimental stage and no estimation has been finalized yet.

References

Ben-Akiva, M.E. and Bierlaire, M. (1999). Discrete choice methods and their application to short-term travel decisions. In: Hall, R. (ed.), Handbook of Transportation Science. Kluwer, pp. 5-34.

Bovy, P.H.L. (2006). Modelling Route Choice Sets in Transportation Networks: Principles and Empirical Validity. Paper for ISTTT 17.

Cascetta, E. (2001). Transportation Systems Engineering: Theory and Methods, Kluwer Academic Publishers

Cascetta, E. and Papola, A. (2001). A model of route perception in urban road networks, *Transp. Res. B* 36, pp. 577-592.

Frejinger, E. and Bierlaire, M.(2006a). Capturing Correlation with Subnetworks in Route Choice Models, forthcoming in Transportation Research part B

Frejinger, E. and Bierlaire, M.(2006b). A latent route choice model in Switzerland, presented at ETC 2006

Nielsen, O.A., Daly, A. and Frederiksen, R.D. (2002). A Stochastic Route Choice Model for Car Travellers in the Copenhagen Region, Network and Spatial Economics 2, 327-346.

Nielsen, O.A. and Jovicic, G. (2003). The AKTA road pricing experiment in Copenhagen. 10th International Conference on Travel Behaviour Research. Proceedings, session 3.2 Valuation/Pricing. Lucerne, Switzerland, August.

Nielsen, O.A. (2004). Behavioural responses to pricing schemes: Description of the Danish AKTA experiment. Journal of Intelligent Transportation Systems. Vol. 8(4). Pp. 233-251. Taylor & Francis

Nielsen, Otto Anker; Würtz Christian & Jørgensen, René Munk (2007). Improved map-matching algorithms for gps-data - - Methodology and test on data from The AKTA roadpricing experiment in Copenhagen. European ITS Conference**.** Paper 2626. Aalborg, Denmark, 20/2

Ramming, M.S. (2002). Network knowledge and route choice, PhD thesis, MIT, Cambridge.

Ruijgrok, C.J. (1979). Disaggregate choice models: an evaluation. In G.R.M Jansen, P.H.L. Bovy, J.P.J.M. van Est and F. Le Clercq (eds), *New Developments in Modelling Travel Demand and Urban Systems.* Saxon House, Westmead.

Swait, J. and Ben-Akiva, M.E. (1985). Analysis of the effects of captivity on travel time and cost elasticities. In: *Behavioural Research for Transport Policy*, Proc. 1985, pp. 119-134.

Swait, J. and Ben-Akiva, M.E. (1985). Incorporating random constraints in discrete models of choice set generation. *Transp. Res. B* 21, 2, pp. 91-102.

Zabic, M. and Nielsen. O. A. (2006) 'A Geographic Information System Analysis of Global Positioning Quality for Road Pricing',In Brebbia and Dolezel (eds), *Urban Transport XII.*, WIT Press, ISBN: 1-84564-179-5, pp. 859-868.