Specification of the cross nested logit model with sampling of alternatives for route choice models

Xinjun Lai [*] Michel Bierlaire [†]

June 2, 2014

Report TRANSP-OR 140602
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*] Sun Yat-sen University, School of Engineering, China, laixinj@mail2.sysu.edu.cn

[†] École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, michel.bierlaire@epfl.ch

Abstract

A novel approach is proposed to use the Cross Nested Logit (CNL) model in route choice when sampling of paths is considered. It adopts the Metropolis-Hasting algorithm to sample the choice sets for the model. A new expansion factor and an approximation method are put forward to calculate the sampled probabilities of alternatives. We build on state-of-the-art results for the Multivariate Extreme Value models and extend then to the route choice context. Case studies on both synthetic data and a real network demonstrate that the new method is valid and practical. This paper thus provides an operational solution to use the CNL model in the route choice context, where the number of alternatives is particularly large.

1 Introduction

A route choice model predicts the probability that one traveler chooses a certain route between a given origin and destination (OD). It is one of the most challenging problems in the transportation research because of its large choice set and its highly correlated structure. In order to address the correlation issue in route choice, the Path Size and C-Logit models are the most frequently used methods due to their simplicity. They both add an additional term to the utility to compensate for the correlation of paths (Cascetta et al., 1996; Ben-Akiva and Bierlaire, 1999). However, the additional terms are convenient approximations and previous research shows that they might be too sensitive to the composition of the choice set (Bovy et al., 2008; Jun et al., 2013). The cross nested logit model explicitly captures the correlation among paths, where alternatives are classified into nests if they share common unobserved attributes (Bierlaire, 2006; Wen and Koppelman, 2001). The link-based Cross Nested Logit (CNL) model has been proposed by Vovsha and Bekhor (1998). It assumes that each link corresponds to a nest and the paths that share the same link belongs to the same nest (Prashker and Bekhor, 1999; Bekhor and Prashker, 2001; Ramming, 2002). However it is difficult to estimate because of the large number of parameters. Therefore, it is common to empirically set their values based on the network topology. The Error Component (Ramming, 2002; Frejinger and Bierlaire, 2007) model is also used in the route choice context. It is a mixture logit model and requires simulation-based estimation, which is cumbersome for large-scale applications.

The models described above require the full choice set which is unfeasible in route choice. In a road network with 80 nodes and 0.4 density, between any OD pairs there are approximately 10^{84} possible routes without loop (Roberts and Kroese, 2007). Such a choice set is impractically large justifying the need to sample a subset of paths with a reasonable size. Frejinger and Bierlaire (2010) analyze several existing algorithms for the generation of the paths. The most desirable features in this framework is firstly the chosen alternative should be included, and secondly the algorithm is able to generate the sampled probabilities that are required by the correction formula. An algorithm named Random Walk is proposed to first address the issue of paths sampling (Frejinger et al., 2009). However this approach cannot avoid loops in the sampled paths. Flötteröd and Bierlaire (2013) present an algorithm based on the Metropolis-Hasting simulation to sample paths between an origin and a destination with pre-specified sampling probabilities. Besides, the algorithm can avoid cyclic paths.

Regarding the estimation of model with sampled alternatives, McFadden, 1978 demonstrated that if the underlying model is Logit, a correction term should be added to the log-likelihood function to achieve consistent parameters estimation with sampled alternatives. There are several applications adopting McFadden's result (Parsons and Kealy, 1992; Sermons and Koppelman, 2001). Bierlaire et al. (2008) have proposed a consistent estimator for the MEV (Multivariate Extreme Value) models with samples of alternatives, which has been made operational by Guevara and Ben-Akiva (2013). They argue that the correlation is represented by the full choice set of the nests, so a sampled subset is also required and it should be different from the first set that used in McFadden's formula. An expansion factor is also provided to compensate the loss of the non-sampled paths of the nests to achieve unbiased estimations. This general method is applied to a real data case of residential location choice as a Nested Logit model. Papola (2004) proposes a conjecture for the variance-covariance matrix of the CNL model, and with the normalization and relationships, the nesting parameters can be determined by solving a system of equations. Another research shows that this conjecture over-estimates the correlation (Abbe et al., 2007). However with the complicated structure of the road network, solving a large system of equations is a heavy burden, besides it does not provide much information on the travelers' behavior. The objective of this paper is to estimate a Cross Nested Logit model of route choice with sampled paths. We adopt the recent methodologies in route choice and the MEV models. The general framework proposed by Guevara and Ben-Akiva (2013) is extended to the route choice context. The Metropolis-Hasting (MH) algorithm is applied because of its advantage in generating non-cyclic paths and its ability to generate paths from an arbitrary distribution. In particular, a new expansion factor is proposed to avoid path enumeration. An approximation method is also presented to calculate the sampled probabilities, which are required in the model but not explicitly provided by the MH algorithm. The validation of the new method is based on synthetic data and the estimated parameters are compared with their true values. Moreover, it is also applied on the real data. All the results show that the new method is feasible and practical, and it is suitable for large network analysis.

The paper is organized as follows. Section 2 describes the methodology, including the sampling of alternatives, the approximation of the cross nested logit model with the sampled paths, and the expansion factors. The validation on synthetic data is presented in section 3. The new method is applied on real data in section 4. Finally conclusions and discussions for future study are given in section 5.

2 Methodology

The CNL route choice model derived in this paper has a similar structure as proposed by Vovsha and Bekhor (1998). Each link of the network is associated with a nest, and each path (that is each alternative) belongs to some extent to each nest corresponding to the link composing the path. A path belongs to several nests, capturing the correlation among alternatives due to the network topology. If there are M links in the network, the probability that one traveler chooses path i from the universal choice set $\mathcal{C}$ is

$$P(i|\mathcal{C}) = \frac{\exp(V_i + \ln G_i(\mathcal{C}))}{\sum_{j\in\mathcal{C}} \exp(V_j + \ln G_j(\mathcal{C}))} \tag{1}$$

where

$$G_i(\mathcal{C}) = \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left(\sum_{j\in\mathcal{C}} \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu-\mu_m}{\mu_m}} \right], \tag{2}$$

where V_i is the deterministic utility for path i, $\mu > 0$ is the scale parameter for the model, μ_m is the scale parameter for nest m, such that $\mu_m \geq \mu$, α_{im} is the inclusive parameter capturing the membership of path i to nest m, G is the MEV generating function, and G_i is its partial derivative with respect to $\exp(V_i)$. The notation $G_i(\mathcal{C})$ emphasized that the sum involved in (2) spans the entire choice set.

In order to make the model operational, we consider now the sampling of alternatives. The idea is to approximate the choice probability by replacing the full choice set $\mathcal{C}$ by a subset $\mathcal{D}$, with the positive conditioning property, that is

$$\Pr(\mathcal{D}|j) > 0, \ \forall j \in \mathcal{D}. \tag{3}$$

In this case, Bierlaire et al. (2008) have shown that the choice probability conditional on the sampled set is given by

$$\Pr(i|\mathcal{D}) = \frac{\exp(V_i + \ln G_i(\mathcal{C}) + \ln \Pr(\mathcal{D}|i))}{\sum_{j\in\mathcal{D}} \exp(V_j + \ln G_j(\mathcal{C}) + \ln \Pr(\mathcal{D}|j))}. \tag{4}$$

The quantity $\ln \Pr(\mathcal{D}|i)$ is the alternative specific correction term to accommodate for errors due to sampling of alternatives.

Eq. (2) indicates that the calculation of G_i requires the summation over the full choice set $\mathcal{C}$, which is not feasible in general. An approximation of G_i based on a sampled set is proposed in this section, building on the work by Guevara and Ben-Akiva (2013). They point out that it is essential to generate two samples of alternatives, one for the denominator of (1) (we denote it by $\mathcal{D}$, as before) and another one ($\mathcal{D}'$) for the approximation of G_i. Note that it is required that $\mathcal{D}$ contains the alternative for which we want to compute the choice probability. It is not required for $\mathcal{D}'$.

We propose to use the Metropolis-Hasting (MH) sampling algorithm (Flötteröd and Bierlaire, 2013) due to its capability to sample a path i with a sampling probability

$$q(i) = b(i)/\sum_{j \in \mathcal{C}} b(j) = b(i)/B, \tag{5}$$

proportional to a "weight" $b(i)$, without involving the normalization term B (which would require a sum over the full choice set $\mathcal{C}$). The input of the algorithm requires the definition of the path's weight and the road network.

Frejinger et al. (2009) provide the formula of the correction term $\ln \Pr(\mathcal{D}|i)$ if the sampling protocol is importance sampling and paths are drawn independently with replacement from the universal choice set $\mathcal{C}$:

$$\Pr(\mathcal{D}|i) = K_{\mathcal{D}} \frac{k_i}{q(i)} \tag{6}$$

where $K_{\mathcal{D}}$ is a proportionality factor independent from i. Therefore, (4) can be written

$$\begin{aligned}
\Pr(i|\mathcal{D}) &= \frac{\exp(V_i + \ln G_i(\mathcal{C}) + \ln K_D + \ln k_i - \ln b(i) + \ln B)}{\sum_{j \in \mathcal{D}} \exp(V_j + \ln G_j(\mathcal{C}) + \ln K_D + \ln k_j - \ln b(j) + \ln B)} \\
&= \frac{\exp(V_i + \ln G_i(\mathcal{C}) + \ln k_i - \ln b(i))}{\sum_{j \in \mathcal{D}} \exp(V_j + \ln G_j(\mathcal{C}) + \ln k_j - \ln b(j))} \\
&= \frac{\exp(V_i + \ln G_i(\mathcal{C}) + \ln \frac{k_i}{b(i)})}{\sum_{j \in \mathcal{D}} \exp(V_j + \ln G_j(\mathcal{C}) + \ln \frac{k_j}{b(j)})},
\end{aligned} \tag{7}$$

as $K_{\mathcal{D}}$ and B are independent from i.

We adopt the approximation of G_i based on the sampled set $\mathcal{D}'$ as proposed by Guevara and Ben-Akiva (2013):

$$G_i(\mathcal{C}) \approx \widehat{G}_i(D', w) = \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left(\sum_{j \in \mathcal{D}'} w_j \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu - \mu_m}{\mu_m}} \right], \tag{8}$$

where w_j is an expansion factor to compensate the loss of the non-sampled paths of the nests. We consider three definitions of the expansion factor.

1. Guevara and Ben-Akiva (2013) propose

$$w_j^G = \frac{k_j}{E[k_j]} = \frac{k_j}{q(j)R} = \frac{k_j B}{b(j)R}, \tag{9}$$

where R is the number of draws used to generate $\mathcal{D}'$. Note that this definition requires $B = \sum_{j \in \mathcal{C}} b(j)$, which cannot be computed due to the need to enumerate all the paths. It is therefore approximated. We have

$$B = \sum_{j \in \mathcal{C}} b(j) = |\mathcal{C}| \frac{\sum_{j \in \mathcal{C}} b(j)}{|\mathcal{C}|} = |\mathcal{C}|\bar{b}, \tag{10}$$

where $|\mathcal{C}|$ is the number of paths in the full choice set, and $\bar{b}$ is the average weight of paths in the full choice set, which is approximated by the average weight of paths in $\mathcal{D}'$, that is

$$\bar{b} = \frac{\sum_{i \in \mathcal{C}} b(i)}{|\mathcal{C}|} \approx \frac{\sum_{i \in \mathcal{D}} b(i)}{|\mathcal{D}|}. \tag{11}$$

For an estimate of $|\mathcal{C}|$, we adopt the simulation method proposed by Roberts and Kroese (2007), summarized in Appendix A.

2. Frejinger et al. (2009) propose a similar expansion factor for the approximation of the Path Size Logit model, that we denote by w_j^F:

$$w_j^F = \begin{cases} 1 & \text{if } b(j)R > B, \\ \frac{B}{b(j)R} & \text{otherwise.} \end{cases} \tag{12}$$

Again, an approximation of B is required here.

3. We propose another definition of the expansion factor, that does not require B. Denote by s the path that has been sampled the most during the generation of $\mathcal{D}'$, that is $k_s \geq k_i$, for each $i \in \mathcal{D}'$. If the sample is large enough, this number is close to the expected value of the sampling protocol, that is $k_s \approx q(s)R$. Therefore, we approximate the expansion factor by Guevara and Ben-Akiva (2013) as follows:

$$w_j^G = \frac{k_j}{q(j)R} \approx \frac{k_j}{q(j)R}\frac{q(s)R}{k_s} = \frac{k_j}{b(j)}\frac{b(s)}{k_s}, \tag{13}$$

and we define the new expansion factor

$$w_j^L = \frac{k_j}{k_s}\frac{b(s)}{b(j)}. \tag{14}$$

Based on this approximation, we obtain the choice model

$$\Pr(i|\mathcal{D},\mathcal{D}',w) = \frac{\exp(V_i + \ln G_i(\mathcal{D}',w)) + \ln \frac{k_i}{b(i)})}{\sum_{j\in\mathcal{D}} \exp(V_j + \ln G_j(\mathcal{D}',w) + \ln \frac{k_j}{b(j)})}. \tag{15}$$

3 Validation with synthetic data

The validation presented here uses synthetic data, so that the true values of the parameters are known to the researcher beforehand, and the quality of the estimated parameters can be assessed. The following features of the proposed methodology are investigated here:

1. the sampling correction;

2. the expansion factors;

3. the configuration of the MH algorithm.

3.1 Synthetic data

The road network shown in Figure 1 is a simplified network inspired from the topology of the city of Borlange in Sweden (Frejinger, 2008). There is only one OD pair from node O to D and the size of the full choice set is $|C| = 170$. The lengths of links are proportional to their lengths on the figure and some links have speed bumps (denoted by SB). The deterministic utility of path i is defined as

$$V_i = \beta_L L_i + \beta_{SB} SB_i, \tag{16}$$

where L_i is the length of path i and SB_i is the number of speed bumps along path i. The "true" value for the parameters are defined as $\beta_L = -0.5$ and $\beta_{SB} = -0.1$. The underlying true model is the link-based CNL model. The inclusive parameter is defined as $\alpha_{im} = \ell_m / L_i$, where ℓ_m is the length of link m and L_i the length of path i. The scale parameters are assumed to be the same for each nest/link: $\mu_m = 1.5$. 3000 synthetic observations of the chosen paths are generated by the simulation of this CNL model. To evaluate the quality of the synthetic data, the observations are used back to estimate the parameters of the CNL model with the full choice set (it is feasible here as the choice set contains only 170 alternatives). Three parameters are to be estimated: β_L, β_{SB} and μ_m. The results and the statistical tests are shown in Table 1. The fourth column is the t-test against zero, and the last column is the test against the true value of the parameter. The results indicate that the estimated parameters are not significantly different from their true value, validating the quality of the synthetic data.

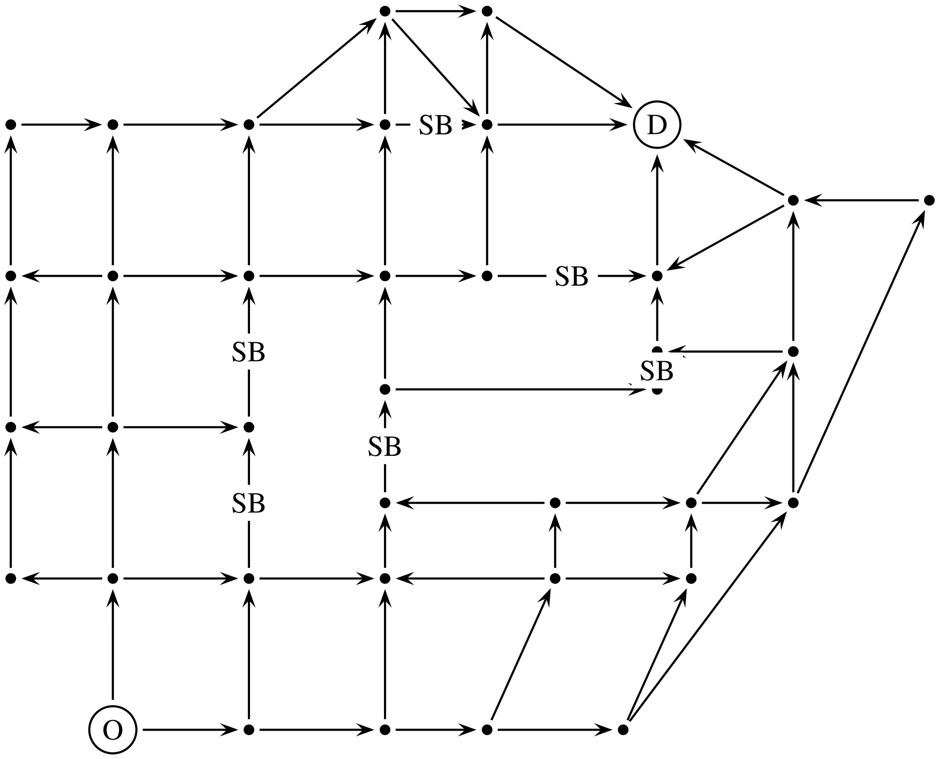

Figure 1: Network with 170 paths from O to D

Parameters	Estimated value	Standard error	t-test (0)	t-test (true value)
β_L	-0.501	0.0118	43.1	0.678
β_{SB}	-0.0910	0.0240	3.19	0.375
μ_m	1.49	0.0269	55.2	0.0535

Table 1: Estimations for the CNL model with the full choice set and synthetic data

Now, we apply the methodology proposed in Section 2. Paths are sampled using the MH algorithm with weights

$$b(i) = \exp(-\theta L_i), \tag{17}$$

where θ is non-negative input scale parameter for the algorithm. The idea is that shorter paths are more likely to be sampled. Large values of θ penalize more the longer paths. A value of $\theta = 0$ assigns equal sampling probability to each path. Figure 2 shows the number of different paths in the sampled choice set with different values for θ. The x-axis is the number of draws and the y-axis is the number of generated paths. Note that if a path is generated several times, it is counted only once.

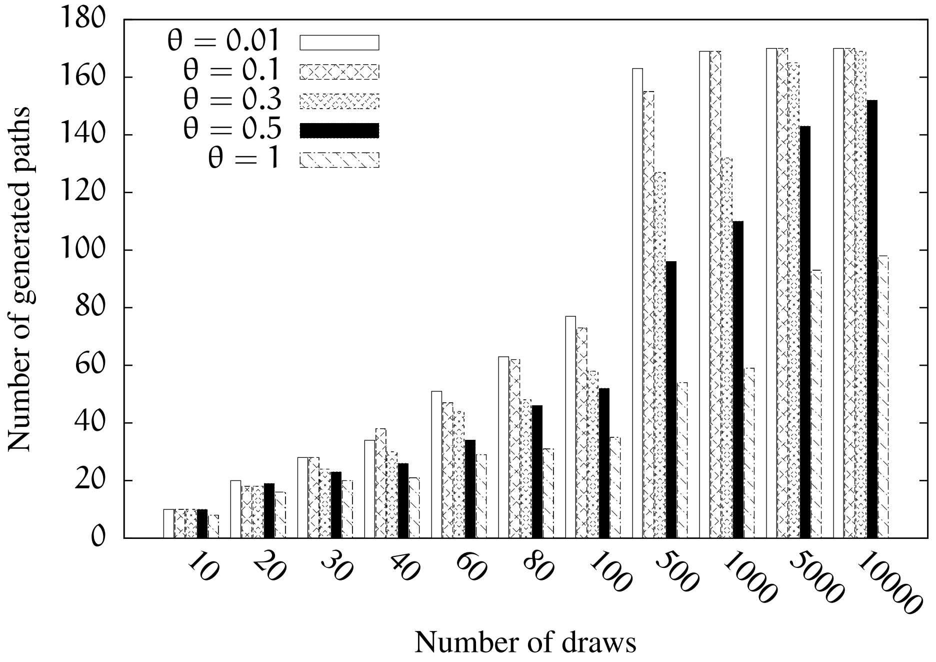

Figure 2: Size of the choice set generated by the MH algorithms for various values of θ

In addition to models (7) and (15), we will also test versions of (15) without the sampling correction nor the expansion factor, that is

$$\Pr(i|\mathcal{D}, \mathcal{D}') = \frac{\exp(V_i + \ln G_i(\mathcal{D}', 1))}{\sum_{j \in \mathcal{D}} \exp(V_j + \ln G_j(\mathcal{D}', 1))}, \tag{18}$$

Model (7) is estimated firstly because it requires only one sampled set. This model allow us to analyze the impact of the configuration of the MH algorithm

Sampling protocol: $\theta = 0.5$				
10 draws	Estimated value	Standard error	t-test(0)	t-test(true value)
β_L (-0.5)	-0.443	0.0163	27.3	3..48
β_{SB} (-0.1)	-0.0647	0.0427	1.51	0.826
μ_m (1.5)	1.56	0.0340	45.8	1.72
Estimation time: 1362 seconds				
40 draws	Estimated value	Standard error	t-test(0)	t-test(true value)
β_L (-0.5)	-0.479	0.0156	30.8	1.34
β_{SB} (-0.1)	-0.0720	0.0393	1.83	0.713
μ_m (1.5)	1.51	0.0322	47.0	0.367
Estimation time: 4648 seconds				
Sampling protocol: $\theta = 0.01$				
10 draws	Estimated value	Standard error	t-test(0)	t-test(true value)
β_L (-0.5)	-0.535	0.0174	30.8	2.01
β_{SB} (-0.1)	-0.132	0.0545	2.42	0.580
μ_m (1.5)	1.41	0.0355	39.8	2.47
Estimation time: 1612 seconds				
40 draws	Estimated value	Standard error	t-test(0)	t-test(true value)
β_L (-0.5)	-0.544	0.0160	33.9	2.76
β_{SB} (-0.1)	-0.130	0.0410	3.16	0.726
μ_m (1.5)	1.41	0.0322	43.8	2.85
Estimation time: 4914 seconds				

Table 2: Estimation results for model (7)

on the model estimation. Secondly, (15) is estimated to evaluate the impact of different expansion factors. It also aims to investigate the role of $\mathcal{D}'$. Finally, (18) is estimated for the comparison with model (15). The scale parameter μ is normalized to 1 for all the cases, so all the estimations are in the same scale to facilitate comparisons.

3.2 Model with the true G_i

We first estimate model (7). We have sampled path using two values for the parameter of the MH algorithm: $\theta = 0.5$ and $\theta = 0.01$. Note that the first one happens to be the same as the true value of the parameter β_L. The second is smaller so as to generate more paths. Estimations are shown in Table 2.

Each θ is evaluated twice, with choice sets generated with 10 and 40 draws. The values between parentheses in the first column are the true values of the corresponding parameter. Results show that the estimation time increases as the number of draws becomes larger (as expected), but no significant difference is observed between the two θ's. When there are 40 draws and $\theta = 0.5$, approximately 25 paths are generated, and we obtain good estimates in the sense that all the values of t-test(true value) are under the critical value 1.96, so that the hypothesis that the parameter is equal to its true value is not rejected by the test. When $\theta = 0.01$, there are approximately 10 and 38 paths in the choice set corresponding to 10 and 40 draws respectively. However we do not obtain good estimates in this case. All the t-tests w.r.t. the true value are larger than 1.96. These results suggest that for the first choice set, it is suitable to use the importance sampling protocol where shorter paths are preferred to longer ones.

3.3 Model with an approximated G_i

We now present estimation results for models (15) and (18). The objective is to analyze the joint impact of sampling correction and expansion factor for the model. To analyze the role of the choice set $\mathcal{D}'$ used for the approximation of G_i, we vary the number of draws, the parameter θ and also the expansion factors w. For the expansion factors w^G (defined by (9)) and w^F (defined by (12)), the value of B is approximated by the simulation procedure described above.

Firstly $\bar{b}$ is approximated by (11), where the set $\mathcal{D}$ is generated with the MH algorithm with 100 draws and $\theta = 0.01$. We repeat the procedure 100 times to obtain 100 estimates of $\bar{b}$, and report the average and standard error in Table 3.

	True value	Mean	Standard error	t-test(true value)
$\bar{b}$	0.688	0.684	0.0023	1.62
$\lvert\mathcal{C}\rvert$	170	169.8	2.52	0.0722

Table 3: Normalization factor B: estimations and t-tests w.r.t. the true values

The second step is to approximate $|\mathcal{C}|$. 10000 paths are drawn with the procedure described in Appendix A. The algorithm is repeated 100 times to obtain 100 approximations of $|\mathcal{C}|$. The means and standard errors are shown in Table 3. Thanks to the relatively small size of the example, the true values of b and $|\mathcal{C}|$ are already known, so the t-tests w.r.t. the true values are also presented. Results indicate that we can obtain good estimates.

Table 4 shows the t-tests w.r.t. the true values for both model (15) and (18). The choice set $\mathcal{D}$ is generated with the MH algorithm with 40 draws and $\theta = 0.5$. For $\mathcal{D}'$ used to approximate G_i, we test $\theta = 0.5$ and $\theta = 0.01$ as before, and the number of draws increases from 100 to 300. The results for different expansion factors used in the calculation of G_i are also compared, including $w = 1$, which corresponds to no expansion factor.

Using $\theta = 0.5$ and 100 draws, the estimation with w^L already has good results (in the sense that the t-test against the true value is less than 1.96), while the other two expansion factors are unable to obtain good estimates for two parameters out of three. When the number of draws increases to 200, w^G allows to obtain good estimates, while w^F does not. As for 300 draws, w^G obtains good estimates for only two parameters out of three, while w^L recovers the three parameters to the desired precision. Regarding the case with w^F and 300 draws, the t-tests for β_L and μ_m are still larger than 1.96 which suggest that they cannot be estimated with sufficient precision. As for $w = 1$, increasing the number of draws shows an improvement in estimation, but at least one parameter has poor estimate. Model (18) with no sampling correction or expansion factor behaves the worst, as expected.

When $\theta = 0.01$ and 300 draws are used, all the expansion factors (except $w = 1$) produce good estimates for all the parameters. With less draws, only parameter β_{SB} is correctly estimated. As for $w = 1$, increasing the number of draws does not help to obtain good estimates, which justifies the need for an expansion factor.

When a small number of draws is used (100 or 200), the results with a $\theta = 0.01$ are worse than for $\theta = 0.5$. The sample is more likely to include long and circuitous paths, which contribute less to a good estimate of the correlation structure of relevant paths.

The estimation time for models (15) and (18) are plotted in Figure 3.

It shows that more time is needed for $\theta = 0.01$ compared to $\theta = 0.5$. Indeed, it generates more distinct paths in the choice set for the same number of draws.

Since the expansion factor w^L generates good estimates with $\theta = 0.5$, starting from 100 draws, these experiments suggest that it is the most suitable for the approximation of G_i. Figure 4 shows the t-tests against the true values of model (15) with w^L and $\theta = 0.5$ for different number of draws. The number of generated paths is indicated above the curve. The estimates of β_{SB} and μ_m are already good with 40 draws, while good estimates of β_L are obtained starting at 100 draws.

4 Application to real data

In order to evaluate the performances of the proposed method and different expansion factors, this section applies the new model on real data. A case study of taxi drivers choosing routes in the city center is presented. The studied city, Guangzhou, is situated in the southern China and has approximately ten million inhabitants. Only the central business district (CBD), the Tianhe region is considered.

The data set for the estimation are GPS traces from taxis carrying passengers.

Sampling protocol for $\mathcal{D}'$: $\theta = 0.5$															
	100 draws					200 draws					300 draws				
	Mod. (15)				Mod. (18)	Mod. (15)				Mod. (18)	Mod. (15)				Mod. (18)
	w^{G}	w^{F}	w^{L}	$w = 1$		w^{G}	w^{F}	w^{L}	$w = 1$		w^{G}	w^{F}	w^{L}	$w = 1$	
β_{L}	2.48	4.34	**1.25**	3.59	19.4	**0.578**	10.5	**0.0374**	3.38	18.9	**0.981**	3.62	**0.703**	**0.981**	19.3
β_{SB}	**0.910**	**0.867**	**0.722**	**0.179**	**0.221**	**0.513**	**0.194**	**0.440**	**0.259**	**0.269**	**0.428**	**1.34**	**0.537**	**0.428**	**0.0052**
μ_m	2.02	3.09	**0.437**	2.98	**1.06**	**1.36**	5.02	**1.34**	3.07	**0.965**	2.28	3.12	**1.70**	2.28	**1.66**

Sampling protocol for $\mathcal{D}'$: $\theta = 0.01$															
	100 draws					200 draws					300 draws				
	Mod. (15)				Mod. (18)	Mod. (15)				Mod. (18)	Mod. (15)				Mod. (18)
	w^{G}	w^{F}	w^{L}	$w = 1$		w^{G}	w^{F}	w^{L}	$w = 1$		w^{G}	w^{F}	w^{L}	$w = 1$	
β_{L}	4.61	4.23	4.48	4.30	18.9	3.51	3.84	2.86	4.37	18.5	**0.809**	**0.0271**	**1.02**	5.05	18.5
β_{SB}	**0.303**	**0.297**	**0.254**	**0.467**	**0.634**	**0.173**	**0.119**	**0.298**	**0.409**	**0.571**	**0.565**	**0.780**	**0.480**	**0.564**	**0.654**
μ_m	4.70	4.71	5.38	4.55	3.63	9.11	8.65	7.19	5.41	3.72	**1.66**	**0.650**	**1.84**	5.19	3.01

Table 4: Estimations with different expansion factors: t-test against true value ($\mathcal{D}$ is constructed with 40 draws and $\theta = 0.5$)

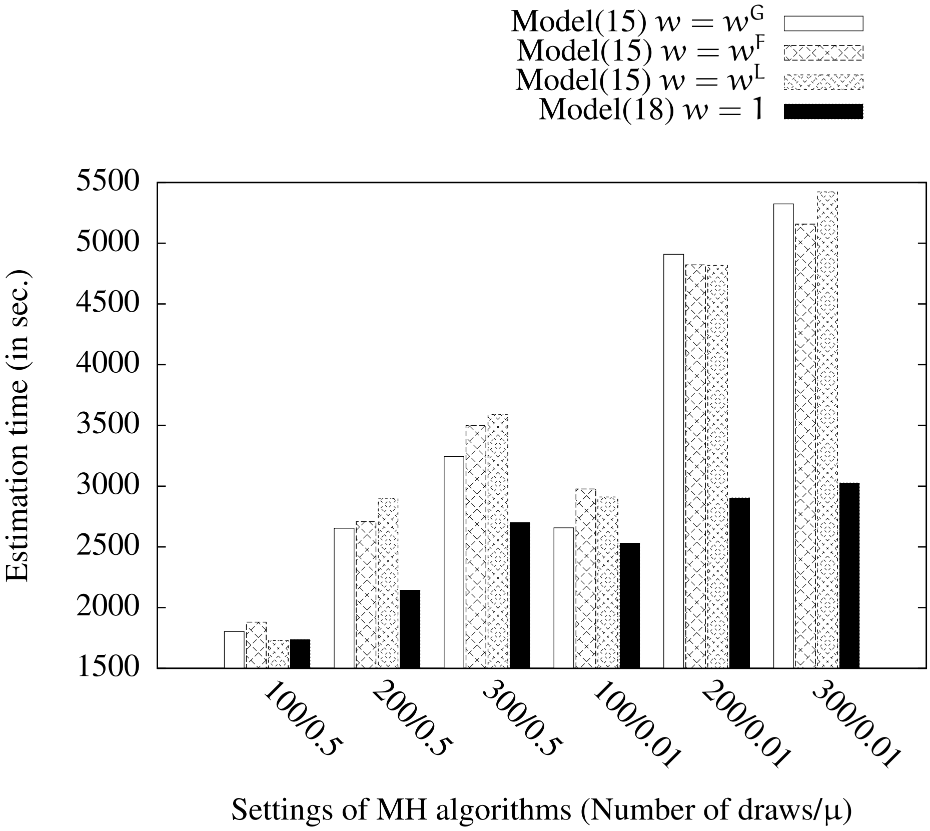

Figure 3: Estimation time of model (15) and (18) with different configurations

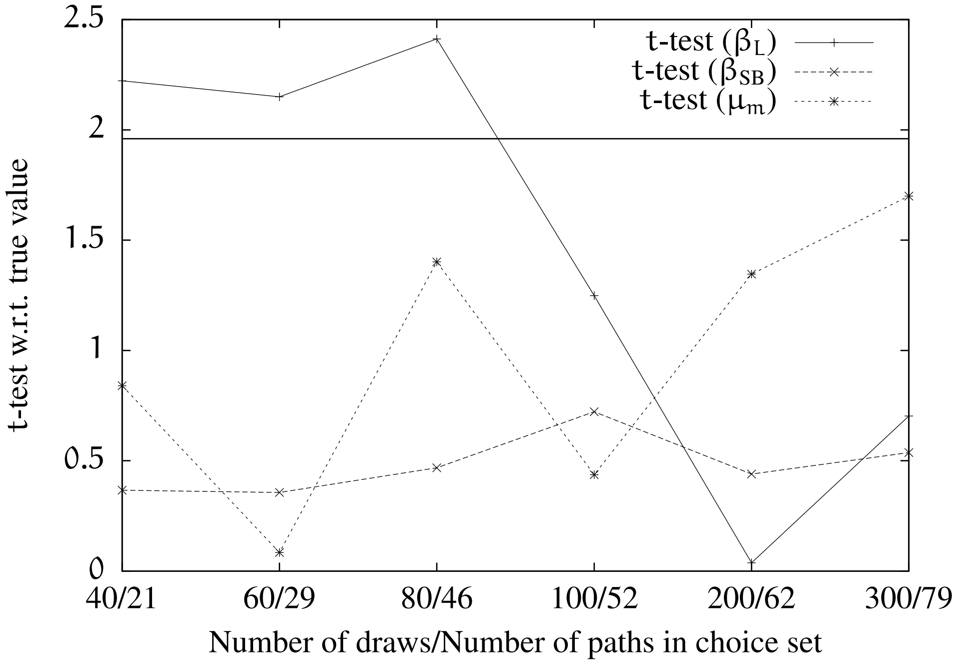

Figure 4: t-tests against the true values of model (15) with $w = w^L$ and $\theta = 0.5$

The data was collected by a management company for monitoring purpose but not for navigation, so the route choice behavior is based on the drivers' own judgment. The vehicles were monitored within a radius of 25 km in the CBD, and 740 trips from seven ODs are collected for case study. The network contains 208 nodes, 662 unidirectional links, 24 major roads, 34 arterial streets, 32 minor streets and 57 signal-controlled intersections.

Attributes	Min	Average	Max
Length (km)	0.447	2.84	8.342
Artery road ratio	0	0.797	1
Number of signal-controlled intersections	0	3.66	14

Table 5: Statistics on routes attributes

Three attributes are chosen for the utility function:

- length (in kilometers),
- the artery road ratio, that is the length of the artery road (major roads and arterial streets) divided by the total length of the trip, and,

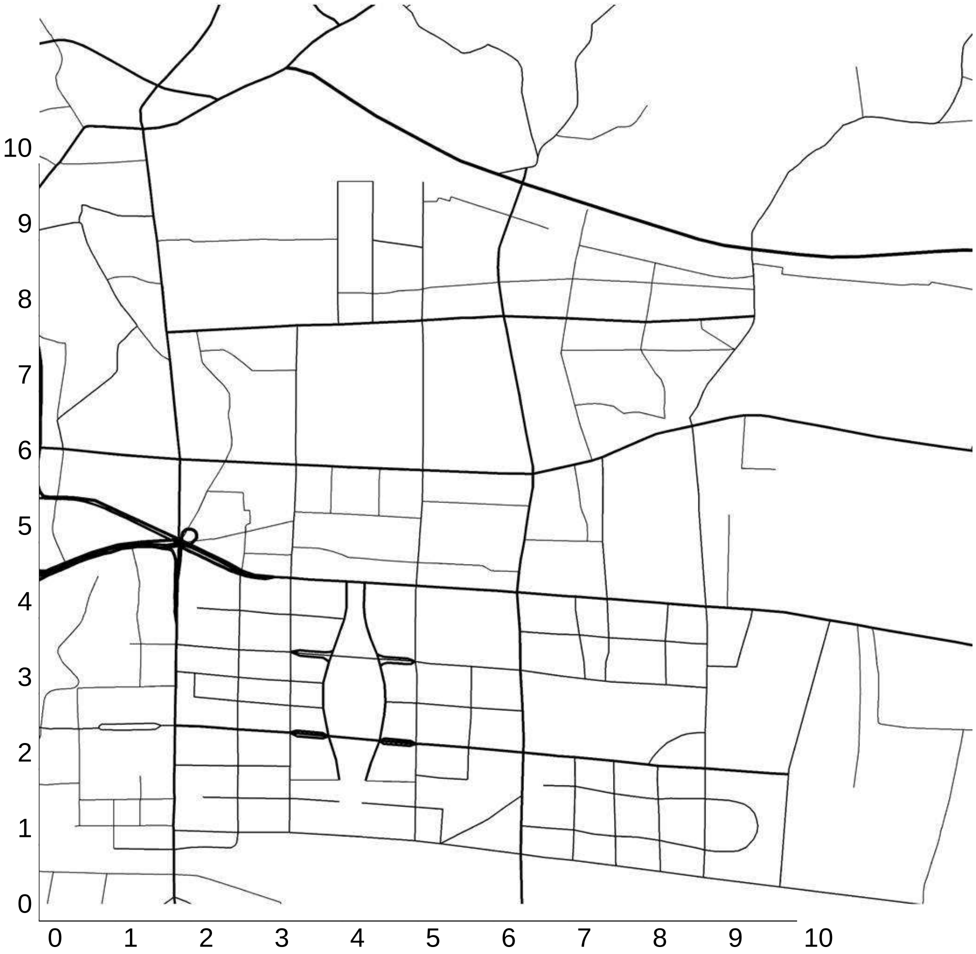

Figure 5: The network of the studied region from Guangzhou city

θ	$\vert\mathcal{D}\vert$	θ	$\vert\mathcal{D}\vert$
0.005	29	0.0025	3813
0.004	54	0.0023	5624
0.003	201	0.002	7766
0.0028	2036	0.001	9836

Table 6: Size of the generated choice sets (with 10000 draws)

- the number of signal-controlled intersections.

The minimum, maximum and average value of these variables in the sample are reported in Table 5.

The specification of the deterministic part of the utility is linear-in-parameter:

$$V_i = \beta_L \text{Length}_i + \beta_{ARR} \text{ArteryRoadRatio}_i + \beta_S \text{Signal}_i. \tag{19}$$

Several values of θ are tested for the MH algorithm. Each of them generate a choice set of different size, as described in Table 6.

The simulation procedure described in Appendix A to approximate $|\mathcal{C}|$ failed on this case study, as it involved very large numbers (larger than 10^{308}) that could not be handled. Therefore, the expansion factors w^G defined by (9), and w^F defined by (12) could not be used. We only use w^L defined by (14), as it does not require B.

To analyze the sensitivity of the new approach with different values of θ for the generation of $\mathcal{D}'$, we choose $\theta = 0.003$ and $\theta = 0.0029$. These values have been chosen to generate enough distinct paths so that we have a choice set with appropriate size to represent the correlation, and at the same time to exclude some extremely irrelevant paths which do not correlate with the chosen path. For each observation, we use 100 draws for the generation of $\mathcal{D}$ and 500 draws for the generation of $\mathcal{D}'$.

We specify a link nested logit model with two nest parameters: μ for links corresponding to non-artery roads and μ_{mA} for those corresponding to artery roads. As before, the inclusive parameter is assumed to be $\alpha_{im} = \ell_m / L_i$, where ℓ_m is the length of link m and L_i the length of path i. All the results are shown in Table 7. Columns 2 to 4 are estimations for the proposed model, and the last three columns are for the model without any correction. The scale parameter μ of the CNL model is normalized to 1 so that the models are compared with the same scale.

For model (15), the signs of the five parameters are all as expected, and the t-test (0) shows that all estimates are significant. The positive value of β_{ARR} indicates that taxi drivers prefer to drive on artery links. As for model (18), the t-test

	$\theta = 0.003$					
	Model (15)			Model (18)		
	Est.	Std. err.	t-test (0)	Est.	Std. err.	t-test (0)
β_L	-1.58	0.0566	27.9	-1.17	0.0642	18.2
β_{ARR}	8.09	0.636	12.7	4.63	1.86	2.49
β_S	-0.513	0.267	1.91	0.199	0.224	0.888
μ_m	3.90	0.117	33.3	3.98	0.166	24.0
μ_{mA}	2.22	0.257	8.62	1.23	0.417	2.95
Number of observations	740 trips from 7 OD					
Null log likelihood	-3.4078e+03			-3.4078e+03		
Final log likelihood	-1.9206e+03			-1.3978e+03		
Estimation time	22.32 hours[1]			39.02 hours[2]		
	$\theta = 0.0029$					
	Model (15)			Model (18)		
	Est.	Std. err.	t-test (0)	Est.	Std. err.	t-test (0)
β_L	-1.69	0.0653	25.9	-1.12	0.0615	18.3
β_{ARR}	7.63	0.648	11.8	4.57	1.27	3.61
β_S	-0.662	0.206	3.21	0.126	0.211	0.596
μ_m	3.99	0.139	28.6	4.88	0.171	28.6
μ_{mA}	2.23	0.835	2.668	1.26	0.198	6.38
Number of observations	740 trips from 7 OD					
Null log likelihood	-3.4078e+03			-3.4078e+03		
Final log likelihood	-1.9297e+03			-1.4181+03		
Estimation time	44.56 hours[2]			44.11 hours[2]		

[1] Computed with Matlab on Intel i5 with 4GB RAM, one processor

[2] Computed with Matlab on AMD FX with 16GB RAM, one processor

Table 7: Estimation results for the taxi drivers' route choice model in Guangzhou, China

	Model (15)	Model (18)
β_L	1.73	0.696
β_{ARR}	0.458	0.0506
β_S	0.149	0.348
μ_m	0.638	*5.24*
μ_{mA}	0.0137	0.148

Table 8: Sensitivity analysis

(0) for β_S indicates that it is not significantly estimated. Other parameters have their expected signs and t-tests are all significant. In both models, μ_m is greater than μ_{mA}, corresponding to a stronger correlation. Note that it is an important advantage of this method to be able to estimate the values of the nest parameters from data instead of postulating them.

The results in Table 7 show that the parameters estimates for model (15) are less sensitive to a modification of the θ parameter, contrarily to Model (18). To show this formally, we report in Table 8 the quantity

$$\left|\frac{\beta_i^{\theta=0.003} - \beta_i^{\theta=0.0029}}{\beta_i^{\theta=0.0029}}\right|. \tag{20}$$

We can perform a t-test to test if the value estimated with $\theta = 0.0029$ is significantly different from the one estimated with $\theta = 0.003$, used as a reference. All values are inferior to 1.96, so that the hypothesis cannot be rejected, except for μ_m with Model (18),which is sensitive to the value of θ.

From the results reported in Table 7, it appears that the estimated parameters are different for the two models. To show this formally, we calculate the quantities

$$\left|\frac{\beta_i^{\text{Mod. (15)}} - \beta_i^{\text{Mod. (18)}}}{\sigma_i^{\text{Mod. (15)}}}\right| \tag{21}$$

and

$$\left|\frac{\beta_i^{\text{Mod. (15)}} - \beta_i^{\text{Mod. (18)}}}{\sigma_i^{\text{Mod. (18)}}}\right| \tag{22}$$

The test (21) takes the value estimated for Model (18) as the reference, and tests if the value estimated for Model (15) is significantly different from it. Eq (22) does the opposite. The results are reported in Table 9. With the exception of μ_m for $\theta = 0.003$, all parameters are significantly different from one model to the next.

This case study shows the feasibility of our approach on real data. Combined with the findings on synthetic, it seems that the proposed model (Model (15)) is appropriate and tractable to estimate a cross nested logit model for route choice.

	(21)		(22)	
	$\theta = 0.003$	$\theta = 0.0029$	$\theta = 0.003$	$\theta = 0.0029$
β_L	7.24	8.72	6.38	9.26
β_{ARR}	5.44	4.72	1.86	2.41
β_S	2.67	3.82	3.18	3.73
μ_m	0.684	6.40	0.483	5.20
μ_{mA}	3.85	1.16	2.37	4.89

Table 9: Testing the difference between the two models

5 Conclusions

This article proposes to use sampling of alternatives with a cross nested logit model in the route choice context, where the universal choice set is difficult or impossible to obtain. In a Multivariate Extreme Value (MEV) model such as the CNL, capturing the correlation requires the enumeration of the full choice set of the nests. This paper proposes an approximation method inspired by the work by Guevara and Ben-Akiva (2013). Using a similar framework, we propose a new expansion factor to compensate for non sampled paths. This paper presents, to the best of our knowledge, the first tractable approach to estimate a MEV model using sampling of alternatives in a route choice context. The results on synthetic data demonstrate that the proposed approach is valid. The comparison shows that the new expansion factor generates good estimates.

The application with real data shows that the proposed method and expansion factor is practical. The results indicate that the proposed methodology generates significantly different results than the approach without corrections. Moreover, these results make more sense, reinforcing our confidence in the approach.

To summarize, the approach proposed in this paper have strong theoretical foundations and behaves well with both the synthetic and real data. It provides a solution to estimate the CNL model with sampled alternatives when the full choice set is inaccessible to the researcher. Moreover, it shows that it is possible to estimate the nesting parameter in a CNL route choice model. A natural extension would be to try and estimate all the nesting parameters instead of assuming them to be equal, because travelers are supposed to have different perception errors on different links according to their lengths and locations. Estimating the parameters α_{im} would be desirable as well. However the computational burden is currently unacceptable. To relieve the estimation burden and at the same time retain the behavioral meaning in the model, the concept of the sub-network (Frejinger and Bierlaire, 2007) may be investigated.

References

Abbe, E., Bierlaire, M. and Toledo, T. (2007). Normalization and correlation of cross-nested logit models, *Transportation Research Part B: Methodological* **41**(7): 795–808.

Bekhor, S. and Prashker, J. (2001). Stochastic user equilibrium formulation for generalized nested logit model, *Transportation Research Record: Journal of the Transportation Research Board* **1752**: 84–90.
URL: *http://dx.doi.org/10.3141/1752-12*

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*, Operations Research and Management Science, Kluwer, pp. 5–34. ISBN:0-7923-8587-X.

Bierlaire, M. (2006). A theoretical analysis of the cross-nested logit model, *Annals of Operations Research* **144**(1): 287–300.

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., Bekhor, S. and Prato, C. (2008). The factor of revisited path size: Alternative derivation, *Transportation Research Record: Journal of the Transportation Research Board* **2076**(-1): 132–140.
URL: *http://dx.doi.org/10.3141/2076-15*

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, *Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Pergamon, pp. 697–711.

Flötteröd, G. and Bierlaire, M. (2013). Metropolis-hastings sampling of paths, *Transportation Research Part B: Methodological* **48**: 53–66.

Frejinger, E. (2008). *Route choice analysis - data, models, algorithms and applications*, PhD thesis, Ecole Polytechnique Fédérale de Lausanne.
URL: *http://infoscience.epfl.ch/record/114777*

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

Frejinger, E. and Bierlaire, M. (2010). On path generation algorithms for route choice models, *in* S. Hess and A. Daly (eds), *Choice Modelling: The State-of-the-Art and the State-of-Practice*, Emerald Group Publishing Limited, pp. 307–315. ISBN:978-1-84950-772-1.

Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling, *Transportation Research Part B: Methodological* **43**(10): 984–994.

Guevara, C. A. and Ben-Akiva, M. E. (2013). Sampling of alternatives in multivariate extreme value (mev) models, *Transportation Research Part B: Methodological* **48**(0): 31–52.
URL: *http://www.sciencedirect.com/science/article/pii/S0191261512001518*

Jun, L., Lai, X. and Yu, Z. (2013). Paired route impedance correction for multinomial logit model based on equivalent impedance, *Transportation Research Board 92nd Annual Meeting*, number 13-2486.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlquist *et al.* (ed.), *Spatial interaction theory and residential location*, North-Holland, Amsterdam, pp. 75–96.

Papola, A. (2004). Some developments on the cross-nested logit model, *Transportation Research Part B: Methodological* **38**(9): 833–851.
URL: *http://www.sciencedirect.com/science/article/pii/S0191261504000049*

Parsons, G. R. and Kealy, M. J. (1992). Randomly drawn opportunity sets in a random utility model of lake recreation, *Land Economics* **68**(1): 93–106.

Prashker, J. and Bekhor, S. (1999). Stochastic user-equilibrium formulations for extended-logit assignment models, *Transportation Research Record: Journal of the Transportation Research Board* **1676**(-1): 145–152.
URL: *http://dx.doi.org/10.3141/1676-18*

Ramming, M. S. (2002). *Network Knowledge and Route Choice*, PhD thesis, Massachusetts Institute of Technology.

Roberts, B. and Kroese, D. P. (2007). Estimating the number of st paths in a graph, *J. Graph Algorithms Appl.* **11**(1): 195–214.

Sermons, M. W. and Koppelman, F. S. (2001). Representing the differences between female and male commute behavior in residential location choice models, *Journal of Transport Geography* **9**(2): 101–110.
URL: *http://www.sciencedirect.com/science/article/pii/S0966692300000478*

Vovsha, P. and Bekhor, S. (1998). Link-nested logit model of route choice: Overcoming route overlapping problem, *Transportation Research Record: Journal of the Transportation Research Board* **1645**(-1): 133–142.
URL: *http://dx.doi.org/10.3141/1645-17*

Wen, C.-H. and Koppelman, F. S. (2001). The generalized nested logit model, *Transportation Research Part B: Methodological* **35**(7): 627–641.
URL: *http://www.sciencedirect.com/science/article/pii/S019126150000045X*

A Approximation of the size of the choice set

We describe here the simulation method proposed by Roberts and Kroese (2007) to estimate $|\mathcal{C}|$. If ν is the set of nodes in the network, A is the adjacent matrix such that $A(i, j) = 1$ if there is a link between node i and node j, and 0 otherwise, the algorithm is described as follows.

Step 1 c denotes the current node and is initialized to the origin $c = O$. Denote by ℓ the likelihood of a path, and initialize it to 1. Initialize the iteration counter $t = 1$.

Step 2 Set $A(\cdot, O) = 0$ to make sure that the algorithm will not return to O.

Step 3 Define ν' as the set of all possible successors of c, that is

$$\nu' = \{k \in \nu | A(c, k) = 1\}. \tag{23}$$

If $\nu' = \emptyset$, the network is not connected. Stop the algorithm.

Step 4 Randomly choose the next node c uniformly within ν',

Step 5 Set $\ell = \ell/|\nu'|$, $A(\cdot, c) = 0$ and $t = t + 1$

Step 6 If $c = D$, stop. Otherwise return to step 3.

Repeat the algorithm N times, and use Eq. (24) to estimate the size of $\mathcal{C}$, where $\ell^{(i)}$ is the likelihood of the path generated by the algorithm during run i:

$$|\mathcal{C}| \approx \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\ell^{(i)}}. \tag{24}$$