

Expanded path size attribute for route choice models including sampling correction

E. Frejinger* M. Bierlaire[†] M. Ben-Akiva[‡]

March 4, 2009

*Results presented in this paper are accepted for publication in
Transportation Research Part B*

Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (to appear). Sampling of Alternatives for Route Choice Modeling, Transportation Research Part B: Methodological (accepted for publication, March 3, 2009).

*Royal Institute of Technology, Centre for Transport Studies, Teknikringen 78B, SE-100 44 Stockholm, Sweden. E-mail: emma.frejinger@infra.kth.se

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

[‡]Massachusetts Institute of Technology, Room 1-181, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. E-mail: mba@mit.edu

Abstract

Recently, we proposed a new paradigm for choice set generation in the context of route choice model estimation. As detailed in Frejinger and Bierlaire (2007), we assume that choice sets contain all paths connecting each origin-destination pair. Although it is behaviorally questionable, this assumption is made in order to avoid bias in the econometric model. 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. In Frejinger and Bierlaire (2007) we derive such a sampling correction for the multinomial logit (MNL) model.

The path size logit model is a MNL model where a path size (PS) attribute is included in the deterministic utilities. The PS attribute should capture the correlation among routes. It is generally computed based on sampled paths only but we argue that it should capture the correlation among all routes (universal choice set). This becomes problematic since the universal choice set is unknown in practice. In this paper we present a generalization of the PS attribute called expanded PS (EPS). It is computed based on sampled paths only but involves an expansion factor that corrects for the sampling.

We present estimation results based on synthetic data that clearly show the strength of this approach. Unbiased parameter estimates are only obtained for models including a sampling correction. Moreover, the results show that EPS is superior to the original PS attribute.

1 Introduction

Route choice modeling is complex for various reasons and involves several steps before the actual 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 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 computationally 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 enu-

meration 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 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.

Several route choice models $P(i|\mathcal{C}_n)$ exist in the literature. Multinomial logit based models; C-logit (Cascetta et al., 1996) and path size logit (Ben-Akiva and Ramming, 1998, and Ben-Akiva and Bierlaire, 1999) are the most frequently used models in practice due to their simple structure. In these models, the utilities are deterministically corrected with an attribute that accounts for correlation. More complex models explicitly capturing the correlation among paths have been proposed in the literature. The link-nested logit (Vovsha and Bekhor, 1998) model has a cross-nested logit structure but is difficult to estimate because of the large number of nesting parameters. Error Component (Bekhor et al., 2002, and Frejinger and Bierlaire, 2007) and multinomial probit (Yai et al., 1997) models have also been proposed which require simulated maximum likelihood estimation.

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

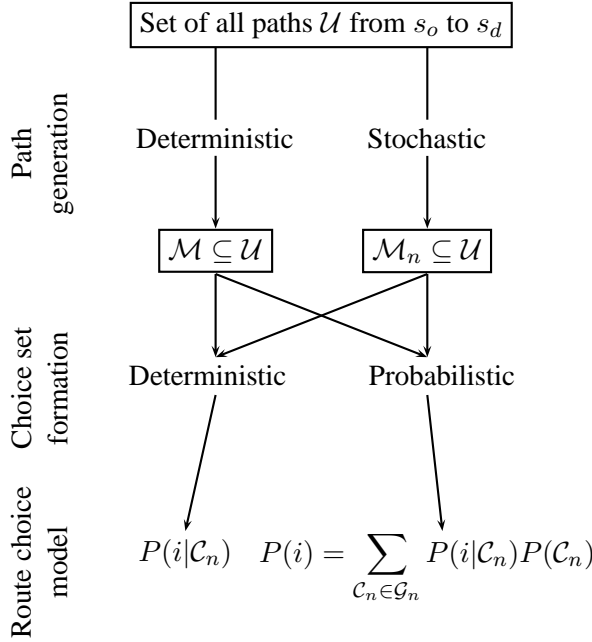


Figure 1: Choice Set Generation Overview

Bekhor (2006), Bekhor and Prato (2006), Van 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. Namely, Ramming (2001) finds at best 91% of the observations by combining various algorithms and Prato and Bekhor (2006) find 91% of the observations using their branch-and-bound algorithm.

Recently, we proposed a new paradigm based on a sampling approach (Frejinger and Bierlaire, 2007, and Frejinger, 2008). In order to avoid bias in the econometric model, we assume that all paths connecting an origin-destination pair belong to the choice set. Since this set is in general impossible to generate, we propose an importance sampling approach and a corresponding correction of the path utilities. Unlike existing choice set generation approaches, which aim at generating actual choice sets, we focus on obtaining unbiased parameter estimates using samples of alternatives.

The main flaw of the approach has so far been how to define the path size attribute that should capture correlation among alternatives. In previous work we propose a heuristic that generates a larger choice set than the sampled one that is intended to approximate the universal choice set. This paper proposes a theoretically more appealing approach that corrects also the path size (PS) attribute according to the used sampling protocol. We call this new PS formulation ex-

panded PS.

In the following section we briefly present the sampling approach on which the expanded PS is based. We discuss the proposed formulation in Section 3 before presenting numerical results. We finish with some conclusions and issues for future research.

2 Sampling Approach

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

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

where μ is a scale parameter and V_{in} is the deterministic utility. It also includes an alternative specific term, $\ln q(\mathcal{C}_n|j)$ that corrects for sampling bias. This correction term is based on the probability $q(\mathcal{C}_n|j)$ of sampling \mathcal{C}_n given that j is the chosen alternative. See for example Ben-Akiva and Lerman (1985) for a more detailed discussion on sampling of alternatives. Bierlaire et al. (2008) have recently shown that multivariate extreme value (also known as generalized extreme value) models can be consistently estimated as well and propose a new estimator.

When using a sampling protocol selecting attractive alternatives with higher probability than unattractive alternatives (importance sampling), the correction terms in (1) do not cancel out. 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.

Frejinger and Bierlaire (2007) derive a sampling correction in the context of route choice

$$q(\mathcal{C}_n|j) = \frac{k_{jn}}{q(j)} \quad \forall j \in \mathcal{C}_n \quad (2)$$

where k_{jn} is the number of times path j was drawn while sampling choice set \mathcal{C}_n and $q(j)$ is the probability of drawing path j . The correction is based on a sample protocol where paths are sampled with replacement and the chosen alternative is always added to the choice set, even if it is sampled. Furthermore, they propose

a biased random walk algorithm that allows to compute $q(\mathcal{C}_n|j)$ in a straightforward way. The algorithm is based on a distribution with parameters that control the random walk more or less towards the shortest path. With this algorithm the probability $q(j)$ of generating a path j is the probability of selecting the ordered sequence of links Γ_j

$$q(j) = \prod_{\ell \in \Gamma_j} q(\ell|\mathcal{E}_v, b_1, b_2). \quad (3)$$

The probability of selecting a link $\ell = (v, w)$ given the set of outgoing links \mathcal{E}_v at node v is defined by

$$q(\ell|\mathcal{E}_v, b_1, b_2) = \frac{\omega(\ell|b_1, b_2)}{\sum_{m \in \mathcal{E}_v} \omega(m|b_1, b_2)} \quad (4)$$

where b_1 and b_2 are the parameters of the distribution and $\omega(\ell|b_1, b_2)$ are weights of each link. The weights are

$$\omega(\ell|b_1, b_2) = 1 - (1 - x_\ell^{b_1})^{b_2} \quad (5)$$

with

$$x_\ell = \frac{SP(v, s_d)}{C(\ell) + SP(w, s_d)} \quad (6)$$

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 . For more details on the correction term and the algorithm we refer to Frejinger and Bierlaire (2007) and Frejinger (2008).

3 Expanded Path Size

As discussed previously, we base our model on the assumption that all paths connecting an origin-destination pair belong to the choice set. We should therefore have a description of the correlation among paths that is consistent with this assumption. We use the path size logit (PSL) model proposed by Ben-Akiva and Ramming (1998) and Ben-Akiva and Bierlaire (1999). It is a multinomial logit model that includes a PS attribute which is intended to correct the path utilities for correlation.

The PS attribute is based on the physical overlap between paths that are in the choice set:

$$\text{PS}_{in}^C = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{M_{an}} \quad (7)$$

where Γ_i is the set of links in path i , L_a is the length of link a and L_i the length of path i . M_{an} is the number of paths in \mathcal{C}_n using link a . That is $M_{an} = \sum_{j \in \mathcal{C}_n} \delta_{aj}$

where δ_{aj} equals one if path j contains link a and zero otherwise. Note that this original formulation of the attribute ignores all paths that are not in the choice set.

We propose a corrected version of the PS attribute, called expanded PS (EPS), where the sum representing the number of paths using a particular link involves an expansion factor that corrects for the sampling:

$$\text{EPS}_{in} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{M_{an}^{\text{EPS}}}, \quad (8)$$

and

$$M_{an}^{\text{EPS}} = \sum_{j \in \mathcal{C}_n} \delta_{aj} \Phi_{jn} \quad (9)$$

where Φ_{jn} is the expansion factor defined by

$$\Phi_{jn} = \begin{cases} 1 & \text{if } \delta_{jc} = 1 \text{ or } q(j)R_n \geq 1 \\ \frac{1}{q(j)R_n} & \text{otherwise.} \end{cases} \quad (10)$$

The definition of the expansion factor is based on the sampling protocol described in the previous section. Recall that we draw paths with replacement and add the chosen alternative with certainty. Paths are included only once in the choice set even if they are sampled several times. In the expansion factor $\Phi_{jn} = 1$ if $\delta_{jc} = 1$ represents that the chosen alternative is always included. Moreover, duplicates are ignored, that is $\Phi_{jn} = 1$ if path j is expected to be drawn more than once, $q(j)R_n \geq 1$. If path j is expected to be drawn less than once, M_{an}^{EPS} is increased ($1/(q(j)R_n) > 1$). Note that the formulation is asymptotically valid; if $R_n \rightarrow \infty$ then $q(j)R_n \geq 1 \forall j \in \mathcal{U}$ and $M_{an}^{\text{EPS}} \approx \sum_{j \in \mathcal{U}} \delta_{aj}$.

Estimation results for this formulation as well as PS based on \mathcal{C}_n and on \mathcal{U} are presented in the following section.

4 Numerical Results

The numerical results presented in this section are based on synthetic data and aim at comparing the original PS and the EPS attributes. We also include a sensitivity analysis of the estimation results with respect to the parameters of the sampling algorithm (biased random walk) and the definition of the postulated model used for generating the data.

The main advantage of using synthetic data is that the true model structure and parameter values are known. Based on such data we can 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 the 5% significance level (critical value: 1.96).

4.1 Synthetic Data

The network is shown in Figure 2 and is composed of 38 nodes and 64 links. Originally, it is a small part of a real network (Borlänge, Sweden) which has been modified so that it contains no loops. 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).

Two sets of observations are generated with two different postulated models. For each data set, we generate 3000 synthetic observations by simulation, associating a choice with the alternative having the highest utility. We use a PSL model and specify a utility function for each alternative i and observation n

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

where $\beta_{PS} = 1$, $\beta_{SB} = -0.1$ and ε_{in} are independently and identically distributed extreme value with scale 1 and location 0. The PS attribute reflects the correlation among all paths and is computed based on \mathcal{U}

$$\text{PS}_i^{\mathcal{U}} = \sum_{a \in \Gamma_i} \frac{L_a}{L_i} \frac{1}{\sum_{j \in \mathcal{U}} \delta_{aj}}. \quad (12)$$

The length attribute is used to compute the shortest path cost for the biased random walk algorithm in (6). We therefore evaluate the influence of the postulated length parameter on the estimation results by using two different values, $\beta_L = -0.3, -1$. The length attribute is more important compared to the other attributes when $\beta_L = -1$ than $\beta_L = -0.3$. Note however that the model using $\beta_L = -1$ does not correspond to a simple shortest path model. Figure 3 shows the probability of the 29 paths with the highest probabilities. The sum of these 29 probabilities is 0.978 for $\beta_L = -1$ and 0.618 for $\beta_L = -0.3$.

4.2 Model Specifications

Five different models are considered in order to evaluate the sampling correction and the different PS attribute formulations:

- PS attribute based on sampled paths only with ($M_{PS(C)}^{\text{Corr}}$) and without ($M_{PS(C)}^{\text{NoCorr}}$) sampling correction,

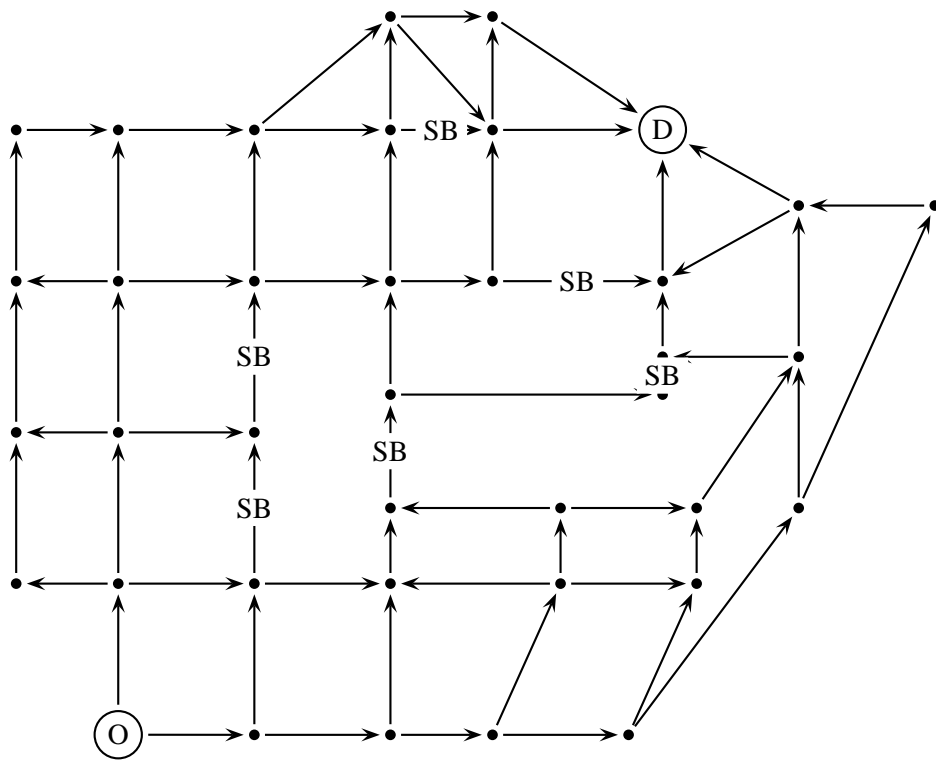


Figure 2: Example Network

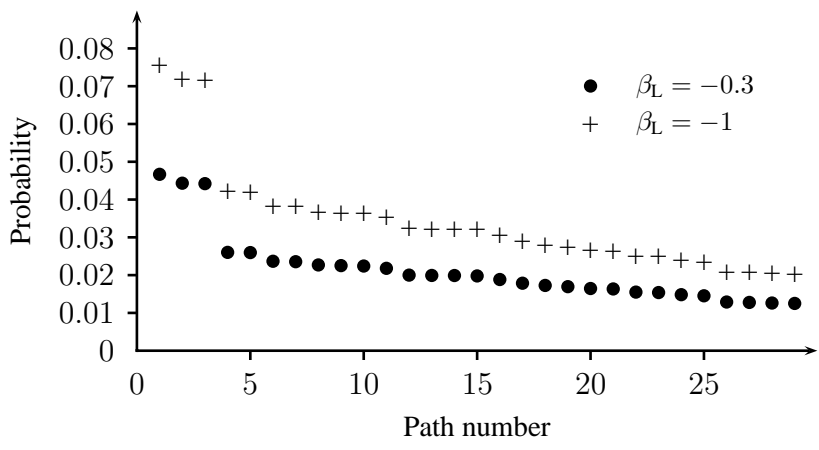


Figure 3: 29 Highest Path Probabilities for the Two Postulated Models

- PS attribute based on the universal choice set with ($M_{PS(\mathcal{U})}^{\text{Corr}}$) and without ($M_{PS(\mathcal{U})}^{\text{NoCorr}}$) sampling correction and
- EPS attribute (M_{EPS}^{Corr}) with sampling correction.

For each of these models we specify the deterministic term of the utility function as follows

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

Note that the two first specifications are based on PS formulation (7), the following two on (12) and finally the last on (8). β_{L} is fixed to the true value (-0.3 or -1 depending on the dataset) 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.

4.3 Estimation Results

In total we have estimated more than 300 models; the previously presented five models have been estimated based on the two data sets with different choice sets. For the sampling of alternatives we vary the number of draws (10, 20, 40, 80, 120, 170, 250) and the random walk parameters ($b_1 = 1, 2, 3, 5, 10, 15, 20$ with b_2 always fixed to one). The higher the value of b_1 the more the random walk is oriented towards the shortest path. In the following a “setting” refers to a combination of dataset, number of sampling draws and value of b_1 .

We start by describing detailed results for one specific setting (the $\beta_{\text{L}} = -1$ dataset using 40 draws and $b_1 = 1$) reported in Table 1. Except when explicitly stated, these interpretations can be generalized for all settings. First we note that the sampling correction is validated by the $M_{PS(\mathcal{U})}^{\text{Corr}}$ model. Except for the sampling correction term this model has the same utility specification as the postulated one, and as expected, the parameter estimates are unbiased. Furthermore, we note that when there is no sampling correction of utilities (models $M_{PS(\mathcal{U})}^{\text{NoCorr}}$ and $M_{PS(\mathcal{C})}^{\text{NoCorr}}$) the

	$M_{PS(u)}^{NoCorr}$	$M_{PS(u)}^{Corr}$	$M_{PS(c)}^{NoCorr}$	$M_{PS(c)}^{Corr}$	M_{EPS}^{Corr}
$\hat{\beta}_{PS}$	-0.108	0.969	0.285	0.397	1.09
Rob. std	0.045	0.0541	0.0785	0.074	0.052
Rob. t -test 1	-24.62	-0.57	-9.11	-8.15	1.73
$\hat{\beta}_{SB}$	-0.547	-0.0849	-0.52	-0.00941	-0.109
Rob. std	0.0322	0.0262	0.0331	0.0261	0.0281
Rob. t -test -0.1	-13.88	0.58	-12.69	3.47	-0.32
$\hat{\mu}$	1.04	0.983	1.05	0.945	1.05
Rob. std	0.0314	0.028	0.0316	0.0264	0.0314
Rob. t -test 1	1.27	-0.61	1.58	-2.08	1.59
Final L-L	-7284.711	-6966.668	-7281.035	-7160.154	-6704.515
Adj. rho bar sq.	0.291	0.322	0.292	0.303	0.348

Null log likelihood: -10283.7, 3000 observations
 $\beta_L = -1$. Algorithm parameters: 40 draws, $b_1 = 1$, $b_2 = 1$, $C(\ell) = L_\ell$
Average size of sampled choice sets: 30.92
BIOGEME (Bierlaire, 2007, and Bierlaire, 2003) has been used for all model estimations

Table 1: Detailed Estimation Results of PSL models

parameter estimates are biased. (For this setting $\hat{\beta}_{PS}$ and $\hat{\beta}_{SB}$ are biased and at least one parameter estimate is biased for all settings). The model fit is significantly better for models that are corrected for sampling than those that are not.

For the setting reported in Table 1 the $M_{PS(c)}^{Corr}$ model has biased parameter estimates while M_{EPS}^{Corr} has not. Moreover, the latter has better model fit than the former. Before analyzing these models in detail for different settings we give some statistics on the choice set sizes. Figure 4 shows the average number of sampled paths as a function of number of draws when $b_1 = 1, 2, 3$. Recall that the higher the value of b_1 the more the random walk is oriented toward the shortest path. Hence, it is expected that the number of sampled paths decrease as b_1 increase. Moreover, the same paths can be drawn several times and this is why we see an attenuation effect as the number of draws increase, this effect is of course more important the higher the value of b_1 .

Figure 5 shows the absolute value of the t -test statistics as a function of number of draws for models M_{EPS}^{Corr} and $M_{PS(c)}^{Corr}$ estimated based on the $\beta_L = -1$ dataset. At least one parameter estimate is biased for both models when $b_1 > 3$ and we therefore only report results for $b_1 = 1, 2, 3$. The results of the M_{EPS}^{Corr} model improves as the number of draws increase. All parameter estimates are unbiased from 40 draws when $b_1 = 1$. The choice sets are larger the lower the value of b_1

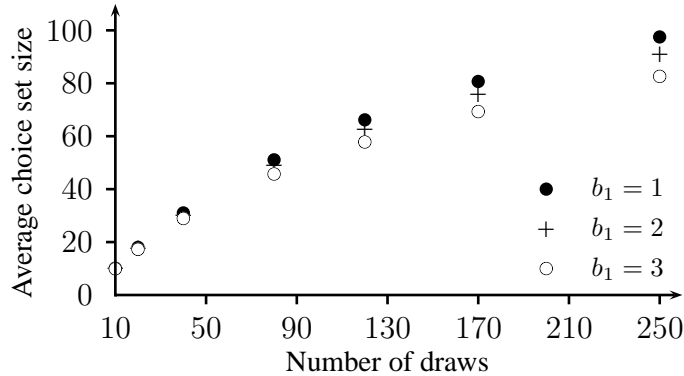


Figure 4: Average number of paths in choice sets

which explains the better results. The $M_{PS(C)}^{\text{Corr}}$ model has on the other hand at least one biased estimate for all settings. Moreover, note from $\hat{\beta}_{\text{SB}}$ that the estimate first rapidly worsen before slightly improving as the number of draws increase.

The same results are presented in Figure 6 for models estimated based on the $\beta_L = -0.3$ dataset. Also for this dataset the results of $M_{\text{EPS}}^{\text{Corr}}$ improve as the number of draws increase. As expected, a higher number of draws is needed in order to obtain unbiased parameter estimates (170 draws). Indeed, the length attribute has lower weight in this dataset and the accuracy of the PS attribute is hence more important. The estimates in the $M_{PS(C)}^{\text{Corr}}$ model do not converge to the true values and remain biased even for a high number of draws.

The results presented in this section clearly show the importance of correcting the utilities for sampling. The correction is robust both with respect to algorithm parameters and the definition of the postulated model. Unbiased estimates are obtained even with a low number of sampling draws. Furthermore, the PS attribute should reflect the correlation among all paths in \mathcal{U} and the results indicate that the original PS formulation (7) is not appropriate. On the other hand, the EPS attribute shows good results; the estimates converge rather rapidly toward the true values. However the formulation is valid only asymptotically and it is therefore important not to have too few paths in the choice sets, like for any results based on samples. Low values of b_1 and high number of draws give therefore the best results.

5 Conclusions and Future Work

This paper presents a new formulation of the path size attribute, called expanded path size, that can be used in route choice models that are corrected for sampling. The expanded path size attribute is defined consistently with the sampling protocol proposed by Frejinger and Bierlaire (2007) (Frejinger, 2008).

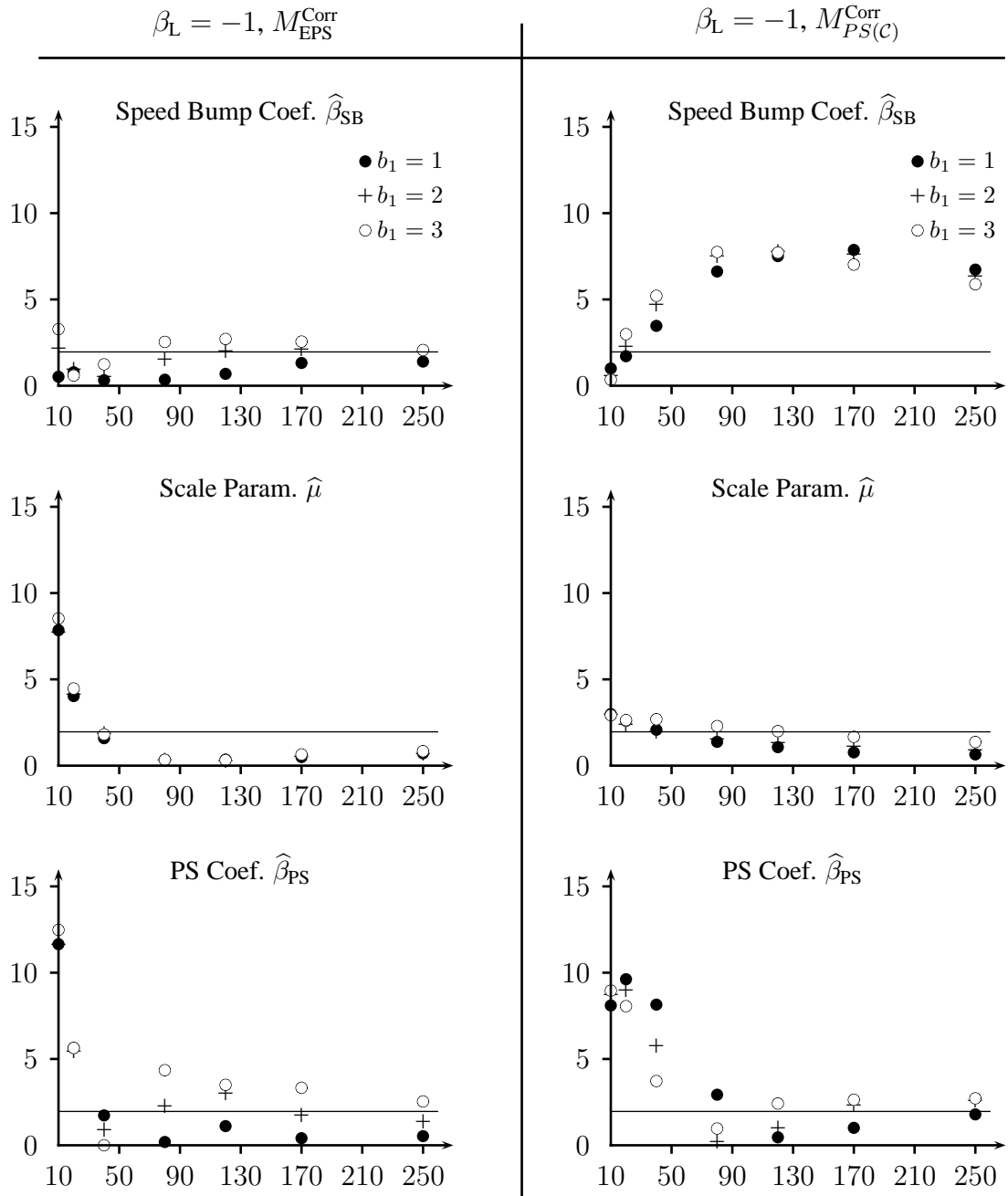


Figure 5: t -test values for M_{EPS}^{Corr} and $M_{PS(C)}^{Corr}$, x-axis: number of draws, y-axis: absolute value of t -test w.r.t. true value

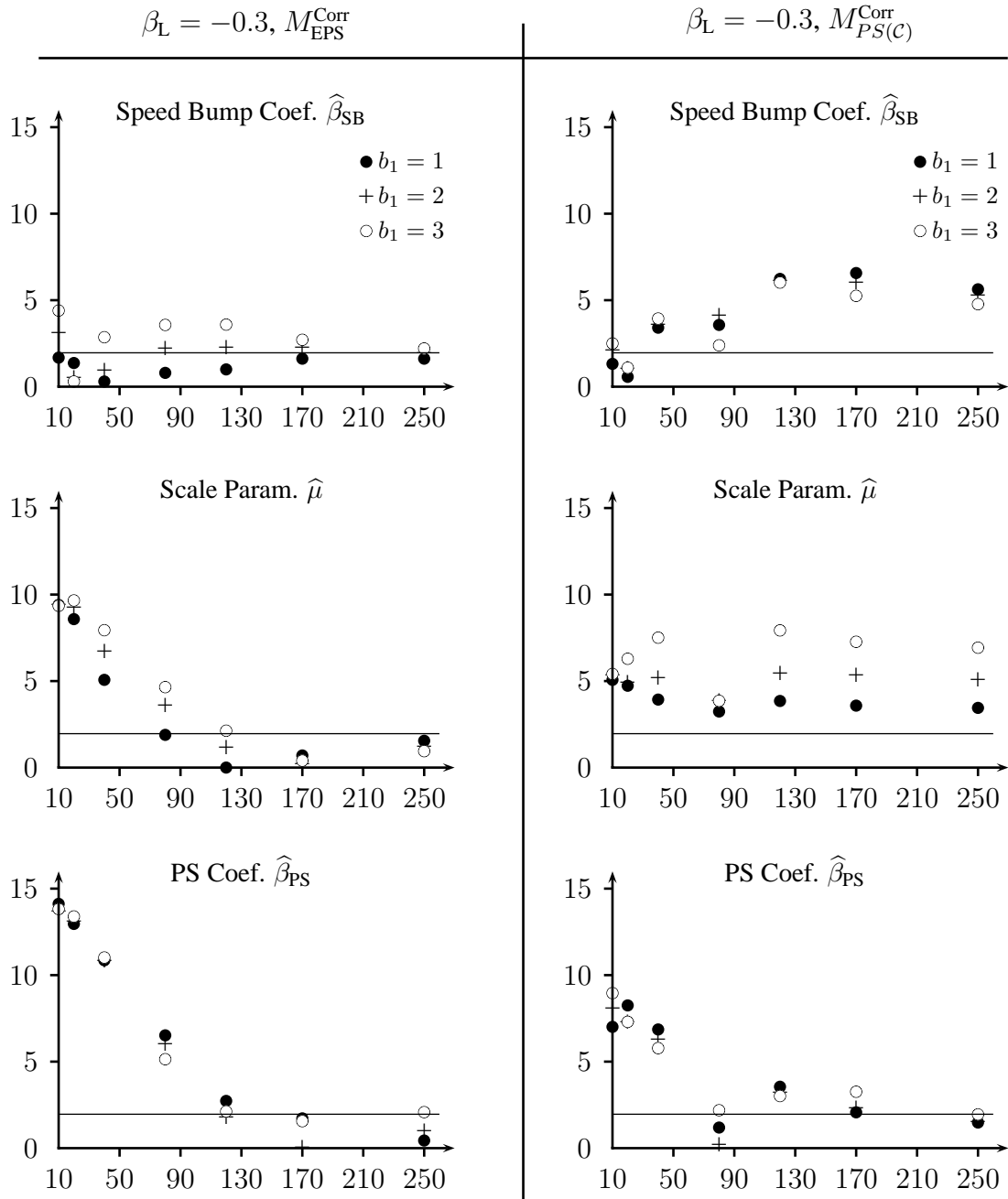


Figure 6: t -test values for M_{EPS}^{Corr} and $M_{PS(C)}^{Corr}$, x-axis: number of draws, y-axis: absolute value of t -test w.r.t. true value

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. Unbiased parameter estimates can be obtained with the expanded path size logit model and it is remarkably better than models with the original path size formulation.

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. In such a setting it would also be interesting to compare results of different choice set generation algorithms, with and without correction for sampling. Moreover, future research should be dedicated to sampling of alternatives for prediction.

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., Ben-Akiva, M. and Ramming, M. (2002). Adaptation of logit kernel to route choice situation, *Transportation Research Record* **1805**: 78–85.
- 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., 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.
- 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. (2008). The estimation of Generalized Extreme Value models from choice-based samples, *Transportation Research Part B: Methodological* **42**(4): 381–394.
- 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: Emerging Technologies* **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: Methodological* **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.
- Fiorenzo-Catalano, S. (2007). *Choice Set Generation in Multi-modal Transportation Networks*, PhD thesis, Delft University of Technology.

- Frejinger, E. and Bierlaire, M. (2007). Sampling of alternatives for route choice modeling, *Technical Report TRANSP-OR 071121*, Transport and Mobility Laboratory, ENAC, EPFL.
- Frejinger, E. (2008). *Route choice analysis: data, models, algorithms and applications*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland.
- Frejinger, E. and Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models, *Transportation Research Part B: Methodological* **41**(3): 363–378.
- Friedrich, M., Hofsäuss, 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, Delft University of Technology.
- 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.
- 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: Methodological* **21**(2): 91–102.

- 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.
- Vovsha, P. and Bekhor, S. (1998). Link-nested logit model of route choice Overcoming route overlapping problem, *Transportation Research Record* **1645**: 133–142.
- Yai, T., Iwakura, S. and Morichi, S. (1997). Multinomial probit with structured covariance for route choice behavior, *Transportation Research Part B: Methodological* **31**(3): 195–207.