STRATEGIC SAMPLING FOR LARGE CHOICE SETS IN ESTIMATION AND APPLICATION

Jason D. Lemp
Cambridge Systematics, Inc.
9015 Mountain Ridge, Suite 210
Austin, TX 78759
jlemp@camsys.com

Kara M. Kockelman
(Corresponding author)
Professor and William J. Murray Jr. Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin – 6.9 E. Cockrell Jr. Hall Austin,
TX 78712-1076
kkockelm@mail.utexas.edu
Phone: 512-471-0210 & FAX: 512-475-8744

The following paper is a pre-print and the final publication can be found in
Transportation Research Part A, 46 (2012):602-613, 2012.

ABSTRACT

Many discrete choice contexts in transportation deal with large choice sets, including destination, route, and vehicle choices. Model estimation with large numbers of alternatives remains computationally expensive. In the context of the multinomial logit (MNL) model, limiting the number of alternatives in estimation by simple random sampling (SRS) yields consistent parameter estimates, but estimator efficiency suffers. In the context of more general models, such as the mixed MNL, limiting the number of alternatives via SRS yields biased parameter estimates. In this paper, a new, strategic sampling scheme is introduced, which draws alternatives in proportion to updated choice-probability estimates. Since such probabilities are not known *a priori*, the first iteration uses SRS among all available alternatives. The sampling scheme is implemented here for a variety of simulated MNL and mixed-MNL data sets, with results suggesting that the new sampling scheme provides substantial efficiency benefits. Thanks to reductions in estimation error, parameter estimates are more accurate, on average. Moreover, in the mixed MNL case, where SRS produces biased estimates (due to the violation of the independence of irrelevant alternatives property), the new sampling scheme appears to effectively eliminate such biases. Finally, it appears that only a single iteration of the new strategy (following the initialization step using SRS) is needed to deliver the strategy's maximum efficiency gains.

INTRODUCTION

A variety of discrete choice models are widely used for transportation applications thanks to their ability to reflect key determinants of individuals' choice behaviors while facilitating model estimation and/or providing a defensible behavioral basis (i.e., random-utility maximization). In many decision contexts, choice sets can be very large. For instance, residential location and trip-destination choices are often modeled with choice alternatives represented by individual dwelling units, parcels or a region's traffic analysis zones (TAZs). In large metropolitan regions there are regularly thousands of TAZs and potentially hundreds of thousands of available homes. Other large-choice-set examples include vehicle acquisition, route selection, activity diaries, and departure times. In the case of travel tours, where two timing alternatives are chosen simultaneously (one each for the outbound and inbound travel legs), the total number of alternatives can grow very large very quickly, particularly as smaller time intervals are sought. No matter the choice context, consideration of the complete alternative choice set in model estimation with a large number of alternatives can be computationally burdensome and sometimes prohibitive. This work develops a sampling scheme for such choice contexts that provides efficient model parameter estimates.

When choice sets are large, researchers typically sample alternatives at random from the complete set. For instance, one may choose a simple random sample of 19 alternatives from a 1000-alternative choice set, along with the chosen alternative, for each observation. (See, e.g., Pozsgay and Bhat 2001, Sermons and Koppelman 2001, Nerella and Bhat 2004, among many others.) McFadden (1978) showed that such sampling satisfies a uniform conditioning property, thereby providing consistent parameter estimates – as long as the model meets the independence of irrelevant alternatives (IIA) property (as in the case of the multinomial logit [MNL], but not in mixed MNL, nested logit and other contexts). However, model performance and estimator efficiency are negatively impacted by smaller sample sizes, and Nerella and Bhat (2004) recommend that one-eighth of the full choice set be used, at a minimum, in the case of the MNL. This can be computationally challenging when the choice set numbers reach into the thousands or more, as they often do.

Others have turned to importance sampling to generate choice sets for model estimation (see, e.g., Ben-Akiva and Lerman 1985, Ben-Akiva and Bowman 1998, Frejinger et al. 2009) and application (see, e.g., Ben-Akiva and Watanatada 1981). Importance sampling more often selects alternatives that are more likely to be utility-maximizing (based on the analyst's prior beliefs, rather than a rigorous approach), while providing consistent parameter estimates, as long as the likelihood function reflects the biased sampling technique (via appropriate weighting of probabilities) and the positive conditioning property is met (McFadden 1978)[1]. By including alternatives that are (thought to be) more likely to be chosen by the decision-maker, more information is provided to the model and parameter estimates should be more efficient (Ben-Akiva and Lerman 1985). While the intuition behind this notion is clear, there is no obvious path for generating such choice set probabilities *a priori*. Previous studies have relied on intuition about the data to determine alternatives that are more likely. For instance, in a destination choice study, one may assume the likelihood of an alternative to be proportional to the number of jobs in a zone and inversely proportional to the distance to the zone. However, such intuition is not necessarily available in all choice contexts; and, even when it is, one cannot ensure the sampling strategy provides more efficient estimates. Moreover, such sampling procedures cannot offer consistent estimates for models other than the MNL. For other GEV-based specifications, Bierlaire et al. (2008) recently developed a weighted conditional maximum likelihood (WCML) estimator. While their results focus on choice-based sampling of observational units (rather than alternatives from the full choice set), Bierlaire et al. also demonstrated the applicability to alternatives-sampling frameworks.

[1] If some choice set is drawn conditional on the actual choice *i*, the positive conditioning property implies that the probability of drawing the same choice set conditional on any other alternative within that choice set be positive.

It turns out that one can generate a much more efficient sampling scheme by sampling alternatives in proportion to the exponential transformation of their true systematic utilities. Of course, the values of such systematic utilities cannot be inferred until the model is estimated. Here a new sampling scheme is proposed. The method iteratively develops probabilities for alternative inclusion in the subset of choices based on the current model's estimates of model coefficients. At each iteration, alternative subsets are generated for each observation and the likelihood maximized. Subset choice probabilities are updated and samples re-generated. The sampling scheme is specifically designed to utilize as much information as is available in the data, resulting in more efficient parameter estimates with very little added effort or expertise.

Simulated data sets are primarily used here to demonstrate the method and compare the efficiency of estimates and accuracy of predictions across different data structures and different sampling schemes for MNL and mixed MNL model types. The mixed MNL model was chosen since it represents a rather general model structure (McFadden and Train 2000) in comparison to many other closed-form GEV models (such as the nested [see, e.g., Williams 1977 and McFadden 1978] and cross-nested logits [see, e.g., Vovsha 1997 and Wen and Koppelman 2001]). In addition, such closed form models present some sampling concerns in that without additional sampling restrictions, it would be possible to draw subsets of alternatives where certain nests are not represented. The following section details the methodology used.

METHODOLOGY

The MNL model's probability of individual i choosing alternative k takes the following form:

$$P_{ik} = \frac{e^{\beta X_{ik}}}{\sum_{j \in C} e^{\beta X_{ij}}} \tag{1}$$

Here β is a parameter vector and X_{ij} is a vector of variables for individual i and alternative j.

In the case where the choice set, C, is very large, McFadden (1978) showed that one can sample alternatives from that choice set and obtain consistent parameter estimates, so long as the sampling scheme meets the positive conditioning property. This requires that if $j \in D_i \subseteq C$ and $\pi(D_i|k) > 0$ (where $\pi(D_i|k)$ is the probability of generating alternative set D_i given actual choice k under the sampling scheme), then $\pi(D_i|j) > 0$. In other words, to meet the positive conditioning property, the sampling scheme must be such that the probability of drawing the subset of alternatives is positive regardless of the actual chosen alternative within that subset. McFadden (1978) also shows that MNL choice probabilities in model estimation must be adjusted as follows:

$$P_{ik} = \frac{e^{\beta X_{ik} + \ln\left(\pi(D_i|k)\right)}}{\sum_{j \in D_i} e^{\beta X_{ij} + \ln\left(\pi(D_i|j)\right)}} \tag{2}$$

However, if SRS (of choice sets, for log likelihood estimation) is used (in which case the probability of generating any choice set of non-chosen alternatives is equal), McFadden's (1978) uniform conditioning property is met, negating the need for choice probability adjustments. In such cases, the MNL can be estimated using the typical MNL likelihoods, but with the sampled choice sets. For many years, researchers have employed this simple sampling scheme to estimate models with large choice sets in a variety of contexts (see, e.g., Ben-Akiva and Bowman 1998, Bhat et al. 1998, Sermons and Koppelman 2001, and Lemp et al. 2007), since it requires no modification of the likelihood function.

This paper explores this SRS strategy and compares it the strategic and iterative sampling procedure proposed above. In the first iteration of this strategic process, SRS of alternatives is used. In each iteration thereafter, alternative inclusion probabilities are set equal to the MNL choice probabilities derived from the previous iteration's parameter estimates. The likelihood function in the second and any

later iterations is updated to include the probability of choice set formation (using weights on alternatives that are proportional to the prior iteration's choice probability estimates).

Mixed Multinomial Logit

As noted earlier, the MMNL allows for very general choice specifications. However, it adds some complexity. Its choice probabilities can be written as follows (see, e.g., McFadden and Train 2000 and Train 2009):

$$P_{ik} = \int_{-\infty}^{\infty} Q_{ik}(\beta)\, f(\beta|\theta) d\beta \tag{3}$$

$$Q_{ik}(\beta) = \frac{e^{\beta X_{ik}}}{\sum_{j \in C} e^{\beta X_{ij}}} \tag{4}$$

where $f(\beta|\theta)$ is the distribution of the parameter vector β and the term $Q_{ik}(\beta)$ represents standard MNL choice probabilities.

In the MMNL, one or more of the parameters in β is assumed to vary over the population (rather than being fixed) with distribution given by $f(\beta|\theta)$[2]. Thus, instead of estimating β itself, the parameters of β's distribution (i.e., θ) are estimated.

Since the MMNL relaxes the IIA assumption (Train 2009), the SRS-based choice probabilities do not simplify in any meaningful way in the likelihood function. That is, Eq. 3 cannot be rewritten using a subset of all choice alternatives in the same way Eq. 1 can be rewritten as Eq. 2 for a subset of standard MNL alternatives. This is due to the integral in the choice probabilities. In the case of SRS, Nerella and Bhat (2004) suggest using an approximation of the true choice probabilities, as follows:[3]

$$Q_{ik}(\beta) \approx \frac{e^{\beta X_{ik}}}{\sum_{j \in D_i} e^{\beta X_{ij}}} \tag{5}$$

Eq. 5's approximation is used here to compute choice probabilities under SRS. However, when the sampling scheme is employed and alternatives are drawn with non-uniform probabilities (proportional to the alternative's choice probabilities from the previous iteration's model coefficients), Eq. 5's approximation would be poor. Instead, choice probabilities are better approximated by replacing Eq. 5's right side with Eq. 2's right side, as follows:

$$Q_{ik}(\beta) \approx \frac{e^{\beta X_{ik} + \ln\left(\pi(D_i|k)\right)}}{\sum_{j \in D_i} e^{\beta X_{ij} + \ln\left(\pi(D_i|j)\right)}} \tag{6}$$

In general, when Eqs. 5 and 6 are used in the MMNL's likelihood, one cannot expect statistically consistent estimators to emerge. However, as the sample size (of sampled alternatives) grows, one expects the approximations to improve by more closely matching estimates one would obtain using the full choice set. In this paper, the reasonableness of these approximations is examined through empirical analysis under a variety of sampling circumstances, both simple and strategic.

EXPERIMENTAL DESIGN

[2] This specification is commonly referred to as the random parameters logit (Train 2009). Another form of the mixed logit is the error components logit, in which error components of the random utility equation are assumed to be stochastic. Mathematically, the two specifications are equivalent (Train 2009).

[3] Nerella and Bhat (2004) found this approximation to perform reasonably, though they caution the use of sample choice sets too small in comparison to total choice sets.

Multinomial Logit Model

A variety of scenarios were examined for both the MNL and MMNL cases. In the case of the MNL, three different simulated data sets were first generated: one with $N = 200$ records, another with $N = 1{,}000$ records, and another with $N = 4{,}000$ records (such that the smaller sets are subsets of the larger sets). This was done to investigate the effect of sample size on sample alternatives estimation efficiency. In addition, for each sample size, three separate data sets with $J = 50$, 500, and 2,000 alternatives were generated. Four explanatory variables were simulated for each data set, all coming from the standard-normal distribution. In each data set, the variable means for the first 50 alternatives (this represents all alternatives in the 50 alternative case) were assumed to be 2, 3, and 4 for the first three variables. For all other alternatives, the means of these variables were assumed to be 1. The mean of the fourth variable (across all alternatives) was assumed to be 1. In order to simulate response data, parameter coefficients on each of these variables were assumed to be 0.5, 0.3, 0.1, and -1.0. As a result of such numeric assumptions, the first 50 alternatives represent more appealing options, on average.[4]

While it is not necessary to simulate data here, to illustrate the sampling scheme, the use of simulated data (in lieu of real data) allows one to vary the number of observations and choice-set alternatives. Another benefit of using simulated data for algorithm evaluation is that the parameter coefficient values on explanatory variables can be varied in systematic ways, for purposeful changes in estimators' precision.

After simulating the data sets described above, the analysis proceeded by implementing the new strategic sampling process. In the first step, the set of sampled alternatives was drawn from a uniform distribution (i.e., SRS was used) and the model's parameters estimated (using R programming language). The strategic (i.e., alternative-weighted) sampling scheme was then implemented for a series of five iterations, each time using parameter estimates from the previous iteration's weighted-likelihood maximization to generate probabilities for inclusion in the new sampled subset. For the 50-alternative dataset, alternative sample sizes of $S = 5$ (10%), $S = 10$ (20%), and $S = 25$ (50%) were used to identify the fraction of overall alternatives needed to reasonably estimate the models parameters. For the 500-alternative dataset, sample sizes of $S = 5$ (1%), 10 (2%), and 50 (10%) were used; and, for the 2,000-alternative dataset, sample sizes of $S = 10$ (0.5%), 20 (1%), and 50 (2.5%) were used.

The sampling scheme proceeds by first generating a SRS sample of alternatives to estimate the model. The strategic sampling process is then iterated five times. Lastly, the whole process was repeated 10 times in order to compute evaluation measures (described later).

In addition to these sampling-based estimation strategies, complete-sample MNL models were estimated for each data set, thus providing another basis for comparison and evaluation of results. And several tests were run using an actual (non-synthetic) data set, for destination choice, further supporting the proposed sampling design.

Mixed Multinomial Logit

For analysis of the sampling scheme for the mixed MNL model, a similar procedure was used, but fewer data sets were generated (due to the much longer estimation times required). Instead of examining 3 separate data sets with 50, 500, and 2,000 alternatives, a single data set, with 500 alternatives, was used. In addition, 200- and 1,000-observation cases were explored. The process of generating four explanatory variables for these data sets remained as above. In order to generate simulated choices, the first two

[4] In the J = 50 case, all alternatives are equally likely, on average. When J = 500 and when J = 2,000, the first 50 alternatives are about 4 times more likely to be selected than the others, on average.

variables were assumed to have fixed coefficients of 0.5 and 0.3, while the later two had random coefficients, with means of 0.1 and -1.0, and standard deviations of 0.4 and 0.7, respectively.[5]

The choice-set sampling process proceeded the same as for the MNL. Sample sizes of S = 5, 10, and 50 alternatives were used, with SRS in the first iteration and 5 subsequent iterations of the strategic sampling process. The whole process was repeated 10 times for each case in order to generate statistics used in evaluation.

Due to the MMNL's intractable likelihood (Eq. 3), standard maximum likelihood procedures cannot be used to estimate the model. Instead, maximum simulated likelihood methods were used here (Train 2009), in R programming language. With different random seeds for initialization, 100 Halton draws with primes of 2 and 3 were generated for each observation. These are bivariate draws in the unit square and are used to approximate Eq. 3's integral (over the model's random coefficients).

Alternative Sampling

In theory, the methods discussed above for sampling alternatives from the full choice set could be used with or without replacement. If sampling with replacement, one may repeat alternatives in the sampled subset, but each repeat offers no additional information to the model, thus somewhat compromising estimator efficiency. Sampling with*out* replacement is ideal, and is presumably the standard approach for large-choice-set logit-model calibrations.

While sampling *without* replacement from a multinomial distribution is not difficult, computing the probabilities of each sampled subset can be very computationally expensive with large S (the number of sampled alternatives). In contrast, sampling with replacement yields simple-to-compute probabilities. (One simply takes the product of all choice probabilities of the sampled alternatives.) Thus, SRS *with* replacement is used here. An examination of how this impacts estimator efficiency may prove a meaningful area for future research. (Of course, the less distinctive the alternatives, in terms of their systematic and thus random utilities, the more likely one will end up with repeats, thereby experiencing presumably notable efficiency reductions.)

Evaluation Criteria

The criterion used here -- to evaluate the strategic-sampling technique's performance (versus the one-shot standard SRS method and the full-choice-set approach) -- is a measure of *bias* in parameter estimates. This mean absolute error (MAE) is computed as follows:

$$MAE_r = \text{abs}\left(\frac{\frac{1}{Q}(\sum_q \hat{\theta}_{rq}) - \theta_r}{\theta_r}\right) \quad (7)$$

where θ_r is the full-choice-set estimate of parameter r, Q is the number of times the sampling process is repeated (equal to 10 in this instance), and $\hat{\theta}_{rq}$ is the estimate of parameter r using either strategic or simple sampling of alternatives for the q^{th} repetition.

This measure was used to evaluate bias in primary parameters, as well as bias in the standard errors obtained for each of these. The latter is important as a measure of estimator-efficiency impacts.

For the MNL model, the MAEs of each parameter were averaged to generate a single, representative bias term. For the MMNL model, MAEs of the fixed parameters, the means of random parameters, and the variances of random parameters were averaged separately, since these represent rather different facets of

[5] These parameter values result in the first 50 alternatives being approximately 5 times more likely than the remaining 450 alternatives.

this more complex model and thus merit added attention. In either setting, the scale of the chosen parameter values are very similar (ranging from 0 to 1), so size effects are not of great concern when aggregating/averaging MAE values in the MNL case.

SIMULATION RESULTS

As discussed previously, the strategic sampling scheme developed in this paper first estimated the model using SRS of alternatives. This first set of parameter estimates was used to estimate the choice probabilities of each alternative, for each observational unit, and then update (via probability-weighted sampling) each unit's sample of alternatives. The model was estimated again, and the improved parameter estimates used in later rounds. Here, the strategic re-sampling was repeated 5 times for each of the 30 cases (24 MNL cases and 6 MMNL cases). Thus, in total 6 parameter sets were generated for each case, with the last of these presumably being the most improved and robust. The whole process was repeated 10 times for each of the 30 cases to generate the evaluation measures described above.

MNL Model Results

As noted earlier, MNL data sets with N = 200, 1,000, and 4,000 were used, with choice set sizes ranging from J = 50 to 2,000, for a total of eight data sets. Model parameters were estimated using small (5 to 10), medium (10 to 20), and large (25 to 50) sample sizes across alternatives.

Figure 1 details the MNL cases' simulated results, as they relate to the (known) parameter values. Each plot in Figure 1 represents a different combination of data set size (N) and complete choice set size (J). Within each plot, the MAEs for the small, medium, and large sampled numbers of alternatives (S) for each of the six strategically-sampled iterations is displayed.

Not surprisingly, after the first iteration, estimator error reductions generally appear (though not in all cases, due to standard simulation errors). As one would expect (and as discussed below) MAE reductions are in concert with the size of the standard error of each estimators (rather than the values of parameter estimates themselves. Interestingly, there does not seem to be much error reduction after the second iteration, on average. That is, additional iterations past the second do not offer much return, as errors essentially oscillate from iterations 2 to 6.

The greatest MAE reductions typically occur when N, J, and S are small. And they rise with choice set size, as one would expect (since it is much easier to select non-competitive SRS samples in such contexts). When compared to the first iteration's SRS-based MAE value, the average MAE over subsequent (strategically-sampled) iterations (2 through 6, which are largely similar) range from roughly 40% to 80% lower (with just two anomalies, out of the 24, as shown in Table 1. Thus, strategic sampling may be expected to provide estimates that are 40% to 80% closer to full-sample-set values, as compared to SRS of alternative choice sets. Meaningfully, this result appears to hold regardless of the number of iterations; so a single iteration of strategic sampling (or two iterations total: SRS, followed by strategic sampling) can be recommended here.

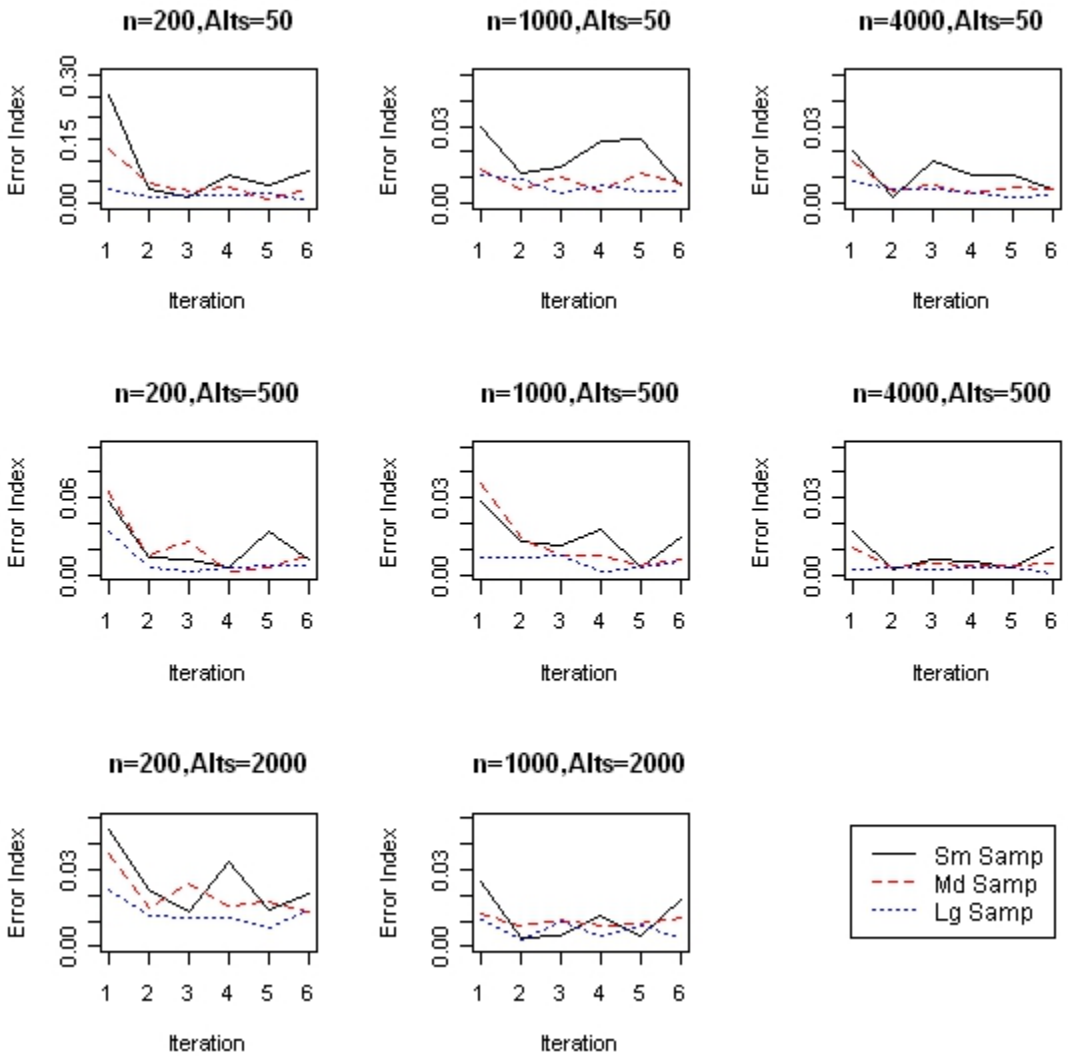

Figure 1: MAEs of Parameter Estimates for the MNL Model (where Iteration 1 relies on SRS, and Iterations 2 through 6 rely on strategic sampling)

Table 1: MNL Model Estimates' MAE Reductions, from Iteration 1 to 2-6

Total Alts. (*J*)	Samp. Alts. (*S*)	#Obs. (*N*)	Model Coefficient Estimates			Model Standard Error Estimates		
			MAE, Iter. 1	Average MAE, Iter. 2 to 6	Reduction in MAE	MAE, Iter. 1	MAE, Iter. 2 to 6	Reduction in MAE
50	5	200	0.254	0.046	81.9%	0.304	0.130	57.3%
50	10	200	0.128	0.031	75.4%	0.150	0.058	61.6%
50	25	200	0.032	0.016	51.6%	0.059	0.023	61.9%
50	5	1,000	0.030	0.016	44.2%	0.316	0.119	62.2%
50	10	1,000	0.014	0.008	42.3%	0.161	0.054	66.2%
50	25	1,000	0.011	0.006	45.3%	0.063	0.020	67.5%
50	5	4,000	0.020	0.009	54.9%	0.298	0.118	60.2%
50	10	4,000	0.016	0.006	66.4%	0.152	0.054	64.4%
50	25	4,000	0.009	0.004	52.7%	0.059	0.021	65.1%
500	5	200	0.058	0.016	72.9%	0.447	0.118	73.5%
500	10	200	0.066	0.013	80.3%	0.246	0.055	77.8%
500	50	200	0.033	0.006	83.1%	0.056	0.011	80.7%
500	5	1,000	0.029	0.012	58.5%	0.432	0.121	72.0%
500	10	1,000	0.036	0.008	77.9%	0.230	0.054	76.4%
500	50	1,000	0.006	0.005	26.9%	0.057	0.011	80.9%
500	5	4,000	0.017	0.005	69.4%	0.421	0.118	71.9%
500	10	4,000	0.011	0.004	65.5%	0.227	0.055	75.9%
500	50	4,000	0.002	0.002	-14.2%	0.056	0.010	81.9%
2,000	10	200	0.046	0.021	55.0%	0.242	0.055	77.1%
2,000	20	200	0.036	0.018	51.7%	0.141	0.027	80.8%
2,000	50	200	0.022	0.011	49.7%	0.066	0.010	84.4%
2,000	10	1,000	0.025	0.009	66.4%	0.213	0.054	74.6%
2,000	20	1,000	0.013	0.009	28.9%	0.118	0.026	77.7%
2,000	50	1,000	0.011	0.006	48.4%	0.055	0.010	81.2%

Table 1 and Figure 2 show MAEs across iterations standard errors on parameter estimates (relative to the full-choice-set estimates). Unlike Figure 1's MAEs of actual parameter values, there is a very strong trend across all data sets, regardless of *N*, *J* and *S*. The MAEs of standard errors fall by a striking 60% to 80% (see Table 1) across all data sets from iteration 1's SRS approach to iteration 2+'s strategic-sampling approach. These estimates are so stable that there is almost no variation in reduction between iterations 2 through 6, strongly supporting the recommendation that analysts need only perform two iterations to obtain the most efficient estimates, at least for the data sets simulated here. As one might expect, MAE reductions rise when more total alternatives (*J*) exist (since SRS have a greater likelihood of providing poor/uncompetitive samples) and as the ratio of *S* to *J* rises. This suggests that when a sampling of alternatives is most needed (i.e., when a very large number of alternatives exists), the benefit of sampling more intelligently is at its greatest.

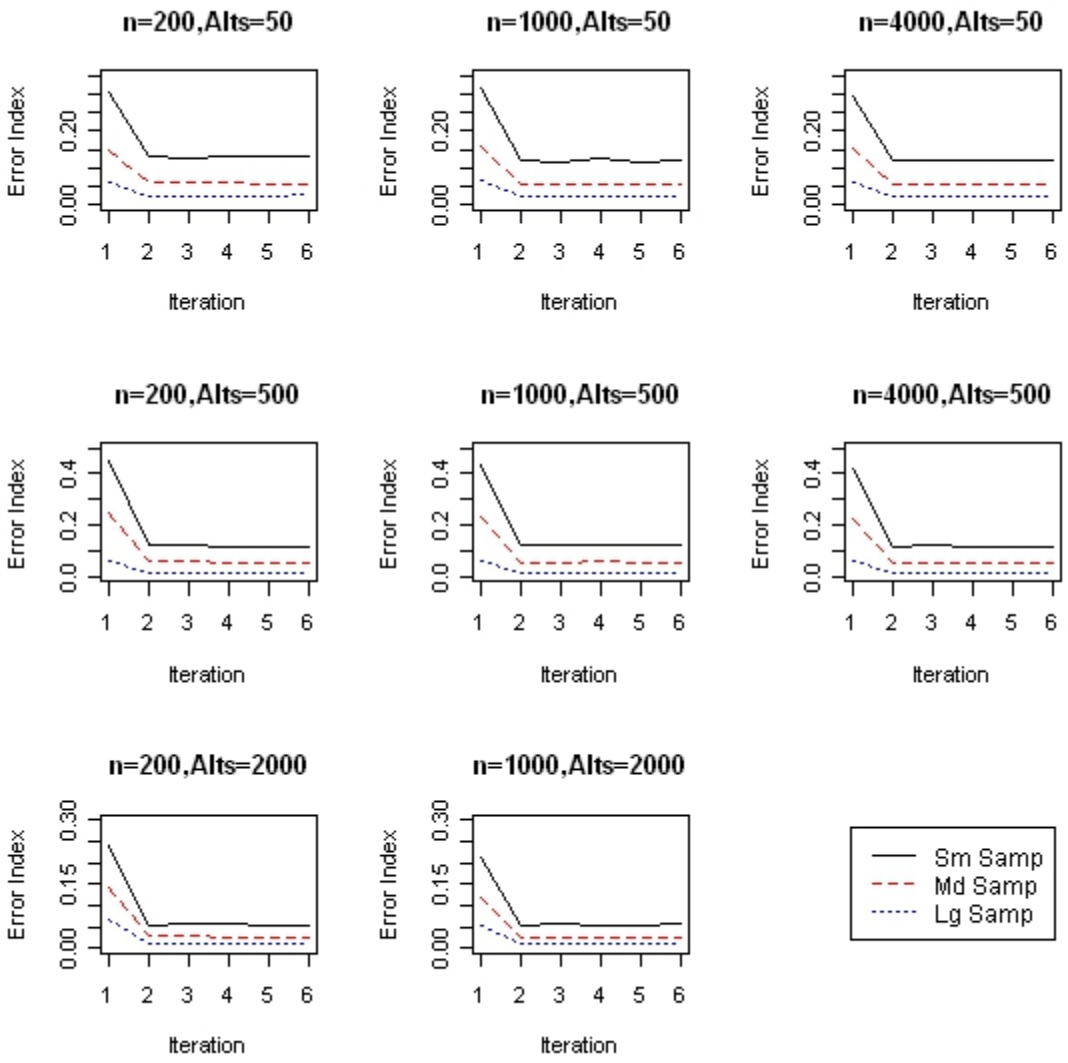

Figure 2: MAEs of Coefficient Standard Errors for the MNL Model

Another striking feature of Table 1 is that the magnitude of the standard errors' MAEs for iterations 2 through 6 is about the same for a fixed number of sampled alternatives (S), regardless of N and J. For instance, all data sets using 10 sampled alternatives have standard error MAEs for iterations 2 through 6 of approximately 0.055. This is likely a product of the way in which data was generated: The first 50 alternatives in each data set were simulated to be more appealing/competitive, on average; so, in each simulation, those are the alternatives most likely to be included in the sampled set of alternatives. Thus, this result is probably specific to the simulation/data-generating process used. Other results may be somewhat assumption-specific as well, so more choice settings may be useful.

Mixed Multinomial Logit Results

For the MMNL model, data sets with 200 and 1,000 observational units were used, each with $J = 500$ total alternatives. Models were estimated using small ($S = 5$), medium ($S = 10$), and large ($S = 50$) alternative-sample sizes.

Figure 3 shows the parameter-estimation results of the MMNL simulations. Figure 3's top row of plots represents the $N = 200$ data set, while the second row presents the $N = 1{,}000$ results. The three plots in

each row correspond to the two fixed coefficients, the two random-coefficient means, and the two random-coefficient variances. Within each plot, the MAEs for the small, medium, and large *S* values for each of the six sampling scheme iterations are displayed.

Like the MNL simulation results, the MMNL simulation results show, on average, MAE reductions in parameter values after the first iteration. This is most evident for the fixed parameters (first column of plots). In addition, the fixed parameters' errors do not appear to improve after the second iteration.

Results for the two random parameter means and variances, however, are a little less clear. When only 5 alternatives are sampled to estimate the models (Fig. 3's solid black lines), errors do not improve much from iteration 1 to the other iterations. However, when 10 and 50 alternatives are sampled (Fig. 3's dashed lines), MAEs typically do fall over the second through sixth iteration. One possible explanation for this is that 5 alternatives (a 1% sample) is simply not enough to obtain reliable parameter estimates from the MMNL model, since the estimation must use an approximation of the true likelihood. Of course, almost no analysts would use such a small sample (though many will use a 1% sample, if J=2000+). Because this work required multiple estimations (60 in total) for each *S*, the MMNL computing times become an issue; with more time and computing resources, one can explore trends beyond $S = 50$. Lastly, as with the fixed parameters, there is no distinguishable improvement in the MAEs beyond the second iteration of (i.e., beyond the first iteration of strategic sampling), which again suggests that just two iterations of the proposed strategic sampling approach will offer as much improvement in estimates as one can achieve.

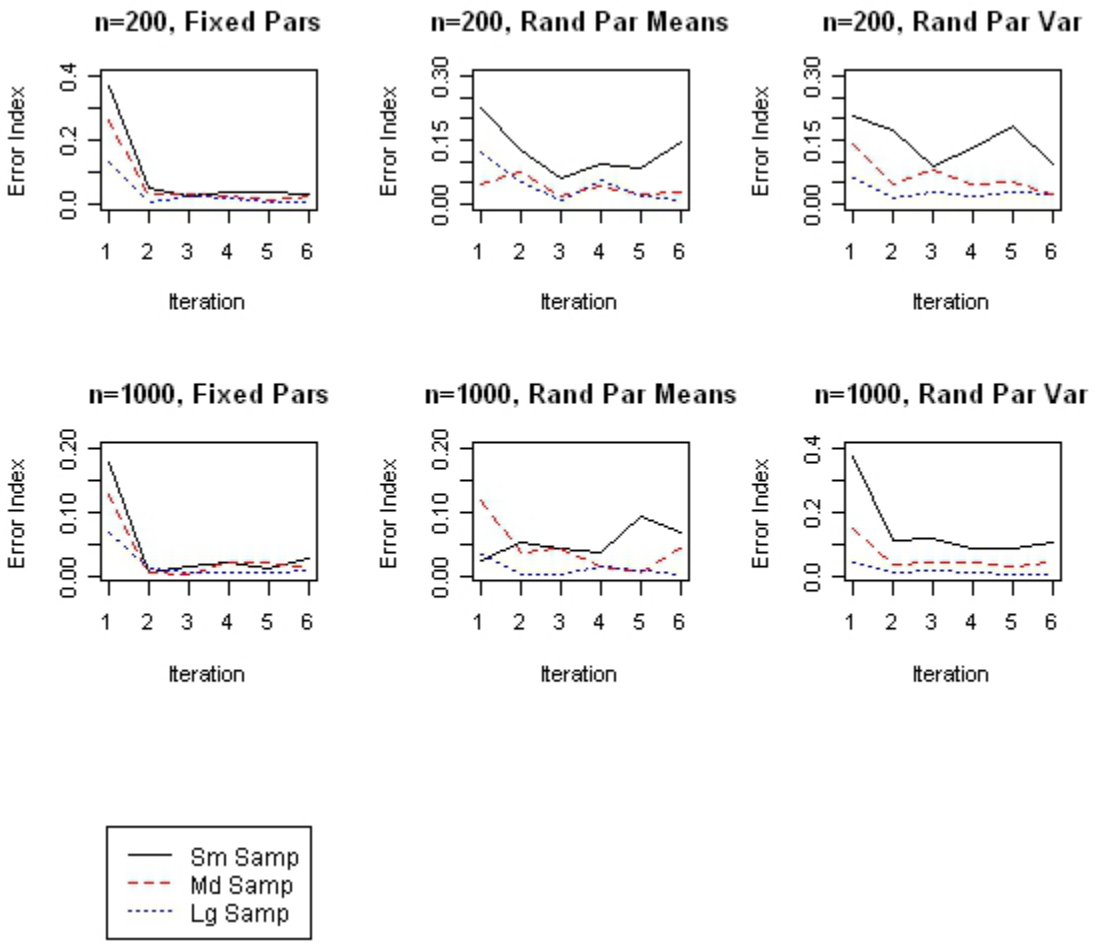

Figure 3: MAEs of Parameter Estimates for Mixed MNL Model

Table 2 lists the MAE reductions across iterations. As Figure 3 makes quite evident, MAE reductions for the fixed-parameters case are very large, reducing error by about 90% in every case. The MAE reductions are a bit more varied in the random-parameter (RP) cases. When only five alternatives are sampled in estimation, this is likely due (in part) to set-size, as discussed above. However, when 50 alternatives are sampled (representing 10% of the total number of alternatives), MAE reductions for mean and variance parameters are all in the range of 60 to 80%. This is substantial, especially in comparison to the MAE reductions found for the comparable MNL cases (with 50 sampled alternatives out of 500 total), which ranged from 83% in the 200-observation case to only 27% in the 1,000-observation case and -14%[6] in the 4,000-observation case. Since sampling alternatives for MMNL estimation does not produce consistent results, in theory and in the limit (see, e.g., Nerella and Bhat 2004), perhaps the estimation bias is smaller under the proposed form of strategic sampling, as opposed to SRS. Essentially, strategic sampling appears to offer far more benefits in MMNL estimation than any loss of consistency.

[6] The -14% is surely an aberration due to simulation error, but the trend across values of *N* is telling.

Table 2: Mixed MNL Model Estimates' MAE Reductions, Iteration 1 vs. Iterations 2 to 6

Total Alts. (*J*)	Samp. Alts. (*S*)	#Obs. (*N*)	Model Coefficient Estimates			Model Std. Error Estimates		
			MAE Redxn., Fixed Param.	MAE Redxn., RP Mean	MAE Redxn., RP Variance	MAE Redxn., Fixed Param.	MAE Redxn., RP Mean	MAE Redxn., RP Variance
500	5	200	90.1%	54.2%	35.1%	44.6%	61.2%	56.1%
500	10	200	90.4%	22.4%	63.7%	56.8%	67.6%	61.1%
500	50	200	89.3%	75.8%	64.2%	77.2%	83.8%	76.1%
500	5	1,000	91.0%	-132.0%	73.3%	45.5%	58.9%	54.0%
500	10	1,000	90.0%	75.4%	74.6%	60.1%	70.5%	63.8%
500	50	1,000	89.2%	82.5%	77.9%	77.1%	83.9%	78.7%

To understand the bias in parameter estimates better, a new error measure was constructed. This new measure will be called the mean error index (or MEI), and is equal to the average parameter estimate over the 10 simulations, divided by the full-choice-set parameter estimate (i.e., when using all alternatives in model estimation). Thus, if the error index is 1.0, the parameter estimate using choice-set sampling equals the full-set parameter estimate. The only difference between MEI and MAE is MEI does not take the absolute value of error, and is indexed around 1.0, rather than 0.

Figure 4 shows how this error measure varies by parameter for the cases with $S = 10$ and 50. As shown, the fixed-parameter estimates are consistently biased high when SRS is used (i.e., after one iteration). However, in each subsequent iteration of the new, strategic sampling scheme, the MEIs oscillate around 1.0, very near the target value. This indicates that SRS for the mixed MNL will result in biased parameter estimates for fixed model parameters, just as one would expect (since they are statistically inconsistent in theory); however, the new sampling scheme effectively reduces this fixed-parameters-estimator bias to near zero, which was not so expected.

In the case of the random-parameter means, it again appears that SRS produces estimates that are biased high, though in one case of the first mean parameter (with $N = 200$ and $S = 50$), the average MEI was quite low (at only 0.8 of the target). Nonetheless, in all other instances, random-parameter means were overestimated, when compared to the true parameter values. As with the fixed parameters, the bias is reduced substantially in iterations 2 through 6, thanks to the simple yet strategic sampling technique recommended here. Similar results are found in the case of random-parameter variance estimates. SRS consistently produces estimates that are biased low, while our strategic sampling technique consistently reduces that bias -- though variance parameter estimates exhibit the greatest amount of variation, as compared to mean parameters and fixed parameters. Thus, the simulation exercise not only suggests how estimator efficiency can be improved, but that results effectively carry over from standard MNL to the far more general case of MMNL.

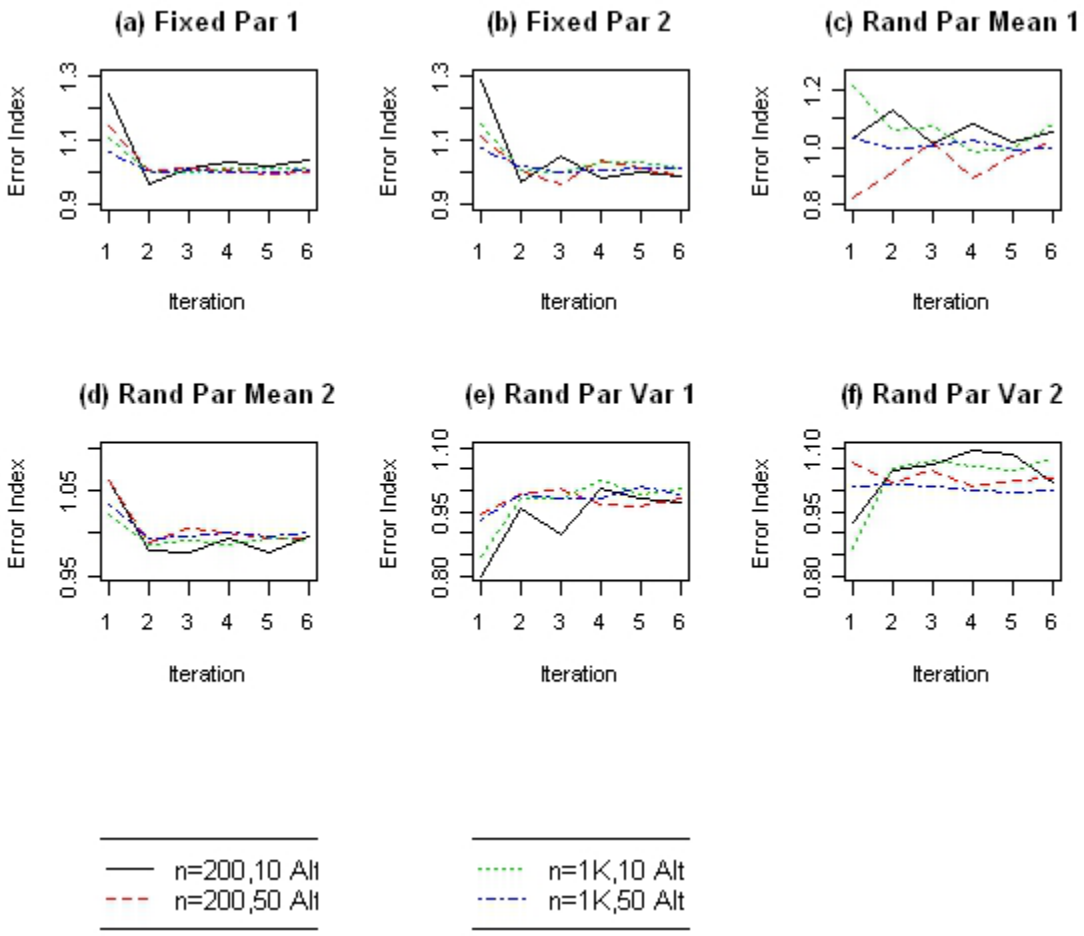

Figure 4: Mean Error Indices (MEIs) of Parameter Estimates for Mixed MNL Model

Figure 5 details MAEs of the MMNL estimators' standard errors. Results are very similar to those of the standard MNL, in that standard errors resulting from SRS (iteration 1) are quite high, and can be reduced substantially using the new sampling method, even for just a single subsequent iteration. Improvement in the standard errors really does not improve beyond the second iteration. Furthermore (and as detailed in the last three columns of Table 2), the reduction in standard error MAEs from iteration 1 to iterations 2 through 6 for the mixed MNL model are magnified as the number of sampled alternatives increases (from 45-60% reductions with a 1% alternative sample size to 75-85% reductions with a 10% alternative sample size).

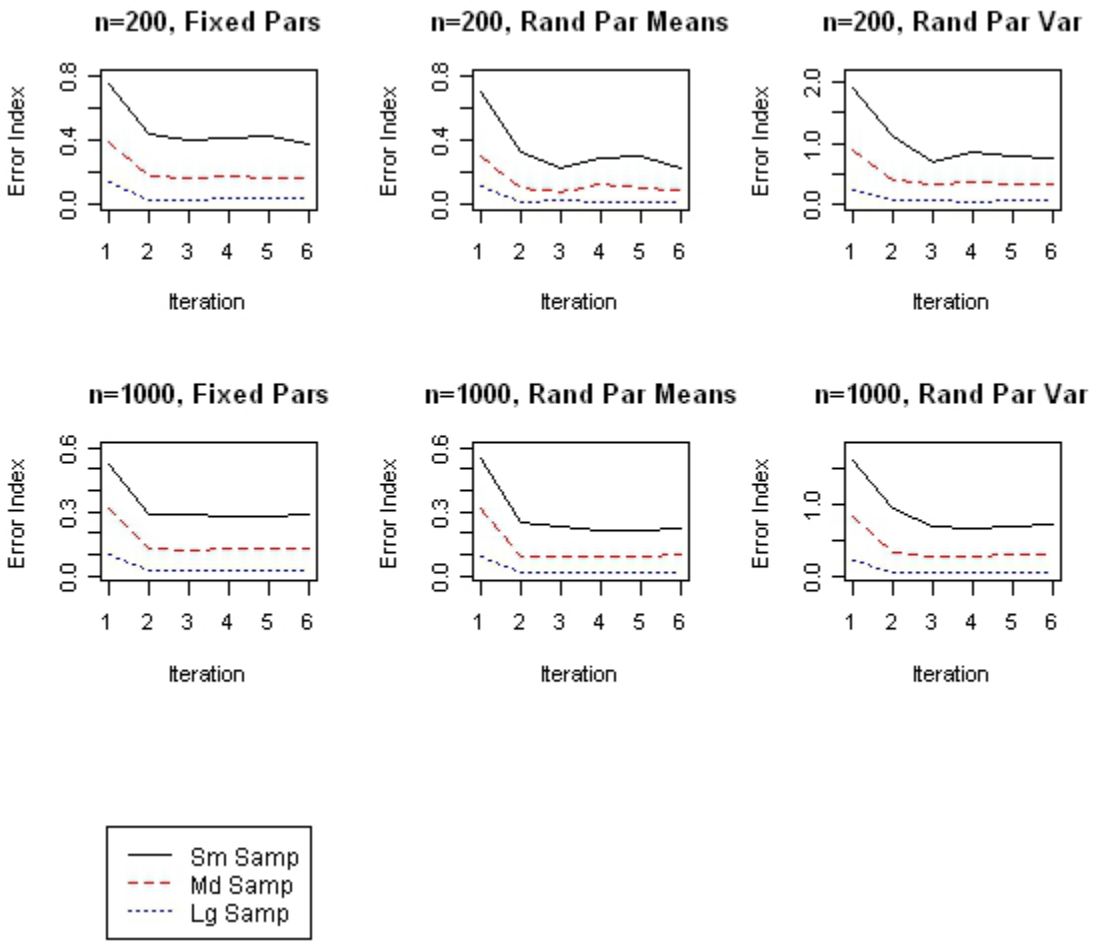

Figure 5: MAEs of Coefficient Standard Errors for Mixed MNL Model (Two Fixed Parameters, Two Random Parameter Means, and Two Random Parameter Variances)

Varying Model Noise

A final simulation exercise examined model noise, or error-term scale, for the MNL. This was done be changing the values of the model's simulated coefficients. A "low noise" scenario was constructed by doubling the parameter values (from 0.5, 0.3, 0.1, and -1.0 to 1.0, 0.6, 0.2, and -2.0). A "high noise" scenario was constructed by halving the simulated parameter values (to 0.25, 0.15, 0.05, and -0.5). By increasing the simulated values in the "low noise" scenario while holding the model's error term variance constant, there is more variation in the simulated systematic utility values. This will improve the fit of the estimated model, increasing the precision of the parameter estimates. Similarly, the "high noise" scenario will reduce the precision of parameter estimates.

Table 3 shows standard error MAE reductions for these two scenarios along with the original results in the 500-alternative cases and 5, 10, or 50 sampled alternatives. As shown, the size of the error terms relative to systematic model components has very important consequences for improvements achieved via strategic sampling. When the error term is large relative to systematic components (high noise), the MAE reduction is smaller (ranging from about 39% to 56%), and when the error term is small, relative to systematic components (low noise), the MAE reduction is much larger (ranging from 91% to 97%). In the high-noise case, there is less difference between a more competitive alternative and a non-competitive alternative. Since an alternative's competitiveness is the basis for strategic sampling, the reduction in

MAE will be lower. However, when uncertainty, unobserved heterogeneity, and thus data noise are reduced, the strategic sampling scheme is more capable of generating a sample of competitive alternatives, which offer more information to the model than non-competitive ones.

Table 3: Standard Error MAE Reductions for Low-Noise and High-Noise MNL Simulations

Total Alts. (*J*)	Samp. Alts. (*S*)	#Obs. (*N*)	Measure	Low Noise	Medium Noise	High Noise
500	5	1,000	MAE, Iter. 1	1.445	0.432	0.197
			MAE, Iter. 2 to 6	0.119	0.121	0.121
			MAE Reduction (%)	91.7%	72.0%	38.8%
500	10	1,000	MAE, Iter. 1	0.904	0.230	0.092
			MAE, Iter. 2 to 6	0.054	0.054	0.056
			MAE Reduction (%)	94.0%	76.4%	39.5%
500	50	1,000	MAE, Iter. 1	0.301	0.057	0.019
			MAE, Iter. 2 to 6	0.010	0.011	0.011
			MAE Reduction (%)	96.5%	80.9%	43.2%

Interestingly, the main driver of Table 3's MAE reductions comes from the higher starting MAE values evident in SRS sampling under the low-noise settings. There is no discernible difference in MAE values when applying strategic sampling across the three noise-level settings. The SRS MAEs are simply much higher when there is less data noise (i.e., relatively lower variance in associated Gumbel error terms). This suggests that uncertainty in the choice behavior will not affect the precision of parameter estimates (relative to estimation with all alternatives) when strategic sampling is used.

The simulation results presented above consistently highlight the potential benefits of the proposed strategic sampling approach. In both the standard and mixed MNL models, standard errors of estimates can be reduced by as much as 95%, with little effort and expertise. This leads to more accurate parameter estimates than SRS can provide, particularly in the case of the MMNL where parameter estimates will be biased.

Such results are echoed by applications of the strategic sampling scheme with actual data, in a destination-choice context, for more than 1,000 alternatives. Due to space limitations and an inability to control for all the variables a synthetic data set allows, the results of these destination-choice data are not presented here, but they are similar to the results obtained for the simulated data with medium noise. The two-step approach appears maximally efficient and the efficiency gains are substantial. The sampling scheme should work for real, imperfect data environments, just as it does for simulated data.

DISCUSSION

The results (using both synthetic and actual data) consistently indicate that the strategic sampling method proposed in this paper offers great value. Compared to a SRS sample, strategic sampling can reduce estimator errors by 40 to 95%, depending on sample size, number of alternatives, number of sampled alternatives, and noise or uncertainty levels. However, strategic sampling is more computationally expensive than SRS sampling. One wonders what the tradeoffs are.

In the case of the MNL (excluding tests with only 50 alternatives, since computing times are very low in these cases anyway), strategic-sampling computation times varied between 3% (when 0.5% of alternatives were sampled) and 30% (when 10% of alternatives were sampled) of the time required for full estimation without sampling. Computation times (as a percentage of full estimation time) were fairly

consistent across scenarios, regardless of the total number of alternatives and sample size. In the case of the MMNL, computation times varied from 2% to 15% of full estimation time when 1% and 10% of alternatives were sampled. Unlike with the MNL setting, strategic sampling times reduced MMNL estimation times further (relative to full estimation) when sample sizes increased.

Computing time also depends on the number of explanatory variables used. In all exercises presented here, four explanatory variables were used. A couple additional MNL simulations -- with 8 and 20 explanatory variables -- were tested, to obtain a better understanding of how this feature can affect computing time. Under this setup, computing times for estimation with complete choice sets (i.e., without sampling) rose approximately three times for a doubling of the number of explanatory variables and 13 times when variables were increased by 5 times (i.e., 4 to 20). In contrast, computing times under strategic sampling increased only by the share of added variables (e.g., times doubled when variable count doubled).

Given these experiences, it is difficult to offer a single rule-of-thumb as to when strategic sampling should be used. Though they never examined cases with more than 200 alternatives, Nerella and Bhat (2004) suggested minimum choice-set SRS sample sizes of one-eighth and one-quarter for the MNL and MMNL setups, respectively. Their recommended shares are one quarter and one half, of all alternatives. Here, in the case of the 500-alternatives MNL case, a 5% sample would likely achieve standard errors within 2% of their true values, and such precision could probably be achieved with just a 2% sample in the 2,000-alternative case. It seems to reason that, as the choice set size increases, one can strategically sample – following the methods proposed here -- at much lower rates that researchers have suggested in the past, while maintaining high levels of precision.

Based on the computing times associated with strategic sampling (versus full choice set use), the method's greatest benefits clearly emerge for very large choice sets (e.g., over 1,000 alternatives). In such cases, an analyst can thoughtfully sample at a low rate, obtain very precise parameter estimates, and gain enormous computational savings (which are further amplified with large numbers of explanatory variables).

For the MMNL, a larger sample size (of alternatives) is really needed. Even when 10% of the choice set was sampled here, model-estimated standard errors still had a 5% margin of error for fixed and mean parameters and over 10% margin of error for variance parameter estimates. Thus, a formal recommendation is difficult to make, but Nerella and Bhat's (2004) minimum bound of 25% sampled alternatives could represent a reasonable choice. Even then, one can achieve noticeable computational benefits by sampling strategically. Further analysis would be helpful to draw additional conclusions on how estimators' precision varies with an MMNL's total choice set size.

CONCLUSIONS

In many transportation choice contexts -- like destination/location, trip pattern, and route choices, choice sets can be quite large. In such cases, employing the complete set of choice alternatives in model estimation can be computationally burdensome and sometimes prohibitive, even for the relatively straightforward MNL specification.[7] Choice-set sampling scales back this burden substantially, but has only been proven to generate consistent estimators for certain model conditions (like the MNL). SRS is most often used, since it meets McFadden's (1978) conditioning property, and thus requires no modification of the likelihood function. In this paper, a strategic sampling scheme has been introduced, specifically to generate more efficient parameter estimates.

[7] In a Bayesian estimation setting, where one typically simulates random utility values (for every alternative), computation with large choice sets is particularly challenging due to added simulation randomness.

A variety of simulated data sets (as well as an actual destination-choice data set) were used to examine the effectiveness of such strategic sampling in cases where it is theoretically consistent (i.e., the standard MNL) and in cases where it is not (the mixed MNL). The empirical evidence strongly indicates that the new approach provides more efficient estimates than its SRS counterpart. Thanks to the improved efficiency, the accuracy of estimates is also typically improved, even in the case of the mixed MNL where SRS of choice alternatives was found to produce biased estimates. The new sampling scheme virtually eliminated bias in these mixed MNL estimates, while also improving estimator accuracy and precision.

Moreover, compared to SRS sampling, the new strategy requires very little additional computing time or analyst capability. A single added iteration of model estimation was found sufficient, offering the maximum benefits of the new sampling scheme. Moreover, the greatest benefits of the proposed strategic sampling scheme can be found in the largest choice set contexts (e.g., over 1,000 alternatives), when one can sample at a low rate, providing tremendous computational savings while maintaining precise model estimates.

Of course, all of these results are specific to the data sets examined here. Drawing universal conclusions would be unwise at this point, and future work would be helpful. For instance, it would be worthwhile to examine how the strategic sampling technique performs when the model specification is imperfect. It would be useful to apply the model to additional real-world data sets. In the case of the MMNL, non-normal parameter distributions should be investigated.

To summarize, the new, strategic sampling strategy proposed and tested here provides data analysts and researchers a new tool to more precisely and robustly characterize choice behaviors in contexts where the numbers of choice alternatives are very large. While it is always ideal to consider all choice-set alternatives when estimating discrete-choice models, there are many instances where such calculations are prohibitive. The strategic sampling approach presented in this paper offers an attractive substitute to the uniform sampling (SRS) technique used almost exclusively in practice and research to date.

REFERENCES

Ben-Akiva, M. and T. Watanatada (1981) Application of a Continuous Spatial Choice Logit Model. In *Structural Analysis of Discrete Data and Econometric Applications* (C.F. Manski and D. McFadden, eds.), MIT Press, Cambridge, MA.

Ben-Akiva, M. and S.R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Ben-Akiva, M. and J.L. Bowman (1998) Integration of an Activity-Based Model System and a Residential Location Model. *Urban Studies*, 35 (7), 1131-1153.

Bhat, C., A. Govindarajan, and V. Pulugurtha (1998) Disaggregate Attraction End Choice Modeling: Formulation and Empirical Analysis. *Transportation Research Record*, 1645, 60-68.

Bierlaire, M., D. Bolduc, and D. McFadden (2008) The Estimation of Generalized Extreme Value Models from Choice-Based Samples. *Transportation Research Part B*, 42 (4), 381-394.

Frejinger, E., M. Bierlaire, and M. Ben-Akiva (2009) Sampling of Alternatives for Route Choice Modeling. *Transportation Research Part B*, 43, 984-994.

Lemp, J.D., L.B. McWethy, and K.M. Kockelman (2007) From Aggregate Methods to Microsimulation: Assessing Benefits of Microscopic Activity-Based Models of Travel Demand.. *Transportation Research Record*, 1994, 80-88.

McFadden, D. (1978) Modeling the Choice of Residential Location. In *Spatial Interaction Theory and Residential Location* (A. Karlqvist, L. Lundqvist, F. Snickbars, J. Weibull, eds.), North-Holland, Amsterdam, 75-96.

McFadden, D. and K. Train (2000) Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, 15 (5), 447-470.

Nerella, S. and C.R. Bhat (2004) Numerical Analysis of Effect of Sampling of Alternatives in Discrete Choice Models. *Transportation Research Record*, 1894, 11-19.

Pozsgay, M.A. and C.R. Bhat (2001) Destination Choice Modeling for Home-Based Recreational Trips: Analysis and Implications for Land Use, Transportation, and Air Quality Planning. *Transportation Research Record*, 1777, 47-54.

Sermons, M.W. and F.S. Koppelman (2001) Representing Differences between Female and Male Commute Behavior in Residential Location Choice Models. *Journal of Transport Geography*, 9 (2), 101-110.

Train, K. (2009) *Discrete Choice Methods with Simulation, 2nd Edition*. Cambridge University Press, New York, NY.

Vovsha, P. (1997) Application of Cross-Nested Logit Model to Mode Choice in Tel Aviv, Israel, Metropolitan Area. *Transportation Research Record*, 1607, 6-15.

Wen, C.H. and F.S. Koppelman (2001) The Generalized Nested Logit Model. *Transportation Research Part B, 35*, 627-641.

Williams, H.C.W.L. (1977) On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit. *Environment and Planning A*, 9, 285-344.