RANDOM SAMPLING OF ALTERNATIVES IN A ROUTE CHOICE CONTEXT

Emma Frejinger
Transport and Mobility Laboratory (TRANSP-OR), EPFL

Abstract

In this paper we present a new point of view on choice set generation and route choice modeling. Choice sets of paths need to be defined when modeling route choice behavior using random utility models. Existing approaches generate paths and assume that actual choice sets are found. On the contrary, we assume that actual choice sets are the sets of all paths connecting each origin-destination pair. These sets are however unknown and we propose a stochastic path generation algorithm that corresponds to an importance sampling approach. The path utilities should then be corrected according to the used sampling protocol in order to obtain unbiased parameter estimates. We derive such a sampling correction for the proposed algorithm. Furthermore, based on the assumption that actual choice sets contain all paths, we argue that Path Size (or Commonality Factor) attributes should be computed on all paths (or as many as possible) in order to reflect the true correlation structure.

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, unbiased estimation results are obtained if the Path Size attribute is computed based on all paths and not on generated choice sets. In real networks the set of all paths is unknown, we therefore study how many paths are needed for the Path Size computation in order to obtain unbiased results. The parameter estimates improve rather rapidly with the number of paths which is promising for real applications.

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 number of paths (intractable if the network contains loops) connect an origin o and destination d. This set, referred to as the universal choice set $\mathcal{U}$, is unknown. In order to estimate a route

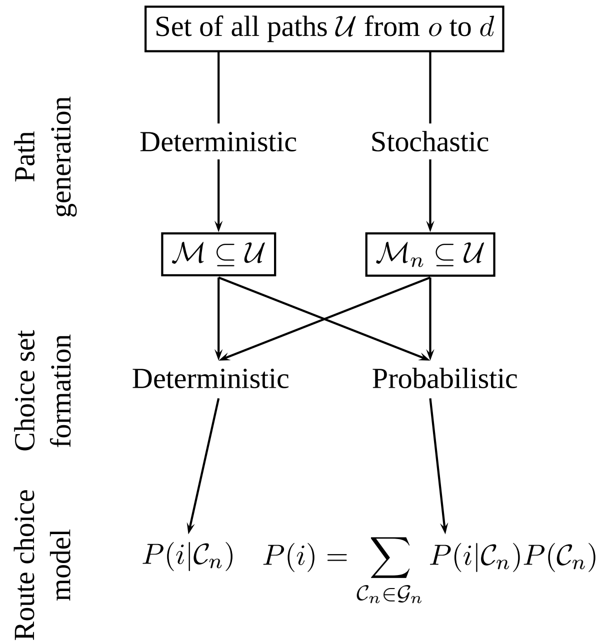

Figure 1: Route choice modeling process

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. The former refers to algorithms always generating the same set $\mathcal{M}$ of paths for a given origin-destination pair whereas an individual (or observation) specific subset $\mathcal{M}_n$ is generated with stochastic approaches. A choice set $\mathcal{C}_n$ for individual n can be defined based on $\mathcal{M}$ (or $\mathcal{M}_n$) 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. $P(i|\mathcal{C}_n)$ is the probability of route i given $\mathcal{C}_n$. Defining choice sets in a probabilistic way in 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, Morikawa, 1996 and Cascetta and Papola, 2001 for more details on the probabilistic approaches.)

In this paper we focus on stochastic path generation and specifically on how to take into account in the route choice model that we limit the analysis to paths in $\mathcal{M}_n$. We view path generation as importance sampling of alternatives and we propose a correction of the path utilities for the sampling approach. This is a substantially different approach from existing ones because we hypothesize that the true choice set is the universal one. This assumption also has implications on how correlation should be modeled which is further investigated with numerical results.

In the following section we give an overview of existing path generation algorithms. An introduction to sampling of alternatives is presented in Section 3. We describe the proposed algorithm in Section 4 and we continue by deriving the sampling correction in Section 5. Numerical results based on synthetic data are presented (Section 6) before some conclusions.

2 PATH GENERATION ALGORITHMS

Many heuristics for generating paths have been proposed in the literature. Most of them are deterministic approaches, for example, labeled paths (Ben-Akiva et al., 1984), link elimination (Azevedo et al., 1993), link penalty (de la Barra et al., 1993), constrained k-shortest paths (e.g. van der Zijpp and Catalano, 2005) and branch-and-bound (Friedrich et al., 2001, Hoogendoorn-Lanser, 2005 and Prato and Bekhor, 2006).

Stochastic approaches are of interest for this paper since we view path generation as sampling of alternatives. Only two stochastic algorithms have been proposed in the literature. Ramming (2001) uses 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 so-called doubly stochastic choice set generation approach. It is similar to the simulation method but the generalized cost function has both random parameters and random attributes.

Bovy (2007) discusses the role of choice set generation in route choice modeling and gives an overview of existing approaches. Bekhor and Prato (2006) analyze empirically the effects of choice set generation on route choice model estimation results. They observe differences in the estimation results for various algorithms. They conclude that the branch-and-bound algorithm (Prato and Bekhor, 2006) performs best.

Existing approaches, both deterministic and stochastic, assume that actual choice sets are generated. Empirical studies suggest however that this is not true since in general not even all observed paths are generated (see e.g. Ramming, 2001, Prato and Bekhor, 2006, Frejinger and Bierlaire, 2007 and Bierlaire and

Frejinger, 2007a). We assume on the contrary that the true choice set is $\mathcal{U}$. This set is however to large to be enumerated and we therefore define a random sample $\mathcal{M}_n$. In order to obtain unbiased estimation results, the path utilities must be corrected according to the used sampling protocol. In the following section we give a brief introduction to sampling of alternatives.

3 SAMPLING OF ALTERNATIVES

The multinomial logit (MNL) model can be consistently estimated on a subset of alternatives. 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, Bolduc and McFadden (2006) show that the more general family of GEV models can also be consistently estimated and propose a new estimator. Here we focus however on the MNL model.

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. Namely, $q(\mathcal{C}_n|i)$ is then equal to $q(\mathcal{C}_n|j)\ \forall\ j \in \mathcal{C}_n$ (uniform conditioning property, McFadden, 1978) and the correction for sampling bias cancels out in Equation (1). This simple random sampling protocol is however not appropriate in a path generation context. First of all, we are unaware of any algorithm generating paths with equal probabilities without first enumerating all paths in $\mathcal{U}$. Second, due to the large (possibly intractable) number of paths, a simple random sample is likely to contain many alternatives that a traveler would never consider. Comparing the chosen path to a set of highly unattractive alternatives would not provide much information on the traveler's route choice. In this context, a simple random sample would need to be prohibitively large. We therefore propose a path generation algorithm that corresponds to an importance sampling approach where attractive paths have higher probability of being sampled than unattractive paths. In this case, the correction terms in Equation (1) do not cancel out and path utilities must be corrected in order to obtain unbiased results.

The Path Size Logit (PSL), proposed by Ben-Akiva and Ramming (1998) (see also Ben-Akiva and Bierlaire, 1999), and the C-Logit (Cascetta et al., 1996) models are the most commonly used MNL models for route choice analysis. An attribute, Path Size (PS) or Commonality Factor respectively, captures the correlation among paths and is added to the deterministic utilities. Up to date, these attributes are computed based on the generated choice sets. Since we assume that the true choice set is $\mathcal{U}$ we hypothesize that these attributes should be computed based on a path set that is as large as possible in order to approximate the true correlation structure. We study this hypothesis numerically in Section 6.

Note that existing stochastic path generation approaches may also be viewed as importance sampling approaches. It is however unclear to us how to compute the sampling correction for these algorithms.

4 A STOCHASTIC PATH GENERATION APPROACH

In this section, we first present a general stochastic approach for generating paths (also described in Bierlaire and Frejinger, 2007b). The approach is flexible and can be used in various algorithms including those presented in the literature. We then describe a biased random walk algorithm that is used in this paper.

This stochastic path generation approach is based on the concept of subpath where a subpath is a sequence of links. (A link is a special case of a subpath.) We associate a probability with a subpath based on its distance to the shortest path. More precisely, its probability is defined by the double bounded Kumaraswamy distribution (proposed by Kumaraswamy, 1980) whose cumulative distribution function is $F(x_s|a,b) = 1 - (1 - {x_s}^a)^b$ for $x_s \in [0,1]$. a and b are shape parameters and for a given subpath s with source node v and sink node w, x_s is defined as

$$x_s = \frac{SP(o,d)}{SP(o,w) + C(s) + SP(w,d)},$$

where $C(s)$ is the cost of s, o the origin, d the destination and $SP(v_1, v_2)$ is the cost of the shortest path between nodes v_1 and v_2. Any generalized cost can be used in this context. Note that x_s equals one if s is part of the shortest path and $x_s \to 0$ as $C(s) \to \infty$. In Figure 2 we show the cumulative distribution function for different values of a when $b = 1$. The probabilities assigned to the subpaths can be controlled by the definition of the distribution parameters. High values of

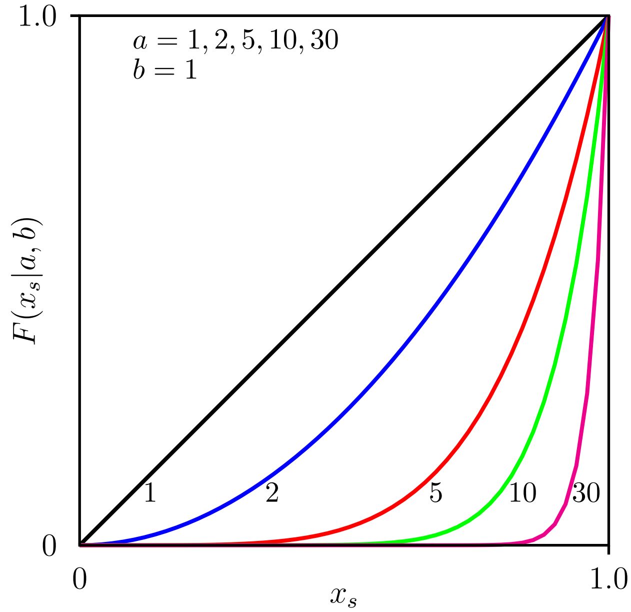

Figure 2: Kumaraswamy distribution - cumulative distribution function

a when $b = 1$ yield low probabilities for subpaths with high cost. Low values of a have the opposite effect.

This approach can be used in various algorithms. For example, in an algorithm similar to link elimination approach but where the choice of subpaths to be eliminated is stochastic. Another example is a gateway algorithm, where a subpath is selected anywhere in the network, using the probability distribution described above. A generated path is then composed of three segments: the shortest path from the origin to the source node of the subpath, the subpath itself, and the shortest path from the sink node of the subpath to the destination. This gateway algorithm is used by Bierlaire, Frejinger and Stojanovic (2006) for modeling long distance route choice behavior in Switzerland.

In this paper, we use a biased random walk algorithm which has some properties which makes it appropriate as an importance sampling approach. First, it can generate potentially any path in $\mathcal{U}$. Second, path selection probabilities can be computed in a straightforward way.

4.1 Biased Random Walk Algorithm

Starting from the origin, this algorithm selects a link using the probability distribution described previously. Another link starting at the sink node of the first one is then selected and this process is applied until the destination is reached and a complete path has been generated. 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 a simple random sample of paths.

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

$$q(j) = \prod_{\ell \in \Gamma_j} q(\ell|\mathcal{E}_v, a, b) \tag{2}$$

where ℓ denotes a link, v its source node and $\mathcal{E}_v$ the set of outgoing links from v. In accordance with the general approach presented previously $q(\ell|\mathcal{E}_v, a, b)$ is defined by the Kumaraswamy distribution using

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

5 CORRECTION FOR SAMPLING IN ROUTE CHOICE MODELS

Importance sampling takes expected choice probabilities into account; paths which are expected to have high choice probabilities have higher sampling probabilities than paths with lower expected choice probabilities. As discussed in Section 3 the correction terms $q(\mathcal{C}_n|j)\ \forall\ j \in \mathcal{C}_n$ must be defined since they do not cancel out for this type of sampling protocol. It is worth mentioning 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. 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.

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 $\mathcal{U}$

and adding the chosen path to it ($|\widetilde{\mathcal{C}}_n| = R+1$). In theory $\mathcal{U}$ can be unbounded, here we assume that paths with many loops have infinitely small sampling probability (due to the importance sampling) and we treat $\mathcal{U}$ as bounded with size J. (In practice J is unknown.) Each path $j \in \mathcal{U}$ has sampling probability $q(j)$ and $K\sum_{j\in\mathcal{U}} q(j) = 1$ where K is a normalizing constant. This constant does not play a role in the same way as $K_{\mathcal{C}_n}$ in Equation (5) and is therefore ignored in the following equations.

The outcome of this protocol is $(\widetilde{k}_1, \widetilde{k}_2, \ldots, \widetilde{k}_J)$ where $\widetilde{k}_j$ is the number of times alternative j was drawn ($\sum_{j\in\mathcal{U}} \widetilde{k}_j = R$). 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

$$P(\widetilde{k}_1, \widetilde{k}_2, \ldots, \widetilde{k}_J) = \frac{R!}{\prod_{j\in\mathcal{U}} \widetilde{k}_j!} \prod_{j\in\mathcal{U}} q(j)^{\widetilde{k}_j}. \tag{3}$$

The number of times alternative j appears in $\widetilde{\mathcal{C}}_n$ is $k_j = \widetilde{k}_j + \delta_{jc}$, 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 ($\mathcal{C}_n = \{j \in \mathcal{U} \mid k_j > 0\}$). 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 was not drawn nor were any duplicates.

Using Equation (3), the probability of drawing $\mathcal{C}_n$ given the chosen alternative i can be defined as

$$q(\mathcal{C}_n|i) = q(\widetilde{\mathcal{C}}_n|i) = \frac{R!}{(k_i-1)!\prod\limits_{\substack{j\in\mathcal{C}_n\\ j\neq i}} k_j!} q(i)^{k_i-1} \prod_{\substack{j\in\mathcal{C}_n\\ j\neq i}} q(j)^{k_j} = K_{\mathcal{C}_n}\frac{k_i}{q(i)} \tag{4}$$

where $K_{\mathcal{C}_n} = \frac{R!}{\prod_{j\in\mathcal{C}_n} k_j!}\prod_{j\in\mathcal{C}_n} q(j)^{k_j}$. We can now define the probability that an individual chooses alternative i in $\mathcal{C}_n$ as

$$P(i|\mathcal{C}_n) = \frac{e^{V_{in}+\ln\left(\frac{k_i}{q(i)}\right)}}{\sum\limits_{j\in\mathcal{C}_n} e^{V_{jn}+\ln\left(\frac{k_j}{q(j)}\right)}}, \tag{5}$$

where $K_{\mathcal{C}_n}$ in Equation (4) does not play a role 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 (2).

Finally we note that by design the observed route is included in $\mathcal{C}_n$. Hence, there is no issue of coverage which has been discussed in the literature (e.g. Ramming, 2001 and Prato and Bekhor, 2006).

6 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 PS attribute; and
- the biased random walk algorithm parameters.

Synthetic data is 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).

6.1 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 PSL model, and we specify a utility function for each alternative $i \in \mathcal{U}$: $U_i = \beta_{\text{PS}}\text{PS}_i^{\mathcal{U}} + \beta_{\text{L}}\text{Length}_i + \beta_{SB}\text{SpeedBumps}_i + \varepsilon_i$, where $\beta_{\text{PS}} = 1$, $\beta_{\text{L}} = -0.3$, $\beta_{\text{SB}} = -0.1$ and ε_i is distributed Extreme Value with scale 1 and location 0. The PS attribute is defined by $\text{PS}_i^{\mathcal{U}} = \sum_{\ell \in \Gamma_i} \frac{L_\ell}{L_i} \frac{1}{\sum_{j \in \mathcal{U}} \delta_{\ell j}}$ where Γ_i is the set of links in path i, L_ℓ is length of link ℓ, L_i length of path i and $\delta_{\ell j}$ equals one if path j contains link ℓ, zero otherwise. Note that we explicitly index $\mathcal{U}$ since later on we compute PS based on sampled choice sets. 3000 observations have been generated by associating a choice with the alternative having the highest utility.

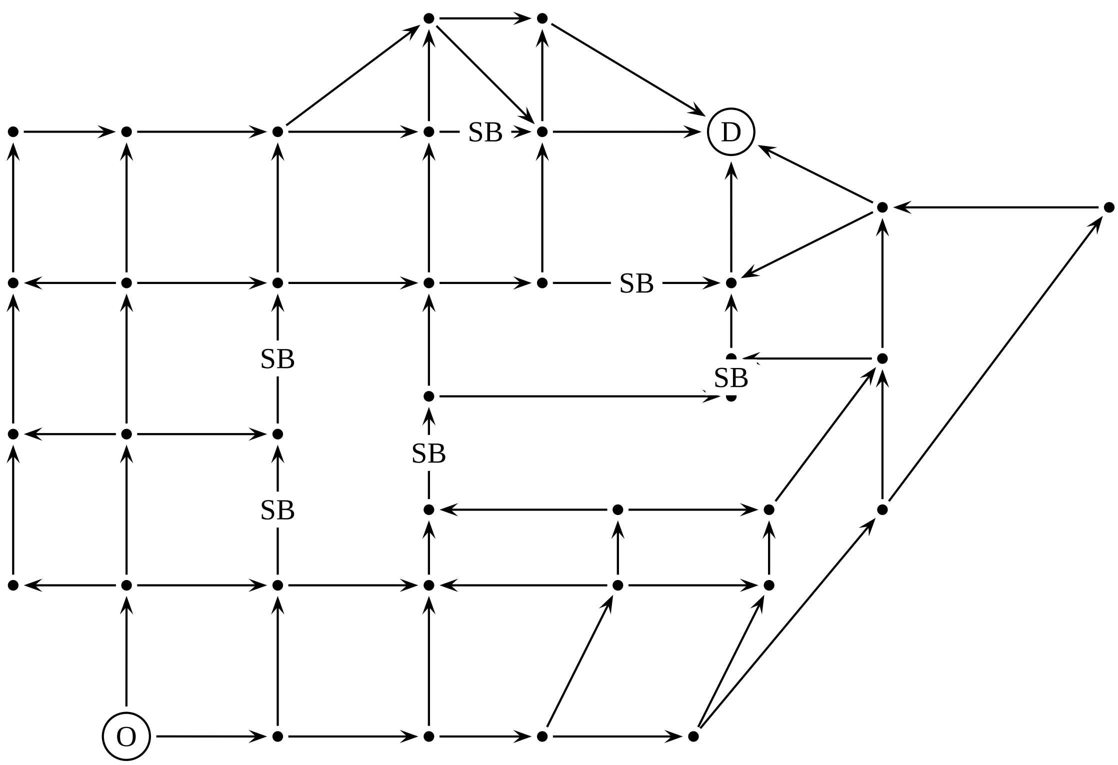

Figure 3: Example Network

6.2 Model Specifications

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

Table 1: Model Specifications

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

$$
\begin{aligned}
M^{\text{NoCorr}}_{PS(\mathcal{C})} \;\; V_{in} =&\; \mu \left(\beta_{\text{PS}}\text{PS}^{\mathcal{C}}_{in} - 0.3\text{Length}_i + \beta_{SB}\text{SpeedBumps}_i\right)\\
M^{\text{Corr}}_{PS(\mathcal{C})} \;\; V_{in} =&\; \mu \left(\beta_{\text{PS}}\text{PS}^{\mathcal{C}}_{in} - 0.3\text{Length}_i + \beta_{SB}\text{SpeedBumps}_i + \ln(\frac{k_i}{q(i)})\right)\\
M^{\text{NoCorr}}_{PS(\mathcal{U})} \;\; V_{i} =&\; \mu \left(\beta_{\text{PS}}\text{PS}^{\mathcal{U}}_{i} - 0.3\text{Length}_i + \beta_{SB}\text{SpeedBumps}_i\right)\\
M^{\text{Corr}}_{PS(\mathcal{U})} \;\; V_{i} =&\; \mu \left(\beta_{\text{PS}}\text{PS}^{\mathcal{U}}_{i} - 0.3\text{Length}_i + \beta_{SB}\text{SpeedBumps}_i + \ln(\frac{k_i}{q(i)})\right).
\end{aligned}
$$

The PS attribute based on sampled paths is defined by $\text{PS}^{\mathcal{C}}_{in} = \sum_{\ell\in\Gamma_i} \frac{L_\ell}{L_i} \frac{1}{\sum_{j\in\mathcal{C}_n} \delta_{\ell j}}$. 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.

6.3 Estimation Results

Table 2 shows 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). The t-test values show that only the model including a sampling correction and PS computed based on $\mathcal{U}$ ($M^{\text{Corr}}_{PS(\mathcal{U})}$) 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 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

	True PSL	$M^{\text{NoCorr}}_{PS(\mathcal{C})}$ PSL	$M^{\text{Corr}}_{PS(\mathcal{C})}$ PSL	$M^{\text{NoCorr}}_{PS(\mathcal{U})}$ PSL	$M^{\text{Corr}}_{PS(\mathcal{U})}$ 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	-6.58	0.0238
t-test w.r.t. -0.1		-7.24	-3.55	0.41	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, Bierlaire, 2003) has been used for all model estimations

Table 2: Path Size Logit Estimation Results

fit is remarkably better for the models with correction compared to those without. Despite of this the model with PS computed based on $\mathcal{C}_n$ ($M^{\text{Corr}}_{PS(\mathcal{C})}$) 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 $\mathcal{U}$ is unknown.

In order to study how many paths are needed for computing the PS attribute we generate an extended choice set $\mathcal{C}_n^{\text{extended}}$ that is only used to compute the PS. In addition to all paths in $\mathcal{C}_n$ we randomly draw (uniform distribution) an extra number of paths from $\mathcal{U}\backslash\mathcal{C}_n$ and add these to $\mathcal{C}_n^{\text{extended}}$. The deterministic utilities for a model including sampling correction are now defined as

$$V_{in} = \mu\left(\beta_{\text{PS}}\text{PS}_i^{\mathcal{C}^{\text{extended}}} - 0.3\text{Length}_i + \beta_{SB}\text{SpeedBumps}_i + \ln(\frac{k_i}{q(i)})\right) \ \forall\ i \in \mathcal{C}_n$$

where $\text{PS}_i^{\mathcal{C}^{\text{extended}}} = \sum_{\ell\in\Gamma_i}\frac{L_\ell}{L_i}\frac{1}{\sum_{j\in\mathcal{C}_n^{\text{extended}}}\delta_{\ell j}}$. The estimation results as a function of the average size of $\mathcal{C}_n^{\text{extended}}$ are shown in Figure 4 where the average number of paths in $\mathcal{C}_n$ (9.66) is the first data point and $|\mathcal{U}| = 170$ the last one. For each parameter estimate we report the absolute value of 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 $\mathcal{C}_n^{\text{extended}}$ where both the speed bump and path size coefficients are unbiased. The scale parameter is unbiased from 80 additional paths. Even though many paths (average number in $\mathcal{C}_n^{\text{extended}}$ approximately $0.5|\mathcal{U}|$) are needed in order for all parameter estimates to be unbiased, we believe that these results are promising from a practical point of view. Indeed, by using more paths than those in $\mathcal{C}_n$ for computing the PS attribute we can have an important improvement in the parameter estimates.

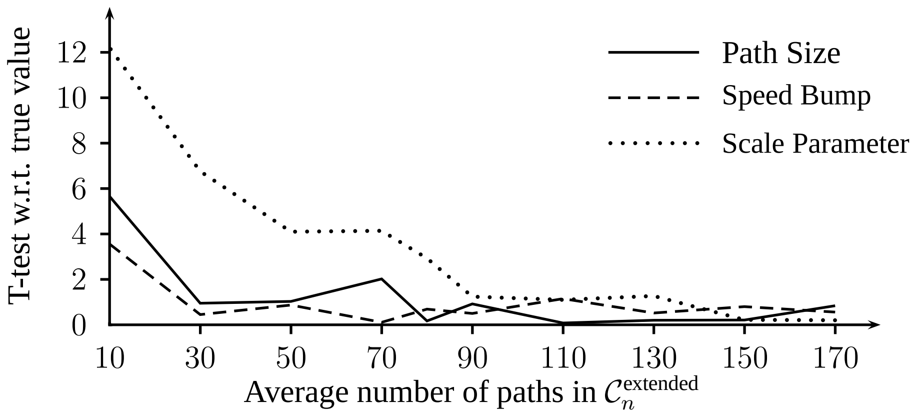

Figure 4: Estimation results for corrected model with PS based on various sizes

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 5 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 shortest path. Figure 6 shows the absolute value of the t-tests w.r.t. the true values for model $M_{PS(\mathcal{U})}^{\text{Corr}}$. With few exceptions the parameters are unbiased for both 10 and 40 draws and for all values of a. (A line is shown at the critical value 1.96.) These results indicate that the estimation results are robust w.r.t. to the algorithm parameter settings.

Finally we note that the other three model specifications ($M_{PS(\mathcal{C})}^{\text{NoCorr}}$, $M_{PS(\mathcal{C})}^{\text{Corr}}$ and $M_{PS(\mathcal{U})}^{\text{NoCorr}}$) have biased estimates for at least one parameter for all values of a and for all number of draws. The detailed results are shown in the Appendix.

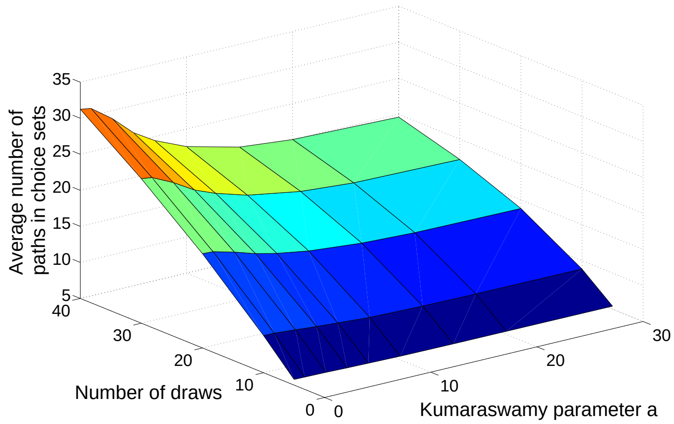

Figure 5: Average number of paths in choice sets

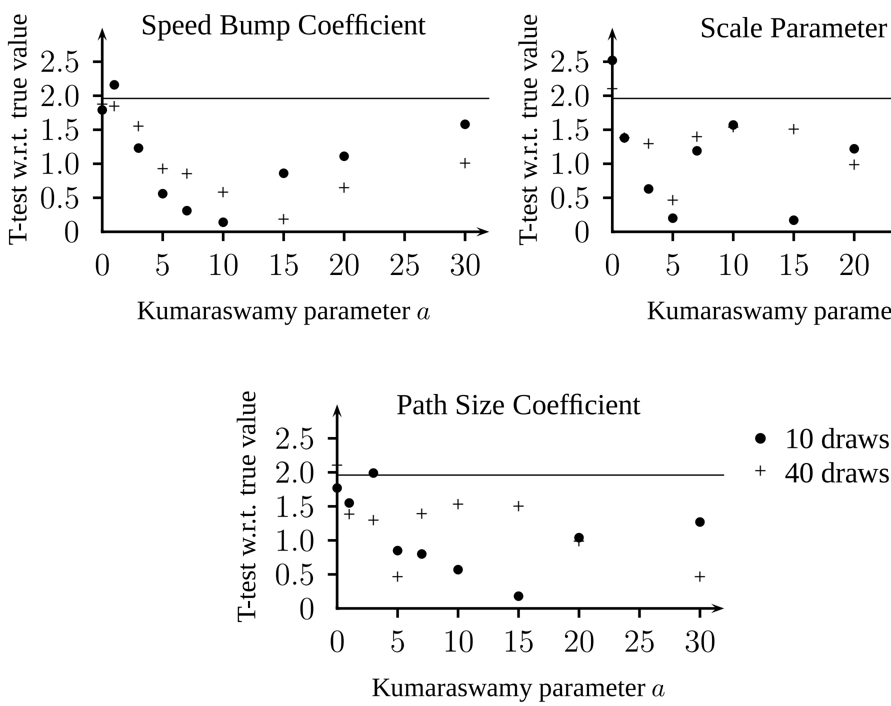

Figure 6: T-test values w.r.t. true values for the coefficients of $M^{\text{Corr}}_{PS(\mathcal{U})}$

7 CONCLUSIONS AND FUTURE WORK

This paper presents a substantially different approach for choice set generation and route choice modeling compared to existing ones. 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 on all path (or as many as possible) so that it reflects the actual 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 praxis where generated choice sets are assumed to correspond to the true ones and Path Size (or Commonality Factor) 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. The results show important improvement of estimates when more paths than the ones in the choice set are used which is promising for real applications.

In the near future we will test the approach on a GPS data set from Sweden and continue the study on the computation of the Path Size attribute.

8 ACKNOWLEDGMENTS

Michel Bierlaire has made important contributions to this paper. I have also 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. and Prato, C. G. (2006). Effects of choice set composition in route choice modelling, *Paper presented at 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. (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.

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. (2006). The estimation of generalized extreme value models from choice-based samples, *Technical Report TRANSP-OR 060810*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

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.

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.

Bierlaire, M., Frejinger, E. and Stojanovic, J. (2006). A latent route choice model in Switzerland, *Proceedings of the European Transport Conference*, Strasbourg, France.

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

Bovy, P. H. and Fiorenzo-Catalano, S. (2006). Stochastic route choice set generation: behavioral and probabilistic foundations, *Paper presented at 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.

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.

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

McFadden, D. (1978). *Modelling the choice of residential location*, Spatial interaction theory and residential location, North-Holland, 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.

Prato, C. 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.

van der Zijpp, N. and Catalano, S. F. (2005). Path enumeration by finding the constrained k-shortest paths, *Transportation Research Part B* **39**(6): 545–563.

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
$\widehat{\beta}_{\text{SB}}$	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
$\widehat{\beta}_{\text{PS}}$	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 3: Model $M_{PS(\mathcal{C})}^{\text{NoCorr}}$ (no convergence for $a > 5$ due to very small $\widehat{\mu}$)

Parameter	Nb. Draws	Kumaraswamy parameter a								
		0	1	3	5	7	10	15	20	30
$\widehat{\beta}_{\text{SB}}$	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
	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
$\widehat{\beta}_{\text{PS}}$	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
	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
$\widehat{\mu}$	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
	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 4: Model $M^{\text{Corr}}_{PS(\mathcal{C})}$

Parameter	Nb. Draws	Kumaraswamy parameter a			
		0	1	3	5
$\widehat{\beta}_{\text{SB}}$	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
$\widehat{\beta}_{\text{PS}}$	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 $M^{\text{NoCorr}}_{PS(\mathcal{U})}$ (no convergence for $a > 5$ due to very small $\widehat{\mu}$)

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
$\widehat{\beta}_{\text{SB}}$	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
$\widehat{\beta}_{\text{PS}}$	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 6: Model $M_{PS(\mathcal{U})}^{\text{Corr}}$