Practical solutions for sampling alternatives in large scale models

Andrew Daly[a,b]
Stephane Hess[a]
Thijs Dekker[c]

a: Institute for Transport Studies, University of Leeds
36-40 University Road
Leeds LS2 9JT
United Kingdom
b: RAND Europe
c: Faculty of Technology, Policy and Management, Delft University of Technology
Jaffalaan 5
2628BX Delft,
Netherlands.

Abstract

Many large scale real world transport applications have choice sets that are so large as to make model estimation and application computationally impractical. The ability to estimate models on subsets of the alternatives is thus of great appeal, and correction approaches have existed since the late 1970s for the simple Multinomial Logit model. However, many of these models in practice rely on Nested Logit specifications, for example in the context of the joint choice of mode and destination. Recent work by Guevara and Ben-Akiva has put forward solutions for such GEV structures, but they remain difficult to apply in practice. This paper puts forward a simplification of their method for use in computationally efficient implementations of Nested Logit. We illustrate the good performance of this approach using simulated data and we also provide additional insights into sampling error with different sampling strategies for Multinomial Logit.

Keywords: *nested logit; sampling of alternatives; GEV; destination choice*

Word count: 6,827 words and 3 tables = 7,577 words

1. Introduction

In various empirical applications it is necessary to estimate choice models with substantial numbers of alternatives. In mode and destination choice models, for example, individuals face a wide range of spatially distributed destinations and a set of possible modes. Since the calculation of choice probabilities requires consideration of all the alternatives in the choice set, tens of thousands or more, such models can impose heavy demands on computer resources, particularly run time, but also potentially storage requirement. This problem increases substantially when making use of more extensive systems of models, as in activity-based modelling (see e.g. Bradley et al., [1]). An option to reduce these computational demands is to use sampling to restrict the number of alternatives actually used in estimation.

McFadden [2] set out the Positive Conditioning (PC) property under which consistent estimates of a Multinomial Logit (MNL) can be obtained using sampling of alternatives. Estimation under sampling with PC sampling procedures requires maximisation of a modified likelihood function with an added correction term in the utility function. Much more recently, Guevara and Ben-Akiva [3], hereafter abbreviated to GBA, extend the work of McFadden [2] to the GEV framework, so that consistent estimates can be obtained for two-level nested logit models, either tree-nested or cross-nested. Since the denominator of the logit formula and the additional GEV term introduced by

Guevara and Ben-Akiva in these nested logit models both contain a logsum calculated over a set of alternatives, sampling needs to be done at two points in the GEV framework. It is important to note that the sampling procedure need not be the same at the two points. In this paper, we further explore the research programme started by GBA for GEV models, in particular looking at making the work practical for large scale modelling.

Sampling of alternatives inevitably leads to increased error in parameter estimation. Nerella and Bhat [4] give an indication of the magnitude of the error, both for MNL and for more complicated models, such as mixed logit. For MNL, they give some guidelines on minimum sample size to achieve stability, but that is with a simple sampling strategy in simulated data and so not likely to be transferable to real data and more efficient sampling procedures. In particular, the efficiency of sampling can vary substantially between contexts and sampling procedures. Sampling alternatives in MNL or GEV inevitably causes noise, but we would not be able to state in advance what that noise would be in a specific situation. GBA can only recommend empirical testing to derive appropriate sampling levels.

In this paper, we extend the current state of knowledge on the impact of alternative PC sampling procedures and the resulting sampling error in MNL and GEV models. We particularly focus on mode and destination choice models. Typically, in these types of models individuals are faced with various modes and destination choices of which the choice probability is heavily affected by the travel accessibility from a specific origin. We investigate the way in which different distributions of choice probability over the alternatives, as would occur with variations in mode choice and trip length for different travel purposes, affects the effectiveness of different PC sampling schemes and estimation procedures. Like Nerella and Bhat [4] we use simulated data, but with a clear focus on applicability. Hence, the aim is to provide more transferable results on the arising sampling error.

Our focus differs from GBA, who apply their framework in a residential location choice model. The properties of mode and destination choice models are different from residential choices, because the choice probability is much more strongly linked to the travel accessibility from a specific origin. A clear aim of the paper is therefore to test the impact of alternative PC sampling in such a setting on the resulting sampling error in the GBA framework. In short, using simulated data we evaluate the efficiency and effectiveness of the Guevara-Ben-Akiva approach by exploring various PC sampling schemes in an attempt to minimise the estimation error for a given computational burden.

A further contribution of this paper is to simplify the approach of Guevara and Ben-Akiva to make it practical for large-scale modelling using existing efficient software. The simplification is achieved by reparameterising the nested logit model. Tests of the approach are made indicating the efficiency of the approach and how the errors vary as a function of model parameters and the level of sampling adopted.

In the following section of the paper, we discuss sampling strategies and how these may be expected to affect both the computation time and the accuracy of the modelling. This topic does not seem to have been discussed at any length in the literature and, as an initial contribution, we give some results on sampling procedures and then on sampling error in MNL, which in turn give indications for methods and errors in more complex models. The following section looks at the issues of sampling in GEV models, drawing on the work of Guevara and Ben-Akiva but taking a more practical and simpler approach intended for large-scale applications. Section 4 presents the results of tests based on simulated mode and destination choice models. The final section presents conclusions and recommendations for applications and future research.

2. Sampling strategies and error in MNL

McFadden [2] set out the PC property under which consistent estimates of a MNL can be obtained with alternative sampling. Specifically, he showed that asymptotically consistent estimates of model parameters can be obtained if we maximise a modified log likelihood function, with a contribution for each observed choice of

$$L = \log \frac{\exp(V_c + \log \pi(D|c))}{\sum_{j \in D} \exp(V_j + \log \pi(D|j))} \quad (1)$$

where V_j is the systematic part of utility for alternative j;
c is the chosen alternative;
D is the sampled set of alternatives, which is a subset of the set of all available alternatives C; and
$\pi(D|j)$ is the probability of sampling D, if j is the chosen alternative.

The PC property, i.e. *positive* conditioning, is the condition that $\pi(D|j) > 0 \; \forall \; j \in D$, which is clearly necessary to evaluate (1). Note that it is also essential that the chosen alternative c is included in D. For estimation purposes, (1) implies working as if D was the complete choice set, not C.

Clearly, if $\pi(D|j)$ is the same for all $j \in D$, then it will cancel out in (1). The system then conforms with the Uniform Conditioning (UC) property also defined in McFadden [2]. The simplicity of UC is attractive, but in many practical cases some alternatives are much more important than others, in the sense of being much more likely to be chosen, so that the general PC approach with unequal π values is more efficient. In particular, in modelling destination choice it is clear that nearer alternatives are each much more likely to be chosen than distant alternatives and common sense suggests they are therefore more relevant for modelling. Some intuition on how 'important' alternatives should be identified is given in Section 2.2.

An important point in practice is that McFadden's PC theorem requires the assumption that the true choice model is MNL. In practice, this will often *not* be the case and the consequence is then that estimation using the amended likelihood function may not give consistent estimates of the parameters that would be obtained when using the full model. However, in this case, neither the base MNL nor the sampled version is estimated correctly or gives a true representation of behaviour. The theorem also requires that each choice observation be treated as independent, an assumption we shall maintain in this paper, though in some important practical cases the assumption may not be appropriate. The assumption will usually be valid if each observed choice is made by a separate individual as is the case in most revealed preference studies and thus large scale models by extension.

2.1 Practical strategies for PC sampling

A practical approach for PC is to use independent sampling, where each unchosen alternative j is included in the sample with probability q_j, making a separate draw for each alternative. Another approach is to sample a fixed number of times from C, with replacement, giving each alternative a probability q_j of being sampled at each draw, then deleting the duplicate sampled alternatives. In each case these strategies yield

$$\pi(D|j) = \frac{1}{q_j} K(D) \quad (2)$$

with $K(D)$ independent of j (Ben-Akiva and Lerman [5], equations 9.22 and 9.23). Examining the log likelihood (1), we see that $K(D)$ cancels out and we are left with

$$L = \log\left(\frac{\exp(V_c - \log q_c)}{\sum_{j\in D}\exp(V_j - \log q_j)}\right) = V_c^M - \log\sum_{j\in D}\exp V_j^M \qquad (3)$$

where $V_j^M = V_j - \log q_j$ is an amended utility function for alternative j.

With independent sampling the expected set size is $\sum_{j\in C} q_j$ but there is quite likely to be some variation around this number. When sampling with replacement the probabilities q_j must sum to 1, of course, but the advantage claimed in Ben-Akiva and Lerman for this method is that the size of the set D varies less than with independent sampling. However, the expected set size is more complicated to determine. In each case we can adjust the sampling rate to obtain a suitable balance between sampling error and computational cost.

Another sampling strategy, stratified sampling, involves division of the choice set into a number of strata and sampling a fixed number of alternatives in each stratum. For efficiency, the relative frequencies of selection would relate approximately to the choice probabilities. Ben-Akiva and Lerman show how to calculate the values of $\pi(D|j)$ when the sample rate is constant in each stratum and indicate that the main advantage of this approach is that fixed set sizes are obtained for D. However, a fixed set size does not necessarily give an important advantage in practical estimation, particularly when the number of alternatives sampled from each stratum provides an inadequate representation of the underlying choice probabilities.

Some aspects of the sampling issues can be tested quite readily by simulation, as we will now illustrate in the context of destination choice among zones in a hypothetical study area. These simulations are based on 10,000 draws of sets of destinations from a total set of 100. For person *n* out of *N* (with *N*=5,000), the travel time by car to destination *j (j=1,…,100*) is assumed to be given by:

$$ttc_{n,j} = u_{n,j,1}10\sqrt{j} \qquad (4)$$

The term $u_{n,j,1}$ is a random component that serves to add variance across individuals in the travel times to given destinations, where $u_{n,j,1}$ is uniformly distributed across individuals and destinations between 0.8 and 1.2. The square root also ensures that there are more opportunities (destinations) at increasing distance.

Each destination is attributed a utility function of

$$V_{n,j} = \beta_{ttc}.ttc_{n,j} + \gamma.\delta_{j=1} + \eta.\delta_{j=62\ldots66} \qquad (5)$$

where $\delta_{j=1}$ indicates a trip to zone 1, perhaps the origin zone of the trip;
$\delta_{j=62\ldots66}$ indicates a trip to the 'central area' formed of zones 62 to 66;
β_{ttc}, γ and η are the assumed parameters of the model.

In these models we set $\gamma = 1$ and $\eta = 1$ and β_{ttc} is set to scale the impact of distance. The advantage of different approaches may vary with the size of the study area relative to mean trip lengths and this is represented in the simulations by varying β_{ttc}: a larger negative value implies shorter trip lengths or, equivalently, a larger study area relative to a given trip length. Two sets of

samples were drawn, each with varying values of β_{ttc}, with five different values, going from -0.03 to -0.11 in steps of 0.02. Independent sampling is undertaken with $q_j = f.\exp V_j / \sum \exp V_k$ and f is set to achieve a roughly uniform sample size. In the first set of 10,000 samples for each β_{ttc}, we aimed for a sample of approximately 1 in 6 destinations, while the second aimed for approximately 1 in 12. Replacement sampling is undertaken $\tilde{J}$ times, with $q_j = \exp V_j / \sum \exp V_k$ and $\tilde{J}$ set to achieve a roughly uniform sample size, again using two samples, one at around 1 in 6 and one at around 1 in 12 destinations. Sampling without replacement was done with 16 or 8 samples each time, with $q_j = \exp V_j / \sum \exp V_k$ and the denominator reduced at each step to account for the sampling at the previous step. We excluded stratified sampling from this analysis after initial results confirmed that it consistently led to the poorest performance overall. The appropriate sampling rate for a particular study will depend on the computational cost and the estimation accuracy required.

The results from this simulation process are summarised in **Error! Reference source not found.**. We also show the coverage of expected choices, which is the expected cumulative choice probability covered by the sampled alternatives. Finally, to avoid dependence on the precise sample sizes, measures of 'effort' and 'variation' were devised as follows:

$$Effort = \frac{E(fraction\ of\ alternatives\ sampled)}{E(cumulative\ choice\ probability\ of\ sampled\ alternatives)} \qquad (6)$$

$$Variation = \frac{St.dev.\ of\ number\ of\ alternatives\ sampled}{E(number\ of\ alternatives\ sampled)} \qquad (7)$$

The effort relates to the relative computation effort required for each of these protocols, calculated as the ratio of the expected percentage of alternatives sampled over the expected cumulative choice probability (i.e. the coverage of expected choices) covered by the sampled alternatives. The variation is simply the coefficient of variation of the sampled set size in the runs, for those strategies where the sample size varies across respondents.

[Table 1 about here]

As can be seen in **Error! Reference source not found.**, with the approximate 1/6 sampling rate, the coverage of expected choices runs from about 30% with a value of –0.03 for β_{ttc} to about 90% with a value of -0.11 for β_{ttc}. Clearly, it is not possible for the analyst to control the model parameters, which are a function of behaviour and the study area, but it is possible to control the sampling rate. To see the effect of changing this rate, we can note that when aiming for a sampling rate of about 1/12, the coverage of expected choices runs from about 20% to 80% as β_{ttc} varies.

Looking at the measures of effort and variation, we can see that for the former, the differences between the sampling approaches are very small and all the procedures become more efficient as β_{ttc} becomes more strongly negative, and the initial small advantage of independent sampling essentially disappears. In the 1/12 runs, with replacement and without replacement sampling give indistinguishable results in terms of effort. For the same value of β_{ttc}, the relative effort is substantially less in the 1/12 runs across all values of β_{ttc}; that is, more than half of the coverage is achieved when we sample half as many alternatives with these protocols.

Turning to the coefficient of variation of the sampled set size for those strategies where the sample size varies (i.e. independent and replacement sampling), we can see that while, for small values of β_{ttc}, replacement sampling gives less variation than independent sampling, as expected, the difference decreases substantially as the parameter increases and for the largest value in the 1/6 runs, independent sampling gives a *less* variable set size, contrary to the expectation of Ben-Akiva and Lerman [5]. Contrary to what we saw for effort, the relative variation in set size is nearly

doubled for both independent and with-replacement sampling in the 1/12 runs, while the advantage in this respect of with-replacement sampling is retained for the full range of β_{ttc} values tested.

These tests show that there is little to choose in efficiency between the three main sampling strategies that might be considered. The impact of varying set sizes is not very important *in practice*, while the simplicity of calculation for independent sampling is very helpful in the calculations we need to make in the remainder of the paper; these involve more issues than just the correction $\pi(D|j)$. For this reason we use independent sampling in the rest of the paper.

2.2 Sampling error in PC sampling

The work of GBA, discussed in more detail in the following section, indicates that the 'sandwich' matrix can be used to obtain the error in the parameter estimates for a GEV model when the alternatives are sampled. It does not seem to have been noted previously in the literature that this result can be applied to MNL. This finding is important and implies that sandwich error estimators should be used in all cases when alternatives are sampled. However, the sandwich matrix includes both the error induced by sampling alternatives and the 'ordinary' error that would be present if the full sample of alternatives were used. To determine how sampling error varies with the size of the sample of alternatives, and therefore what a suitable sample size might be in any given context, requires more specific analysis.

A simple approach to assessing sampling error is to calculate the coverage or the fraction of expected choices that are captured by D:

$$P_D = \frac{\sum_{j\in D} \exp(V_j)}{\sum_{j\in C} \exp(V_j)} \tag{8}$$

This approach is taken by Miller *et al.* [6] in the context of sampling alternatives for application. It is intuitively clear that as the fraction increases then the approximation will generally improve. But this is a rough measure of sampling error.

The standard MNL gives the log probability of the observed choice c for a single individual by

$$L = \log p_c = V_c - \log \sum_{j\in C} \exp(V_j) \tag{9}$$

The objective of sampling alternatives is to save time by not evaluating the logsum, the last term in (9), in full. That is, writing $\exp(V_j) = w_j$ for economy of notation, we want to estimate $W = \sum_{j\in C} w_j$ by $\widetilde{W} = \sum_{j\in D} \frac{w_j}{q_j}$, where D is the set of j that have been sampled from C and q_j is the expansion factor for each j. That is, we approximate

$$L \cong V_c - \log \widetilde{W} = V_c - \log \sum_{j\in D} \frac{\exp(V_j)}{q_j} \tag{10}$$

If we apply independent sampling, so that q_j is simply the sampling probability for j, the variance of $\widetilde{W}$ over different samples is given by

$$\mathrm{var}(\widetilde{W}) = \sum_{j\in C} \left(\frac{w_j}{q_j}\right)^2 q_j \left(1 - q_j\right) = \sum_{j\in C} {w_j}^2 \left(\frac{1}{q_j} - 1\right) \tag{11}$$

The variance can obviously be reduced by increasing q_j and clearly becomes zero when $q_j = 1$ for all j. However, the calculation cost is proportional to the number of alternatives sampled and the expectation of this number is given by $\sum_{j \in C} q_j$. Holding this expected calculation cost fixed, it is quite easy to see that the variance (11) is minimised when $q_j = k.w_j$ for a constant k (see also Hammersley and Handscomb [7]). This result gives a strong indication that the intuitive attribution of sampling probability as approximately proportional to 'importance', measured by $\exp(V_j) = w_j$, i.e. roughly proportional to choice probability, is a reasonable way to minimise error. Of course, we cannot calculate the true $\exp V_j$ in advance of estimating the model, so importance sampling has to be performed with an approximate proxy used for w_j. This will usually be done using information from previous studies and therefore does not introduce endogeneity with the estimates to be made on the basis of the sample drawn.

The calculations above apply for independent sampling. For replacement sampling we might expect that an equation like (11) could be developed, though it would not necessarily be so simple. For other sampling protocols the formulae are likely to be even more complicated. This is the chief reason we have adopted independent sampling for the current paper, although we showed above that it was also likely to be a good approach on criteria of efficiency.

The error in the likelihood-contribution calculation (9), to which $\widetilde{W}$ contributes the sampling error, can be estimated as

$$\mathrm{var}(L) \cong \left(\frac{\partial L}{\partial \widetilde{W}}\right)^2 \mathrm{var}\left(\widetilde{W}\right) = \frac{\mathrm{var}(\widetilde{W})}{\widetilde{W}^2} \cong \frac{\mathrm{var}(\widetilde{W})}{W^2} \qquad (12)$$

The final approximation follows because $\widetilde{W}$ is an estimate of W and, using (11) to obtain $\mathrm{var}(\widetilde{W})$, we are then able to make a calculation of the expected error variance (12) in terms of quantities that are known before any sampling is done.

Thus, for MNL, the error in the log likelihood is equal to the square of the coefficient of variation of $\widetilde{W}$, which in turn is a function of simple statistics of the set D. It may be noted that the calculations needed to derive the expected value of $\mathrm{var}(L)$, i.e. (11) and (12), can be made in advance, so that a sample size can be set a priori to obtain an appropriate balance between likelihood error and sample effort.

In order to check that (11) and (12) give a good expectation of the likelihood variation to be found in practice with models estimated on samples, the relevant statistics were derived for a series of simulation runs. These runs used the same independent sampling approach described in the previous section, with approximate 1/12 samples and with the same settings in terms of five different values for β_{ttc} and corresponding values for f to ensure the 1/12 rate. The results are shown in **Error! Reference source not found.**, where, alongside (11) and (12), we also include the crude measure from (8).

The first measure shows that the coverage given by the sampling increases substantially as distance becomes more important, i.e. with increases in β_{ttc}. That is, importance sampling is more effective when there is a strong impact of distance in determining choice. The second measure shows that, in parallel, the variation in the estimated logsum $\widetilde{W}$ reduces substantially. Finally, the calculation results for the third measure show that likelihood variation with sampling of alternatives reduces dramatically as the distance effect grows. Clearly, if these measures connect to errors in parameter estimates then sampling can be reduced more when the distance effect is greater.

While these measures are indicative, further work is also needed to determine how the error in likelihood calculation (12) carries through to error in parameter estimation. It seems likely that there would be a proportionality relationship. Meanwhile we can use the approximate error indicators P_D and var(L), the latter calculated using (11) and (12), to obtain some insight into error.

3. Sampling in GEV models

In this section we move from the basic MNL model to the GEV framework, where we start with a brief overview of the literature of sampling in a GEV framework, go on to summarise the parts of the GBA work most relevant for our study and close by outlining the developments we propose.

3.1 Previous research

It is important to distinguish between sampling of alternatives, which is the focus of the current work, and sampling of observations, an important issue but one which is not directly related to the sampling of alternatives. The distinction is not always made clear in literature reviews. For example, Koppelman and Garrow [8] do not discuss sampling alternatives at all. Another paper that is mentioned in literature reviews is Mabit and Fosgerau [9], but again this does not mention the sampling of alternatives.

Bierlaire *et al*. [10]) is also aimed chiefly at the issue of sampling observations. However, "for the sake of completeness", they give some attention to sampling alternatives, deriving results that foreshadow somewhat the work of GBA. However, the latter work is more complete and more directly focussed on our topic of interest. Frejinger *et al*. [11] do not deal with models beyond MNL except through the 'path size' correction and the PC correction is therefore sufficient for their work. Similarly, Train [12] does not go beyond the results given in McFadden [2].

Lee and Waddell [13] claim to provide the first consistent estimator for tree-nested logit with sampling of alternatives. The formula (their equation 5) is simple, the logsum used in the higher (unsampled) level is

$$V_m = \left(\frac{1}{\mu}\right) \log\left(\sum_{i \in m} \left(\frac{1}{R}\right) \exp(\mu V_i)\right) \qquad (13)$$

where R is the sampling rate "which only applies to the sampled non-chosen alternatives", so they apply a rate of 1 to the chosen alternative. The estimate of the logsum is therefore a function of the chosen alternative. When $\mu = 1$, i.e. the model is MNL, (13) is different from McFadden's PC, so that it appears that the Lee and Waddell procedure is incorrect and, indeed, simple simulations can serve to confirm that a bias is introduced.

Our investigation takes the work of GBA as a starting point. Their work is more complete than e.g. Bierlaire et al. [10] and directly focussed on our topic of interest.

3.2 Guevara and Ben-Akiva work

GBA give the theorem that consistent estimation of a GEV model (McFadden [2] introduces the GEV family) based on a sample of alternatives D can be achieved by a correction of the logit utility function

$$V_i^* = V_i + \log G_i(D^*) + \log \pi(D|i) \qquad (14)$$

where $\pi(D|i)$ is the probability of selecting the reduced choice set D, given that i is the chosen alternative; we note that this is reassuringly the standard McFadden PC correction;
G_i is the derivative with respect to its i^{th} argument of the GEV generating function G; here we note that it is calculated over a restricted choice set D^*. This set has to be chosen to give an unbiased estimate of the true G_i calculated over C and exactly how this is to be done is discussed further below.

The theorem also gives the error in the parameter estimates as asymptotically normal, with covariance equal to the well-known 'sandwich' matrix (cf. Huber [14]), subject to technical conditions.

In an MNL model, $G_i = 1$ for all the alternatives, so that this term disappears from the function and we return to the standard McFadden MNL PC formulation. However, in more general GEV, such as nested logit, this term does not disappear. Ben-Akiva and Lerman [5] show that (14) can be used (without sampling, i.e. without the π term) to represent any GEV model, so that the GBA theorem using (14) represents an intuitive extension of both McFadden sampling and the Ben-Akiva/Lerman finding.

For two-level tree-nested logit (i.e. excluding the possibility of cross-nesting), GBA obtain the formula:

$$\log G_i = \left(\frac{\mu}{\mu_{m(i)}} - 1\right) logsum(m(i)) + \log \mu + \left(\mu_{m(i)} - 1\right) V_i \qquad (15)$$

where $\mu_{m(i)}$ is the nesting coefficient for the nest $m(i)$ that contains alternative i.

In this formula, the logsum has to be approximated, in addition to the standard logsum in the choice probability, as otherwise we need to make calculations for all the alternatives, defeating the objective of saving calculation time. The estimator GBA propose for the logsum for nest m is given by:

$$logsum(m) \approx \log \sum_{j \in D^*(m)} \frac{\tilde{n}_j}{E(j)} \exp\left(\mu_m V_j\right) \qquad (16)$$

where $D^*(m)$ is the set of sampled alternatives within nest m;
n_j is the number of times alternative j is actually sampled and
$E(j)$ is the expectation of this number.

It is shown by GBA that the term $\tilde{n}_j / E(j)$ is exactly the expansion factor required to obtain an unbiased estimate of the logsum. It is important to note that the sampling procedure used to obtain D^* to estimate the logsum *need not be the same* as the procedure used to sample the set D. A key consideration is that the set D must contain the chosen alternative, so that the probability that it is selected depends on the choice probabilities and hence on the parameters of the model. GBA discuss two alternative procedures and show them to work well on simulated data.

1. The same sampling can be used, i.e. $D^* = D$, but in this case, because of the dependence of D on the chosen alternative, the expansion factors depend on the model parameters and the model must be estimated iteratively or by approximate methods.
2. Using separate sampling procedures, such that the sampling for D^* for the logsum approximation does not depend on the chosen alternative, does not require iterative estimation or further approximations.

Clearly, procedure 2 without iteration is more convenient in practice. Further, it also appears that in existing software, procedure 2 is easier to implement. Finally, GBA give no guarantee that the iterative process 1 converges, although no problems are reported from their tests. Procedure 2 is therefore to be recommended, provided the analyst has sufficient access to the data to make the required manipulations.

3.3 Implementation in practice

For practical work, equations (15) and (16), as presented by GBA, are inconvenient and not immediately suited to implementation in practical software suitable for large-scale modelling. However, by making some simple changes we can make that implementation, as we now show.

If we apply independent sampling for D^*, $\tilde{n}_j = 1$ for $j \in D^*(m)$ and $E(j) = q_j$, so that (16) can be written as:

$$logsum(m) \approx \log \sum_{j \in D^*(m)} \exp\left(\mu_m V_j - \log q_j\right) \qquad (17)$$

The GBA equations (15), (16) and (17) are written for a tree logit specification as used in e.g. Ben-Akiva & Lerman [5], which divide utilities in the nest-specific choice probabilities by the structural parameter. This is the specification also sometimes referred to as RU2 – see Hensher et al. [15]. For practical implementation it is easier to use the version which lacks this normalisation, referred to as RU1 by Hensher et al., [15] and as implemented in ALOGIT, where consistency with utility maximisation is ensured through an equality constraint between structural parameters on a given level. We have:

$$\mu_m = 1, \text{ for all } m \quad \text{and, to simplify further,} \quad \mu = \phi + 1 \qquad (18)$$

which gives the much simpler equation, replacing (15):

$$\log G_i = \phi . logsum(m) + \log(\phi + 1) \qquad (19)$$

Moreover, the term $\log(\phi + 1)$ is constant across the alternatives and can therefore be omitted from the practical calculations. Thus, if we are using independent or with-replacement sampling for D, and using the brief notation $V_j^M = V_j - \log q_j$ introduced in (3), noting that q_j is a constant in the estimation process, we can implement (14) as

$$V_i^* = V_i^M + \phi . \log \sum_{j \in D^*(m(i))} \exp V_j^M \qquad (20)$$

This is the form we use in our practical tests. Note that $-1 < \phi \leq 0$ to be consistent with the usual constraints on structural parameters in nested logit. A simple approach, which we have used, is to use $D = D^* \cup \{c\}$, i.e. creating D by adding the chosen alternative if it is not already selected by the sampling procedure for D^*.

As mentioned above, the constraint that μ_m is constant across the nests m is necessary to apply this specification consistently with Random Utility theory without introducing multiple levels of nesting. In mode-destination choice modelling this constraint would usually be applied.

4. Empirical tests

This section presents the results of an empirical analysis on simulated data aimed at providing empirical support to the discussions in Section 3. We also include runs using MNL models, where, in contrast with Section 2, we now look at bias in estimates while the focus before was on sampling error. This practical testing is aimed at modelling mode-destination choice, a context that involves different considerations from those studied by GBA, who looked at residential location choice. Indeed, destination choice is much more strongly dependent on separation that residential choice, so that the PC approach is more important.

4.1 Set up for practical testing

The practical tests reported here relate to modelling the choice of mode and destination, an important practical issue arising in travel demand forecasting studies. A simple approach to sampling alternatives in these studies is to make a sample of destinations, including in the sampled choice set all of the modes that are relevant for the sampled destinations.

An interesting feature of the GBA result is that, if sampling is such that no logsums require approximation, then no correction is required. For example, if we have a mode-destination choice model with destinations 'above' modes (i.e. the destination utility contains a logsum over modes) and we sample destinations but not modes, then there is no approximation of logsums. That is, in (20), D^* is always the complete set of alternatives (modes) in each nest corresponding to a destination that is sampled. But if modes are 'above' destinations (the mode utility includes a logsum over destinations) then a sampling correction is required. A practical investigation of mode and destination choice should investigate both these possibilities. The testing in this paper, however, investigates only the case of modes above destinations, where correction is required.

For the implementation of (20), two approaches can be considered.

1. The logsum term $\log\left(\sum_{j\in D^*(m)} \exp\left(V_j - \log q_j\right)\right)$ can be pre-calculated using preliminary estimates of the model parameters inside V. These logsums can then be used in a simple MNL model to obtain an estimate of ϕ and new estimates of the parameters inside V, which then permit an updated calculation of the logsums. This begins an iterative procedure which, perhaps with luck, will converge.
2. A notional tree structure can be set up, with each of the alternatives in D^* appearing in a nest feeding into each of the alternatives in D. This formulation is of course more complicated than approach 1, but does not require iteration and does not require an appeal to luck to converge.

In the practical tests we focus on the second approach.

The setup for our empirical tests reuses the sampling basic ideas from Section 2.1 with 100 destinations. We now however add a second mode, public transport, with:

$$ttPT_{n,j} = \frac{4}{3} u_{n,j,2} 10\sqrt{j} \qquad (21)$$

The term $u_{n,j,2}$ is again a random component that serves to add variance across individuals in the travel times to given destinations, where $u_{n,j,2}$ is uniformly distributed across individuals and destinations between 0.8 and 1.8, creating more variance than for car travel; the additional $\frac{4}{3}$ multiplier gives slower speeds for public transport than car. The utility function for public transport follows the same approach as for car, with:

$$V_{n,j} = \beta_{ttPT}.ttPT_{n,j} + \gamma.\delta_{j=1} + \eta.\delta_{j=62\ldots66}, \quad (22)$$

with no mode specific constants. In our simulation work, we choose values of *f* to ensure sampling rates of roughly 1/12, reusing the same five 'true' values for β_{ttc} (-0.03; -0.05; -0.07; -0.09 and -0.11) and *f* (8; 9; 10; 12; 15) as in Section 2.1, with $\beta_{ttPT} = \frac{7}{3}\beta_{ttc}$ to lead to a much higher travel time sensitivity on public transport. We use independent sampling in our empirical work, with ten sets of choices simulated for each sample and ten different samples of destinations drawn, for each set of true parameter values. This thus leads to a total of 500 samples from which models are estimated.

4.2 Estimation results

Three different versions of the 500 samples were generated for empirical testing. In the first set, we made use of a MNL model in simulation but focussed on destination choice only, under the assumption that car is chosen. The full choice set thus includes 100 alternatives, where we use the parameter values given in Section 4.1 in simulating the data. The second set of 500 samples were generated once again using a MNL model where we now however looked jointly at mode and destination choice thus using 200 alternatives in the full choice set. Finally, we also simulated choices for a Nested Logit model, using the coefficient and F values from Section 4.1 and with nesting by mode, with a true value of 0.5 for the structural parameter, giving a true value of -0.5 for ϕ in (20), and true values of 1 for γ and η. The estimation results are summarised in **Error! Reference source not found.**. In presenting the results, we average first across the ten sets of choices in each sample, and then give the means and standard deviations of the relevant measures across the ten samples of alternatives. In this way we focus on the variation over alternative sampling, which is the aim of the paper, and attempt to reduce the impacts of simulating the data and choices.

The presentation of results focusses on measures of bias, with the actual estimates not reported here (being directly related to the reported bias). We first note very high stability across the ten samples of alternatives within each set of simulations, with coefficients of variation for all estimates and t-ratios never going above 0.07 in absolute values. This is a strong indication of the stability of the results across samples meaning that the sole interest now is in the reliability of the resulting estimates.

The mean percentage bias in absolute value across all models and all settings is a mere 1.6%, with the highest bias being 8% (η in setting 4 for the mode destination MNL). This includes the estimation of ϕ in the Nested Logit model, suggesting that the proposed approach in (20) could be an attractive solution for large scale Nested Logit applications. In addition, we compared the bias in each model to the standard errors for the associated parameter, and the table reports the mean values in these t-ratios across samples, as well as the variation. There are only two cases with moderate levels of significance for the bias, both in the Nested Logit runs using the lowest values for the time coefficients, with an average t-ratio of -1.99 for the bias in β_{ttPT} and an average t-ratio of -1.84 for the bias in φ. The actual associated levels of bias remain very small, at 5.5% and 6.1%, respectively.

In principle, we would hope to be able to relate the variation in the coefficient estimates across the alternative-sampling runs to the analytic measures of error derived in Section 2.2, which are repeated in **Error! Reference source not found.**. However, the variation in the parameters, even at this relatively low level of sampling, is so small that such a connection is not possible.

[Table 3 about here]

The simulation runs were made (in part) on an industry-standard laptop computer. On this machine, the estimations (to convergence) averaged 1.3 seconds for the MNL destination choice model, 2.4 seconds for the MNL mode-destination choice model and 39.4 seconds for the nested model in the ALOGIT software. While the extension of the model with the notional nest in (20) to accommodate the logsum estimation increases the run time substantially, it remains entirely practical.

Our results are not directly comparable to those of the two-level nested logit model presented by GBA. Apart from applying a different model specification and normalization, they explore the bias associated with i) alternative methods for sampling of alternatives, at ii) alternative sampling rates. Our analysis has explored the resampling approach, identified by GBA as being the most well-behaved method, at a fixed sampling rate over a range of varying parameter values. Like GBA, our t-ratios for the bias are sufficiently low to assume that our proposed RU1 modification of the GBA approach can offers substantial computational benefits for practitioners of large scale nested choice models.

5. Conclusions

This paper has addressed the issue of sampling alternatives for practical work in large-scale models, with particular reference to the mode-destination choice models used in transport planning. The early work of McFadden [2], applicable only to MNL, remained the only approach for 30 years. Even for McFadden's approach, little was known about the error introduced by sampling and the best approach to minimise that error.

The paper investigates the potential sampling approaches and their efficiency, concluding that independent and with replacement sampling offer efficiency and simple correction procedures. Without-replacement sampling is a little less efficient and more complex to correct, while stratified sampling is not efficient.

In practical work, we start from the theorem of GBA that has recently taken sampling of alternatives forward significantly by allowing consistent estimation for GEV models that go beyond MNL. However, their formulae are not suitable for practical work and we present simplified formulae for tree-nested logit models.

In practical testing we present simulation results from MNL models of destination and mode-destination choice, showing that the variation of parameter estimates is small and bias is nearly absent. Similarly, we apply the simplified GBA formulae to estimate tree-nested models, again showing that bias is absent and variation across samples of alternatives is small.

We conclude that this approach represents a worthwhile possibility for practical implementation. Tests using real data should next be undertaken to test its practicality.

Acknowledgement

This work was conducted as part of the ACTUM project financed by the Danish Strategic Research Council.

We are grateful to Angelo Guevara for clarifying points concerning the alternative sampling procedures in his work, but we remain responsible for any errors in this work.

References

[1] Bradley, Mark, John L. Bowman and Bruce Griesenbeck (2010) SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution, Journal of Choice Modelling 3(1), pp. 5-31.

[2] McFadden, D.L. (1978), Modelling the choice of residential location, in Karlqvist, A., Lundqvist, L., Snickars, F. and Weibull, J., Spatial interaction theory and residential location, North-Holland, pp. 75-96.

[3] Guevara, C.A. and Ben-Akiva, M. (2013), Sampling of alternatives in multivariate extreme value (MEV) models, Transportation Research Part B 48, pp. 31–52

[4] Nerella, S. and Bhat, C. (2004), A numerical analysis of the effect of sampling of alternatives in discrete choice models, TRB.

[5] Ben-Akiva, M. and Lerman, S. (1985), Discrete Choice Analysis: theory and application to travel demand, MIT Press, see pp. 261-269 (Estimation of choice models with a sample of alternatives).

[6] Miller, S., Daly, A., Fox, J. and Kohli, S. (2007), Destination sampling in forecasting: application in the PRISM model for the UK West Midlands Region, presented to European Transport Conference, Noordwijkerhout.

[7] Hammersley, J. and Handscomb, D. (1964), Monte Carlo Methods, Chapman and Hall, pp. 57-59 (Importance Sampling).

[8] Koppelman, F. and Garrow, L. (2005), Efficiently estimating nested logit models with choice-based samples: Example applications, Transportation Research Record 1921: 63-69.

[9] Mabit, S. and Fosgerau, M. (2006) unpublished note, Danish Technical University (extract from Mabit's thesis).

[10] Bierlaire, M., Bolduc, D. and McFadden, D. (2008), The estimation of generalized extreme value models from choice-based samples, Trans. Res. B, **42**, pp. 381-394.

[11] Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009), Sampling of alternatives for route choice modelling, Trans. Res. B, **43**, pp. 984-994.

[12] Train, K. (2009), Discrete Choice Methods with Simulation, second edition, Cambridge University Press, Cambridge, MA.

[13] Lee, B.H. and Waddell, P (2010), Residential mobility and location choice: a nested logit model with sampling of alternatives, Transportation, **37**, pp. 587-601.

[14] Huber, P. J. (1967). "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. I, pp. 221–33.

[15] Hensher, D.A., Rose, J.M. & Greene, W.H. (2005), Applied Choice Analysis: A Primer, Cambridge University Press, Cambridge, MA.

List of Tables

Table 4: Results of simulation experiments for destination choice

		Independent sampling					Replacement sampling					Without replacement	
	β_{ttc}	f	average sample	average coverage	effort	variation	$\tilde{J}$	average sample	average coverage	effort	variation	average coverage	effort
Aiming for 1/6	–0.03	17	16.28	31.99%	0.51	0.21	20	16.66	30.51%	0.55	0.09	29.44%	0.54
	–0.05	19	16.25	53.03%	0.31	0.18	25	16.97	50.19%	0.34	0.12	48.32%	0.33
	–0.07	25	16.72	73.33%	0.23	0.16	33	16.56	68.46%	0.24	0.14	67.33%	0.24
	–0.09	36	16.64	85.58%	0.19	0.14	50	16.74	82.57%	0.20	0.14	81.29%	0.20
	-0.11	57	16.75	92.58%	0.18	0.13	80	16.79	90.83%	0.18	0.14	89.91%	0.18
Aiming for 1/12	–0.03	8	7.99	18.55%	0.43	0.32	9	8.26	16.95%	0.49	0.10	16.38%	0.49
	–0.05	9	8.26	35.27%	0.23	0.29	10	8.25	32.22%	0.26	0.14	31.16%	0.26
	–0.07	10	8.16	52.56%	0.16	0.27	12	8.23	49.91%	0.16	0.18	48.56%	0.16
	–0.09	12	8.23	68.60%	0.12	0.24	16	8.5	65.49%	0.13	0.19	63.68%	0.13
	-0.11	15	8.24	79.66%	0.10	0.22	20	8.25	76.07%	0.11	0.20	75.13%	0.11

Table 5: Statistics on sampling error from simulated MNL data

		β_{ttc}				
	equation	-0.03	-0.05	-0.07	-0.09	-0.11
Mean across people of P_D	(8)	28.8%	47.5%	63.1%	75.6%	83.9%
Mean across people of sd of $\widetilde{W}$	(11)	27.652	21.110	16.397	11.751	8.330
Mean across people of sd of L	(12)	6.313	2.341	1.031	0.474	0.231

Table 6: Estimation results from simulated mode (M) - destination (D) case studies

	Setting 1	Setting 2	Setting 3	Setting 4	Setting 5
True β_{ttc}	-0.03	-0.05	-0.07	-0.09	-0.11
True β_{ttPT}	-0.07	-0.12	-0.16	-0.21	-0.26

		Results summarised across ten sampling runs for each setting									
		mean	sd.	mean	sd.	mean	sd.	mean	sd.	mean	sd.
	sample size	8.19	0.04	8.54	0.02	8.35	0.02	8.44	0.02	8.39	0.02
	P_D	0.29	0.00	0.47	0.00	0.63	0.00	0.76	0.00	0.84	0.00
MNL (dest.)	MNL(D) β_{ttc} bias.	-0.0006	0.0002	-0.0004	0.0002	-0.0005	0.0002	-0.0003	0.0002	-0.0005	0.0002
	β_{ttc} bias. t-rat.	-0.86	0.23	-0.45	0.27	-0.48	0.19	-0.24	0.17	-0.29	0.13
	γ bias.	-0.0013	0.0078	0.0149	0.0074	0.0202	0.0043	0.0160	0.0037	0.0132	0.0027
	γ bias. t-rat.	-0.02	0.14	0.34	0.17	0.51	0.11	0.42	0.10	0.34	0.07
	η bias.	0.0035	0.0147	0.0030	0.0246	0.0834	0.0388	-0.0207	0.0479	-0.0266	0.0678
	η bias. t-rat.	0.08	0.23	0.05	0.27	0.67	0.28	-0.04	0.21	-0.03	0.18
MNL (mode – destination)	MNL(M+D)β_{ttc} bias.	-0.0001	0.0001	0.0000	0.0002	0.0006	0.0002	0.0007	0.0002	0.0002	0.0002
	β_{ttc} bias. t-rat.	-0.17	0.19	0.02	0.24	0.59	0.17	0.48	0.15	0.15	0.13
	β_{ttPT} bias.	0.0000	0.0001	0.0014	0.0002	0.0005	0.0002	0.0006	0.0002	0.0012	0.0002
	β_{ttPT} bias. t-rat.	0.03	0.07	0.62	0.07	0.15	0.05	0.18	0.05	0.25	0.04
	γ bias.	0.0113	0.0057	0.0012	0.0064	0.0240	0.0042	0.0133	0.0028	-0.0013	0.0025
	γ bias. t-rat.	0.23	0.11	0.03	0.15	0.61	0.11	0.34	0.07	-0.04	0.07
	η bias.	0.0001	0.0149	-0.0199	0.0183	0.0109	0.0198	-0.0758	0.0224	0.0137	0.0299
	η bias. t-rat.	0.02	0.22	-0.21	0.21	0.15	0.17	-0.39	0.14	0.13	0.14
Nest Logit (mode – destination)	Nested Logit - β_{ttc}	0.0003	0.0002	0.0002	0.0002	0.0000	0.0002	-0.0004	0.0002	0.0001	0.0004
	β_{ttc} bias. t-rat.	0.45	0.22	0.25	0.26	0.02	0.20	-0.27	0.14	0.08	0.20
	β_{ttPT} bias.	-0.0039	0.0003	-0.0009	0.0004	-0.0047	0.0018	-0.0001	0.0003	-0.0036	0.0038
	β_{ttPT} bias. t-rat.	-1.99	0.17	-0.20	0.11	-0.71	0.08	0.09	0.03	-0.22	0.03
	γ bias.	-0.0283	0.0096	-0.0075	0.0090	-0.0141	0.0056	-0.0128	0.0033	0.0008	0.0059
	γ bias. t-rat.	-0.55	0.19	-0.17	0.21	-0.34	0.13	-0.33	0.09	-0.02	0.07
	η bias.	-0.0494	0.0175	-0.0064	0.0258	0.0444	0.0403	0.0065	0.0437	0.0566	0.0576
	η bias. t-rat.	-0.65	0.24	-0.04	0.25	0.33	0.27	0.22	0.18	0.22	0.15
	ϕ bias	-0.0304	0.0021	-0.0151	0.0021	-0.0169	0.0111	-0.0002	0.0008	-0.0004	0.0202
	ϕ bias t-rat.	-1.84	0.14	-0.81	0.10	-0.93	0.12	-0.11	0.03	-0.31	0.03