Sampling of Alternatives for Route Choice Modeling*

E. Frejinger † M. Bierlaire †

November 21, 2007

Report TRANSP-OR 071121
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*This research is supported by the Swiss National Science Foundation grant 200021-107777/1

†École Polytechnique Fédérale de Lausanne, Transport and Mobility Laboratory, Station 18, CH-1015 Lausanne, Switzerland. E-mail: {emma.frejinger, michel.bierlaire}@epfl.ch

Abstract

This paper presents a new paradigm for choice set generation in the context of route choice. We assume that the choice sets contain all paths connecting each origin-destination pair. These sets are in general impossible to generate explicitly. Therefore, we propose an importance sampling approach to generate subsets of paths suitable for model estimation. Using only a subset of alternatives requires the path utilities to be corrected according to the sampling protocol in order to obtain unbiased parameter estimates. We derive such a sampling correction for the proposed algorithm.

Estimating models based on samples of alternatives is straightforward for some types of models, in particular the Multinomial Logit (MNL) model. In order to apply MNL for route choice, the utilities must also be corrected to account for the correlation using, for instance, a Path Size (PS) formulation. We show that the PS should be computed based on the full choice set. Again, this is not feasible in general, and we propose an operational solution, called the Extended PS.

We present numerical results based on synthetic data. The results show that models including a sampling correction are remarkably better than the ones that do not. Moreover, the Extended PS appears to be a good approximation of the true one.

1 Introduction

Route choice models play an important role in many transport applications. The modeling is complex for various reasons and involves several steps before the actual route choice model estimation. We start by giving an overview of the modeling process in Figure 1. In a real network a very large set of paths (actually infinitely many if the network contains loops) connect an origin s_o and a destination s_d. This set, referred to as the universal choice set $\mathcal{U}$, cannot be explicitly generated. In order to estimate a route choice model, a subset of paths needs to be defined and path generation algorithms are used for this purpose. There exist deterministic and stochastic approaches for generating paths.

Deterministic methods always generate the same set $\mathcal{M}$ of paths for a given origin-destination pair. Most of them are based on some form of repeated shortest path search. This type of approach is computation-

ally appealing thanks to the efficiency of shortest path algorithms. Examples are link elimination (Azevedo et al., 1993), link penalty (de la Barra et al., 1993) and labeled paths (Ben-Akiva et al., 1984). Instead of performing repeated shortest path searches, a constrained enumeration approach referred to as branch-and-bound has recently been proposed. Friedrich et al. (2001) present an algorithm for public transport networks, Hoogendoorn-Lanser (2005) for multi-modal networks and Prato and Bekhor (2006) for route networks.

Stochastic methods generate an individual (or observation) specific subset $\mathcal{M}_n$. Actually, most of the deterministic approaches can be made stochastic by using random generalized cost for the shortest path computations. Ramming (2001) proposes a simulation method that produces alternative paths by drawing link costs from different probability distributions. The shortest path according to the randomly distributed generalized cost is calculated and introduced in the choice set. Recently, Bovy and Fiorenzo-Catalano (2006) proposed the doubly stochastic choice set generation approach. It is similar to the simulation method but the generalized cost functions are specified like utilities and both the parameters and the attributes are stochastic. They also propose to use a filtering process such that, among the generated paths, only those satisfying some constraints are kept in the choice set.

Once $\mathcal{M}$ (or $\mathcal{M}_n$) has been generated, a choice set $\mathcal{C}_n$ for individual n can be defined in either a deterministic way by including all feasible paths, $\mathcal{C}_n = \mathcal{M}$ (or $\mathcal{C}_n = \mathcal{M}_n$), or by using a probabilistic model $P(\mathcal{C}_n)$ where all non-empty subsets $\mathcal{G}_n$ of $\mathcal{M}$ (or $\mathcal{M}_n$) are considered. Defining choice sets in a probabilistic way is complex due to the size of $\mathcal{G}_n$ and has never been used in a real size application. See Manski (1977), Swait and Ben-Akiva (1987), Ben-Akiva and Boccara (1995) and Morikawa (1996) for more details on probabilistic choice set models. Cascetta and Papola (2001) (Cascetta et al., 2002) propose to simplify the complex probabilistic choice set models by viewing the choice set as a fuzzy set in a implicit availability/perception of alternatives model.

The formal evaluation of the relevance and realism of generated choice sets is difficult in practice since the actual choice sets in general are unknown to the modeler. Several researchers, including Ramming (2001), Hoogendoorn-Lanser (2005), Bekhor et al. (2006), Bovy and Fiorenzo-Catalano (2006), Prato and Bekhor (2006), Bekhor and Prato (2006), van

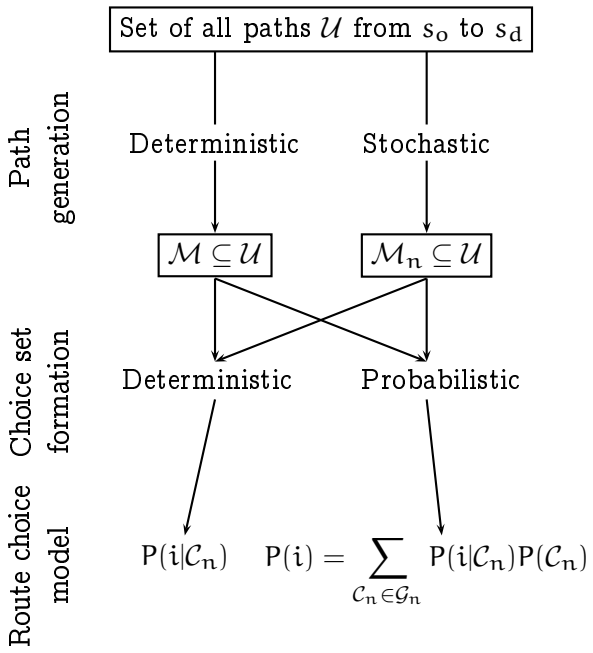

Figure 1: Choice Set Generation Overview

Nes et al. (2006), Bovy (2007) and Fiorenzo-Catalano (2007), have proposed various measures of quality of the generated sets. Empirical analysis show that no choice set generation algorithm is able to fully reproduce observed paths.

In the context of our new paradigm based on sampling from the universal choice set, these measures do not apply, as all possible paths belong to the choice set. Moreover, the observed path is always in the sample by design. The validation of our approach is based on the verification that unbiased estimates of the parameters are obtained.

In the following section we give an introduction to sampling of alternatives. We describe the proposed algorithm in Section 3 and we continue by deriving the sampling correction in Section 4. In Section 5 we present numerical results based on synthetic data and describe the heuristic for computing the Extended Path Size attribute. Finally we present conclusions and issues for future research.

2 Sampling of Alternatives

The Multinomial Logit model can be consistently estimated on a subset of alternatives (McFadden, 1978) using classical conditional maximum likelihood estimation. The probability that an individual n chooses an alternative i is then conditional on the choice set $\mathcal{C}_n$ *defined by the modeler*. This conditional probability is

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in} + \ln q(\mathcal{C}_n|i)}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn} + \ln q(\mathcal{C}_n|j)}} \tag{1}$$

and includes an alternative specific term, $\ln q(\mathcal{C}_n|j)$, correcting for sampling bias. This correction term is based on the probability of sampling $\mathcal{C}_n$ given that j is the chosen alternative, $q(\mathcal{C}_n|j)$. See for example Ben-Akiva and Lerman (1985) for a more detailed discussion on sampling of alternatives. Bierlaire et al. (to appear) have recently shown that Multivariate Extreme Value (also called Generalized Extreme Value) models can also be consistently estimated and propose a new estimator.

Importance sampling of alternatives has been used in the literature. For example, Ben-Akiva and Watanatada (1981) use samples of destinations for prediction and Train et al. (1987) sample alternatives for the estimation of local telephone service choice models. A sampling of alternatives approach has however never been used for route choice modeling, to the best of our knowledge.

If all alternatives have equal selection probabilities, the estimation on the subset is done in the same way as the estimation on the full set of alternatives. Indeed, $q(\mathcal{C}_n|i)$ is equal to $q(\mathcal{C}_n|j)\ \forall\ j \in \mathcal{C}_n$ and the corrections for sampling bias cancel out in Equation (1). A simple random sampling protocol is however not efficient if the full set of alternatives is very large. The sample should include attractive alternatives since comparing a chosen alternative to a set of highly unattractive alternatives would not provide much information on the choice. In order to ensure that attractive alternatives are included, the sample would need to be prohibitively large.

When using a sampling protocol selecting attractive alternatives with higher probability than unattractive alternatives (importance sampling), the correction terms in Equation (1) do not cancel out. Note that if alternative specific constants are estimated, all parameter estimates except

the constants would be unbiased even if the correction is not included in the utilities (Manski and Lerman, 1977). In a route choice context it is in general not possible to estimate alternative specific constants due to the large number of alternatives and the correction for sampling is therefore essential. Therefore, the key element of our approach consists in designing a stochastic path generation algorithm such that the probability $q(\mathcal{C}_n|i)$ can easily be derived. We propose a simple example in the next section.

3 A Stochastic Path Generation Approach

This stochastic path generation approach is flexible and can be used in various algorithms including those presented in the literature. We start by describing the general approach and then focus on a specific instance based on a biased random walk.

For a given origin-destination pair (s_o, s_d), the general approach associates a weight with each link $\ell = (v, w)$ based on its distance to the shortest path according to a given generalized cost. More precisely, the weight $\omega(\ell|a, b)$ is defined by the double bounded Kumaraswamy distribution (proposed by Kumaraswamy, 1980), that is

$$\omega(\ell|a, b) = 1 - (1 - x_\ell{}^a)^b. \tag{2}$$

a and b are shape parameters and $x_\ell \in [0, 1]$ represents a measure of distance to the shortest path and is defined as

$$x_\ell = \frac{SP(s_o, s_d)}{SP(s_o, v) + C(\ell) + SP(w, s_d)}, \tag{3}$$

where $C(\ell)$ is the generalized cost of link ℓ, and $SP(v_1, v_2)$ is the generalized cost of the shortest path between nodes v_1 and v_2. Note that x_ℓ equals one if ℓ is part of the shortest path and $x_\ell \to 0$ as $C(\ell) \to \infty$. In Figure 2 we show the cumulative distribution function for different values of a when $b = 1$. The weights assigned to the links can be controlled by the definition of the distribution parameters. High values of a when $b = 1$ yield low weights for links with high cost. Low values of a have the opposite effect.

Note that other distributions with suitable properties can be used. It is also worth mentioning that this idea presents similarities in its nature with the approach proposed by Dial (1971).

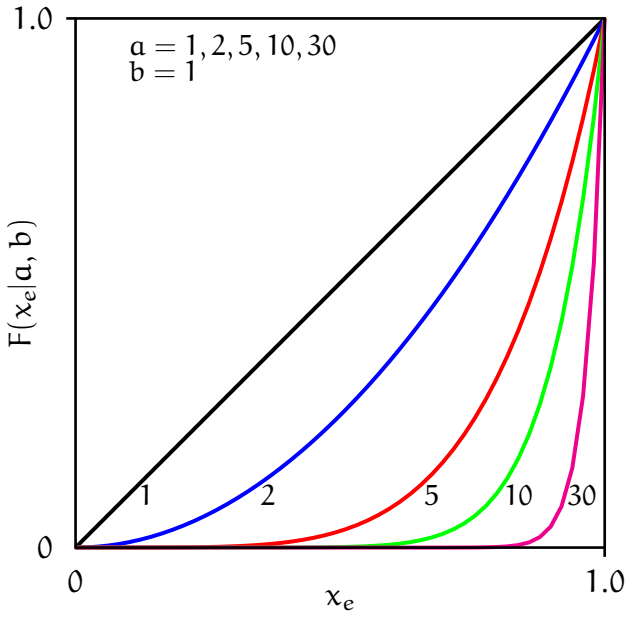

Figure 2: Kumaraswamy Distribution: Cumulative Distribution Function

Once a weight has been assigned to each link, various methods can be applied. Bierlaire and Frejinger (2007b) propose a gateway approach, used by Bierlaire and Frejinger (2007a) for modeling long distance route choice behavior in Switzerland. Note also that the method can be generalized to subpaths instead of links, in order to better reflect behavioral perceptions (see Frejinger and Bierlaire, 2007 and Frejinger, 2008).

In this paper, we use a biased random walk algorithm which is appropriate for the importance sampling approach. First, it generates any path in $\mathcal{U}$ with non-zero probability. Second, path selection probabilities can be computed in a straightforward way.

Given an origin s_o and a destination s_d, an ordered set of links Γ is generated as follows:

Initialize $v = s_o$, $\Gamma = \emptyset$

Loop While $v \neq s_d$ perform the following

Weights For each link $\ell = (v, w) \in \mathcal{E}_v$, where $\mathcal{E}_v$ is the set of outgoing links from v, we compute the weights based on (2) where x_ℓ is defined by

$$x_\ell = \frac{SP(v, s_d)}{C(\ell) + SP(w, s_d)}. \tag{4}$$

Note that this is equivalent to (3) where $s_o = v$.

Probability For each link $\ell = (v, w) \in \mathcal{E}_v$, we compute

$$q(\ell|\mathcal{E}_v, a, b) = \frac{\omega(\ell|a, b)}{\sum_{m \in \mathcal{E}_v} \omega(m|a, b)} \tag{5}$$

Draw Randomly select a link (v, w^*) in $\mathcal{E}_v$ based on the above probability distribution.

Update path $\Gamma = \Gamma \cup (v, w^*)$

Next node $v = w^*$.

The algorithm biases the random walk towards the shortest path in a way controlled by the parameters of the distribution. The algorithm corresponds to a simple random walk if a uniform distribution (special case of Kumaraswamy distribution with $a = 0$ and $b = 1$) is used. Note however that a simple random walk does not generate paths with equal probability.

The probability $q(j)$ of generating a path j is the probability of selecting the ordered sequence of links Γ_j

$$q(j) = \prod_{\ell \in \Gamma_j} q(\ell|\mathcal{E}_v, a, b), \tag{6}$$

where $q(\ell|\mathcal{E}_v, a, b)$ is defined by (5).

With this algorithm, it is easy to compute path selection probabilities and it is not computationally demanding since at most $|\mathcal{V}|^2$ shortest path computations are needed for any number of observations, where $\mathcal{V}$ is the number of nodes in the network.

Note that existing stochastic path generation approaches may also be viewed as importance sampling approaches. We are however unaware of how to compute the sampling correction in a straightforward way for these algorithms.

4 Sampling Correction

As discussed in Section 2, the correction terms $q(\mathcal{C}_n|j) \ \forall \ j \in \mathcal{C}_n$ must be defined for this type of sampling protocol in order to obtain unbiased parameter estimates.

We define a sampling protocol for path generation as follows: a set $\widetilde{\mathcal{C}}_n$ is generated by drawing R paths with replacement from the universal set of paths $\mathcal{U}$ using the biased random walk method described before, and then adding the chosen path to it ($|\widetilde{\mathcal{C}}_n| = R + 1$). We assume without loss of generality that $\mathcal{U}$ is bounded with size J. Note that J is unknown in practice. Each path $j \in \mathcal{U}$ has sampling probability $q(j)$ defined by (6).

The outcome of this protocol is $(\widetilde{k}_{1n}, \widetilde{k}_{2n}, \ldots, \widetilde{k}_{Jn})$ where $\widetilde{k}_{jn}$ is the number of times alternative j is drawn ($\sum_{j\in\mathcal{U}} \widetilde{k}_{jn} = R$). Following Ben-Akiva (1993) we derive $q(\mathcal{C}_n|j)$ for this sampling protocol. The probability of an outcome is given by the multinomial distribution

$$P(\widetilde{k}_{1n}, \widetilde{k}_{2n}, \ldots, \widetilde{k}_{Jn}) = \frac{R!}{\prod_{j\in\mathcal{U}} \widetilde{k}_{jn}!} \prod_{j\in\mathcal{U}} q(j)^{\widetilde{k}_{jn}}. \tag{7}$$

The number of times alternative j appears in $\widetilde{\mathcal{C}}_n$ is $k_{jn} = \widetilde{k}_{jn} + \delta_{jc}$, where c denotes the index of the chosen alternative and δ_{jc} equals one if $j = c$ and zero otherwise. Let $\mathcal{C}_n$ be the set containing all alternatives corresponding to the R draws ($\mathcal{C}_n = \{j \in \mathcal{U} \mid k_{jn} > 0\}$). The size of $\mathcal{C}_n$ ranges from one to $R + 1$; $|\mathcal{C}_n| = 1$ if only duplicates of the chosen alternative were drawn and $|\mathcal{C}_n| = R + 1$ if the chosen alternative is not drawn nor were any duplicates.

The probability of drawing $\mathcal{C}_n$ given the chosen alternative i (randomly drawn $k_{in} - 1$ times) can be defined using Equation (7) as

$$q(\mathcal{C}_n|i) = q(\widetilde{\mathcal{C}}_n|i) = \frac{R!}{(k_{in}-1)!\prod_{\substack{j\in\mathcal{C}_n \\ j\neq i}} k_{jn}!} q(i)^{k_{in}-1} \prod_{\substack{j\in\mathcal{C}_n \\ j\neq i}} q(j)^{k_{jn}} \tag{8}$$

where the products now are over all elements in $\mathcal{C}_n$ since the terms for alternatives that are not drawn ($k_{jn} = 0$) equal one. Equation (8) can be reformulated as

$$q(\mathcal{C}_n|i) = \frac{R!}{\frac{1}{k_{in}}\prod_{j\in\mathcal{C}_n} k_{jn}!} \frac{1}{q(i)} \prod_{j\in\mathcal{C}_n} q(j)^{k_{jn}} = K_{\mathcal{C}_n} \frac{k_{in}}{q(i)} \tag{9}$$

where

$$K_{\mathcal{C}_n} = \frac{R!}{\prod_{j\in\mathcal{C}_n} k_{jn}!} \prod_{j\in\mathcal{C}_n} q(j)^{k_{jn}}.$$

Note that the positive conditioning property is trivially verified, that is

$$q(\mathcal{C}_n|i) > 0 \implies q(\mathcal{C}_n|j) > 0 \; \forall j \in \mathcal{C}_n.$$

We can now define the probability (1) that an individual chooses alternative i in $\mathcal{C}_n$ as

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in}+\ln\left(\frac{k_{in}}{q(i)}\right)}}{\sum_{j\in\mathcal{C}_n} e^{V_{jn}+\ln\left(\frac{k_{jn}}{q(j)}\right)}}, \tag{10}$$

where $K_{\mathcal{C}_n}$ in Equation (9) cancels out since it is constant for all alternatives in $\mathcal{C}_n$. When using the previously presented biased random walk algorithm we consequently only need to count the number of times a given path j is generated as well as its sampling probability given by Equation (6) which are both straightforward to compute.

5 Numerical Results

The numerical results presented in this section aim at evaluating the impact on estimation results of

- the sampling correction,
- the definition of the Path Size (PS) attribute and
- the biased random walk algorithm parameters.

Synthetic data are used for which the true model structure and parameter values are known. Based on this data we then evaluate different model specifications with the t-test values of the parameter estimates with respect to (w.r.t.) their corresponding true values. In the following we refer to a parameter estimate as biased if it is significantly different from its true value at 5% significance level (critical value: 1.96).

5.1 Synthetic Data

The network is shown in Figure 3 and is composed of 38 nodes and 64 links. It is a network without loops and the universal choice set $\mathcal{U}$ can therefore be enumerated ($|\mathcal{U}| = 170$). The length of the links is proportional to the length in the figure and some links have a speed bump (SB).

Observations are generated with a postulated model. In this case we use a Path Size Logit (PSL) model (Ben-Akiva and Ramming, 1998 and

Ben-Akiva and Bierlaire, 1999), and we specify a utility function for each alternative i and observation n:

$$U_{in} = \beta_{PS} \ln PS_i^{\mathcal{U}} + \beta_L \text{Length}_i + \beta_{SB} \text{NbSB}_i + \varepsilon_{in}, \tag{11}$$

where $\beta_{PS} = 1$, $\beta_L = -0.3$, $\beta_{SB} = -0.1$ and ε_{in} are i.i.d. Extreme Value with scale 1 and location 0. The PS attribute is defined by

$$PS_i^{\mathcal{U}} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in \mathcal{U}} \delta_{aj}} \tag{12}$$

where Γ_i is the set of links in path i, L_a is the length of link a, L_i the length of path i and δ_{aj} equals one if path j contains link a, zero otherwise. Note that we explicitly index $\mathcal{U}$ to emphasize on which path set it is computed. 3000 synthetic observations have been generated by simulation, associating a choice with the alternative having the highest utility.

5.2 Model Specifications

		Sampling Correction	
		Without	With
Path Size	$\mathcal{C}$	$M_{PS(\mathcal{C})}^{\text{NoCorr}}$	$M_{PS(\mathcal{C})}^{\text{Corr}}$
	$\mathcal{U}$	$M_{PS(\mathcal{U})}^{\text{NoCorr}}$	$M_{PS(\mathcal{U})}^{\text{Corr}}$

Table 1: Model Specifications

Table 1 present the four different model specifications that are used in order to evaluate both the PS attribute and the sampling correction. For each of these models we specify the deterministic term of the utility function as follows

$$\begin{aligned}
M_{PS(\mathcal{C})}^{\text{NoCorr}} \quad V_{in} &= \mu \left(\beta_{PS} \ln PS_{in}^{\mathcal{C}} - 0.3 \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) \\
M_{PS(\mathcal{C})}^{\text{Corr}} \quad V_{in} &= \mu \left(\beta_{PS} \ln PS_{in}^{\mathcal{C}} - 0.3 \text{Length}_i + \beta_{SB} \text{NbSB}_i + \ln(\frac{k_{in}}{q(i)}) \right) \\
M_{PS(\mathcal{U})}^{\text{NoCorr}} \quad V_i &= \mu \left(\beta_{PS} \ln PS_i^{\mathcal{U}} - 0.3 \text{Length}_i + \beta_{SB} \text{NbSB}_i \right) \\
M_{PS(\mathcal{U})}^{\text{Corr}} \quad V_{in} &= \mu \left(\beta_{PS} \ln PS_i^{\mathcal{U}} - 0.3 \text{Length}_i + \beta_{SB} \text{NbSB}_i + \ln(\frac{k_{in}}{q(i)}) \right).
\end{aligned}$$

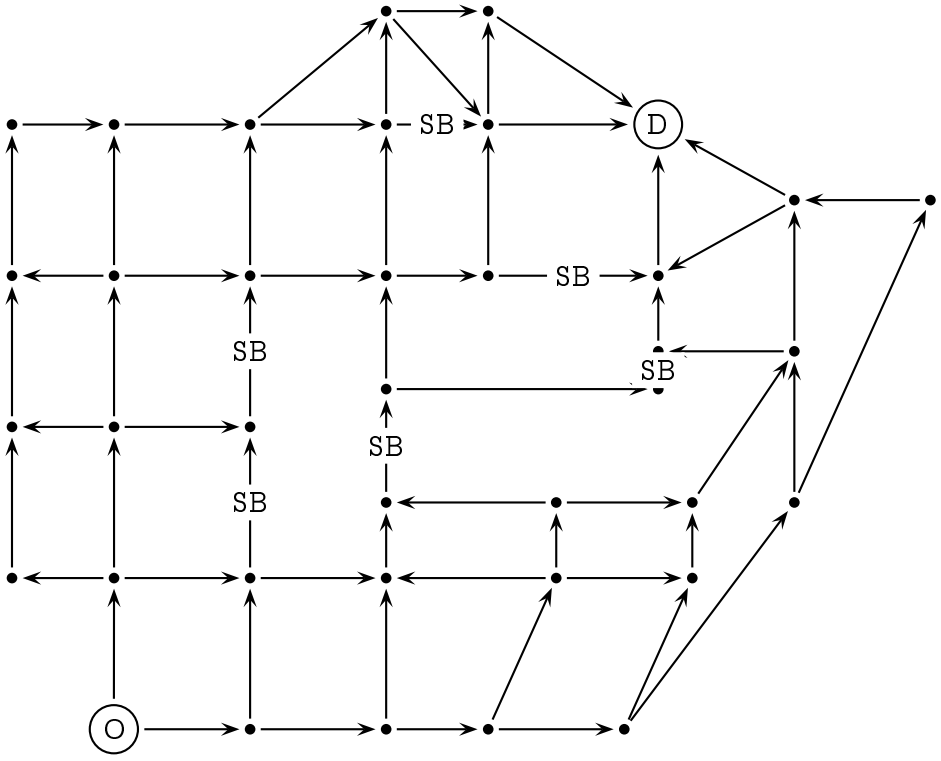

Figure 3: Example Network

The PS attribute based on sampled paths is defined by

$$PS_{in}^{\mathcal{C}} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in \mathcal{C}_n} \delta_{aj}}. \tag{13}$$

Note that the two first specifications are based on (13) and the two last on (12). β_L is fixed to its true value and we estimate μ, β_{PS} and β_{SB}. In this way the scale of the parameters is the same for all models and we can compute the t-tests w.r.t. the corresponding true values.

5.3 Estimation Results

For a specific parameter setting of the biased random walk algorithm (10 draws, Kumaraswamy parameters $a = 5$ and $b = 1$, length is used as generalized cost for the shortest path computations), we generate one choice set per observation and estimate the models. The corresponding estimation results are reported in Table 2. The t-test values show that only the model including a sampling correction and PS computed based on $\mathcal{U}$ ($M^{Corr}_{PS(\mathcal{U})}$) has unbiased parameter estimates.

The models including sampling correction have smaller variance of the random terms compared to the models without correction. (Recall that μ^2 is inversely proportional to the variance.) The standard errors of the parameter estimates are also in general smaller indicating more efficient estimates. Moreover, the model fit is remarkably better for the models with correction compared to those without. Despite of this the model with PS computed based on sampled choice sets ($M^{Corr}_{PS(\mathcal{C})}$) has biased parameter estimates. Hence, these results support the hypothesis that the PS should be computed based on the true correlation structure, otherwise the attribute biases the results. In a real application it is however not possible to compute PS based on the true correlation structure since $\mathcal{U}$ cannot be explicitly generated. This is further discussed in the following section.

We now analyze the estimation results as a function of two of the biased random walk algorithm parameters: the Kumaraswamy distribution parameter a and the number of draws. First we note from Figure 4 that, as expected, the number of generated paths increase with the number of draws but decrease as a increase. Recall from Figure 2 that the higher the value of a the more the biased random walk is oriented towards the

	True PSL	$M^{NoCorr}_{PS(\mathcal{C})}$ PSL	$M^{Corr}_{PS(\mathcal{C})}$ PSL	$M^{NoCorr}_{PS(\mathcal{U})}$ PSL	$M^{Corr}_{PS(\mathcal{U})}$ PSL
$\beta_{\mathbf{L}}$ fixed	**-0.3**	**-0.3**	**-0.3**	**-0.3**	**-0.3**
$\widehat{\mu}$	**1**	**0.182**	**0.724**	**0.141**	**0.994**
standard error		0.0277	0.0226	0.0263	0.0286
t-test w.r.t. 1		-29.54	-12.21	-32.64	-0.2
$\widehat{\beta}_{\mathbf{PS}}$	**1**	**1.94**	**0.411**	**-1.02**	**1.04**
standard error		0.428	0.104	0.383	0.0474
t-test w.r.t. 1		2.20	-5.66	-5.27	0.84
$\widehat{\beta}_{\mathbf{SB}}$	**-0.1**	**-1.91**	**-0.226**	**-2.82**	**-0.0867**
standard error		0.25	0.0355	0.428	0.0238
t-test w.r.t. -0.1		-7.24	-3.55	-6.36	0.56
Final log likelihood		-6660.45	-6082.53	-6666.82	-5933.98
Adj. rho-square		0.018	0.103	0.017	0.125
Null log likelihood: -6784.96, 3000 observations					
Algorithm parameters: 10 draws, $a = 5$, $b = 1$, $C(\ell) = L_\ell$					
Average size of sampled choice sets: 9.66					
BIOGEME (Bierlaire, 2007, and Bierlaire, 2003) has been used for all model estimations					

Table 2: Path Size Logit Estimation Results

shortest path. Figure 5 shows the absolute value of the t-tests w.r.t. the true values for the $M^{Corr}_{PS(\mathcal{U})}$ model. With few exceptions the parameters are unbiased for both 10 and 40 draws and for all values of α. (A line is shown at the critical value 1.96.) These results indicate that for this example the estimation results are robust w.r.t. to the algorithm parameter settings.

The other three model specifications ($M^{NoCorr}_{PS(\mathcal{C})}$, $M^{Corr}_{PS(\mathcal{C})}$ and $M^{NoCorr}_{PS(\mathcal{U})}$) have biased estimates for at least one parameter for all values of α and for all number of draws. The detailed results are presented in the Appendix.

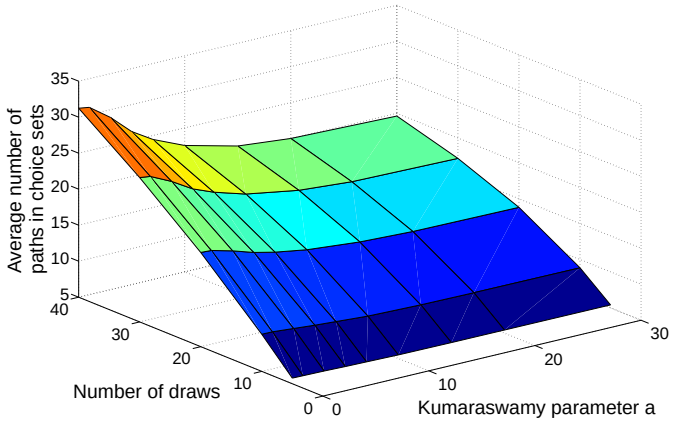

Figure 4: Average Number of Paths in Choice Sets

5.4 Heuristic for Extended Path Size

In a real application where $\mathcal{U}$ cannot be generated it is not possible to compute the PS attribute on the true correlation structure. It is important, though, to compute it based on a set of paths larger than the sampled set $\mathcal{C}_n$. It is therefore interesting to first study, for the previous example, how many paths are needed in order to obtain unbiased parameter estimates. Second, we propose a heuristic for computing a PS attribute that approximates the true correlation structure.

We generate an *extended choice set* $\mathcal{C}_n^{extended}$ for each observation in the network shown in Figure 3. This choice set is only used for computing the PS attribute. In addition to all paths in $\mathcal{C}_n$ we randomly draw (uniform distribution) a number of paths from $\mathcal{U} \backslash \mathcal{C}_n$ and add these to $\mathcal{C}_n^{extended}$. The

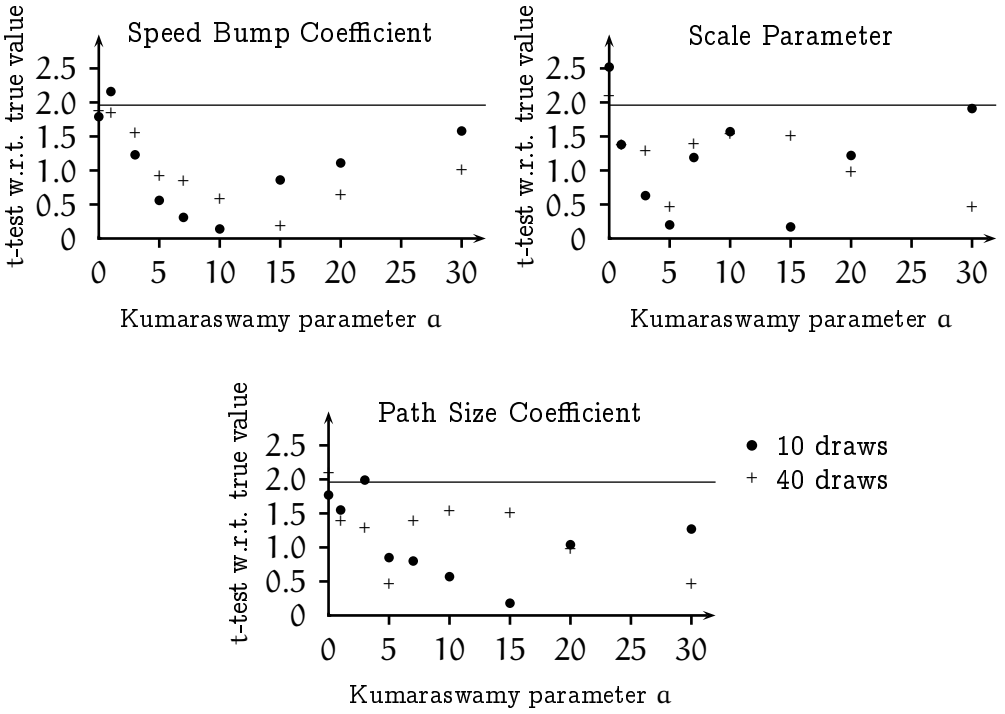

Figure 5: t-test Values w.r.t. True Values for the Coefficients of $M^{\text{Corr}}_{\text{PS}(\mathcal{U})}$

deterministic utilities for a model including sampling correction are now defined as

$$V_{in} = \mu \left(\beta_{PS} \ln PS_{in}^{\mathcal{C}^{extended}} - 0.3 Length_i + \beta_{SB} NbSB_i + \ln(\frac{k_{in}}{q(i)}) \right) \quad (14)$$

where

$$PS_{in}^{\mathcal{C}^{extended}} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in \mathcal{C}_n^{extended}} \delta_{aj}}.$$

The estimation results as a function of the average size of $\mathcal{C}_n^{extended}$ are shown in Figure 6 where x-axis ranges from the average number of paths in $\mathcal{C}_n$ (9.66) up to $|\mathcal{U}| = 170$. For each parameter estimate we report the absolute value of the t-test w.r.t. its true value. An important improvement of the t-test values can be noted after only 20 additional paths in $\mathcal{C}_n^{extended}$ where both the speed bump and PS coefficients are unbiased. The scale parameter is unbiased from 80 additional paths. Even though many paths (average number in $\mathcal{C}_n^{extended}$ approximately $0.5|\mathcal{U}|$) are needed in order for all parameter estimates to be unbiased, we can improve significantly the estimates by using an extended choice set for the PS computation.

Note that the purpose of the results presented in Figure 6 is to have an indication of the parameter estimates when the PS attribute is computed on more paths than those in $\mathcal{C}_n$. Each data point correspond to one random sample of paths. More samples would be needed in order to perform a deeper analysis, but this is already a clear indication on the need for using larger sets for computing the PS attribute.

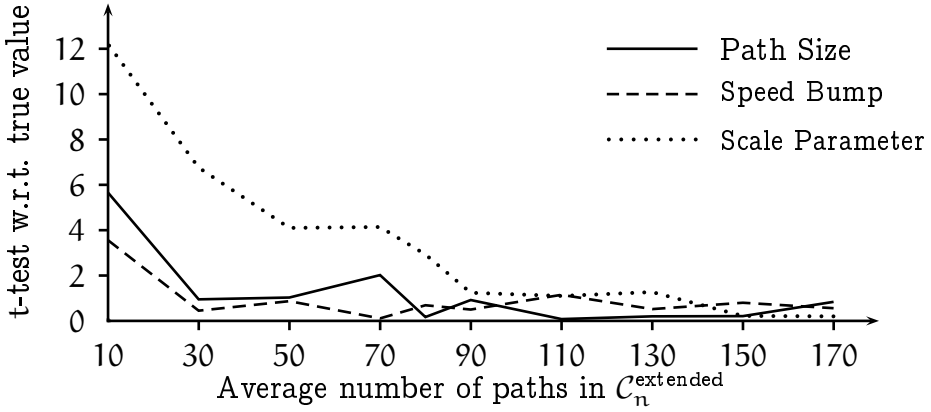

Figure 6: Estimation Results for Corrected model as a Function of the $\mathcal{C}_n^{extended}$ Average Size

In order to use an extended choice set for the PS computation in a real network, we need to generate paths such that the true correlation structure is approximated. That is, the number of paths in the extended choice set using each link in the network should reflect the number of paths in $\mathcal{U}$ using each link. For this purpose we propose a *recursive gateway* algorithm that uses the general stochastic approach presented in Section 3. An extended choice set $\mathcal{C}^{\text{extended}}$ is defined for each origin-destination pair as follows:

- For each link in the network we generate a path and add it to $\mathcal{C}^{\text{extended}}$ if it is not already present.
- A path is generated by recursively drawing links based on weights defined by (2) and (3).
- In order to avoid selecting links scattered over the network, we update s_o, s_d, v and w in Equation (3) each draw so that higher probabilities are assigned to links close to already selected links than those further away, as illustrated below.

The Extended PS attribute for alternative j and observation n is then computed based on $\mathcal{C}_n^{\text{extended}} = \mathcal{C}^{\text{extended}} \cup \mathcal{C}_n$.

We illustrate the heuristic with a small network in Figure 7 where we generate a path (dashed links in part IV) for link $(2, D)$ (bold link in part I). The weight for a link $\ell = (v, w)$ in the first iteration is given by (we use $a = b = 1$):

$$\omega(\ell) = \frac{SP(O, 2)}{SP(O, v) + C(\ell) + SP(w, 2)}$$

and the first link to be drawn is $(O, 3)$ (part II). The weights are then updated according to

$$\omega(\ell) = \frac{SP(3, 2)}{SP(3, v) + C(\ell) + SP(w, 2)}$$

where only one link is possible, namely $(3, 2)$ (part III).

The heuristic has been tested on the example network (Figure 3) and the average size of $\mathcal{C}_n^{\text{extended}}$ is 57 paths. The estimation results, with deterministic utility specifications given by Equation (14), are reported in Table 3 where the reference model $M^{\text{corr}}_{\text{PS}(\mathcal{C})}$ from Table 2 is also shown. $\widehat{\mu}$ and $\widehat{\beta}_{\text{SB}}$ are comparable to the ones obtained by randomly sampling from

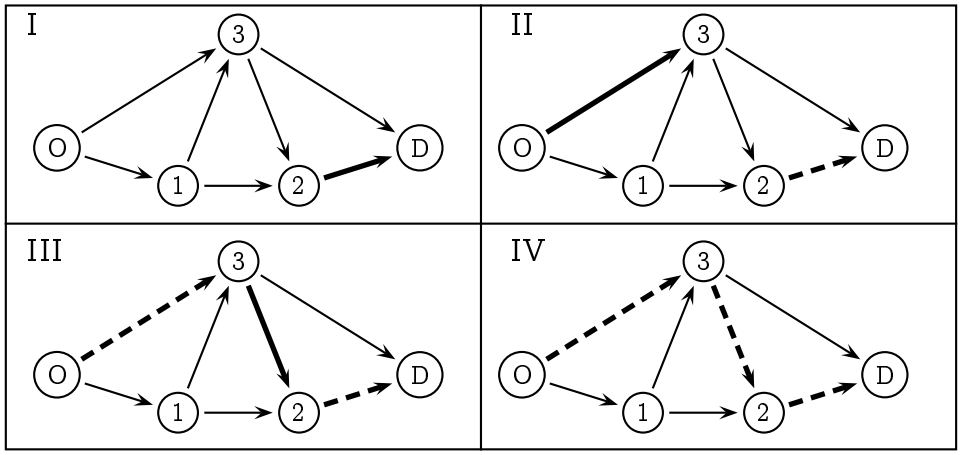

Figure 7: Illustration of Heuristic for Extended Path Size

$\mathcal{U}\backslash\mathcal{C}_n$ (Figure 6) with the same average size of $\mathcal{C}_n^{\text{extended}}$; the scale parameter estimate $\widehat{\mu}$ is improved in $M^{\text{corr}}_{\text{PS}(\mathcal{C}^{\text{extended}})}$ compared to $M^{\text{corr}}_{\text{PS}(\mathcal{C})}$ but remains biased and the speed bump coefficient is unbiased in $M^{\text{corr}}_{\text{PS}(\mathcal{C}^{\text{extended}})}$. The PS coefficient is biased, this is however expected since $\mathcal{C}_n^{\text{extended}}$ is only an approximation of $\mathcal{U}$. Moreover, this approximation does not have the nice properties of a simple random sample and poorer $\widehat{\beta}_{\text{PS}}$ than the results reported in Figure 6 seems reasonable. Finally we note that the model fit is remarkably better for $M^{\text{corr}}_{\text{PS}(\mathcal{C}^{\text{extended}})}$.

6 Conclusions and Future Work

This paper presents a new paradigm for choice set generation and route choice modeling. We view path generation as an importance sampling approach and derive a sampling correction to be added to the path utilities. We hypothesize that the true choice set is the set of all paths connecting an origin-destination pair. Accordingly, we propose to compute the Path Size attribute based on an approximation of the true correlation structure.

We present numerical results based on synthetic data which clearly show the strength of the approach. Models including a sampling correction are remarkably better than the ones that do not. Moreover, unbiased estimation results are obtained if the Path Size attribute is computed based on all paths and not on generated choice sets. This is completely different from route choice modeling practice where generated choice sets are assumed to

	True PSL	$M^{corr}_{PS(\mathcal{C}^{extended})}$ PSL	$M^{corr}_{PS(\mathcal{C})}$ PSL
$\widehat{\beta}_{\mathbf{L}}$ fixed	**-0.3**	**-0.3**	**-0.3**
$\widehat{\mu}$	**1**	**0.885**	**0.724**
Standard error		0.0259	0.0266
t-test w.r.t. 1		-4.43	-12.21
$\widehat{\beta}_{\mathbf{PS}}$	**1**	**1.52**	**0.411**
Standard error		0.102	0.104
t-test w.r.t. 1		5.10	-5.66
$\widehat{\beta}_{\mathbf{SB}}$	**-0.1**	**-0.131**	**-0.266**
Standard error		0.0281	0.0355
t-test w.r.t. -0.1		-1.10	-3.55
Adj. Rho-Squared		0.114	0.103
Final Log-likelihood		-6006.96	-6082.53

Table 3: Estimation Results for Extended Path Size

correspond to the true ones and Path Size (or Commonality Factor for the C-Logit model Cascetta et al., 1996) is computed on these generated path sets. Since it is not possible in real networks to compute these attributes on all paths, we study how many paths are needed in order to obtain unbiased estimates and we propose a heuristic for generating *extended choice sets*.

It is important to note that the proposed sampling approach can be used with Multinomial Logit (MNL) based models (Path Size Logit and C-Logit). A consistent estimator for mixture of MNL (MMNL) models based on samples of alternatives does not exist but is available for Multivariate Extreme Value models (see Nerella and Bhat, 2004, for an empirical study of the bias in MMNL models when estimated on samples of alternatives).

Since the purpose of this paper is to illustrate the proposed methodology, it is appropriate to use synthetic data for which the actual model is known. This allows to test the parameter estimates against their true values. A natural next step is to test the approach on real data. Moreover, future research can be dedicated to sampling of alternatives for prediction.

Acknowledgments

We have benefited from discussions with Moshe Ben-Akiva, Piet Bovy and Mogens Fosgerau.

References

Azevedo, J., Costa, M. S., Madeira, J. S. and Martins, E. V. (1993). An algorithm for the ranking of shortest paths, *European Journal of Operational Research* **69**: 97–106.

Bekhor, S., Ben-Akiva, M. E. and Ramming, S. (2006). Evaluation of choice set generation algorithms, *Annals of Operations Research* **144**(1).

Bekhor, S. and Prato, C. G. (2006). Effects of choice set composition in route choice modelling, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Ben-Akiva, M. (1993). Lecture notes on large set of alternatives. Massachusetts Institute of Technology.

Ben-Akiva, M., Bergman, M., Daly, A. and Ramaswamy, R. (1984). Modeling inter urban route choice behaviour, *in* J. Vollmuller and R. Hamerslag (eds), *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, VNU Science Press, Utrecht, Netherlands, pp. 299–330.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions, *in* R. Hall (ed.), *Handbook of Transportation Science*, Kluwer, pp. 5–34.

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

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Massachusetts.

Ben-Akiva, M. and Ramming, S. (1998). Lecture notes: Discrete choice models of traveler behavior in networks. Prepared for Advanced Methods for Planning and Management of Transportation Networks. Capri, Italy.

Ben-Akiva, M. and Watanatada, T. (1981). Application of a continuous spatial choice logit model, *in* C. F. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data and Econometric Applications*, MIT Press, Cambridge.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of discrete choice models, *Proceedings of the 3rd Swiss Transport Research Conference*, Ascona, Switzerland.

Bierlaire, M. (2007). An introduction to BIOGEME version 1.5. http://biogeme.epfl.ch.

Bierlaire, M., Bolduc, D. and McFadden, D. (to appear). The estimation of Generalized Extreme Value models from choice-based samples, *Transportation Research Part B* . Accepted for publication, doi:10.1016/j.trb.2007.09.003.

Bierlaire, M. and Frejinger, E. (2007a). Route choice modeling with network-free data, *Transportation Research Part C* . doi:10.1016/j.trc.2007.07.007 (article in press).

Bierlaire, M. and Frejinger, E. (2007b). Technical note: A stochastic choice set generation algorithm, *Technical report TRANSP-OR 070213*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Bovy, P. H. L. (2007). Modeling route choice sets in transportation networks: A preliminary synthesis, *Proceedings of the Sixth Triennial Symposium on Transportation Analysis (TRISTAN)*, Phuket, Thailand.

Bovy, P. H. L. and Fiorenzo-Catalano, S. (2006). Stochastic route choice set generation: behavioral and probabilistic foundations, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A. (1996). A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks, *in* J. B. Lesort (ed.), *Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France*.

Cascetta, E. and Papola, A. (2001). Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand, *Transportation Research Part C* **9**(4): 249–263.

Cascetta, E., Russo, E., Viola, F. and Vitetta, A. (2002). A model of route perception in urban road network, *Transportation Research Part B* **36**: 577–592.

de la Barra, T., Pérez, B. and Añez, J. (1993). Mutidimensional path search and assignment, *Proceedings of the 21st PTRC Summer Meeting*, pp. 307–319.

Dial, R. (1971). A probabilistic multipath traffic assignment algorithm which obviates path enumeration, *Transportation Research* **5**(2): 83–111.

Fiorenzo-Catalano, S. (2007). *Choice Set Generation in Multi-modal Transportation Networks*, PhD thesis, Deft University of Technology.

Frejinger, E. (2008). *Route choice analysis: data, models, algorithms and applications*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne.

Frejinger, E. and Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models, *Transportation Research Part B* **41**(3): 363–378.

Friedrich, M., Hofsäss, I. and Wekeck, S. (2001). Timetable-based transit assignment using branch and bound, *Transportation Research Record* **1752**.

Hoogendoorn-Lanser, S. (2005). *Modelling Travel Behaviour in Multi-modal Networks*, PhD thesis, Deft University of Technology.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes, *Journal of Hydrology* **46**: 79–88.

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

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples, *Econometrica* **45**(8): 1977–1988.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull (eds), *Spatial Interaction Theory and Residential Location*, North-Holland, Amsterdam, pp. 75–96.

Morikawa, T. (1996). A hybrid probabilistic choice set model with compensatory and noncompensatory choice rules, *Proceedings of the 7th World Conference on Transport Research*, Vol. 1, pp. 317–325.

Nerella, S. and Bhat, C. R. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models, *Transportation Research Record* **1894**: 11–19.

Prato, C. G. and Bekhor, S. (2006). Applying branch and bound technique to route choice set generation, *Presented at the 85th Annual Meeting of the Transportation Research Board*.

Ramming, M. (2001). *Network Knowledge and Route Choice*, PhD thesis, Massachusetts Institute of Technology.

Swait, J. and Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation, *Transportation Research Part B* **21**(2): 91–102.

Train, K., McFadden, D. and Ben-Akiva, M. (1987). The demand for local telephone service: A fully discrete model for residential calling patterns and service choice, *The RAND Journal of Economics* **18**(1): 109–123.

van Nes, R., Hoogendoorn-Lanser, S. and Koppelman, F. (2006). On the use of choice sets for estimation and prediction in route choice, *Proceedings of the 11th International Conference on Travel Behaviour Research*, Kyoto, Japan.

A Estimation Results

The following tables show the absolute value of t-test values for the four different models discussed in the paper.

Parameter	Nb. Draws	Kumaraswamy parameter a			
		0	1	3	5
$\widehat{\beta}_{SB}$	5	24.68	21.99	17.12	6.65
	10	24.20	20.61	16.68	7.24
	20	21.31	18.10	12.76	7.71
	30	19.11	15.03	10.52	6.93
	40	15.99	14.17	8.92	5.89
$\widehat{\beta}_{PS}$	5	5.17	5.11	0.22	2.46
	10	5.08	3.98	2.18	2.20
	20	6.93	5.23	0.30	3.52
	30	6.93	3.93	0.22	3.28
	40	4.97	5.12	0.10	3.38
$\widehat{\mu}$	5	0.66	6.52	18.7	29.35
	10	0.27	6.47	18.34	29.54
	20	0.06	5.92	18.01	27.49
	30	0.53	5.75	17.45	26.51
	40	0.31	5.38	16.93	25.66

Table 4: Model $M^{NoCorr}_{PS(C)}$ (no convergence for $a > 5$ due to $\widehat{\mu}$ close to zero)

Parameter	Nb. Draws	Kumaraswamy parameter a			
		0	1	3	5
$\widehat{\beta}_{SB}$	5	28.02	24.67	18.92	5.63
	10	29.06	25.26	19.90	6.35
	20	28.38	24.93	18.78	8.20
	30	28.02	23.96	17.71	9.31
	40	26.81	22.88	16.47	9.83
$\widehat{\beta}_{PS}$	5	36.35	28.19	15.18	5.34
	10	37.07	28.12	14.69	5.29
	20	35.01	25.84	12.05	3.98
	30	32.31	23.04	9.81	2.26
	40	29.17	20.50	7.80	0.94
$\widehat{\mu}$	5	3.06	4.54	19.25	31.3
	10	3.69	4.65	19.23	32.64
	20	3.56	4.43	19.68	32.41
	30	3.75	4.41	19.15	31.65
	40	3.37	4.38	18.77	30.99

Table 5: Model $M^{NoCorr}_{PS(\mathcal{U})}$ (no convergence for $a > 5$ due to $\widehat{\mu}$ close to zero)

Parameter	Nb. Draws	Kumaraswamy parameter α								
		0	1	3	5	7	10	15	20	30
	5	1.99	2.10	3.54	4.67	4.73	4.45	2.22	1.34	0.50
	10	0.48	0.17	3.31	3.56	2.93	2.45	0.72	0.13	1.40
$\widehat{\beta}_{\mathrm{SB}}$	20	1.58	1.56	0.06	0.73	1.82	1.22	0.37	0.78	1.98
	30	2.98	3.76	2.11	0.19	0.95	0.35	0.36	1.48	2.56
	40	5.19	4.17	3.63	1.31	0.01	0.48	0.70	1.47	2.56
	5	4.62	4.87	2.66	3.49	4.36	3.91	4.23	4.70	3.05
	10	3.93	3.45	5.82	5.66	4.80	3.51	2.81	3.01	3.34
$\widehat{\beta}_{\mathrm{PS}}$	20	4.72	4.57	4.22	5.02	6.86	6.40	3.95	3.40	4.18
	30	3.85	2.99	3.99	5.48	4.64	7.21	5.26	4.39	4.19
	40	1.62	3.60	3.39	5.25	7.66	7.09	5.75	5.33	4.80
	5	8.78	10.18	12.56	11.14	12.04	8.12	3.88	2.12	3.28
	10	8.35	10.03	12.69	12.21	11.66	10.08	5.48	2.86	1.65
$\widehat{\mu}$	20	8.26	8.21	10.95	11.26	12.01	10.86	7.05	4.06	1.83
	30	8.06	6.92	8.03	11.02	11.97	10.38	8.03	3.72	2.03
	40	7.22	6.84	6.53	10.03	11.97	10.38	8.03	3.72	2.03

Table 6: Model $\mathrm{M}^{\mathrm{Corr}}_{\mathrm{PS}(\mathcal{C})}$

Parameter	Nb. Draws	Kumaraswamy parameter a								
		0	1	3	5	7	10	15	20	30
	5	1.22	1.94	1.34	0.19	0.46	0.22	1.53	1.17	1.17
	10	1.79	2.16	1.23	0.56	0.31	0.14	0.86	1.11	1.58
$\widehat{\beta}_{SB}$	20	2.32	2.33	1.42	0.93	0.52	0.60	0.66	0.29	1.08
	30	1.94	2.08	1.70	0.82	0.82	0.60	0.26	0.65	1.23
	40	1.85	1.82	1.53	0.90	0.83	0.56	0.16	0.62	0.98
	5	2.04	1.67	1.45	0.60	1.31	0.02	0.23	1.85	1.32
	10	1.77	1.55	1.99	0.85	0.80	0.57	0.18	1.04	1.27
$\widehat{\beta}_{PS}$	20	1.37	1.41	1.59	0.88	1.04	0.79	0.19	0.34	0.94
	30	1.16	0.95	1.41	0.88	1.07	0.61	0.57	0.24	0.92
	40	1.17	0.93	0.94	0.67	0.87	0.62	0.58	0.24	0.80
	5	1.70	1.27	0.48	0.41	1.35	0.36	1.48	1.62	1.16
	10	2.52	1.38	0.63	0.20	1.19	1.57	0.17	1.22	1.91
$\widehat{\mu}$	20	2.03	2.31	0.40	0.07	1.54	2.03	0.83	0.35	0.84
	30	1.78	2.37	1.55	0.63	1.37	1.51	1.48	0.96	0.44
	40	2.08	1.36	1.27	0.44	1.37	1.51	1.48	0.96	0.44

Table 7: Model $M^{Corr}_{PS(\mathcal{U})}$