Sampling of alternatives for spatial choice modelling

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

a: Institute for Transport Studies, University of Leeds
b: RAND Europe

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 modes by which they can travel to the respective destination. Since the calculation of choice probabilities requires consideration of all the alternatives in the choice set, tens of thousands or more, such models can make heavy demand on computer resources, particularly run time, but also potentially storage requirement. This problem increases substantially when making use of more advanced types of models, such as random coefficients models, or systems of models, as in activity-based (AB) modelling. Independently of the specific model structure used, an option to reduce these demands is to use sampling to restrict the number of alternatives actually used in the estimation. In AB models, we model numerous activities for the same person, often involving a destination component with a very large number of alternatives, and interactions between these spatial choices and the activity scheduling component. Given the reliance on standard software, practical solutions to sampling are needed.

McFadden (1978) 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 of alternatives 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 (2010) extend the work of McFadden (1978) to the GEV framework, so that consistent estimates can be obtained for two-level nested logit models, being 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 model. 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 Guevara and Ben-Akiva (2010) for GEV models, but addressing the issues for obtaining consistent parameter estimates for random coefficients models falls beyond the scope of this paper. It seems from recent work by Guevara (2012) that a definitive resolution of the mixed logit issue will be difficult, although the results of a simple approach are promising.

Sampling of alternatives inevitably leads to increased error in parameter estimation, but it is not completely straightforward to calculate the magnitude of this additional error. Nerella and Bhat (2004) 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. For mixed logit models, Nerella and Bhat (2004) show that the results are less accurate, but in this case the estimates are not necessarily (proven to be) consistent and would not usually be used in practice for that reason. Summarising the current state of the literature, we have to conclude that 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; sampling alternatives in mixed logit causes more noise and bias may also be present, but we are not able to say how much of either.

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 travel demand modelling, based 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 (2004) 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 Guevara and Ben-Akiva (2010), 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. This is in clear contrast to residential choice models, where accessibility is of less importance as individuals move home on an infrequent basis. 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 Guevara and Ben-Akiva (2010) 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 reparametrising 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 multinomial logit models (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

In his 1978 paper, McFadden 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 individual 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 exactly 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, equation (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 equation (1). The system then conforms with the Uniform Conditioning (UC) property also defined in McFadden (1978). 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 and will be used in this note. In particular, in modelling destination choice it is clear that nearer alternatives are each very 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 multinomial logit. 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.[1] However, in this case, neither the base MNL nor the sampled version is correct. 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.

2.1 Practical strategies for PC sampling

A simple 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,

[1] That is, the validity of McFadden's theorem depends on behaviour being truly in accordance with the MNL formula. It should also be noted that the theorem applies only to large samples.

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

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

with $K(D)$ independent of j (Ben-Akiva and Lerman, 1985, equations 9.22 and 9.23). Examining the log likelihood equation (1), we see that $K(D)$ cancels out and we are left with

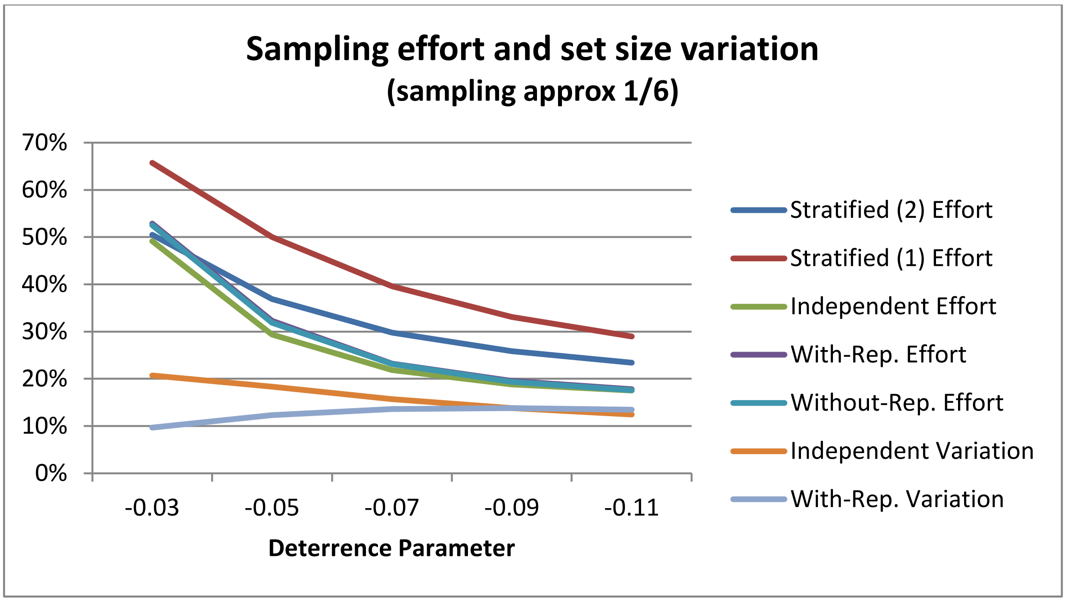
$$L = \log\left(\frac{\exp(V_c - \log q_c)}{\sum_{j \in D} \exp(V_j - \log q_j)}\right) = V_c^M - \log \sum_{j \in D} \exp V_j^M \tag{3}$$

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. A special case of stratified sampling is to treat the whole set of alternatives as one stratum and to sample uniformly from this set without replacement; this leads to the sampling satisfying UC and therefore not requiring correction.

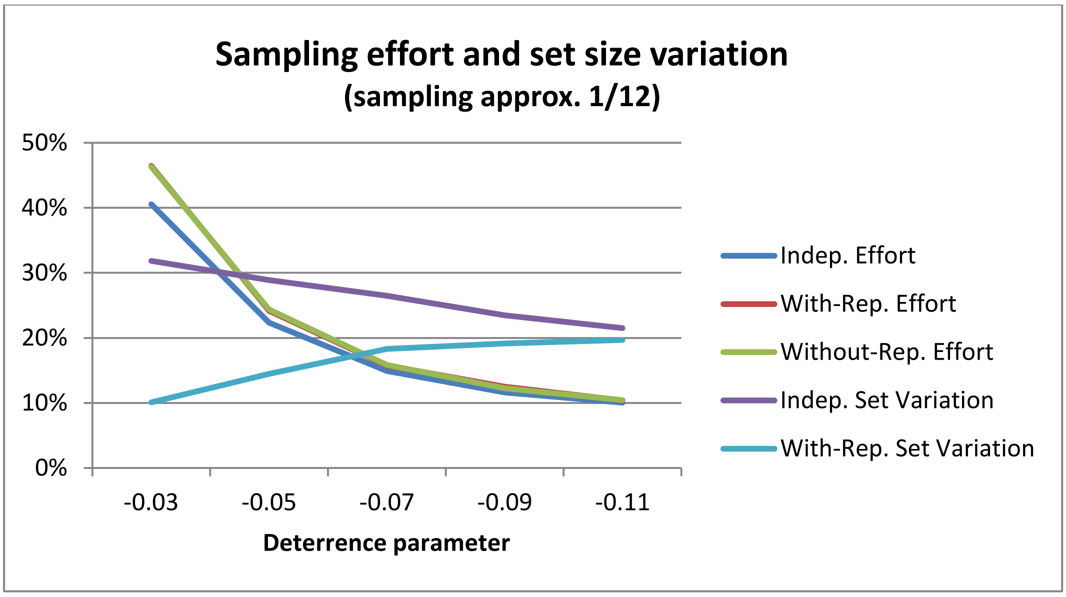
Some aspects of the sampling issues can be tested quite readily by simulation, as illustrated in the graphs below. Details of the simulations we have made are given in the Appendix. The simulations are set in the context of destination choice among zones in a hypothetical study area. The advantage of 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 the deterrence parameter: a larger negative deterrence parameter implies shorter trip lengths or, equivalently, a larger study area relative to a given trip length. Two levels of sampling are tested, the first corresponding to approximately 1 in 6 alternatives, the second to about 1 in 12. The appropriate sampling rate for a particular study will depend on the computational cost and the estimation accuracy required. These issues are considered in the following section, but here we include two levels of sampling to illustrate how the choice of sampling rate affects the choice of sampling protocol.



The five upper lines in the first graph show the relative computation effort required for each of these protocols, shown 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. We see that the differences between several of the sampling approaches in this respect are very small and that all the procedures become more efficient as the deterrence parameter becomes more strongly negative. However, stratified sampling seems to perform generally worse than the other sampling methods. This is not surprising, because stratified sampling usually reduces the correlation of the sample with the predicted probabilities, i.e. the alternatives sampled within a stratum are not necessarily those with the highest choice probabilities. The first stratified sampling approach, in which the sampling is more closely connected to the choice probabilities (see the Appendix), is less inefficient. Independent sampling performs a little better than with-replacement or without-replacement sampling, but these differences are small.

The lower lines in the first graph show the coefficient of variation of the sampled set size in these runs, for those strategies where the sample size varies. Although for small values of the deterrence parameter replacement sampling gives less variation than independent sampling, as expected, but the difference decreases substantially as the parameter increases and for the largest value independent sampling gives a *less* variable set size, contrary to the expectation of Ben-Akiva and Lerman (1985).

As indicated in the Appendix, with the approximate 1/6 sampling rate of these runs, the coverage of expected choices runs from about 30% with a deterrence factor of –0.03 to about 90% with a deterrence parameter of –0.11. Clearly, it is not possible for the analyst to control the deterrence parameter, which is a function of the behaviour and study area, but it is possible to control the sampling rate. To see the effect of changing this rate, simulated samples were made with a sampling rate of about 1/12, i.e. a little over 8%. In this case, the coverage of expected choices runs from about 20% to 80% as the deterrence parameter varies. The relative performance of the various sampling protocols is shown in the graph below. Stratified sampling was not tested at this level, because of its poor performance in the previous tests.



In these tests, we see again that the difference in relative effort for the three candidate sampling protocols is very slight[2], with independent sampling again the best by a small margin, and the relative effort is substantially less than in the previous tests across all values of the deterrence parameter, i.e. more than half of the coverage is achieved when we sample half as many alternatives with these protocols. However, the relative variation in set size is nearly doubled for both independent and with-replacement sampling, while the advantage in this respect of with-replacement sampling is retained for the full range of deterrence parameter values tested.

It is also important to note that the efficiency of PC sampling, relating sampling probabilities to the importance (i.e. approximate choice probability) of each alternative, is greater when the link to travel accessibility is stronger. The increase of efficiency as the deterrence parameter increases shows that for large study areas and/or short trips the gain from PC sampling also increases.

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 believed to be very important, while the simplicity of calculation for independent sampling is very helpful in the calculations we need to make in the remainder of the paper. For this reason we use independent sampling in the rest of the work described here.

2.2 Sampling error in PC sampling

The work of Guevara and Ben-Akiva (2010), 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. It does not seem to have been noted previously in the literature that this result can be applied to an MNL model, which of course is a particular GEV model. 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

2 With-replacement and without-replacement sampling give indistinguishable results.

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:

$$P_D = \frac{\sum_{j \in D} \exp(V_j)}{\sum_{j \in C} \exp(V_j)} \tag{4}$$

This approach is taken by Miller *et al.* (2007) in the context of sampling alternatives for application. It is intuitively clear that as the fraction increases then the approximation will generally improve. After all, when the fraction reaches 1 there is no approximation. 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

$$L = \log p_c = V_c - \log \sum_{j \in C} \exp(V_j) \tag{5}$$

The basic objective of sampling alternatives is to save time by not evaluating the logsum in (5) 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 and q_j is the expansion factor for each j. That is, we approximate

$$L \cong V_c - \log \widetilde{W} = V_c - \log \sum_{j \in D} \frac{\exp(V_j)}{q_j} \tag{6}$$

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

$$\mathrm{var}(\widetilde{W}) = \sum_{j \in C} \left(\frac{w_j}{q_j}\right)^2 q_j \left(1 - q_j\right) = \sum_{j \in C} {w_j}^2 \left(\frac{1}{q_j} - 1\right) \tag{7}$$

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 (7) is minimised when $q_j = k.w_j$ for a constant k (see also Hammersley and Handscomb, 1964). 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 reasonable[3].

[3] 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 (7) could be developed, though it would not necessarily be so simple. For other sampling protocols the formulae are likely to be even more complicated.

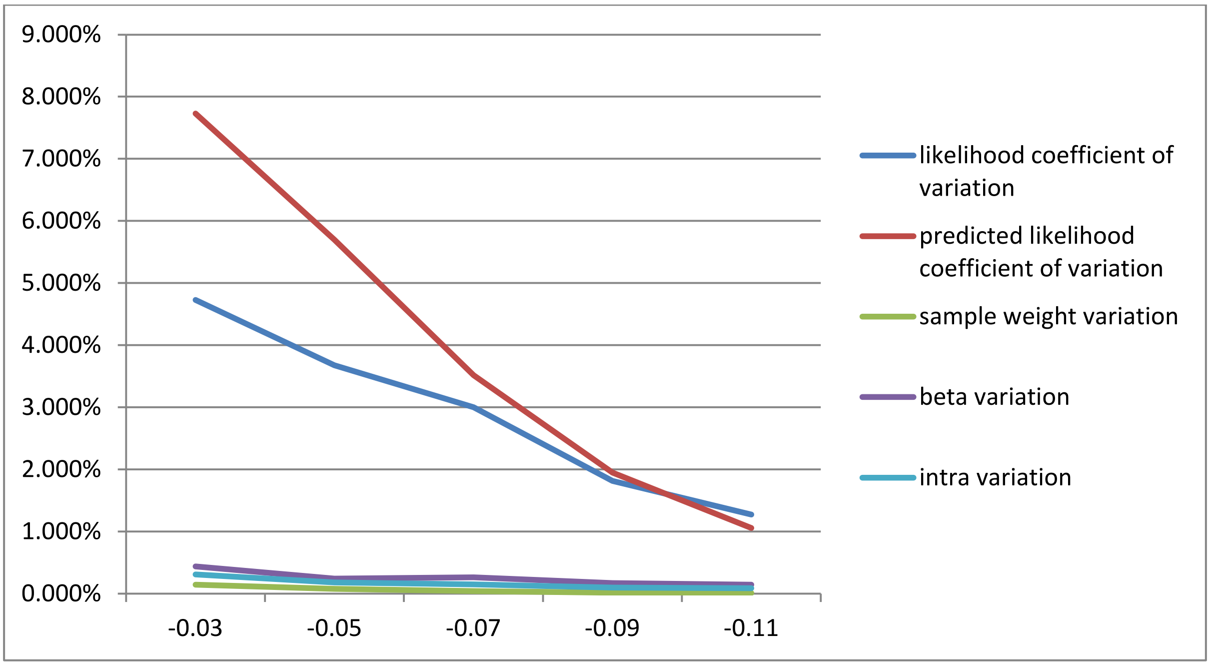
The error in the likelihood-contribution calculation (5), to which $\widetilde{W}$ contributes the sampling error, can be estimated as

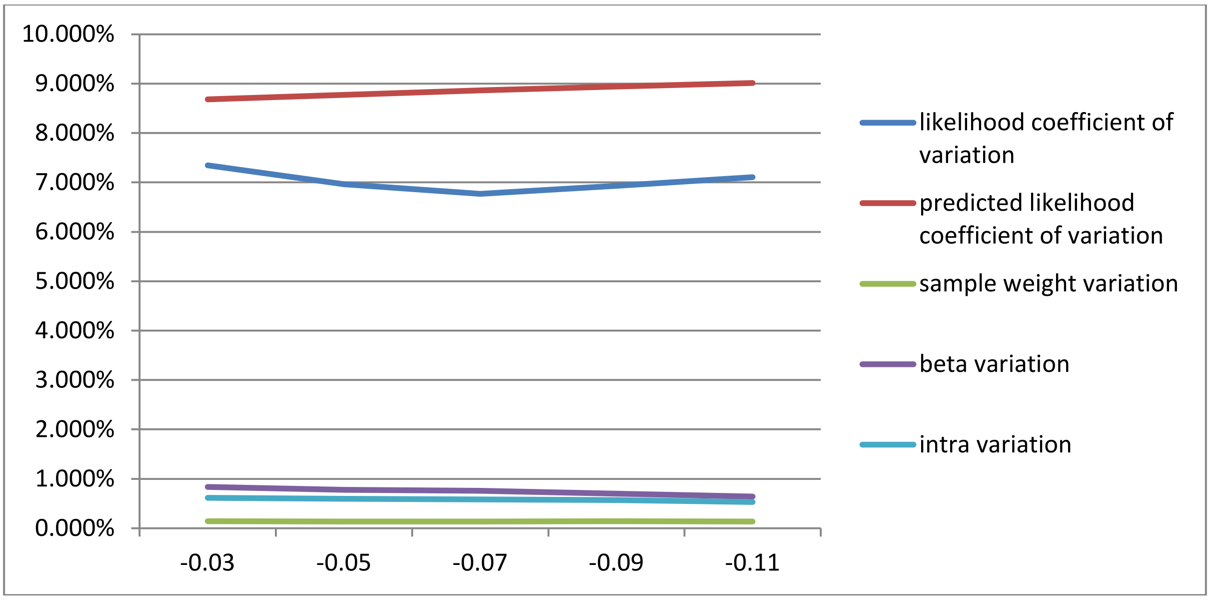
$$\mathrm{var}(L) \cong \left(\frac{\partial L}{\partial \widetilde{W}}\right)^2 \mathrm{var}(\widetilde{W}) = \frac{\mathrm{var}(\widetilde{W})}{\widetilde{W}^2} \cong \frac{\mathrm{var}(\widetilde{W})}{W^2} \qquad (8)$$

The final approximation follows because $\widetilde{W}$ is an estimate of W and, using (7) to obtain $\mathrm{var}(\widetilde{W})$, we are then able to make a calculation of the expected error variance (8) 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. equations (7) and (8), can be made in advance, so that a sample size can be set to obtain an appropriate balance between likelihood error and sample effort.

In order to check that equations (7) and (8) give a good expectation of the likelihood variation to be found in practice with models estimated on samples, a series of simulation runs was made. These runs used the same independent sampling approach described in the previous section, but adding the chosen alternative where necessary to make estimation possible. The results are shown in the following graphs.





In the first graph we see that the prediction exceeds the output likelihood substantially for low values of β, but this is corrected for higher values, whereas in the second series of simulations, with lower overall samples, the difference is more moderate but is consistent across the range. While some noise is to be expected from these runs involving random sampling, the pattern observed here is difficult to explain.

Further work is also needed to determine how the error in likelihood calculation (7) 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 $\mathrm{var}(L)$, the latter calculated using equations (7) and (8), to obtain some insight into error.

3. Sampling in GEV models

Our investigation takes the work of Guevara and Ben-Akiva (2010) as a starting point but first we discuss briefly other papers in this area.

3.1 Previous research

It is important to distinguish between the sampling of alternatives, which is the focus of the current work, and the 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, which often mention papers on sampling observations. For example, Koppelman and Garrow (2005) do not discuss sampling alternatives at all. Another paper that is mentioned in reviews is Mabit and Fosgerau (2006), but again this does not mention the sampling of alternatives.

Bierlaire *et al*. (2008) 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 Guevara and Ben-Akiva (2010). However, the latter work is more complete and more directly focussed on our topic of interest and we shall base our discussion on those publications.

Frejinger *et al*. (2009) do not deal with models beyond MNL except through the 'path size' correction and that the PC correction is therefore sufficient for their work. Similarly, Train (2009) does not go beyond the results given in McFadden (1978).

Lee and Waddell (2010) 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

$$V_m = \left(\frac{1}{\mu}\right) \log \left(\sum_{i \in m} \left(\frac{1}{R}\right) \exp(\mu V_i) \right) \tag{8}$$

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, this is different from McFadden's PC, so that it appears that the Lee and Waddell procedure is incorrect. Simple simulations confirm that a bias is introduced.

3.2 Guevara and Ben-Akiva work

Guevara and Ben-Akiva (2010, abbreviated as GBA) give the theorem that consistent estimation of a GEV[4] model based on a sample of alternatives D can be achieved by a correction of the logit utility function

$$V_i^* = V_i + \log G_i(D^*) + \log \pi(D|i) \tag{9}$$

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, 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 (1985) show that equation (10) can be used (without sampling, i.e. without the π term) to represent any GEV model, so that the GBA theorem using (10) 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:

[4] GBA use the term MEV to describe the models introduced by McFadden as GEV (also in the remarkable 1978 paper).

$$\log G_i = \left(\frac{\mu}{\mu_{m(i)}} - 1\right) logsum(m(i)) + \log\mu + \left(\mu_{m(i)} - 1\right)V_i \qquad (10)$$

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 estimated as otherwise we need to make calculations for all the alternatives, defeating the objective of saving calculation time. The estimator GBA propose is, for the logsum for nest m:

$$logsum(m) \approx \log \sum_{j\in D^*(m)} \frac{\tilde{n}_j}{E(j)} \exp\left(\mu_m V_j\right) \qquad (11)$$

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 recommend two alternative procedures[5].

1. Using separate sampling procedures, such that the sampling for D^* for the logsum approximation does not depend on the chosen alternative, works well in the GBA simulations and does not require iterative estimation.
2. 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. This iterative procedure also works well in the GBA simulations.

Clearly, procedure 1 without iteration is more convenient in practice. Further, it appears that in existing software procedure 1 is also easier to implement. Finally, GBA give no guarantee that the iterative process 2 converges, although no problems are reported from their tests.

If we apply independent sampling for D^*, $\tilde{n}_j = 1$ for $j \in D^*(m)$ and $E(j) = q_j$, so that equation (11) can be written

$$logsum(m) \approx \log \sum_{j\in D^*(m)} \exp\left(\mu_m V_j - \log q_j\right) \qquad (12)$$

The GBA equations (11) and (12) are written for a tree logit specification that is normalised at the top, as used in Ben-Akiva & Lerman (1985) and implemented in BIOGEME (Bierlaire, 2005). This is the specification also sometimes referred to as RU2 – see Hensher et al. (2005). For practical implementation it is easier to use the version which uses normalisation at the bottom, as in Train (2009), referred to as RU1 by Hensher et al., (2005). For ease of estimation in large scale models, a *non-normalised* version of this model can be used, as implemented in ALOGIT, where consistency with utility maximisation is ensured through an equality constraint between structural parameters on a given level. We have:

[5] We are grateful to Angelo Guevara for clarifying these points.

$$\mu_m = 1, \text{ for all } m \quad \text{and, to simplify further,} \quad \mu = \phi + 1 \tag{13}$$

which gives the much simpler equation, replacing (10):

$$\log G_i = \phi . logsum(m) + \log(\phi + 1) \tag{14}$$

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[6], and using the brief notation $V_j^M = V_j - \log q_j$ introduced in equation (3), noting that q_j is a constant in the estimation process, we can implement equation (9) as

$$V_i^* = V_i^M + \phi . \log \sum_{j \in D^*(m(i))} \exp V_j^M \tag{15}$$

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.

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.

3.3 Set up for practical testing

The practical tests reported in the following section of the paper 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 equation (16), 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 proper investigation of mode and destination choice should investigate both these possibilities. The practical testing in this paper, however, investigates only the case of modes above destinations, where the correction is required.

For the implementation of equation (16), 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 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.

[6] adding the chosen alternative if it is not already selected by the sampling procedure.

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.

References

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).

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

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.

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

Guevara, C.A. (2012), presentation at International Association of Travel Behaviour Research, Toronto.

Guevara, C.A. and Ben-Akiva, M. (2010), Sampling of alternatives in multivariate extreme value (MEV) models, WCTR, Lisbon.

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

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

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

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.

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

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.

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.

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

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

Appendix: Details of Simulations

A1: Independent and with and without-replacement sampling

These simulations are based on 10000 draws of sets of destinations from a total set of 100. Destinations are located at a 'distance' from the origin of $d_j = 10\sqrt{j}$, with j the destination number, the square root function giving a representation of roughly uniform distribution of destinations in space. Each destination is attributed a utility function of

$$V_j = \beta . d_j + \gamma . \delta_1$$

where δ_1 indicates an 'intrazonal' trip, i.e. with destination in zone 1;
β, γ are the assumed parameters of the model.

In these models we set $\gamma = 1$ and β is set to scale the impact of distance. In effect the variation of β models the variation of the average trip length relative to the size of the study area. Two sets of samples were drawn, each with varying values of β, one sampling approximately 1 in 6 destinations, the other sampling approximately 1 in 12.

Independent sampling is undertaken with $q_j = f . \exp V_j / \sum \exp V_k$ and f set to achieve a roughly uniform sample size.

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.

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.

The settings of β, γ and $\tilde{J}$, the sample sizes achieved and the average coverage are shown in the tables below.

Series 1: Sample Rate about 1:6

	Independent sampling			Replacement sampling			Without replacement
β	f	average sample	average coverage	$\tilde{J}$	average sample	average coverage	average coverage
–0.03	17	16.13	32.8%	23	16.58	31.3%	30.4%
–0.05	19	16.08	54.8%	32	16.64	51.5%	50.2%
–0.07	25	16.32	74.7%	50	16.17	69.6%	69.2%
–0.09	36	16.22	86.3%	50	16.27	83.3%	82.7%
-0.11	57	16.27	92.9%	80	16.24	91.2%	90.8%

Series 2: Sample Rate about 1:12

	Independent sampling			Replacement sampling			Without replacement
β	f	average sample	average coverage	$\tilde{J}$	average sample	average coverage	average coverage
–0.03	8	7.98	19.7%	9	8.23	17.7%	17.3%
–0.05	9	8.21	36.8%	10	8.21	33.7%	32.9%
–0.07	10	8.01	53.8%	12	8.01	51.1%	50.4%
–0.09	12	8.09	69.7%	16	8.09	66.5%	65.3%
-0.11	15	8.05	80.3%	20	8.05	76.7%	76.5%

To avoid dependence on the precise sample sizes, a measure of ‘Effort’ was devised and shown in the graphs in the main text. This measure is calculated by

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

The measure of variation shown in those graphs is

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

A.2 Stratified sampling

Stratified sampling was done only at the level of 1 in 6 destinations.

Stratified sampling series 1 was done by creating 4 strata. Alternative 1 was always selected and the other three strata consisted respectively of the following ranges of alternatives: 2-10; 11-30; 31-100. From each block respectively 4, 5 and 6 alternatives were sampled without replacement. For stratified sampling strategy 2 we increased the number of strata to 6. Still the first alternative was always selected. The other strata comprised respectively alternatives 2-20; 21-40; 41-60; 61-80; 81-100. From each of these strata 3 alternatives were sampled without replacement. Both sampling strategies resulted in a constant number of 16 alternatives. Note that the numbering of the alternatives means that the choice probabilities decline with the alternative index number. Stratified sampling series 1 retains some correlation of sampling probability with choice probability, but this is reduced in series 2.